Universidade de Aveiro - Departamento de Matemática

Exploração e Visualização de Dados

Tiago Ferreira 78106

João Diogo 89340

João Acácio 98040

Índice

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 armazenamentode 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. 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.

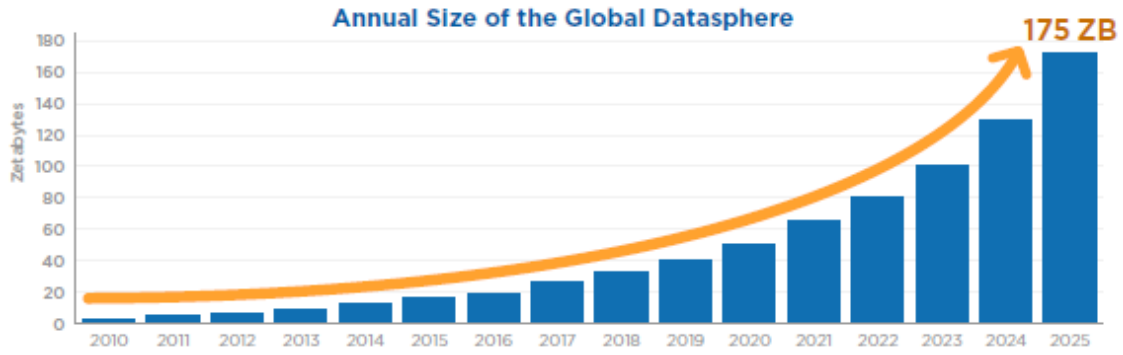

Figura 1 - Tamanho Anual do Global Datasphere, de 2010 até 2025.

Fonte: [1]

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. 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. 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.

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 aquique vamos determinar o que são os dados, melhorar a sua qualidade, padronizar, consolidar e transformá-los paraque 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].

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].

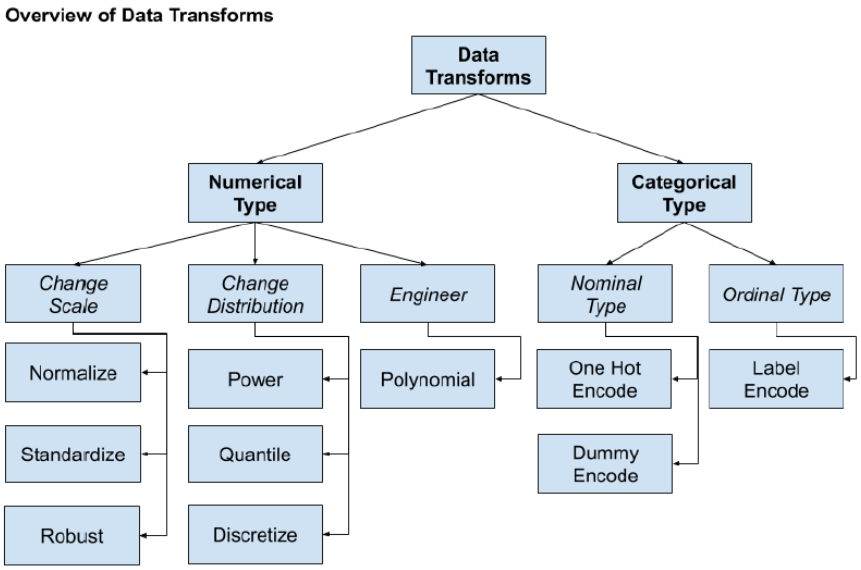

Figura 2 - - Visão geral das tecnicas de transformação de dados.

Fonte: [6]

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].

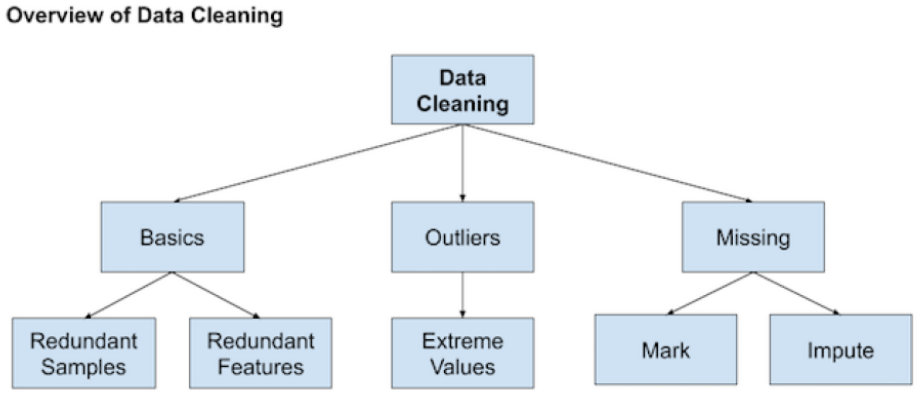

Figura 3 - Visão geral das técnicas de limpeza de dados.

Fonte: [6]

3. Análise exploratória de dados

A análise exploratória de dados (AED) é uma etapa fundamental do processo de análise de dados, que tem como objetivos principais otimizar a extração de conhecimentos do conjunto de dados, visualizar a existência de potencias relações entre variáveis, verificar anomalias e outliers, desenvolver modelos eficientes de predição ou de explicação e criar ou extrair variáveis relevantes[10].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.

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. Algumas ferramentas usadas na AED são: estatísticas descritivas, gráficos estatísticos, análise de correlação, regressão, clusterização e técnicas de projeção.

- Estatística descritiva

A estatística descritiva tem como objetivosumarizar e descrever as características básicas de um conjunto de dados,calculando medidas estatísticas básicas, como média, mediana, desvio padrão, mínimo e máximo. Mensurando assim a tendência central, a dispersão e a distribuição dos valores, o que dá uma visão geral dos dados. Com estas análises é possível entender a estrutura dos dados, identificar outliers e tomar decisões informadas[11], [12].

- Análise de correlação

A análise de correlação é uma técnica estatística usada para medir o grau de relação entre duas variáveis, revelando a existência de uma relação linear entre as variáveis, caso esta exista, mostrando a direção e a força dessa relação. A medida mais usual para calcular a correlação é o coeficiente de correlação de Pearson, que varia de -1 a 1. Um valor próximo de 1 indica uma forte correlação positiva, um valor próximo de -1 indica uma forte correlação negativa, e um valor próximo de zero indica uma correlação fraca ou inexistente. Logo é possível identificar padrões e relações entre variáveis, auxiliando na compreensão dos dados e na tomada de decisões embasadas em dados empíricos[13].

- Regressão

A regressão é uma técnica estatística que modela e investiga a relação entre uma variável dependente e uma ou várias variáveis independentes. Assim podemos estimar a natureza e a intensidade dessa relação, bem como fazer previsões com base nos dados disponíveis. Existem diferentes tipos de regressão, como regressão linear, regressão logística e regressão polinomial, cada uma adequada para diferentes tipos de dados e objetivos de análise. A regressão envolve a estimativa de parâmetros, como coeficientes de regressão, que quantificam a influência das variáveis independentes na variável dependente. Ela desempenha um papel fundamental na previsão e na compreensão das relações entre as variáveis em estudo.

- Clusterização

A clusterização é uma técnica que agrupa dados similares, tem como objetivo encontrar estruturas ou padrões nos dados sem a necessidade de rótulos pré-existentes. A clusterização é útil para explorar e entender a organização dos dados, identificando grupos naturais. Algoritmos de clusterização, como o k-means, hierárquico ou DBSCAN, são aplicados para atribuir instâncias a clusters com base nas suas características[14].

- Técnicas de projeção

As técnicas de projeção têm como objetivo a redução de dimensionalidade, ou seja,diminuído o nosso conjunto de dados sem perder informação relevante. Estas técnicas revelam-se úteisem dados com muitas variáveis, o que pode causar problemas de dimensionalidade e dificultar a análise. Existem técnicas lineares como a análise de componentes principais (PCA), que identifica combinações lineares de variáveis que capturam a maior parte da variabilidade dos dados. Também existem técnicas não lineares como o mapeamento de características isométricas (Isomap), que consiste no uso de geodésicase grafos. Estas técnicas permitem visualizar os dados num espaço de menor dimensão, facilitando a interpretação, a visualização e a análise posteriores[3], [11].

4. 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 ilustrações gráficas. O problema reside como escolhê-los. Existem várias técnicas comuns de visualização de forma clara e compreensiva como:

- 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 paravisualizar comportamentos de variáveis ao longo do tempo[15].

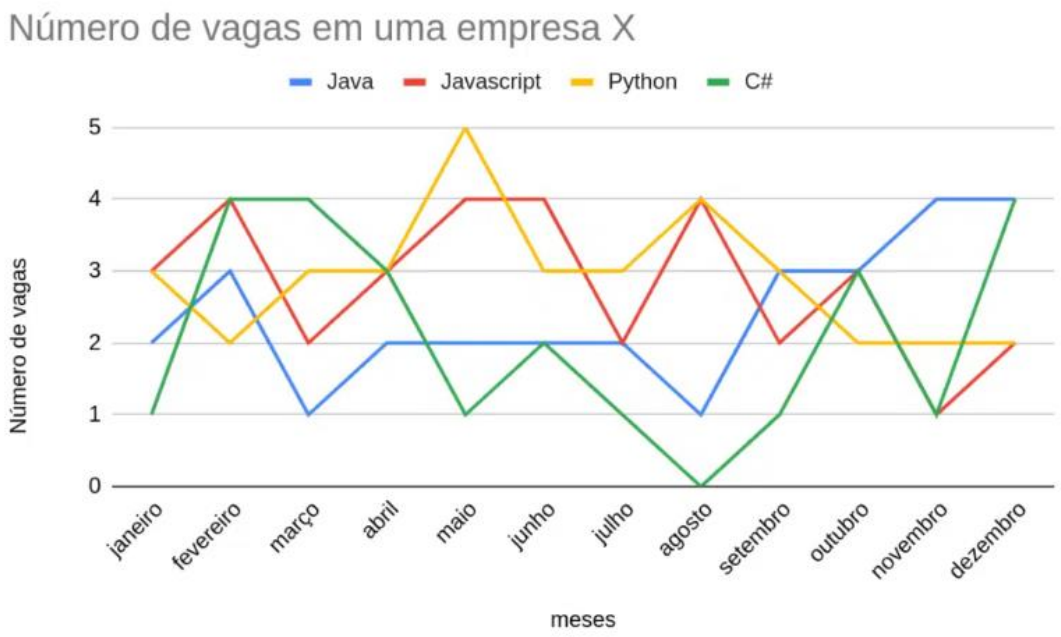

Figura 4 - Exemplo de gráfico de linhas, que exibe as vagas de linguagens de programação de uma hipotética empresa X ao longo dos meses do ano. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico).

- Gráficos de barras

Estes tipos de gráficos usam barras para representar valores numéricos, estessão 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[15].

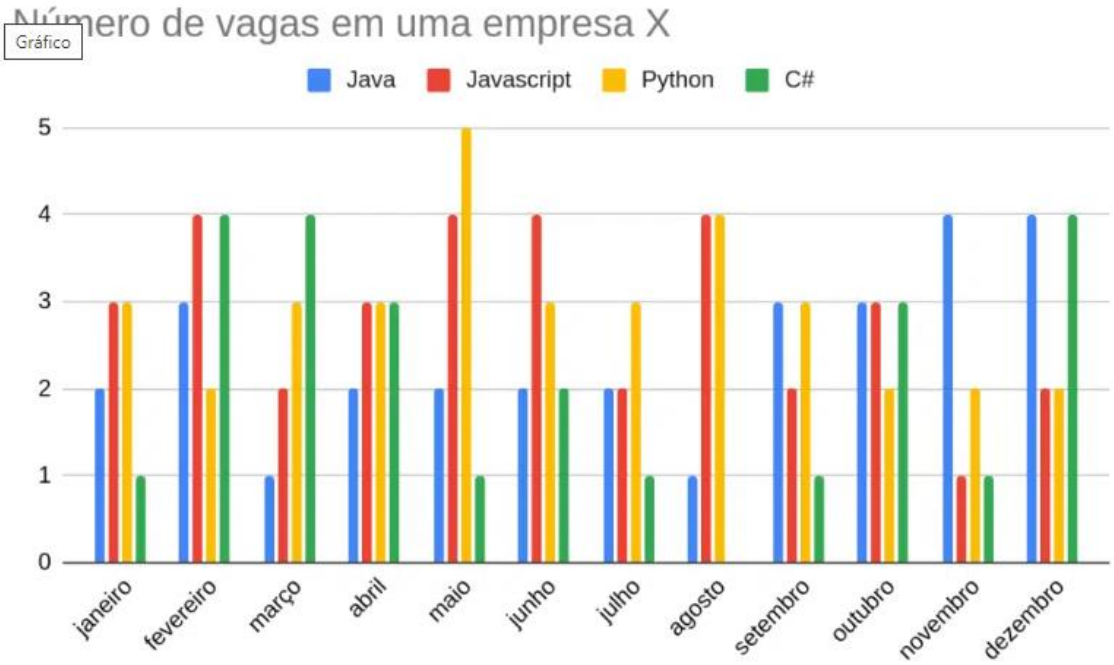

Figura 5 - Exemplo de gráfico de barras, que exibe as vagas de linguagens de programação de uma hipotética empresa X ao longo dos meses do ano. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico).

- 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[15].

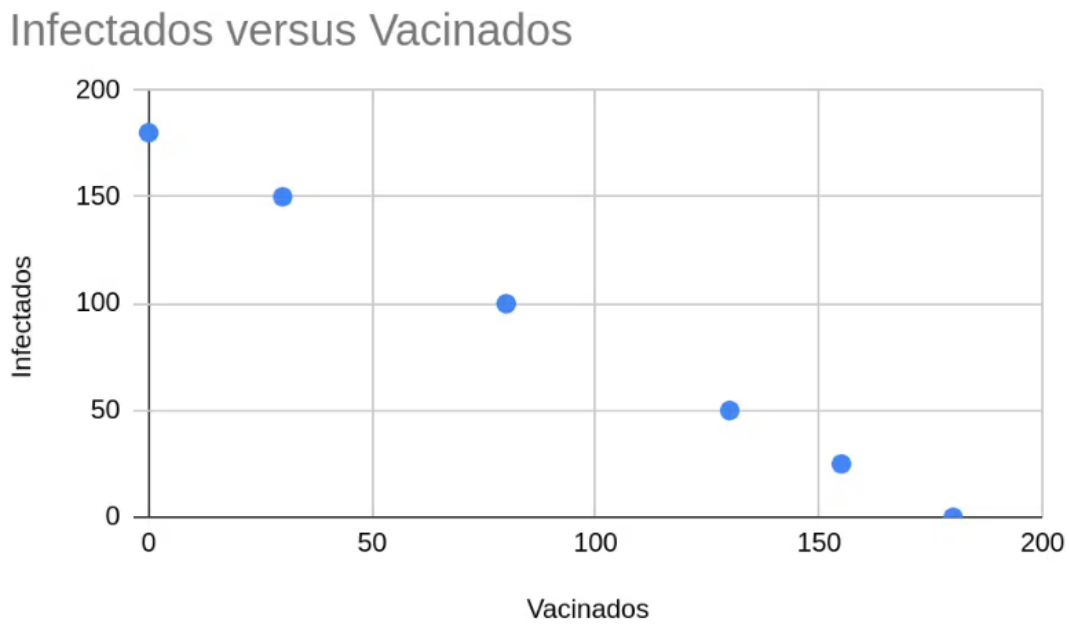

Figura 6 - Exemplo de gráfico de dispersão, que exibe a relação sobre infetados e vacinados de uma hipotética doença. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico).

- 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[15].

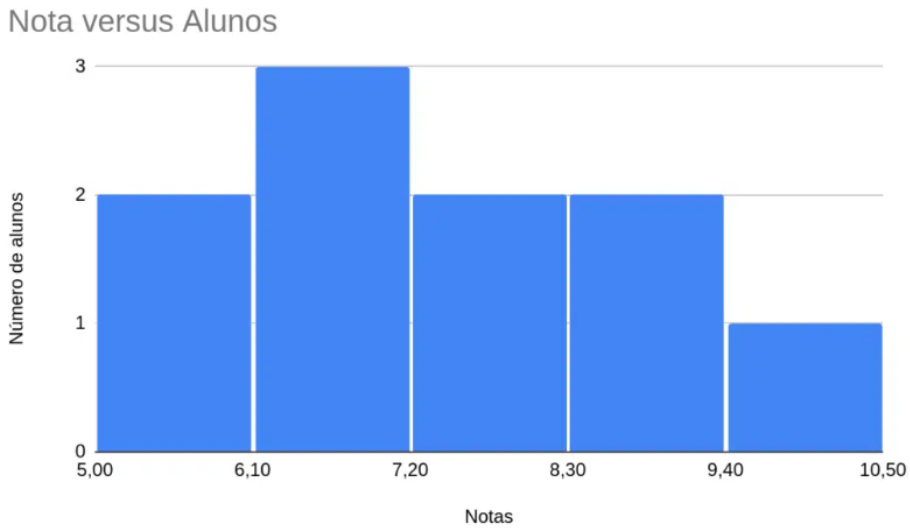

Figura 7 - Exemplo de Histograma, que exibe a frequência das notas de uma turma, em que os dados ordenados indicam a quantidade de alunos que tiraram notas em um determinado intervalo. (Retirado de https://blog.betrybe.com/estatistica/principais-tipos-de-grafico)

- Caixa de bigodes

Este tipo de gráfico, também conhecido como boxplot, é composto por uma caixa retangular que representa o intervalo interquartil, com uma linha vertical no meio que indica a mediana. Os "bigodes" são linhas que se estendem a partir da caixa e representam os valores mínimo e máximo ou, em alguns casos, os limites estatísticos. O seu uso é importante para a estatística descritiva por este fornecer uma visão rápida e eficiente das características do conjunto de dados, como a dispersão, assimetria e presença de valores atípicos[15].

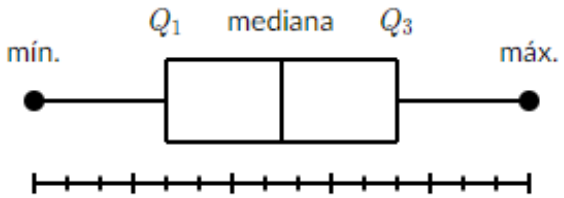

Figura 8 - Exemplo de uma caixa de bigodes (retirado de https://pt.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/box-plot-review)

- Gráfico de Bolhas

Este tipo de gráfico representações visuais que combinam coordenadas cartesianas para exibir três dimensões de dados. Cada bolha no gráfico é uma marcação circular com tamanho proporcional a um valor numérico específico. As coordenadas x e y determinam a posição horizontal e vertical da bolha no gráfico, enquanto o tamanho da bolha representa a terceira dimensão do dado. Esses gráficos são úteis para visualizar relações entre variáveis e identificar padrões, correlações ou clusters nos dados. Com tudo quando existe demasiada informação pode tornar-se difícil a sua leitura, esta dificuldade pode ser ultrapassada com o uso de ferramentas de interatividade[15].

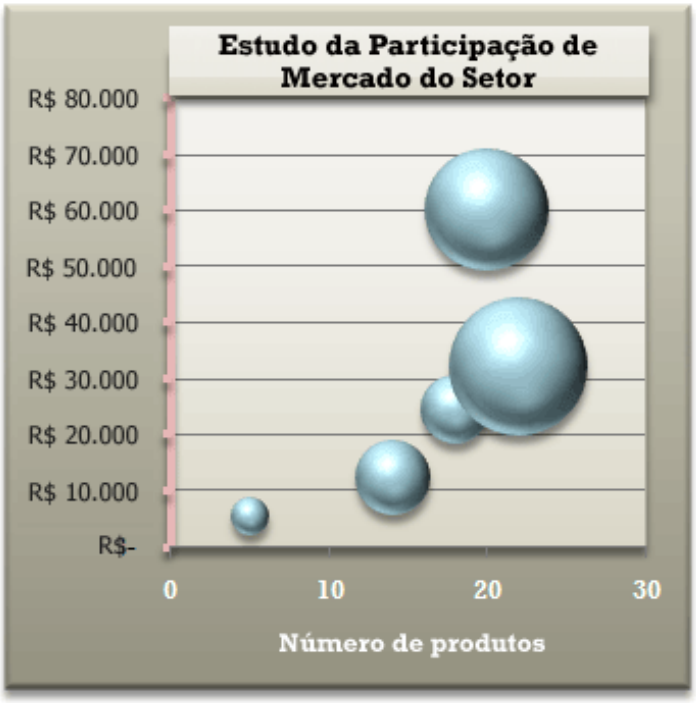

Figura 9 - Exemplo de um Gráfico de bolhas (retirado de https://support.microsoft.com/pt-br/office/apresentar-seus-dados-em-um-gr%C3%A1fico-de-bolhas-424d7bda-93e8-4983-9b51-c766f3e330d9).

Existe uma variedade de modelos de gráficos, o desafio será em escolher. Para isso temos de ter em conta o tipo de dados com que nos deparamos, e objetivo que temos. Por exemplo se pretendemos mostrar uma variação em um conjunto de dados, o diagrama de caixa pode ser o melhor, contudo se o objetivo passa a ser a distribuição percentual de várias variáveis, o gráfico circular revelasse ser mais adequando.

5. 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 usada para a análise e visualização de dados, conhecida pela sua facilidade de uso e capacidade de criar visualizações interativas;
- Excel - Software de folhas de cálculomuito utilizado que oferece recursos para análise e visualização de dados, como 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 source 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 source 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 pelas suas visualizações interativas;
- SAP BusinessObjects - Ferramenta paga que oferece ferramentas para gerenciamento de dados, análise e geração de relatórios;
- SAS - Ferramenta paga que oferece uma ampla gama de recursos para análise de dados, incluindo manipulação de dados, estatísticas e geração de relatórios;
- Stata - Software estatístico com recursos abrangentes para análise e manipulação de dados, muito utilizado em pesquisa e análise académica.

Figura 8 - Logótipo Tableau

Figura 9 - Logótipo Excel

Figura 10 - Logótipo PowerBi

Figura 11 - Logótipo Python

Figura 12 - Logótipo R

Figura 13 - Logótipo QlikView

Figura 14 - Logótipo SAP BO

Figura 15 - Logótipo SAS

Figura 16 - Logótipo Stata

6. Casos de uso de exploração e visualização de dados

No processo de análise de dados é essencial explorar e visualizar dados, algo que ajuda nas tomadas de decisão, identificação de padrões e demonstração de resultados. Logo podemos perceber a relevância e o contributo que esta área tem para outras tais como: negócios e análise de mercado,finanças,saúde, pesquisa científica,ciênciassociais, ciências de dados e marketing.

Estes acabam por ser apenas algumas das áreas em que a exploração e visualização de dados é utilizada. No fundo, as técnicas mencionadas anteriormente podem e devem ser utilizadas em praticamente qualquer campo em que haja uma recolha de dados.Tendo isto em conta podemos identificar casos mais práticos de uso de exploração e visualização de dados verificados em diversos artigos como por exemplo:

- Análise de dados topológicos, aqui, a ideia é utilizar a noção matemática da topologia para visualizar e explorar conjuntos de dados complexos e de elevada dimensão. Esta análise já foi utilizada em diversos campos ligados à saúde, mais precisamente na biologia ou química. A ideia geral é considerar um conjunto como uma amostra ou nuvem retirada de um espaço de elevada dimensão e posteriormente utilizar os dados recolhidos para construir simplicidades para serem por sua vez utilizados mais eficientemente[16].
- Deteção de fraude financeira aplicando técnicas de “data mining”. O “data mining” que no fundo é uma abordagem utilizada para extrair uma quantidade de dados significativos de um determinado conjunto maior de dados, utilizando técnicas estatísticas, de aprendizagem automática ou de inteligência artificial. Neste caso são aplicados diversos métodos de exploração e visualização de dados tais como a máquina de vetores de suporte, o modelo oculto de Markov, a regressão logística, entre outros[17].
- Análise de dados recolhidos durante experiências de difração de raios X (técnica para analisar a estrutura atómica dos materiais). Neste caso, é essencial a recolha de imagens com ajuda de detetores mais rápidos e eficientes. Durante estas experiências é emitida uma quantidade enorme de dados e para o estudo e utilização dos mesmos torna-se essencial a utilização de um software que ajude à redução e exploração de dados[18].
- Extensão da análise exploratória de dados à cartografia de forma que seja possível facilitar e automatizar a visualização cartográfica com uso de dados recolhidos previamente. Para isto são usados “softwares” que automatizam

os dados recolhidos com o objetivo de criar mapas e manipulá-los interactivamente. Esta automatização na construção de mapas torna a tarefa mais fácil e rápida[19].

- Auxiliar a visualização de dados financeiros de forma que sejam oferecidas oportunidades aos analistas de mercado e investidores para extrair novos conhecimentos de forma pragmática, e, com essas ferramentas, poderem tomar decisões informadas. Tradicionalmente, os analistas utilizam ferramentas de análise visual baseados em métodos estatísticos de forma a visualizar gráficos de linhas ou gráficos de velas (os mais populares no ramo financeiro)[20].

É de ter em conta que os casos de uso que são aqui referidos representam apenas uma pequena parte do potencial da exploração e visualização de dados. Estas noções e técnicas podem ser aplicadas a muitos outros ramos.

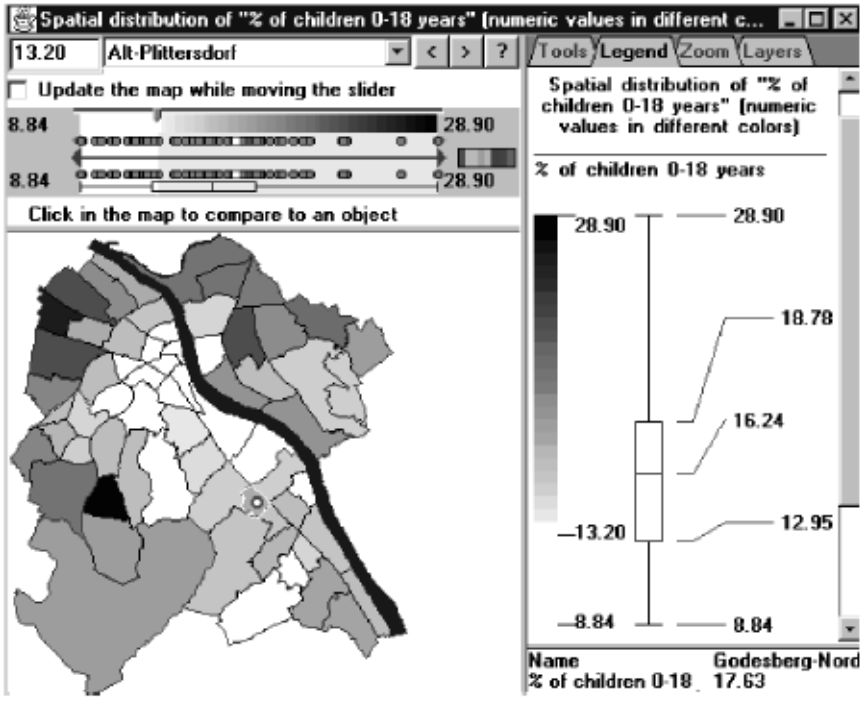

Figure 1. An example map window with dynamic manipulation tools.

Figura 17 - Exemplo de um caso de uso onde a visualização interativa de dados é utilizada após a sua exploração.

Fonte: [19]

7. Desafios e tendências

A exploração e visualização de dados é cada vez mais importante no mundo moderno, sendo impulsionada pelo avanço da tecnologia e crescimento da quantidade de dados disponíveis. Todavia, apesar do potencial fornecido pelas diversas técnicas de exploração e visualização de dados, surgem alguns desafios, sejam estes lidar com maiores volumes de dados ou garantir a sua veracidade. Por outro lado, as diferentes tendências também se vão alterando com o surgir de novas técnicas como a incorporação de inteligência artificial.Assim, torna-se fundamental entender os desafios da exploração e visualização de dados bem como seguir a evolução das suas tendências.

7.1. Desafios

Um dos maiores desafios na exploração e visualização de dados são os métodos tradicionais serem inadequados para lidar com grandes volumes de dados. A nova visualização de dados tem que encontrar formas de processar, analisar e visualizar quantidades maiores de dados. As novas ferramentas e técnicas de visualização devem ser capazes de identificar valores em falta, errados ou duplicados. Um dos fatores que afeta o que foi referido anteriormente é a perceção humana, o olho humano tem dificuldade em extrair informações significativas quando os dados se tornam extremamente grandes. Outro fator é que a maioria dos sistemas de visualização está concebido para tratar dados com apenas uma determinada informação, torna-se assim um desafio ultrapassar limitações de memória e processamento em tempo real. Em seguida, temos a interatividade que amplia as vantagens da visualização de dados e nos ajuda a compreender melhor e com maior perceção um conjunto de dados. No entanto, é preciso tempo para processar e analisar dados antes da respetiva visualização, como exemplo, a consulta de grandes quantidades de dados ou algoritmos complexos pode perturbar a interação fluente [21].

Outros desafios que se encontram estão relacionados com a própria recolha de dados e o seu armazenamento, que podem gerar problemas de privacidade e segurança dos mesmos. Além disso, a qualidade dos dados é fundamental para apurar resultados corretos, maior quantidade de dados pode não ser traduzida em maior qualidade, é necessário encontrar um limite no chamado “volume de dados” [15].

7.2. Tendências

Uma das tendências que podemos identificar na visualização de dados é a visualização interativa que se define como um conjunto de ferramentas e processos

para produzir uma visualização interativa de dados que pode ser explorada e analisada dentro da visualização em si. Este conceito é focado em representações de dados que melhorem a maneira como interagimos com a informação, esta questão torna-se vital visto que a facilidade com que alguém consegue interpretar uma representação de dados é importante para a tiragem de conclusões[22].

É importante referir também o papel da inteligência artificial e do *"machinelearning"* na exploração e visualização de dados. Com o rápido crescimento em tamanho e número de dados, é essencial desenvolver métodos novos para trabalhar com as quantidades enormes de dados armazenados. O progresso na tecnologia e automação nas atividades produz um fluxo cada vez maior de dados e é aqui que entra a inteligência artificial, visando a facilitar uma tarefa cada vez mais complicada. Aplicar técnicas de *"machinelearning"* para descoberta de conhecimento em bases de dados enormes tornou-se assim essencial[23].

8. Conclusão

No momento atual a exploração e a visualização de dados têm um papel importante na compreensão de um conjunto de dados. No nosso relatório foram exploradas as diferentes noções deste tema.

Começando pela preparação dos dados, este é o primeiro passo para criar uma análise. Usando diferentes ferramentas, é possível trabalhar com um conjunto de dados ao proveito do utilizador. É importante utilizar a análise exploratória dos dados com o objetivo de entender os dados disponíveis, queremos aqui identificar padrões para que surja uma compreensão melhor do caso de estudo.

Por outro lado, é necessário trabalhar os dados obtidos de forma a que sejam visualizados de forma eficiente, aqui entram as técnicas de visualização de dados. Pela forma de métodos como gráficos é possível interpretar as informações recolhidas. Foram também pesquisados casos de uso em diversas áreas, algo que ajuda a entender como é realmente utilizado na prática o tema do nosso relatório.

O principal obstáculo tem sido como lidar com o aumento da quantidade de dados a analisar. É também importante seguir as tendências que vão surgindo, desde o uso da inteligência artificial ou a visualização interativa de dados, estas têm o objetivo de melhorar o modo como trabalhamos com dados obtidos.

Pelo exposto, entendemos que a exploração e visualização de dados são técnicas essenciais para trabalhar os dados de forma correta. Utilizando os métodos adequados é possível desenvolver com mais facilidade uma imensidão de áreas. Na revolução tecnológica em que vivemos, a relevância da exploração e visualização de dados continua a aumentar, sendo esta uma área fundamental.

Por fim, juntamente com este relatório foi feito um pequeno trabalho prático acerca do tema que foi usado para a apresentação e este é encontrado no repositório https://github.com/ferreira-trc/Relatorio_seminario.

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] R. Pirracchio, *Secondary Analysis of Electronic Health Records*. 2016. doi: 10.1007/978-3-319-43742-2_20.

[11] M. Li Vigni, C. Durante, and M. Cocchi, *Exploratory Data Analysis*, 1st ed., vol. 28. Copyright © 2013 Elsevier B.V. All rights reserved., 2013. doi: 10.1016/B978-0-444-59528-7.00003-X.

[12] A. Unwin, “Exploratory Data Analysis,” *Int. Encycl. Educ. Third Ed.*, vol. 23, pp. 156–161, 2009, doi: 10.1016/B978-0-08-044894-7.01327-0.

[13] D. B. F. Filho and J. A. D. S. Júnior, “Desvendando os mistérios do coeficiente de correlação de Pearson (r),” *Rev. Política Hoje*, vol. 18, no. 1, pp. 115–146, 2009.

[14] T. Hastie, R. Tibshirani, and F. Jerome, *The Elements of Statiscal Learning*, vol. 27, no. 2. 2001. [Online]. Available: http://www.springerlink.com/index/D7X7KX6772HQ2135.pdf

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

[16] M. Offroy and L. Duponchel, “Topological data analysis: A promising big data exploration tool in biology, analytical chemistry and physical chemistry,” *Anal. Chim. Acta*, vol. 910, pp. 1–11, 2016, doi: 10.1016/j.aca.2015.12.037.

[17] K. G. Al-Hashedi and P. Magalingam, “Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019,” *Computer Science Review*,

vol. 40. Elsevier Ireland Ltd, May 01, 2021. doi: 10.1016/j.cosrev.2021.100402.

[18] C. Prescher and V. B. Prakapenka, “DIOPTAS: A program for reduction of two-dimensional X-ray diffraction data and data exploration,” *High Press. Res.*, vol. 35, no. 3, pp. 223–230, Jul. 2015, doi: 10.1080/08957959.2015.1059835.

[19] G. L. Andrienko and N. V. Andrienko, “Interactive maps for visual data exploration,” *Int. J. Geogr. Inf. Sci.*, vol. 13, no. 4, pp. 355–374, 1999, doi: 10.1080/136588199241247.

[20] S. Ko *et al.*, “A Survey on Visual Analysis Approaches for Financial Data,” *Comput. Graph. Forum*, vol. 35, no. 3, pp. 599–617, Jun. 2016, doi: 10.1111/cgf.12931.

[21] R. Agrawal, A. Kadadi, X. Dai, and F. Andres, “Challenges and opportunities with big data visualization,” *7th Int. ACM Conf. Manag. Comput. Collect. Intell. Digit. Ecosyst. MEDES 2015*, no. October, pp. 169–173, 2015, doi: 10.1145/2857218.2857256.

[22] P. Godfrey, J. Gryz, and P. Lasek, “Interactive visualization of large data sets,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 8, pp. 2142–2157, 2016, doi: 10.1109/TKDE.2016.2557324.

[23] D. A. Keim, C. Panse, J. Schneidewind, M. Sips, M. C. Hao, and U. Dayal, “Pushing the limit in visual data exploration: Techniques and applications,” *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.*, vol. 2821, pp. 37–51, 2003, doi: 10.1007/978-3-540-39451-8_4.