LADATA
liga acadêmica de ciência de dados

Autores do material: Pedro Rocha; Eduarda Mascarenhas; Kely Murta.

Área: Ciência de Dados.

Palavras-chave: Análise de Dados; Estatística Descritiva; Visualização de Dados; Python.

Análise de Dados: Conceitos, Técnicas e Aplicações

11 de Novembro de 2024, São Cristóvão - SE

Sumário

1. Introdução

A **análise de dados** é o processo sistemático de examinar, organizar e interpretar informações para extrair *insights*, identificar padrões e auxiliar na tomada de decisões. Esse campo, essencial em um mundo cada vez mais movido por dados, não se restringe a uma definição única, pois suas aplicações e metodologias variam conforme o contexto. Desde a estatística e probabilidade até ferramentas computacionais modernas, a análise de dados combina diferentes áreas do conhecimento, como matemática, ciência da computação e disciplinas específicas dos domínios em que é aplicada, como economia, saúde e ciências sociais.

No amplo espectro da tríade dos dados, que engloba engenharia, ciência e análise, cada área desempenha um papel distinto, mas interligado. Enquanto a **engenharia de dados** lida com a coleta, armazenamento e organização eficiente das informações, e a **ciência de dados** foca em métodos avançados como aprendizado de máquina e modelagem preditiva, a **análise de dados** assume a função de transformar dados brutos em conhecimento prático, muitas vezes atuando como ponte entre as outras duas.

No entanto, no dia a dia de quem trabalha com dados, essas fronteiras frequentemente se misturam. Não é incomum que um profissional de análise se envolva em atividades de ciência, como treinar modelos preditivos simples, ou de engenharia, ao preparar dados para análise. Essa flexibilidade faz com que a área seja caracterizada por um aspecto generalista, onde o foco principal é o impacto gerado pelo uso inteligente dos dados.

2. Conceitos

2.1. Pensamento Analítico

Antes de mergulhar em ferramentas e técnicas, é essencial desenvolver uma forma de pensar que esteja alinhada com os princípios da análise de dados. O pensamento analítico envolve olhar para os dados com uma mentalidade investigativa, buscando não apenas o “como” das coisas, mas também o “porquê”. Trata-se de estruturar o raciocínio para identificar problemas, formular hipóteses e interpretar resultados de maneira lógica e fundamentada.

Na prática, isso significa questionar os dados:

- **Qual é a origem deles?**
- **O que representam?**
- **Como podem ser utilizados para responder perguntas específicas?**

Esse tipo de pensamento também requer a habilidade de conectar padrões observados com ações práticas, seja para tomar decisões estratégicas, validar teorias ou melhorar processos.

2.2. Estatística

A estatística desempenha um papel essencial na análise de dados, oferecendo métodos estruturados para explorar, descrever e inferir informações. Dividida em dois principais ramos – **estatística descritiva** e **estatística inferencial** – ela proporciona ferramentas para resumir conjuntos de dados e realizar generalizações baseadas em amostras.

- **Estatística Descritiva**

A estatística descritiva é o ramo da estatística dedicado à organização, apresentação e resumo dos dados de maneira clara e compreensível. Seu objetivo é transformar números brutos em informações que permitam uma compreensão inicial do comportamento dos dados, evidenciando padrões, tendências e características principais. Para isso, utiliza-se os seguintes conceitos:

Média: É a soma de todos os valores dividida pelo número total de valores. Representa o valor central ou típico de um conjunto de dados.

Exemplo:
Notas de uma prova: 6, 7, 8, 9, 10
Média = soma das notas/quantidade de notas = (6 + 7 + 8 + 9 + 10) / 5 = 8
A média é 8.

Moda: É o valor que mais se repete em um conjunto de dados. Pode haver mais de uma moda (bimodal ou multimodal).

Exemplo:
Idades de um grupo: 18, 19, 18, 20, 21, 18, 20
Moda = 18 (aparece 3 vezes)

Mediana: É o valor que divide os dados em duas partes iguais quando estão organizados em ordem crescente. Se houver um número par de valores, a mediana é a média dos dois valores centrais.

Exemplo:
Salários de uma empresa: R$ 1500,00; R$ 2000,00; R$ 2500,00; R$ 3000,00; R$ 3500,00
Mediana = R$ 2500,00 (o valor central na ordem crescente)

Quantis: Dividem os dados em partes iguais. O quartil, por exemplo, divide os dados em 4 partes iguais, enquanto os percentis dividem em 100 partes.

Exemplo:
Altura (em cm) de 10 crianças: 100, 102, 104, 106, 108, 110, 112, 114, 116, 118
O primeiro quartil (Q1) é o valor que divide os 25% menores: 104.
Q1 = 104.

Variância: Mede o quanto os dados estão espalhados em relação à média. É a média dos quadrados das diferenças entre cada valor e a média.

Exemplo:
Pesos (em kg): 50, 52, 53, 55, 60
Média = (50+52+53+55+60)/5 = 54
Diferenças ao quadrado: $(50-54)^2$, $(52-54)^2$, $(53-54)^2$, $(55-54)^2$, $(60-54)^2$
Variância = (16 + 4 + 1 + 1 + 36) / 5 = 11.6 kg

Desvio Padrão: É a raiz quadrada da variância. Representa, em média, o quanto os valores se desviam da média.

Exemplo:
Com base no exemplo anterior:
Desvio padrão = $\sqrt{11.6} \approx 3.4$ kg

Assimetria: Mede o grau de simetria da distribuição dos dados. Pode ser positiva (cauda à direita), negativa (cauda à esquerda) ou neutra (simétrica). A figura 1 mostra cada tipo de assimetria.

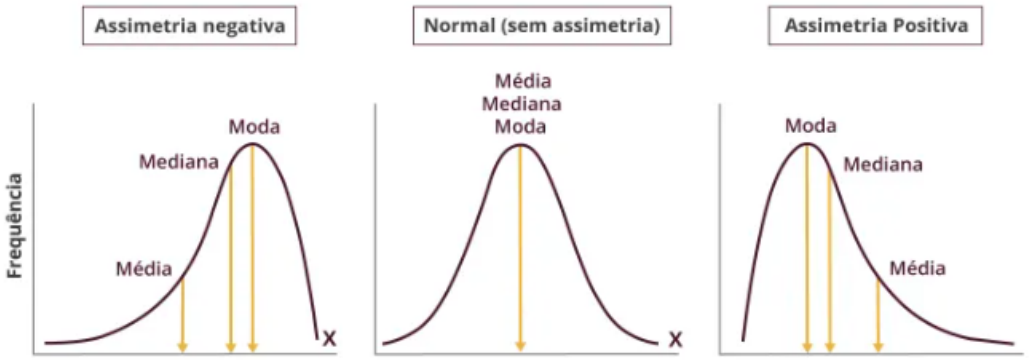

Figura 1. Demonstração dos tipos mais comuns de assimetria. Fonte: *DYNAMOX*, 2023.

Exemplo:
Altura de crianças: 100, 102, 104, 105, 120 (valor muito alto).
Assimetria positiva: a cauda está à direita devido ao valor 120.

Curtose: Mede a concentração dos dados nos extremos em relação ao centro.
- Curtose alta: muitos valores nos extremos (caudas pesadas);
- Curtose baixa: distribuição achatada.

A figura 2 mostra como são as curtoses:

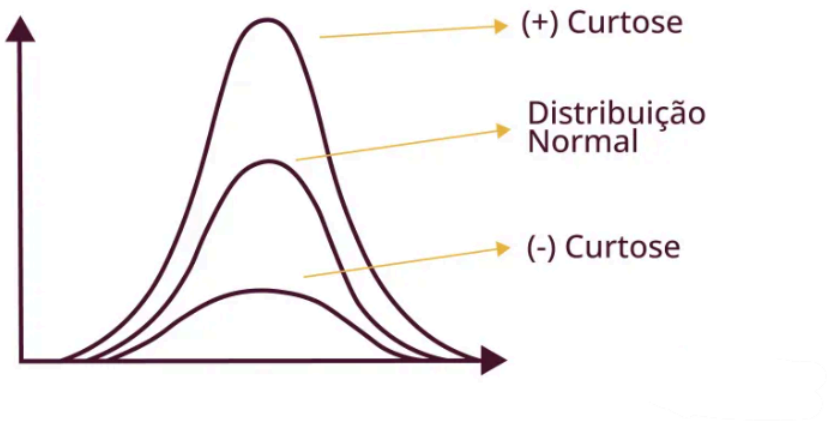

Figura 2. Demonstração dos tipos de curtoses.. Fonte: *DYNAMOX*, 2023.

Exemplo:
Notas de duas turmas:
- Turma A: 6, 6, 6, 7, 7 (curtose baixa – pouca variação).
- Turma B: 0, 10, 6, 9, 1 (curtose alta – mais valores extremos).

- **Estatística Inferencial**

A estatística inferencial expande os horizontes da análise ao permitir que conclusões sejam tiradas sobre populações inteiras com base em amostras representativas. Diferente da estatística descritiva, que se limita a descrever o que é observado, a inferencial trabalha com incertezas e probabilidades, usando métodos que permitem generalizações fundamentadas. Conceitos importantes para o entendimento da estatística inferencial são:

Amostragem: É o processo de selecionar um subconjunto de indivíduos ou itens de uma população maior para fazer inferências sobre ela. Uma amostra deve ser representativa da população para garantir resultados confiáveis.

Exemplo:
Uma empresa quer saber a satisfação de seus clientes. Em vez de perguntar para todos os 10.000 clientes, ela seleciona uma amostra de 500 clientes de forma aleatória. Com base nas respostas dessa amostra, a empresa pode inferir o nível de satisfação da população inteira.

Distribuição: Refere-se ao padrão de como os dados de uma variável se distribuem. A distribuição pode ser representada graficamente (como um histograma, figura 3) e matematicamente (como a normal, figura 4; binomial, figura 5, ou Poisson, figura 6).

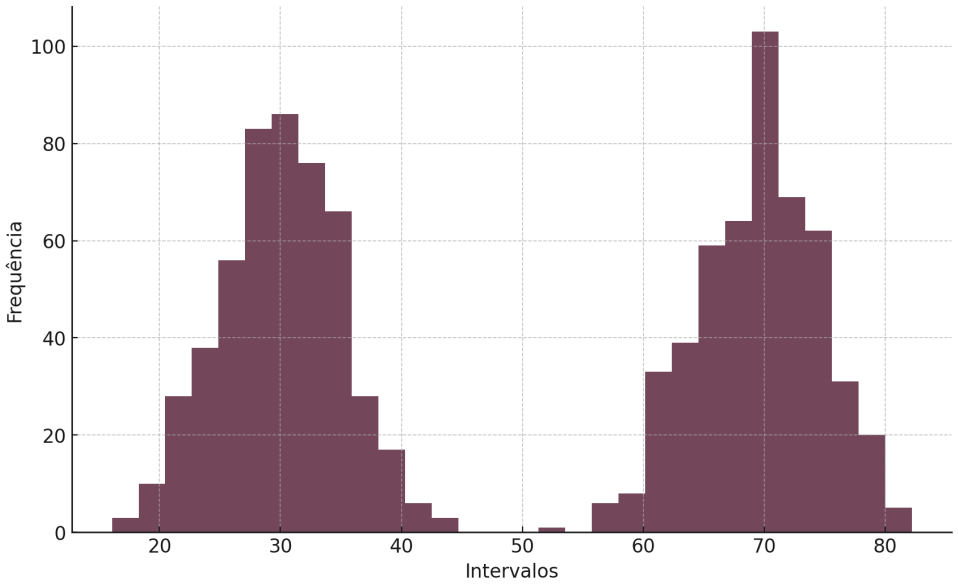

Figura 3. Histograma. Fonte: Elaboração própria.

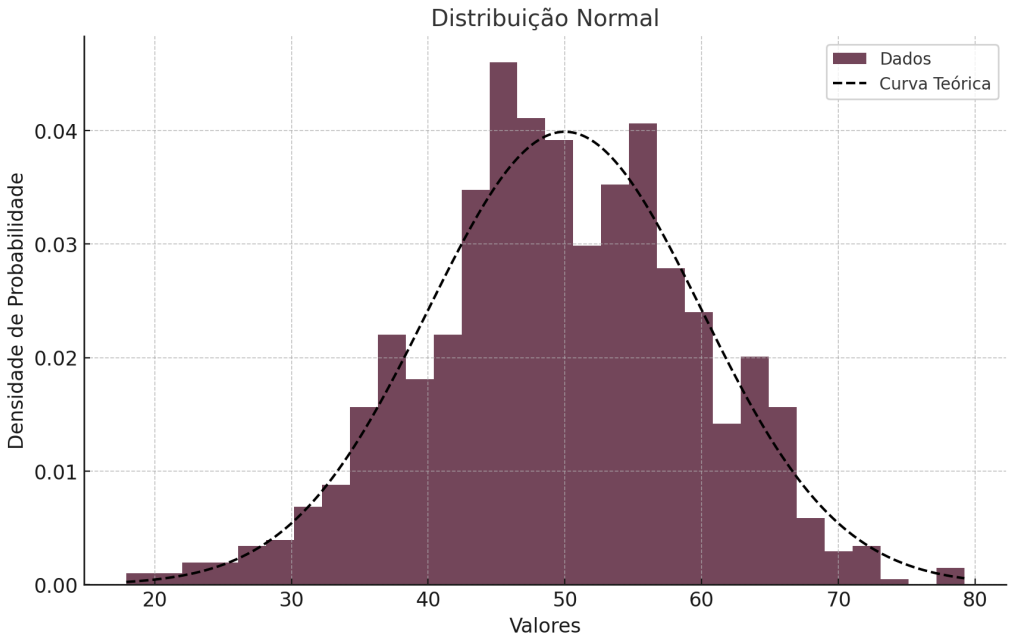

Figura 4. Distribuição normal. Fonte: Elaboração própria.

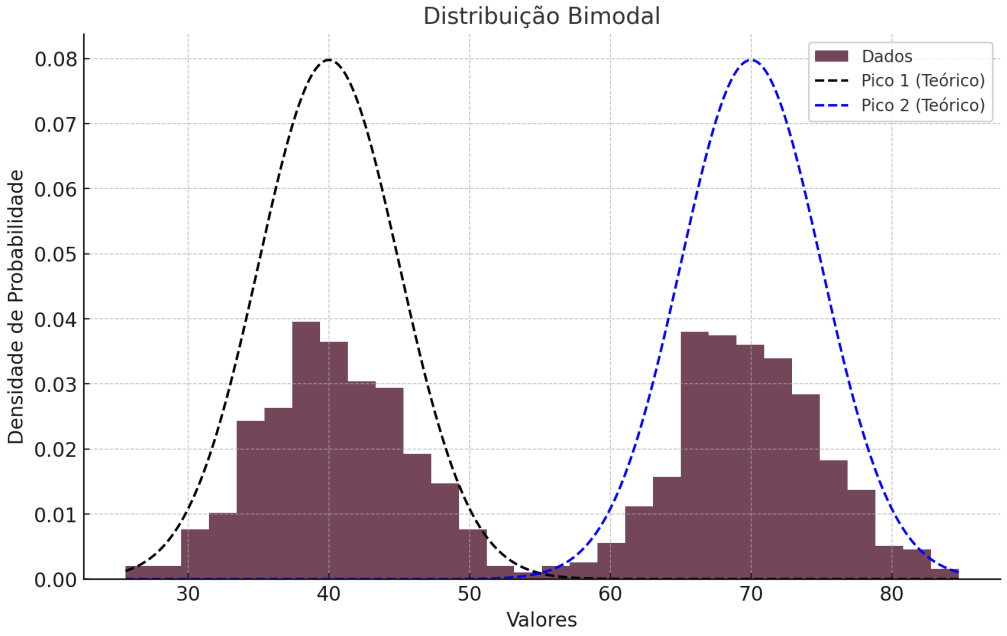

Figura 5. Distribuição bimodal. Fonte: Elaboração própria.

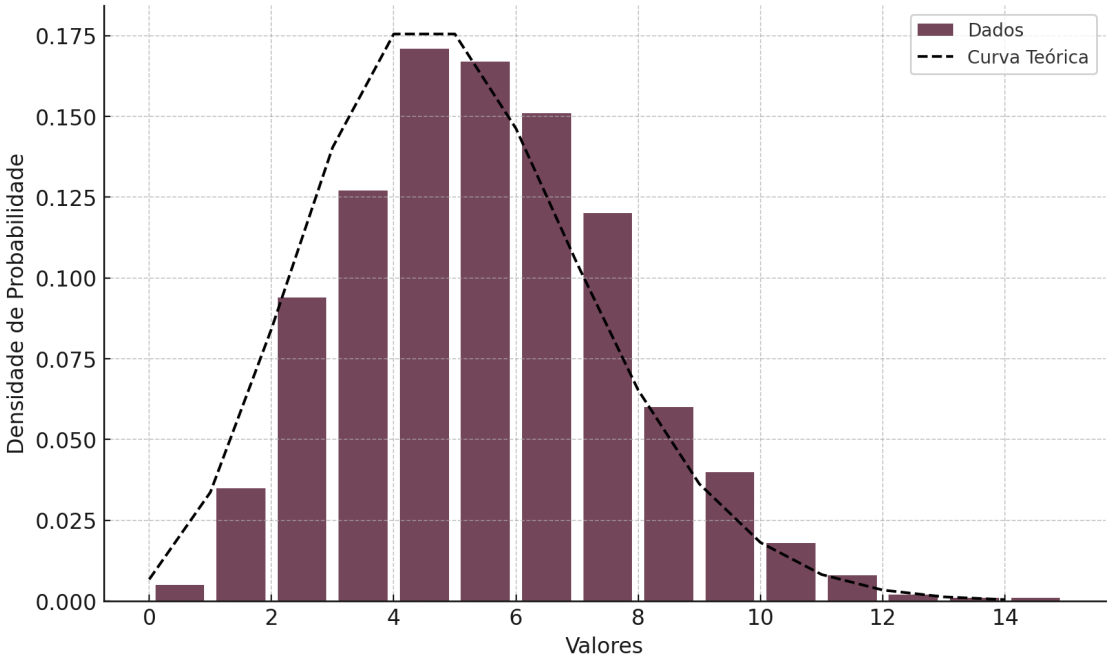

Figura 6. Distribuição de Poisson. Fonte: Elaboração própria.

Intervalo de Confiança: É uma faixa de valores usada para estimar um parâmetro desconhecido da população, como a média. Ele vem acompanhado de um nível de confiança (ex.: 95%), indicando a probabilidade de que o parâmetro esteja dentro do intervalo.

Exemplo:
Uma pesquisa sobre o peso médio de uma população mostrou que a média da amostra é 70 kg, com um intervalo de confiança de 95% entre 68 kg e 72 kg. Isso significa que estamos 95% confiantes de que o peso médio da população está entre 68 kg e 72 kg.

Teste de Hipóteses: É uma técnica usada para avaliar uma suposição (hipótese) sobre um parâmetro populacional com base nos dados de uma amostra. O teste gera uma hipótese nula (H_0), que será aceita ou rejeitada com base nos resultados.

Exemplo:
Uma farmácia quer saber se um novo medicamento é mais eficaz do que o atual.

- H_0 (Hipótese nula): O novo medicamento não é mais eficaz.
- H_1 (Hipótese alternativa): O novo medicamento é mais eficaz.
  Após conduzir um teste com 100 pacientes, os resultados mostram que o novo medicamento tem um efeito significativamente melhor (p-valor < 0,05). Assim, a farmácia rejeita H_0 e conclui que o novo medicamento é mais eficaz.

Análise de Variância: A Análise de Variância (ANOVA) é um método estatístico usado para comparar as médias de dois ou mais grupos e determinar se há diferenças estatisticamente significativas entre elas. A ANOVA pode ser realizada com o pacote scipy usando a função `f_oneway`. Etapas básicas:

- Organizar os dados: Separe os valores das variáveis em grupos distintos.
- Realizar o teste: Use `scipy.stats.f_oneway(grupo1, grupo2, ...)` para comparar as médias.
- Interpretar os resultados. O teste retorna dois valores:

  statistic: valor F (razão da variância entre grupos sobre a variância dentro dos grupos).

  p-value: se for menor que o nível de significância (ex.: 0.05), rejeitamos a hipótese de médias iguais.

ANOVA assume que os dados seguem uma distribuição normal e possuem variâncias similares entre os grupos.

2.3. Visualização de Dados

Os gráficos são ferramentas poderosas de visualização e interpretação de dados. Eles ajudam a transformar informações complexas em representações visuais que facilitam a identificação de padrões, tendências e anomalias. Isso não só torna os dados mais acessíveis, como também permite a comunicação clara e eficiente dos *insights* obtidos.

Os gráficos são ferramentas poderosas de visualização e interpretação de dados. Eles ajudam a transformar informações complexas em representações visuais que facilitam a identificação de padrões, tendências e anomalias. Isso não só torna os dados mais acessíveis, como também permite a comunicação clara e eficiente dos *insights* obtidos. Além disso, os gráficos são essenciais para a tomada de decisões estratégicas. Ao transformar dados em narrativas visuais, eles revelam oportunidades, facilitam a colaboração e garantem que as informações sejam compreendidas e aplicadas de forma eficaz, impulsionando o sucesso de qualquer iniciativa.

Cada gráfico tem seu papel específico na análise de dados, dependendo do que se deseja explorar. Aqui estão alguns tipos de gráficos:

- **Gráficos de dispersão**

Mostram a relação entre duas variáveis numéricas, plotadas em um sistema de coordenadas cartesianas. Cada ponto representa um par de valores (x, y). Eles servem para identificar padrões, tendências, correlações e *outliers*. São úteis para analisar a relação entre variáveis como idade e salário, altura e peso, etc.

Exemplos de uso: Análise de correlação entre variáveis, identificação de *outliers*, modelagem de regressão.

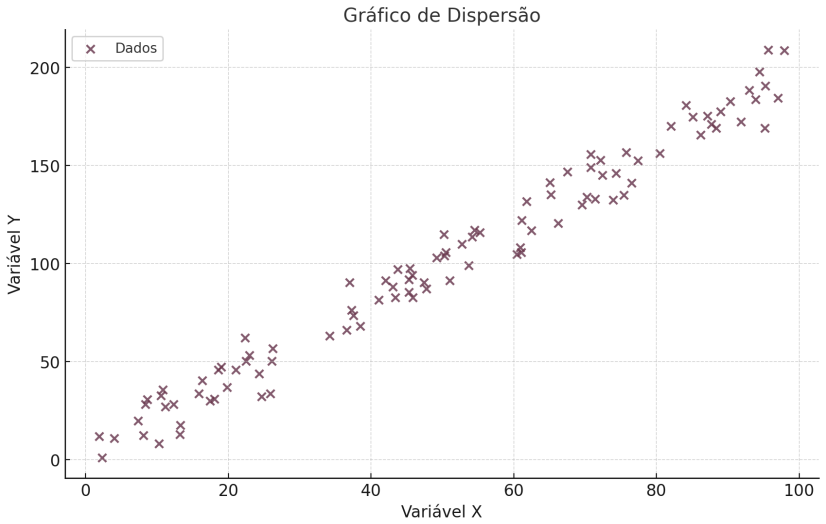

Figura 7. Gráfico de dispersão. Fonte: Elaboração própria.

- **Histogramas**

Representam a distribuição de frequência de uma variável numérica contínua. Os dados são divididos em intervalos (classes) e a altura de cada barra indica a frequência dos dados dentro daquele intervalo. Eles são úteis para visualizar melhor a forma da distribuição dos dados, identificar a centralidade (média, mediana), dispersão (desvio padrão) e assimetria. A figura 3 mostra a forma de um histograma.

Exemplos de uso: Análise da distribuição de idades, salários, notas de provas, etc.

- **Gráficos de barras**

Utilizados para comparar categorias ou grupos. As barras representam as categorias e a altura de cada barra indica a frequência ou valor de uma variável numérica associada a cada categoria. Eles são usados para comparar frequências, proporções ou médias entre diferentes categorias. São úteis para visualizar dados categóricos ou discretos.

Exemplos de uso: Comparação de vendas entre produtos, distribuição de gênero, preferências de consumidores, etc.

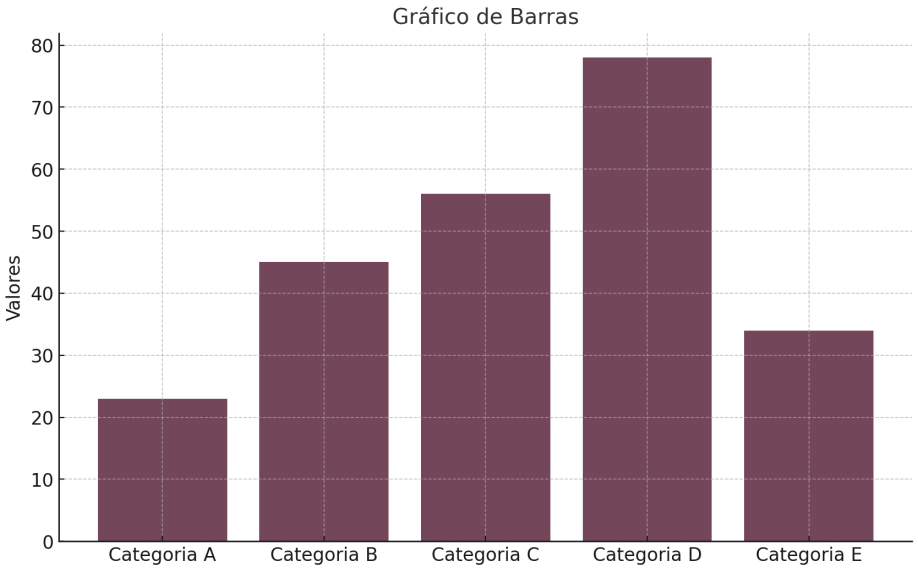

Figura 8. Gráfico de barras. Fonte: Elaboração própria.

- **Boxplots**

Resumo gráfico da distribuição de um conjunto de dados. Mostram a mediana, quartis (Q1 e Q3), amplitude interquartil (IQR) e *outliers*. Eles servem para comparar a distribuição de várias variáveis ou grupos, identificar *outliers* e avaliar a variabilidade dos dados.

Exemplos de uso: Comparar a distribuição de salários entre diferentes setores, identificar valores atípicos em um conjunto de dados, etc.

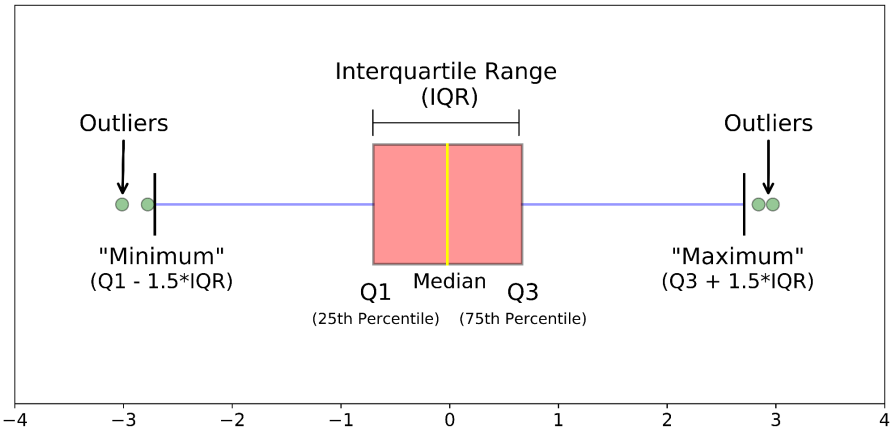

Figura 9. Boxplot. Fonte: Elaboração própria.

- **Heatmaps**

Representam dados em uma matriz, utilizando cores para indicar a intensidade de um valor numérico. As cores mais intensas indicam valores mais altos e as cores mais claras, valores mais baixos. Eles são utilizados para visualizar a relação entre duas variáveis categóricas ou numéricas, identificar padrões e agrupamentos. São úteis para analisar grandes volumes de dados. Exemplos de uso: Analisar correlações entre variáveis, visualizar matrizes de confusão em classificação, analisar dados de expressão gênica, etc.

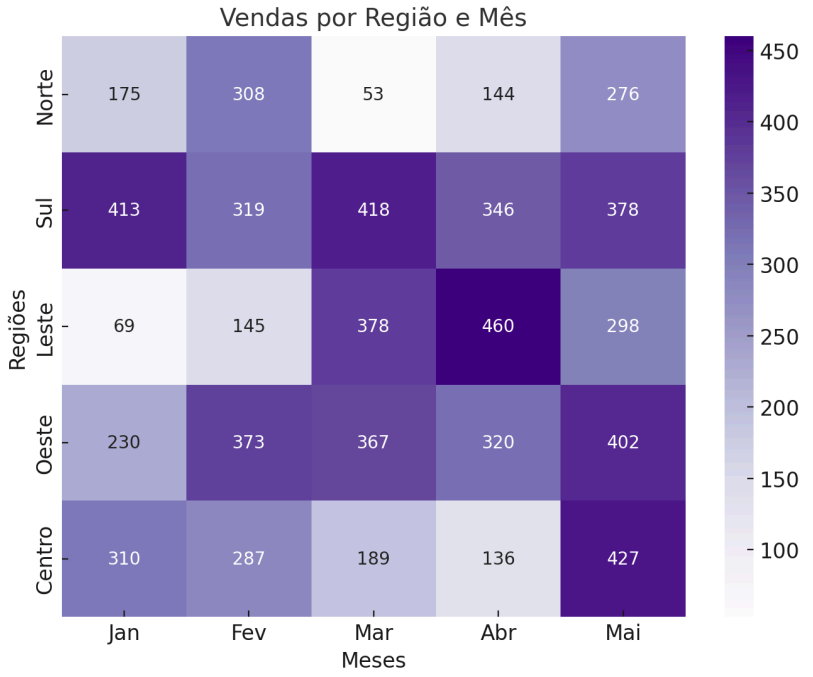

Figura 10. Heatmap de vendas por região. Fonte: Elaboração própria.

A escolha do gráfico adequado depende do tipo de dados que você possui e da pergunta que deseja responder.

2.4. Computação

A computação é a força que potencializa a análise de dados nos dias atuais. Enquanto a estatística fornece as ferramentas conceituais, os *softwares* e linguagens de programação possibilitam o processamento e a análise de grandes volumes de dados de maneira eficiente e replicável.

Ferramentas como Excel, SPSS, e R são conhecidas por suas funcionalidades estatísticas, mas é com **Python** que a análise de dados ganha flexibilidade e poder, sendo capaz de integrar estatísticas, visualizações e aprendizado de máquina em um único ambiente. Na prática, bibliotecas como **Pandas**, **NumPy** e **Matplotlib** são essenciais. Elas permitem carregar, manipular e visualizar dados rapidamente, **Jupyter Notebooks** são amplamente utilizados por sua interatividade e por combinarem código, texto explicativo e visualizações em um único arquivo. Uma ótima alternativa ao Jupyter é o **Google Colab.**

Essa oficina utilizará exatamente esse ecossistema – As bibliotecas Pandas, NumPy, Matplotlib e Seaborn, integradas em um *notebook* interativo. Esse ambiente permitirá exemplificar, de maneira prática, as técnicas e conceitos apresentados, implementando-os diretamente em código Python.

3. Materiais

Para consolidar os conceitos e técnicas apresentados nesta oficina, foram desenvolvidos dois *notebooks* interativos: um focado na elucidação dos conceitos e métodos ensinados, e outro estruturado como um desafio prático final para reforçar o aprendizado. Ambos os *notebooks* estão disponíveis publicamente em nosso repositório no GitHub, com acesso livre a todos que tiverem interesse em explorar mais a fundo o mundo da análise de dados.

Link: https://github.com/ladata-ufs/X-SEMAC/tree/main

O notebook principal cobre desde os primeiros passos com exploração de dados e estatísticas descritivas até a aplicação de técnicas mais avançadas, enquanto o notebook de desafio propõe um problema prático para os participantes testarem suas habilidades recém-adquiridas.

4. Finalização

Ao longo desta oficina, os fundamentos da análise de dados foram apresentados de forma prática e integrada, desde conceitos básicos até aplicações reais com ferramentas modernas. Cada etapa demonstrou como a análise de dados é um

processo interligado. O primeiro contato permitiu compreender as características iniciais dos dados. A limpeza garantiu a confiabilidade das informações enquanto a visualização auxiliou no entendimento de diversos comportamentos do conjunto de dados e a análise de avançada trouxe à tona conexões que orientam decisões e estratégias.

Através do conteúdo, reforça-se a importância da análise de dados em diferentes áreas do conhecimento. Seja na resolução de problemas, no apoio à tomada de decisões ou na geração de insights, ela se destaca como uma ferramenta versátil e essencial. Ao abordar os dados com rigor e criatividade, é possível identificar oportunidades, otimizar processos e gerar impacto em cenários variados.

Independentemente da área de atuação, a aplicação dos conceitos e técnicas apresentados geram oportunidades de interpretar o mundo. Contanto que haja dados, a análise não se limita a um campo específico; ela está presente em negócios, saúde, educação, pesquisa e muitos outros setores.

Analisar dados não é apenas sobre números, mas sobre transformar informações em conhecimento e conhecimento em impacto. Ao dominar esse processo, cria-se o potencial de influenciar decisões e moldar realidades com base em evidências.

Glossário

Insights:

- **Definição:** Conhecimento profundo e perspicaz sobre algo, obtido a partir da análise de dados ou de uma experiência. São descobertas significativas que podem levar a novas ideias, estratégias ou soluções.

Outliers:

- **Definição:** Valores que se desviam significativamente da maioria dos outros dados em um conjunto. São observações que parecem não pertencer ao padrão geral.

Software:

- **Definição:** Conjunto de programas e dados que permitem a um computador executar tarefas específicas.

Notebook:

- **Definição:** Ambiente de desenvolvimento interativo que permite combinar código, texto e visualizações em um único documento. É muito utilizado para análise de dados e machine learning.

Referências

ATLASSIAN. **A complete guide to bar charts.** Disponível em: https://www.atlassian.com/data/charts/bar-chart-complete-guide. Acesso em: 21 nov. 2024.

ATLASSIAN. **A complete guide to scatter plots.** Disponível em: https://www.atlassian.com/data/charts/what-is-a-scatter-plot. Acesso em: 21 nov. 2024.

DYNAMOX. **Métricas de análise de vibração: Curtose e Skewness.** Disponível em: https://dynamox.net/blog/metricas-de-analise-de-vibracao-curtose-e-skewness. Acesso em: 21 nov. 2024.

FÁVERO, Luiz Paulo; BELFIORE, Patrícia. **Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata®.** Elsevier Brasil, 2017.

HERMES, Thais Schmidt Vitali et al. Criança diabética do tipo 1 e o convívio familiar: repercussões no manejo da doença. **Saúde em Debate**, v. 42, p. 927-939, 2018.

KDNUGGETS. **Understanding Boxplots.** Disponível em: https://www.kdnuggets.com/2019/11/understanding-boxplots.html. Acesso em: 21 nov. 2024.

LAERD STATISTICS. **Histograms.** Disponível em: https://statistics.laerd.com/statistical-guides/understanding-histograms.php. Acesso em: 21 nov. 2024.

LO, Andrew W.; WANG, Jiang. Trading volume: definitions, data analysis, and implications of portfolio theory. **The Review of Financial Studies**, v. 13, n. 2, p. 257-300, 2000.

MORETTIN, Pedro A.; SINGER, Julio M. **Estatística e Ciência de Dados**. Grupo Gen-LTC, 2022.

TECHTARGET. **What is a heat map (heatmap)?**. Disponível em: https://www.techtarget.com/searchbusinessanalytics/definition/heat-map. Acesso em: 21 nov. 2024.

TUKEY, John W. **Exploratory data analysis.** Reading/Addison-Wesley, 1977.