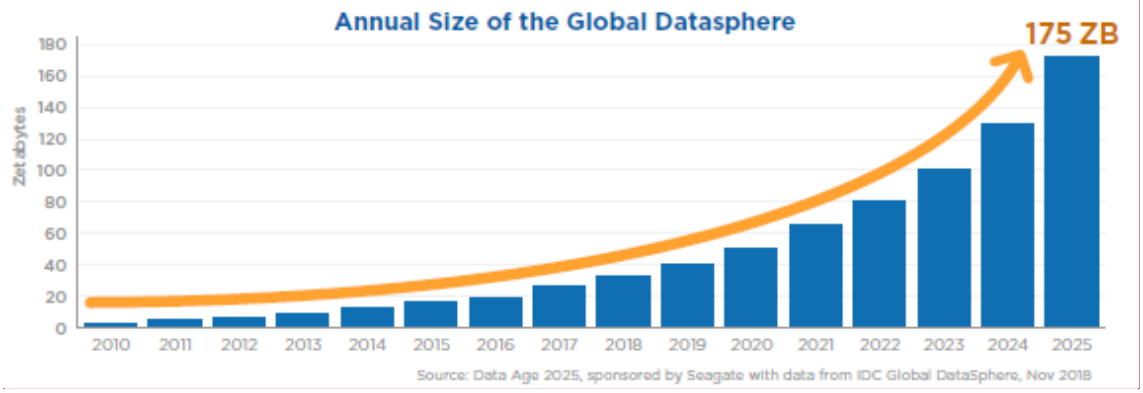
Índice

1. Introdução 3
2. Preparação de dados 4
   2.1. Extração 4
   2.2. Limpeza 4
   2.3. Seleção de variáveis 4
   2.4. Transformação 5
   2.5. Organização 5
   2.6. Técnicas 5
3. Ferramentas de exploração e visualização de dados 6
4. Análise exploratória de dados 7
5. Técnicas de visualização de dados 8
6. Casos de uso de exploração e visualização de dados 13
7. Desafios e tendências 13
8. Conclusão 13
9. Bibliografia 13

1. Introdução

Com o crescimento das novas tecnologias, o surgimento da internet das coisas, a rapidez com que a informação circula e a capacidade crescente de criação, assim como, com o armazenamento de dados, estima-se que em 2025 o tamanho da Esfera de Dados Global seja de 175 ZB [1] [2]. Tendo em conta que a exploração e visualização de dados é utilização de técnicas para apresentar e analisar informação, por vezes extremamente complexa, tanto pela sua estrutura como pela sua quantidade, como referido anteriormente, de forma transparente[TF1]. Isso faz com que a exploração de dados e a sua visualização, nos próximos anos, assuma um papel cada vez mais importante por esta ser uma ferramenta poderosa para explorar, entender e comunicar informação complexa.

Figure 1 – Annual Size of the Global Datasphere



[TF2]

Como referido anteriormente, devido a grande quantidade de dados que possuímos atualmente, pode ser difícil extrair conhecimento significativo apenas olhando para números e tabelas. É aí que a visualização de dados entra em jogo, permitindo que os usuários transformem informações em gráficos e diagramas facilmente compreensíveis.[TF3] Além de ajudar a entender os dados, a visualização também tem o poder de revelar padrões, tendências e anomalias que podem não ser aparentes em formatos de dados brutos [3]. Ao visualizar dados, podemos descobrir relações e informação que não seriam detetáveis de outra forma.[TF4] Ao longo deste relatório iremos abordar a exploração e a visualização de dados, em três partes: preparação de dados, exploração de dados, e visualização de dados. Estes temas estão intimamente ligados sendo difícil de os separar em processos destintos, segundo a biografia. Será também apresentado algumas das ferramentas de exploração e visualização de dados usadas nos dias atuais, como exemplos reais onde estas técnicas são usadas[TF5]. [TF6]

2. Preparação de dados

Muitas vezes os dados que possuímos são derivados de texto, tabelas ou base de dados e são nos apresentado em bruta, isto é, com valores em falta, erros, distorções, entre outros problemas. É por isso que a preparação de dados é um passo importante para a exploração e visualização de dados, pois é nesta fase que se trabalha a qualidade dos dados, que por vezes pode ser um processo demoroso e chato. É aqui que vamos determinar o que são os dados, melhorar a sua qualidade, padronizar, consolidar e transformá-los para que estes sejam uteis para posterior analise. A preparação de dados pode ser dividida por vários passos [4]–[7]:

2.1. Extração

A extração é onde os dados são obtidos, seja através da recolha de dados brutos, importação de dados de fontes externas ou através de arquivos armazenados em sistemas de gestão de base de dados. É importante garantir que a fonte dos dados esteja fiável e que os dados sejam recolhidos de forma consistente [7], [8].

2.2. Limpeza

Após a extração, vem a fase de limpeza, onde os dados são submetidos a uma série de tratamentos para eliminar dados duplicados, corrigir valores incorretos, preencher valores em falta e eliminar valores discrepantes. Este é um passo crucial na preparação de dados, pois dados incorretos ou incompletos podem levar a análises imprecisas e conclusões erradas [4]–[8].[TF7]

2.3. Seleção de variáveis

Na seleção de variáveis, o foco é identificar as variáveis que são relevantes para a análise, eliminando as que não têm importância ou são redundantes. Isso ajuda a simplificar a análise, tornando-a mais eficiente [9].

2.4. Transformação

A transformação de dados pode incluir a normalização, discretização e agregação de dados, bem como a conversão de formatos de dados. Este passo é importante para garantir que os dados sejam comparáveis e possam ser analisados em conjunto [4]–[8].

2.5. Organização

Por fim, a organização dos dados é feita para garantir que os dados estejam num formato que seja fácil de trabalhar, incluindo a ordenação dos dados, a definição de tipos de dados e a criação de tabelas de base de dados. Esta fase é importante para garantir que os dados sejam acessíveis e possam ser facilmente integrados em ferramentas de análise e visualização [4]–[7].

2.6. Técnicas

Extração	Limpeza	Seleção de variáveis	Transformação	Organização
Web scraping	Remoção de Dados Duplicados	Análise de correlação	Normalização	Ordenação
Exportação de banco de dados	Remoção de Outliers[9]	Análise de importância	Discretização	Agregação
	Preenchimento de valores em falta		Transformação de variáveis não numéricas[TF8]	

3. Ferramentas de exploração e visualização de dados

As ferramentas de exploração e visualização de dados são fundamentais para ajudar os profissionais a entenderem e apresentarem os dados de forma clara e compreensível, poupando-lhes muito tempo. Existem várias opções no mercado, das quais [1]:

- Tableau - Ferramenta paga e popular para análise e visualização de dados, conhecida pela sua facilidade de uso e capacidade de criar visualizações interativas;
- Excel - Software de folhas de calculo amplamente utilizado que oferece recursos para análise e visualização de dados, como filtros, tabelas dinâmicas e gráficos;
- Power BI - Ferramenta paga da Microsoft que oferece recursos de análise e visualização de dados, como visualizações interativas, relatórios e painéis;
- Python - Linguagem de programação open sorce que pode ser usada para análise e visualização de dados, com várias bibliotecas disponíveis, tais como Pandas e Matplotlib. Atualmente, é das linguagens mais utilizadas;
- R - Linguagem de programação estatística open sorce com muitas bibliotecas disponíveis para análise e visualização de dados;
- QlikView - Ferramenta paga, conhecida por sua capacidade de lidar com grandes conjuntos de dados e por suas visualizações interativas;
- SAP BusinessObjects - Ferramenta paga que oferece ferramentas para gerenciamento de dados, análise e geração de relatórios[TF9][TF10].

4. Análise exploratória de dados

A análise exploratória de dados (AED) é uma etapa fundamental do processo de análise de dados, que tem como objetivo principal entender e explorar os dados disponíveis antes de se aplicar qualquer modelo estatístico ou técnicas machine learning. Esta envolve a aplicação de ferramentas estatísticas e algoritmos para identificar características, padrões, distribuições e relacionamentos nos dados, bem como para verificar suposições sobre as distribuições dos dados. É aqui que podemos verificar tendências e anomalias. De realçar que a AED está intimamente ligada com a visualização de dados, decidimos então neste relatório abordá-la separadamente na próxima secção.

Para realizar uma análise exploratória de dados, é importante fazer a preparação dos mesmos, como foi falado anteriormente. Após a verificação da qualidade dos dados, a próxima etapa é a análise exploratória propriamente dita.

Nessa etapa, é possível utilizar diversas técnicas para entender os dados disponíveis. Algumas das técnicas mais comuns incluem a criação de gráficos de frequência e histogramas para entender a distribuição dos dados, abordaremos estes mais a frente na secção 5. Técnicas de visualização de dados, a aplicação de medidas descritivas como média, desvio padrão e correlação para entender a relação entre variáveis e a identificação valores extremos.

Outras técnicas comuns incluem a análise de séries temporais que nos permitem entender tendências e variações ao longo do tempo e a aplicação de técnicas de clusterização usadas para identificar grupos de dados similares e a análise de componentes principais que nos permite entender a estrutura dos dados.[10]

É importante ressaltar que a análise exploratória de dados é uma etapa importante do processo de análise de dados, mas não é suficiente para tirar conclusões definitivas sobre os dados. É preciso combinar as informações obtidas na AED com técnicas estatísticas ou de machine learning para realizar uma análise mais completa.

5. Técnicas de visualização de dados

Devido ao poder do olho humano em detetar padrões, a visualização de dados é uma ferramenta importante para a análise e comunicação de informações complexas. Por vezes, a melhor maneira de entender grandes conjuntos de dados é através de gráficos. O maior desafio será saber como escolhê-los. Existem várias técnicas comuns de visualização de dados que podem ser usadas para apresentar dados de forma clara e eficaz.

As técnicas mais comuns de visualização de dados incluem gráficos de barras, gráficos de linhas, gráficos circulares, gráficos de dispersão, gráficos de bolhas, mapas de calor, histogramas e diagramas de caixa. Cada técnica tem sua própria vantagem, dependendo do tipo de dados que está sendo apresentado.

- ### Gráficos em linhas

Estes tipos de gráficos usam linhas para representar valores numéricos, usualmente usa-se um sistema de coordenadas cartesianas, estes são ótimos para visualizar tendências e relações, entre variáveis continuas, mas também pode ser usado para variáveis discretas. Por estas características este torna-se ideal para visualizar comportamentos de variáveis ao longo do tempo [11].

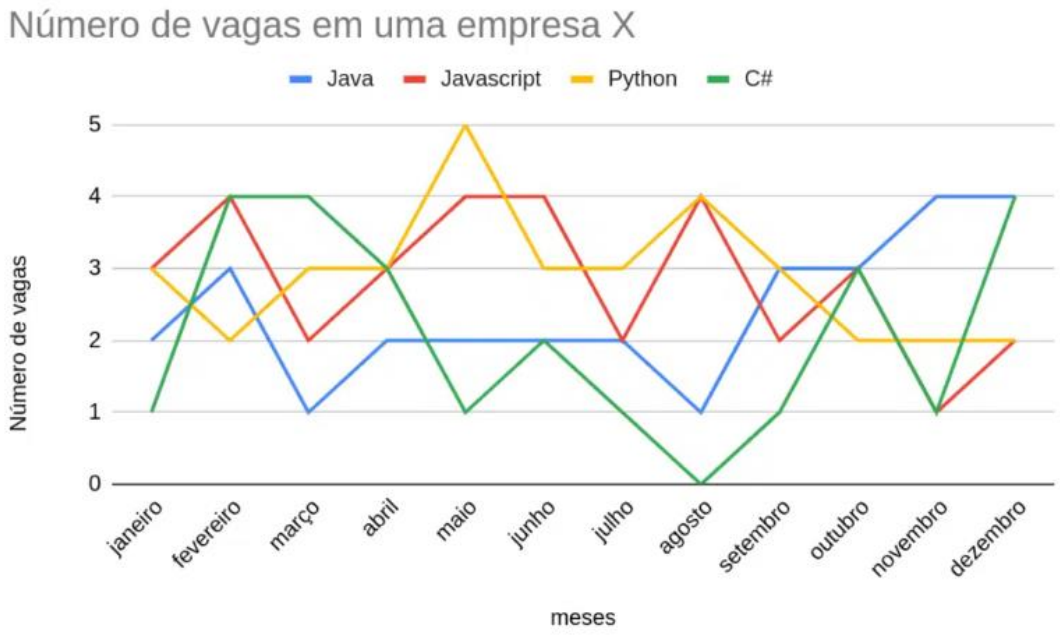


Figura 1 - Exemplo de gráfico de linhas, que exibe as vagas de linguagens de programação de uma hipotética empresa X ao longo dos meses do ano. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico)

- Gráficos de barras

Estes tipos de gráficos usam barras para representar valores numéricos, estes são ótimos para comparar valores numéricos ou discretos, da mesma categoria. Existem autores que diferenciam gráfico de barra de gráfico de colunas, o gráfico de barra tem os valores no eixo horizontal, e as informações comparativas, no eixo vertical. Já o gráfico de coluna apresenta as informações comparativas no seu eixo horizontal, e os valores, no eixo vertical [11].

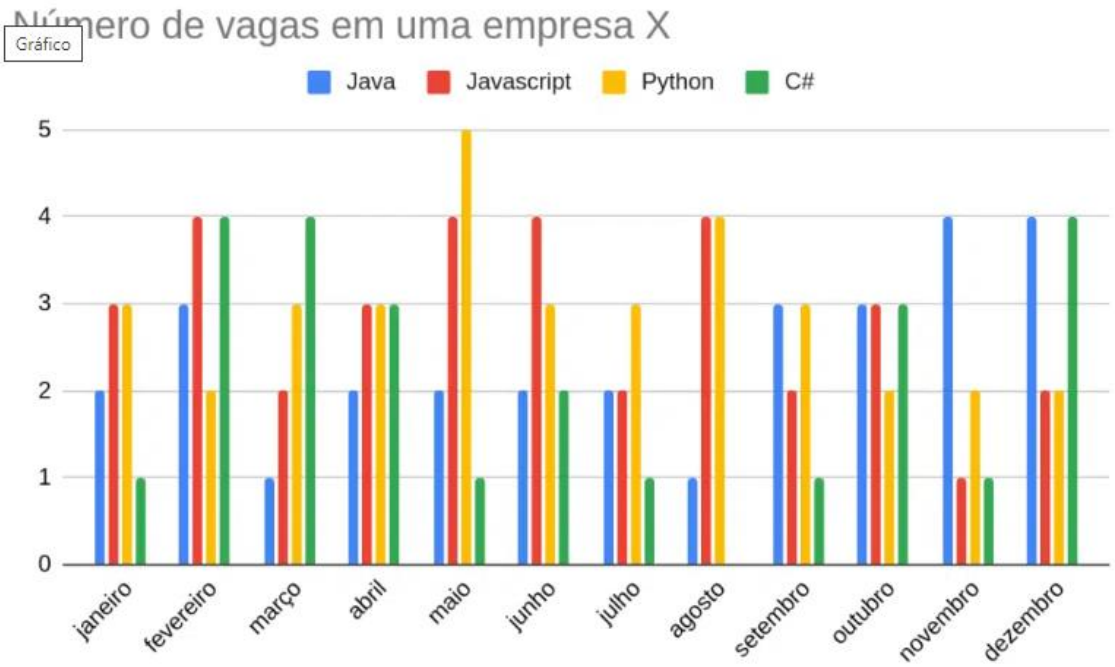


Figura 2 - Exemplo de gráfico de barras, que exibe as vagas de linguagens de programação de uma hipotética empresa X ao longo dos meses do ano. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico)

- Gráfico de dispersão

Estes tipos de gráficos usam o sistema de coordenadas cartesianas, usando um conjunto de pontos, sendo estes mapeados no gráfico através dos valores de duas variáveis. É atribuído um variável a cada eixo, permitindo assim verificar se existe algum tipo de relação entre as mesmas. Podem surgir padrões de relação como: valores positivos (aumentam em conjunto), negativos (valores aumentam enquanto outros diminuem), nulos (não há correlação), lineares e exponenciais. São ideais para verificar a dependência ou independência entre duas variáveis[11].

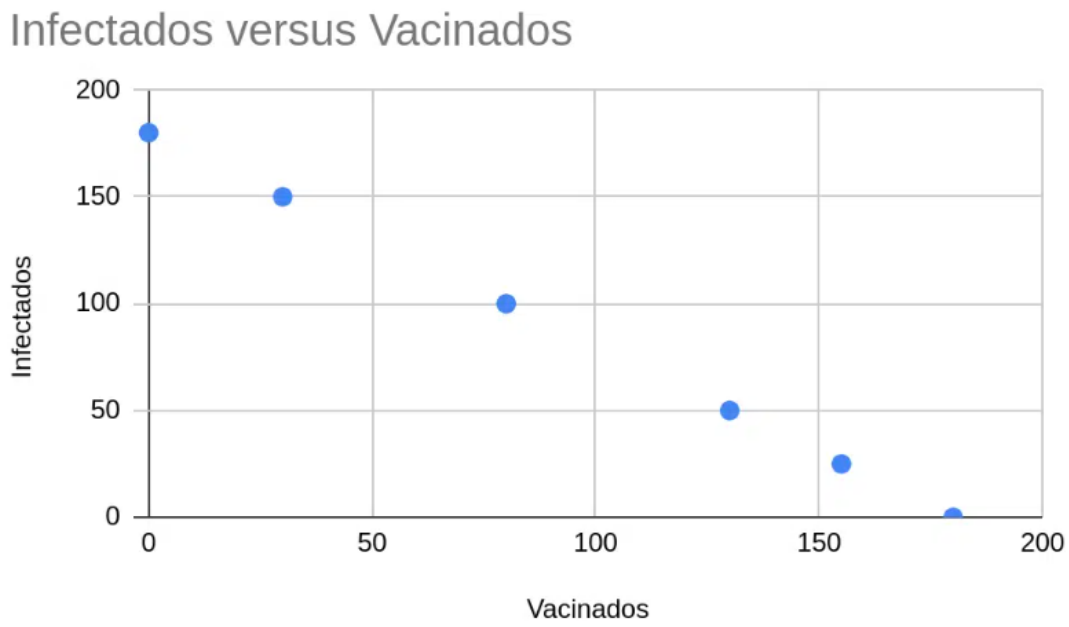


Figura 3 - Exemplo de gráfico de dispersão, que exibe a relação sobre infetados e vacinados de uma hipotética doença. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico)

- Histograma

Estes tipos de gráficos, usam barras na vertical sem espaços entre elas, com o objetivo de mostrar a distribuição de frequência de uma variável continua ou discreta. A forma do histograma revela a distribuição dos dados, como normal, assimétrica ou binomial. A partir destes podemos retirar informações como centralidade, dispersão e forma da distribuição dos valores, como referido anteriormente[11].

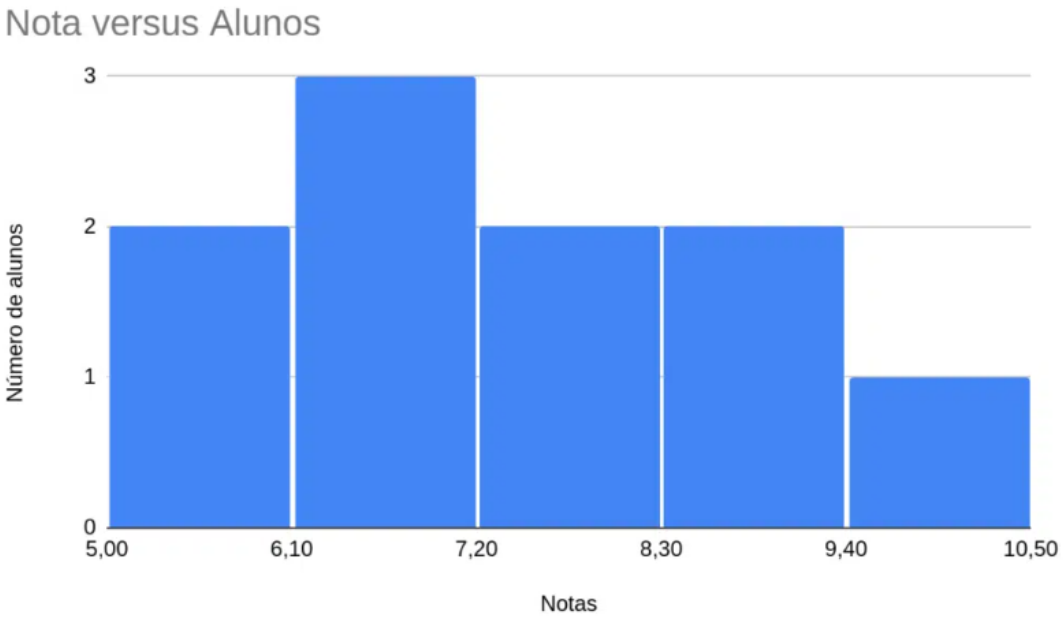


Figura 4 - Exemplo de Histograma, que exibe a frequência das notas de uma turma, em que os dados ordenados indicam a quantidade de alunos que tiraram notas em um determinado intervalo. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico)

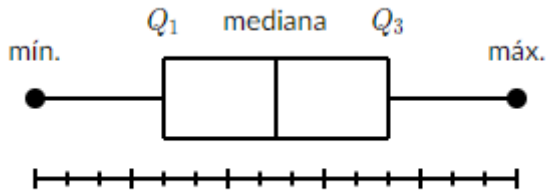
[TF11]

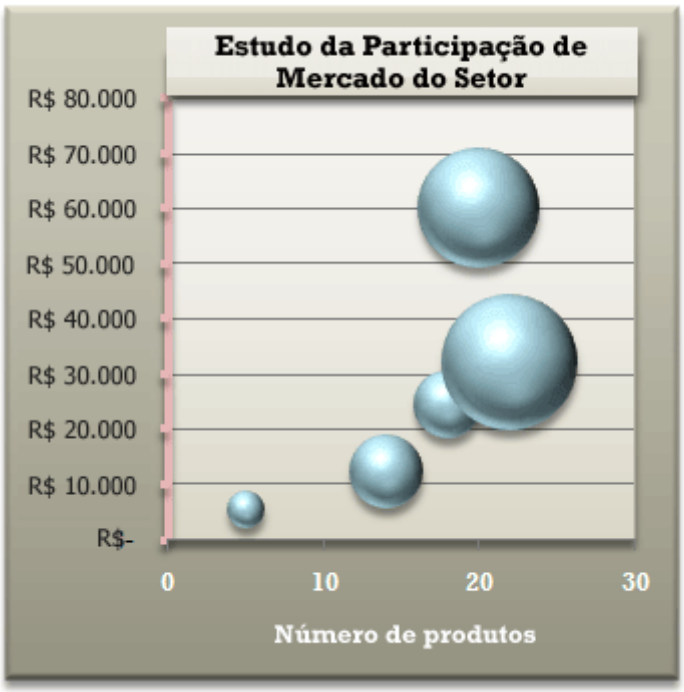
[TF12]

[TF13]

[TF14]

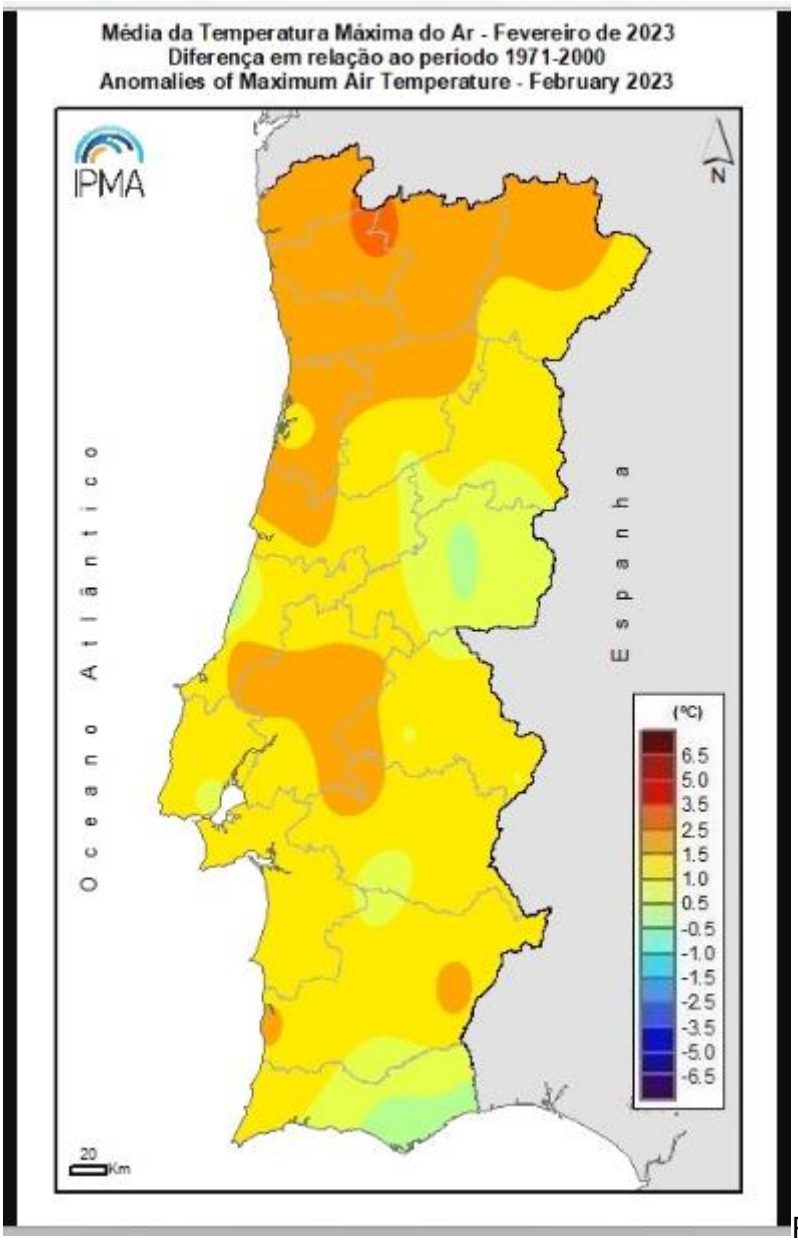
Os gráficos de barras são úteis para comparar valores entre diferentes categorias, enquanto os gráficos de linhas são ideais para mostrar tendências ao longo do tempo. Os gráficos de circulares são excelentes para mostrar a distribuição percentual de uma variável categórica. Já os gráficos de dispersão e bolhas são ideais para mostrar a relação entre duas variáveis. Mapas de calor são úteis para mostrar a intensidade de uma variável geográfica. Histogramas são excelentes para mostrar a distribuição de uma variável contínua. Por fim, diagramas de caixa são ideais para mostrar a variação em um conjunto de dados.





Para escolher a melhor

técnica de visualização de acordo com o tipo de dado, é importante levar em consideração a natureza dos dados. Por exemplo, se desejar mostrar a variação em um conjunto de dados, um diagrama de caixa pode ser a melhor opção. Por outro lado, se desejar mostrar a distribuição percentual de uma variável categórica, um gráfico circular pode ser mais adequado.



Para criar visualizações eficazes e atraentes, é importante seguir algumas diretrizes. Em primeiro lugar, é importante manter as visualizações simples e limpas, evitando o uso de gráficos desnecessários ou excessivamente complexos. Em segundo lugar, é importante escolher uma gama de cores adequada, que facilite a compreensão das informações apresentadas. Em terceiro lugar, é importante adicionar legendas claras e rótulos precisos para garantir que a visualização seja compreensível e informativa.

6. Casos de uso de exploração e visualização de dados

7. Desafios e tendências

8. Conclusão

9. Bibliografia

[1] C. S. Rosa, “Estudo sobre as técnicas e métodos de análise de dados no contexto de Big Data,” Patos de Minas, 2018.

[2] D. Reinsel, J. Gantz, and J. Rydning, “The Digitization of the World From Edge to Core,” 2018.

[3] F. S. Tsai and K. L. Chan, “Dimensionality Reduction Techniques for Data Exploration.”

[4] D. Stodder, “Improving Data Preparation for Business Analytics Applying Technologies and Methods for Establishing Trusted Data Assets for More Productive Users BEST PRACTICES REPORT Q3 2016,” 2016.

[5] G. Mansingh, K. M. Osei-Bryson, L. Rao, and M. McNaughton, “Data preparation: Art or science?,” in *Proceedings of the 2016 International Conference on Data Science and Engineering, ICDSE 2016*, Institute of Electrical and Electronics Engineers Inc., Jan. 2017. doi: 10.1109/ICDSE.2016.7823936.

[6] J. Brownlee, “Data Preparation for Machine Learning,” 2020.

[7] B. R. Santos, P. T. Fonseca, M. Barata, R. A. Ribeiro, and P. A. C. Sousa, “New data preparation process - A case study for an exomars drill,” in *2006 World Automation Congress, WAC’06*, IEEE Computer Society, 2006. doi: 10.1109/WAC.2006.376041.

[8] M. João and G. Cardoso, “Ferramentas de Extração e Exploração de DadosparaBusiness Intelligence,” 2018.

[9] L. Soibelman, M. Asce, and H. Kim, “Data Preparation Process for Construction Knowledge Generation through Knowledge Discovery in Databases,” 2002, doi: 10.1061/ASCE0887-3801200216:139.

[10] S. Idreos, O. Papaemmanouil, and S. Chaudhuri, “Overview of data exploration techniques,” in *Proceedings of the ACM SIGMOD International Conference on*

Management of Data, Association for Computing Machinery, May 2015, pp. 277–281. doi: 10.1145/2723372.2731084.

[11] F. Pereira, “Big Data e Data Analysis - Visualização de Informação,” 2015.