Secretaria
de Planejamento, Gestão
e Desenvolvimento
Regional

GOVERNO DE PERNAMBUCO
ESTADO DE MUDANÇA

Achando padrões

Curso Ciência de Dados para a Gestão Pública

Júlia Barrêto

14 de Maio de 2025

Achando padrões

1. Exploração gráfica e numérica.
2. Os 5 números de Tukey.
3. Outliers e anomalias.
4. Agrupamento.
5. Lei de Benford.

Utilização do GitHub

- Importante para programação em equipe
- Gestão de conhecimento
- Armazenamento de códigos
- Controle de versionamento
- Link do GitHub do curso:
- https://github.com/juliabarretocp/curso_sefaz/tree/main
- Prazo de uma semana pós finalização do curso

Exploração gráfica e numérica.

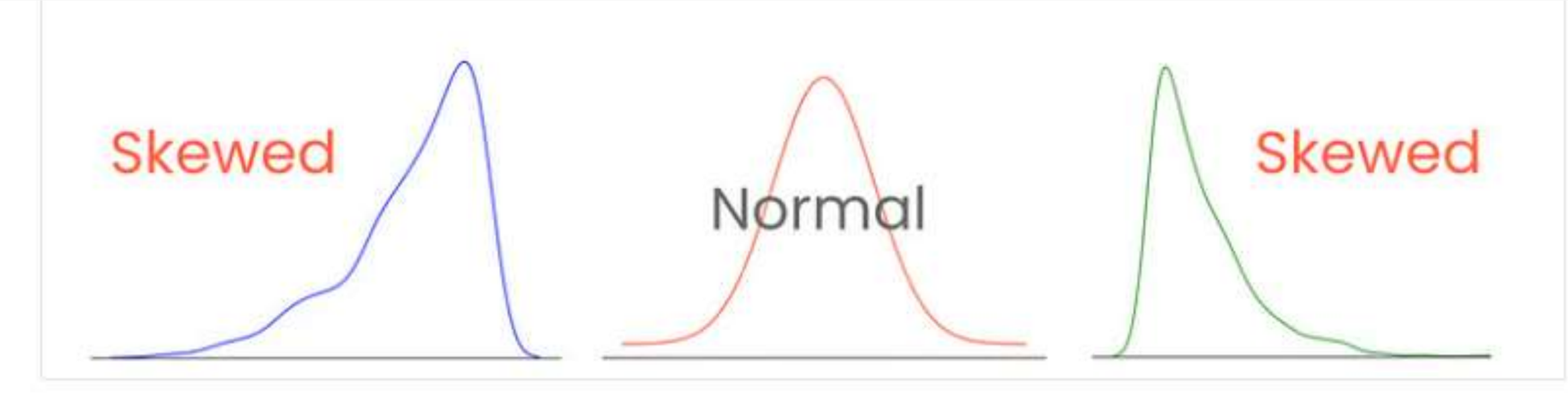
- A Análise Exploratória de Dados (AED) é um conjunto de técnicas estatísticas utilizadas para resumir, visualizar e entender os principais aspectos de um conjunto de dados antes de aplicar modelos mais complexos.
- Permite identificar padrões, detectar anomalias, testar hipóteses e verificar suposições.

Exploração gráfica e numérica.

- Exploração Numérica: calcular medidas estatísticas que resumem as características dos dados
- Medidas de Tendência Central:
- Média: soma dos valores dividida pelo número de observações.
- Mediana: valor central que divide o conjunto de dados ordenado em duas partes iguais.
- Moda: valor que ocorre com maior frequência
- Medidas de Dispersão:
- Desvio Padrão: mede a dispersão dos dados em relação à média.
- Variância: quadrado do desvio padrão.
- Amplitude: diferença entre o maior e o menor valor.
- Intervalo Interquartil (IQR): diferença entre o terceiro quartil (Q3) e o primeiro quartil (Q1).

Exploração gráfica e numérica.

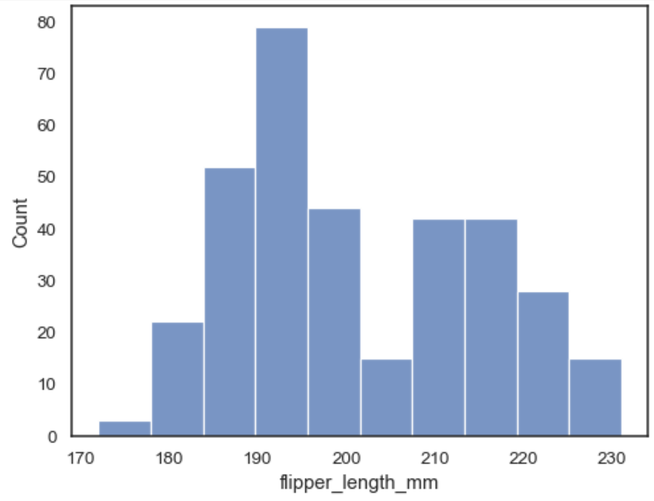
- Medidas de Forma:
- Assimetria (Skewness): indica a simetria da distribuição dos dados.
- Curtose (Kurtosis): mede a "pontiagudez" da distribuição.

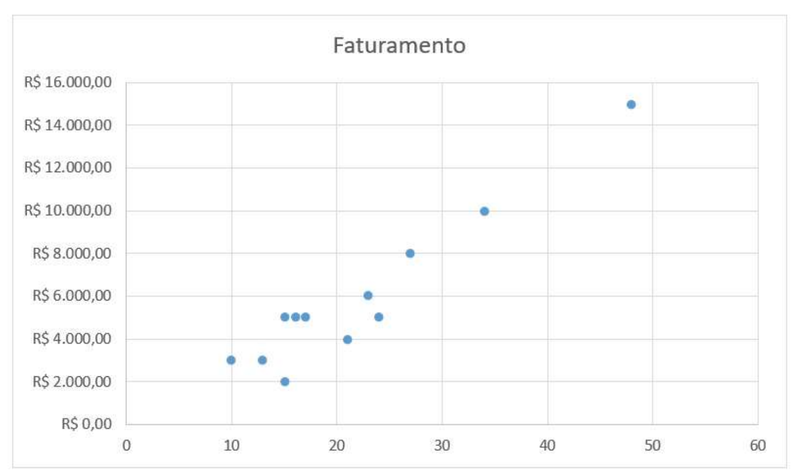


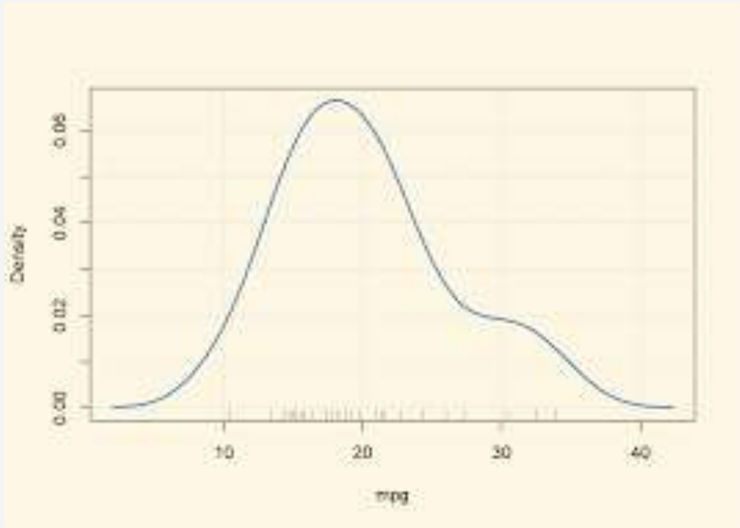
Exploração gráfica e numérica.

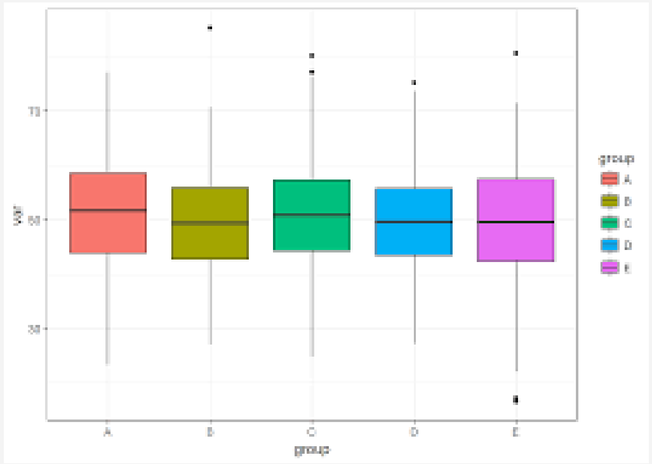
- Visualizações ajudam a compreender melhor os dados:
- Histograma: representa a distribuição de frequências de uma variável contínua.
- Boxplot (Diagrama de Caixa): exibe o resumo de cinco números e identifica outliers.
- Gráfico de Dispersão (Scatter Plot): mostra a relação entre duas variáveis quantitativas.
- Gráficos de Barras e Setores (Pizza): usados para variáveis categóricas.
- Gráfico de Densidade: estimativa suavizada da distribuição de dados contínuos

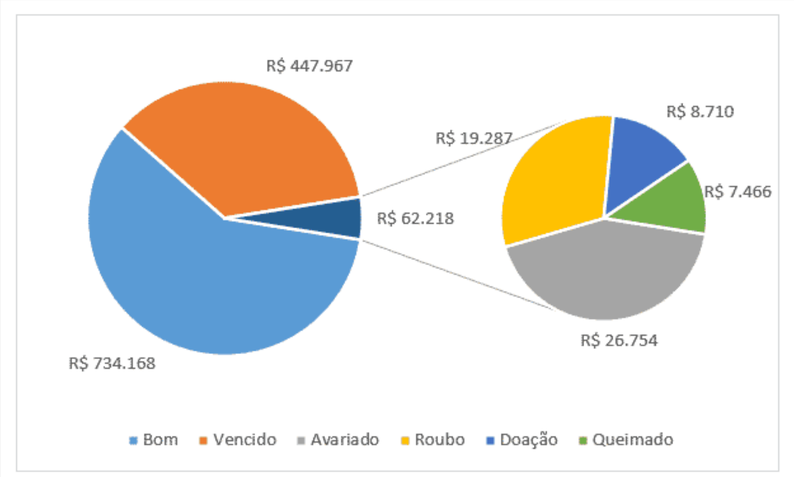
Exploração gráfica e numérica.







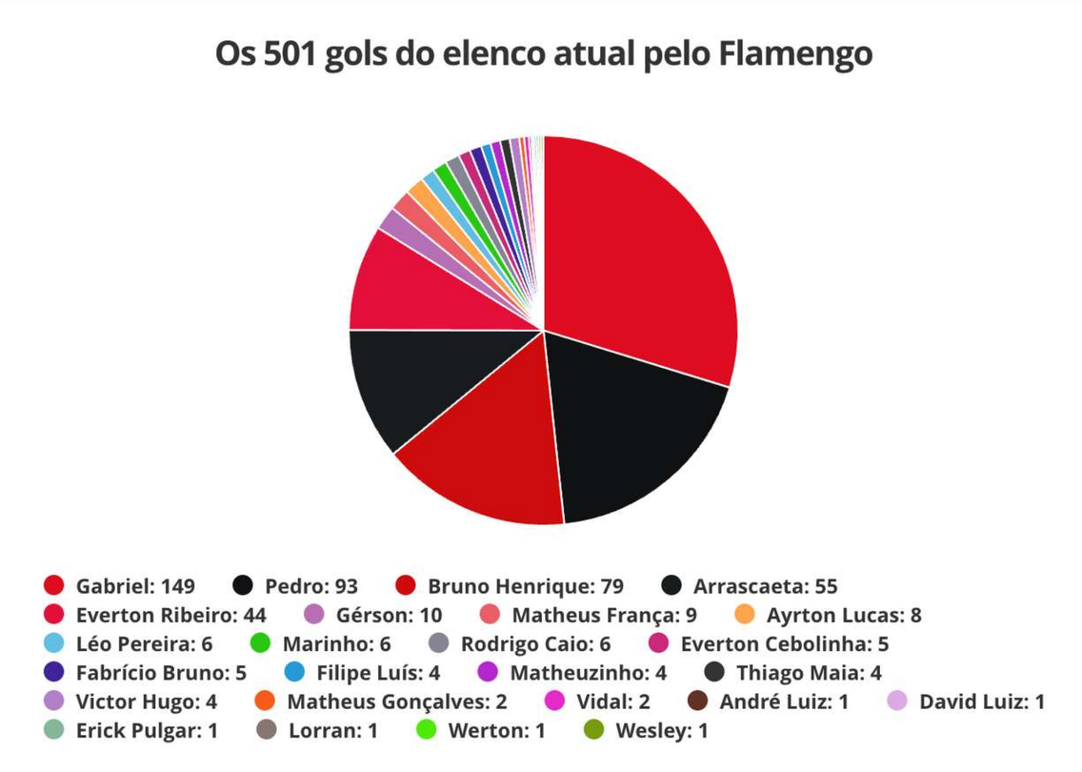




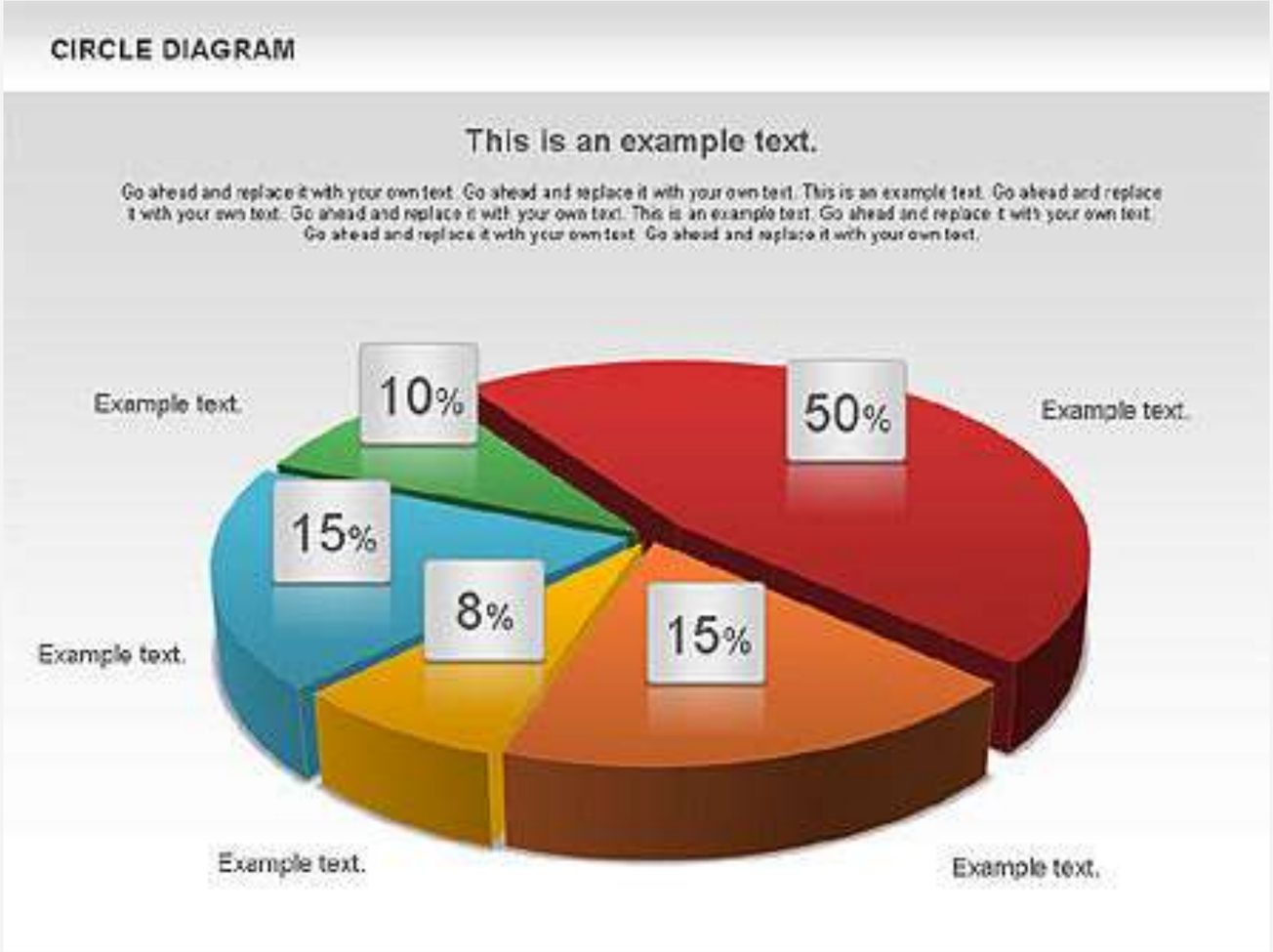
GOVERNO DE PERNAMBUCO
ESTADO DE MUDANÇA

O gráfico de pizza

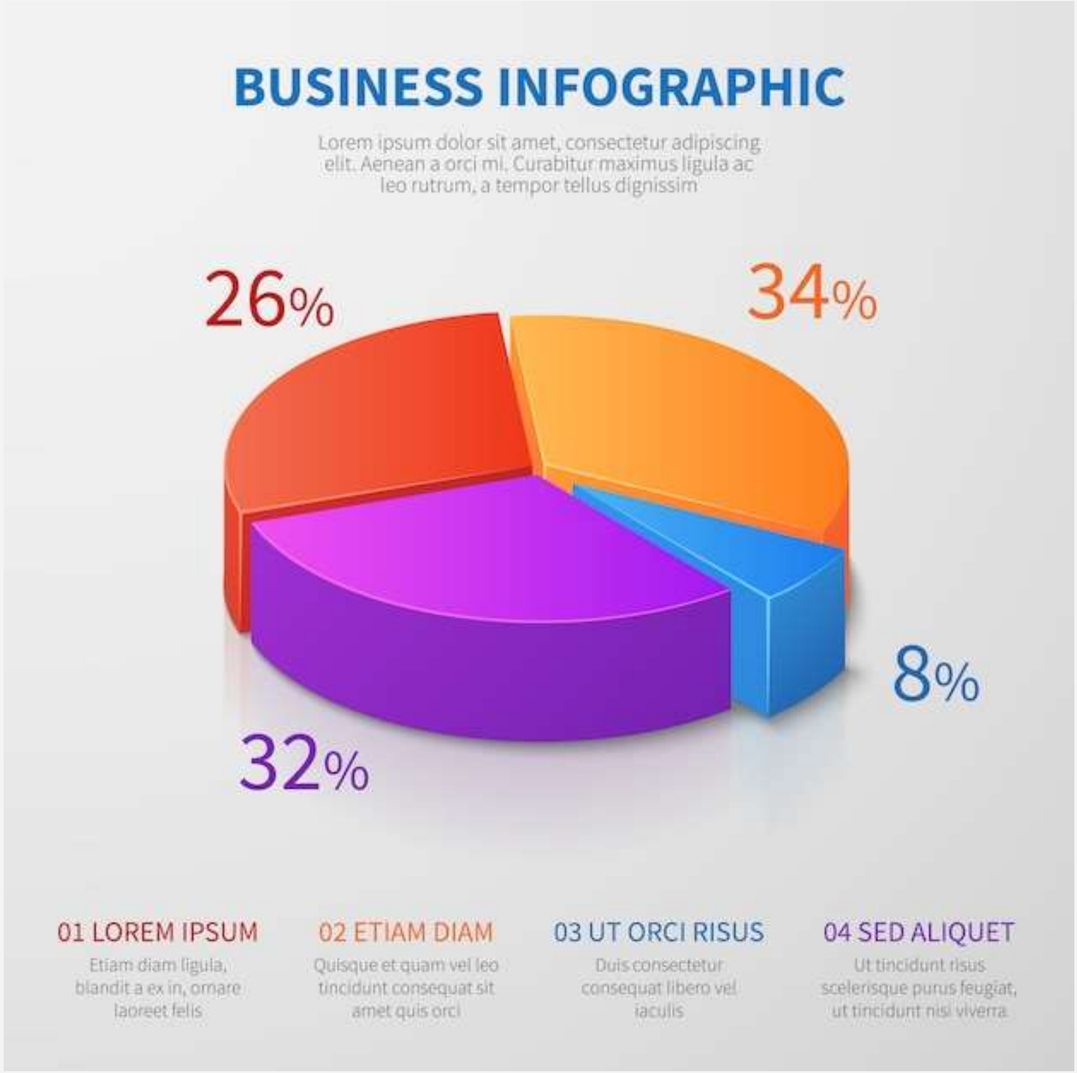
- são difíceis de usar quando se pretende comparar valores próximos, têm dificuldade em representar pequenas porcentagens e podem ser confundidos com outras visualizações devido à forma como representam os dados



O gráfico de pizza



O gráfico de pizza

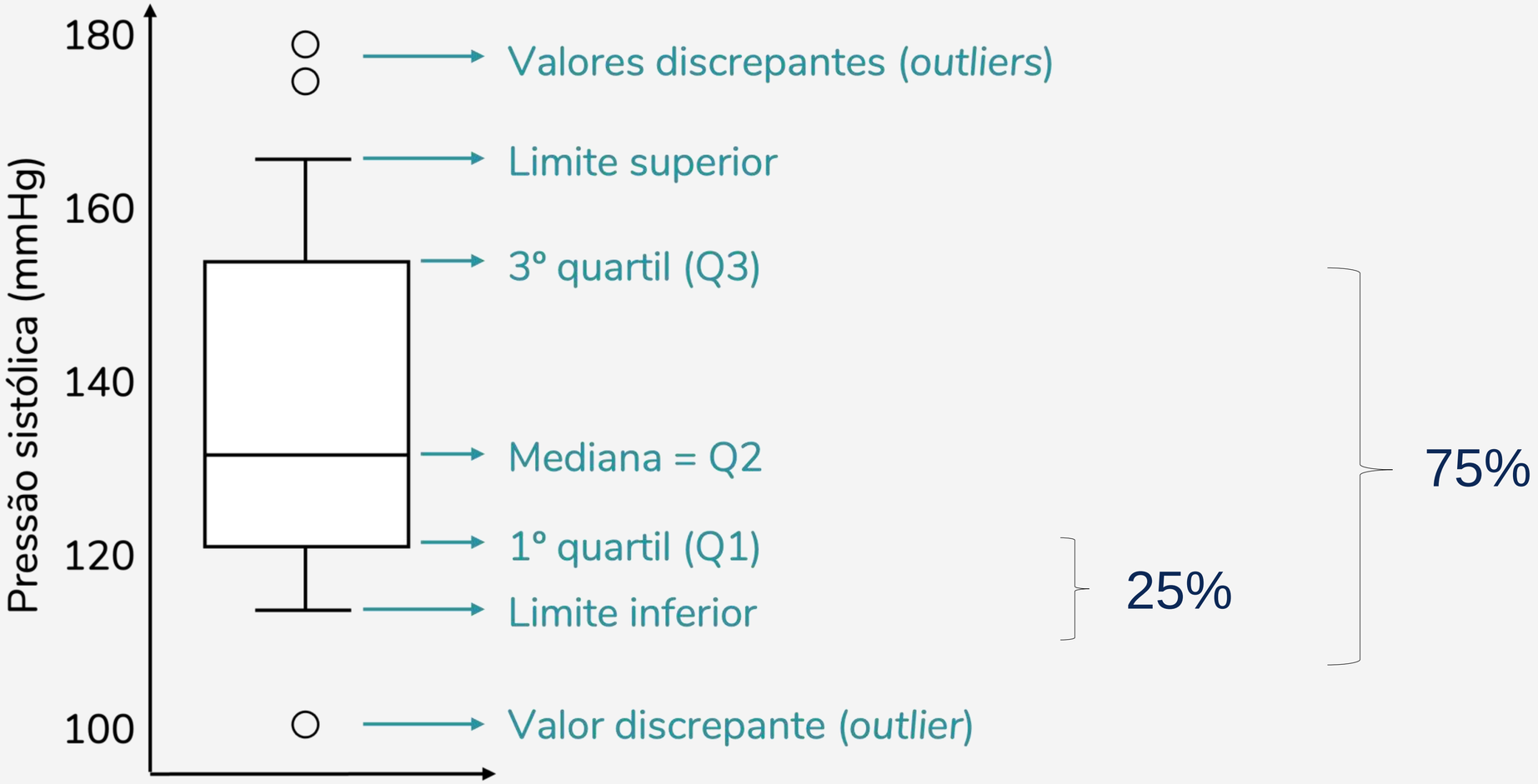


GOVERNO DE PERNAMBUCO
ESTADO DE MUDANÇA

Os 5 Números de Tukey

- O resumo dos cinco números, proposto por John Tukey, é uma estatística descritiva que resume a distribuição de um conjunto de dados:
- Mínimo: menor valor observado.
- Primeiro Quartil (Q1): 25% dos dados estão abaixo desse valor.
- Mediana (Q2): valor central dos dados.
- Terceiro Quartil (Q3): 75% dos dados estão abaixo desse valor.
- Máximo: maior valor observado.
- Esse resumo é frequentemente representado por um boxplot, facilitando a visualização da dispersão e identificação de outliers.

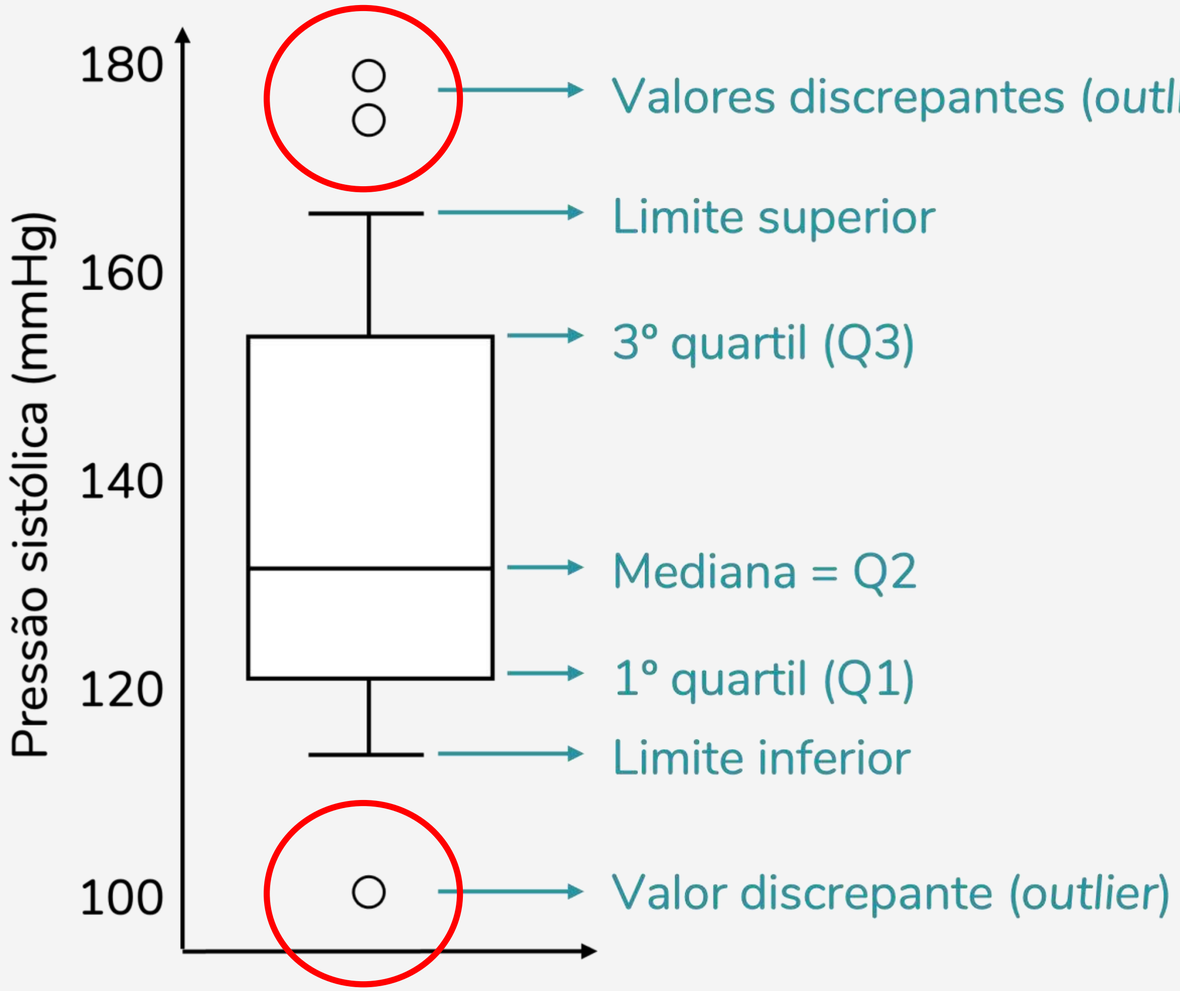
Os 5 Números de Tukey



Outliers e Anomalias

- Outliers são valores que se diferenciam significativamente dos demais dados. Podem indicar variabilidade natural, erros de medição ou fenômenos interessantes.
- Métodos de Detecção:
- Intervalo Interquartil (IQR):Valores abaixo de Q1 - 1,5×IQR ou acima de Q3 + 1,5×IQR são considerados outliers.
- Z-Score: Medida estatística que descreve a relação de um valor com a média de um grupo de valores.
- Calcula quantos desvios padrão um valor está da média.
- Valores com |Z| > 3 são geralmente considerados outliers
- Se um escore Z for 0, indica que a pontuação do ponto de dados é idêntica à pontuação média. Um escore Z de 1,0 indica um valor que representa um desvio-padrão da média.

Outliers



Outliers e Anomalias

- Ele nos diz se um dado é típico ou incomum em comparação com o restante do grupo
- Anomalias: Anomalias são padrões nos dados que não se conformam ao comportamento esperado. Podem ser causadas por erros, fraudes ou eventos raros
- Técnicas de Detecção:
- Modelos Estatísticos: assumem uma distribuição dos dados e identificam desvios significativos.
- Machine Learning em conjuntos mais complexos

Outliers e Anomalias

Conceito	Outlier	Anomalia
Definição	Um ponto de dado que **se afasta significativamente** dos demais valores em uma distribuição.	Um ponto ou padrão que **não se ajusta ao comportamento esperado** dos dados.
Origem	Pode surgir de **variabilidade natural**, erro de medição ou entrada rara.	Pode indicar um **problema ou evento interessante**, como uma fraude ou falha no sistema.
Base Teórica	É definido com base em **propriedades estatísticas** (ex: IQR, Z-score, distância).	Envolve **interpretação contextual** ou modelos mais complexos para detecção.
Exemplo	Um salário de R$ 1.000.000 entre outros de R$ 2.000 a R$ 10.000.	Uma transação de R$ 20.000 feita por um cartão de crédito que normalmente tem gastos de até R$ 500.
Relevância	Pode ser **removido ou tratado** em modelagens, se for considerado ruído.	Deve ser **investigado** porque pode representar um evento importante.

Outliers e Anomalias

- Outlier: Você mede a altura de 100 pessoas adultas. Um valor de 2,40 m aparece, enquanto a maioria está entre 1,55 m e 1,95 m. Isso pode ser:
- Um erro de digitação (talvez era 1,40 m)
- Uma pessoa realmente muito alta
- Uma leitura incorreta
- É um outlier porque quebra o padrão estatístico dos dados.

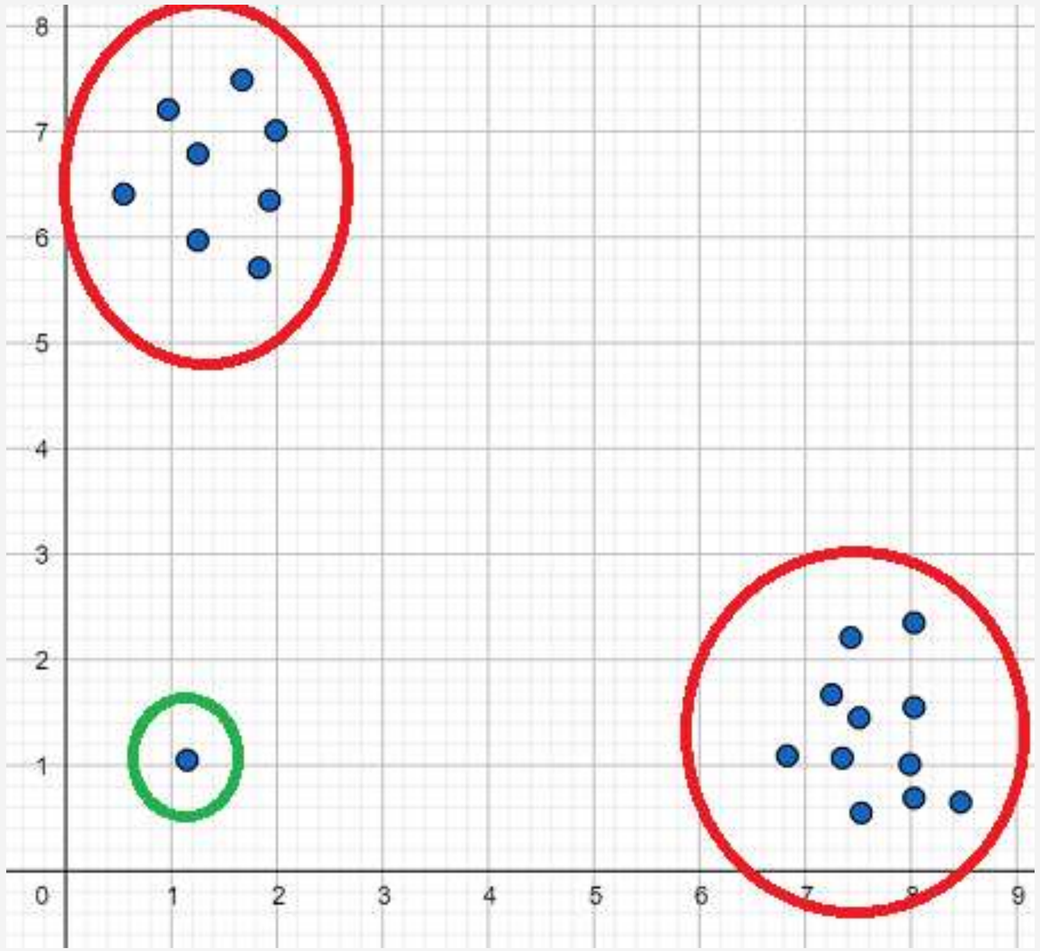
Outliers e Anomalias

- Anomalia: Você monitora as vendas de um e-commerce. Em um dia normal, são feitas 200 vendas. De repente, em um dia há apenas 5 vendas. Isso pode ser:
- Um bug no sistema
- Um problema nos servidores
- Um evento externo (apagão, feriado inesperado)
- É uma anomalia, pois representa um comportamento inesperado que foge da norma do sistema.
- A diferença está no contexto, na causa e no impacto do valor incomum.

Agrupamento (Clustering)

- O agrupamento é uma técnica de aprendizado não supervisionado que visa organizar dados em grupos (clusters) de acordo com a similaridade entre eles
- Algoritmos Comuns:
- K-Means: Divide os dados em K clusters, minimizando a variância dentro de cada grupo. Requer a definição prévia do número de clusters.
- Hierárquico: Cria uma árvore de clusters (dendrograma), permitindo visualizar a formação dos grupos em diferentes níveis de similaridade.
- DBSCAN (Density-Based Spatial Clustering of Applications with Noise):Identifica clusters de forma arbitrária com base na densidade dos dados. Não requer a definição do número de clusters e é eficaz na detecção de outliers.

Agrupamento (Clustering)

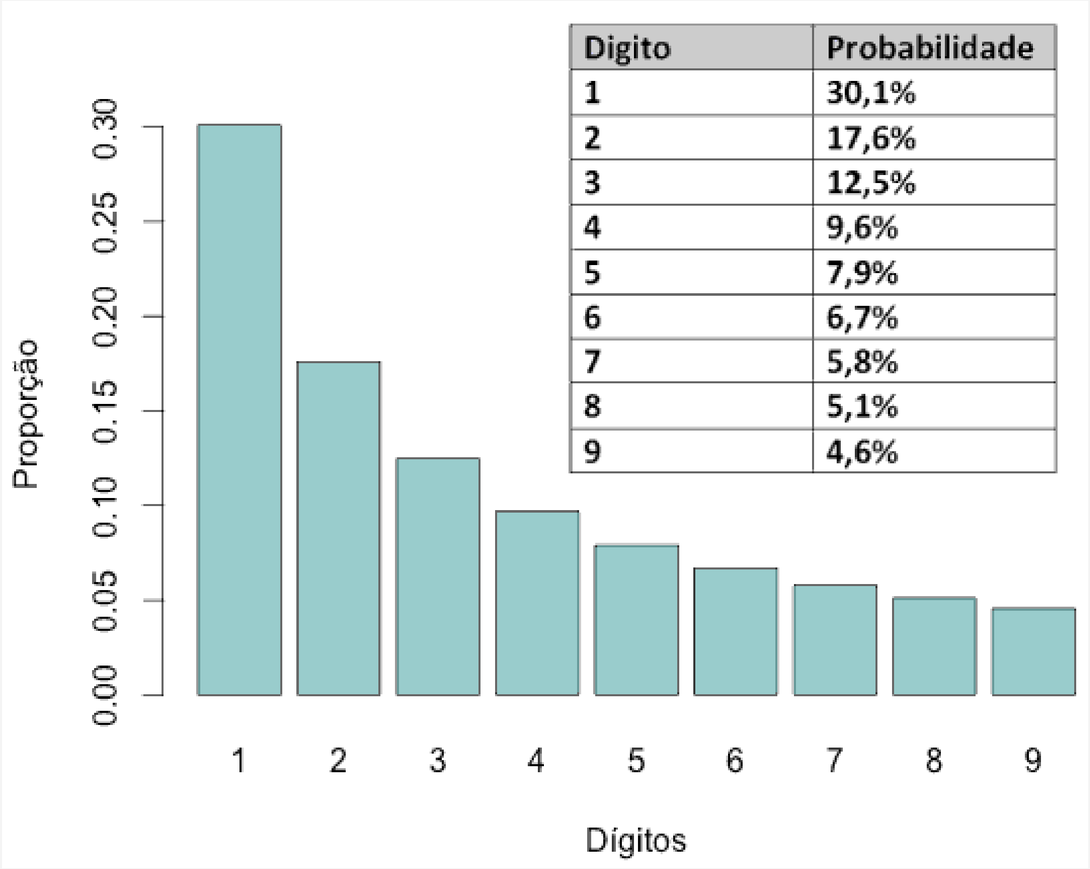


Lei de Benford

- Lei do Primeiro Dígito
- A Lei de Benford descreve a frequência de ocorrência dos dígitos iniciais em muitos conjuntos de dados reais. Segundo essa lei, o dígito 1 aparece como primeiro dígito cerca de 30,1% das vezes, enquanto dígitos maiores ocorrem com menor frequência.
- Ou seja, os dígitos menores aparecem mais
- Mais “conveniente” pro ser humano usar “1 em cada 2” do que 50%.
- Mais fácil 1 metro do que “997 milímetros”

Lei de Benford

- Em determinado conjunto de dados que contém a metragem de terrenos, qual a probabilidade do primeiro dígito ser 1? E 2?
- “Fura” o conceito da probabilidade
- 30,1% de probabilidade de ser 1
- 4,6% de probabilidade de ser 9
- Utilizada na detecção de fraudes em grandes volumes de dados -> se os números de uma planilha não seguem a distribuição esperada, pode ser sinal de fraude



Digito	Probabilidade
1	30,1%
2	17,6%
3	12,5%
4	9,6%
5	7,9%
6	6,7%
7	5,8%
8	5,1%
9	4,6%

Lei de Benford

- Aplicações:
- Detecção de Fraudes: usada por auditores para identificar manipulações em dados financeiros e contábeis.
- Análise de Dados Eleitorais: verifica a autenticidade de resultados de eleições.
- Verificação de Conformidade: em conjuntos de dados científicos e econômicos.
- Não é adequada para dados com valores mínimos e máximos definidos ou para conjuntos de dados pequenos.

GOVERNO DE
PERNAMBUCO
ESTADO DE MUDANÇA