



*Estratégia baseada em Realidade Aumentada e  
LLM para Tratamento de Dados Ausentes:  
estudo de caso usando Unity e ChatGPT*  
***Enio Vicente de Limas***  
março / 2024

Dissertação de Mestrado em Ciência da  
Computação

***Estratégia baseada em Realidade Aumentada e LLM para Tratamento de Dados Ausentes: estudo de caso usando Unity e ChatGPT***

Dissertação apresentada ao Programa de Mestrado em Ciência da Computação da UNIFACCAMP – Centro Universitário Campo Limpo Paulista.

Campo Limpo Paulista, 24 de março de 2024.

Enio Vicente de Limas

Prof. Dr. Marcelo de Paiva Guimarães (Orientador)

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) - Código de Financiamento 001.

## FICHA CATALOGRÁFICA

Ficha catalográfica elaborada pela  
Biblioteca Central da Unifaccamp

L696e

Limas, Enio Vicente de

Estratégia baseada em realidade aumentada e LLM para tratamento de dados ausentes: estudo de caso usando *Unity* e *ChatGPT* / Enio Vicente de Limas. Campo Limpo Paulista, SP: Unifaccamp, 2024.  
100 p.: il.

Orientador: Prof. Dr. Marcelo de Paiva Guimarães

Dissertação (Programa de Mestrado Profissional em Ciência da Computação) – Centro Universitário Campo Limpo Paulista – Unifaccamp.

1. Realidade aumentada. 2. Dados ausentes. 3. Imputação de dados. 4. Tratamento de dados. 5. Modelos de linguagens grandes. I. Guimarães, Marcelo de Paiva. II. Centro Universitário Campo Limpo Paulista. III. Título.

CDD – 001.640

## **Resumo**

*Por mais planejado que seja uma base de dados, é comum a presença de dados ausentes. Quando não é possível a recuperação dos mesmos e deseja-se utilizar esses dados, é importante encontrar uma forma de tratá-los, isso significa que é necessário preencher os valores ausentes utilizando algum método, ou até mesmo eliminar os registros ou colunas com dados ausentes. Dados ausentes comprometem a análise de um conjunto específico, impactando negativamente a qualidade e a confiabilidade dos resultados. O foco principal deste trabalho é introduzir uma estratégia voltada para a facilitação do tratamento de dados ausentes, enfatizando a visualização de dados. Essa abordagem inclui a utilização de visualizações em 3D com Realidade Aumentada, além do tratamento de dados propriamente dito. Ambos os aspectos são apoiados pela geração de códigos-fonte através do ChatGPT, uma ferramenta baseada em Modelos de Linguagem Grande (LLM). Visualizações de dados com Realidade Aumentada, proporcionam uma compreensão mais clara e intuitiva das características dos dados ausentes, permitindo, assim, uma escolha informada e eficaz dos métodos de tratamento mais adequados. A estratégia foi desenvolvida principalmente para ajudar Analistas de Dados. Para concretizar esta estratégia, desenvolvemos o aplicativo DMT-AR (Data Missing Treatment with Augmented Reality). Ao utilizar o DMT-AR os usuários podem explorar visualmente os dados, mensurando anomalias como, por exemplo, dados ausentes. Os botões virtuais integrados ao aplicativo permitem aos usuários aplicar diferentes métodos de tratamento de dados no mesmo conjunto de dados. Nossa investigação conclui que a abordagem imersiva e interativa adotada para a exploração e compreensão de conjuntos de dados facilita uma análise profunda e eficaz na detecção e tratamento de anomalias. Contudo, identificamos um desafio associado à diminuição do desempenho e complexidade das visualizações conforme a escala dos conjuntos de dados aumenta.*

*“Ensinar não é transferir conhecimentos, mas criar as possibilidades para a sua própria produção ou construção.” (Paulo Freire)*

**Palavras-chave:** *Realidade Aumentada, Dados Ausentes, Imputação de Dados, Tratamento de Dados, Modelos de Linguagens Grandes.*

## ***Abstract***

*As planned as a database might be, the presence of missing data is common. When it is not possible to recover these data and there is a desire to utilize them, it is important to find a way to handle them. This means filling in missing values using some method, or even eliminating records or columns with missing data. Missing data compromises the analysis of a specific dataset, negatively impacting the quality and reliability of the results. The main focus of this work is to introduce a strategy aimed at facilitating the treatment of missing data, with an emphasis on data visualization. This approach includes the use of 3D visualizations with Augmented Reality, in addition to the actual data treatment. Both aspects are supported by the generation of source code through ChatGPT, a tool based on Large Language Models (LLMs). Data visualizations with Augmented Reality provide a clearer and more intuitive understanding of the characteristics of missing data, thus allowing for an informed and effective choice of the most suitable treatment methods. The strategy was developed primarily to assist Data Analysts. To implement this strategy, we developed the DMT-AR (Data Missing Treatment with Augmented Reality) application. By using DMT-AR, users can visually explore data, measuring anomalies such as missing data. The virtual buttons integrated into the application allow users to apply different data treatment methods to the same dataset. Our research concludes that the immersive and interactive approach adopted for the exploration and understanding of datasets facilitates deep and effective analysis in detecting and treating anomalies. However, we identified a challenge associated with the decrease in performance and complexity of visualizations as the scale of datasets increases.*

*To teach is not to transfer knowledge, but to create the possibilities for its own production or construction.*

*(Paulo Freire)*

***Keywords:*** *Augmented Reality, Data Missing , Data Imputation, Data Treatment, Large Language Models.*

## **Dedicatória**

Eu dedico esta dissertação inicialmente a Deus, que constantemente me dá força durante os períodos desafiadores. À minha esposa, família e aos professores que estiveram ao meu lado e me forneceram apoio e estabilidade durante este período de imersão nos estudos.

## **Agradecimentos**

Agradeço primeiramente a Deus por me dar discernimento, me sustentar e conduzir para a realização deste trabalho. Ao meu orientador professor Dr. Marcelo de Paiva Guimarães pela dedicação, paciência, compreensão, insistência, motivação e ensinamentos durante todo o processo de desenvolvimento deste trabalho, um verdadeiro professor, amigo e orientador, obrigado por ajudar a melhorar sempre, agradecimentos eternos. À minha esposa, pelo sustento durante todo o processo de estudos até aqui, pela perseverança, motivação que me trouxeram, minha eterna gratidão e amor a você. Aos professores e funcionários do programa de mestrado em Ciência da Computação do Centro Universitário Campo Limpo Paulista. A todos que contribuíram para a realização deste trabalho.

“Na adversidade, uns desistem,  
enquanto outros batem recordes.”

(Ayrton Senna)



## Sumário

<b>1. INTRODUÇÃO .....</b>	<b>01</b>
<b>1.1. Contexto e Motivação .....</b>	<b>02</b>
<b>1.2. Problemática e Justificativa .....</b>	<b>03</b>
<b>1.3. Objetivo, Contribuições e Métodos .....</b>	<b>03</b>
<b>1.4. Objetivo Específico .....</b>	<b>04</b>
<b>1.5. Estrutura da Proposta .....</b>	<b>04</b>
<b>2. REFERENCIAL TEÓRICO E METODOLÓGICO .....</b>	<b>05</b>
<b>2.1. Tratamento de Dados Ausentes .....</b>	<b>05</b>
<b>2.1.1. Métodos para tratamento de dados ausentes .....</b>	<b>06</b>
2.1.1.1 Método de exclusão de registros .....	07
2.1.1.2 Método de imputação única .....	08
2.1.1.3 Método de imputação múltipla .....	12
<b>2.2. Realidade Aumentada .....</b>	<b>13</b>
<b>2.3. Processamento de Linguagem Natural .....</b>	<b>15</b>
<b>2.3.1. Uso de uma ferramenta <i>LLM</i> para geração de código-fonte .....</b>	<b>16</b>
<b>2.3.2. Uso de uma ferramenta <i>LLM</i> para gerar visualizações ....</b>	<b>19</b>
<b>3. REVISÃO DA LITERATURA .....</b>	<b>23</b>
<b>3.1. Resultados da Revisão da Literatura .....</b>	<b>29</b>
<b>3.2. Trabalhos Relacionados .....</b>	<b>40</b>
<b>3.2.1. Análise comparativa dos trabalhos relacionados .....</b>	<b>44</b>
<b>3.3. Conclusão da Revisão da Literatura .....</b>	<b>45</b>

<b>4. ESTRATÉGIA PARA O TRATAMENTO DOS DADOS AUSENTES</b>	<b>46</b>
4.1. Introdução à estratégia .....	46
4.2. Visão geral da estratégia .....	46
4.2.1. Subciclo Análise de Dados .....	49
4.2.2. Subciclo Tratamento dos Dados .....	50
<b>5. APLICATIVO DE REALIDADE AUMENTADA</b> .....	<b>52</b>
5.1. Funcionamento do Aplicativo .....	53
5.1.1. Módulo de Análise dos Dados .....	53
5.1.2. Módulo de Tratamento de Dados .....	55
5.2. Arquitetura do <i>DMT-AR</i> .....	58
5.3. O <i>ChatGPT</i> e a Geração de Códigos-Fonte .....	64
5.3.1. Padrões de <i>prompts</i> aplicados .....	65
5.3.2. Conclusão da colaboração do <i>ChatGPT</i> .....	77
<b>6. ESTUDO DE CASO DO DMT-AR</b> .....	<b>78</b>
<b>7. CONSIDERAÇÕES FINAIS</b> .....	<b>80</b>
<b>REFERÊNCIAS</b> .....	<b>83</b>
<b>APÊNDICE A: Arquivo C# que gera o gráfico virtual</b> .....	<b>96</b>
<b>APÊNDICE B: Arquivo C# responsável por transferir os dados para o gráfico virtual</b> .....	<b>97</b>
<b>APÊNDICE C: Arquivo C# responsável pelo tratamento dos dados e atualização da base de dados</b> .....	<b>99</b>

<b><i>APÊNDICE D: Arquivo PHP responsável pelos cálculos da moda</i></b>	<b>100</b>
<b><i>APÊNDICE E: Prompts enviados ao ChatGPT do estudo de caso</i></b>	<b>102</b>

## **Lista de siglas**

- AR - Augmented Reality*
- DMT-AR - Data Missing Treatment with Augmented Reality*
- DXR - Development Reality Extended*
- GPT3 - Generative Pre-trained Transformer 3*
- GUI - Graphical User Interface*
- HMD - Head Mounted Display*
- IEEE - Institute of Electrical and Electronics Engineers*
- IP - Internet Protocol*
- LLM - Large Language Model*
- MCAR - Missing Completely at Random*
- MAR - Missing Aleatório*
- ML - Machine Learning*
- MR - Mixed Reality*
- NMAR - Missing Not at Random*
- NLP - Natural Language Processing*
- SQL - Structured Query Language*
- SVR - Simpósio de Realidade Virtual e aumentada*
- UML - Unified Modeling Language*
- URL - Uniform Resource Locator*
- XML - Extensible Markup Language*

## Lista de Tabelas

<b>Tabela 1: <i>String</i> de busca, <i>url</i> e resultado na base de dados <i>IEEE/Xplore</i> .....</b>	<b>25</b>
<b>Tabela 2: <i>String</i> de busca, <i>url</i> e resultado na base de dados <i>ScienceDirect</i> .....</b>	<b>25</b>
<b>Tabela 3: Artigos selecionados nas bases de dados <i>IEEE/Xplore</i> e <i>ScienceDirect</i> .....</b>	<b>26</b>
<b>Tabela 4: Trabalhos relacionados .....</b>	<b>41</b>
<b>Tabela 5: Análise comparativa .....</b>	<b>44</b>
<b>Tabela 6: Documento de visão do <i>DMT-AR</i> .....</b>	<b>61</b>
<b>Tabela 7: Exemplo de informações estatísticas .....</b>	<b>67</b>
<b>Tabela 8: Exemplo de estatísticas para dados categóricos .....</b>	<b>68</b>
<b>Tabela 9: Exemplo para geração de tipo de dados .....</b>	<b>69</b>
<b>Tabela 10: Exemplo de verificação da existência de dados ausentes ..</b>	<b>69</b>
<b>Tabela 11: Verificando características dos dados ausentes .....</b>	<b>70</b>
<b>Tabela 12: Exemplo de geração de gráficos 3D.....</b>	<b>71</b>
<b>Tabela 13: Exemplos de técnicas de tratamento de dados ausentes ..</b>	<b>74</b>

## Lista de Figuras

<b>Figura 1: Gráfico de regressão com correlação positiva .....</b>	<b>10</b>
<b>Figura 2: Gráfico de regressão com correlação negativa .....</b>	<b>11</b>
<b>Figura 3: Gráfico de regressão para os conjuntos X e Y .....</b>	<b>12</b>
<b>Figura 4: Exemplo para contagem de dados ausentes .....</b>	<b>18</b>
<b>Figura 5: Exemplo para rastreamento de dados ausentes .....</b>	<b>19</b>
<b>Figura 6: Exemplo de geração de gráfico 2D .....</b>	<b>20</b>
<b>Figura 7: Exemplo de visualização através de gráfico 2D .....</b>	<b>21</b>
<b>Figura 8: Exemplo da geração de um gráfico 3D .....</b>	<b>22</b>
<b>Figura 9: Exemplo da visualização através de gráfico 3D .....</b>	<b>22</b>
<b>Figura 10: Etapas da revisão da literatura .....</b>	<b>24</b>
<b>Figura 11: Visão geral da estratégia .....</b>	<b>49</b>
<b>Figura 12: Subciclo Análise de dados .....</b>	<b>50</b>
<b>Figura 13: Subciclo Tratamento de dados .....</b>	<b>51</b>
<b>Figura 14: Módulo Análise de dados em execução .....</b>	<b>55</b>
<b>Figura 15: Módulo Tratamento de dados em execução .....</b>	<b>57</b>
<b>Figura 16: Gráfico virtual após o tratamento de dados .....</b>	<b>58</b>

<b>Figura 17: Diagrama de Caso de Uso para o <i>DMT-AR</i> .....</b>	<b>59</b>
<b>Figura 18: Diagrama de Sequência para o <i>DMT-AR</i> .....</b>	<b>60</b>
<b>Figura 19: Diagramas de Pacotes para o <i>DMT-AR</i> .....</b>	<b>63</b>
<b>Figura 20: Fluxograma do <i>DMT-AR</i> .....</b>	<b>64</b>
<b>Figura 21: Cabeçalho do arquivo csv de exemplo .....</b>	<b>66</b>
<b>Figura 22: Cálculos de estatísticas de dados numéricos .....</b>	<b>67</b>
<b>Figura 23: Exemplo de contagem de valores únicos .....</b>	<b>68</b>
<b>Figura 24: Exemplo de geração de tipo de dados .....</b>	<b>69</b>
<b>Figura 25: Exemplo de detecção de dados ausentes .....</b>	<b>70</b>
<b>Figura 26: Exemplo de verificação de tipos de dados ausentes .....</b>	<b>71</b>
<b>Figura 27: Código-fonte para geração de gráfico 3D .....</b>	<b>72</b>
<b>Figura 28: Visualização de dados através de gráfico 3D.....</b>	<b>73</b>
<b>Figura 29: Código-fonte para o tratamento de dados ausentes ....</b>	<b>75</b>
<b>Figura 30: Cabeçalho do arquivo csv com os dados tratados .....</b>	<b>76</b>
<b>Figura 31: Visualização parcial de dados ausentes na tabela exemplo</b>	<b>78</b>
<b>Figura 32: Visualização parcial de dados tratados na tabela exemplo</b>	<b>79</b>

## 1. Introdução

Conforme Schafer e Graham (2002), dados ausentes (*data missing*) significam dados que deveriam ter sido informados, mas, por algum motivo, isso não ocorreu. Isso pode acontecer devido a fatores humanos intencionais (não saber responder, recusar-se a responder, desconsideração, etc.) ou não intencionais (distração, defeitos técnicos, esquecimento, etc.). Para exemplificar um motivo intencional, suponhamos uma coleta de dados onde observou-se uma tendência significativa de recusa em responder à pergunta sobre renda familiar. Com relação a motivos não intencionais, podemos exemplificar um problema técnico na interface de um questionário online.

Por mais planejada que seja uma base de dados, é inevitável que surjam dados ausentes. Quando se torna necessário recuperá-los e isso se torna impossível, é crucial encontrar uma forma de tratá-los. Isso significa utilizar métodos para preencher os valores desses dados ausentes ou, se necessário, eliminar esses registros ou colunas (Liu e Hauswirth, 2020). Conforme observado por esses autores, a maioria dos métodos de tratamento de dados baseia-se em técnicas de imputação de dados (substituição dos valores ausentes). A necessidade do tratamento é crucial, pois dados ausentes comprometem a análise de um determinado conjunto de dados, afetando diretamente a qualidade e a confiança que são essenciais para a utilização deles, especialmente quando são empregados por algoritmos de aprendizagem de máquina que não são projetados para lidar com dados ausentes.

Diversos métodos de tratamento de dados ausentes já foram propostos (por exemplo: imputação única, imputação múltipla, árvores de imputação e baseados em padrões de semelhanças). Entretanto, ainda não há um método consolidado para tratar todos os tipos de dados ausentes (Liu e Hauswirth, 2020).

Alguns métodos, como por exemplo, imputações pela média ou moda, são muito eficientes em tempo de processamento (em comparação com métodos mais complexos como aprendizado de máquina), mas podem produzir um conjunto de dados pré-processados com características muito alteradas em relação ao conjunto de dados original, quando este tiver os



valores grandes de espalhamentos<sup>1</sup>. O método de imputação é dependente da base de dados e dos critérios estabelecidos (Othman e Yahia, 2018).

Este trabalho apresenta uma estratégia para apoiar o tratamento de dados ausentes, principalmente, para analista de dados, visto a complexidade técnica envolvida em todo o processo. Esta estratégia utiliza visualizações 3D com Realidade Aumentada (*Augmented Reality - AR*) e geração de códigos-fonte com auxílio de ferramentas de Modelos de Linguagens Grandes (*LLM - Large language model*). No entanto, é importante destacar que a integração entre *LLM* e *AR* não ocorre diretamente. Em vez disso, ocorre o uso de visualizações de dados apoiados pela *AR* e um suporte da *LLM* na geração de códigos-fonte, necessários para as manipulações dos dados e geração dos gráficos 3D.

Um *LLM*, ainda na sua forma bruta e pré-treinada antes de qualquer ajuste fino é uma rede neural com o objetivo de prever a próxima palavra (*token*) dada uma sequência de palavras (*tokens*) como contexto (Shanahan, McDonell e Reynolds, 2023).

A *AR* É um sistema que combina elementos virtuais com elementos reais, é interativo em tempo real e registra ou alinha objetos reais e virtuais em uma única cena, dando a sensação ao usuário que os objetos virtuais fazem parte do ambiente real (Bianchi *et al.*,2021).

A capacidade de uma *LLM* resolver tarefas para desenvolvedores de software, incluindo a geração de códigos-fonte e a habilidade de integrar objetos virtuais em ambientes reais, habilitada pela *AR*, são fundamentais em nossa estratégia.

## 1.1. Contexto e motivação

A existência de dados ausentes cria dificuldades nos resultados pretendidos, como baixa qualidade e/ou baixa confiabilidade, especialmente na análise de dados, uma vez que a maioria dos procedimentos não são projetados para lidar com dados ausentes, como por exemplo algoritmos de aprendizagem de máquina. Como consequência, uma análise de dados pode gerar resultados errôneos quando existem dados ausentes em um determinado conjunto de dados (Liu e Hauswirth, 2020).

---

<sup>1</sup> Espalhamento é medido pelo grau de dispersão: o quanto os dados são semelhantes em relação ao elemento central – após ordenar os dados.

A *AR* oferece recursos avançados para a visualização de dados em 3D, proporcionando uma visão em profundidade. Além disso, a *AR* desperta no usuário um maior engajamento na visualização de dados, pois permite uma experiência imersiva e interativa. No entanto, a codificação necessária para implementar essas visualizações pode ser complexa, o que motiva a exploração do uso de ferramentas *LLM* para a geração de códigos-fonte destinados a esse tipo de visualização, bem como para a manipulação e análise dos dados.

## **1.2. Problemática e justificativa**

Uma eficiente análise de dados depende da imputação de valores, quando da ocorrência de dados ausentes. Antes da imputação, é necessário conhecer as características do conjunto de dados, sendo que a visualização 3D pode apoiar esta tarefa.

Dessa forma, abre-se a possibilidade de escolher critérios que serão utilizados para avaliar a eficácia de um método de imputação (ou a agregação de vários métodos) e, como resultado, a abordagem a ser adotada

Nossa hipótese é que a estratégia apresentada neste trabalho apoiará o tratamento de dados ausentes. Para isso, será usada a *AR* para facilitar o entendimento das características dos dados ausentes, através de visualizações através de gráficos virtuais e *LLM* para gerar os códigos-fonte para a manipulação dos dados e geração de gráficos 3D.

## **1.3 Objetivo, contribuições e métodos**

O objetivo deste trabalho é apresentar uma estratégia para auxiliar o tratamento de dados ausentes. Para isso, propomos disponibilizar gráficos virtuais em *AR* para que analistas de dados possam obter uma compreensão aprofundada das características de um conjunto de dados. Com essa compreensão aprimorada, torna-se mais simples para os analistas de dados selecionarem o método de tratamento mais adequado a ser aplicado aos dados ausentes.

A estratégia envolve a criação de objetos virtuais e manipulação dos dados utilizando códigos-fonte gerados por uma ferramenta *LLM*. Desta forma, esses códigos-fonte, após serem adaptados ao nosso contexto, permitirão as interações e visualizações dos dados.

#### 1.4. Objetivo específico

Desenvolver e implementar uma interface em *AR* que permita a analistas de dados compreender as características de um conjunto de dados através de visualizações em gráficos virtuais e interagir com objetos virtuais para acionar métodos de tratamento de dados ausentes, além de, caso necessário, permitir adicionar novos gráficos virtuais e novos métodos de tratamento de dados, utilizando códigos-fonte gerados totalmente ou parcialmente por uma ferramenta *LLM*.

#### 1.5. Estrutura da proposta

O restante desta dissertação está estruturada da seguinte forma:

- Capítulo 2: descreve o referencial teórico e metodológico da dissertação, com foco principal na visualização virtual de dados e no tratamento destes dados, ambos com auxílio de *LLM*;
- Capítulo 3: revisão da literatura referente às estratégias de visualizações de dados e imputações para lidar com dados ausentes, além de examinar os estudos sobre técnicas de *AR*;
- Capítulo 4: apresenta a estratégia para responder à pergunta que norteia esta pesquisa: “A Realidade Aumentada com o apoio de uma ferramenta de *LLM* podem auxiliar no tratamento de dados ausentes? ”.
- Capítulo 5: apresenta o aplicativo que concretiza a estratégia descrita no capítulo anterior;
- Capítulo 6: estudo de caso do *DMT-AR*;
- Capítulo 7: considerações finais.

## 2. Referencial teórico e metodológico

Neste capítulo, é apresentado o referencial teórico utilizado na estratégia desenvolvida nesta dissertação.

A Seção 2.1 mostra o conceito de tratamento de dados ausentes e principais métodos de tratamento. A Seção 2.2 discute o conceito e exemplos de aplicações de *AR*. Na Seção 2.3 é apresentado o conceito de processamento de linguagem natural, com foco na ferramenta *ChatGPT*, que é uma implementação de *LLM*.

### 2.1. Tratamento de dados ausentes

As empresas possuem uma vasta quantidade de dados em sua posse, abrangendo um extenso volume de informações coletadas ao longo de suas operações. Segundo Sharma *et al.* (2021), diariamente cerca de 2.5 quintilhões de bytes em forma de texto, imagens, vídeos e áudios são produzidos. Com este crescimento 'explosivo de dados', métodos e técnicas para analisá-los foram aprimorados. Isso se tornou necessário, uma vez que as empresas podem tomar decisões após análises destes dados. Para uma análise eficiente, os métodos e técnicas exigem que os conjuntos de dados possuam qualidade. Caso contrário, os resultados apresentarão viés.

Dados ausentes, comumente conhecidos como valores ausentes, são reconhecidos pelos pesquisadores de qualidade de dados como uma questão importante a ser considerada (Othman e Yahia, 2018). Trata-se de um problema bem conhecido que surge em inferências estatísticas e análise de dados.

Dados ausentes são valores de variáveis que deveriam ter sido preenchidos, mas por algum motivo não foram. Segundo Hauswirth e Liu (2020), estes motivos são classificados pela literatura em três tipos, de acordo com o padrão de preenchimento de uma determinada instância: *Missing Completely at Random - MCAR* (*Missing Completamente Aleatório*), *Missing at Random - MAR* (*Missing Aleatório*), *Missing Not at Random - NMAR* (*Missing Não Aleatório*).

Conforme Hauswirth e Liu (2020) as diferenças entre *MCAR*, *MAR* e *NMAR* são:

- *MCAR*: ocorre quando os dados ausentes acontecem completamente ao acaso, não apresentam padrão e não tem nenhuma relação com o sujeito, com o item ou com as respostas aos outros itens. Isto é, quando os dados não são informados por um “esquecimento” ou “erro”. Neste caso, não há nenhuma causa que justifique a geração dessas ocorrências. Um exemplo ilustrativo seria um conjunto de dados com 10.000 registros, onde o campo 'prefere inverno ou verão' está sem resposta em 100 desses registros. Nesse caso, não encontramos nenhuma relação que identifique as causas dessas ausências.
- *MAR*: ocorre quando a ausência de resposta não tem relação direta com o próprio item, mas está relacionada com alguma variável externa ou interna do conjunto de dados. Há uma relação sistemática entre os dados ausentes e alguma informação coletada sobre os dados. Por exemplo, sabemos que os homens são mais propensos a responder perguntas sobre a sua idade e sobre o seu peso do que as mulheres. Portanto, em alguns conjuntos de dados podemos encontrar uma relação entre os dados ausentes e a idade/sexo do entrevistado.
- *NMAR*: ocorre por uma questão específica do item. Portanto, o motivo da ausência é o conteúdo do próprio item. Por exemplo em uma pesquisa existe um item que revela alguma intimidade do entrevistado e neste caso muitos deles optam em não responder. Então, há uma razão dentro do próprio item que faz com que ocorram ausências.

### **2.1.1 Métodos para tratamento de dados ausentes**

A escolha do método de tratamento de dados ausentes depende do contexto específico e dos dados disponíveis. É importante considerar as características deles, a quantidade de dados ausentes e o impacto potencial nas análises e nos resultados. Conforme Othman e Yahia (2018), existem três abordagens principais para lidar com dados ausentes:

- Excluir registros com dados ausentes: consiste em excluir registros com valores ausentes do conjunto de dados. Também é válido decidir em excluir os atributos com dados ausentes;

- Completar dados por imputação: significa preencher os valores ausentes com valores obtidos, por exemplo, aplicando desde métodos ingênuos (inserir valores sem estabelecer algum critério) até métodos que preenchem os valores através de cálculos estatísticos, por exemplo média, moda, mediana, regressão linear, interpolações e relacionamento entre os dados. Segundo Othman e Yahia (2018), a imputação é de fato a estratégia mais empregada para o problema de dados ausentes.
- Usar um algoritmo de mineração de dados ajustado para lidar com valores ausentes: os algoritmos existentes para lidar com a presença de valores ausentes se adequam ou são combinados entre si.

#### 2.1.1.1 Método de exclusão de registros

Existem várias técnicas que implementam o método de exclusão de registros com dados ausentes. No entanto, é importante ressaltar que nem sempre é recomendada porque existe o risco do descarte de informações potencialmente valiosa (Pavithrakannan *et al.*, 2021). Assim, métodos que se utilizam de imputação ou métodos ajustados são mais apropriadas para preservar informações e evitar o viés nos resultados na maioria das situações. Não é recomendado usar método de exclusão de registros quando a porcentagem de valores ausentes é alta (Pavithrakannan *et al.*, 2021). A seguir alguns exemplos de técnicas que envolvem a exclusão de registros com dados ausentes:

a) **Exclusão completa de registros (*listwise*)**: esta técnica envolve remover completamente os registros que contêm, pelo menos, um valor ausente. Embora seja simples de implementar, pode levar à perda de informações, especialmente se houver um número significativo de registros com valores ausentes. Recomenda-se essa técnica quando os mecanismos que causaram a falta de dados são do tipo *MCAR* (Makaba e Dogo, 2018).

b) **Exclusão parcial de registros (*pairwise*)**: esta técnica consiste em excluir apenas os atributos específicos nos quais ocorrem valores ausentes nos registros. Por exemplo, suponhamos um determinado registro com os atributos A, B e C. Neste registro, apenas a variável B tem valor ausente. Então, opta-se por excluir apenas essa, mantendo-se o registro.

Esta técnica é mais indicada quando os mecanismos que causaram os dados ausentes são dos tipos *MCAR* ou *MAR* (Pavithrakannan *et al.*, 2021).

c) **Exclusão baseada em limite (*threshold-based deletion*)**: Nesta técnica, define-se um limite máximo de valores ausentes permitidos em um registro. Se um registro exceder este limite, então ele será excluído. Por exemplo, se o limite for estabelecido em 50% de valores ausentes, qualquer registro com mais de 50% de valores ausentes será excluído. Esta técnica é indicada para qualquer um dos três mecanismos (*MCAR*, *NMAR* e *MAR*) de causa de dados ausentes (Pavithrakannan *et al.*, 2021).

### 2.1.1.2 Método de imputação única

Neste método, são utilizadas técnicas que substituem os dados ausentes uma única vez. A vantagem é a sua simplicidade, mas a desvantagem é a característica de ignorar completamente a própria informação, pois não é levado em consideração o padrão de respostas dessa informação (Othman e Yahia, 2018). Assim, a informação é a única que não está sendo considerada durante a substituição de seus dados ausentes. Exemplos de técnicas:

**a) Imputação pela média simples:** é uma técnica bastante simples e amplamente utilizada devido à sua facilidade de implementação. Aplica-se a média aritmética nos dados preenchidos (não se consideram os dados ausentes para este cálculo); logo depois, preenchem-se todos os dados ausentes com o valor obtido. Esta técnica não leva em consideração a estrutura dos dados ou possíveis relações entre as variáveis. Portanto, pode não ser a melhor opção em conjuntos de dados onde existem padrões complexos ou relacionamentos entre as variáveis (Psychogyos, Ilias e Askounis, 2022).

**b) Imputação pela moda:** é uma técnica que substitui os valores ausentes pelo valor mais frequente (moda) da variável em questão. Assim como a imputação pela média simples, a imputação pela moda também pode reduzir a variabilidade dos dados, já que todos os valores ausentes são substituídos pelo mesmo valor, o mais frequente. Também não leva em consideração a estrutura dos dados ou possíveis relações entre as variáveis (Psychogyos, Ilias e Askounis, 2022);

**c) Imputação pela mediana:** segundo Crespo (2009), o cálculo da mediana é realizada nos seguintes passos: primeiro coloca-se os valores em ordem (crescente ou decrescente). Na sequência, considera-se que  $N$  seja a quantidade de instâncias da variável em questão, então encontra-se a posição dentro desta coleção através da fórmula:  $posição = ((N - 1) * 2/4 + 1)$ . Por último encontra-se o valor desta posição (mediana) através da fórmula:  $mediana = valor\_posição + ((valor\_posição + 1) - valor\_posição) * parte\_decimal$ . Para exemplo usamos os valores  $\{ 6, 2, 4, 10, 8, 12\}$ . Colocando-se em ordem:  $\{ 2,4,6,8,10,12\}$ .  $N = 6$ . Então aplicando a fórmula  $posição = ((6 - 1) * 2/4 + 1)$  obtemos  $posição = 3,5$ . Em seguida:  $mediana = 6 + ((8 - 6) * 0.5) = 7$ . Segundo Psychogyos, Ilias e Askounis (2022), a imputação pela mediana também não leva em consideração a estrutura dos dados ou possíveis relações entre as variáveis;

**d) Regressão linear simples:** segundo Bangdiwalaa (2018), apesar desta técnica também realizar uma imputação única, os valores imputados são diferentes para cada inserção. Ainda conforme o autor, o objetivo é obter uma reta de ajuste, ou seja, uma reta que melhor se aproxima dos pontos do gráfico cartesiano<sup>2</sup> (X Y). O objetivo é considerar determinada variável em função de outra. Ou seja, o comportamento de uma variável influenciará o comportamento de outra. Análise de regressão significa descrever, através de um modelo matemático, a relação entre duas variáveis, partindo de  $n$  observações das mesmas (Crespo, 2009, pág. 150). Ainda, segundo este autor, o objetivo da regressão linear é prever um valor Y baseado em um valor X. Entretanto, o modelo somente será útil (para fazer previsões) se houver uma relação matemática entre estas duas variáveis que será medida, ou seja, se há uma correlação forte entre as duas variáveis, em estudo (Bangdiwalaa, 2018). Esta correlação será medida (forte ou fraca), por exemplo, pela variável R. Esta variável mostrará a força (o quanto estão relacionados) e a direção (linha de ajuste), R varia de -1 a 1, quanto mais próxima de -1 e 1 mais forte a relação é, quanto mais próxima de zero mais fraca é esta relação. O nível de confiança no modelo é medido pelo coeficiente de determinação  $R^2$  que varia de 0 a 1. Quanto mais próximo de 1 mais é confiável. Por exemplo, os gráficos das figuras 1 e 2 possuem um  $R^2$

<sup>2</sup> Gráfico Cartesiano é um objeto matemático plano e composto por duas retas numéricas perpendiculares que formam 90 graus (possuem apenas um ponto em comum chamada origem).



= 0.86, isto significa que 86% da variável dependente consegue ser explicada pela variável independente. Além disso, A Figura 1 mostra um coeficiente de correlação positivo (indicando a direção da reta de ajuste). A Figura 2 mostra um coeficiente de correlação negativo (indicando a direção da reta de ajuste). As mesmas figuras mostram também uma previsão para Y. A fórmula para encontrar R(Coeficiente de *Person*) é:  $R = (\Sigma((X - \bar{X}) * (Y - \bar{Y}))) / \sqrt{\Sigma(X - \bar{X})^2} * \sqrt{\Sigma(Y - \bar{Y})^2}$ , onde os valores observados, X e Y são as médias aritméticas e *SQRT* significa raiz quadrada. Quando ocorre dados ausentes se faz necessário definir qual variável será considerada dependente (Y) e qual independente (X). Entre as variáveis de um conjunto de dados verifica-se quais são as mais correlacionadas. Depois, entre estas, a que possuir maior número de dados ausentes será escolhida como variável dependente e a outra independente.

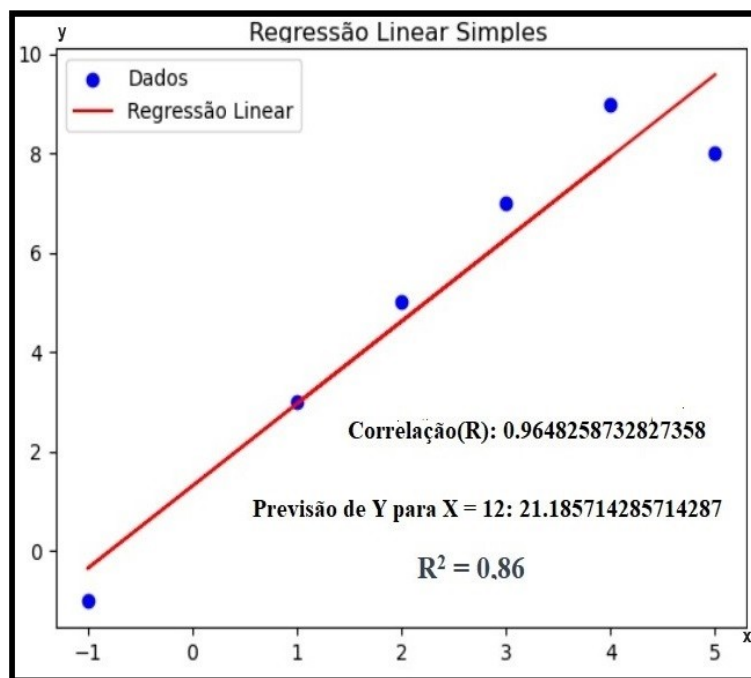


Figura 1: Gráfico de regressão com correlação positiva

A reta de ajuste (regressão linear) é alcançada através da relação:  $Y = A + B * X$ , onde A é o intercepto (ponto que toca o eixo X) e B é o coeficiente angular (valor fixo). Para encontrar o valor de B utiliza-se a fórmula:  $B = ((n * (\Sigma (X * Y)) - (\Sigma X * \Sigma Y)) / ((n * \Sigma X^2) - (\Sigma X)^2))$ . Para encontrar o valor de A a fórmula a ser utilizada é:  $A = ((\Sigma Y - B * \Sigma X) / n)$ . Onde,

n é o número de observações, X a variável independente e Y a variável dependente (Crespo, 2009, pág.151). Supomos que temos o conjunto  $X = \{ 3, 2, -1, 4, 1,5\}$  e  $Y = \{ 7, 5, -1, 9, NA, NA\}$ . 'NA' representa dados ausentes. Então teremos  $B = ((4 * (68) - (8 * 20)) / ((4 * 30) - 64))$ , logo  $B = 2$ . Para A teremos  $A = ((20 - 2 * 8) / 4)$ , então,  $A = 1$ . Portanto, a Regressão Linear(linha de ajuste) será indicada pela relação  $Y = 1 + 2 * X$ . Desta forma os dados ausentes em Y serão, respectivamente: 3 e 11. Além disso o valor de  $R = 1$  e  $R^2 = 1$ , assim o modelo é praticamente 100% confiável.

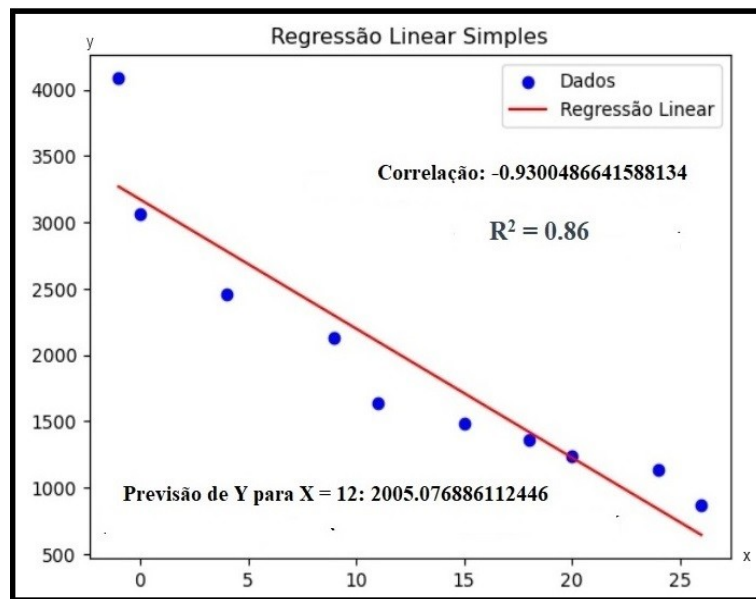


Figura 2: Gráfico de regressão com correlação negativa

A Figura 3 mostra o Gráfico de Regressão deste exemplo. Quando existirem duas ou mais variáveis independentes teremos a Regressão Linear Múltipla. Esta é definida pela relação:  $Y = A + B1 * X1 + B2 * X2 + \dots + Bn * Xn$ , onde X são as variáveis independentes e B os coeficientes angulares.

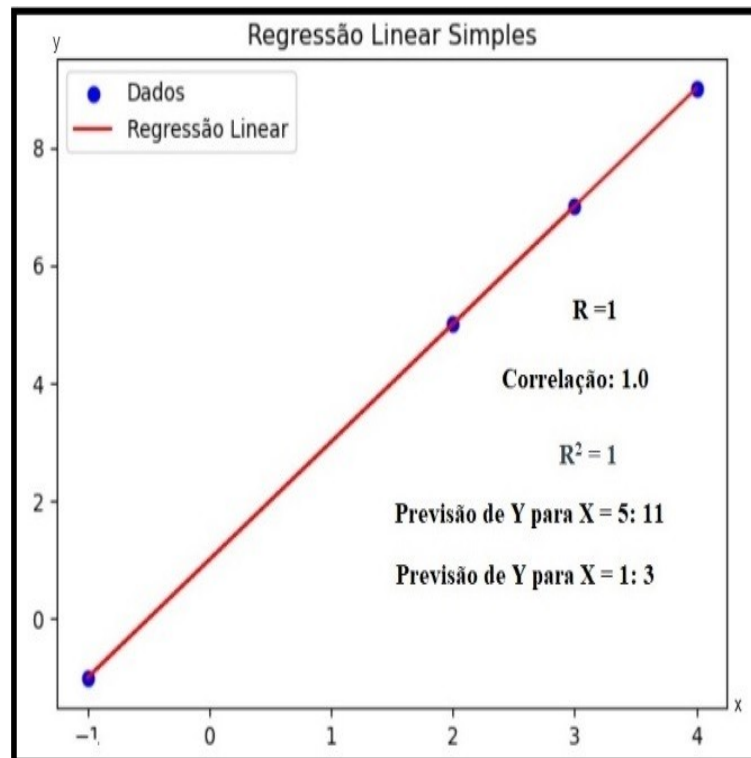


Figura 3: Gráfico de regressão para os conjuntos X e Y

### 2.1.1.3 Método de imputação múltipla

A imputação múltipla foi proposta por *D. B. Rubin* nos anos 70, entretanto, apenas recentemente esse método vem sendo mais utilizado devido aos desenvolvimentos computacionais para a sua implementação (Klück, Nunes e Fachel, 2009). Segundo essas autoras, após identificar e compreender os motivos dos dados estarem ausentes, aplica-se o processo de imputação múltipla, que consiste em três passos, descritos a seguir:

1) a partir do banco de dados original são obtidos  $n$  bancos de dados com os dados ausentes preenchidos (o valor de  $n$  é definido pelo pesquisador), ou seja, cria-se uma amostragem múltipla. São criadas várias cópias do conjunto de dados original, cada uma das  $n$  cópias com valores imputados de maneiras diferentes, como por exemplo, pela média, moda, mediana ou regressão linear, para as observações ausentes. Cada uma das  $n$  cópias, agora sem dados ausentes, são chamados de conjuntos de dados imputados, tratados como uma réplica do conjunto de dados original, cujos valores ausentes são preenchidos, opcionalmente, com técnicas diferentes ou iguais;

2) cada banco de dados é analisado separadamente. Para as análises pode-se utilizar técnicas como estatísticas, coeficientes de regressão, cálculo de quartis e outras. A razão para realizar essas análises em cada banco de dados é capturar as incertezas associadas à imputação como, por exemplo, discrepâncias, viés e imprecisões. Portanto, analisa-se como as diferentes imputações podem afetar os resultados;

3) os resultados, obtidos através dos passos anteriores, são combinados entre todas as  $n$  versões dos bancos de dados imputados. Estas combinações podem ser realizadas calculando a média, mediana ou qualquer outra técnica mais avançada de combinação, dependendo da natureza dos dados e objetivo da análise.

## **2.2. Realidade Aumentada**

A *AR* é a combinação de objetos virtuais com objetos reais em um mesmo ambiente, proporcionando ao usuário final a impressão de que esses objetos virtuais também pertencem ao ambiente real. Meiguins *et al.* (2021) define a *AR* como um sistema que combina elementos virtuais com elementos reais, é interativo em tempo real e registra ou alinha objetos reais e virtuais em uma única cena, dando a sensação ao usuário que os objetos virtuais fazem parte do ambiente real.

Ainda, conforme esses autores, a *AR* enriquece uma cena, com textos, imagens, vídeos e outros elementos virtuais, pois existe um vasto espaço visual para apresentação desses elementos. Um exemplo seria um estudante que, ao adentrar em uma biblioteca e se aproximar das estantes, visualiza informações virtuais através do aparelho celular ou outro dispositivo móvel, a respeito do acervo.

Como as tecnologias de aparelhos móveis desenvolveram-se rapidamente e foram amplamente difundidas, estes passaram a fornecer uma vasta variedade de aplicativos e serviços para o benefício de seus usuários. Assim, maneiras de interação entre o ser humano e os dispositivos móveis têm recebido muita atenção por parte de pesquisadores. Um dos meios para enriquecer esta interação é a *AR*, no qual o conteúdo virtual é colocado em cima da visão de câmera, no mundo real (Kim *et al.*, 2014).

Desta forma, os dispositivos físicos desempenham um papel crucial na entrega da experiência de sobreposição de elementos virtuais ao ambiente real. Existem diversos dispositivos utilizados para esse propósito, cada um com suas características e funcionalidades específicas.

Conforme Galvão, Carmona e Ricarte (2018), alguns dos principais dispositivos físicos utilizados em *AR* são:

- *Smartphones e tablets*: as câmeras destes capturam o ambiente real e sobrepõem elementos virtuais na tela;
- Óculos de *AR (Head Mounted Display HMD)*: possuem câmeras, sensores e exibições integradas que permitem aos usuários interagirem com elementos virtuais;
- *Headsets de AR*: é semelhante ao óculos de *AR*, mas possuem áudio e exibição estereoscópica, o que torna a experiência de *AR* mais rica e interativa;
- *Smart glasses*: além das funcionalidades de *AR* possuem um foco mais amplo, incluindo notificações, informações contextuais e outras capacidades.

Esses são apenas alguns exemplos de dispositivos físicos utilizados em sistemas de *AR*. Com o avanço da tecnologia, é provável que novos dispositivos e formas de interações sejam desenvolvidos para proporcionar experiências de *AR* ainda mais imersivas e inovadoras.

Segundo Xia (2022), com relação às interações com os usuários os sistemas de *AR* podem oferecer aos usuários:

- Interações com os elementos virtuais por meio de toques na tela e/ou movimentando o dispositivo físico;
- Rastreamento dos movimentos dos usuários através de sensores, assim poderão interagirem com os objetos virtuais por meio de movimentos físicos;
- Interações usando comandos de voz;

Com base às localizações dos objetos virtuais, Lim, Kim e Park (2016), fazem a seguinte classificação:

- *AR* em tela: os objetos virtuais são sobrepostos à tela do dispositivo físico utilizado;
- *AR* no ambiente: os objetos virtuais são integrados ao ambiente físico, por exemplo em uma parede.

Sobre o tipo de rastreamento, segundo Lim, Kim e Park (2016), os sistemas de *AR* classificam-se em:

- Baseados em marcadores: são ativados com o reconhecimento de marcadores visuais, com códigos de respostas rápidas (*QR code*) ou marcadores impresso, que servem de referencial para sobrepor os objetos virtuais;
- Sem marcadores: não são necessários marcadores físicos. Ao invés disso, é utilizado sensores, câmeras e algoritmos para rastrear o ambiente e posicionar os objetos virtuais.

Neste trabalho, a *AR* é utilizada para a visualização de dados é baseada em marcadores, possibilitando a mensuração e localização dos dados ausentes em um determinado conjunto de dados.

### **2.3. Processamento de Linguagem Natural**

A linguagem natural é a ferramenta que os humanos usam para se comunicar entre si. Extrair informações significativas desse tipo de dados e manipulá-las usando computadores está dentro do campo do Processamento de Linguagem Natural (*Natural Language Processing - NLP*) (Rodrigues, 2022). O surgimento de modelos de linguagens pré-treinadas deu início a uma nova era no campo do *NLP*, permitindo-nos construir modelos de linguagens poderosas. Nesse contexto, os algoritmos de *LLM* tornaram-se uma abordagem relevante para lidar com volumes grandes e diversificados de dados e extrair informações úteis. Portanto, *LLM* refere-se a modelos de linguagem que possuem um grande número de parâmetros e foram treinados em uma grande quantidade de dados textuais.

As aplicações mais comuns de *LLM* são do tipo “Perguntas e Respostas” que orientam alguém que está tentando resolver um problema, agindo como sistemas especialistas e

tradutores. Um *LLM* tem numerosos casos de uso e pode ser solicitado a exibir uma ampla variedade de comportamentos, incluindo o diálogo (Shanahan, McDonell e Reynolds, 2023).

No uso contemporâneo, o termo *LLM* tende a ser reservado para ferramentas baseadas em *transformers* que possuem bilhões de parâmetros e são treinadas em bilhões de *tokens*, como por exemplo o *GPT-3 (Generative Pre-trained Transformer 3)*. Ferramentas de *LLM* como essas são o componente central dos agentes de diálogo, incluindo o *ChatGPT* da *OpenAI*, o *Bing Chat* da *Microsoft* e o *Bard* do *Google* (Paes, Vianna e Rodrigues, 2023).

Segundo essas autoras, na arquitetura dos *transformers*, para resolver tarefas, o mais comum é considerar dois grandes componentes (duas Redes Neurais Profundas<sup>3</sup>). O primeiro, chamado de *encoder* ou codificador, que é responsável por processar a sequência de entrada e codificá-la como um vetor de números (*embedding*), levando em conta o *token* e sua contextualização. O segundo componente, que é chamado de *decoder* ou decodificador, é responsável por receber e processar o vetor de número, também conhecido como vetor de contexto, e transformá-lo na sequência de saída. Portanto, o *encoder* processa a entrada e codifica-a em um vetor de contexto, enquanto o *decoder* recebe esse vetor de contexto e gera a sequência de saída.

Cada camada do *encoder* possui dois componentes: o *Feed Forward* que faz a conexão com a próxima camada e *self-attention* que computa pesos para diferentes partes da entrada dinamicamente, dependendo do contexto.

O *decoder*, além desses dois componentes possui um terceiro, chamado de *attention encoder-decoder*, que executa os mesmos cálculos do componente *self-attention*, mas levando em consideração somente a sequência de entrada que vem do *encoder*.

### **2.3.1. Uso de uma ferramenta *LLM* para geração de código-fonte**

O uso de uma ferramenta de *LLM* pode significativamente potencializar a produtividade de desenvolvedores em várias frentes. Uma delas é seu papel como uma fonte rápida de informações e respostas para dúvidas cotidianas relacionadas ao trabalho de programação. Em vez de investir o tempo pesquisando manualmente em fóruns ou

---

<sup>3</sup> Redes Neurais Profundas: Redes Neurais com mais de duas camadas

documentos, os desenvolvedores podem formular perguntas diretamente à ferramenta *LLM* e receber respostas em tempo real. Isso não apenas economiza tempo, mas também permite que os desenvolvedores se concentrem mais em tarefas de maior valor agregado, aumentando assim a eficiência geral do processo de desenvolvimento de software.

Entretanto, segundo Khoury *et al.* (2023), as ferramentas *LLM* atualmente podem gerar códigos-fonte, mas não robustos, pois a geração de códigos-fonte está relacionada diretamente com o contexto da solução pretendida. Conforme esses autores, apesar dessa limitação, esta abordagem tem grande potencial para acelerar os processos de desenvolvimento de softwares, pois além de gerar códigos-fonte, é capaz de detectar e localizar *bugs*, otimizar códigos-fonte, fazer conclusões sobre vulnerabilidades e fornecer explicações sobre os códigos-fonte.

A ferramenta *LLM* utilizada neste trabalho, para a geração de códigos-fonte é o *ChatGPT 3.5*, que foi desenvolvida pela *OpenAI*. O *ChatGPT 3.5* possui bilhões de parâmetros e foi treinado em uma vasta quantidade de dados textuais, permitindo uma compreensão profunda e uma capacidade de gerar respostas contextualmente relevantes em linguagem natural ([Introduction - OpenAI API](#)).

O *ChatGPT* pode ser uma ferramenta útil para gerar códigos-fonte inicial, mas é importante ter um conhecimento sólido de programação e realizar revisões rigorosas para garantir que o código atenda aos padrões e requisitos específicos do projeto. Desta forma, para assegurar a qualidade nas respostas geradas pelo *ChatGPT*, é necessário refinar os *prompts*. Em outras palavras, é crucial padronizá-los, pois fornecem diretrizes reutilizáveis para contextos específicos (White *et al.*, 2023).

Exemplo de um *prompt*: “gere código-fonte para retornar a quantidade de valores ausentes de um arquivo csv, na primeira coluna deve aparecer o nome de cada coluna do arquivo csv e na segunda coluna a quantia de dados ausentes destas colunas. Crie um outro arquivo csv com essas informações”. A Figura 4 mostra o retorno fornecido pela ferramenta a esta solicitação.



```
import pandas as pd

# Carregue o arquivo CSV
nome_arquivo = 'seuarquivo.csv'
df = pd.read_csv(nome_arquivo)

# Obtenha uma contagem de valores ausentes em cada coluna
valores_ausentes = df.isnull().sum()

# Crie um novo DataFrame para armazenar os resultados
resultados = pd.DataFrame({'Nome da Coluna': valores_ausentes.index,
'Quantidade de Valores Ausentes': valores_ausentes.values})

# Salve os resultados em um novo arquivo CSV
resultados.to_csv('valores_ausentes.csv', index=False)

print("Valores ausentes salvos em 'valores_ausentes.csv'")
```

Figura 4: Exemplo para contagem de dados ausentes

Refinando essa solicitação, podemos solicitar o total de dados de uma coluna, a quantidade de dados ausentes e a porcentagem de dados ausentes em uma coluna específica de um arquivo csv. Por exemplo, o seguinte *prompt* pode ser utilizado: "Como obter o total de linhas do arquivo csv?" e "Como obter a quantidade de dados ausentes de cada coluna, neste mesmo arquivo csv?" e "Como obter a porcentagem de dados ausentes em relação ao total de linhas deste mesmo arquivo csv?". A Figura 5 mostra o retorno fornecido pela referida ferramenta, a essa solicitação.

```

import pandas as pd
# Criar um DataFrame com os dados
df = pd.read_csv('seuarquivo.csv')
nome_da_coluna = 'suacoluna'
quantidade_de_valores_ausentes
df[nome_da_coluna].isnull().sum()
quantidade_de_linhas = len(df)
porcentagem_de_valores_ausentes
(quantidade_de_valores_ausentes/quantidade_de_linhas)
100

porcentagem_de_valores_ausentes
round(porcentagem_de_valores_ausentes, 2)
dados = {'quantidade_de_linhas': [quantidade_de_linhas],
        'quantidade_de_valores_ausentes':
[quantidade_de_valores_ausentes],
        'Porcentagem_de_valores_ausentes':
[porcentagem_de_valores_ausentes]}
df_resultado = pd.DataFrame(dados)
# Especifique o nome do arquivo CSV de destino
nome_arquivo_destino = 'resultado.csv'
# Salvar o DataFrame em um arquivo CSV
df_resultado.to_csv(nome_arquivo_destino, index=False)

print(f'Dados salvos em {nome_arquivo_destino}')

```

Figura 5: Exemplo para rastreamento de dados ausentes

Uma funcionalidade adicional, altamente benéfica para desenvolvedores, é a obtenção de informações sobre a integração de *API's - Application Programming Interface* e bibliotecas populares. O *ChatGPT* tem a capacidade de fornecer exemplos práticos e de uso para bibliotecas ou *frameworks* relevantes.

### 2.3.2. Uso de uma ferramenta *LLM* para gerar visualizações

A utilização de uma ferramenta *LLM*, em conjunto com outras ferramentas, como as plataformas de programação *R*, *Python* e planilhas eletrônicas, proporciona diversos benefícios para o desenvolvimento de projetos de visualização de dados, uma vez que essas ferramentas oferecem bibliotecas para tal integração (Buscemi, 2023).

Ao integrar uma ferramenta *LLM*, como o *ChatGPT*, ao processo de visualização de informações, os profissionais de análise de dados podem otimizar o uso de recursos preciosos

e reduzir o tempo gasto, possibilitando que eles se dediquem à tomada de decisões do negócio (Biswas *et.al.*, 2023).

Entretanto, segundo esses autores, é preciso ficar atento às limitações na compreensão de conceitos, pois a ferramenta *LLM* utilizada pode não entender completamente as complexidades dos *prompts* fornecidos.

Para exemplificar o auxílio de *LLM* na visualização e análise de dados podemos dar continuidade aos *prompts* do *ChatGPT* anteriores: “Crie um gráfico 2D que mostre o nome de cada coluna e a quantidade de dados ausentes de cada coluna” O código-fonte para a geração do referido gráfico é apresentado na Figura 6.

```
import pandas as pd
import matplotlib.pyplot as plt

# Carregue os dados do arquivo CSV
nome_arquivo = 'valores_ausentes.csv'
df = pd.read_csv(nome_arquivo)

# Crie um gráfico de barras
plt.figure(figsize=(10, 6))
plt.bar(df['Nome da Coluna'], df['Quantidade de Valores Ausentes'])
plt.xlabel('Nome da Coluna')
plt.ylabel('Quantidade de Valores Ausentes')
plt.title('Quantidade de Valores Ausentes por Coluna')
plt.xticks(rotation=90) # Rotacione os rótulos do eixo x
para melhor legibilidade
plt.tight_layout()

# Exiba o gráfico
plt.show()
```

Figura 6: Exemplo de geração de gráfico 2D

Executando o código-fonte da Figura 6 é gerado o gráfico que pode ser visualizado na Figura 7. Através desse gráfico é possível a visualização do nome das colunas e as respectivas quantidades de dados ausentes.

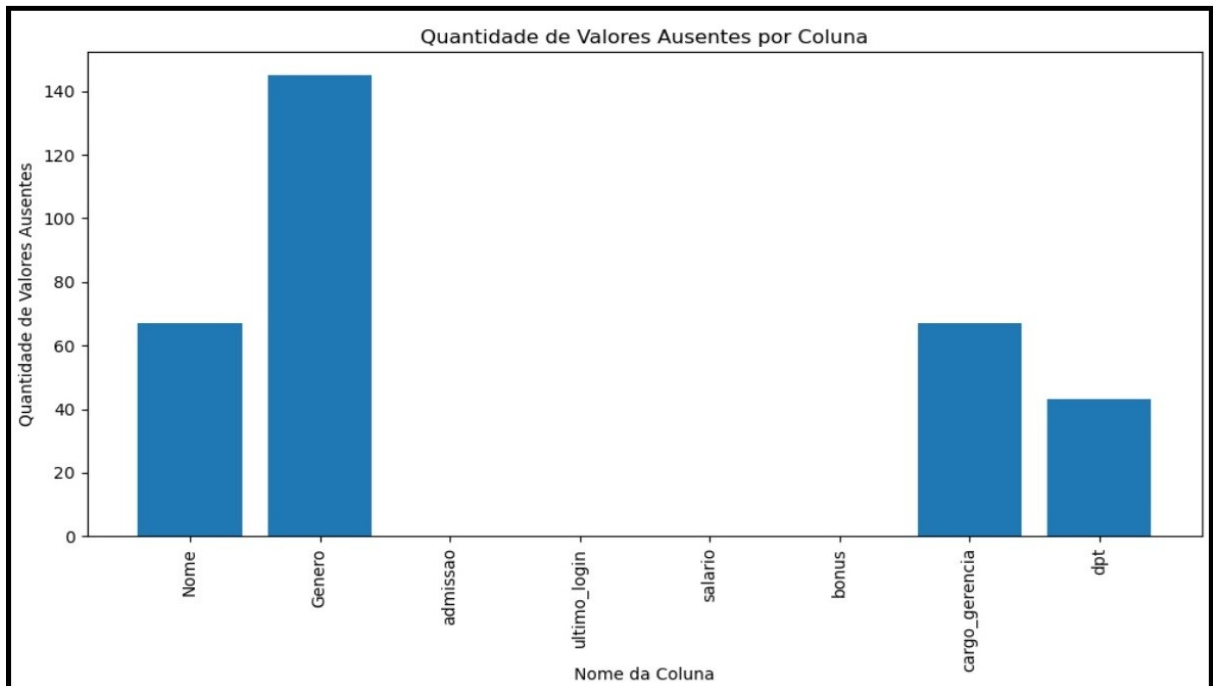


Figura 7: Exemplo de visualização através de gráfico 2D

Para completar essa análise, solicitamos a geração de um gráfico 3D. Por exemplo, fornecemos ao *ChatGPT* o *prompt*: “gere código-fonte para gerar um gráfico 3D. Na primeira dimensão deverá informar o total de dados de uma única coluna, na segunda dimensão o total de dados ausentes desta coluna e na terceira dimensão a porcentagem da segunda coluna em relação a primeira. A referida coluna pertence a um arquivo csv”. A Figura 8 mostra o código gerado.

```

from mpl_toolkits.mplot3d import Axes3D
import pandas as pd
# Carregue o arquivo CSV em um DataFrame
df = pd.read_csv('seuarquivo.csv')
# Especifique a coluna da qual você deseja analisar os
dados
coluna_alvo = 'nome_da_sua_coluna' # Substitua pelo
nome da sua coluna
# Calcule o total de dados na coluna
total_de_dados = len(df[coluna_alvo])
# Calcule o total de dados ausentes na coluna
dados_ausentes = df[coluna_alvo].isnull().sum()
# Calcule a porcentagem de dados ausentes em relação ao
total
porcentagem_ausentes = (dados_ausentes
total_de_dados) * 100
# Crie um gráfico 3D
figura = plt.figure()
eixo = figura.add_subplot(1, 1, 1, projection='3d')
eixo.scatter(total_de_dados, dados_ausentes,
porcentagem_ausentes)
eixo.set_xlabel('Total de Dados')
eixo.set_ylabel('Dados Ausentes')
eixo.set_zlabel('Porcentagem de Dados Ausentes (%)')
plt.title(f'Análise da Coluna: {coluna_alvo}')
plt.show()

```

Figura 8: Exemplo da geração de um gráfico 3D

Executando o código-fonte da figura 8 é gerado o gráfico 3D que pode ser visualizado na Figura 9.

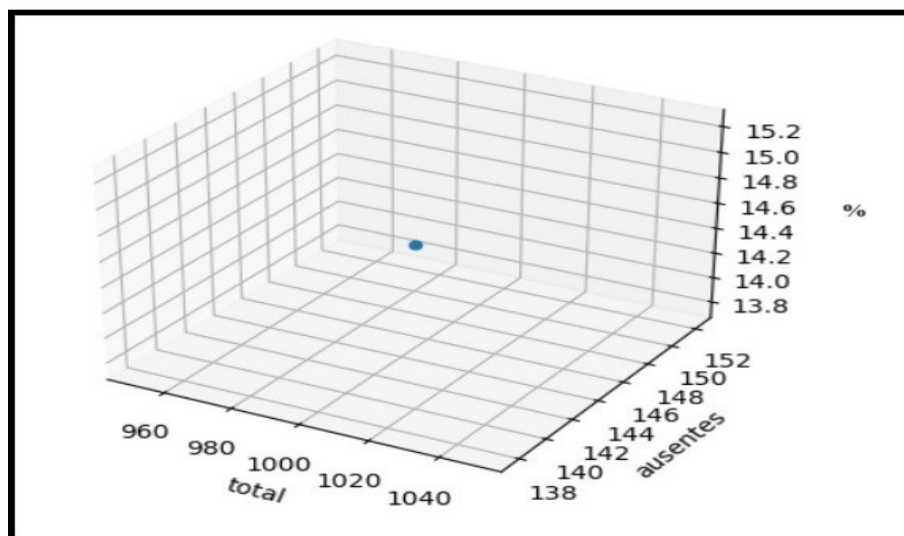


Figura 9: Exemplo da visualização através de gráfico 3D

### 3. Revisão da literatura

Realizamos uma revisão sistemática na literatura científica que implementou as seguintes etapas: definição de uma pergunta norteadora; critérios de buscas; palavras chaves e base de dados selecionados; análises dos trabalhos selecionados; resumo dos trabalhos selecionados e conclusão. Conforme a Figura 10, as etapas realizadas foram:

- 1) **Definição da pergunta que norteou a pesquisa:** “A Realidade Aumentada com o apoio de uma ferramenta de *LLM* podem auxiliar no tratamento de dados ausentes?”.
- 2) **Identificação das palavras chaves:** as referidas palavras chaves são *augmented reality* e *data missing*, que serão identificadas nos títulos e nos resumos das publicações.
- 3) **Os critérios de inclusão dos artigos selecionados foram:** artigos que tratam de realidade aumentada e tratamento de dados ausentes. Somente artigos publicados entre os anos de 2017 a 2022 e publicados na língua inglesa.
- 4) **Analisar criticamente e avaliar todos os estudos incluídos na revisão:** desta forma certificar-se que todos os artigos importantes ou que possam ter algum impacto na conclusão da revisão sejam incluídos;
- 5) **Preparar um resumo crítico:** síntese das informações encontradas nos artigos que foram incluídas na revisão;
- 6) **Apresentar uma conclusão da revisão:** informações a respeito das evidências sobre a hipótese desta pesquisa.

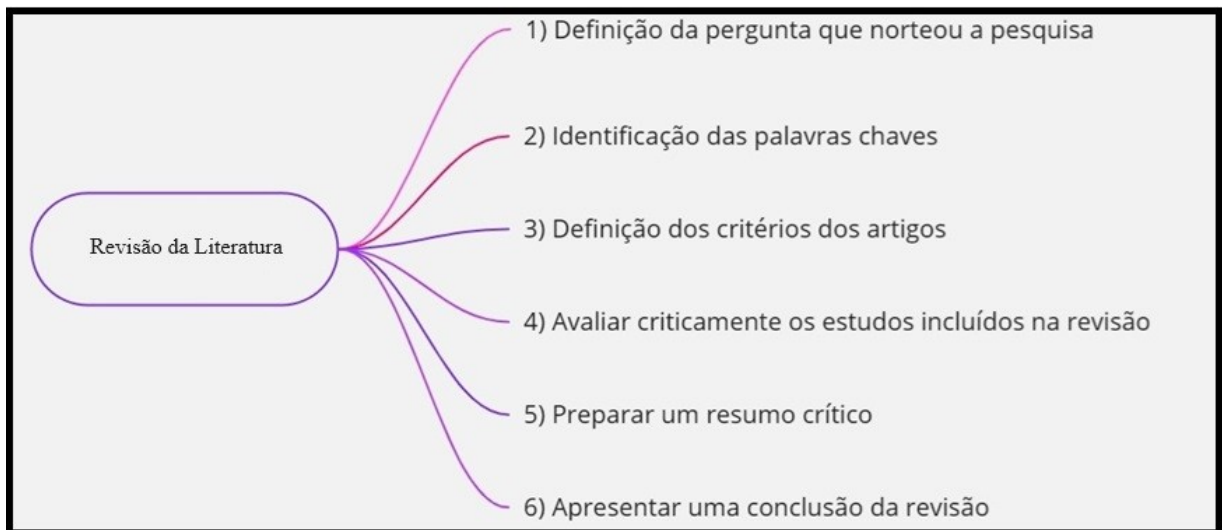


Figura 10: Etapas da revisão da literatura

Iniciamos a referida revisão da literatura, definindo uma *string* de busca que encapsula os conceitos centrais de nosso interesse. A *string* de busca formulada foi: ("*data missing*" AND "*augmented reality* ") or ("*data visualization*" and "*augmented reality*") que retornou 4.178 artigos relacionados à nossa pergunta norteadora, sendo 1.695 da *IEEE/Xplore* e 2.473 da *ScienceDirect*. Contudo, a amplitude dessa resposta inicial revelou-se excessiva para a análise detalhada que buscamos realizar. Consequentemente, reconhecemos a necessidade de refinar nossa pesquisa a fim de focalizá-la em aspectos mais específicos e relevantes para nossos objetivos de estudo. Desta forma, formulamos as seguintes *strings* de buscas: ("*missing data*" AND "*augmented reality*"), que retornou 45 trabalhos, sendo 26 *IEEE/Xplore* e 19 na *ScienceDirect* e a *string* ("*data visualization*" and "*augmented reality*"), que retornou 427 trabalhos, sendo 267 da *IEEE/Xplore* e 160 da *ScienceDirect*. As Tabelas 1 e 2 resumem as referidas buscas à época da escrita desta dissertação. Além disso, é importante ressaltar que outras bases foram consultadas, e embora os artigos encontrados nelas não tenham sido incluídos nesta revisão, eles são referenciados ao longo da dissertação.

Tabela 1– *Strings* de busca, *url* e resultado na base de dados *IEEE/Xplore*

<b>String de busca</b>	<b>Resultado</b>
<i>("data missing" AND "augmented reality ") or ("data visualization" and "augmented reality")</i>	<i>1695 results</i>
<i>("missing data" AND "augmented reality")</i>	<i>26 results</i>
<i>("data visualization" ) AND ("augmented reality")</i>	<i>267 results</i>

Tabela 2 – *Strings* de busca, *url* e resultado na base de dados *ScienceDirect*

<b>URL</b>	<b>Resultado</b>
<i>("data missing" AND "augmented reality") or ("data visualization" and "augmented reality")</i>	<i>2473 results</i>
<i>("data missing" AND "augmented reality")</i>	<i>19 results</i>
<i>("data visualization" ) AND ("augmented reality")</i>	<i>160 results</i>

Para focar em resultados mais pertinente, considerando os trabalhos retornados com estas filtragens, fizemos uma leitura criteriosa e crítica de seus resumos. Com base nas análises dos resumos, foram selecionados 26 artigos, sendo 25 pertencentes à base de dados do



*IEEE/Xplore* e 1 pertencente à base de dados da *ScienceDirect*. Encontrou-se resultados positivos referente a utilização de *AR* para auxiliar a análise de dados. Apesar dos trabalhos selecionados a este tema terem finalidades diferentes ao nosso propósito, verificou-se um conjunto de técnicas que podemos associar a nossa pesquisa. A Tabela 3 sintetiza os 26 trabalhos, com seus objetivos e resultados.

Tabela 3 – Artigos selecionados nas bases de dados *IEEE/Xplore* e *ScienceDirect*

Nº	Título da Publicação	Autor/Ano/Base	Objetivo
1	<i>Volume Rendering: An Analysis based on the HoloLens Augmented Reality Device.</i>	Rezende et. al. 2020/IEEE	Verificar a viabilidade de visualização de dados volumétricos em tempo real com apoio da <i>AR</i> .
2	<i>Augmented Reality for Easy Sailing</i>	Laera 2020/IEEE	Projetar um aplicativo em <i>AR</i> para apoiar a navegação em veleiros.
3	<i>Augmented Reality for Big Data Visualization: A Review</i>	Chanda, Jamiy e Reza 2019/IEEE	Levantar na literatura científica o estado da arte do uso da <i>AR</i> como tecnologia para a análise de dados.
4	<i>Look Inside: Understanding Thermal Flux Through Augmented Reality</i>	Kiss, Schmidt e Knierim 2018/IEEE	Desenvolvimento de um aplicativo <i>AR</i> que permite aos alunos de física visualizarem dados virtuais representando o fluxo de calor projetado em uma haste.
5	<i>Look Inside: Understanding Thermal Flux Through Augmented Reality</i>	Gayanika e Ekanayake 2022/IEEE	Analisar o panorama atual das pesquisas direcionadas à criação de visualizações virtuais de dados com o objetivo de aprimorar a experiência pedagógica.
6	<i>Card-Based Interaction to Design Visualizations in Augmented Reality Environments</i>	Meiguins et. al. 2019/IEEE	Visualizar dados virtualmente, através de cartões físicos.
7	<i>An approach towards data visualization based on AR principles</i>	Hirve et. al. 2017/IEEE	Unir <i>AR</i> e <i>Big Data</i> para gerar visualizações virtuais.

8	<i>Connecting User Experience to Learning in an Evaluation of an Immersive, Interactive, Multimodal Augmented Reality Virtual Diorama in a Natural History Museum &amp; the Importance of Story</i>	Harrington 2020/IEEE	Apresentar o <i>AR Perpetual Garden App</i> . Trata-se de um aplicativo que tem uma finalidade educacional.
9	<i>MARVisT: Authoring Glyph-Based Visualization in Mobile Augmented Reality</i>	Chen et. al. 2020/IEEE	Apresentar o aplicativo <i>MARVisT</i> , assim permitir que usuários leigos criem aplicações em <i>AR</i> .
10	<i>CDVVAR: VR/AR Collaborative Data Visualization Tool</i>	Yassien, Emad e Abdennadher 2021/IEEE	Apresentar a ferramenta <i>CDVVAR</i> . Esta permite a visualização de dados tanto com a <i>AR</i> como também com a Realidade Virtual.
11	<i>Using the augmented reality sandbox for advanced learning in geoscience education</i>	Kundu, Muhammad e Sattar 2017/IEEE	Construção do aplicativo <i>SandBox</i> para alcançar o aprendizado avançado em Geociência utilizando <i>AR</i> .
12	<i>Coordinate: A Spreadsheet-Programmable Augmented Reality Framework for Immersive Map-Based Visualizations</i>	Shaikh 2019/IEEE	Apresentar o <i>Coordinate</i> , uma ferramenta para sobreposições virtuais em <i>AR</i> sobre mapas utilizando aparelhos móveis.
13	<i>Personal Augmented Reality for Information Visualization on Large Interactive Displays</i>	Respschlager, Flemisch e Dachelt 2021/IEEE	Utilizar <i>AR</i> em conjunto com telas grandes, para apoiar a análise de dados.
14	<i>Augmented Reality for Teaching Data Structures in Computer Science</i>	Narman et. al. 2020/IEEE	Apresentar um protótipo, em <i>AR</i> , para estudar conteúdos práticos de estrutura de dados.
15	<i>Visualizing Toronto City Data with HoloLens: Using Augmented Reality for a City Model</i>	Zhang et. al. 2018/IEEE	Discutir o uso do <i>Microsoft HoloLens</i> para visualizações de dados virtuais sobre o mapa físico da cidade de Toronto.
16	<i>DXR: A Toolkit for Building Immersive Data Visualizations</i>	Sicat et. al. 2019/IEEE	Apresentar o aplicativo <i>DXR</i> que permite a construção de objetos virtuais.

17	<i>When Augmented Reality meets Big Data</i>	Bermejo et. al. 2017/IEEE	Investigar o potencial de valor para a convergência entre <i>AR</i> e <i>Big Data</i> .
18	<i>Mid-Air Fingertip-Based User Interaction in Mixed Reality</i>	Dani et. al. 2018/IEEE	Discutir uma maneira mais econômica para visualização de dados com o uso de <i>AR</i> e Realidade Virtual em conjunto, cujas interações são realizadas através de gestos. Uma alternativa a preços exorbitantes de outras abordagens.
19	<i>Design of a Node Status Visualizing Software Utilizing the AR Technology for Multihop Wireless Networks</i>	Otha, Ito e Kakuda 2017/IEEE	Visualizar dados virtuais da topologia de redes e dados virtuais dos estados dos nós desta rede.
20	<i>Application of 3D tracking and registration in exhibition hall navigation interaction</i>	Xin et. al. 2020/IEEE	Apresentar o desenvolvimento e a execução de um <i>software</i> em <i>AR</i> , em um dispositivo móvel <i>Android</i> , que atua como um guia de virtual de um salão de exposição.
21	<i>Visualizing the "Hidden" Variables in Robot Programs</i>	Shepherd, Kraft e Francis 2019/IEEE	Investigar o uso de <i>AR</i> para apoiar a programação de robôs industriais, a fim de reduzir custos de programação.
22	<i>A Usability Assessment Of Augmented Situated Visualization</i>	Guarese et. al. 2020/IEEE	Apresentar um aplicativo de <i>AR</i> que é instalado no <i>HMD Microsoft HoloLens</i> para visualização de dados virtuais e interações com os mesmos, para facilitar tomadas de decisões.
23	<i>The Use of Mixed-Realities Techniques for the Representation of Islamic Cultural Heritage</i>	Mohamed 2017/IEEE	<i>Estuda a possibilidade da criação de um sistema em AR que permita aos visitantes visualizarem dados virtuais de um museu islâmico.</i>

24	<i>Analysis of XR Research in Brazil from 21 Years of SVR Publications</i>	Teixeira e Peres 2020/IEEE	Uma análise das edições de 1997 até 2019 do Simpósio de Realidade Virtual e Aumentada (SVR).
25	<i>Visualizing Missing Values</i>	Sjobergh e Tanaka 2017/IEEE	Construir o sistema de visualização para análise exploratória de conjuntos de dados que possuem dados ausentes. Este sistema apresenta diferentes maneiras de visualizações em um único campo de visão.
26	<i>Collaborative program comprehension via software visualization in extended reality</i>	ScienceDirect 2022/IEE	Apresentar o <i>ExplorViz</i> . Trata-se de uma implementação para visualização virtual e colaborativa para o desenvolvimento de programas de computação.

### 3.1 Resultados da revisão da literatura

A seguir, apresenta-se uma análise resumida dos artigos selecionados, os quais estão listados na Tabela 3 e são identificados pelo número da linha da referida tabela.

Um estudo (1) foi realizado por Rezende *et al.* (2020) para verificar a viabilidade de visualização de dados volumétricos em tempo real com apoio da *AR* utilizando-se o *HMD (Head Mounted Display) Microsoft HoloLens*. Foi desenvolvido um protótipo, que por meio da *AR*, gerou visualizações holográficas no ambiente real. No entanto, a baixa capacidade de memória desses dispositivos foi identificada como um fator limitante. Apenas visualizações que não excediam a capacidade de memória do dispositivo foram geradas.

Em outro estudo (2) realizado por Laera (2020), buscou-se projetar um aplicativo em *AR* para apoiar a navegação em veleiros. Dessa maneira, os velejadores poderiam visualizar informações náuticas virtuais (25 variáveis) sobrepostas no ambiente, correspondentes aos fenômenos físicos. Os estudos conduzidos levaram à conclusão de que a utilização da *AR* para

visualização dessas informações durante as viagens é perfeitamente possível, resultando, conseqüentemente, no aumento da segurança.

Em uma revisão literária (3), Chandra, Jamiy e Reza (2019) realizaram um levantamento da literatura científica para compreender o estado da arte do uso da *AR* como tecnologia para a análise de dados de *Big Data*. Segundo os autores, faltam ferramentas e técnicas convencionais de visualização de dados para lidar com essas análises. Estes autores destacam que a *AR* por si só, permite que a visualização virtual seja incorporada ao mundo real para fornecer informações mais significativas de dados. Desta forma técnicas imersivas e interativas para visualizações multidimensionais levam a uma melhor experiência quando comparadas aos *desktops* tradicionais. O artigo retornou alguns trabalhos de pesquisa em *AR* (medicina, ciência, engenharia, etc) que estão sendo usados como ferramenta primária ou complementar para visualizar *Big Data*. As visualizações baseadas em *AR* são acessíveis e aplicadas em vários domínios. Entretanto, este tipo de visualização ainda possui limitações (como *hardware*, integração e usabilidade). Isto é uma motivação, para estes autores, em trabalhar nas minimizações destas preocupações.

Os estudos conduzidos (4) por Knierim, Kiss e Schmidt (2018) os levaram ao desenvolvimento de um aplicativo em *AR* que permite aos alunos do ensino médio, especialmente os de física, observarem dados virtuais representando o fluxo de calor projetado em uma haste metálica, além de um gráfico virtual que mostra as variações de temperatura em função do tempo. Inicialmente, um protótipo foi criado utilizando a ferramenta de apoio ao desenvolvimento de *AR Vuforia*, integrada à plataforma de desenvolvimento *Unity*. Os resultados revelaram um pequeno, mas positivo, efeito da *AR* no desempenho desses alunos em relação ao aprendizado de termodinâmica.

Uma revisão da literatura (5) foi realizada por Ekanayake e Gayanika (2022) para identificar o estado da arte da pesquisa em *AR*, com foco em proporcionar uma melhor experiência pedagógica. O estudo abrangeu todos os níveis de ensino, desde o Ensino Fundamental até a Pós-Graduação, e buscou responder a três questões principais: “Quais são os grupos-alvo mais comuns em que a maioria das pesquisas foi utilizada?”; “Quais são os tipos de aplicativos preferidos para criações lúdicas, destinados à utilização durante as aulas?”;

“Quais são as ferramentas tecnológicas usadas para criar aplicativos de *AR*?”; “Quais são os assuntos/áreas mais comuns que foram abordados?”. Este estudo concluiu que a maioria dos trabalhos contribuem para a inserção de *AR* no Ensino Superior e de Pós-Graduação (65,9%). Com relação aos tipos de dispositivos físicos utilizados, 48,7% dos artigos revisados pelos autores, usam dispositivos móveis, seguidos de 17,03% que usam o *HMD Microsoft Hololens*. Também levantou-se informações sobre as ferramentas para desenvolver as aplicações em *AR*, na área de Educação. A referida revisão concluiu que 40% dos artigos utilizam-se da combinação *Unity* e *Vuforia* para esta finalidade. Por fim, na questão, áreas de assuntos verificou-se que 90% dos artigos relatavam trabalhos para aplicações em disciplinas científicas (matemática, física e química) e o restante para disciplinas não científicas (artes, dança, comércio e economia).

Um protótipo destinado a estudar o suporte da *AR* para a visualização, manipulação e seleção de dados foi apresentado no artigo (6) de Meiguins *et al.* (2019). A técnica aplicada no protótipo consiste em utilizar cartões tangíveis (referenciais) para criar, manipular e interagir com as visualizações. São utilizados 13 cartões divididos em três categorias: cartões de visualizações, responsáveis pela apresentação de sete tipos de visualizações; cartões de funções, que mostram aos usuários as ações que podem ser realizadas; e cartões do grupo de interações, que permitem a interação dos usuários com as visualizações. A combinação de *AR* com cartões mostrou-se uma tecnologia de fácil utilização e compreensão quando bem projetada. No entanto, foram identificadas necessidades de melhorias devido a algumas limitações encontradas, como restrições nas tarefas relacionadas à seleção de objetos virtuais e filtragem de informações virtuais. Os autores constataram que a criação de menus virtuais poderia ser uma possível solução para superar essas limitações. Esses serão os objetivos de estudos futuros dos autores.

Uma pesquisa (7) foi conduzida por Hirve, Kunjir, Shaikh e Shah (2017) sobre a viabilidade de combinar *AR* e *Big Data*. Os autores observam que, com o avanço da tecnologia nos dispositivos móveis, a possibilidade de visualização de *Big Data* apoiada por *AR* tornou-se promissora. Por fim, os autores apresentaram exemplos de aplicações que utilizam a *AR* para a

visualização de dados em diversos campos de negócios, tais como educação, entretenimento, economia, medicina, entre outros.

Foi publicado por Harrington (2020) um artigo (8) que apresenta o *AR Perpetual Garden App*. Trata-se de um aplicativo lançado em 2018 que tem uma finalidade educacional. Este aplicativo oferece interações em ambientes virtuais através de um diorama de uma floresta real com dioramas virtuais desta mesma floresta. As interações ocorrem através de botões para disponibilizações de áudio que reproduz a história da floresta, visualização das informações que se alternam entre duas visualizações: floresta controlada e floresta não controlada. Também é possível obter informações virtuais sobre as plantas presentes na cena. O usuário pode levar o *Ipad* (com o aplicativo instalado) mais próximo da planta virtual para ampliar/reduzir a visão panorâmica das plantas. Também é possível ouvir sons de pássaros que são naturais do tipo da floresta visualizada. Os autores comentam que existem muitos estudos sobre o uso de RA no campo educacional, porém os estudos sobre os resultados reais de aprendizagem são raros. Desta forma, fez-se um estudo sobre o impacto no aprendizado de estudantes. Foram recrutados 56 estudantes para os experimentos. Estes visualizaram informações utilizando o diorama com o uso do aplicativo de *AR* e sem o aplicativo (cartilhas físicas com as mesmas informações do aplicativo). Os testes mostraram não existir diferenças entre o aprendizado real entre as duas abordagens. Entretanto, foram encontradas diferenças significativas, em favor da utilização do aplicativo em *AR*, nas percepções de aprendizagem, nas permanências e no engajamento dos usuários.

No artigo (9) de Chen *et al.* (2020) é apresentado o aplicativo *MARVisT* que tem como objetivo permitir que usuários leigos criem aplicações em *AR*. Assim as visualizações de informações virtuais podem ser personalizadas a critério do usuário. Os autores consideram que plataformas de desenvolvimento de *AR* tradicionais, como a *Unity* não são adequados para usuários leigos. O aplicativo é baseado em glifos que foram nomeados como *ARGVis*. Estes glifos servem como referenciais. Portanto, a ferramenta *MARVisT* é destinada a não especialistas em criações de objetos virtuais. O *MARVisT* é projetado para ser utilizado em dispositivos móveis populares como *smartphones* e *tablets*. Foi realizado um estudo com usuários não experientes em *AR* e/ou visualização de informações para verificar a usabilidade,

facilidade de uso e utilidade do *MARVisT*. Os resultados foram otimistas pois a usabilidade, facilidade e utilidade foram classificadas como bom por 95% dos usuários. Com relação ao engajamento (despertar desejo pela criação) teve uma avaliação positiva de 100% dos participantes. Como sugestões destes participantes apareceram: melhorias nas inicializações, pois às vezes não se sabe por onde começar e colocar glifos em locais diferentes aos oferecidos pela aplicação.

O artigo (10) de Yassien, Emad e Abdennadher (2021) apresenta a ferramenta *CDVVAR (Collaborative Data Visualization)*. Esta permite a visualização de informações, através de gráficos virtuais, tanto com a *AR* como também com a realidade virtual. Nesta o usuário imerge-se no gráfico, naquela são orientados (através de uma linha amarela) a direcionar o dispositivo móvel em direção a uma superfície plana (referencial), para obterem a renderização dos gráficos virtuais. Foi realizado um teste de usabilidade com 24 usuários. Percebeu-se que o tempo de renderização dos resultados é menor com Realidade Virtual, 58% foi esta diferença. Entretanto, a medida que os usuários realizavam mais visualizações, nos dois modos, notou-se que esta diferença de tempo é diminuía. Assim se concluiu que a prática dos usuários levariam os mesmos a terem melhores desempenhos com visualizações utilizando a *AR*.

Os autores Kundu, Muhammad e Sattar (2017) discutem (11) a utilização da *AR* nas aulas de Geociências. Propõe-se a construção do chamado '*SandBox*' para alcançar um aprendizado avançado em Geociência. Desenvolvido na Universidade da *Califórnia*, o *SandBox* é composto por uma caixa de areia onde a topografia pode ser moldada de acordo com o assunto que será exposto em aula. Em seguida, essa moldura é visualizada em tempo real por meio da *AR*. Além da caixa de areia, o sistema é complementado por uma câmera *Kinect 3D* posicionada acima da caixa de areia, que detecta a profundidade (distância até a areia). Essas imagens são enviadas ao computador hospedeiro do *software* responsável por codificar os contornos e cores. Posteriormente, são transmitidas de volta a um projetor de tela que lança a imagem de volta à caixa de areia. Todas as visualizações são realizadas com interações dos usuários. O estudo revelou um profundo potencial para o aprendizado, reforçando a compreensão e o raciocínio no processo de aprendizado.



No artigo (12) publicado por Shaikh *et al* (2019) é apresentado o *Coordinate*, uma ferramenta para sobreposições virtuais em *AR* de mapas utilizando aparelhos móveis (*IOS*, *Android*, *Hololens*). Esta ferramenta oferece a possibilidade de usuários leigos programarem as sobreposições virtuais em *AR* sobre mapas (acessando uma conta *web*). Desta forma múltiplos usuários poderão obterem visualizações virtuais, diferentes entre si, deste mapa (com os objetos virtuais que foram inseridos por estes usuários) através de seus próprios aparelhos móveis. O próprio mapa físico é o referencial para estes dispositivos. O projeto do *Coordinate* é composto, inicialmente, por uma interface *web* (o usuário acessa sua conta). Esta interface apresenta uma planilha onde cada usuário poderá programar suas apresentações em *AR*. Os resultados são diferentes visões virtuais e reais dos dados e experiência, opcionalmente, compartilhada.

O artigo (13) de Reipschlagel, Flemisch e Dachselt (2021) relata o uso de uma aplicação (protótipo) de *AR* com o uso de *HMD's* em conjunto com telas grandes físicas para análises de dados, unindo dados reais apresentados em telas físicas com dados virtuais. Essas mesmas telas físicas são os referenciais para a aplicação em *AR*. As intenções desta união são as questões de percepções (visão periférica da tela, distorção devido a ângulos de visão e longas distâncias); gerências de densidades e complexidades (espaço apropriado e falta de espaço); suporte a múltiplos usuários (abstração, distração e visualizações pessoais). Portanto, o objetivo principal com este trabalho é enfrentar estes 3 desafios. Criou-se um protótipo (*framework*) que permite as visualizações simultâneas: na forma tradicional (*displays* físicos) e em 3D (*AR*) que fica livre no espaço, visualizada por *HMD's*. Para os testes, as visualizações de dados ocorreram utilizando uma base de dados de filmes. Estas visualizações ocorreram em relação à duração dos filmes e como ocorreram mudanças nestas durações durante o passar dos anos. Então, os analistas puderam analisar as informações através de telas físicas apoiadas por telas virtuais (na frente, em cima, ao lado e abaixo das telas físicas). Assim, ficou constatado que esta abordagem traz muitos benefícios. Comparando com as tradicionais formas de visualizações, verificou-se um alto desempenho na análise de dados.

O artigo (14) de Narman *et al.* (2020) apresenta um protótipo, em *AR*, que visa mostrar conteúdos para aulas práticas de estrutura de dados. As contribuições esperadas são:

desenvolver um aplicativo que permita apresentações de estrutura de dados com o uso de *AR* proporcionando além desta apresentação a interatividade e analisar os efeitos da apresentação de *AR* comparando-a com as maneiras tradicionais de apresentação (lousa e *VisuAlg*). No final das comparações foi constatado que 54% dos alunos gostaram mais do uso de *AR*, 84% consideraram o aprendizado de estrutura de dados mais interessante e proveitoso com *AR* e 38% informaram entender melhor com a apresentação em *AR*. Portanto concluiu-se que o uso de *AR* é mais interessante aos alunos de estrutura de dados. A *AR* pode ser mais uma ajuda para explicar assuntos complexos.

Os autores Zhang *et al.* (2018) publicaram um artigo (15) que discute o uso do *Microsoft HoloLens* para projetar com *AR* dados virtuais sobre o mapa físico da cidade de Toronto, no Canadá. Segundo estes autores o *Microsoft HoloLens* integra todos os componentes necessários para se tornar um sistema em RA completo. Ou seja, diferente de seus concorrentes, daquela época, não precisa estar conectado com máquinas externas. O *HoloLens* utiliza o Sistema Operacional *Windows 10*. Usou-se como exemplo visualizações de edifícios daquela cidade, tendo como referencial um mapa físico. Após o usuário olhar para uma localização, com gestos do dedo no ar o usuário seleciona as informações (em um menu virtual) que serão sobrepostas sobre o edifício escolhido. O edifício que estiver sendo analisado terá uma projeção (holograma) na cor vermelha (para destacar-se aos demais edifícios presentes no mapa). Foi realizada uma comparação com 5 voluntários executando as mesmas tarefas com visualizações destes dados em *AR* e em *laptops*. As referidas tarefas quando executadas com RA obtiveram as preferências destes voluntários.

Foi publicado por Sicut *et al.* (2019) um artigo (16) que descreve o *DXR* – *Development Reality Extended*. O *DXR* é baseado na plataforma de desenvolvimento *Unity* (é uma biblioteca a ser adicionada) e é voltado à construção de aplicativos para visualizações de dados para realidade estendida (*AR* + Realidade Virtual). O *DXR* fornece uma interface em alto nível para construir (coordenadas  $x, y, z$  – cores – posição – tamanho) e adaptar visualizações configuradas. Possui integração com o *vega-lite* (linguagem declarativa para a criação de visualizações de dados interativas). Este aplicativo é de código aberto e está disponível para *download* gratuito em ([DXR \(google.com\)](http://DXR.google.com)). O *DXR* utiliza uma coleção de

objetos da Unity cujas propriedades como posição, cor, tamanho e outras, são mapeadas por atributos de dados. Esse mapeamento pode ocorrer em tempo de execução ou em tempo de programação, ambos pela interface *GUI (Graphical User Interface)*. O *DXR* consiste em *prefabs* (objetos criados anteriormente) que fornecem uma interface em alto nível para construir *GameObjects* (objetos virtuais disponíveis na *Unity* ou em alguma das bibliotecas integradas ao *Unity*) orientados a dados em uma cena *Unity*. Uma visualização em *DXR* é representada como um *prefab*, os quais são chamados de *vis-prefabs* que são adicionados por meio da *Unity* como quaisquer outros *gameObject*. O *vis-prefab*, adicionado lê o mapeamento através de um arquivo *json*. A filosofia do *DXR* é que usuários sem experiência também possam criar as visualizações usando a *Unity*. O *DXR* encontra-se em produção, é gratuito e possui vários exemplos prontos.

O artigo (17) de Bermejo *et al.* (2017) explora o potencial de valor para a convergência entre *AR* e *Big Data*. Os autores comentam que para as pessoas em geral é mais fácil entender e interagir com as informações, quando elas são sobrepostas em forma de conteúdo virtual e têm visão do mundo ao nosso redor. Desta forma, precisamos repensar como misturar os dados digitais com o mundo físico para apresentar as informações aos usuários. Entretanto, existem ainda diversas barreiras como por exemplo a exibição intrusiva, duração de baterias e dados fragmentados. No entanto, existem esforços para superar estas limitações. Uma delas é o uso de *AR Markup Language (ARML)*, um padrão de dados para descrever e interagir com cenas de *AR* desenvolvido pelo *Open Geospatial Consortium - OGC*. Consiste em uma gramática *xml (Extensible Markup Language)* para descrever a localização e a aparência de objetos virtuais.

No artigo (18) de Dani *et al.* (2018), os autores discutem uma maneira mais econômica para visualização de dados com o uso de *AR* e Realidade Virtual em conjunto (a literatura denomina esta combinação de *Mixed Reality - MR*), cujas interações são realizadas através de gestos (dos dedos). Segundo estes autores, os trabalhos atuais (em sua maioria) são realizados com dispositivos de *hardware* proprietários que possuem preços exorbitantes. Desta forma, propõem-se uma solução também com interações através de gestos, mas utilizando dispositivos mais econômicos. Alguns exemplos são o *Google Carboard* e o *VRBOX* (óculos

de realidade virtual). Este artigo apresenta um *framework* para visualização de dados contextuais através de *HMD's* (mas também é possível utilizar em *smartphone android*). Além disso apresenta uma forma de interação com o usuário por meio de gestos das mãos (dedos). Para localização das pontas dos dedos o dispositivo móvel envia quadros de vídeo para um servidor. Este calcula as coordenadas para as pontas dos dedos e envia de volta ao dispositivo móvel. Foi realizado uma avaliação com 30 usuários (visualização e interação). A classificação média foi de 4.1 e 4.5 respectivamente. Entretanto alguns usuários consideraram cansativa o uso do aplicativo, mas a maioria respondeu terem gostado e apreciaram a ideia de visualização e interação com o aplicativo.

É proposto, no artigo (19) de Ohta, Ito e Kakuda (2017), o projeto de um *software* para visualizar a topologia de redes e os estados dos nós desta rede com o apoio da *AR*. Para testar a viabilidade do projeto, foi apresentado um protótipo com três nós móveis, os quais estavam relacionados a marcadores para identificação dos mesmos. Existe um *software* sendo executado por um *tablet* e os *status* dos nós e a topologia da rede (em *AR*) é visualizada pela câmera deste *tablet*. Cada nó da rede envia uma mensagem de *log* para este *software*, por outra conexão. Linhas (vermelha e azul) virtuais são sobrepostas para identificar as conexões e círculos brancos mostram quem está enviando mensagens no momento. Além disso, informações a respeito dos dispositivos, como por exemplo o *Internet Protocol - IP* dos mesmos foram sobrepostos virtualmente no ambiente real.

Através do artigo (20) de Xin *et al.* (2020), é apresentado o desenvolvimento e a execução de um *software em AR*, em um dispositivo móvel *Android*, o qual obteve sucesso como guia de um salão de exposição. Dez imagens diferente foram utilizadas nos testes. Estes testes obtiveram precisão de 100% nos resultados de execução do referido aplicativo. Com relação às interações pode-se utilizar gestos (dedos) para ampliar/reduzir e girar os 10 modelo em qualquer área. Também obteve 100% de eficiência no rastreamento (informações de coordenadas precisas em tempo real incluindo o mundo real através do dispositivo de aquisição de imagens para renderizações da cena virtual no mundo real) e registro (posicionamento e rastreamento da cena virtual e do mundo real.) dos objetos virtuais.

Propõe-se, no artigo (21) de Shepherd, Kraft e Francis (2019), o uso de *AR* para apoiar a programação de robôs industriais, a fim de reduzir custos de programação. Através da *AR*, os locais e caminhos que os robôs executam podem ser visualizados juntos. Na visualização tradicional, o robô é movido para cada local um a um. Assim, a principal limitação é que apenas um local por vez pode ser mostrado. Sendo assim, propõe-se uma sobreposição virtual dos caminhos e locais dos robôs, colocando-os neste espaço virtual. Isso permite interações com os usuários. Um caso de uso foi realizado, onde um robô físico é inserido no ambiente virtual e comandos de movimentos são adicionados a ele. Estes movimentos vão sendo armazenados para posterior uso no ambiente real. No final verificou-se que as variáveis de localizações e caminhos (*AR*) foram visualizadas com sucesso. Assim, os usuários puderam interagir com os robôs em ambiente real com segurança.

O artigo (22) de Guarese *et al.* (2020), apresentou um aplicativo de *AR* que deve ser instalado no *HMD Microsoft HoloLens* para visualização de dados e interações com os mesmos (através de gestos). O objetivo é sobrepor informações virtuais para a tomada de decisões. Um estudo de caso utilizou uma sala de reuniões onde cada cadeira possui uma pontuação com relação à sua posição. O usuário pode, através da interface em *AR*, alterar as pontuações dos assentos quando comparados a várias dimensões como temperatura, proximidade do ar-condicionado, proximidade da saída, tomadas, *wifi*, etc. Para efeitos de comparação, as mesmas tarefas foram realizadas em telas 2D, como *tablets*. Analisando o desempenho dos usuários (16), os testes mostraram que com a aplicação em *AR*, os usuários caminharam 12,7% menos. A precisão foi de 84,37% para 2D e 86,46% para a aplicação em *AR*. Depois disso, os usuários deram notas, sendo ambas elevadas, 2D = 85,9 e *AR* = 73,8. Como resultado, ambos aplicativos tiveram resultados satisfatórios, pois as tarefas foram concluídas com 84% de acerto. Portanto, o objetivo principal do trabalho foi alcançado: validar o uso de *AR* para processos de análise de dados.

Uma análise (23) das edições de 1997 até 2019 do Simpósio de Realidade Virtual e Aumentada (SVR) é apresentada através do artigo de Teixeira e Peres (2020). Os autores apresentaram o que foi desenvolvido, o que está sendo desenvolvido e o que esperar do futuro em *AR* e realidade virtual. Segundo os autores, a maioria das publicações do *SVR* está focada

no uso das tecnologias para resolver problemas aplicados e menos em novas tecnologias. Os temas mais frequentes nos eventos são: A) sistemas – estruturas – kits de ferramentas com 41,65%; B) impactos sociais/econômicos/técnicos de *AR* e realidade virtual com 11,49%; C) técnicas de computação gráfica, que incluem renderizações, visualizações e simulações com 10,77%; D) uso de aplicações em redes multiusuários e distribuídas com 8,98%; E) interações 3D com os usuários e interações espaciais com 7,9%; F) representação virtual de usuários com 5,39%; G) rastreamento/detecção 4,49%; H) dispositivos para entrada/saída para fins de uso em *AR* e realidade virtual, 2,3%; I) tato/áudio/outras interfaces 1,97%; K) telepresença 1,49%. Com esta pesquisa, os autores concluíram que essas tecnologias são consideradas maduras e prontas para solucionar problemas relacionados à visualização com *AR* e Realidade Virtual.

O artigo (24) de Elrawi (2017) mostra um esboço sobre a possibilidade da criação de um sistema que permita aos visitantes visualizarem informações históricas de um museu islâmico. Entretanto, este sistema disponibiliza as opções de visualização por meio de *AR*, realidade virtual ou *desktop*. A intenção é que os visitantes possam ter visualizações de personagens e eventos do passado oriundos de vários arquivos. O autor comenta que o principal objetivo da visualização é trazer a compreensão dos dados. A tarefa é apresentar informações complexas da maneira mais abrangente. Considerando artefatos arquitetônicos, o processo de visualização é principalmente focado na compreensão das relações espaciais e no reconhecimento de um estilo e formas particulares. A maneira mais natural de transmitir essas informações é construir um modelo tridimensional. Este autor apresentou vários exemplos de aplicações de *AR* e realidade virtual já em uso em outros locais de demonstrações históricas. Então, o mesmo salienta que a abordagem em *AR* elimina a necessidade de ter controles separados para ver as coisas e filtra as distrações do mundo real. Sobre as experiências dos usuários, é importante ressaltar que estes consigam olhar em volta de forma totalmente natural, sem usar nenhum controle. É uma abordagem intuitiva para descobrir a história oculta de um fragmento do patrimônio cultural.

No trabalho (25) de Sjobergh e Tanaka, é discutido a utilização de um sistema para visualização de dados de diferentes maneiras pelas quais os valores ausentes podem ser visualizados em uma estrutura interativa de exploração de dados. Para isso, o sistema

apresenta visualizações de dados ausentes em diferentes gráficos, por exemplo, gráficos de barras e gráficos de dispersão, onde os dados ausentes possuem uma coloração diferente.

O artigo (26) de Krause-Glau et al. (2022) apresenta o *ExplorViz*. Trata-se de uma implementação para visualização colaborativa para o desenvolvimento de programas de computação (*Unified Modeling Language – UML*). Nesta ferramenta o usuário escolhe em utilizar entre 3 modos: *desktop*, *AR* e realidade virtual. O *ExplorViz* se baseia em uma estrutura composta por telefones celulares e marcadores visuais para permitir a exploração colaborativa de visualizações de *softwares* em *AR*. Segundo os autores o *ExplorViz* é uma alternativa ao *IslandViz*. O *IslandViz* usa *hardware* especializado e caro para *AR*. O *ExplorViz* fornece aos usuários um modelo de paisagem, que fornece uma visão geral do sistema de *software*, e modelos de aplicativos adequados para uma exploração de *software* mais detalhada. Esses modelos podem ser explorados independentemente um do outro, pois podem ser colocados em diferentes marcadores. Os autores concluem que *ExplorViz* permite ótimas visualizações inspiradas em diagramas *UML* de maneira compartilhada (em locais diferentes).

### **3.2. Trabalhos relacionados**

Nesta seção, apresentamos alguns trabalhos anteriores sobre tratamento de dados ausentes, que se aproximam da nossa estratégia. A Tabela 4 resume os trabalhos selecionados, destacando suas metodologias de tratamento de dados ausentes para fins de comparações com a nossa estratégia. Posteriormente, são fornecidas explicações detalhadas, identificadas pelo número da linha, de forma resumida para cada linha desta tabela, oferecendo uma análise das estratégias aplicadas nesses trabalhos de tratamento de dados ausentes. Na próxima subseção apresentamos as comparações.

Tabela 4 – Trabalhos relacionados

Nº	Publicação	Autor/ano	Objetivos
1	<i>A Method for Classification of Missing Values using Data Mining Techniques</i>	Panda e Adhikari 2020	Propor o modelo de método de atribuição eficaz baseado na atribuição de valor ausente usando correlação.
2	<i>Using data mining to handle missing data in multi-hop sensor network applications</i>	Gruenwald et.al. 2010	Descreve uma abordagem chamada <i>MASTER-M</i> para lidar com o tratamento de dados ausentes oriundos de sensores <i>multi-hop</i> .
3	<i>Principled MissingData Treatments</i>	Lang e Little 2016	Comparar diversos métodos de tratamento de dados ausentes. É estudado as vantagens e desvantagens entre eles.
4	<i>A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms</i>	Makaba e Dogo 2019	Realizar uma análise comparativa e avaliar o impacto de quatro métodos de tratamento de dados ausentes em relação a seis algoritmos classificadores de aprendizagem de máquina.
5	<i>Imputation Analysis of Central Tendencies for Classification</i>	Pavithrakannan <i>et.al.</i> 2021	Os autores propõem um projeto de aprendizagem de máquina para a análise do impacto do uso de cada tendência central para diferentes distribuições de dados.
6	<i>Visualizing Missing Values</i>	Sjobergh e Tanaka 2017/IEEE	Construir o sistema de visualização 2D para análise exploratória de conjuntos de dados que possuem dados ausentes. Diferentes visualizações dos dados.

Em um artigo (1) os autores propõem a utilização de um modelo de método de imputação baseado na atribuição de valor ausente usando correlação. Segundo os autores, o método proposto lida com os valores ausentes durante a classificação, aumentando a precisão,



em tempo real. A correlação mede o grau e a direção da relação linear entre duas variáveis. Ela varia de -1 a 1, onde -1 indica uma correlação negativa perfeita, 1 indica uma correlação positiva perfeita e quanto mais próximo a 0 indica correlação fraca ou ausência de correlação.

Os autores apresentam uma abordagem (2) para o tratamento de dados ausentes em base de dados originados de sensores *multi-hop* (onde os dados dos sensores chegam à base de dados por meio de outros sensores). O algoritmo consiste em três etapas: agrupamento de sensores *online*; captura de regras de associação entre sensores dentro de cada *cluster* e estimativa dos valores dos dados ausentes usando as regras de associação obtidas. O valor da distância inicial é calculado usando as primeiras rodadas de dados, e o valor da distância consequente é atualizado incrementalmente, quando ocorrem mudanças nos fenômenos. As associações são realizadas através dos cálculos de suporte e confiança. Por fim, as estimativas são realizadas com base nestas associações.

A publicação (3) traz uma revisão dos principais métodos de tratamento de dados ausentes, classificando os mesmos entre métodos ultrapassados e métodos modernos. Os autores argumentam que pesquisadores devem utilizar os métodos de tratamento de dados modernos. Entre estes eles destacam o método de tratamento de dados ausentes por imputação múltipla e o de máxima verossimilhança com informação completa. Estes métodos são enfatizados como superiores, contrastando-os com os métodos ultrapassados como, por exemplo, técnicas do método baseado em exclusão e técnicas do método de imputação única, apesar de serem muito utilizados atualmente. Segundo estes autores, a principal vantagem dos dois métodos, considerados modernos, é incluir os valores imputados que refletirão de forma mais geral possível, os verdadeiros padrões dos dados, em vez de amplificarem falsamente as associações hipotéticas. A desvantagem é o custo computacional.

No artigo (4) apresenta-se uma análise comparativa entre quatro métodos de tratamento de dados ausentes em relação a seis algoritmos classificadores de aprendizado de máquina. Os autores conduziram um estudo no qual compararam os seguintes métodos de tratamento de dados ausentes: média/moda, k-vizinhos (*KNN*), maximização da expectativa (*EM*) e múltiplas imputações por equações encadeadas (*MICE*), em relação a seis algoritmos classificadores de aprendizado de máquina: regressão logística (*LR*), *KNN*, máquina de vetores de suporte

(*SVM*), floresta aleatória (*RF*), naive bayes e redes neurais artificiais. No que concerne aos dados numéricos, geralmente todos os classificadores apresentaram bom desempenho. Em relação aos dados categóricos, observou-se que o naive bayes demonstrou ter um desempenho ligeiramente superior nas estratégias modais utilizadas, em comparação com os outros classificadores. De maneira geral, os autores observaram que os resultados obtidos variaram de acordo com o classificador, o tipo de dado (numérico ou categórico) e o percentual de dados ausentes. Isso indica que nenhum método de tratamento de dados ausentes é superior ou se adapta a todos os problemas de tipos de conjuntos de dados.

Um projeto de aprendizado de máquina é apresentado (5), com o objetivo de analisar o desempenho de técnicas de imputação por tendência central em arquivos de entrada. Este arquivo de entrada inclui detalhes como o nome do arquivo e informações sobre o conjunto de dados necessárias para treinar o modelo de aprendizado de máquina. Em seguida, o arquivo é analisado e os dados ausentes são imputados com valores obtidos a partir de medidas de tendência central. Três novos arquivos são gerados, sem dados ausentes, sendo imputados com medidas de tendências centrais diferentes. Posteriormente, esses arquivos são submetidos aos modelos de aprendizado de máquina para treinamento e obtenção das métricas de desempenho. Todos os três arquivos são treinados por diferentes algoritmos de aprendizagem de máquina, tais como k-vizinhos e regressão logística. Segundo os autores, treinar e avaliar os dados com esses diferentes algoritmos, cada um utilizando métodos distintos de previsão, auxiliará na compreensão de qual tendência central apresenta melhor desempenho em qual algoritmo.

No artigo (6), é apresentado o funcionamento de um sistema no qual os dados ausentes são visualizados por meio de diferentes tipos de gráficos 2D, tais como gráficos de barras, histogramas e gráficos de dispersão, dispostos lado a lado. Os dados ausentes são destacados nos gráficos por meio de colorações distintas, proporcionando ao usuário uma identificação clara e precisa da localização e mensuração desses dados. Por exemplo, as mensurações de dados ausentes podem ser representadas pela coloração vermelha, enquanto as mensurações de dados presentes são destacadas em azul. Essa abordagem facilita a compreensão e análise dos



Comparações de desempenho entre métodos	Não	Não	Sim	Sim	Sim	Não	Não
---	-----	-----	-----	-----	-----	-----	-----

### 3.3. Conclusão da revisão da literatura

Através dos artigos selecionados e descritos nesta revisão literária concluímos que a *AR* pode apoiar a análise e tratamento de dados. Estes artigos retornam trabalhos de pesquisas em diversas áreas (medicina, engenharias, educação, etc) que utilizam a *AR* como ferramenta complementar ou primária para apoiar a visualização e análise dos dados.

Entretanto, com relação à pergunta específica: “A Realidade Aumentada com o apoio de uma ferramenta de *LLM* podem auxiliar no tratamento de dados ausentes?”, não localizou-se nenhum trabalho abordando exatamente este tema. Isso nos motivou a pesquisar a viabilidade de aplicar as técnicas e métodos de *AR* e tratamento de dados utilizados nesses artigos.

## 4. Estratégia para o tratamento dos dados ausentes

Este capítulo apresenta a principal contribuição desta pesquisa, que é uma estratégia que auxilia o tratamento de dados ausentes com apoio de visualizações com *AR* e com o auxílio de uma ferramenta *LLM*. Esta estratégia foi desenvolvida principalmente para ajudar analistas de dados. Representa uma síntese de conhecimentos provenientes de diversas áreas, incluindo banco de dados, visualização de dados, tratamento de dados ausentes, *AR* e *LLM*, todas convergindo para uma solução que tem o potencial de ser uma alternativa na forma de abordar o tratamento de dados ausentes.

Inicialmente, este capítulo oferece uma visão geral da estratégia, expondo os princípios fundamentais que norteiam a abordagem. Em seguida, detalharemos os componentes que a compõem, permitindo um tratamento eficaz e preciso.

### 4.1. Introdução à estratégia

A estrutura da estratégia foi desenvolvida com o propósito de fornecer recursos para a análise exploratória de dados e assim facilitar a escolha do método de tratamento de dados ausentes. Assim, seus principais objetivos são identificar, localizar e mensurar os dados ausentes. Além disso, facilitar a escolha de métodos para tratamento de outras anomalias nos dados, como, por exemplo, os *outliers* e inconsistência de dados. Desta forma, destaca-se a importância desta estratégia como uma possível resposta à pergunta norteadora introduzida no contexto desta dissertação: “A Realidade Aumentada com o apoio de uma ferramenta de *LLM* podem auxiliar no tratamento de dados ausentes?”.

### 4.2. Visão geral da estratégia

A estratégia desenvolvida utiliza visualizações virtuais de dados por meio da *AR* e métodos de tratamento de dados ausentes que são acionados por objetos virtuais. Essas visualizações de dados e o tratamento dos dados são executados através de códigos-fonte gerados totalmente ou parcialmente por uma ferramenta *LLM*. As visualizações são geradas de forma a representar os dados de forma interativa e imersiva. Os métodos de tratamento dos dados, quando acionados por objetos virtuais, permitem a interação com os dados durante a

visualização. Os códigos-fonte, após serem adaptados ao nosso contexto, são vinculados aos objetos virtuais, permitindo a execução das funcionalidades de visualização e tratamento de dados durante a interação com esses objetos.

A estratégia consiste em um ciclo, que inicia-se a partir de um determinado banco de dados, mas pode estender-se a outros tipos de fonte de dados como planilhas eletrônicas e arquivos texto. Este ciclo é composto de dois subciclos.

No primeiro subciclo, denominado Análise de dados, são realizadas as análises visuais dos dados para identificar a presença de dados ausentes, desvios de padrões nos dados (*outliers*), dados duplicados, erros de formatação e outras exceções, no referido banco de dados. Neste subciclo realiza-se uma análise exploratória através de visualizações dos dados em gráficos virtuais em *AR*. Essa abordagem permite uma exploração dos dados e oferece uma perspectiva visual imersiva.

O segundo subciclo, denominado Tratamento de dados, concentra-se na manipulação dos dados, proporcionando recursos para a seleção dos métodos de tratamento de dados, através de um menu virtual em *AR*, a serem aplicados na correção do conjunto de dados. Escolher o método de preenchimento de dados ausentes é um pré-requisito significativo para melhorar a qualidade dos dados. Portanto, nesse subciclo está incluído recursos para facilitar esta escolha, como por exemplo a apresentação dos tipos de dados. Segundo, Othman e Yahia (2018), é importante considerar os tipos dos dados, a quantidade de dados ausentes e o impacto potencial nas análises e nos resultados, antes de aplicar o método de tratamento.

Caso as opções de visualizações e tratamento de dados definidas não abranger as preferências do usuário, ele tem a opção de incluir novos gráficos virtuais e/ou novos métodos de tratamento de dados. Desta forma, o mesmo poderá solicitar à ferramenta *LLM* ajuda para a geração de códigos-fontes e vincular estes códigos-fontes aos novos objetos virtuais.

Ambos os subciclos são executados até que as condições de visualizações e/ou tratamento de dados ausentes estejam satisfeitas. Portanto, a estratégia visa melhorar a

eficiência de visualizações de dados e tratamento de dados ausentes, aproveitando tecnologias disponíveis. Conforme a Figura 11 a estratégia é constituída dos seguintes passos:

- 1) Inicia-se o primeiro subciclo, Análise dos dados, quando o usuário aponta um cartão físico para a câmera do dispositivo. Assim os dados são apresentados em um gráfico virtual, permitindo ao usuário a análise dos dados (localização, mensuração e características dos dados);
- 2) O usuário decide entre gerar novas visualizações ou seguir para o módulo de Tratamento de dados;
- 3) Um novo gráfico virtual é gerado e volta-se ao passo 1;
- 4) É realizado, novamente, visualização dos dados;
- 5) É apresentado ao usuário o menu virtual com os métodos de tratamento de dados disponíveis;
- 6) Opcionalmente, o usuário poderá verificar as características dos dados, como por exemplo os tipos dos dados, padrões e relações entre os dados. Desta forma, o usuário terá orientação para a escolha do método de tratamento mais apropriado;
- 7) O usuário decide em aplicar o método de tratamento de dados selecionado. Se sim o tratamento de dados é aplicado e aplicação é encerrada;
- 8) Caso o usuário opte em não confirmar a aplicação do referido método de tratamento de dados, significa que o mesmo não deseja utilizar nenhum dos métodos disponíveis, então solicita à ferramenta *LLM* novo(s) método(s) de tratamento de dados;
- 9) O(s) novo(s) método(s) de tratamento de dados gerados, pela ferramenta *LLM*, é vinculado a um novo cartão físico e volta-se ao passo 1;
- 10) O ciclo é reiniciado e é executado até que o usuário confirme a aplicação do método escolhido (passo 7) e a aplicação é encerrada.

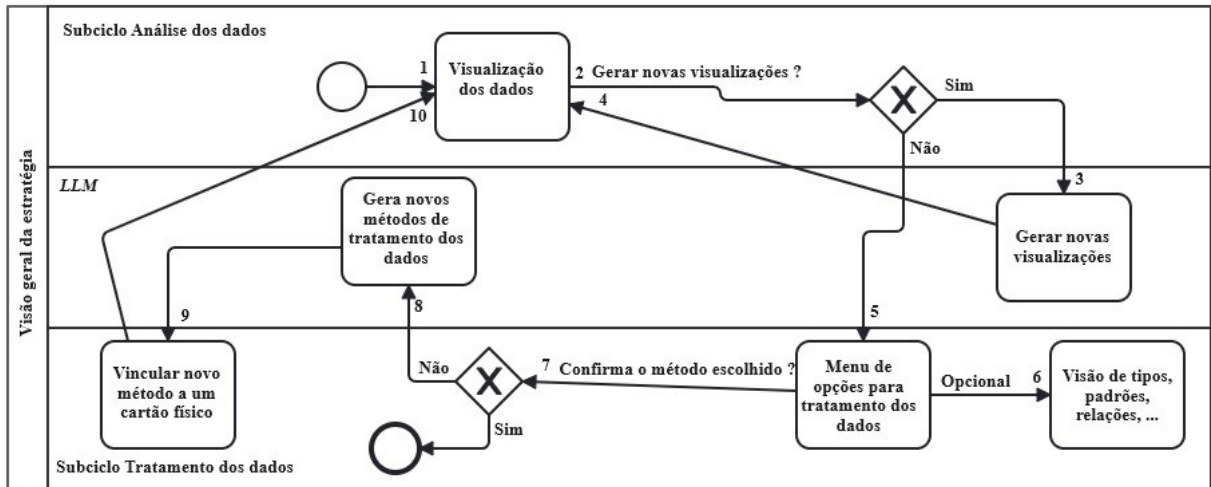


Figura 11: Visão geral da estratégia

#### 4.2.1. Subciclo Análise de dados

O objetivo deste subciclo é oferecer visualizações, através de gráficos virtuais em *AR*, para identificação, avaliação, localização e gerenciamento de valores ausentes em um determinado banco de dados. Entretanto, com esta estratégia, é possível visualizar outros tipos de anomalias nos dados, além de dados ausentes. Conforme (Nunes, Kluck e Fachel, 2009), informações sobre os dados ausentes, em um conjunto de dados, é uma etapa crítica na preparação e análise de dados, pois a presença de dados ausentes pode afetar significativamente a qualidade e a validade das análises estatísticas e dos modelos de dados.

Este subciclo inicia quando o usuário aponta um determinado cartão físico, para a câmera do dispositivo físico. Cada um dos cartões se refere a uma consulta no banco de dados, previamente construída. Em seguida, as informações geradas com a referida consulta são transferidas para um gráfico 3D virtual, gerado com tecnologia de *AR*. As informações visualizadas são, por exemplo: total de dados, total de dados ausentes e a porcentagem de dados ausentes em relação ao total de dados. Assim é possível mensurar e localizar os dados ausentes.

Caso o usuário esteja satisfeito com a visualização ele pode encerrar este subciclo, ou solicitar à ferramenta *LLM* novo gráfico virtual. Este novo gráfico virtual será vinculado a um novo cartão físico, reiniciando o subciclo.



Quando satisfeitas as análises dos dados, o usuário, caso assim desejar, pode passar para o próximo subciclo: Tratamento dos dados. A Figura 12 mostra o subciclo Análise de dados.

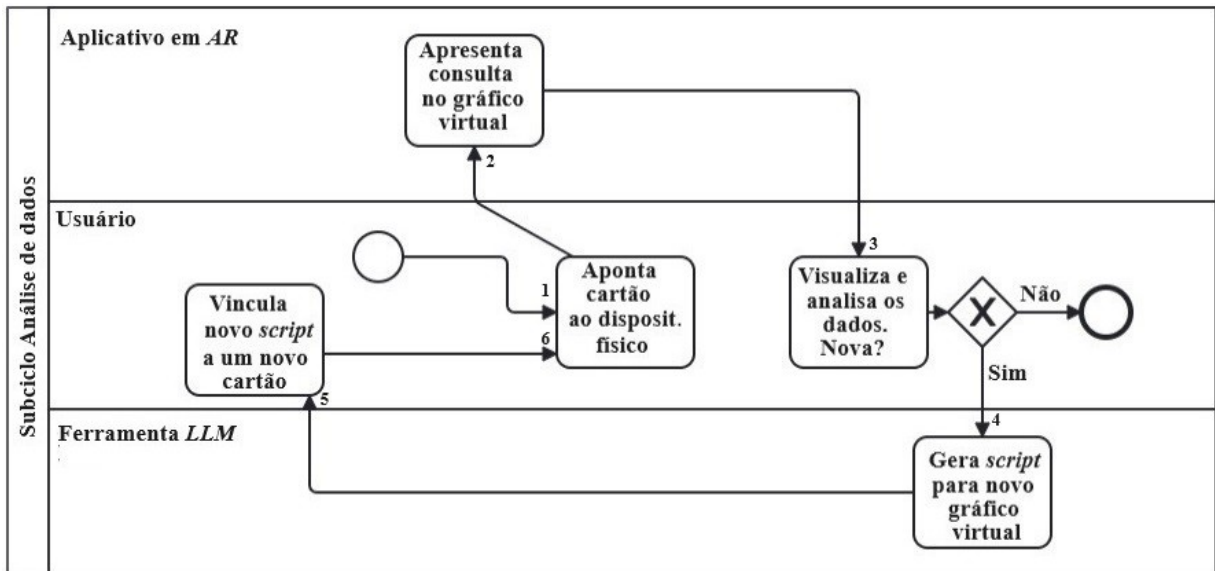


Figura 12: Subciclo Análise de dados

#### 4.2.2. Subciclo Tratamento dos dados

Após concluídas as análises dos dados no subciclo anterior, é iniciado o tratamento de dados. Este subciclo é iniciado quando o usuário, através de um menu virtual em *AR*, seleciona o método de tratamento de dados ausentes desejado.

Antes do usuário escolher o método de tratamento de dados, o usuário opcionalmente, neste menu virtual, pode visualizar as características dos dados, como por exemplo os tipos destes dados, relacionamentos entre os dados e padrões de ausências. Assim o mesmo terá informações necessárias para facilitar escolha do método de dados mais adequado.

Após a escolha do método de tratamento de dados ausentes, um artefato dedicado à execução do método correspondente é ativado. Esse artefato, além de executar o tratamento dos dados ausentes, também atualiza o banco de dados. A comunicação entre este artefato e o banco de dados pode ser realizada por qualquer protocolo compatível com transmissão de dados.

Caso o usuário esteja satisfeito com o tratamento de dados aplicado aos dados ele pode encerrar o subciclo, confirmando o tratamento realizado, ou, solicitar à ferramenta *LLM* um novo método de tratamento de dados. Então esta ferramenta gera o *script* para o novo(s) método(s). Na sequência o usuário insere o(s) novo(s) método(s) no menu virtual e o subciclo reinicia. A Figura 13 ilustra este subciclo.

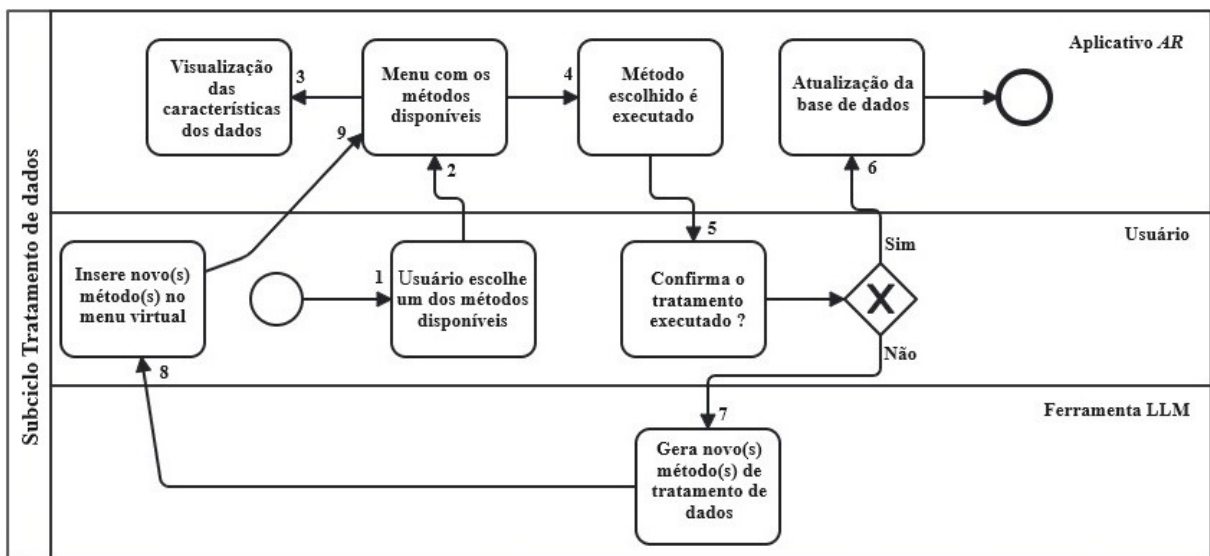


Figura 13: Subciclo Tratamento de dados

Neste subciclo, além do acesso às alternativas entre os métodos de tratamento de dados ausentes, pode-se aplicar métodos para tratamento de outros tipos de anomalias nos dados, como outliers e dados com valores inapropriados.

## 5. O aplicativo de Realidade Aumentada

O presente capítulo apresenta o *DMT-AR* (*Data Missing Treatment with Augmented Reality*), um aplicativo de *AR* implementado para concretizar a estratégia descrita no capítulo anterior. O *DMT-AR* representa uma abordagem dinâmica para identificar, corrigir e preencher as lacunas nos dados. Além do tratamento de dados ausentes, o *DMT-AR* também abrange diversas funcionalidades essenciais para a concretização da nossa estratégia, tais como acesso a bases de dados, visualizações e análises dos dados em tempo real de um determinado conjunto de dados.

Este aplicativo incorpora funcionalidades que foram desenvolvidas por meio de interações com uma ferramenta *LLM*, sendo o *ChatGPT* esta ferramenta utilizada para gerar parcialmente ou integralmente os códigos-fonte.

Dentre esses códigos-fonte, destacam-se arquivos dedicados à geração de visualizações de dados através de gráficos em *AR*, códigos-fonte específicos voltados ao tratamento de dados ausentes e à atualização da base de dados. Os códigos-fonte resultantes das interações com o *ChatGPT* foram ajustados para se adequarem ao nosso contexto e foram associados aos objetos virtuais, com o objetivo de fornecer visualizações e realizar o tratamento dos dados. Esta abordagem colaborativa entre o desenvolvedor e o *ChatGPT* tem como finalidade implementar de forma eficiente e personalizada as operações de visualização e manipulação de dados, o que contribui para a robustez e versatilidade do aplicativo.

Utilizamos a *Unity* para a geração dos objetos virtuais. Esta é uma plataforma de desenvolvimento de jogos e aplicativos de *AR*. A preferência pela *Unity* surgiu devido à disponibilidade da licença gratuita conhecida como "*Unity Personal*", que abrange recursos completos para o processo de desenvolvimento. Esta plataforma permite criar experiências interativas em 2D, 3D e *AR*. Para isso, a *Unity* integra o *Vuforia* e a linguagem de programação *C#*. O *Vuforia* é responsável por detectar e rastrear objetos físicos e, em seguida sobrepõe os elementos virtuais criados no ambiente, como gráficos, animações, informações contextuais e outros objetos 3D ([Vuforia Engine 10 API | Vuforia Library](#)), criando uma fusão entre

o ambiente físico e virtual através da câmara de dispositivos. A linguagem de programação *C#* permite que as interações com os objetos virtuais ocorram.

### **5.1. Funcionamento do aplicativo**

O *DMT-AR* foi desenvolvido em dois módulos distintos, cada um desempenha de maneira especializada um dos subciclos mencionado na descrição da estratégia (capítulo 4), proporcionando uma organização clara e modularizada para a execução precisa e sequencial das operações de análise e tratamento dos dados. Essa divisão funcional visa otimizar a eficácia dele na consecução de seus objetivos.

O primeiro módulo deste aplicativo é dedicado à geração de visualizações de dados para a análise dos dados, através de um gráfico virtual, oferecendo ao usuário uma plataforma para localizar e mensurar os dados ausentes em tempo real. Essa funcionalidade é fundamental para embasar as decisões sobre o tratamento adequado aos dados ausentes.

O segundo módulo introduz um menu virtual de opções de métodos de tratamento de dados ausentes. Com a ajuda deste menu virtual, o usuário pode escolher entre diversos métodos de tratamento de dados ausentes, baseando suas decisões em evidências visuais obtidas a partir do primeiro módulo. Além das opções de métodos de tratamento de dados, este menu virtual conta com uma opção de visualizar as características dos dados, como por exemplo o tipo dos dados, assim, o analista de dados, tem informações a respeito dos dados para auxiliá-lo na tomada de decisão sobre qual o método de tratamento de dados deverá ser aplicado. Esta abordagem interativa e orientada por dados virtuais é desenvolvida para simplificar o processo de tratamento de dados ausentes.

#### **5.1.1. Módulo de Análise dos dados**

A execução do *DMT-AR* inicia quando o usuário aponta um determinado cartão físico para a câmera do dispositivo físico. É importante ressaltar que a função deste cartão físico é servir como marcador para iniciar a sobreposição de elementos virtuais na cena real. Então, se o módulo *Vuforia* reconhecer com sucesso este cartão (é comparada a imagem do cartão com as da base de imagens), então é sobreposto, à cena real, o gráfico virtual correspondente, ainda desprovido dos dados. Os cartões possuem identificações sobre a qual conjunto de dados se

referem. Após o reconhecimento, um arquivo *C#* é acionado, estabelecendo uma conexão com a base de dados e realizando uma consulta *SQL* predefinida. As informações resultantes são então transferidas e apresentadas no referido gráfico virtual. Estas informações são, por exemplo: total de dados, a quantidade de dados ausentes e a porcentagem de dados ausentes. Para consultar diferentes bases de dados, basta trocar para o cartão correspondente ou criar um novo.

Desta forma, promove-se uma compreensão instantânea do *status* dos dados associados a cada cartão físico. Ao visualizarem estas informações os usuários se beneficiarão significativamente de uma compreensão instantânea da integridade de seus conjuntos de dados. A representação visual dos dados ausentes em gráficos virtuais 3D sobrepostos a cartões físicos oferece uma abordagem intuitiva, permitindo aos usuários identificar, localizar e mensurar áreas específicas de lacunas de informações em tempo real.

No entanto, entendemos que a visualização de dados pode variar de acordo com as preferências e necessidades do usuário. Caso o usuário não esteja satisfeito com a apresentação ou considere a visualização do gráfico virtual inadequada, reiniciamos este módulo. Nesse caso, o usuário solicita ao *ChatGPT* a geração de códigos-fonte para outro(s) modelo(s) de gráfico(s) virtuais mais adequado(s) às expectativas. Então, o usuário vincula esses códigos-fontes a novos cartões físicos.

A Figura 14 exemplifica a execução deste módulo, apresentando um gráfico virtual que ilustra a tabela *Assunto\_Consulta* (esta tabela pertence à base de dados da biblioteca da Universidade Federal da Fronteira Sul). Neste gráfico virtual, os nomes das colunas (*Descrição*, *Desc\_Titulo* e *Desc\_Titulo\_Formatado*) são exibidos na dimensão de profundidade(Z). Na dimensão horizontal(X) são representados o total de dados de cada coluna, a quantidade de dados nulos da coluna e a porcentagem de dados nulos em relação ao total de dados. Na dimensão vertical(Y), encontram-se os valores correspondentes à dimensão horizontal.

Neste exemplo, na coluna *Descrição* (cor verde) , há 917 instâncias, destes, 105 são dados nulos que representa 11,45% do total de dados da referida coluna. Na coluna

Desc\_Titulo (cor preta), há 917 instâncias, destas 207 são dados ausentes, representando 21,92% em relação ao total de dados desta coluna. Na coluna Desc\_Titulo\_Formatado (cor azul), há 917 instâncias, com 207 dados ausentes que representa 21,92% em relação ao total de dados desta coluna. À medida que o usuário rotaciona o cartão físico, a qualidade das visualizações é aprimorada de acordo com a orientação desejada.

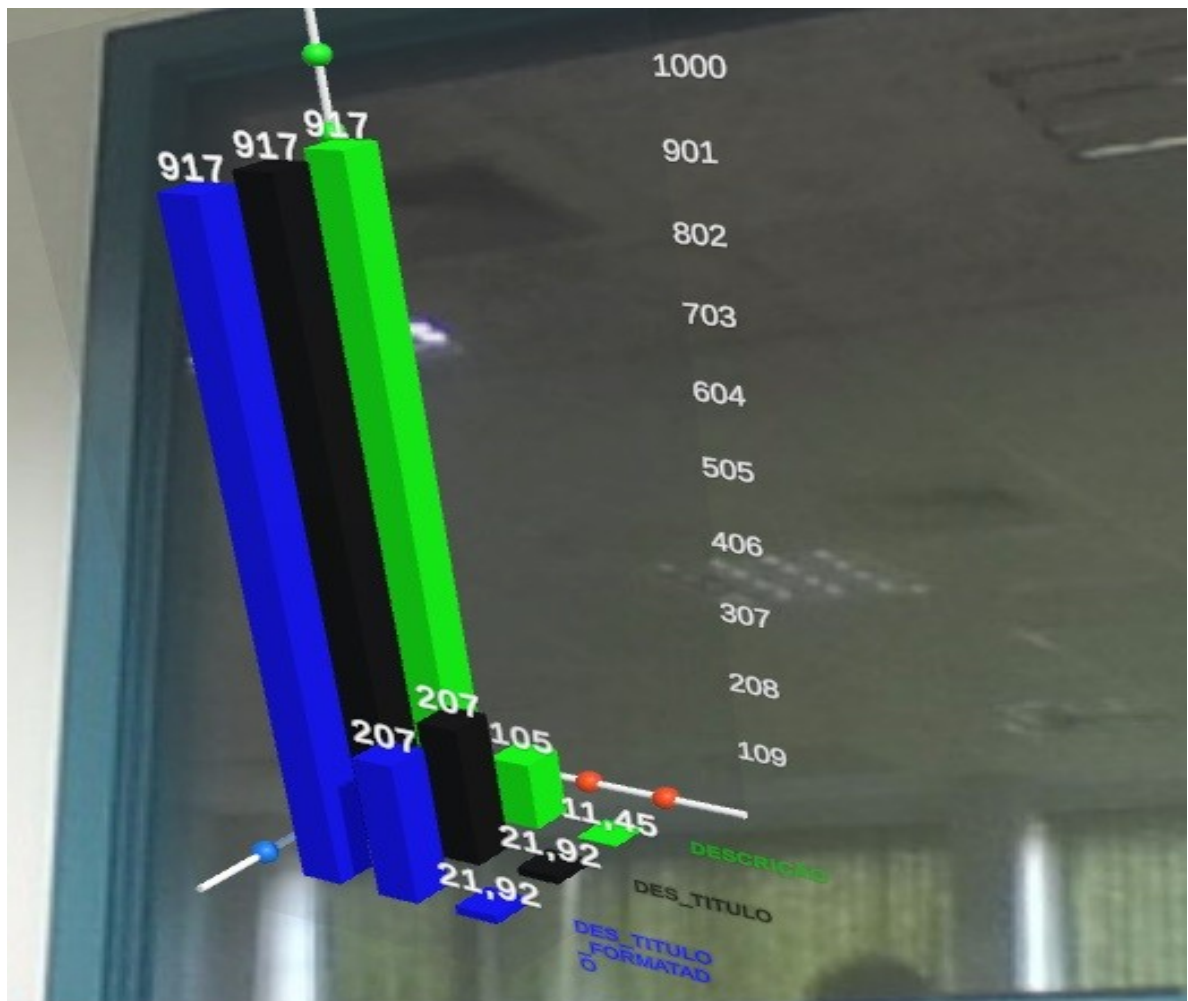


Figura 14: Módulo Análise de dados em execução

### 5.1.2. Módulo de Tratamento dos dados

Este módulo inicia-se quando o usuário faz a escolha do método de tratamento de dados que será aplicado aos dados ausentes, através do menu virtual. Feita a escolha os

dados ausentes são tratados através de um arquivo *C#*, vinculado a este menu virtual. Este mesmo arquivo *C#* também atualiza a base de dados. Cada uma das opções está vinculada a um arquivo *C#* diferente.

Conforme ilustra a Figura 15 o menu virtual é composto, inicialmente, por cinco opções: Média, Mediana, Moda, Eliminação de Valores Atípicos (*BoxPlot*) e Regressão Linear Simples. Além destas opções, o menu virtual conta com a opção para visualializar as características dos dados (opção Tipo dos dados), como por exemplo, tipo dos dados de cada coluna e relações entre os dados. Desta forma, o usuário terá um embasamento melhor quando da escolha do método de tratamento de dados a ser aplicado, pois saber os tipos de dados, referente ao conjunto de dados onde estão inseridos os dados ausentes, ajuda a escolher método mais apropriado.

Se o usuário desejar adicionar outros métodos de tratamento de dados, então o fluxo do módulo é desviado para que o usuário solicite ao *ChatGPT* a geração de um novo arquivo *C#* correspondente ao novo método de tratamento de dados ausentes. Então, o usuário vincula este novo arquivo *C#* ao menu virtual.

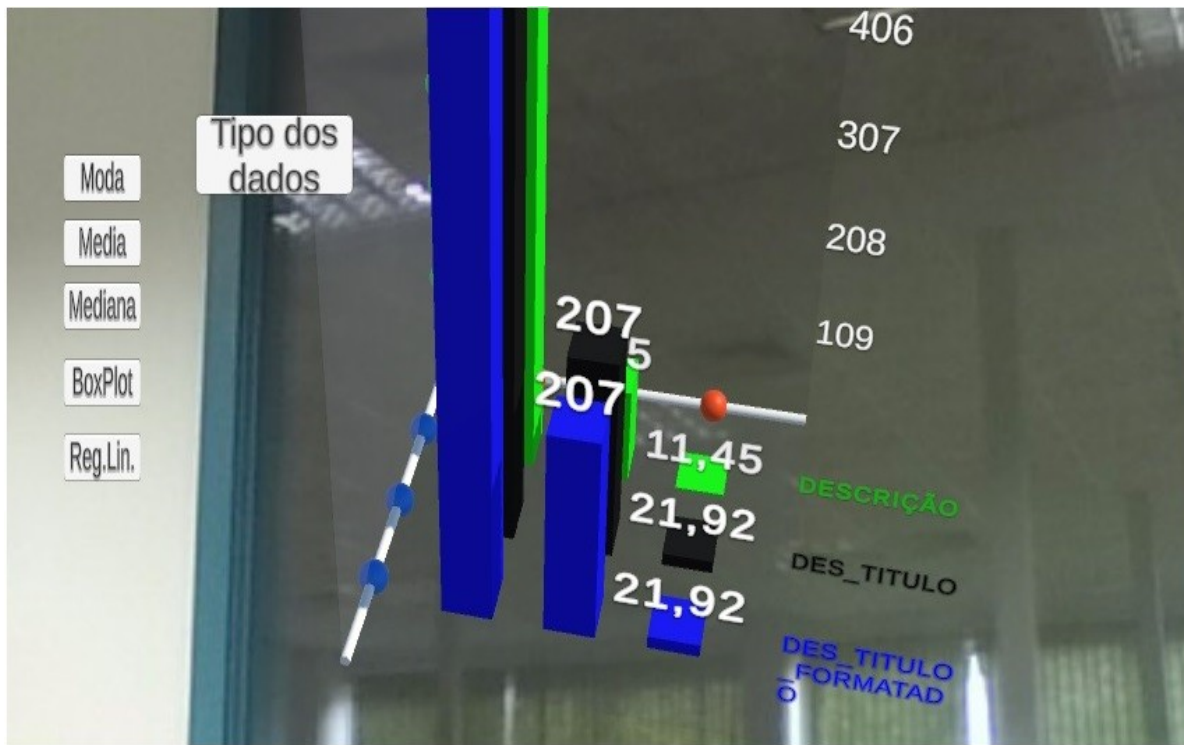


Figura 15: Módulo Tratamento de dados em execução

A Figura 16 ilustra o gráfico virtual após o tratamento de dados ausentes realizado. A referida figura mostra o total de dados de cada coluna, o total de dados ausentes e a porcentagem de dados ausentes, as duas últimas com valor zero. Fechando, assim, o ciclo de funcionamento do *DMT-AR*.



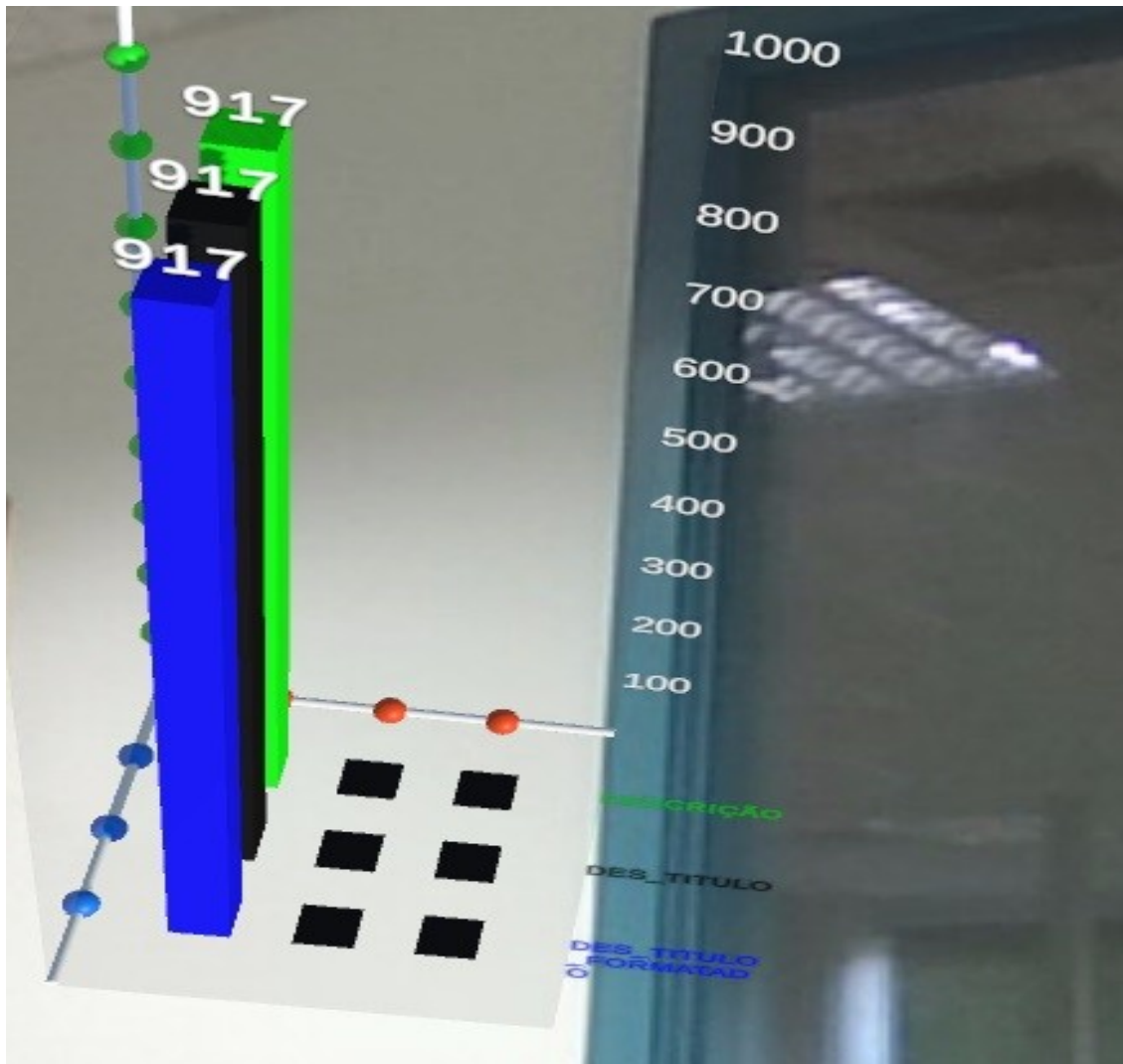


Figura 16: Gráfico virtual após o tratamento dos dados

## 5.2. Arquitetura do *DMT-AR*

A arquitetura do *DMT-AR* é uma construção que suporta de maneira eficaz as funcionalidades e interações que tornam este aplicativo uma solução abrangente para o tratamento de dados ausentes.

Uma das características do *DMT-AR* é a sua flexibilidade de extensão, que o torna acessível a uma ampla gama de usuários. Este aplicativo foi projetado de modo a poder ser executado em qualquer dispositivo móvel comum. Isso permite que os analisistas de dados

usem de maneira conveniente, sem a necessidade de hardware especializado ou configurações complexas.

Conformer mostra o Diagrama de Caso de Uso apresentado na Figura 17, os atores envolvidos são o usuário e a ferramenta *ChatGPT*. O processo inicia após o usuário apontar o cartão físico para a câmera do dispositivo móvel, disparando o primeiro módulo, que consiste na visualização do gráfico virtual. Com esta visualização disponível o usuário então realiza análises visuais do gráfico apresentado. Caso o usuário desejar um outro gráfico, correspondente ao mesmo conjunto de dados, ele, opcionalmente, solicita ao *ChatGPT* a geração de novo gráfico. O usuário, em um novo fluxo, vincula o novo gráfico a um novo cartão físico.

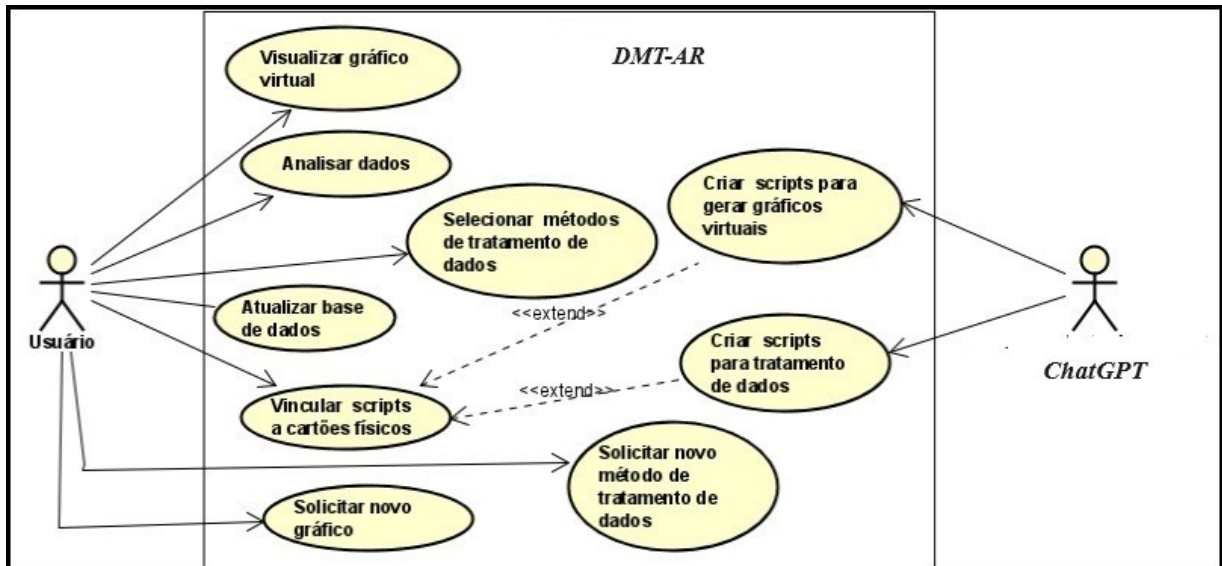


Figura 17: Diagrama de Caso de Uso para o *DMT-AR*

A Figura 18 apresenta o Diagrama de Sequência que ilustra detalhadamente as interações entre os atores (usuário e *ChatGPT*) e o sistema durante a execução deste caso de uso.

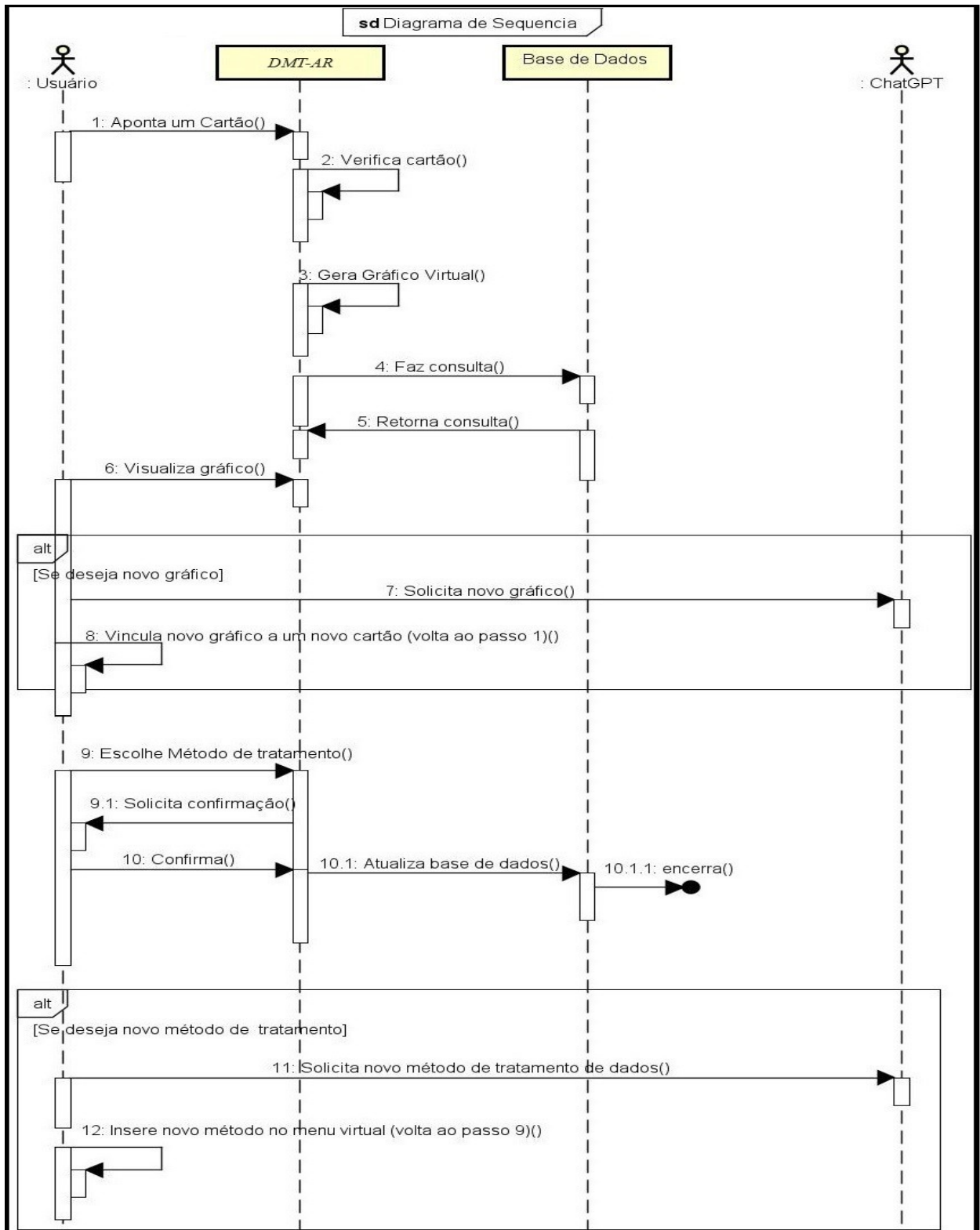


Figura 18 – Diagrama de sequência para o DMT-AR

Este diagrama oferece uma representação visual e detalhada das mensagens trocadas entre os objetos, as ordens de execução das operações e os pontos de decisão ao longo do processo.

As linhas de vida representadas no Diagrama de Sequência, da Figura 17, incluem o usuário, *DMT-AR*, a base de dados e o *ChatGPT*. Essas linhas de vida oferecem uma visão clara das entidades envolvidas e suas interações ao longo do processo de visualização, análise e tratamento de dados.

Na Tabela 6, é apresentado o documento de visão que fornece um passo a passo detalhado do referido Diagrama de Sequência, complementando a compreensão do processo e suas interações.

Tabela 6 – Documento de visão do *DMT-AR*

<b>Passos</b>	<b>Fluxos</b>
1	O usuário aponta um determinado cartão físico para a câmera do dispositivo físico.
2	O <i>DMT-AR</i> , através do módulo <i>Vuforia</i> , verifica se o cartão está presente na base do <i>Vuforia</i> . Se o reconhecimento ocorrer com sucesso, o processo continua.
3	Após o reconhecimento no passo anterior, é gerado um gráfico virtual, ainda desprovido de dados e ainda não visível pelo usuário (cada cartão, corresponde a um gráfico virtual).
4	O <i>DMT-AR</i> , dispara uma consulta ao banco de dados (vinculada ao cartão apresentado no passo 1).
5	O banco de dados retorna esta consulta e os dados são transferidos ao gráfico virtual gerado no passo 3.
6	O usuário visualiza o gráfico virtual e assim faz as análises dos dados.
7	Se o usuário desejar outro gráfico virtual, correspondente a mesma consulta, solicita ao <i>ChatGPT</i> Caso não desejar passa ao passo 9.

8	O usuário vincula este novo gráfico virtual a um novo cartão e o processo reinicia no passo 1.
9	O usuário, através de um menu virtual, escolhe um método de tratamento de dados.
10	Após esta escolha (passo 9) o <i>DMT-AR</i> , solicita confirmação.
10.1	O <i>DMT-AR</i> faz a atualização (com os dados tratados) da base de dados.
10.1.1	O sistema é encerrado.
11	Se o usuário desejar um novo método de tratamento de dados, ele solicita ao <i>ChatGPT</i> .
12	Após o <i>ChatGPT</i> gerar o novo método de tratamento de dados, o processo reinicia no passo 9.

A Figura 19 ilustra o Diagrama de Pacotes para o *DMT-AR*. De acordo com esta Figura a aplicação inicia-se com o acesso à base de dados. Quando o usuário direciona o cartão físico à câmera do dispositivo é acionado um artefato C# que faz a respectiva consulta à base de dados. Em seguida, procede-se à visualização dos dados (classe *Visualiza\_Dados*) e ao tratamento dos mesmos (classe *Trata\_Dados*). No caso de o usuário manifestar interesse em novos gráficos de visualização de dados ou em novos métodos de tratamento de dados, ele pode requisitar ao *ChatGPT* a geração de scripts para a produção desses novos artefatos, os quais são associados, por meio de generalização, às classes *Visualiza\_Dados* e *Trata\_Dados*, respectivamente.

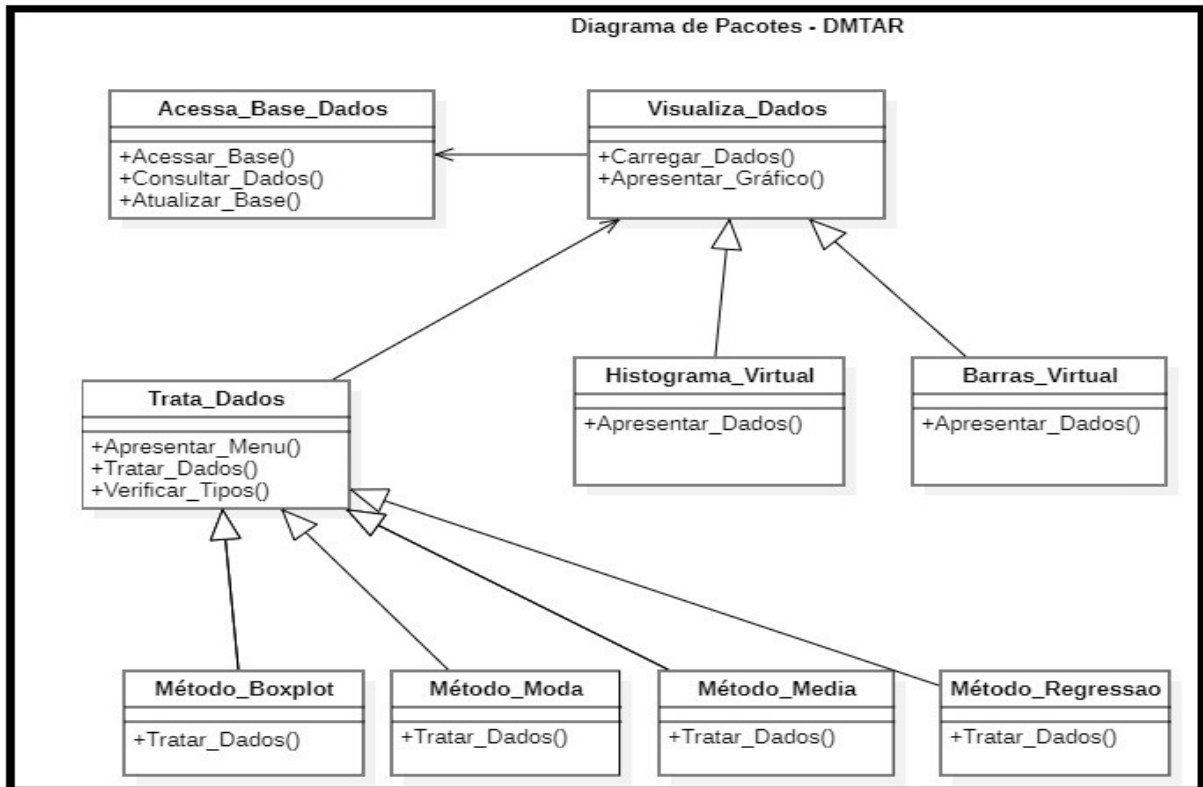


Figura 19 – Diagrama de Pacotes para o *DMT-AR*

A Figura 20 ilustra o Fluxograma do *DMT-AR*. Desta forma, o referido Fluxograma ilustra de maneira abrangente todos os passos do *DMT-AR*, proporcionando uma visão ampla de todo o seu fluxo. Essa representação visual permite uma compreensão mais completa e detalhada das etapas envolvidas no processo, desde a interação inicial do usuário com o dispositivo até a finalização das operações de visualização e tratamento de dados. Por meio dessa abordagem gráfica, torna-se possível identificar de forma clara e sequencial as diferentes fases e subprocessos que compõem o funcionamento do sistema, contribuindo assim para uma melhor apreciação e análise de sua estrutura e funcionamento.

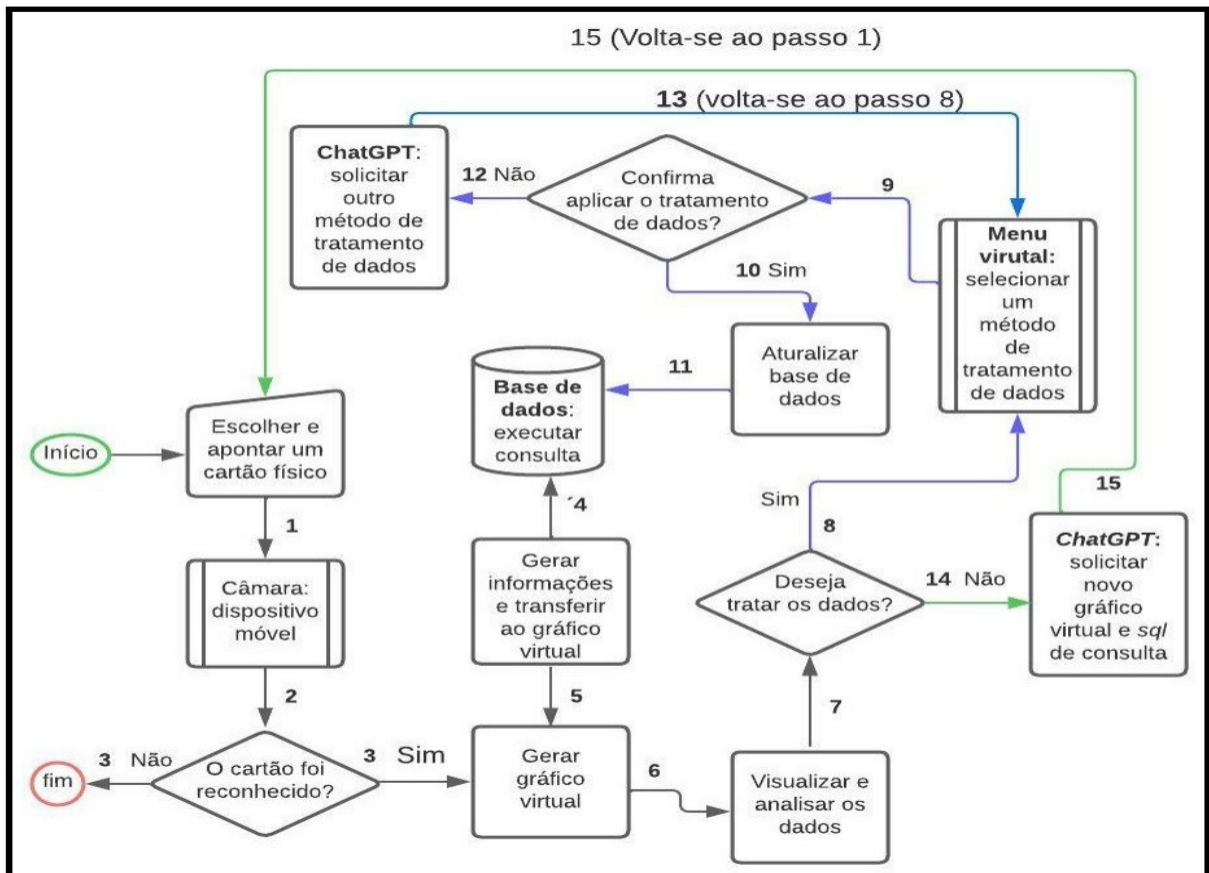


Figura 20 – Fluxograma do DMT-AR

### 5.3. O ChatGPT e a geração dos códigos-fonte

Nesta seção, exploraremos em detalhes como o *ChatGPT* foi empregado como uma ferramenta *LLM* no processo de geração dos códigos-fonte utilizados no desenvolvimento do nosso aplicativo. O *ChatGPT* é uma implementação do modelo *GPT (Generative Pre-trained Transformer)* desenvolvido pela empresa *OpenAI*.

O objetivo principal do uso do *ChatGPT* é que o analista de dados obtenha códigos-fonte para a visualização e tratamento de dados. O analista também solicitará, opcionalmente, códigos-fonte para acrescentar visualizações e métodos de tratamento de dados na aplicação. O *ChatGPT* não acessa dados em arquivos ou base de dados. Entretanto, gera códigos-fonte para estas finalidades (Withe *et. al.*, 2023).

A interação com o *ChatGPT* foi realizada através de *prompts* (perguntas) procurando obter *responses* (respostas ou *insights*). Posteriormente, estas perguntas foram sendo refinadas para obter resultados mais otimizados, uma vez que o *prompt* apresentado pode afetar significativamente as respostas obtidas. Ou seja, as entradas (perguntas), as palavras-chaves e o contexto fornecidos podem influenciar a compreensão e saída do modelo.

Dentro desse contexto, conforme Withe *et. al.* (2023), um *prompt* é um conjunto de instruções fornecidas a um software de processamento de linguagem natural. Ainda, segundo esses autores, Engenharia de *Prompts* é um termo mais amplo que se refere ao processo de criar e otimizar *prompts* para modelos de linguagens como os modelos de geração de textos baseados em inteligência artificial, por exemplo o *ChatGPT*.

### 5.3.1. Padrões de *prompts* aplicados

Esta subseção mostra alguns exemplos de *prompts* que foram aplicados ao *ChatGPT* e respostas obtidas que foram adaptadas ao nosso contexto, durante a construção do *DMT-AR*. Para isso utilizamos um arquivo no formato csv (*empregados.csv*) para este estudo de caso. O arquivo possui 1000 linhas e 8 colunas sendo 6 do tipo categórico e 2 do tipo numérico. Além disso, é possível observar, através da Figura 21, que o arquivo possui dados ausentes, destacados em vermelho. Esta Figura mostra o cabeçalho (colunas e primeira página) do arquivo csv utilizado como exemplo.

O processo iterativo de refinamento dos *prompts* ao longo do tempo visa garantir que as respostas do *ChatGPT* se alinhem cada vez mais com o contexto em questão. Essa abordagem faz sentido, pois permite ajustar e adaptar o modelo para atender às necessidades específicas de cada projeto. Segundo Withe *et. al.* (2023), o refinamento contínuo dos *prompts* ajuda a melhorar a precisão e relevância das respostas, tornando o modelo mais eficaz para interagir com os usuários e fornecer informações relevantes.

Houve cerca de 12 interações com o *ChatGPT* para cada tentativa de alinhamento de contexto com o nosso propósito. Ou seja, os *prompts* foram testados e ajustados para garantir que as respostas alcancem qualidade e precisão.



Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
Douglas	Male	8/6/1993	12:42 PM	97308	6945	true	Marketing
Thomas	Male	3/31/1996	6:53 AM	61933	4.17	true	
Maria	Female	4/23/1993	11:17 AM	130590	11858	false	Finance
Jerry	Male	3/4/2005	1:00 PM	138705	9.34	true	Finance
Larry	Male	1/24/1998	4:47 PM	101004	1389	true	Client Services
Dennis	Male	4/18/1987	1:35 AM	115163	10125	false	Legal
Ruby	Female	8/17/1987	4:20 PM	65476	10012	true	Product
	Female	7/20/2015	10:43 AM	45906	11598		Finance
Angela	Female	11/22/2005	6:29 AM	95570	18523	true	Engineering
Frances	Female	8/8/2002	6:51 AM	139852	7524	true	Business Development
Louise	Female	8/12/1980	9:01 AM	63241	15132	true	
Julie	Female	10/26/1997	3:19 PM	102508	12637	true	Legal
Brandon	Male	12/1/1980	1:08 AM	112807	17492	true	Human Resources
Gary	Male	1/27/2008	11:40 PM	109831	5831	false	Sales
Kimberly	Female	1/14/1999	7:13 AM	41426	14543	true	Finance
Lillian	Female	6/5/2016	6:09 AM	59414	1256	false	Product
Jeremy	Male	9/21/2010	5:56 AM	90370	7369	false	Human Resources
Shawn	Male	12/7/1986	7:45 PM	111737	6414	false	Product
Diana	Female	10/23/1981	10:27 AM	132940	19082	false	Client Services
Donna	Female	7/22/2010	3:48 AM	81014	1894	false	Product
Lois		4/22/1995	7:18 PM	64714	4934	true	Legal
Matthew	Male	9/5/1995	2:12 AM	100612	13645	false	Marketing
Joshua		3/8/2012	1:58 AM	90816	18816	true	Client Services
	Male	6/14/2012	4:19 PM	125792	5042		
John	Male	7/1/1992	10:08 PM	97950	13873	false	Client Services
	Male	10/8/2012	1:12 AM	37076	18576		Client Services
Craig	Male	2/27/2000	7:45 AM	37598	7757	true	Marketing
Scott		7/11/1991	6:58 PM	122367	5218	false	Legal
Terry	Male	11/27/1981	6:30 PM	124008	13464	true	Client Services
Benjamin	Male	1/26/2005	10:06 PM	79529	7008	true	Legal
Christina	Female	8/6/2002	1:19 PM	118780	9096	true	Engineering
Joyce		2/20/2005	2:40 PM	88657	12752	false	Product
	Male	8/21/1998	2:27 PM	122340	6417		
Jean	Female	12/18/1993	9:07 AM	119082	16.18	false	Business Development
Jerry	Male	1/10/2004	12:56 PM	95734	19096	false	Client Services
Theresa	Female	10/10/2006	1:12 AM	85182	16675	false	Sales
Rachel	Female	2/16/2009	8:47 PM	142032	12599	false	Business Development
Linda	Female	10/19/1981	8:49 PM	57427	9557	true	Client Services
Stephanie	Female	9/13/1986	1:52 AM	36844	5574	true	Business Development
	Male	1/29/2016	2:33 AM	122173	7797		Client Services
Michael	Male	10/10/2008	11:25 AM	99283	2665	true	Distribution
Christine		6/28/2015	1:08 AM	66582	11308	true	Business Development
Beverly	Female	9/9/1998	8:26 PM	121918	15835	false	Legal
Marilyn	Female	12/7/1980	3:16 AM	73524	5207	true	Marketing
Cynthia	Female	11/16/1988	6:54 PM	145146	7482	true	Product
Roger	Male	4/17/1980	11:32 AM	88010	13886	true	Sales
Bruce	Male	11/28/2009	10:47 PM	114796	6796	false	Finance
Kathy	Female	6/22/2005	4:51 AM	66820	9.0	true	Client Services
Clarence	Male	3/26/1996	5:57 AM	93581	6083	true	Business Development
Chris		1/24/1980	12:13 PM	113590	3055	false	Sales
Nancy	Female	9/23/2000	8:05 AM	94976	13.83	true	Engineering
		12/17/2011	8:29 AM	41126	14009		Sales
Todd	Male	2/18/1990	2:41 AM	49339	1695	true	Human Resources

Figura 21: Cabeçalho do arquivo csv de exemplo

Na primeira solicitação ao *ChatGPT* visou-se obter estatísticas do arquivo csv utilizado. A Tabela 7 mostra o diálogo utilizado, após vários refinamentos até à versão final, apresentada na referida Tabela.

Tabela 7 – Exemplo de informações estatísticas

Nº	Prompts	Resposta do <i>ChatGPT</i>
1	Gere informações a respeito de estatísticas sobre os dados numéricos de um arquivo. O nome do arquivo é empregados.csv	Não tenho acesso a arquivos externos e não posso realizar operações em tempo real, vou fornecer um exemplo de como você pode realizar essa análise usando <i>Python</i> e a biblioteca <i>pandas</i> .

Como se pode observar, na resposta 1 da Tabela 7, o *ChatGPT* não executa a operação solicitada. Entretanto, ele gera o código-fonte necessário para a execução solicitada. A Figura 22 apresenta o código gerado pelo *ChatGPT* e o resultado da execução. Como observado nesta Figura, a resposta gerada contém o código-fonte (método *describe()*) que, quando executado, retorna as principais informações estatísticas dos dados numéricos do arquivo csv. As informações incluem: contagem de observações, média, desvio padrão, valor mínimo, primeiro percentil, mediana, terceiro percentil e valor máximo.

<pre>import pandas as pd  # Carregar o arquivo CSV caminho_arquivo_csv = "empregados.csv" dados = pd.read_csv(caminho_arquivo_csv)  # Selecionar apenas as colunas numéricas dados_numericos = dados.select_dtypes(include=['number'])  # Calcular estatísticas descritivas para as colunas numéricas estatisticas_numericas = dados_numericos.describe()  # Exibir estatísticas descritivas para as colunas numéricas print("Estatísticas descritivas para as colunas numéricas:") print(estatisticas_numericas)</pre>	<pre>Estatísticas descritivas para as colunas numéricas:               Salary      Bonus % count    1000.000000    1000.000000 mean     90662.181000    9196.417150 std      32923.693342    6115.879002 min       35013.000000         1.020000 25%      62613.000000    3767.250000 50%      90428.000000    9054.500000 75%     118740.250000   14304.750000 max     149908.000000   19944.000000</pre>
---	--

Figura 22: Cálculos de estatísticas de dados numéricos

No diálogo anterior (Tabela 7) as informações apresentadas estavam relacionadas apenas aos dados numéricos do arquivo csv. Para obter informações sobre os dados categóricos no arquivo, solicitamos ao *ChatGPT* calcular e apresentar estatísticas relevantes para os dados categóricos, conforme apresentado na Tabela 8.

Tabela 8 – Exemplo de estatísticas para dados categóricos

Nº	<i>Prompts</i>	<i>Resposta do ChatGPT</i>
1	Crie um <i>script</i> para mostrar contagem de valores únicos para as variáveis categóricas.	Entendido, você deseja obter apenas as informações estatísticas e a contagem de valores únicos para as colunas categóricas no arquivo "empregados.csv". Vou ajustar o código para atender a essa solicitação.

A Figura 23 apresenta o código gerado pelo *ChatGPT* e as respostas fornecidas após os diálogos da Tabela 8. Como observado nesta figura, as respostas geradas contêm o código-fonte que, quando executado, retorna as principais informações relacionadas aos dados categóricos do arquivo.

<pre>import pandas as pd  # Carregando o arquivo CSV df = pd.read_csv('empregados.csv')  # Identificando as colunas categóricas colunas_categoricas = df.select_dtypes(include=['object'])  contagem_valores_unicos_categoricas = colunas_categoricas.nunique()  print("contagem de valores unicos para colunas categóricas:") print(contagem_valores_unicos_categoricas)</pre>	<pre>contagem de valores unicos para colunas categóricas: Name                200 Gender                2 Start Date           972 Last Login Time      720 Senior Management    2 Team                 10 dtype: int64</pre>
---	---

Figura 23: Exemplo de contagem de valores únicos

A compreensão dos tipos de dados em um conjunto de dados é um passo fundamental antes de prosseguir com a criação de gráficos e análises de dados. Segundo *Meiguins et al.* (2019), as capacidades de interações e manipulações de dados diferem conforme os tipos de dados e influenciam na escolha do tipo de visualização de dados.

A Tabela 9 mostra a solicitação sobre tipos de dados enviados ao *ChatGPT* e a respectiva resposta, após diversos refinamentos.

Tabela 9 – Exemplo para geração de tipo de dados

Nº	Prompts	Resposta do ChatGPT
1	Crie um <i>script</i> para mostrar os tipos de dados das colunas para o arquivo empregados.csv	Este código carrega o arquivo "empregados.csv" e exibe os tipos de dados das colunas no console. Certifique-se de que o arquivo "empregados.csv" esteja na mesma pasta do código ou ajuste o caminho conforme necessário.

Na Figura 24 pode-se visualizar o código-fonte gerado e o resultado da execução.

<pre># Carregando o arquivo CSV df = pd.read_csv('empregados.csv')  # Exibindo os tipos de dados das colunas tipos_de_dados = df.dtypes print(tipos_de_dados)</pre>	<pre>Name          object Gender         object Start Date     object Last Login Time object Salary         int64 Bonus %        float64 Senior Management object Team           object dtype: object</pre>
---	---

Figura 24: Exemplo de geração dos tipos de dados

A verificação de dados ausentes é um aspecto crítico em qualquer análise de dados, pois a presença de valores ausentes pode distorcer os resultados e levar a interpretações imprecisas. Identificar e lidar com dados ausentes de forma eficaz é fundamental para garantir a qualidade e a confiabilidade das análises (Makaba e Dogo, 2020).

A Tabela 10 ilustra o *prompt* e *response*, depois dos devidos refinamentos, para gerar código-fonte voltado para descobrir se há dados ausentes no arquivo.

Tabela 10 – Exemplo de verificação da existência de dados ausentes

Nº	Prompts	Resposta do ChatGPT
1	Crie um script para carregar e verificar se existem dados ausentes no arquivo empregados.csv. Se tiver dados ausentes preciso que o <i>script</i> calcule a quantia de dados de cada coluna, a quantia de dados ausentes de cada coluna e a porcentagem de dados ausentes de cada coluna em relação a quantia de dados	Você pode usar o seguinte <i>script Python</i> para carregar o arquivo "empregados.csv" e verificar se existem dados ausentes. Se houver dados ausentes, o <i>script</i> calculará a quantidade de dados de cada coluna, a quantidade de dados ausentes de cada coluna e a porcentagem de dados ausentes de cada coluna em relação à quantidade de dados da respectiva coluna. Se não houver dados ausentes, ele



da respectiva coluna. Se no arquivo inexistir dados ausentes retorne esta informação.	informará que não foram encontrados dados ausentes.
---	---

Na Figura 25 pode-se visualizar o código-fonte gerado pelo *ChatGPT* e o resultado da execução.

<pre>df = pd.read_csv('empregados.csv') # Verificando se existem dados ausentes if df.isnull().values.any():     # Calculando a quantidade de dados de cada coluna     quantidade_de_dados = df.count()      # Calculando a quantidade de dados ausentes de cada     # coluna     dados_ausentes = df.isnull().sum()      # Calculando a porcentagem de dados ausentes em     # relação a quantidade de dados de cada coluna     porcentagem_ausentes = (dados_ausentes /     quantidade_de_dados) * 100      info_ausentes = pd.DataFrame({         'Quantidade de Dados': quantidade_de_dados,         'Dados Ausentes': dados_ausentes,         'Porcentagem de Ausentes (%)':         porcentagem_ausentes     })      print("\nInformações sobre dados ausentes:\n")     print(info_ausentes)</pre>	<table border="1"> <thead> <tr> <th></th> <th>Quantidade de Dados</th> <th>Dados Ausentes</th> </tr> </thead> <tbody> <tr><td>Name</td><td>933</td><td>67</td></tr> <tr><td>Gender</td><td>855</td><td>145</td></tr> <tr><td>Start Date</td><td>1000</td><td>0</td></tr> <tr><td>Last Login Time</td><td>1000</td><td>0</td></tr> <tr><td>Salary</td><td>1000</td><td>0</td></tr> <tr><td>Bonus %</td><td>1000</td><td>0</td></tr> <tr><td>Senior Management</td><td>933</td><td>67</td></tr> <tr><td>Team</td><td>957</td><td>43</td></tr> </tbody> </table> <table border="1"> <thead> <tr> <th></th> <th>Porcentagem de Ausentes (%)</th> </tr> </thead> <tbody> <tr><td>Name</td><td>7.181136</td></tr> <tr><td>Gender</td><td>16.959064</td></tr> <tr><td>Start Date</td><td>0.000000</td></tr> <tr><td>Last Login Time</td><td>0.000000</td></tr> <tr><td>Salary</td><td>0.000000</td></tr> <tr><td>Bonus %</td><td>0.000000</td></tr> <tr><td>Senior Management</td><td>7.181136</td></tr> <tr><td>Team</td><td>4.493208</td></tr> </tbody> </table>		Quantidade de Dados	Dados Ausentes	Name	933	67	Gender	855	145	Start Date	1000	0	Last Login Time	1000	0	Salary	1000	0	Bonus %	1000	0	Senior Management	933	67	Team	957	43		Porcentagem de Ausentes (%)	Name	7.181136	Gender	16.959064	Start Date	0.000000	Last Login Time	0.000000	Salary	0.000000	Bonus %	0.000000	Senior Management	7.181136	Team	4.493208
	Quantidade de Dados	Dados Ausentes																																												
Name	933	67																																												
Gender	855	145																																												
Start Date	1000	0																																												
Last Login Time	1000	0																																												
Salary	1000	0																																												
Bonus %	1000	0																																												
Senior Management	933	67																																												
Team	957	43																																												
	Porcentagem de Ausentes (%)																																													
Name	7.181136																																													
Gender	16.959064																																													
Start Date	0.000000																																													
Last Login Time	0.000000																																													
Salary	0.000000																																													
Bonus %	0.000000																																													
Senior Management	7.181136																																													
Team	4.493208																																													

Figura 25: Exemplo da detecção de dados ausentes

Com a execução do código-fonte da Figura 25 concluímos que há 4 colunas com dados ausentes. Então solicitamos ao *ChatGPT* que nos mostre quais os tipos de dados destas colunas, após refinamentos consecutivos, conforme mostrado no diálogo da Tabela 11.

Tabela 11 – Verificando características dos dados ausentes

Nº	<i>Prompts</i>	<i>Resposta do ChatGPT</i>
1	Mostre os tipos de dados das colunas que possuem dados ausentes no arquivo empregados.csv.	Para mostrar os tipos de dados das colunas que possuem dados ausentes no arquivo "empregados.csv", você pode usar o seguinte código em Python.

Diferentes tipos de dados ausentes podem exigir abordagens de tratamento diferentes. Por exemplo, dados numéricos ausentes podem ser preenchidos com a média, mediana ou um valor específico, enquanto dados categóricos podem ser preenchidos com a moda. Saber os

tipos de dados, referente ao conjunto de dados, onde estão inseridos os dados ausentes, ajuda a escolher a estratégia mais apropriada. A Figura 26 apresenta o código-fonte gerado, após refinamentos, referente ao diálogo 1 da Tabela 11, além do resultado de execução do mesmo.

<pre># Carregando o arquivo CSV df = pd.read_csv('empregados.csv')  # Identificando colunas com dados ausentes colunas_com_ausentes = df.columns[df.isnull().any()]  # Exibindo os tipos de dados das colunas com dados ausentes tipos_de_dados_ausentes = df[colunas_com_ausentes].dtypes  print("\nTipos de dados das colunas com dados ausentes: \n") print(tipos_de_dados_ausentes)</pre>	<pre>Tipos de dados das colunas com dados ausentes:  Name                object Gender              object Senior Management   object Team                object dtype: object</pre>
---	--

Figura 26: Exemplo de verificação de tipos de dados ausentes

Conforme apresentado nos diálogos da Tabela 12, solicitamos ao *ChatGPT* a geração de gráficos 3D para visualizações dos dados ausentes no arquivo usado neste caso de uso. O *ChatGPT* produziu o código-fonte destinado à elaboração do gráfico de barras empilhadas 3D, atendendo à nossa solicitação, realizada na linha 2 desta Tabela.

Tabela 12 – Exemplo de geração de gráficos 3D

Nº	Prompts	Resposta do ChatGPT
1	Como foi mostrado anteriormente, existem dados ausentes em 4 colunas do tipo categóricos no arquivo empregados.csv. Quais tipos de gráficos 3D posso utilizar para apresentar a quantia total de dados de cada uma destas colunas, o total de dados ausentes destas colunas e a porcentagem de dados ausentes em relação a quantia total de dados em cada uma destas colunas, em um único gráfico 3D.	1) Gráfico de barras empilhadas 3D; 2) Gráfico de Dispersão 3D; 3) Superfície 3D; 4) Gráfico de Linhas 3D; 5) Histograma 3D; 6) <i>Wireframe</i> 3D.
2	Crie um <i>script</i> utilizando Gráfico de barras empilhadas 3D (1) para visualizarmos o total de dados ausentes de cada coluna, o total de dados ausentes destas colunas e a porcentagem de dados ausentes em relação a quantia total de dados em cada uma destas.	Para criar gráficos que visualizem o total de dados ausentes, o total de dados não ausentes e a porcentagem de dados ausentes em relação ao total de dados em cada coluna categórica do arquivo "empregados.csv", você pode usar as bibliotecas pandas e matplotlib em Python. Aqui está um script que cria seis tipos diferentes.

A figura 27 exibe uma parte do código-fonte resultante dessa requisição.

```
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D
.
.
.
.
.
# Criando uma figura 3D
fig = plt.figure(figsize=(10, 6))
ax = fig.add_subplot(111, projection='3d')

# Adicionando barras empilhadas 3D
ax.bar(colunas_com_ausentes, dados_ausentes, bottom=0,
zdir='y', color='red', alpha=0.7, label='Total de dados
ausentes')
ax.bar(colunas_com_ausentes, total_dados,
bottom=dados_ausentes, zdir='y', color='green', alpha=
0.7, label='Total de dados')
# Adicionando a porcentagem de dados ausentes como
anotações
for i, valor in enumerate(porcentagem_ausentes):
    ax.text(i, total_dados[i] + dados_ausentes[i] / 2, 0,
f'{valor:.2f}%', ha='center',
va='center').

# Exibindo o gráfico
plt.show()
```

Figura 27: Código-fonte para geração de gráfico 3D

Após a execução deste código-fonte, observou-se a geração eficaz do Gráfico de Barras Empilhadas 3D, mostrado na Figura 28. O referido gráfico 3D mostra na dimensão X: o total de dados de cada coluna, o total de dados ausentes de cada coluna e a porcentagem de dados ausentes de cada coluna; na dimensão Z os nomes das colunas e na dimensão Y os totais referente a dimensão X.

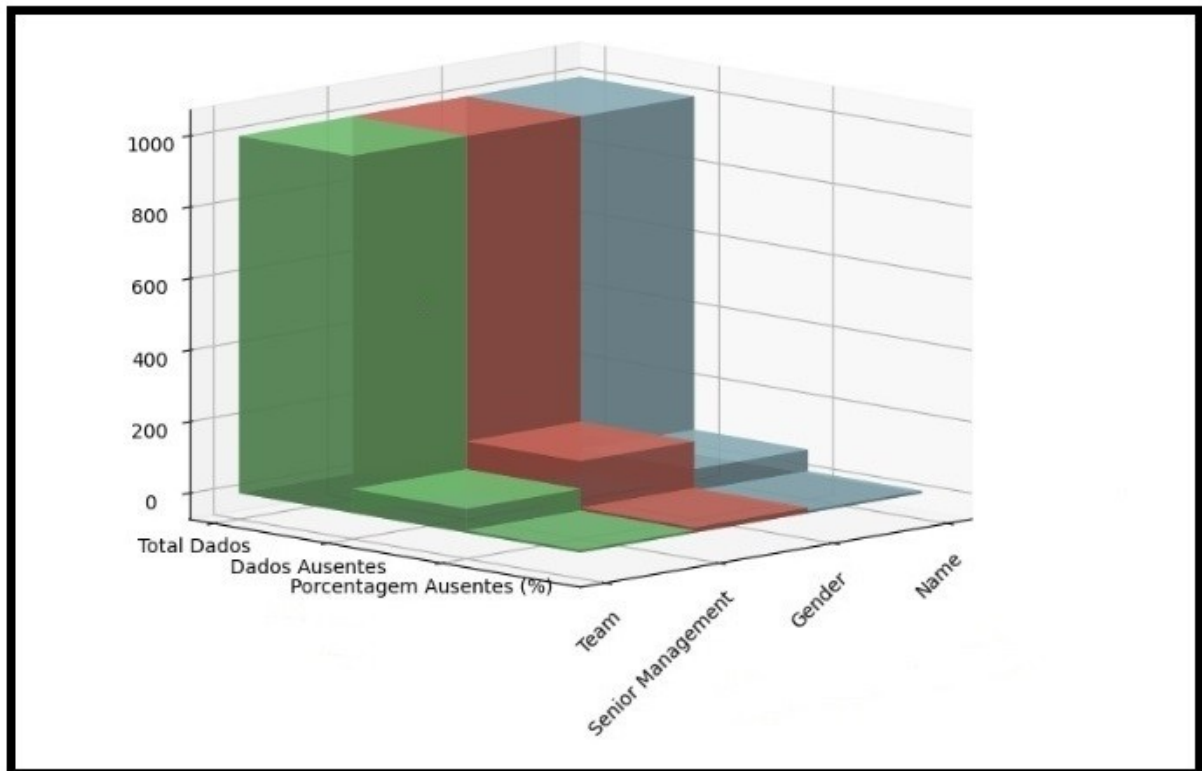


Figura 28: Visualização de dados através de gráfico 3D

Após as gerações dos gráficos para a análise da situação atual do conjunto de dados, a próxima etapa é o tratamento dos dados ausentes nos conjuntos de dados. O tratamento é uma parte fundamental no processo de extração de informações, pois ferramentas que fazem identificação de padrões, previsões ou classificações, como exemplo aprendizagem de máquina, não são projetadas para lidar com anomalias nos dados.

Como já identificamos as colunas do arquivo "empregados.csv" que possuem dados ausentes, todas do tipo categórico, solicitamos ao *ChatGPT* auxílio para tratar os dados. A Tabela 13 exibe os diálogos relacionados a essas solicitações, após os respectivos refinamentos.



Tabela 13 – Exemplos de técnicas de tratamento de dados ausentes

Nº	<i>Prompts</i>	Respostas do <i>ChatGPT</i>
1	Quais são as técnicas para tratar colunas categóricas com dados ausentes?	1) Preenchimento com a moda; 2) Preenchimento com um valor específico; 3) Preenchimento com base em informações de outras colunas; 4) Preenchimento com base em informações de outras colunas; 5) Codificação de rótulos ausentes; 6) Remoção de linhas; 7) Modelagem de aprendizado de máquina.
2	Crie um <i>script</i> para substituir as colunas categóricas com dados ausentes utilizando o preenchimento pela moda (item 1 acima)	Claro, aqui está um <i>script Python</i> para substituir as colunas categóricas com dados ausentes utilizando o preenchimento da moda (valor mais frequente).

Conforme mostra parte do código-fonte mostrado na Figura 29 os dados ausentes foram tratados com sucesso. Ainda, como mostra esta figura, o arquivo original foi preservado e com o código executado foi criado um novo arquivo com os dados imputados: “empregados\_modificado.csv”.

```

# Substituindo dados ausentes pela moda nas colunas
categoricas
for coluna in colunas_categoricas_ausentes:
    moda = df[coluna].mode().iloc[0] # Calcula a moda
    df[coluna].fillna(modas, inplace=True) # Substitui os
valores ausentes pela moda
. . . . .|
if dados_ausentes_apos_preenchimento == 0:
    print("Substituicao concluída. Nao ha mais dados
ausentes.")
else:
    print(f"Ainda existem
{dados_ausentes_apos_preenchimento} dados ausentes apos
substituicao.")
    print("Verifique as colunas categoricas para garantir
que todas foram preenchidas corretamente.")

# Salvando o DataFrame modificado em um novo arquivo CSV
df.to_csv('empregados_modificado.csv', index=False)

```

Figura 29: Código-fonte para o tratamento de dados ausentes

O cabeçalho deste arquivo, agora com os dados ausentes imputados, destacados com a cor verde, é apresentado na figura 30.

Name	Gender	Start Date	Last Login Time	Salary	Bonus %	Senior Management	Team
Douglas	Male	8/6/1993	12:42 PM	97308	6945.0	True	Marketing
Thomas	Male	3/31/1996	6:53 AM	61933	4.17	True	Client Services
Maria	Female	4/23/1993	11:17 AM	130590	11858.0	False	Finance
Jerry	Male	3/4/2005	1:00 PM	138705	9.34	True	Finance
Larry	Male	1/24/1998	4:47 PM	101004	1389.0	True	Client Services
Dennis	Male	4/18/1987	1:35 AM	115163	10125.0	False	Legal
Ruby	Female	8/17/1987	4:20 PM	65476	10012.0	True	Product
Marilyn	Female	7/20/2015	10:43 AM	45906	11598.0	True	Finance
Angela	Female	11/22/2005	6:29 AM	95570	18523.0	True	Engineering
Frances	Female	8/8/2002	6:51 AM	139852	7524.0	True	Business Development
Louise	Female	8/12/1980	9:01 AM	63241	15132.0	True	Client Services
Julie	Female	10/26/1997	3:19 PM	102508	12637.0	True	Legal
Brandon	Male	12/1/1980	1:08 AM	112807	17492.0	True	Human Resources
Gary	Male	1/27/2008	11:40 PM	109831	5831.0	False	Sales
Kimberly	Female	1/14/1999	7:13 AM	41426	14543.0	True	Finance
Lillian	Female	6/5/2016	6:09 AM	59414	1256.0	False	Product
Jeremy	Male	9/21/2010	5:56 AM	90370	7369.0	False	Human Resources
Shawn	Male	12/7/1986	7:45 PM	111737	6414.0	False	Product
Diana	Female	10/23/1981	10:27 AM	132940	19082.0	False	Client Services
Donna	Female	7/22/2010	3:48 AM	81014	1894.0	False	Product
Lois	Female	4/22/1995	7:18 PM	64714	4934.0	True	Legal
Matthew	Male	9/5/1995	2:12 AM	100612	13645.0	False	Marketing
Joshua	Female	3/8/2012	1:58 AM	90816	18816.0	True	Client Services
Marilyn	Male	6/14/2012	4:19 PM	125792	5042.0	True	Client Services
John	Male	7/1/1992	10:08 PM	97950	13873.0	False	Client Services
Marilyn	Male	10/8/2012	1:12 AM	37076	18576.0	True	Client Services
Craig	Male	2/27/2000	7:45 AM	37598	7757.0	True	Marketing
Scott	Female	7/11/1991	6:58 PM	122367	5218.0	False	Legal
Terry	Male	11/27/1981	6:30 PM	124008	13464.0	True	Client Services
Benjamin	Male	1/26/2005	10:06 PM	79529	7008.0	True	Legal
Christina	Female	8/6/2002	1:19 PM	118780	9096.0	True	Engineering
Joyce	Female	2/20/2005	2:40 PM	88657	12752.0	False	Product
Marilyn	Male	8/21/1998	2:27 PM	122340	6417.0	True	Client Services
Jean	Female	12/18/1993	9:07 AM	119082	16.18	False	Business Development
Jerry	Male	1/10/2004	12:56 PM	95734	19096.0	False	Client Services
Theresa	Female	10/10/2006	1:12 AM	85182	16675.0	False	Sales
Rachel	Female	2/16/2009	8:47 PM	142032	12599.0	False	Business Development
Linda	Female	10/19/1981	8:49 PM	57427	9557.0	True	Client Services
Stephanie	Female	9/13/1986	1:52 AM	36844	5574.0	True	Business Development
Marilyn	Male	1/29/2016	2:33 AM	122173	7797.0	True	Client Services
Michael	Male	10/10/2008	11:25 AM	99283	2665.0	True	Distribution
Christine	Female	6/28/2015	1:08 AM	66582	11308.0	True	Business Development
Beverly	Female	9/9/1998	8:26 PM	121918	15835.0	False	Legal
Marilyn	Female	12/7/1980	3:16 AM	73524	5207.0	True	Marketing
Cynthia	Female	11/16/1988	6:54 PM	145146	7482.0	True	Product
Roger	Male	4/17/1980	11:32 AM	88010	13886.0	True	Sales
Bruce	Male	11/28/2009	10:47 PM	114796	6796.0	False	Finance
Kathy	Female	6/22/2005	4:51 AM	66820	9.0	True	Client Services
Clarence	Male	3/26/1996	5:57 AM	93581	6083.0	True	Business Development
Chris	Female	1/24/1980	12:13 PM	113590	3055.0	False	Sales
Nancy	Female	9/23/2000	8:05 AM	94976	13.83	True	Engineering
Marilyn	Female	12/17/2011	8:29 AM	41126	14009.0	True	Sales
Todd	Male	2/18/1990	2:41 AM	49339	1695.0	True	Human Resources

Figura 30: Cabeçalho do arquivo csv com os dados tratados

### 5.3.2. Conclusão da colaboração do *ChatGPT*

A ferramenta *LLM* usada neste trabalho, o *ChatGPT*, desempenhou um papel fundamental na aceleração do processo de análise de dados e tratamento de informações em um conjunto de dados. Ele forneceu suporte em várias etapas do processo, incluindo códigos-fonte para cálculo de estatísticas, tratamento de dados ausentes e geração dos gráficos.

O auxílio desta ferramenta proporcionou uma economia de tempo e recursos, pois gerou soluções personalizadas de acordo com as necessidades do projeto. Além disso, apesar da importante ajuda do *ChatGPT*, é relevante mencionar que alguns dos códigos fonte gerados precisaram de ajustes adicionais para se adaptar completamente ao nosso contexto, mesmo após vários refinamentos do *prompt* enviado ao mesmo.

Durante as interações com o *ChatGPT*, observou-se que embora este não produza diretamente representações gráficas 2D ou 3D, ele é capaz de gerar códigos-fonte em diversas linguagens de programação que, por sua vez, são capazes de criar os mencionados gráficos.

Portanto, nosso objetivo nesta seção: estudar a viabilidade de usar abordagens de *LLM*, para auxiliar na geração de códigos-fonte de linguagens de programação foi alcançado, pois obtivemos os referidos códigos. No entanto, é importante ressaltar que alguns ajustes se mostraram necessários para garantir a qualidade e precisão dos códigos gerados. Ou seja, apesar de precisar realizar, esporadicamente, alguns ajustes nos códigos-fonte retornados, o objetivo foi alcançado, pois o processo de desenvolvimento foi acelerado.

## 6. Estudo de caso do *DMT-AR*

Para este estudo de caso, a tabela "assunto\_consulta" é utilizada como fonte de dados primária. A Figura 31 apresenta uma visão parcial da tabela mencionada, evidenciando os dados ausentes em determinados campos. Conforme, apresentado pelo gráfico virtual no capítulo anterior, seção 5.1.1, a referida tabela possui 917 linhas, destas, a coluna "descricao" apresenta 105 dados ausentes, a coluna "desc\_titulo" 207 e a coluna "desc\_titulo\_formatado" possui 207 dados ausentes também.

COD_AUTORIDADE	DESCRICAO	DESC_TITULO	DESC_TITULO_FORMATADO
814	CRIANCAS E VIOLENCIA	NULL	NULL
816	JUVENTUDE E VIOLENCIA	NULL	NULL
815	DISCIPLINA ESCOLAR BRASIL	NULL	NULL
460	PRATICAS DE ENSINO	NULL	NULL
817	ESCOLAS ORGANIZACAO E ADMINISTRACAO BRASIL	NULL	NULL
817	ESCOLAS ORGANIZACAO E ADMINISTRACAO BRASIL	NULL	NULL
859	ALFABETIZACAO	NULL	NULL
874	COMUNICACAO DE MASSA	NULL	NULL

Figura 31: Visualização parcial de dados ausentes na tabela exemplo

Neste caso deseja-se tratar os dados ausentes, há a opção de selecionar um dos métodos de tratamento disponíveis no menu virtual, apresentado no capítulo anterior na seção 5.1.2. Cada opção, neste menu virtual é vinculado a um arquivo *C#*. Cada um destes arquivos *C#* corresponde a um método específico de tratamento de dados ausentes e atualização da base de dados correspondente. Para este estudo de caso, os dados ausentes da tabela "assunto\_consulta" foram imputados pelo cálculo da moda.

A Figura 32 ilustra, parcialmente, a visualização da tabela Assunto\_Consulta, após a imputação dos dados pelo cálculo da moda. apresentado parcialmente parte dos dados ausentes que foram tratados pelo cálculo estatístico moda, portanto os dados imputados se repetem.



COD_AUTORIDADE	DESCRICAO	DESC_TITULO	DESC_TITULO_FORMATADO
814	CRIANCAS E VIOLENCIA	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
816	JUVENTUDE E VIOLENCIA	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
815	DISCIPLINA ESCOLAR BRASIL	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
460	PRATICAS DE ENSINO	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
817	ESCOLAS ORGANIZACAO E ADMINISTRACAO BRASIL	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
817	ESCOLAS ORGANIZACAO E ADMINISTRACAO BRASIL	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
859	ALFABETIZACAO	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003
874	COMUNICACAO DE MASSA	Vida dos grande brasileiros , A / 2003	VIDA DOS GRANDE BRASILEIROS A 2003

Figura 32: Visualização parcial os dados tratados na tabela exemplo

Nos apêndices estão disponíveis os códigos-fonte utilizados neste estudo de caso:

- Apêndice A: Arquivo *C#* que gera o gráfico virtual;
- Apêndice B: Arquivo *C#* responsável por transferir os dados para o gráfico virtual;
- Apêndice C: Arquivo *C#* responsável pelo tratamento dos dados e atualização da base de dados;
- Apêndice D: Arquivo *PHP* responsável pelos cálculo da moda.
- Apêndice E é demonstrado os últimos *prompts* (refinamentos) enviados ao *ChatGPT* neste estudo de caso. Os código-fontes gerados com os referidos *prompts* foram adaptados ao nosso contexto.

## 7. Considerações finais

Nesta dissertação, apresentamos a estratégia desenvolvida para lidar com dados ausentes. Nossa abordagem consiste em utilizar visualizações de dados e acionar métodos de tratamento de dados, ambos por meio de objetos virtuais com *AR*. Um componente essencial dessa estratégia é a utilização de uma ferramenta de *LLM* para gerar códigos-fonte. Além disso, a estratégia pode ser utilizada no tratamento de outras anomalias nos dados, tais como *outliers*, dados duplicados e erros de formatação. A estratégia descrita compreende um ciclo que tem como ponto de partida o banco de dados. Este ciclo se divide em dois subciclos distintos: um voltado para análises de dados e outro para tratamento de dados. Em ambos subciclos o usuário poderá acessar uma ferramenta *LLM*, solicitar a geração de scripts para outros tipos de gráficos virtuais e outros tipos de métodos de tratamento de dados, além dos já disponíveis. Depois disso o mesmo poderá vincular estes *scripts* a novos cartões físicos ou objetos virtuais.

Para efetivar a estratégia desenvolvemos o *DMT-AR*, um aplicativo de *AR* concebido para implementar integralmente a estratégia. O *DMT-AR* representa uma abordagem dinâmica para identificação, correção e preenchimento de dados ausentes, ampliando suas funcionalidades para acesso a bases de dados, visualizações de dados e análises de dados em tempo real. O aplicativo incorpora códigos-fonte gerados pela ferramenta *ChatGPT*, abrangendo a criação de visualizações em gráficos 3D por meio da *AR* e tratamento de dados ausentes, além da atualização da base de dados. Essa colaboração entre o analista de dados e uma ferramenta *LLM* mostrou-se parcialmente eficaz às operações de visualização e manipulação de dados. No entanto, embora tenhamos realizado diversos refinamentos até alcançar o propósito desejado, em alguns casos, foi necessário ajustar manualmente o código fonte gerado pela ferramenta *LLM*, para adaptar ao nosso contexto, o que justifica o termo 'parcialmente'.

O aplicativo *DMT-AR* enfrenta fatores limitadores significativos. O primeiro refere-se à perda de desempenho nas visualizações conforme a quantidade de dados aumenta. O segundo está relacionado à falta de medição de desempenho entre os métodos de tratamento de dados disponíveis. Embora o aplicativo ofereça diferentes métodos para tratar dados ausentes, a

avaliação comparativa do desempenho desses métodos ainda não está presente. O terceiro é a falta de controles mais precisos dos gráficos virtuais. Por fim a limitação de tratar apenas dados textuais e numéricos.

A revisão realizada revelou que o diferencial distintivo do presente estudo reside em sua abordagem de extensibilidade. Esta característica é fundamental pois permite a incorporação de novas visualizações e novos métodos de tratamento de dados à aplicação, aumentando assim sua capacidade de adaptação e sua relevância contínua em contextos diversos. Além disso, destaca-se o valor adicional proporcionado pela visualização 3D disponibilizada pela *AR*. Essa funcionalidade oferece ao usuário uma experiência de visualização intuitiva e envolvente, permitindo uma compreensão profunda dos dados apresentados. Em uma comparação com gráficos 2D, os gráficos virtuais 3D, tendem a serem mais complexos, são mais difíceis de interpretação e exigem mais recurso computacional. Mesmo com estes desafios, os gráficos 3D oferecem vantagens que compensam. Eles proporcionam um entendimento espacial superior em comparação com os gráficos 2D, permitindo uma exploração mais detalhada em diferentes ângulos. Essa capacidade de representação tridimensional é especialmente útil para visualizar estruturas de dados tornando-os uma escolha gratificante para visualização destas estruturas.

Apesar dos resultados indicarem a existência de limites a serem superados, alcançamos com êxito o objetivo principal deste trabalho: validar o tratamento de dados ausentes com o auxílio da *AR* e apoio de uma ferramenta *LLM*. Limitações identificadas em nossa estratégia, especialmente ao desenvolver o *DMT-AR*, fornecerão diretrizes para futuras investigações. Estas incluem a degradação na qualidade das visualizações à medida que a quantidade de informações apresentadas nos gráficos virtuais aumenta, a necessidade de controles mais precisos dos objetos virtuais e a limitação de tratar apenas dados textuais e numéricos, não incluindo outros tipos de dados. Outros trabalhos futuros incluirão a apresentação virtual dos dados ausentes, pois o *DMT-AR*, apesar de realizar medições, localizações e tratamento dos dados ausentes, ainda não os apresenta ao usuário estes dados; a interação por comando de voz, para uma melhor interação com os objetos virtuais; o refinamento do mecanismo de informações visuais para auxiliar o usuário na tomada de



decisão sobre qual método de tratamento de dados utilizar, pois atualmente, para este fim, o *DMT-AR* apresenta apenas o tipo de dados do conjunto de dados relacionado; e, por último, uma avaliação com analistas de dados para uma avaliação da usabilidade do *DMT-AR*.

## REFERÊNCIAS

- A. Buscemi, A Comparative Study of Code Generation using ChatGPT 3.5 across 10 Programming Languages. 2023 Disponível em: [\[2308.04477\] A Comparative Study of Code Generation using ChatGPT 3.5 across 10 Programming Languages \(arxiv.org\)](https://arxiv.org/abs/2308.04477) [Acessado em: 02/07/2023].
- A. A. Crespo. Estatística Fácil. 19ª Edição. Editora Saraiva. ISBN: 978-85-02-08106-2. São Paulo – SP. 2009.
- A. Indraprastha, "An Interactive Augmented Reality Architectural Design Model : A Prototype for Digital Heritage Preservation," *2019 International Conference on Advanced Computer Science and information Systems (ICACSIS)*, Bali, Indonesia, 2019, pp. 83-88, doi: 10.1109/ICACSIS47736.2019.8979767. Disponível em: <https://ieeexplore.ieee.org/document/8979767>. [Acessado em: 29/03/2023].
- A. Krause-Glau., M. HANSEN and W. HASSELBRING. Collaborative program comprehension via software visualization in extended reality. 2022. Disponível em: <https://www.sciencedirect.com/science/article/pii/S095058492200132X?ref=crajs%20challengefr=RR-1>. [Acessado em: 08/04/2023].
- A. Vaswani, et.al. *Attention is all you need. CoRR, abs/1706.03762, 2017*. Disponível em [1706.03762.pdf \(arxiv.org\)](https://arxiv.org/abs/1706.03762). Acessado em [25 de julho de 2023].
- A. N. Ramaseri Chandra, F. El Jamiy and H. Reza, "Augmented Reality for Big Data Visualization: A Review," *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Las Vegas, NV, USA, 2019, pp. 1269-1274, doi: 10.1109/CSCI49370.2019.00238. Disponível em: <https://ieeexplore.ieee.org/document/9071328>. [Acessado em: 07/03/2023].
- A. Paes, D,Vianna e J, Rodrigues. "Modelos deLinguagens". Disponível em: [Processamento de Linguagem Natural - 15 Modelos de Linguagem \(brasileiraspln.com\)](https://brasileiraspln.com.br/15-Modelos-de-Linguagem). [Acessado em: 11/02/2024].

- A. Shaikh *et al.*, "Coordinate: A Spreadsheet-Programmable Augmented Reality Framework for Immersive Map-Based Visualizations," *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, San Diego, CA, USA, 2019, pp. 134-1343, doi: 10.1109/AIVR46125.2019.00028. Disponível em: <https://ieeexplore.ieee.org/document/8942377>. [Acessado em: 18/03/2023].
- A. Yassien, Y. Emad and S. Abdennadher, "CDVVAR: VR/AR Collaborative Data Visualization Tool," *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Lisbon, Portugal, 2021, pp. 599-600, doi: 10.1109/VRW52623.2021.00182. Disponível em: <https://ieeexplore.ieee.org/document/9419141>. [Acessado em: 14/03/2023].
- B. Bach, R. Sicat, J. Beyer, M. Cordeil and H. Pfister, "The Hologram in My Hand: How Effective is Interactive Exploration of 3D Visualizations in Immersive Tangible Augmented Reality?," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 457-467, Jan. 2018, doi: 10.1109/TVCG.2017.2745941. Disponível em: <https://ieeexplore.ieee.org/document/8019876>. [Acessado em 15/03/2023].
- B. G., R. Ch, K. M., L. B. A. and M. C. V., "Revelation of Geospatial Information using Augmented Reality," *2021 Sixth International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, India, 2021, pp. 303-308, doi: 10.1109/WiSPNET51692.2021.9419459. Disponível em: <https://ieeexplore.ieee.org/documChent/9419459>. [Acessado em: 20/03/2023].
- B. S. Panda and R. Kumar Adhikari, "A Method for Classification of Missing Values using Data Mining Techniques," *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, Gunupur, India, 2020, pp. 1-5, doi: 10.1109/ICCSEA49143.2020.9132935. Disponível em: [A Method for Classification of Missing Values using Data Mining Techniques | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9132935). [Acessado em: 08/02/2024].
- C. E. Faraco. *Grámatica Reformulada*. Editora Ática. ISBN: 978-85-08-10628-8. São Paulo – SP.

- C. Bermejo, Z. Huang, T. Braud and P. Hui, "When Augmented Reality meets Big Data," *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, Atlanta, GA, USA, 2017, pp. 169-174, doi: 10.1109/ICDCSW.2017.62. Disponível em: <https://ieeexplore.ieee.org/document/7979812>. [Acessado em 26/03/2023].
- C. Lim, C. Kim, J. -I. Park and H. Park, "Mobile Augmented Reality Based on Invisible Marker," *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, Merida, Mexico, 2016, pp. 78-81, doi: 10.1109/ISMAR-Adjunct.2016.0045. Disponível em: [Mobile Augmented Reality Based on Invisible Marker | IEEE Conference Publication | IEEE Xplore](#). [Acessado em: 22/03/2022].
- C. Norris. *Large Language Models Like ChatGPT in ABME: Author Guidelines*. 2023. *Ann Biomed Eng* 51, 1121–1122 (2023). Disponível em: <https://doi.org/10.1007/s10439-023-03212-2>. [Acessado em: 05/06/2023].
- D. Aschenbrenner *et al.*, "Mirrorlabs - creating accessible Digital Twins of robotic production environment with Mixed Reality," *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, Utrecht, Netherlands, 2020, pp. 43-48, doi: 10.1109/AIVR50618.2020.00017. Disponível em: <https://ieeexplore.ieee.org/document/9319071>. [Acessado em: 05/04/2023].
- D. C. Shepherd, N. A. Kraft and P. Francis, "Visualizing the "Hidden" Variables in Robot Programs," *2019 IEEE/ACM 2nd International Workshop on Robotics Software Engineering (RoSE)*, Montreal, QC, Canada, 2019, pp. 13-16, doi: 10.1109/RoSE.2019.00007. Disponível em: <https://ieeexplore.ieee.org/document/8823708>. [Acessado em: 03/04/2023].

- D. M. Korngiebel and S.D. Mooney. Considering the possibilities and pitfalls of Generative Pretrained Transformer 3 (GPT-3) in healthcare delivery. 2021. NPJ Digit Med. 2021 Jun 3;4(1):93. doi: 10.1038/s41746-021-00464-x. PMID: 34083689; PMCID: PMC8175735 Disponível em: [Considering the possibilities and pitfalls of Generative Pre-trained Transformer 3 \(GPT-3\) in healthcare delivery - PubMed \(nih.gov\)](#). [Acessado 29 de janeiro de 2023].
- F. B. Rodrigues. Neural Information Extraction Pipeline For Cyber Forensics With Pre-Trained Language Models. Dissertação de Mestrado Profissional submetida ao Departamento de Engenharia da Universidade de Brasília. 2022. Disponível em: [2022\\_FillipeBarrosRodrigues.pdf \(unb.br\)](#). [Acessado em 30 de janeiro de 2023].
- F. Laera, "Augmented Reality for Easy Sailing," *2020 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Recife, Brazil, 2020, pp. 312-313, doi: 10.1109/ISMAR-Adjunct51615.2020.00088. Disponível em: <https://ieeexplore.ieee.org/document/9287823>. [Acessado em: 05/03/2023].
- G. Caggianese, V. Colonnese and L. Gallo, "Situated Visualization in Augmented Reality: Exploring Information Seeking Strategies," *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Sorrento, Italy, 2019, pp. 390-395, doi: 10.1109/SITIS.2019.00069. Disponível em: <https://ieeexplore.ieee.org/document/9067881>. [Acessado em: 10/03/2022].
- G. Caggianese, V. Colonnese and L. Gallo, "Situated Visualization in Augmented Reality: Exploring Information Seeking Strategies," *2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, Sorrento, Italy, 2019, pp. 390-395, doi: 10.1109/SITIS.2019.00069. Disponível em: [Situated Visualization in Augmented Reality: Exploring Information Seeking Strategies | IEEE Conference Publication | IEEE Xplore](#). [Acessado em 17/03/2023].

- H. S. Narman *et al.*, "Augmented Reality for Teaching Data Structures in Computer Science," *2020 IEEE Global Humanitarian Technology Conference (GHTC)*, Seattle, WA, USA, 2020, pp. 1-7, doi: 10.1109/GHTC46280.2020.9342932. Disponível em: <https://ieeexplore.ieee.org/document/9342932>. [Acessado em 20/03/2023].
- I. Ekanayake and S. Gayanika, "Data Visualization Using Augmented Reality for Education: A Systematic Review," *2022 7th International Conference on Business and Industrial Research (ICBIR)*, Bangkok, Thailand, 2022, pp. 533-537, doi: 10.1109/ICBIR54589.2022.9786403. Disponível em: <https://ieeexplore.ieee.org/document/9786403>. [Acessado em: 08/03/2023].
- I. Victor Costa, V. Favacho Queiroz, B. Pinto Miranda, A. Abreu de Freitas, C. Gustavo Resque dos Santos and B. Serique Meiguins, "A Card-Based Interaction to Design Visualizations in Augmented Reality Environments," *2019 23rd International Conference in Information Visualization ??? Part II*, Adelaide, SA, Australia, 2019, pp. 52-55, doi: 10.1109/IV-2.2019.00019. Disponível em: <https://ieeexplore.ieee.org/document/8811944>. [Acessado em: 09/03/2023].
- J. M. Teixeira and F. F. F. Peres, "Analysis of XR Research in Brazil from 21 Years of SVR Publications," *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, Porto de Galinhas, Brazil, 2020, pp. 220-231, doi: 10.1109/SVR51698.2020.00043. Disponível em: <https://ieeexplore.ieee.org/document/9262681>. [Acessado em: 05/04/2023].
- J. Withe, et al. A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT . 2023. Disponível em: [2302.11382.pdf \(arxiv.org\)](https://arxiv.org/abs/2302.11382). [Acessado em 12 de março de 2023].
- J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art. *Psychol Methods*." 2002. Jun; 7(2):147-77. PMID: 12090408. Disponível em: [Missing data: our view of the state of the art - PubMed \(nih.gov\)](https://pubmed.ncbi.nlm.nih.gov/12090408/) [Acessado em 07 de janeiro de 2023].
- J. Sjöbergh and Y. Tanaka, "Visualizing Missing Values," *2017 21st International Conference Information Visualisation (IV)*, London, UK, 2017, pp. 242-249, doi: 10.1109/iV.2017.12. Disponível em: [Visualizing Missing Values | IEEE Conference Publication | IEEE Xplore.](https://ieeexplore.ieee.org/document/8000000) [Acessado em 07/02/2024].

- K.M. Lang e T.D. Little."Principled Missing Data Treatments". doi: 10.1007/s11121-016-0644-5. 2016. Disponível em: [Principled Missing Data Treatments | Prevention Science \(springer.com\)](https://www.springer.com). [Acessado em 13/02/2024].
- K., Kim. G. Lee, J. Kea. AR-based visualization tool for visualizing the high resolution 3D structure of brain cells. 2019. Disponível em: <https://www.sciencedirect.com/science/article/pii/S2451830119316875?via%3Dihub>. [Acessado em: 08/04/2023].
- K. Psychogyios, L. Ilias and D. Askounis, "Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset," *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Ioannina, Greece, 2022, pp. 1-5, doi: 10.1109/BHI56158.2022.9926882. Disponível em: [Comparison of Missing Data Imputation Methods using the Framingham Heart study dataset | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/9926882). [Acessado em: 23/03/202].
- L. Gruenwald *et.al.* "Using data mining to handle missing data in multi-hop sensor network applications".2010. doi: 10.1145/1850822,1850825. Disponível em [Using data mining to handle missing data in multi-hop sensor network applications | Proceedings of the Ninth ACM International Workshop on Data Engineering for Wireless and Mobile Access](https://doi.org/10.1145/1850822,1850825) . [Acessado em: 12/02/2024].
- L. A. Sulisty, T. F. Kusumasari and E. N. Alam, "Implementation of Data Cleansing Pattern Module for Data Quality Management Application using Open Source Tools," 2020 3rd International Conference on Computer and Informatics Engineering (IC2IE), Yogyakarta, Indonesia, 2020, pp. 7-12, doi: 10.1109/IC2IE50715.2020.9274628. Disponível em: [Implementation of Data Cleansing Pattern Module for Data Quality Management Application using Open Source Tools | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/9274628). Acessado em 24/03/2023.

- L. B. Othman and S. B. Yahia, "A multiple criteria evaluation technique for missing values imputation," *2018 12th International Conference on Research Challenges in Information Science (RCIS)*, Nantes, France, 2018, pp. 1-12, doi: 10.1109/RCIS.2018.8406659. Disponível em: <https://ieeexplore.ieee.org/abstract/document/8406659> [Acessado em 07 de janeiro de 2023].
- L. N. Nunes, J.M.G. Fachael, M.M. Kluck, "Uso da imputação múltipla de dados ausentes: uma simulação utilizando dados epidemiológicos 2009." Disponível em: <https://www.scielo.br/j/csp/a/XW3NwV7T5VL77WN7d7TJ3ZR/abstract/?lang=pt> [Acessado em 24/01/2023].
- L. S. O. Rezende, P. H. M. Sá, M. C. F. Macedo, A. L. Apolinário, I. Winkler and M. A. Moret S. G., "Volume Rendering: An Analysis based on the HoloLens Augmented Reality Device," *2020 22nd Symposium on Virtual and Augmented Reality (SVR)*, Porto de Galinhas, Brazil, 2020, pp. 35-38, doi: 10.1109/SVR51698.2020.00021. Disponível em: <https://ieeexplore.ieee.org/document/9262664>. [Acessado em: 05/03/2023].
- L. Sharma, S. Anand, N. Sharma and S. K. Routry, "Visualization of Big Data with Augmented Reality," *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)*, Madurai, India, 2021, pp. 928-932, doi: 10.1109/ICICCS51141.2021.9432383. Disponível em: <https://ieeexplore.ieee.org/document/9432383>. [Acessado em: 25/03/2023].
- L. Zhang, S. Chen, H. Dong and A. El Saddik, "Visualizing Toronto City Data with HoloLens: Using Augmented Reality for a City Model," in *IEEE Consumer Electronics Magazine*, vol. 7, no. 3, pp. 73-80, May 2018, doi: 10.1109/MCE.2018.2797658. Disponível em: <https://ieeexplore.ieee.org/document/8332919>. [Acessado em: 22/03/2023].



- M. C. R. Harrington, "Connecting User Experience to Learning in an Evaluation of an Immersive, Interactive, Multimodal Augmented Reality Virtual Diorama in a Natural History Museum & the Importance of Story," *2020 6th International Conference of the Immersive Learning Research Network (iLRN)*, San Luis Obispo, CA, USA, 2020, pp. 70-78, doi: 10.23919/iLRN47897.2020.9155202. Disponível em: <https://ieeexplore.ieee.org/document/9155202>. [Acessado em: 13/03/2023].
- M. Dani, G. Garg, R. Perla and R. Hebbalaguppe, "Mid-Air Fingertip-Based User Interaction in Mixed Reality," *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Munich, Germany, 2018, pp. 174-178, doi: 10.1109/ISMAR-Adjunct.2018.00061. Disponível em: <https://ieeexplore.ieee.org/document/8699224>. [Acessado em: 29/03/2023].
- M. Karthika, J. Pearl and J. Tian, S. Feiner, S. Julier and B. MacIntyre, "Graphical models for inference with missing data" in *Advances in neural information processing systems*. Volume 16. 2013. Disponível em: [https://escholarship.org/content/qt1464x5ph/qt1464x5ph\\_noSplash\\_ff1eb38e5d96d0572d0c2c0049368168.pdf](https://escholarship.org/content/qt1464x5ph/qt1464x5ph_noSplash_ff1eb38e5d96d0572d0c2c0049368168.pdf) [Acessado em 10 de janeiro de 2023].
- M. S. Hounsell e R. Tori. *Introdução a realidade virtual e aumentada*. ISBN: 978-85-7669-446-5. 2018. Disponível em: [Introdução a Realidade Virtual e Aumentada \(ufpb.br\)](https://www.ufpb.br/~inf/realidade-virtual-e-aumentada/) [Acessado em 20 de janeiro de 2023].
- M. Shanahan, K. McDonell e L. Reynolds. *Role play with large language models*. 2023. Disponível em: [Role play with large language models | Nature](https://www.nature.com/articles/d41586-023-00000-0) [Acessado em 07 de dezembro de 2023].
- M. Tatzgern. *Situated Visualization in Augmented Reality*. Doctoral Thesis. 2015. Disponível em: [Situated Visualization for Augmented Reality \(print-version\) \(researchgate.net\)](https://www.researchgate.net/publication/309111111_Situated_Visualization_for_Augmented_Reality) [Acessado em 20 de janeiro de 2023].

- O. M. Elrawi, "The Use of Mixed-Realities Techniques for the Representation of Islamic Cultural Heritage," *2017 International Conference on Machine Vision and Information Technology (CMVIT)*, Singapore, 2017, pp. 58-63, doi: 10.1109/CMVIT.2017.16. Disponível em: <https://ieeexplore.ieee.org/abstract/document/7878715>. [Acessado em: 08/04/2023].
- P. Knierim, F. Kiss and A. Schmidt, "Look Inside: Understanding Thermal Flux Through Augmented Reality," *2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Munich, Germany, 2018, pp. 170-171, doi:10.1109/ISMAR-Adjunct.2018.00059. Disponível em: <https://ieeexplore.ieee.org/document/8699216>. [Acessado em: 08/03/2023].
- P. Reipschlagel, T. Flemisch and R. Dachsel, "Personal Augmented Reality for Information Visualization on Large Interactive Displays," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1182-1192, Feb. 2021, doi: 10.1109/TVCG.2020.3030460. Disponível em: <https://ieeexplore.ieee.org/document/9223669>. [Acessado em 20/03/2023].
- P. Vateekul and K. Sarinnapakorn, "Tree-Based Approach to Missing Data Imputation," *2009 IEEE International Conference on Data Mining Workshops*, Miami, FL, USA, 2009, pp. 70-75, doi: 10.1109/ICDMW.2009.92. Disponível em: [Tree-Based Approach to Missing Data Imputation | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/5188888) [Acessado em 25 de janeiro de 2023].
- Q. Liu and M. Hauswirth, "A Provenance Meta Learning Framework for Missing Data Handling Methods Selection," *2020 11th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, New York, NY, USA, 2020, pp. 0349-0358, doi: 10.1109/UEMCON51285.2020.9298089. Disponível em: <https://ieeexplore.ieee.org/document/9298089> [Acessado em 07 de janeiro de 2023].

- R. Azuma, Y. Baillet, R. Behringer, S. Feiner, S. Julier and B. MacIntyre, "Recent advances in augmented reality," in *IEEE Computer Graphics and Applications*, vol. 21, no. 6, pp. 34-47, Nov.-Dec. 2001, doi: 10.1109/38.963459. Disponível em: <https://ieeexplore.ieee.org/abstract/document/963459> [Acesso em 02 janeiro de 2023].
- R. Guarese *et al.*, "A Usability Assessment Of Augmented Situated Visualization," *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, Atlanta, GA, USA, 2020, pp. 636-637, doi: 10.1109/VRW50115.2020.00169. Disponível em: <https://ieeexplore.ieee.org/document/9090420>. [Acessado em: 05/04/2023].
- R. Pavithrakannan, N. B. Fenn, S. Raman, V. Kalyanaraman, V. K. Muruganathan and J. Janarthanan, "Imputation Analysis of Central Tendencies for Classification," *2021 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, Toronto, ON, Canada, 2021, pp. 1-7, doi: 10.1109/IEMTRONICS52119.2021.9422507. Disponível em: [Imputation Analysis of Central Tendencies for Classification | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/document/9422507) [Acessado em 11 de janeiro de 2023].
- R., Raphael. A. R. Avila. J. Brunelle. B.M. Camara. How Secure is Code Generated by ChatGPT?. 2023. Disponível em: [\[2304.09655\] How Secure is Code Generated by ChatGPT? \(arxiv.org\)](https://arxiv.org/abs/2304.09655). [Acessado em: 08/04/2023].
- R. Sicat *et al.*, "DXR: A Toolkit for Building Immersive Data Visualizations," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 715-725, Jan. 2019, doi: 10.1109/TVCG.2018.2865152. Disponível em: <https://ieeexplore.ieee.org/document/8440858>. [Acessado em: 25/03/2023].
- S. N. Kundu, N. Muhammad and F. Sattar, "Using the augmented reality sandbox for advanced learning in geoscience education," *2017 IEEE 6th International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, Hong Kong, China, 2017, pp. 13-17, doi: 10.1109/TALE.2017.8252296. Disponível em: <https://ieeexplore.ieee.org/document/8252296>. [Acessado em 14/03/2023].

- S. A. Hirve, A. Kunjir, B. Shaikh and K. Shah, "An approach towards data visualization based on AR principles," *2017 International Conference on Big Data Analytics and Computational Intelligence (ICBDAC)*, Chirala, Andhra Pradesh, India, 2017, pp. 128-133, doi: 10.1109/ICBDACI.2017.8070822. Disponível em: <https://ieeexplore.ieee.org/document/8070822>. [Acessado em: 10/03/2023].
- S. L. Kim, H. J. Suk, J. H. Kang, J. M. Jung, T. H. Laine and J. Westlin, "Using Unity 3D to facilitate mobile augmented reality game development," *2014 IEEE World Forum on Internet of Things (WF-IoT)*, Seoul, Korea (South), 2014, pp. 21-26, doi: 10.1109/WF-IoT.2014.6803110. Disponível em: <https://ieeexplore.ieee.org/document/6803110> [Acessado em 01 de fevereiro de 2023].
- S. S. Biswas. *Potential Use of Chat GPT in Global Warming*. *Annals of Biomedical Engineering*, Vol. 51, No. 6, June 2023 ( 2023) pp. 1126–1127. Disponível em: [Potential Use of Chat GPT in Global Warming \(springer.com\)](https://www.springer.com) [Acessado em 12 de agosto de 2023].
- S. Shigeyuki et al. *A Bayesian missing value estimation method for gene expression profile data*. *Bioinformatics*, v. 19, n. 16, p. 2088-2096. 2003. Disponível em: [Bayesian missing value estimation method for gene expression profile data | Bioinformatics | Oxford Academic \(oup.com\)](https://academic.oup.com/bioinformatics/article/19/16/2088/2096) [Acessado em 12 de fevereiro de 2023].
- S. Xin, S. Qingting, L. Zhiqiang and C. Tengfei, "Application of 3D tracking and registration in exhibition hall navigation interaction," *2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*, Chongqing, China, 2020, pp. 109-113, doi: 10.1109/ICICAS51530.2020.00030. Disponível em: <https://ieeexplore.ieee.org/document/9402915>. [Acessado em 01/04/2023].
- S. Zollmann, T. Langlotz, R. Grasset, W. H. Lo, S. Mori and H. Regenbrecht, "Visualization Techniques in Augmented Reality: A Taxonomy, Methods and Patterns," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 9, pp. 3808-3825, 1 Sept. 2021, doi: 10.1109/TVCG.2020.2986247. Disponível em: <https://ieeexplore.ieee.org/document/9060980>. [Acessado em: 18/03/2023].

- T. Lin, Y. Yang, J. Beyer and H. Pfister, "Labeling Out-of-View Objects in Immersive Analytics to Support Situated Visual Searching," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 3, pp. 1831-1844, 1 March 2023, doi: 10.1109/TVCG.2021.3133511. Disponível em: <https://ieeexplore.ieee.org/document/9645242> [Acessado em: 31/03/2023].
- T. Andrade and D. Bastos, "Extended Reality in IoT scenarios: Concepts, Applications and Future Trends," *2019 5th Experiment International Conference (exp.at'19)*, Funchal, Portugal, 2019, pp. 107-112, doi: 10.1109/EXPAT.2019.8876559. Disponível em: <https://ieeexplore.ieee.org/document/8876559>. [Acessado em: 03/04/2023].
- T. Makaba and E. Dogo, A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms, 2019 International Multidisciplinary Information Technology and Engineering Conference (IMITEC), Vanderbijlpark, South Africa, 2019, pp. 1-7, doi: 10.1109/IMITEC45504.2019.9015889. Disponível em: [A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms | IEEE Conference Publication | IEEE Xplore](#) [Acessado em 12 de janeiro de 2023].
- T. Ohta, R. Ito and Y. Kakuda, "Design of a Node Status Visualizing Software Utilizing the AR Technology for Multihop Wireless Networks," *2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC)*, Turin, Italy, 2017, pp. 270-271, doi: 10.1109/COMPSAC.2017.101. Disponível em: <https://ieeexplore.ieee.org/document/8029934>. [Acessado em: 01/04/2023].
- Z. Chen, Y. Su, Y. Wang, Q. Wang, H. Qu and Y. Wu, "MARVisT: Authoring Glyph-Based Visualization in Mobile Augmented Reality," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 8, pp. 2645-2658, 1 Aug. 2020, doi: 10.1109/TVCG.2019.2892415. Disponível em: <https://ieeexplore.ieee.org/document/8611113>. [Acessado em: 13/03/2023].

- Z. Xia, "An interactive algorithm for augmented reality experiment," *2022 2nd International Conference on Computer Graphics, Image and Virtualization (ICCGIV)*, Chongqing, China, 2022, pp. 208-211, doi: 10.1109/ICCGIV57403.2022.00048. Disponível em: [An interactive algorithm for augmented reality experiment | IEEE Conference Publication | IEEE Xplore](#) [Acessado em: 17/03/2023].
- W. Willett, Y. Jansen and P. Dragicevic, "Embedded Data Representations," in *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 461-470, Jan. 2017, doi: 10.1109/TVCG.2016.2598608. Disponível em: <https://ieeexplore.ieee.org/document/7539328>. [Acessado em: 07/04/2023].

## Apêndice A:Arquivo C# que gera o gráfico virtual

```

//este código-fonte é disponibilizado pelo criador do gráfico virtual de forma gratuita em:
//3D Interactive Barchart | GUI Tools | Unity Asset Store [Acessado em 06/03/2024]

using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using System;
using System.Linq.Expressions;
using BarGraph.VittorCloud;

public class BarGraphExample : MonoBehaviour
{
    // public data set for inserting data into the bar graph
    public List<BarGraphDataSet> exampleDataSet;
    BarGraphGenerator barGraphGenerator;

    void Start() {
        barGraphGenerator = GetComponent<BarGraphGenerator>();
        //if the exampleDataSet list is empty then return.
        if (exampleDataSet.Count == 0)
        {
            Debug.LogError("ExampleDataSet is Empty!");
            return;
        }
        barGraphGenerator.GeneratBarGraph(exampleDataSet); }
    //call when the graph starting animation completed,
    // for updating the data on run time
    public void StartUpdatingGraph() {

        StartCoroutine(CreateDataSet()); }
    IEnumerator CreateDataSet() {
        // yield return new WaitForSeconds(3.0f);
        while (true) {
            GenerateRandomData();
            yield return new WaitForSeconds(2.0f); } }

```

## Apêndice B: Arquivo C# responsável por transferir os dados para o gráfico virtual

```

using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using System;
using System.Linq.Expressions;
using BarGraph.VittorCloud;
using UnityEngine.Networking;
public class BarGraphExemple1 : MonoBehaviour {
public List<BarGraphDataSet> exampleDataSet;
    BarGraphGenerator barGraphGenerator;
    //conecta a base dados e captura os dados
    string url = "http://localhost/dbagenda/unity_assunto_consulta.php";
    public string[] userData;
    public string[] dadosBanco;
    public string[] dadosBanco2;
    public float[] dados = new float[] { };
    public IEnumerator Start2() {
        WWW users = new WWW(url);
        yield return users;
        string userDataString = users.text;
        userData = userDataString.Split(';');
        for (int i = 0; i < userData.Length; i++)
        {
            Debug.Log(userData[i] + " eh o valor esperado "); } }

IEnumerator GetRequest(string uri) {
    using (UnityWebRequest webRequest = UnityWebRequest.Get(uri)) {
        // Request and wait for the desired page.
        yield return webRequest.SendWebRequest();
        string[] userData = uri.Split(';');
        int page = userData.Length - 1;
        switch (webRequest.result) {
            case UnityWebRequest.Result.ConnectionError:
            case UnityWebRequest.Result.DataProcessingError:
                Debug.LogError(userData[page] + ": Error: " + webRequest.error);
                break;
            case UnityWebRequest.Result.ProtocolError:
                Debug.LogError(userData[page] + ": HTTP Error: " + webRequest.error);
                break;
            case UnityWebRequest.Result.Success:
                Debug.Log(userData[page] + "\n Dados recebidos do BD: " +
                    webRequest.downloadHandler.text);
                dadosBanco = new String[] { webRequest.downloadHandler.text };
                dadosBanco2 = dadosBanco[0].Split(';');
                dados = new float[dadosBanco2.Length];

                for (int i = 0; i < dadosBanco2.Length - 1; i++) {
                    dados[i] = float.Parse(dadosBanco2[i]); }
                barGraphGenerator = GetComponent<BarGraphGenerator>();

```



```
exampleDataSet.Clear();
GenerateRandomData();
//if the exampleDataSet list is empty then return.
if (exampleDataSet.Count == 0) {
    Debug.LogError("ExampleDataSet is Empty!"); }
barGraphGenerator.GeneratBarGraph(exampleDataSet);
break; } } }
```

## Apêndice C: Arquivo C# responsável pelo tratamento dos dados e atualização da base de dados

```
using System.Collections;
using System.Collections.Generic;
using UnityEngine;
using System;
using System.Linq.Expressions;
using BarGraph.VittorCloud;
using UnityEngine.Networking;

public class atualizaDados : MonoBehaviour {
    string url = "http://localhost/dbagenda/moda_assunto_consulta.php";
    IEnumerator GetRequest(string uri) {
        using (UnityWebRequest webRequest = UnityWebRequest.Get(uri))
            yield return webRequest.SendWebRequest(); }
    // Start is called before the first frame update

    void Start() {
        Debug.Log("botao apertado - metodo start"); } //apenas para testar botao
    void Update() { }
    public void atualiza_Tabela_Unity () {
        StartCoroutine(GetRequest("http://localhost/dbagenda/moda_assunto_consulta.php"));
        Debug.Log("botao apertado -atualizaDados na tabela unity_assunto_consulta"); }}
```

Apêndice D: Arquivo *PHP* responsável pelos cálculos da moda

```

<?php // moda_assunto_consulta.php acessado pelo C# anterior
//moda_assunto_consulta.php
include 'connection.php';
function calcularModa($array) {
    $contagem = array_count_values($array);
    arsort($contagem);
    return key($contagem);}
// Query para obter os valores das três colunas e filtrar os valores nulos
$query = "SELECT cod_autoridade, desc_titulo, desc_titulo_formatado FROM assunto_consulta
WHERE cod_autoridade IS NOT NULL AND desc_titulo IS NOT NULL AND desc_titulo_formatado IS NOT NULL";
$result = mysqli_query($connect, $query);
// Inicializa arrays para armazenar os valores das colunas
$cod_autoridade_values = array();
$desc_titulo_values = array();
$desc_titulo_formatado_values = array();
// Obtém os valores das colunas
while ($row = mysqli_fetch_assoc($result)) {
    $cod_autoridade_values[] = $row['cod_autoridade'];
    $desc_titulo_values[] = $row['desc_titulo'];
    $desc_titulo_formatado_values[] = $row['desc_titulo_formatado'];}
// Calcula a moda para cada coluna
$moda_cod_autoridade = calcularModa($cod_autoridade_values);
$moda_desc_titulo = calcularModa($desc_titulo_values);
$moda_desc_titulo_formatado = calcularModa($desc_titulo_formatado_values);
// Substitui os valores nulos pelas modas correspondentes
$query = "UPDATE assunto_consulta SET cod_autoridade = COALESCE(cod_autoridade,
'$moda_cod_autoridade'), desc_titulo = COALESCE(desc_titulo, '$moda_desc_titulo'), desc_titulo_formatado
= COALESCE(desc_titulo_formatado, '$moda_desc_titulo_formatado')";

// Executa a query para substituir os valores nulos pelas modas correspondentes
if (mysqli_query($connect, $query)) {
    echo "Valores nulos substituídos pelas modas correspondentes com sucesso. <br>";
} else {
    echo "Erro ao executar a query para substituir valores nulos: " . mysqli_error($connect) . "<br>";
}
// Query para atualizar a tabela unity_assunto_consulta trocando todos os valores por 0
$query = "UPDATE unity_assunto_consulta SET cod_autoridade = 0, desc_titulo = 0,
desc_titulo_formatado = 0";
// Executa a query para atualizar a tabela unity_assunto_consulta
if (mysqli_query($connect, $query)) {
    echo "Tabela unity_assunto_consulta atualizada com sucesso. <br>";
} else {
    echo "Erro ao atualizar a tabela unity_assunto_consulta: " . mysqli_error($connect) . "<br>";
}

// Substituindo todos os valores da tabela unity_assunto_consulta por 0    $sql_replace_zero =
"UPDATE unity_assunto_consulta SET cod_autoridade = 0, desc_titulo = 0, desc_titulo_formatado = 0";

if ($connect->query($sql_replace_zero) === TRUE) {        echo "Valores substituídos por 0 na
tabela unity_assunto_consulta.\n";    } else {

```

```
echo "Erro ao substituir valores por 0 na tabela unity_assunto_consulta: " . $connect->error; }  
} else { echo "Nenhum resultado encontrado.";}  
mysqli_close($connect);  
?>
```

Apêndice E: Prompts enviados ao *ChatGPT* do estudo de caso

Gere um arquivo php que acessa a tabela `assunto_consulta`. A conexão é feita pelo arquivo `connection.php`. Retorne os dados as colunas `cod_autoridade`, `desc_titulo` e `desc_titulo_formatado` onde os dados não possuem valores nulos. Depois calcule a moda para cada uma das colunas e apresente os valores destas modas. Na sequencia onde os dados forem nulos substitua pelo valor da moda calculado, para cada coluna.

\*\*\*\*\*

crie um arquivo `c#` para acessar o endereço: `http://localhost/dbagenda/moda_assunto_consulta.php`

\*\*\*\*\*

Gere um arquivo php que acessa a tabela `assunto_consulta`. A conexão é feita pelo arquivo `connection.php` com a função `connect()`.

Agora acrescente: retorne os dados as colunas `cod_autoridade`, `desc_titulo` e `desc_titulo_formatado` onde os dados não possuem valores nulos. Depois calcule a moda para cada uma das colunas e apresente os valores destas modas.

Acrescente: onde os dados forem nulos substitua pelo valor da moda calculado, para cada coluna. Depois substitua todos os valores da tabela `unity_assunto_consulta` por 0. Os valores da moda de cada coluna deverá ser inserida na tabela `unity_assunto_consulta`.