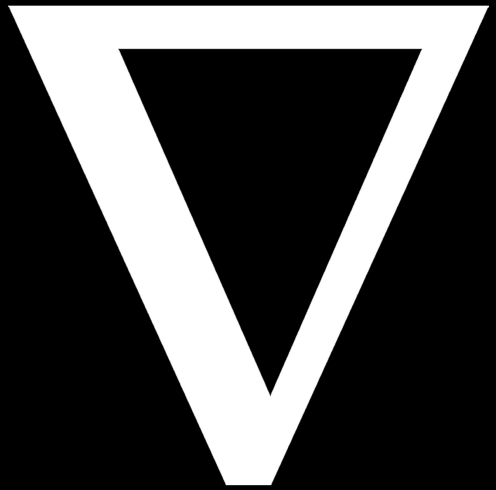


Análise Exploratória de Dados (EDA)

João Pedro (Dora) Mattos • 24/02/2021
@joaopedromattos

Motivação

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

"Existem dados nulos na coluna?"

"Existem outliers nessa feature?"

"Quantas strings diferentes existem na coluna?"

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

"Essa feature tem poder discriminativo?"

"Essa feature é correlacionada com alguma outra?"

Análise Exploratória de Dados


Aquisição
Pré-processamento / Limpeza
Análise Exploratória de Dados
Treinamento
Avaliação

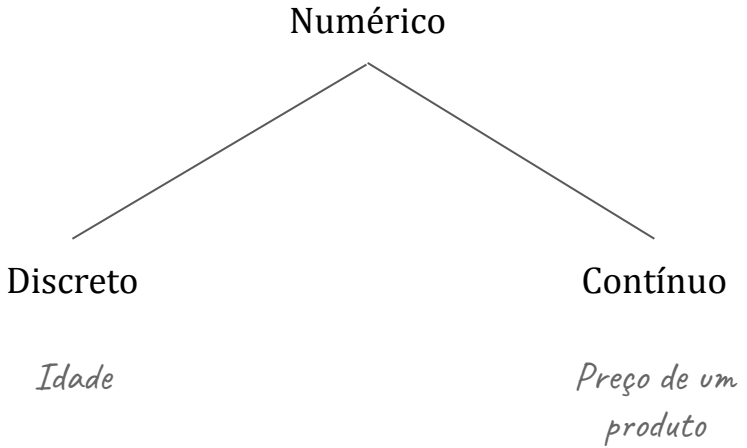
AVISO

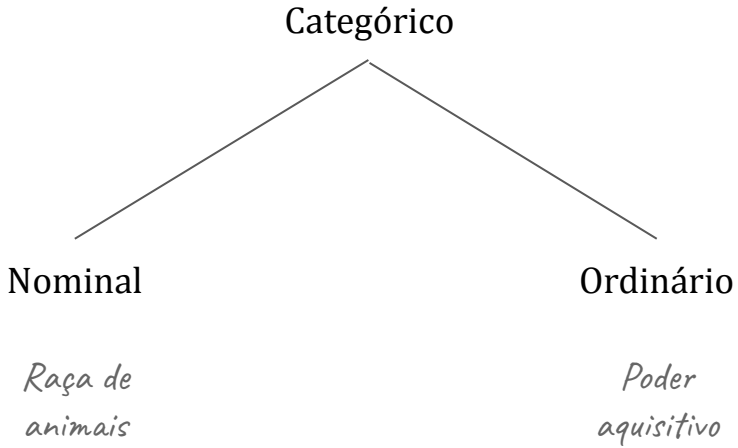
Você não vai aprender a fazer um **bom** EDA nessa aula.

Ainda nesta aula...

- Tipos de dados
- Introdução à estatística descritiva
- Medidas de Correlação
- Boas práticas

Tipos de dados





dúvidas?

Estatística descritiva

Medidas de Centralidade

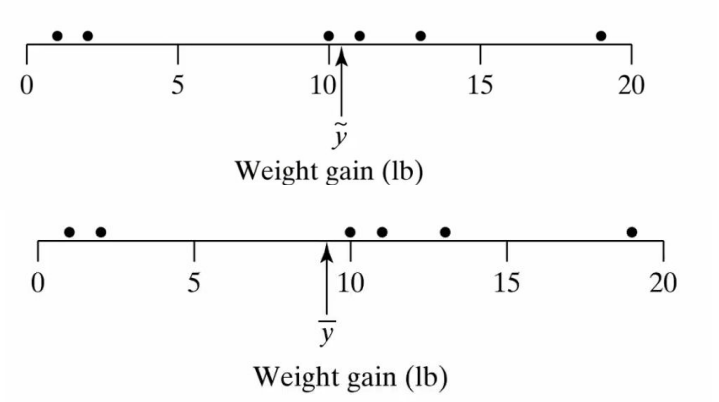
- Mediana
- Média
- Moda

Medidas de Dispersão

- IQR - Intervalo
- Variância / Desvio Padrão

Medidas de Centralidade

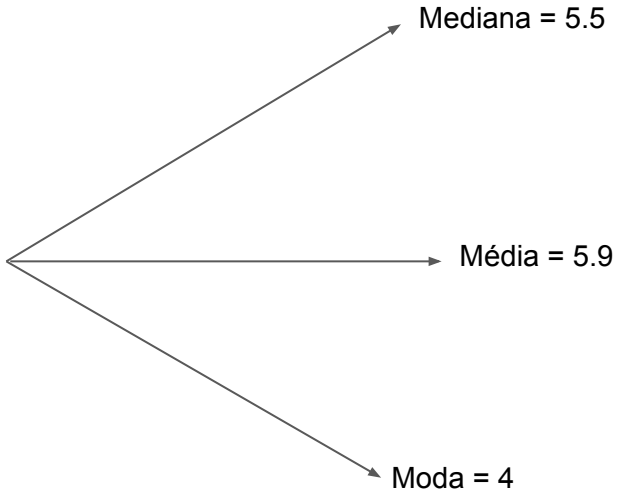
- Mediana
  - Centro do conjunto de dados ordenado
  - Por definição divide o conjunto de dados ao meio
- Média
  - Soma-se todos os elementos do conjunto e divide-se pelo número de elementos.
  - Sensível a outliers e distribuições assimétricas

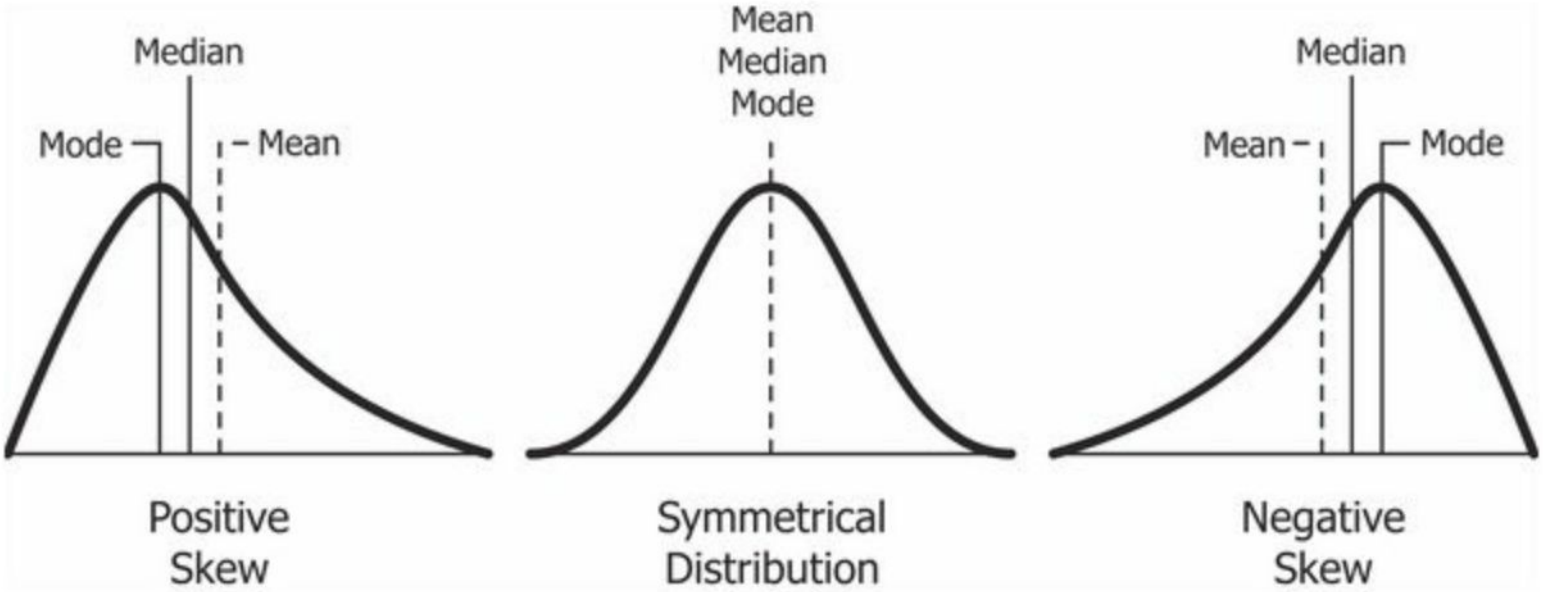


Medidas de Centralidade

- Moda
  - Elemento mais frequente do conjunto de dados
  - Geralmente utilizado para dados categóricos!

[4, 1, 9, 6, 2, 4, 8, 5, 9, 11]



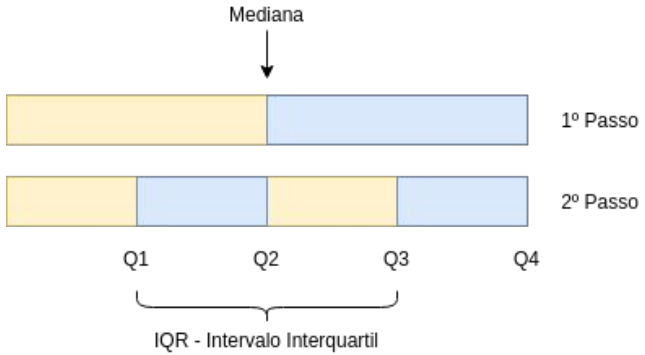


https://en.wikipedia.org/wiki/File:Relationship_between_mean_and_median_under_different_skewness.png

Medidas de Dispersão

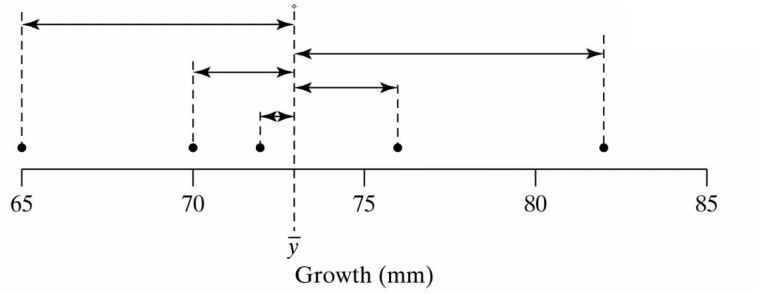
- Intervalo Interquartil (IQR)

  IQR = Q3 - Q1



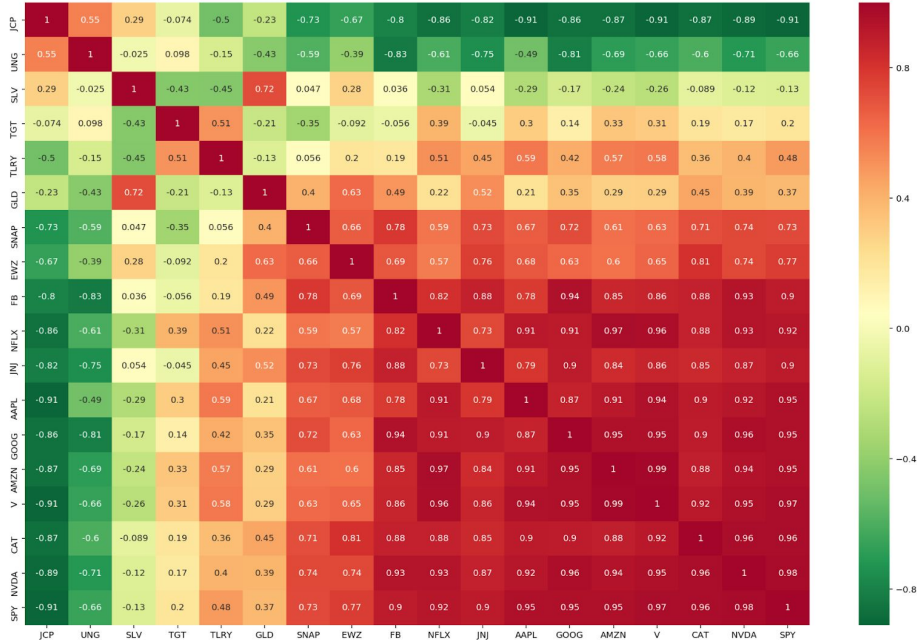
- Desvio padrão
  - Variância = σ^2

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$



dúvidas?

Medidas de Correlação



Medidas de Correlação

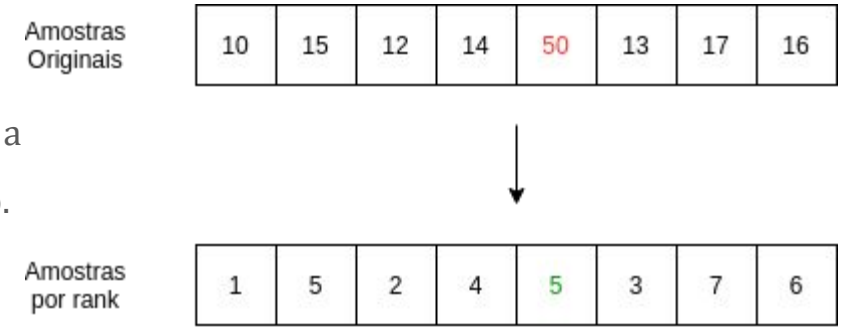
- Coeficiente de Pearson

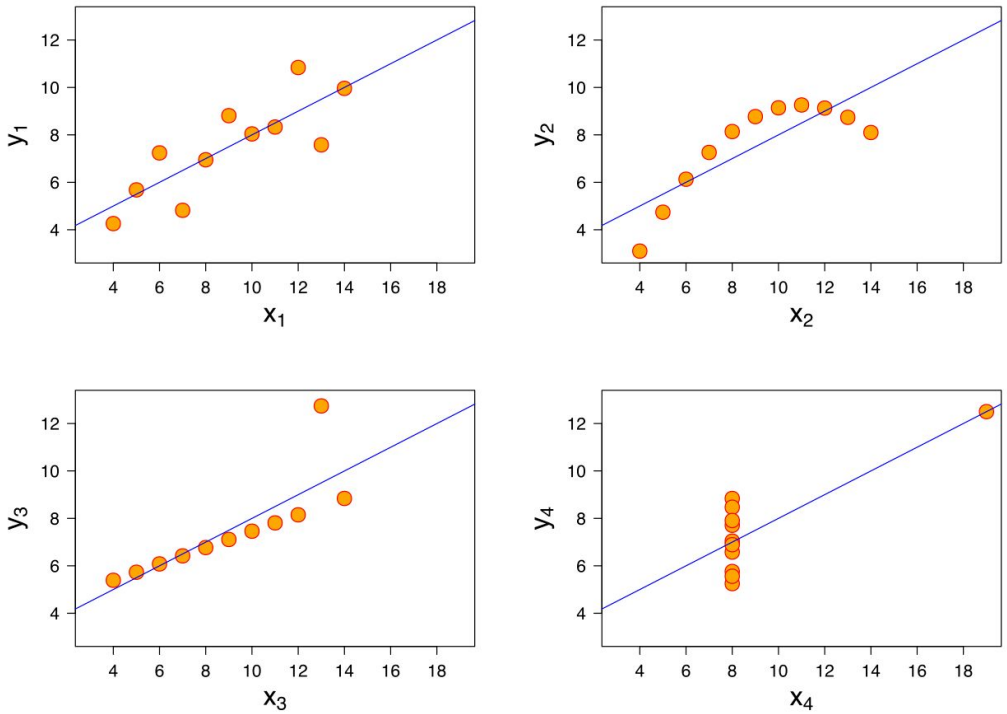
$$\frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

"Como variam em conjunto"

"Produto dos desvios independentes"

- Coeficiente de Spearman
  - Cálculo do coeficiente de Pearson a partir do "rank" de cada elemento.
    Veja exemplo →

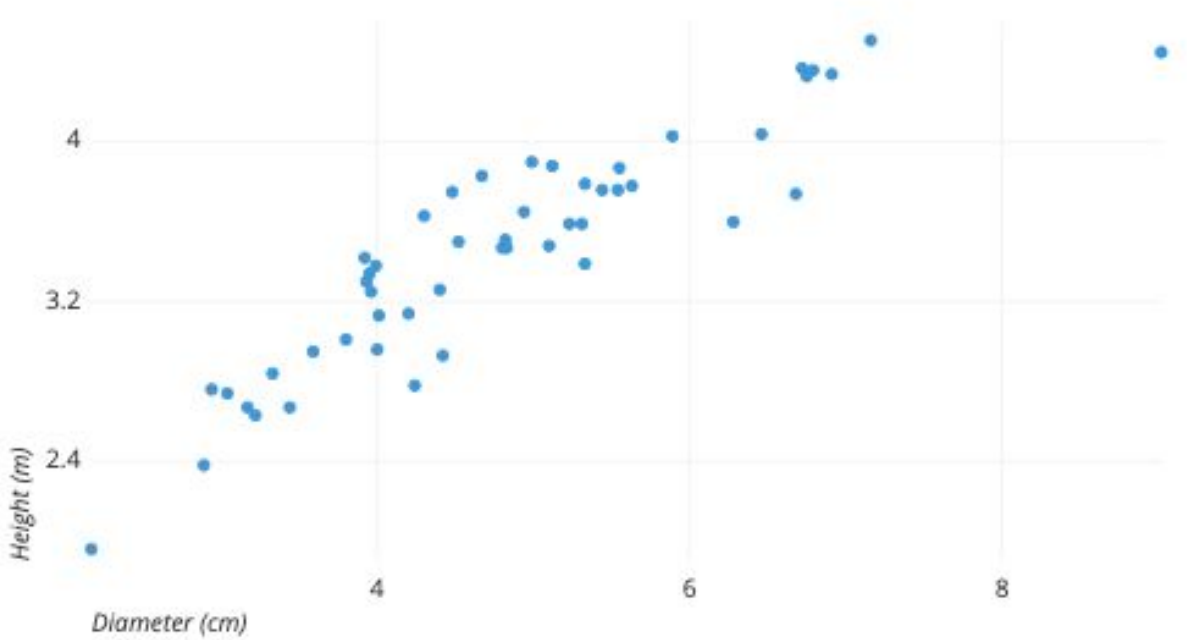



y1
x1
y2
x2
y3
x3
y4
x4

dúvidas?

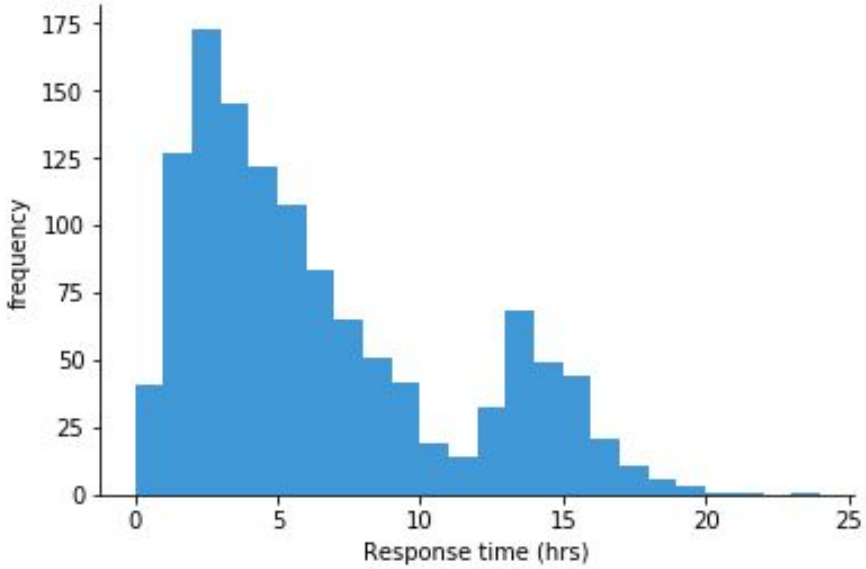
Visualização (Data Visualization)

- Scatter Plot



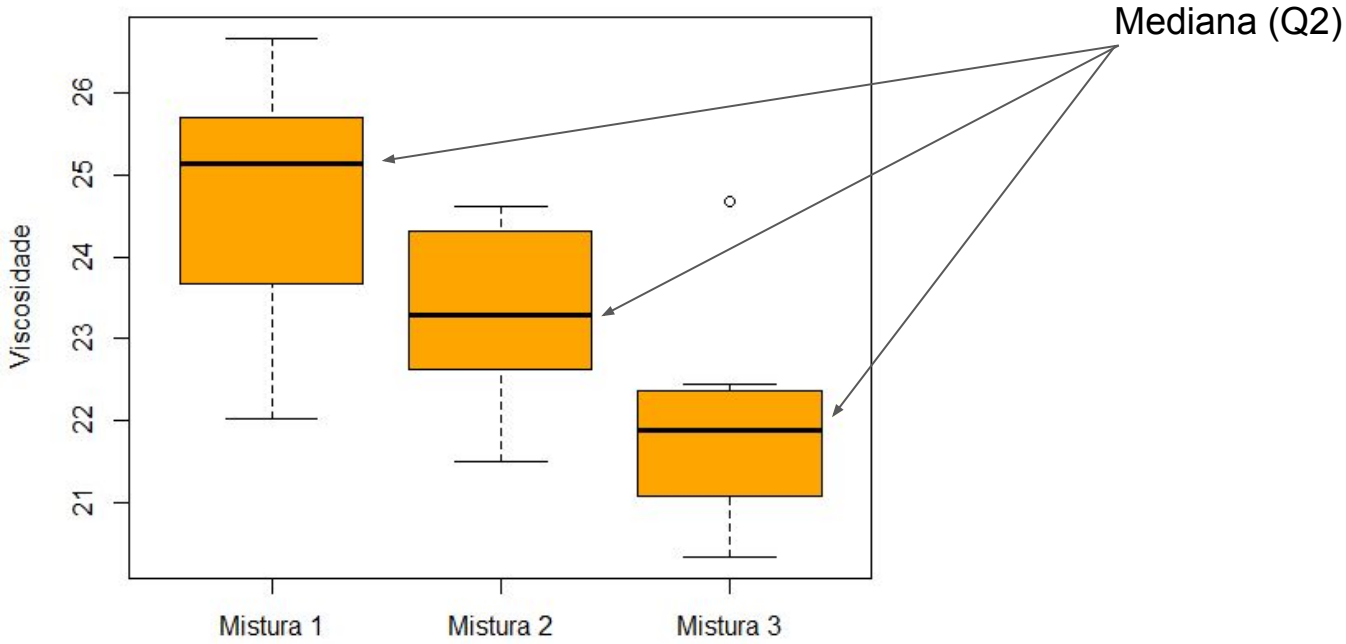
Visualização (Data Visualization)

- Gráfico de frequência



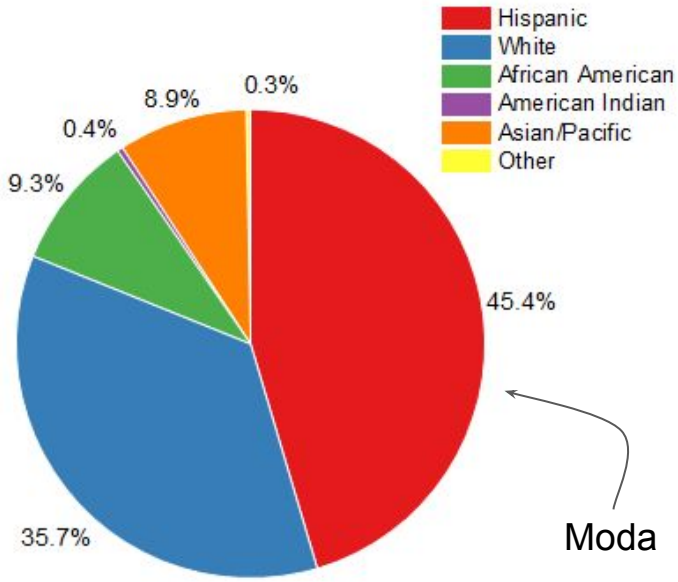
Visualização (Data Visualization)

- Box-plot



Visualização (Data Visualization)

- Pie-Chart (Gráfico por setores)



dúvidas?

Boas práticas

- Sempre procurar por dados nulos
- Checar se existem variáveis altamente correlacionadas
- Pesquisar sobre o domínio do problema durante todo o processo de EDA
- Usar os gráficos adequados para seu tipo de dado
- No caso de classificação, checar o balanceamento de classes
- **Questionar se os resultados fazem sentido!**

Material Complementar - (Altamente recomendado)

- Aula 03 “Summary Statistics” - Curso da Universidade de Tübingen
- Aula 3 e Aula 4 do Curso de Ciência de Dados do professor Francisco Rodrigues