DATA SCIENCE

AULA 2 - Análise Exploratória

Prof. Gabriel Resende Machado

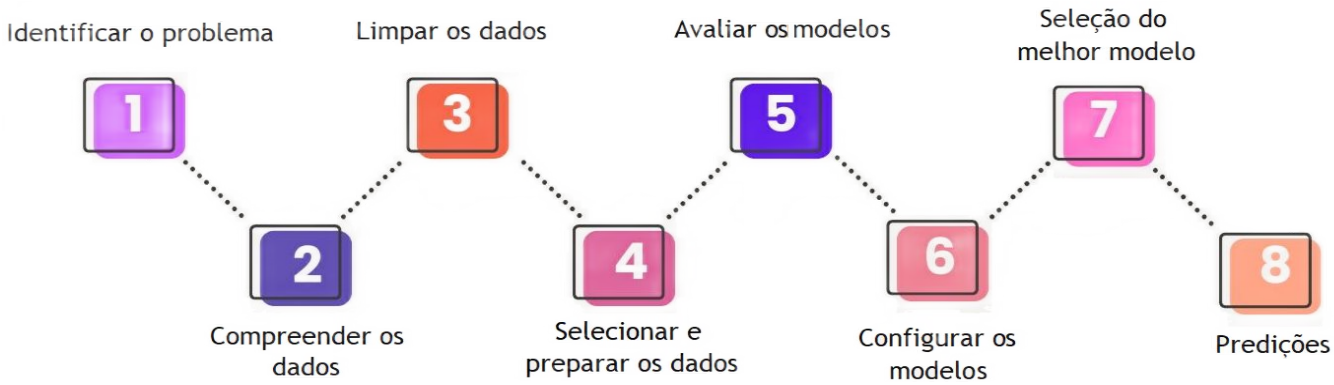
gabrielmachado@unifeso.edu.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science

O que é a Exploração dos Dados (EDA)?

- Etapa **crucial** que antecede qualquer tarefa que envolva Ciência de Dados;
- Ajuda na **compreensão dos dados**, seu **contexto**, **principais atributos** e suas **características**;
- Contribui na **detecção de padrões e anomalias**;
- Auxilia na **seleção de atributos** e na **geração de modelos preditivos**.



Adaptado de https://shorturl.at/iJPT6.

Ferramentas para EDA

- **Estatística Descritiva:**
  - **Medidas de tendência central:** média, mediana e moda;
  - **Medidas de posição:** quantis, quartis e percentis.
  - **Medidas de dispersão:** variância, desvio padrão, IQR;
  - **Medidas de distribuição:** momento;
  - **Medidas de relação:** covariância e correlação.
- **Visualização dos atributos**:
  - Histogramas;
  - *Boxplots;*
  - *Kernel Density Estimation* (KDE).
  - *Scatter plots;*

Estatística Descritiva - Tendência Central

- Definem **pontos centrais** de referência em análise de **dados univariados** (apenas um atributo);
- Define-se um conjuntos de valores para um atributo $x^j = \{x_1, x_2, \cdots, x_n\}$;
  - **Média:** ponto central calculado a partir da equação $\mu_{x^j} = \frac{1}{n}\sum_{i=1}^{n} x_i$;
    - A média pode ser representada também por $\bar{x}$ ou $E(x)$;
    - É suscetível a valores extremos (*outliers*);
  - **Mediana:** alternativa à média, por ser uma estatística mais robusta a *outliers*;
    - A partir de um conjunto ordenado de forma crescente, calcula-se:

$$med(x^j) = \begin{cases} \frac{1}{2}(x_r + x_{r+1}) & \text{se n for par } (n = 2r) \\ x_{r+1} & \text{se n for impar } (n = 2r+1) \end{cases}$$

  - **Moda:** mais voltada a valores categóricos, calcula o valor com maior ocorrência.

Estatística Descritiva - Tendência Central

- Variações da média para cálculo de tendências centrais:
  - **Média ponderada:** calculada a partir de pesos atribuídos aos valores de w^j a partir de x^j :

$$\mu_{x_j,w_j} = \sum_{i=1}^{j} \frac{w_i x_i}{w_i}$$

  - **Média harmônica:** ideal para cálculo de medidas inversamente proporcionais (*e.g.* velocidade/tempo, *F1-score*).

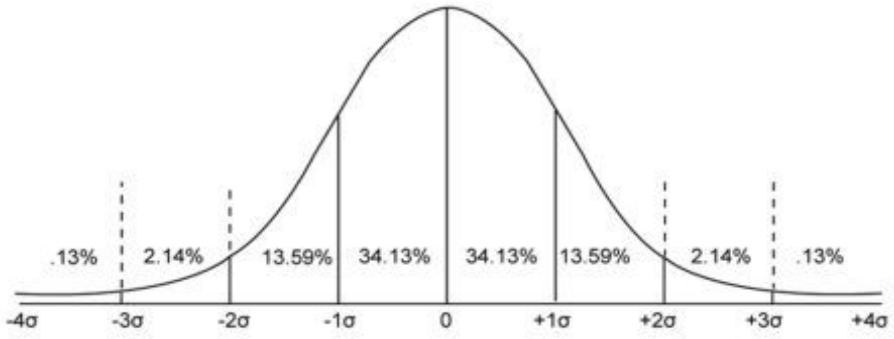
$$H_{x_j} = \sum_{i=1}^{n} \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}} = \frac{1}{\sum_{i=1}^{n} \frac{1}{i}}$$

Estatística Descritiva - Posição

- Se caracterizam pelas estatísticas de **quantis**, **quartis** e **percentis** em sequências de dados **ordenados de forma crescente**;
- **Quantil** é um termo mais geral utilizado para descrever divisões de dados em ***n*** partes iguais;
- **Quartis** repartem o conjunto de dados em quatro partes de tamanhos aproximadamente iguais. *E.g.*, o valor *Q(1)* de uma sequência possui 25% dos valores abaixo dele;
- **Percentis** são semelhantes aos quartis, contudo dividem a sequência de dados em 100 partes, ao invés de apenas 4;
- A **mediana** pode ser calculada a partir do quartil *Q(2)* ou pelo percentil *P(50)*.

Estatística Descritiva - Dispersão I

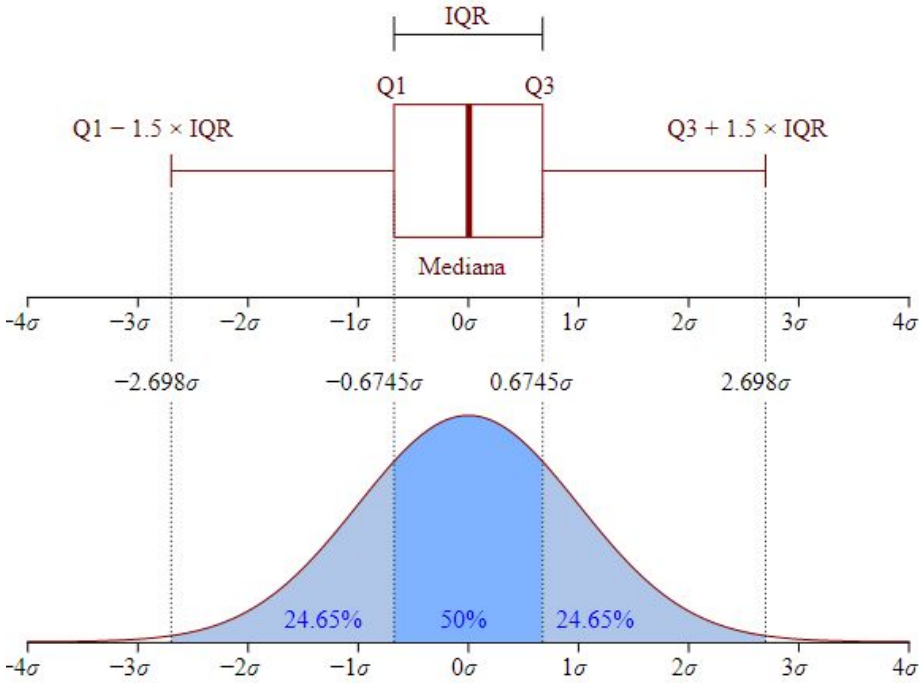
- Tem como objetivo calcular o **nível de dispersão** dos dados no entorno de uma medida de tendência central;
- As estatísticas mais utilizadas são a variância $S^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ e o desvio padrão $\sigma = \sqrt{S^2}$;
- O desvio padrão é comumente mais adotado por estar na mesma escala dos dados.



Função densidade de probabilidade (PDF) N(μ=0, σ=1).

Estatística Descritiva - Dispersão II

- **IQR (*Interquartile Range*):** mede a dispersão dos dados em torno da tendência central;
    - Basicamente, $IQR = Q(3) - Q(1)$.

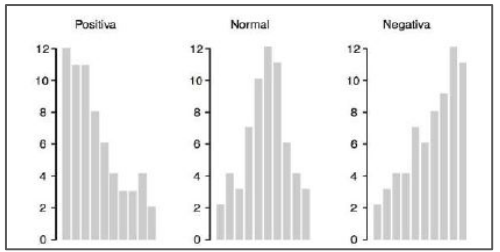


Estatística Descritiva - Distribuição

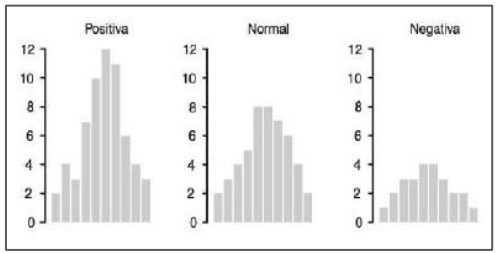
- As medidas que são definidas em torno da média de um conjunto de valores são, em sua maioria, oriundas de instâncias de uma estatística chamada **momento.**

$$momento_{k>2}(x^j) = \frac{\sum_{i=1}^{n}(x_i - \mu_{x^j})^k}{n * \sigma^k}$$

- Quando *k = 1*, considera-se a média como primeiro momento central;
- Quando *k = 2*, considera-se a variância como segundo momento central;
- Quando *k = 3*, **obliquidade** (***skewness***). Mede a simetria em torno da média:
  - obliquidade = 0: aproximadamente simétrica;
  - obliquidade > 0: a distribuição se concentra mais à esquerda;
  - obliquidade < 0: a distribuição se concentra mais à direita;
- Quando *k = 4*, **curtose** (***kurtosis***), representa o achatamento da PDF:
  - curtose = 0: achatamento aprox. de uma distribuição normal;
  - curtose > 0: menor achatamento em comparação com a distr. normal;
  - curtose < 0: maior achatamento em comparação com a distr. normal.



Diferentes obliquidades.



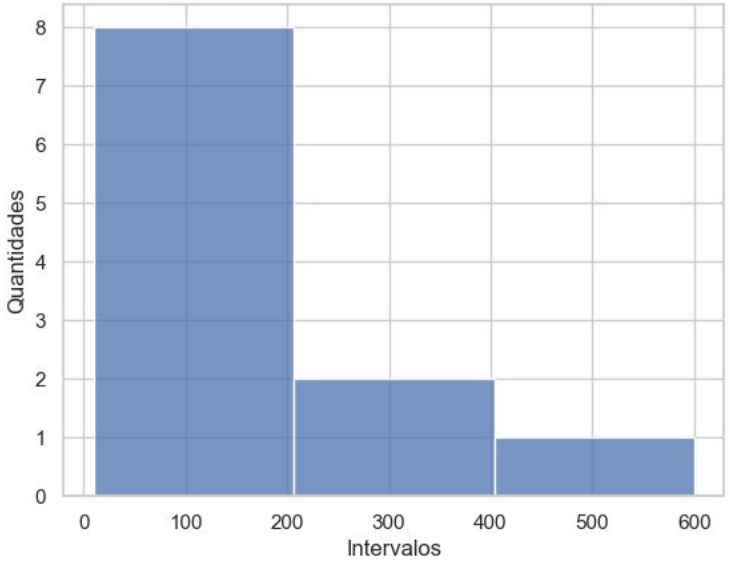
Diferentes curtoses.

Estatística Descritiva - Relação

- Tem como objetivo calcular o **nível de relacionamento** ou **dependência multivariada**;
- As estatísticas mais utilizadas são **covariância, correlação de Pearson e correlação de Spearman**;
  - A covariância $Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \mu_{x^j})(y_i - \mu_{y^j})}{n}$ mede o grau com que os atributos variam juntos, considerando uma relação linear. Contudo, a correlação de Pearson é utilizada com mais frequência por definir os valores possíveis em um intervalo de [-1, 1].
  - Dentre os valores que a correlação de Pearson $Corr(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$ pode assumir, há três valores relevantes:
    - ***Corr(x, y) = 0***: não há um relacionamento linear;
    - ***Corr(x, y) = 1***: presença de relacionamento linear positivo diretamente proporcional (*e.g.* peso e altura);
    - ***Corr(x, y) = -1***: presença de relacionamento linear negativo inversamente proporcional (*e.g.* oferta e demanda);

Estatística Descritiva - Visualização

- **Histogramas:** organizam os dados em grupos chamados ***bins*****,** como forma de visualização da distribuição dos dados;
- É importante frisar que um histograma **não é** um gráfico tradicional de colunas.
- Exemplo de histograma para a sequência {10, 20, 20, 30, 40, 50, 60, 100, 250, 400, 600}:



Estatística Descritiva - Visualização

- ***boxplots***: apresentam uma outra visão da distribuição dos dados, desta vez utilizando as estatísticas referentes aos quartis *Q(1), Q(2)* (mediana) e *Q(3)*.
- ***outliers***: valores extremos representados por pontos, definidos a partir das fórmulas definidas abaixo. Existem variações mais "conservadoras" de *boxplots* para a definição de *outliers*.

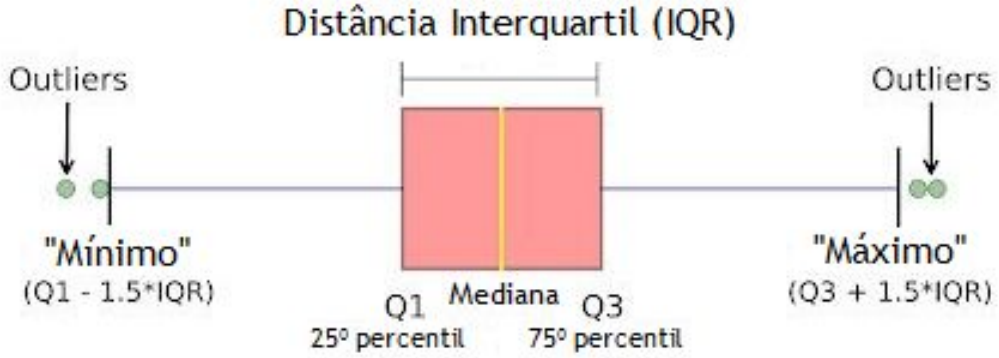
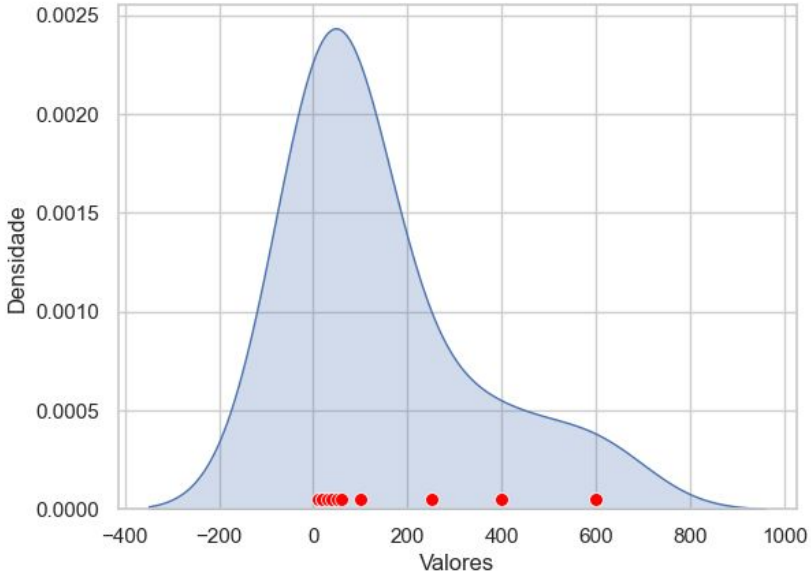


Imagem adaptada de https://shorturl.at/amyIK.

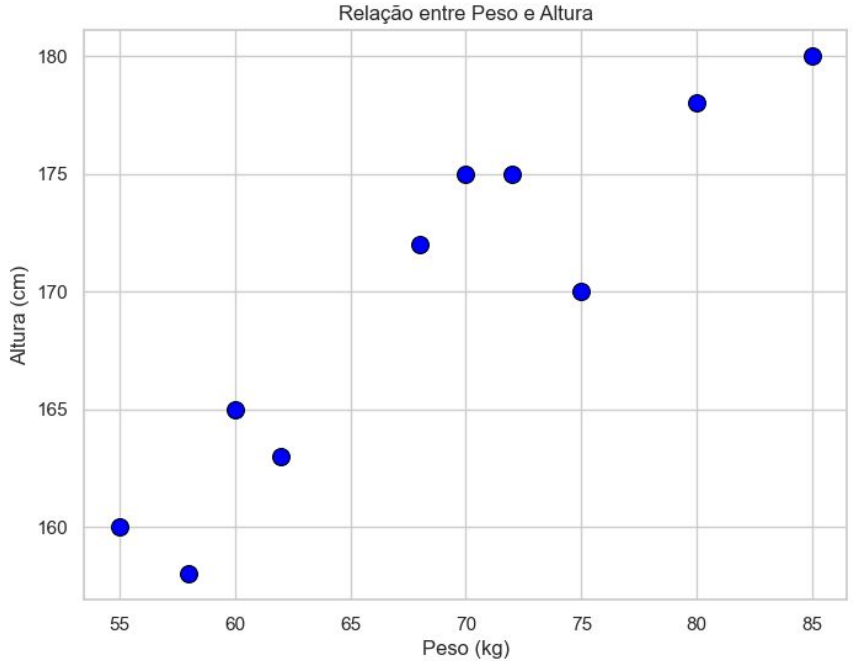
Estatística Descritiva - Visualização

- ***KDE* (*Kernel Density Estimation*):** mostram a densidade dos pontos de um conjunto de dados;
  - Essas visualizações ajudam a estimar aproximadamente se os dados seguem alguma distribuição estatística a partir da concentração dos seus pontos.
  - Exemplo de KDE para a sequência {10, 20, 20, 30, 40, 50, 60, 100, 250, 400, 600}:



Estatística Descritiva - Visualização

- ***Scatter plots*:** mostram a relação dos pontos entre dois atributos de um conjunto de dados;
    - Essas visualizações também ajudam a visualizar correlações entre os dados a partir da concentração dos seus pontos.



DATA SCIENCE

AULA 1 - Introdução
Duvidas e/ou perguntas?

gabriel.rmachado10@gmail.com

https://www.linkedin.com/in/machadogabriel

https://github.com/UNIFESO-Gabriel/data-science