

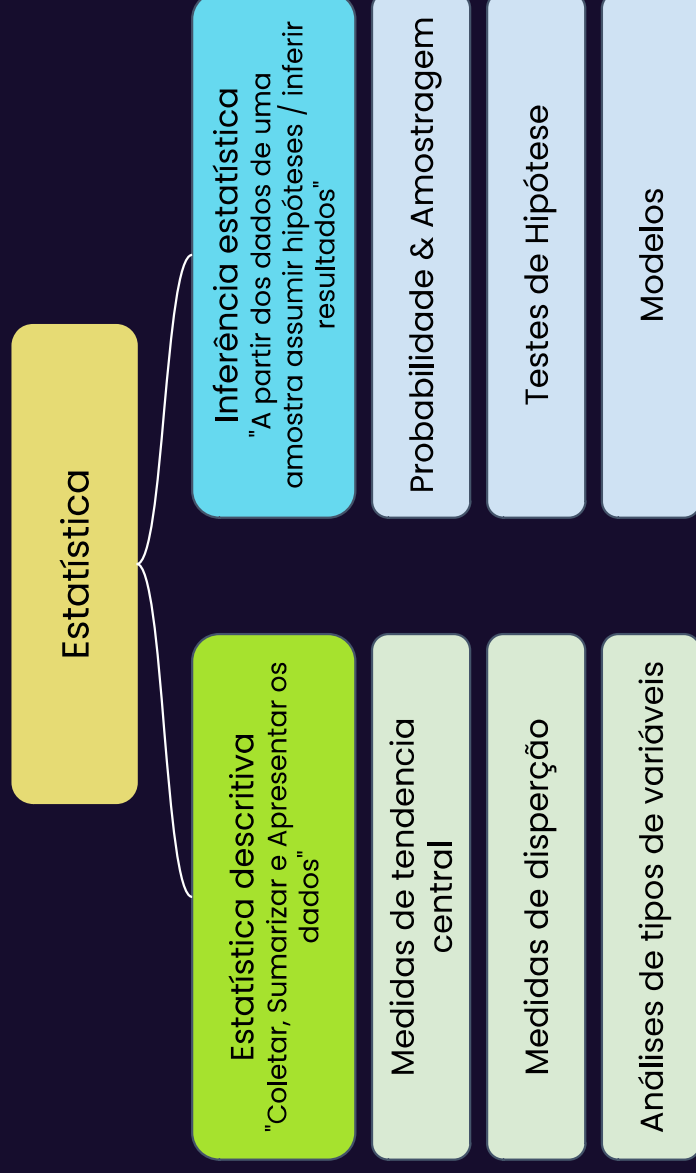
Bootcamp Data Analytics

Estatística Frequências e Medidas

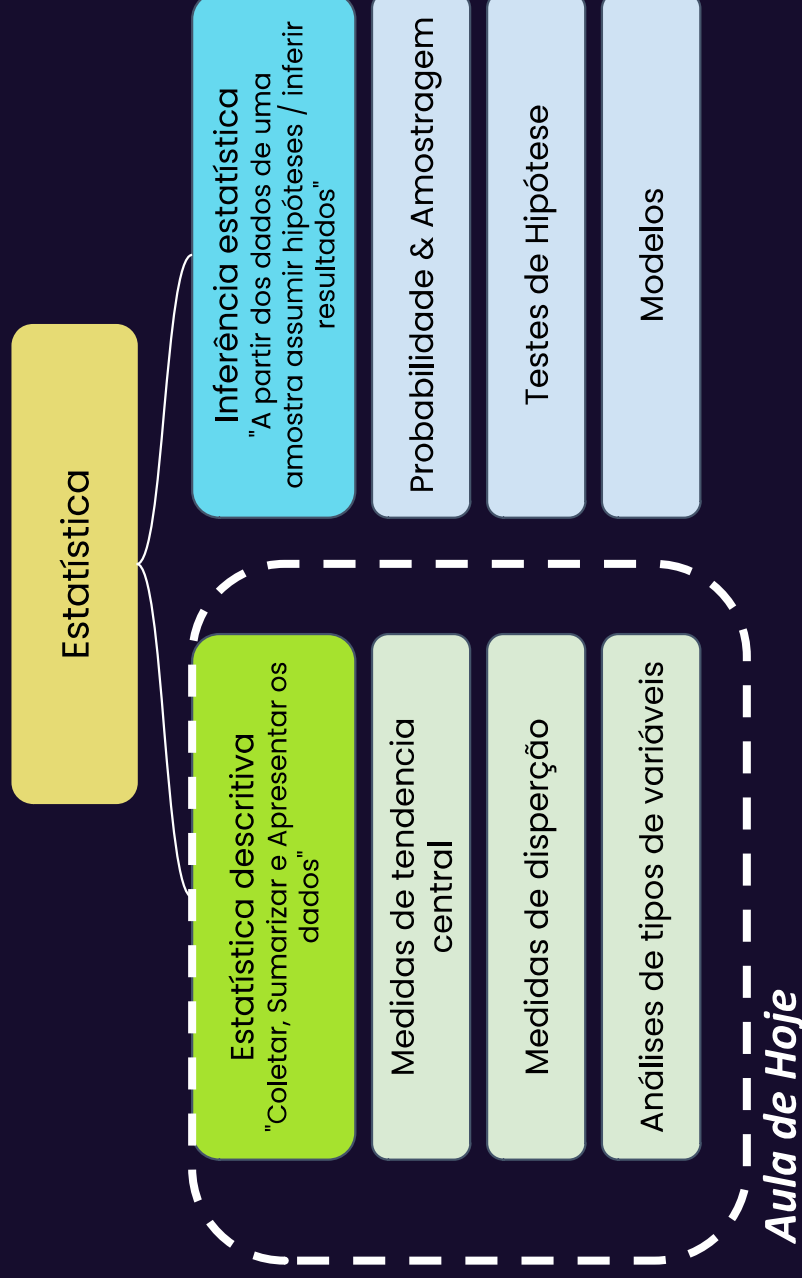
Ana Luiza Pessoa



Em estatística vamos estudar: "A organização, descrição, análise e a interpretação dos dados"



Em estatística vamos estudar: "A organização, descrição, análise e a interpretação dos dados"



Estatística : Frequências e Medidas

Tipos de variáveis



Tipos de variáveis na análise estatística

Numéricas

São variáveis mensuráveis: que possuem valores numéricos ou resultante de contagens.

1. Variáveis numéricas discretas:
Assumem valores finitos e contáveis (Ex: números inteiros)

2. Variáveis numéricas contínuas
Assumem valores contínuos, mensurados a partir de algum instrumento.

Categóricas

São variáveis qualitativas, classificam grupos ou indivíduos

1. Variáveis qualitativas ordinais
Existe uma ordenação nas categorias. (Ex: Ruim, Médio, Bom)

2. Variáveis qualitativas nominais
Variáveis em que não existe ordenação. (Ex: Azul, Vermelho)

Estatística : Frequências e Medidas

Medidas de tendência central



Medidas de Tendência Central

- ▷ O que são? São medidas que buscam visam "resumir" ou descrever os dados refletindo o ponto de equilíbrio ou dados.
- ▷ Exemplo: Em uma entrevista cada uma das 10 candidatas respondeu uma pergunta em um tempo, xi, específico minutos: 2, 3, 1, 4, 2, 2, 3, 1, 4, 2.
- ▷ Ordenando temos: 1, 1, 2, 2, 2, 2, 3, 3, 4, 4

Média

A média será a soma de todos os valores xi dividido pelo numero total de candidatas (Ex: $24/10 = 2,4$)

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Mediana

É o número que divide uma distribuição **ordenada** de dados em 2 partes.

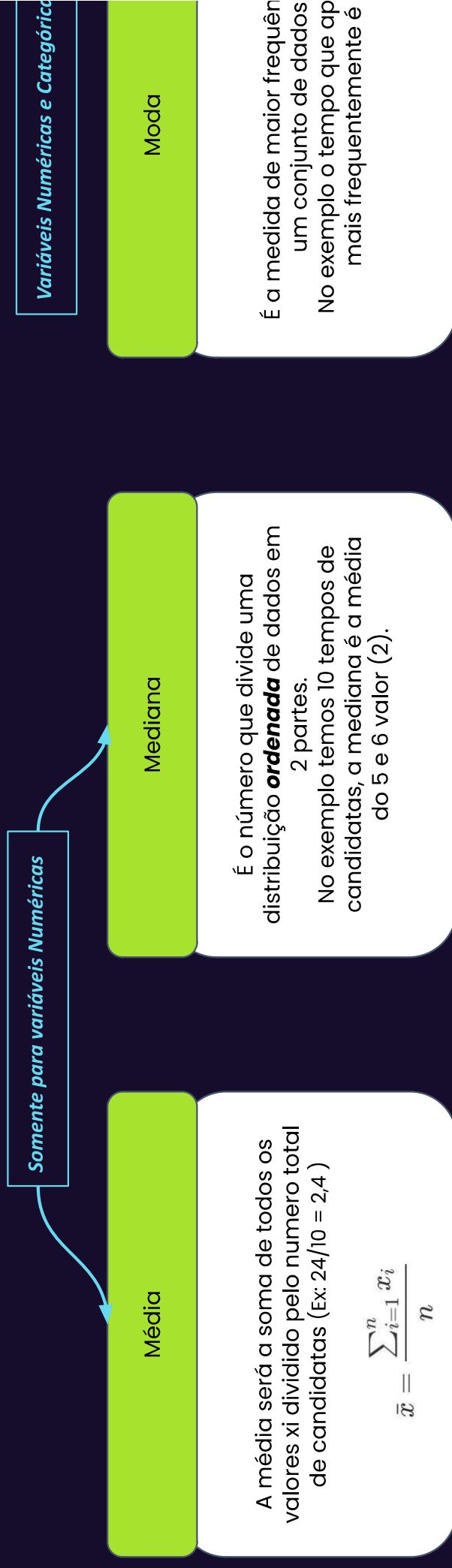
$$= \begin{cases} X\left[\frac{n+1}{2}\right] & \text{Se Impar} \\ \frac{X\left[\frac{n}{2}\right] + X\left[\frac{n}{2} + 1\right]}{2} & \text{Se PAR} \end{cases}$$

Moda

É a medida de maior frequência em um conjunto de dados
No exemplo o tempo que as mais frequentemente é

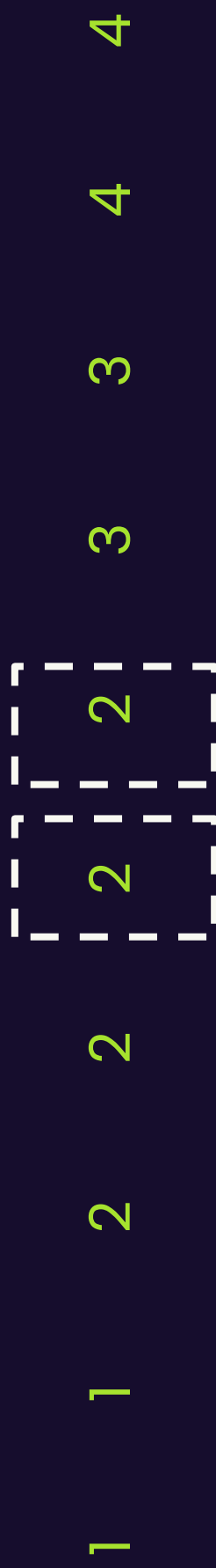
Medidas de Tendência Central

- ▷ O que são? São medidas que buscam visam "resumir" ou descrever os dados refletindo o ponto de equilíbrio ou dados.
- ▷ Exemplo: Em uma entrevista cada uma das 10 candidatas respondeu uma pergunta em um tempo, xi, específico minutos: 2, 3, 1, 4, 2, 2, 3, 1, 4, 2.
- ▷ Ordenando temos: 1, 1, 2, 2, 2, 2, 3, 3, 4, 4



Media, Mediana e Moda: Interpretação visual

Respostas das 10 candidatas ordenadas:



A mediana será a média do quinto e sexto termo, 2,, pois com isso dividimos igualmente a distribuição igualmente (4 termos abaixo, 4 acima)

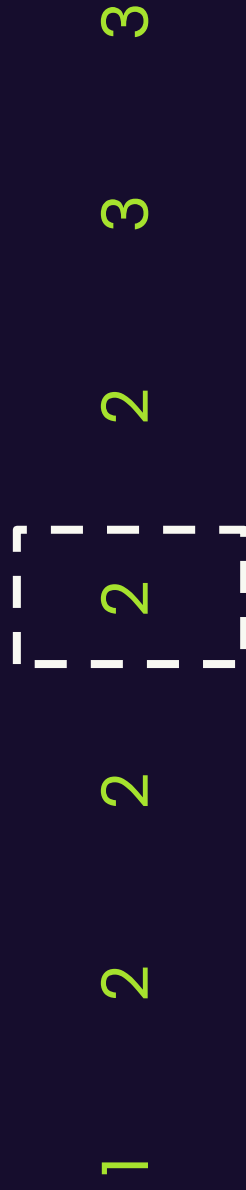
A média será a soma dos termos pelo total de candidatas : $24/10 = 2,4$

A moda será o termo mais frequente: 2



Media, Mediana e Moda: Interpretação visual

E se fossem 7 candidatas com os tempos abaixo?



A mediana será exatamente o quarto termo, 2,, pois com isso dividirmos igualmente a distribuição igualmente (3 termos abaixo, 3 termos acima)

A média será a soma dos termos pelo total de candidatas : $15/7 = 2,14$

A moda será o termo mais frequente: 2



Estatística : Frequências e Medidas

Análise de Dispersão e Outliers



Análise de dispersão de variáveis

- ▷ O que são? São medidas que buscam "resumir" como os dados estão distribuídos; o quão concentrado estão em determinados intervalos e o grau de variação das informações. (somente para variáveis numéricas)
- ▷ Exemplo: Em uma entrevista cada uma das 10 candidatas respondeu uma pergunta em um tempo, xi, em segundos. Ex: 1, 1, 2, 2, 2, 2, 3, 3, 4, 4. Ordenando temos: 1, 1, 2, 2, 2, 2, 3, 3, 4, 4

Amplitude	Variância	Desvio Padrão
<p>É a diferença entre o maior e o menor valor dos dados.</p> <p>Ex: $4 - 1 = 3$ é a amplitude do tempo de resposta.</p>	<p>É um número que nos diz o quão distante da média os dados estão.</p> <p>Ex: Apx 1</p> $V^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + (X_3 - \bar{X})^2 + \dots + (X_n - \bar{X})^2}{n}$	<p>Também nos diz o quão distantes da média estão os dados. É a raiz quadrada da variância (preserva a unidade de medida original)</p> <p>Ex: Apx 1.</p>

Variância Populacional vs Amostral

- ▷ Em muitos casos podemos encontrar medidas de dispersão "Populacionais" ou "Amostrais".
- ▷ As medidas Populacionais podem ser usadas quando estamos analisando dados de uma população com "sem margem de erro" ;
- ▷ Já as medidas Amostrais são as mais frequentemente utilizadas no dia a dia. Vamos utilizar elas no caso de uma amostra de dados para analisar.
- ▷ Mas qual a diferença de fórmula delas?
 - A diferença das fórmulas está no denominador. No caso amostral dividimos por $N-1$, para aplicar uma correção e no caso populacional por N .

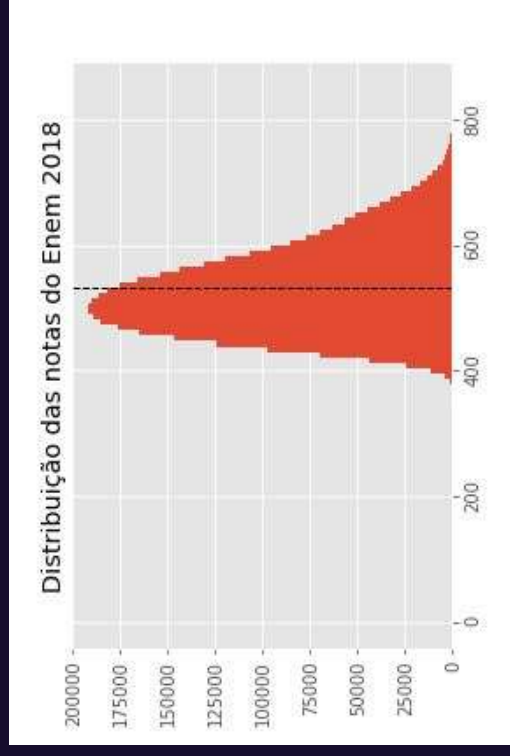
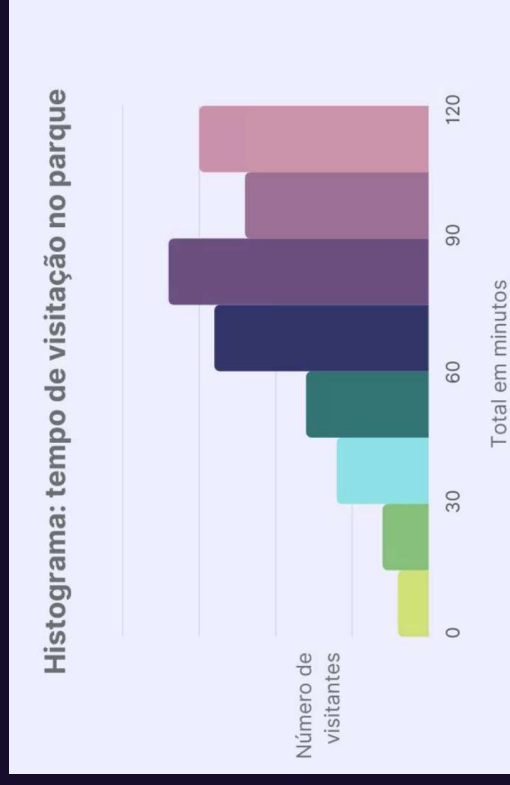
Amplitude, Variância e Desvio Padrão: Interpretando o Histograma

- ▷ O Histograma é um dos gráficos mais utilizados para se analisar a distribuição e a dispersão de variáveis numéricas. Ele é composto por:

No eixo x: intervalos da distribuição dos dados, o número de intervalos é chamado de bins

No eixo y: a contagem dos dados naquele intervalo.

Exemplos:

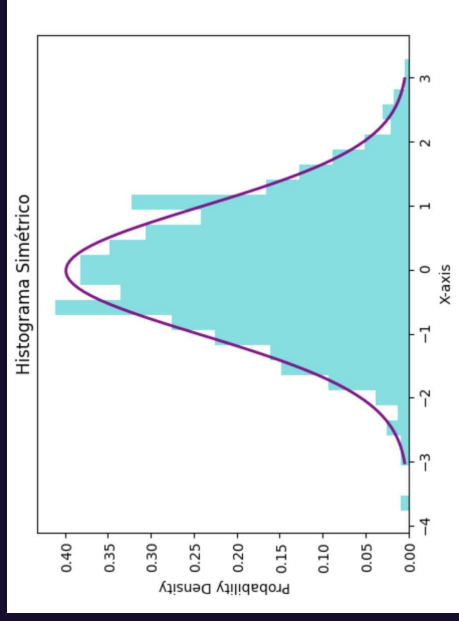


Amplitude, Variância e Desvio Padrão: Interpretando o Histograma

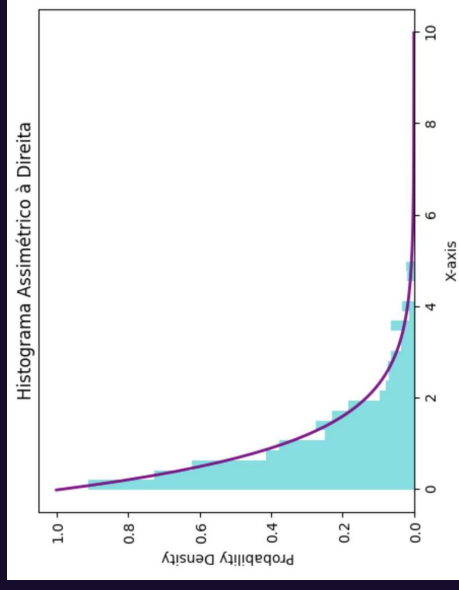
A Análise do histograma nos permite verificar como os dados se distribuem. A **simetria**, a **centralidade** e a **amplitude** dos dados são características importantes nesse tipo de análise.

Exemplos:

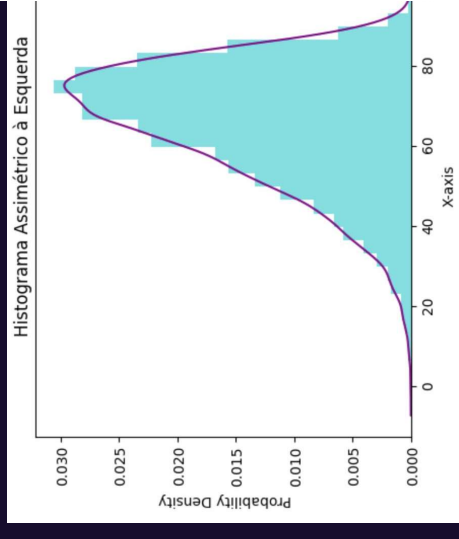
Simétrico



Assimétrico à direita (ou positivo)



Assimétrico à esquerda (ou negativo)

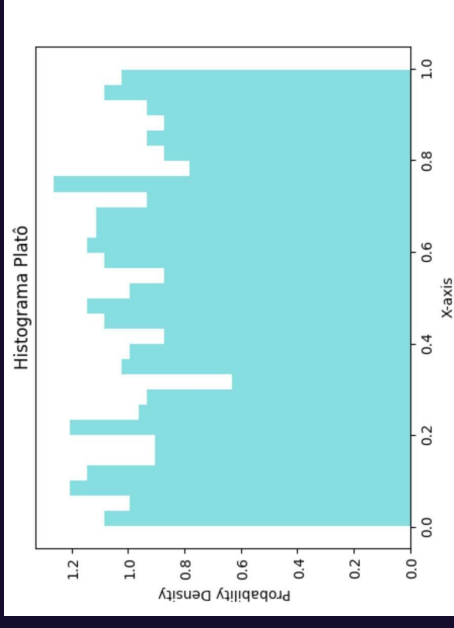


Amplitude, Variância e Desvio Padrão: Interpretando o Histograma

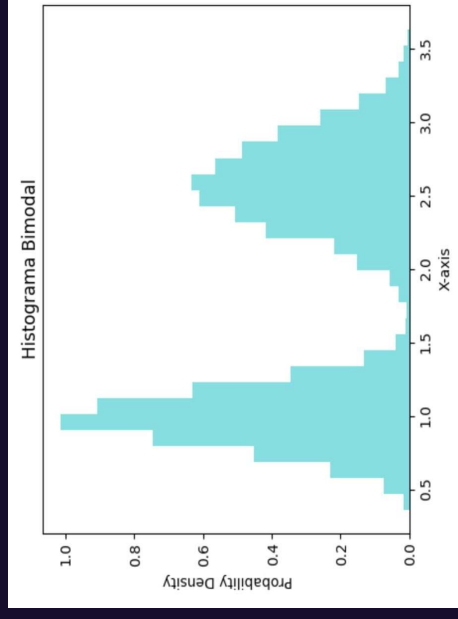
A Análise do histograma nos permite verificar a **simetria**, a **centralidade** e a **amplitude** dos dados.

Exemplos:

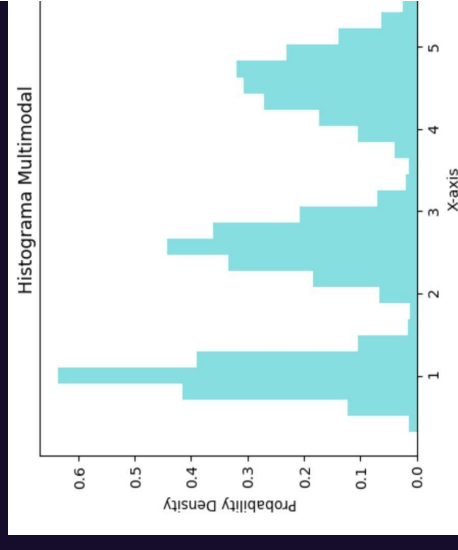
Platô



Bimodal (2 picos)

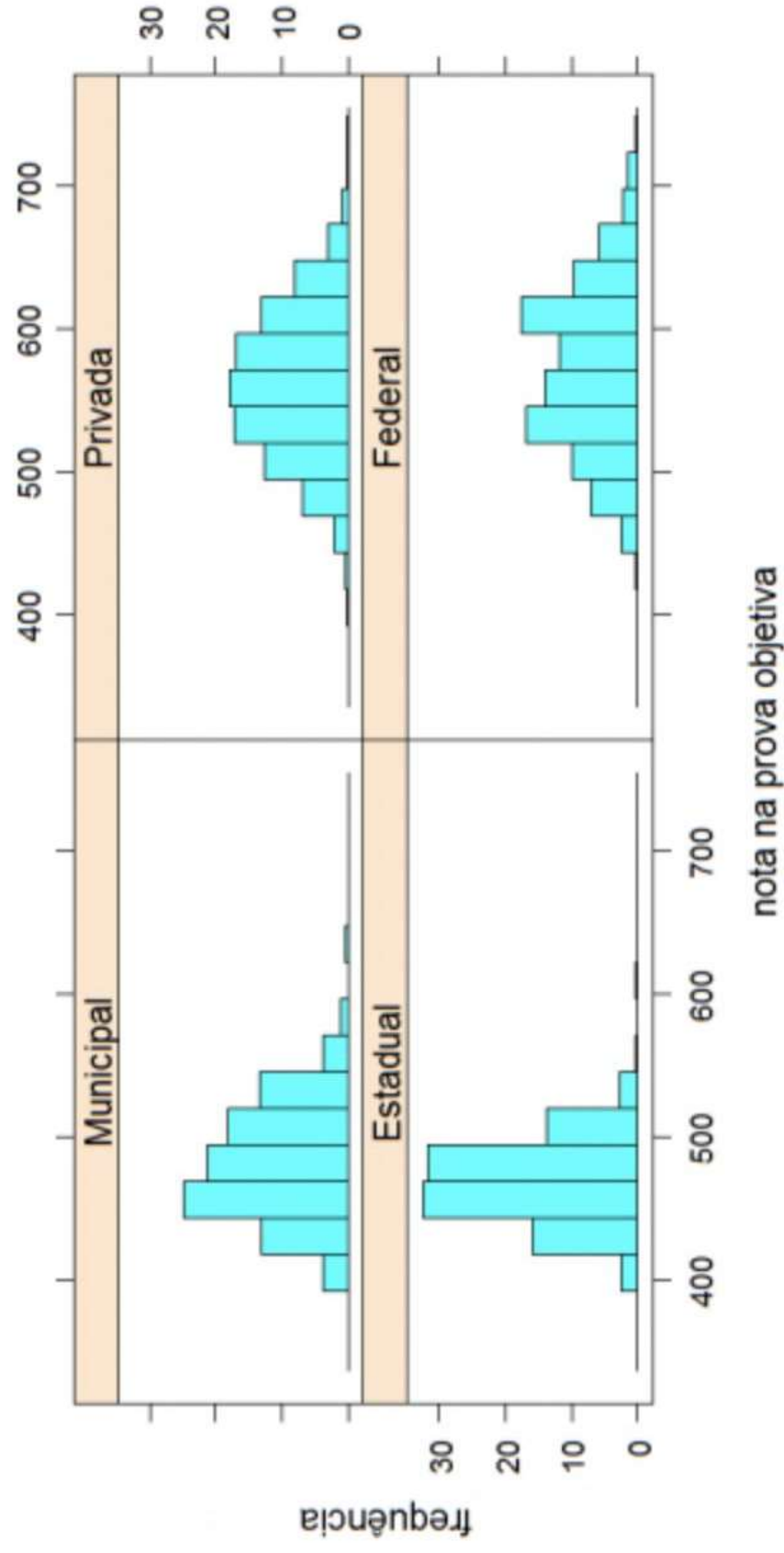


Multimodal (+ de 2 picos)



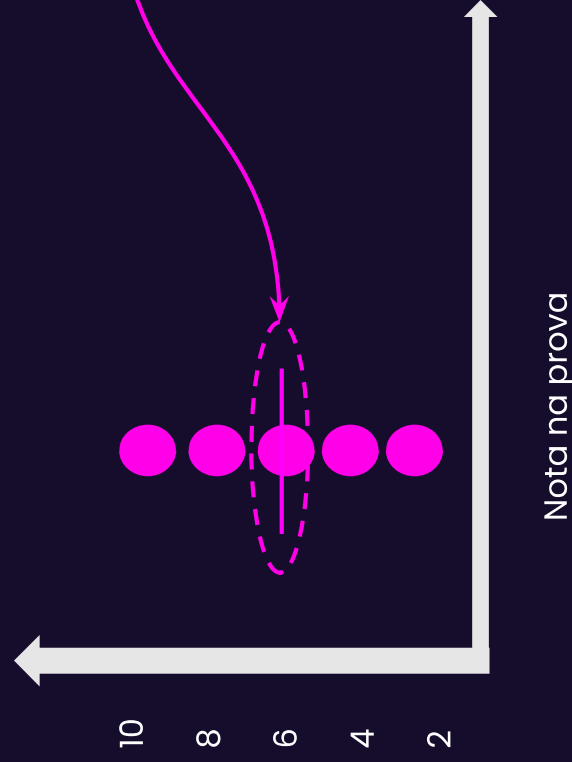
Exemplos do ENEM

Distribuição das notas das escolas por tipo



Medidas Separatrizes

- Medidas separatrizes na estatística são métricas que dividem ou "separam" os dados em partes iguais. Suponha que estamos olhando os dados da Nota de 5 alunos em uma prova::



A mediana é uma medida separatriz , pois ela separa os dados em grupos com o mesmo numero de dados (2 dados cima e 2 abaixo)

Para encontrá-la podemos usar a seguinte fórmula: $X[(5+1)/2] = X[3]$

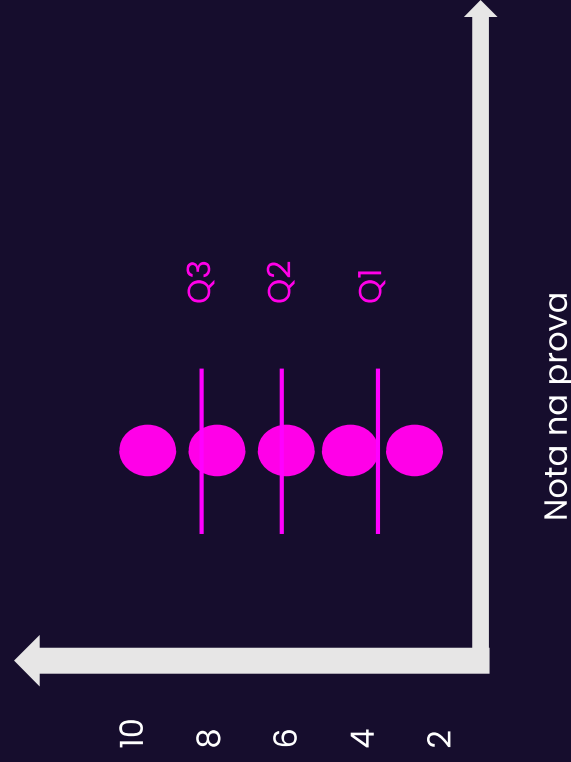
$$\begin{cases} X\left[\frac{n+1}{2}\right] & \text{Se } n \text{ é ímpar} \\ \frac{X\left[\frac{n}{2}\right] + X\left[\frac{n}{2} + 1\right]}{2} & \text{Se } n \text{ é Par} \end{cases}$$

As vezes podemos chamar a mediana de P50 . pois divide os dados em 50%

Medidas Separatrizes

- Medidas separatrizes na estatística são métricas que dividem ou "separam" os dados em partes iguais.

Suponha que estamos olhando os dados da Nota de 5 alunos em uma prova::



Agora suponha que queremos encontrar outras medidas separatrizes separam os dados em 25% , 75% , 50% (mediana) e 100%. Ou seja em grupos de mesmo tamanho

Vamos chamar essas medidas de Quartis. Pois separam os dados em partes.

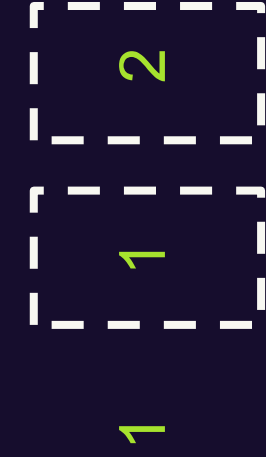
O Q1, ou primeiro quartil será o valor da distribuição em 4 grupos de tamanho igual de modo que 25% dos dados estão abaixo dele.

O Q3,, ou terceiro quartil será o valor da distribuição em 4 grupos de tamanho igual de modo que 75% dos dados estão abaixo dele.

Podemos também dividir a distribuição em 100 partes iguais e obter uma medida chamada Percentil. A mediana será o percentil 50 ou 50º será o percentil 25 ou 25º e Q3 será o percentil 75 ou 75º

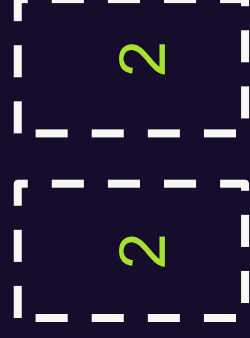
Medidas Separatrizes

No exemplo das candidatas:

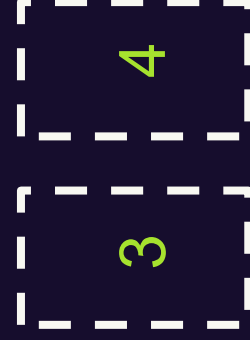


$$Q1 = P25 = X[11/4] = X[2.75]$$

Nesse caso como o termo 2 é 1 e o termo 3 é 2 teremos que $Q1 = 1 + 0.75 * [2-1] = 1.75$



$$Q2 = \frac{X[10/2] + X[11/2]}{2} = 2$$



$$Q3 = P75 = X[33/4] = X[8.25]$$

Nesse caso como o termo 8 é 3 e o termo 9 é 4 teremos que $Q3 = 3 + 0.25 * (4-3) = 3.25$

**Fórmula
para obter
os quartis**

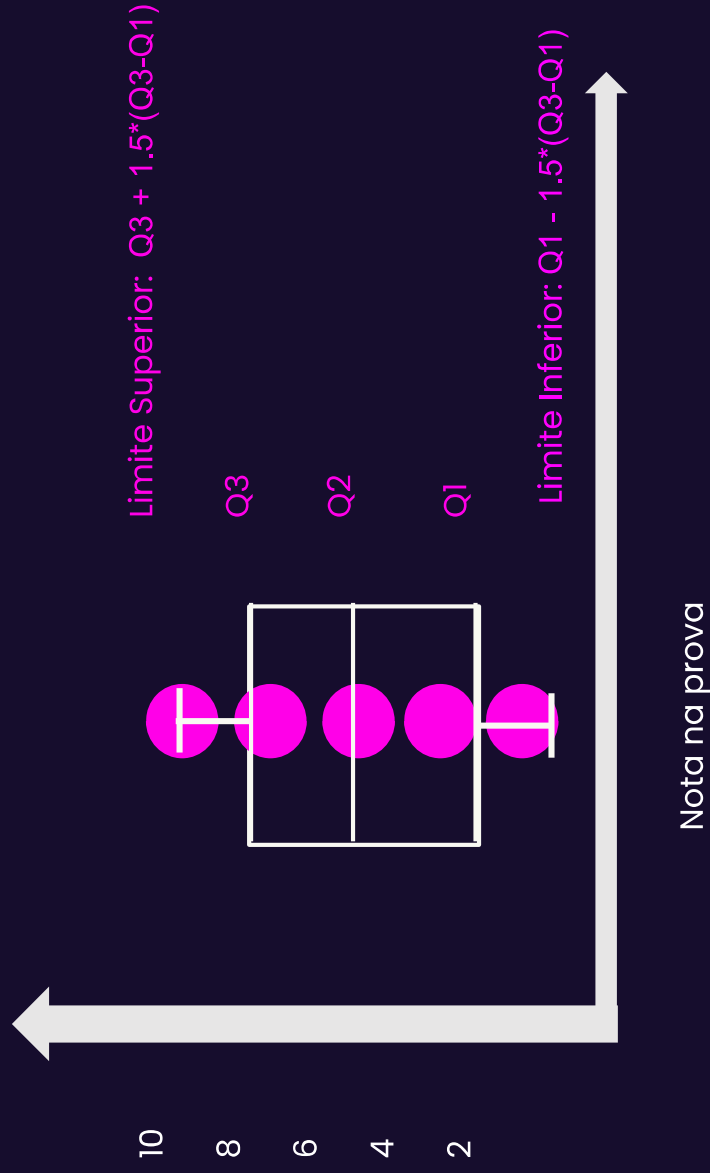
$$Q1 = \frac{1}{4} (n + 1)^{\text{th}} \text{ Termo}$$

$$Q3 = \frac{3}{4} (n + 1)^{\text{th}} \text{ Termo}$$

Boxplot

O Boxplot é um gráfico de caixa em que traçamos visualmente algumas das medidas separatrizes mais importantes de um conjunto de dados.

- Nele vamos ter uma caixa entre Q1 e Q3 e que nos mostra limites superiores

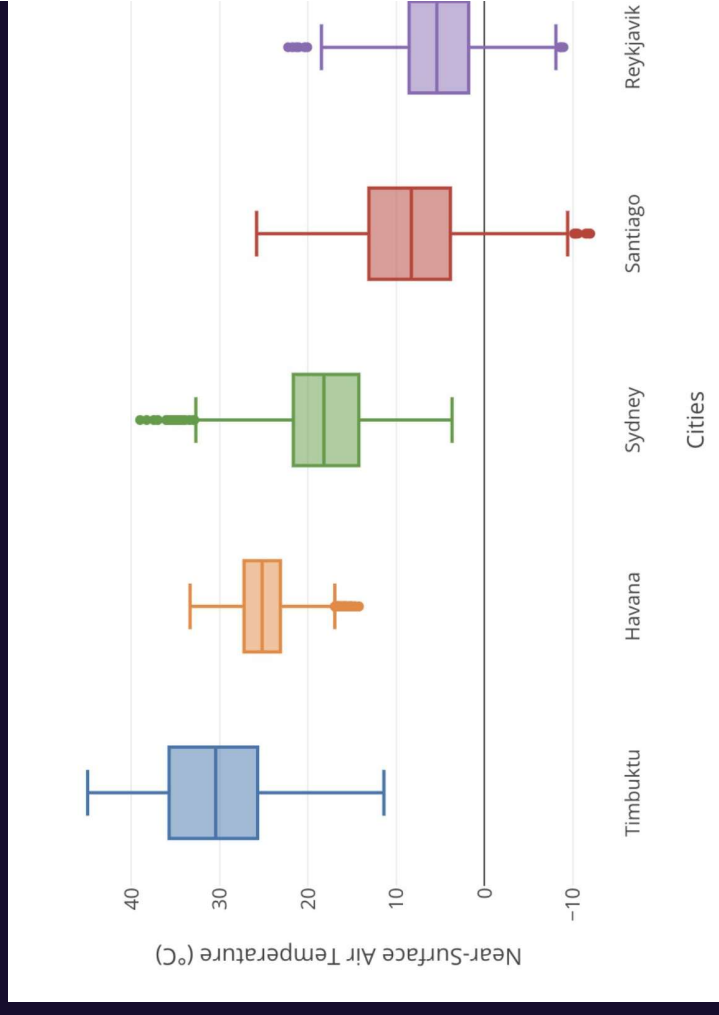


Interpretando o boxplot: Temperatura

O gráfico da direita nos mostra os gráficos de caixa da variável temperatura em diferentes cidades do mundo em um período

Interpretacao: Gráfico de caixa da temperatura

- Pela posição e tamanho das caixas, percebemos que Timbuktu, na África possui temperaturas concentradas em valores mais altos. Q1 a Q3 de 25 a 35 graus, enquanto que Reykjavik, na Islandia, possui temperaturas concentradas em valores mais baixos (2 a 9 graus)
- Havana é a cidade em que as temperaturas têm menor oscilação. (concentradas de 22 a 27 graus), vs Santiago é a que possui maior oscilação
- Sydney é a cidade com mais outliers, com dias com temperaturas atípicas bem elevadas, prox a 40 graus.



Outliers

Vimos que Outliers são valores atípicos ou anômalos dos dados. Mas como podemos definir se um determinado dado ou não? Como podemos identificar outliers em amostras ou bases de dados?

1. Métodos de definição de Outliers:

Método Z-score

O método z-score utiliza como referencia a quantos desvios padrões a informação está da média:

$$Z\text{-score} = (x - \text{média})/\text{std}$$

Se $Z\text{-score} > 3$ ou $Z\text{-score} < -3$ desvios padrões geralmente se pode considerar um outlier.

Método IQR

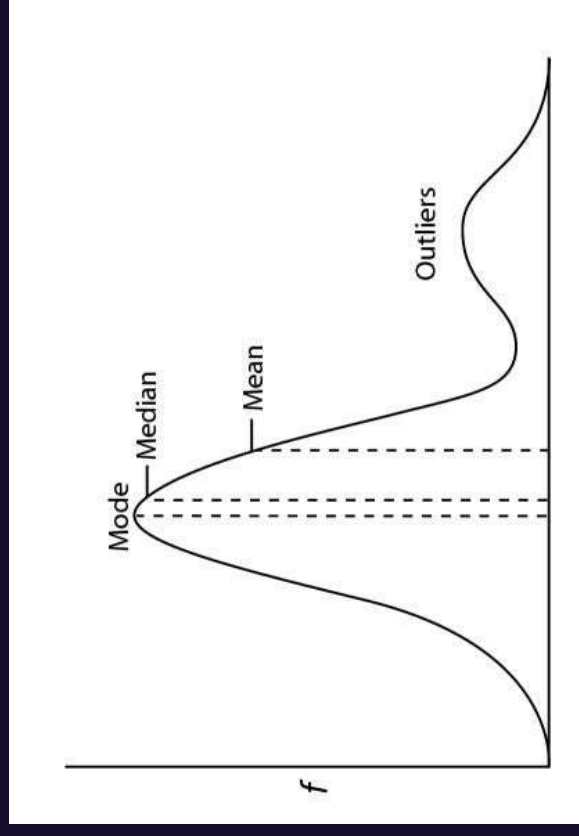
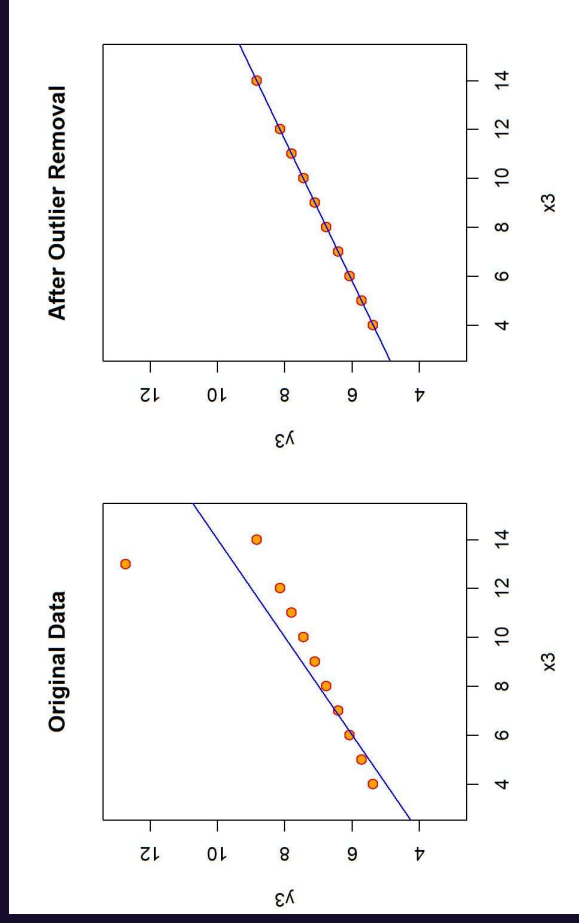
O método IQR (distância inter-quartil em inglês) vê a distância da amplitude dos quartis, também conhecido por método de Tukey:

$$IQR = Q3 - Q1$$

Se $x > 1.5 * IQR$ ou $x < -1.5 * IQR$ se pode considerar um outlier.

A importância dos Outliers

- Outliers são importantes pois a existência deles já nos gera uma informação. Eles podem significar uma nova categoria padrão nos dados.
 - Mas a presença de outliers pode gerar ruído em modelos e análises estatísticas. Isso acontece porque métricas que e distância podem ser sensíveis a outliers. Ex: Modelo de regressão abaixo (Imagem 1)
 - A média é uma das métricas mais sensíveis aos outliers, portanto devemos preferir utilizar a mediana no caso de outliers
- 2)



Tratamento dos Outliers

- Para evitar o viés gerado por outliers em modelos e análises estatísticas podemos utilizar técnicas de tratamento de outliers. As mais conhecidas são:
 1. Caso o outlier não represente um sentido de negócio (ex: fraude) temos as seguintes possibilidades:
 - a. Remover os dados outliers da análise
 - b. Substituir o valor dos outliers pela média ou mediana
 2. Caso o outlier tenha um sentido de negócio:
 - a. caso categórico: criar uma categoria nova para o outlier
 - b. caso numérico: inputar um valor para outliers, ex: -999, -1, etc para dados positivos.

Estatística : Frequências e Medidas

Análise de Variáveis Numéricas x Categóricas



Análise Exploratória de Variáveis Numéricas

Quando vamos analisar dados é importante começarmos por uma análise exploratória dos dados pensando nas frequências e medidas dos dados.

Para as variáveis numéricas podemos analisar as seguintes características:

1. Média, mediana, moda
2. Variância e desvio padrão
3. Valores mínimos e máximos
4. Análise do histograma
5. Análise do boxplot
6. Verificar outliers
7. Verificar valores nulos ou faltantes

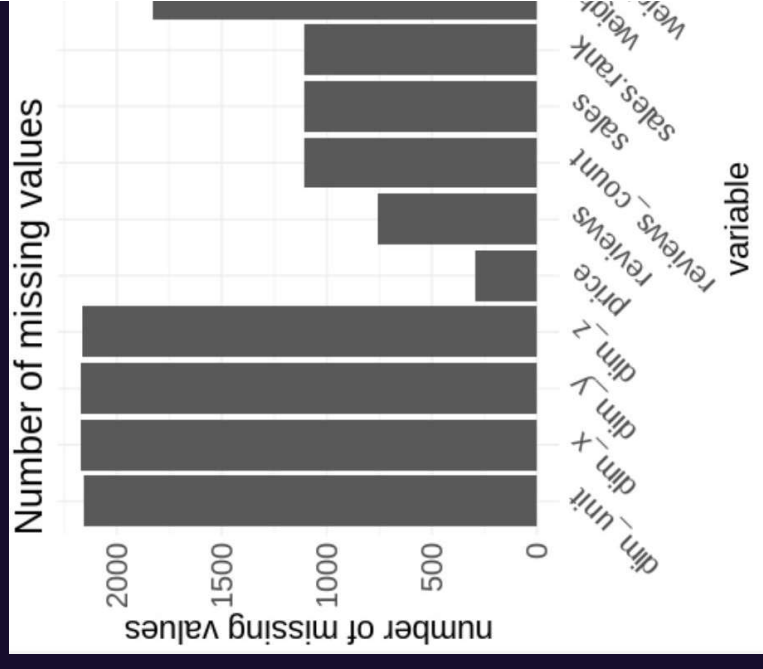
Valores nulos e variáveis numéricas

A análise dos valores nulos é uma etapa importante na análise exploratória dos dados. Em muitos casos no dia a dia podemos ter variáveis importantes e pouco preenchidas. Nesse caso quando devemos substituir os valores nulos? e por quais valores substituir.

A substituição vai a critério do analista e da análise.

Mas uma boa prática é:

- Verificar se a informação faltante, assim como o outlier, possui sentido de negócio.
- Substituir pela média, mediana ou moda.
- Fazer um gráfico de barras de quantidade ou % de valores faltantes para cada variável.



Análise Exploratória de Variáveis Categóricas

Para as variáveis categóricas podemos analisar as seguintes características:

1. Moda dos dados
2. Análise da frequência das categorias ou cardinalidade:
 - a. Quantas categorias existem?
 - b. Qual a frequência de cada uma das categorias (gráfico de barras)?
3. Agrupamento de variáveis de alta cardinalidade.
4. Substituição de valores nulos: ou pela Moda ou criar a categoria "Nulo"

Estatística : Frequências e Medidas

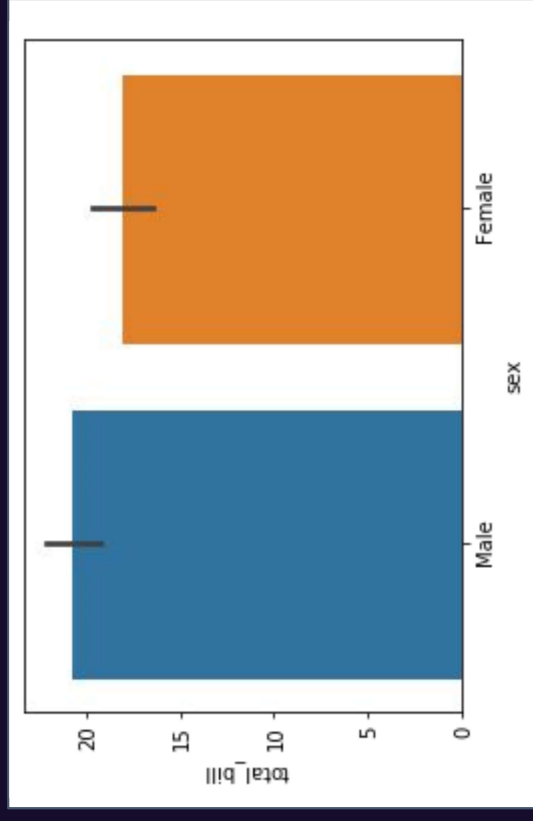
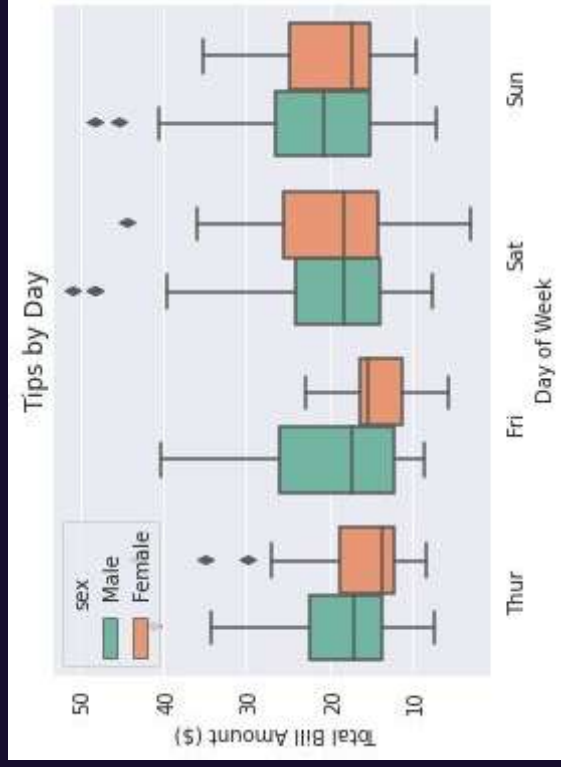
Análises cruzadas



Análise Exploratória Variáveis numéricas x categóricas

Em muitas análises estamos interessadas em entender padrões de variáveis numéricas e categóricas combinadas, por exemplo:

- Análises de distribuição de receita , vendas, gorjetas por : Dia da semana e por Gênero



Estatística : Frequências e Medidas

Vamos Praticar!

