Analises Exploratória de Dados (EDA)

Prof. Dr. Jorge Zavaleta

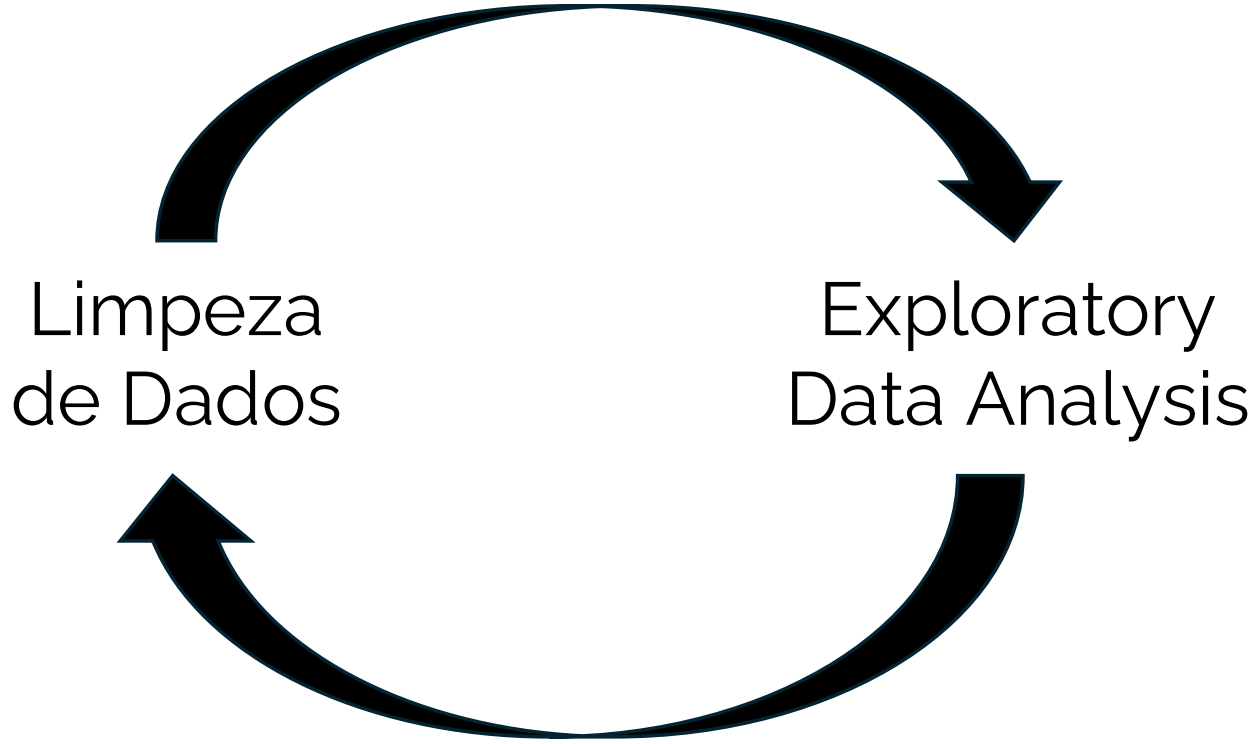
Departamento de Ciências Ambientais (DCA)

Universidade Federal Rural do Rio de Janeiro (UFRRJ)

Pesquisador de Pós-doutorado (PDJ/CNPq)

Ciclo de Trabalho do Cientista de Dados

Quanto tempo?



Tempo
30%
40%
50%
60%
70%

Análise Exploratória de Dados

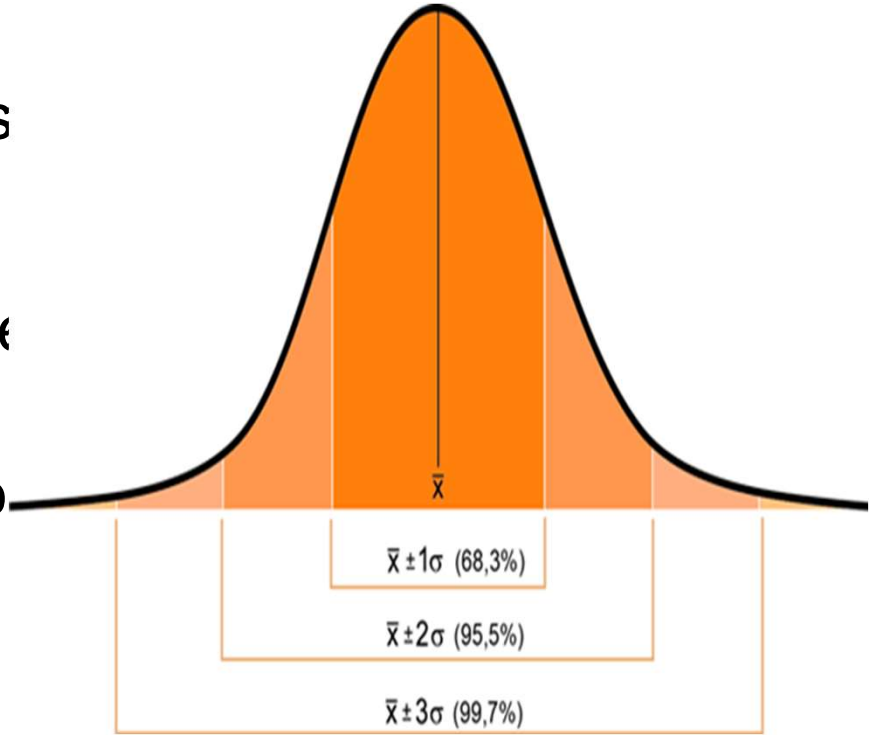
- A análise exploratória de dados (EDA) é uma etapa crucial no processo de análise de dados, que envolve explorar e entender a natureza dos dados antes de aplicar qualquer modelo estatístico ou algoritmo de machine learning.
- Desta forma o analista consegue um entendimento básico de seus dados e das relações existentes entre as variáveis analisadas.
- A linguagem Python, junto da biblioteca **Pandas**, oferece uma ampla gama de ferramentas poderosas para realizar EDA de forma eficiente e eficaz.

Etapas da EDA

- Preparar os dados para serem acessíveis a qualquer técnica estatística;
- Realizar um exame gráfico da natureza das variáveis individuais a analisar e uma análise descritiva que permita quantificar alguns aspectos gráficos dos dados;
- Realizar um exame gráfico das relações entre as variáveis analisadas e uma análise descritiva que quantifique o grau de inter-relação entre elas;
- Identificar os possíveis casos atípicos (outliers);
- Avaliar, se for necessário, a presença de dados ausentes (missing);
- Avaliar, se for necessário, algumas suposições básicas, como normalidade, linearidade e homoscedasticidade.

Etapas da EDA

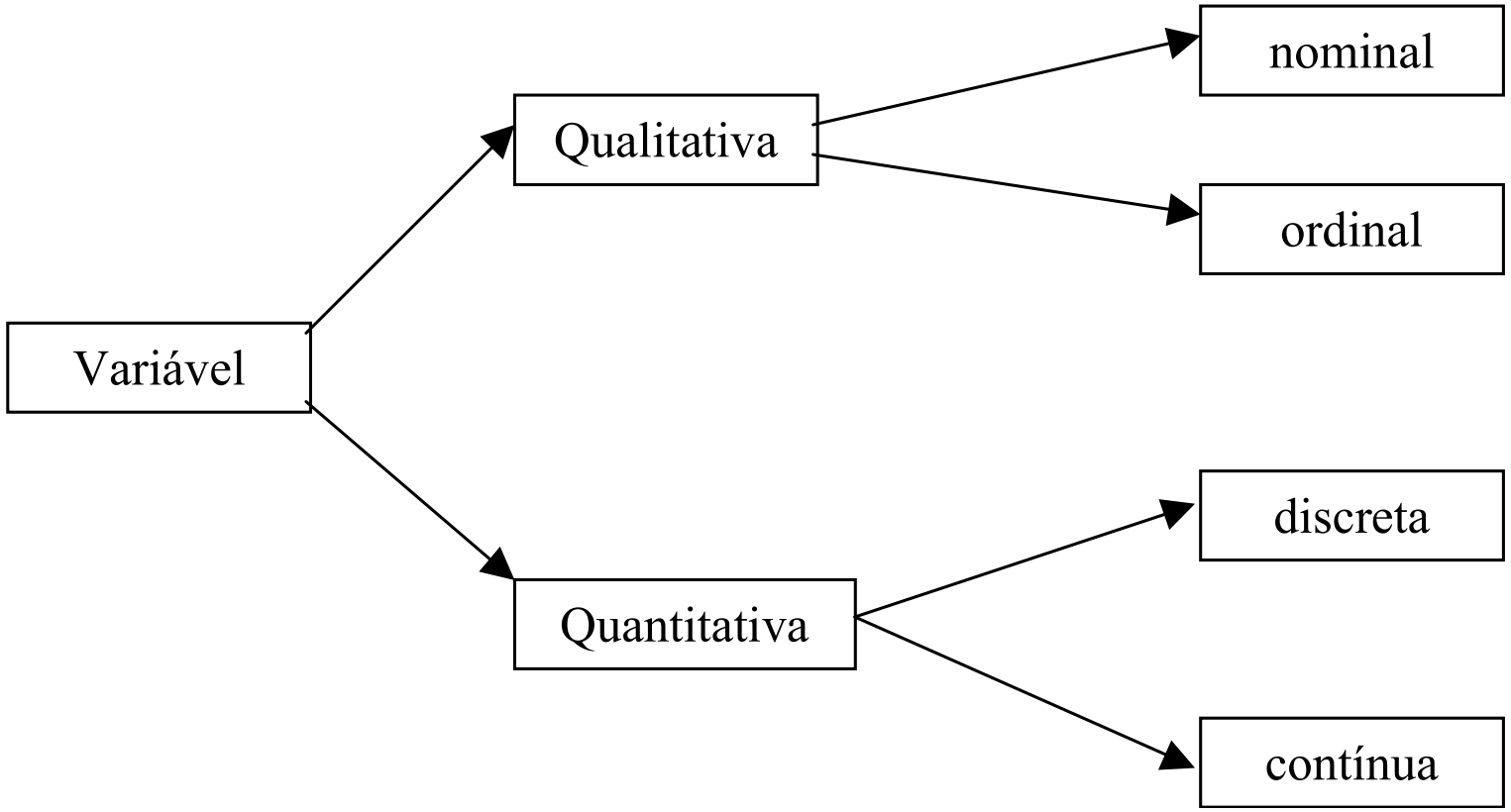
- A EDA extrai informações de um conjunto de dados sem o peso das suposições de um modelo probabilístico. As técnicas gráficas desempenham um importante pape nesta forma de abordagem
- Para entender a EDA é necessário ter uma boa compreensão de:
  - Estratégia de análise da Estatística Clássica,
  - Estatística Bayesiana



Tipos de variáveis

- **Variável** é uma característica, propriedade ou atributo de uma unidade da população, cujo valor pode variar entre as unidades da população.
- **Tipos de Variáveis**
  - **Variáveis Qualitativas ou Categóricas**: Quando os possíveis valores assumem atributos ou qualidades.
    - Exemplo: sexo, cor, escolaridade, doença, condição do ar, condição da água, etc.
  - **Variáveis Quantitativas ou de Medidas:** Quando seus valores são expressos em números.
    - Exemplo: altura, peso, número de filhos, pH, concentração do reagente, etc .

Tipos de variáveis - classificação



Tipos de variáveis - qualitativas

- As **variáveis qualitativas** podem ser classificadas ainda como:
- **Ordinais**: quando o atributo tem uma ordenação natural, indicando intensidade crescente de realização.
  - Exemplo: grau de escolaridade, classe social, condição do ar, condição da água, estado clínico, etc.
- **Nominais**: quando o atributo não se estabelece ordem.
  - Exemplo: sexo, cor, raça, doença, etc.

Tipos de variáveis - quantitativas

- As **variáveis quantitativas** podem ser classificadas ainda como:
- **Discretas**: resultantes de contagens, assumindo assim, em geral valores inteiros.
  - Exemplo: número de filhos, número de peças defeituosas, nº de pessoas doentes na região, etc.
- **Contínuas**: assumem valores em intervalos de números reais e geralmente, são provenientes de uma mensuração.
  - Exemplo: peso, altura, pH, concentração do reagente, etc..

Variáveis Quantitativas

- **Medidas de posição**: valor ao redor do qual os dados estão distribuídos.
  - Máximo: a maior observação
  - Mínimo: a menor observação
  - Moda: valor (atributo) que ocorre com maior frequência
  - Média: soma de todos os valores dividida pelo número de observações
  - Mediana: valor que deixa 50% das observações à sua esquerda
  - Quartis: divide um conjunto de valores dispostos em forma crescente em quatro partes.
    - Primeiro Quartil (Q1): valor que deixa 25% das observações à sua esquerda.
    - Terceiro Quartil (Q3): valor que deixa 75% das observações à sua esquerda.

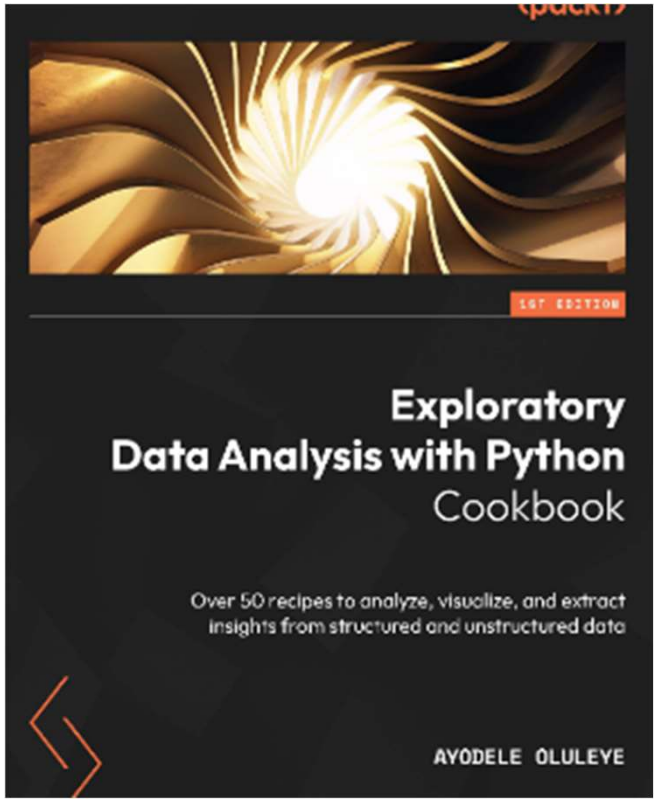
Variáveis Quantitativas

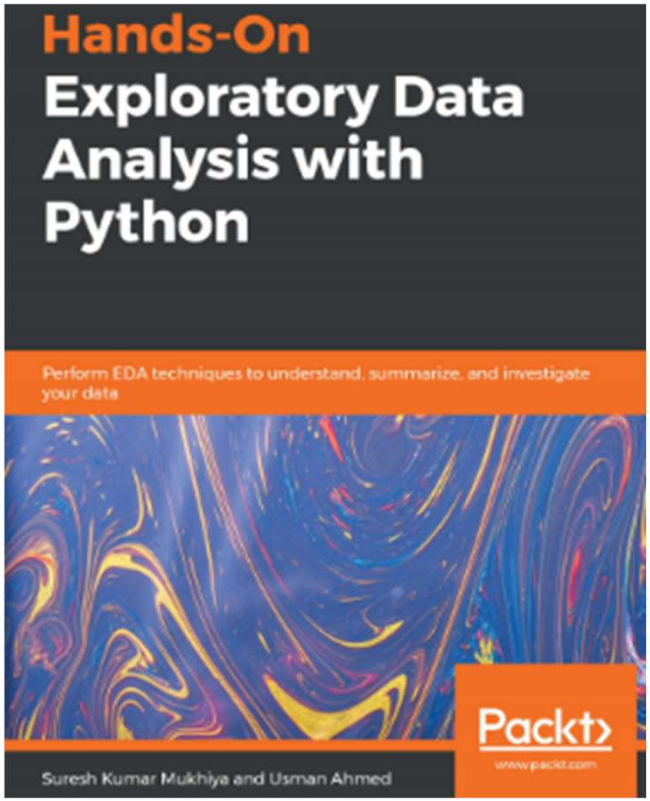
- **Medidas de dispersão:** A finalidade é encontrar um valor que resuma a variabilidade de um dataset.
  - **Amplitude**: diferença entre o valor máximo e o valor mínimo
  - **Intervalo-Interquartil**: É a diferença entre o terceiro quartil e o primeiro quartil, ou seja, Q3 - Q1
  - **Variância**: média dos quadrados dos desvios em relação à média aritmética
  - **Desvio Padrão**: mede a variabilidade independentemente do número de observações e com a mesma unidade de medida da média
  - **Coeficiente de Variação**: mede a variabilidade numa escala percentual independente da unidade de medida ou da ordem de grandeza da variável

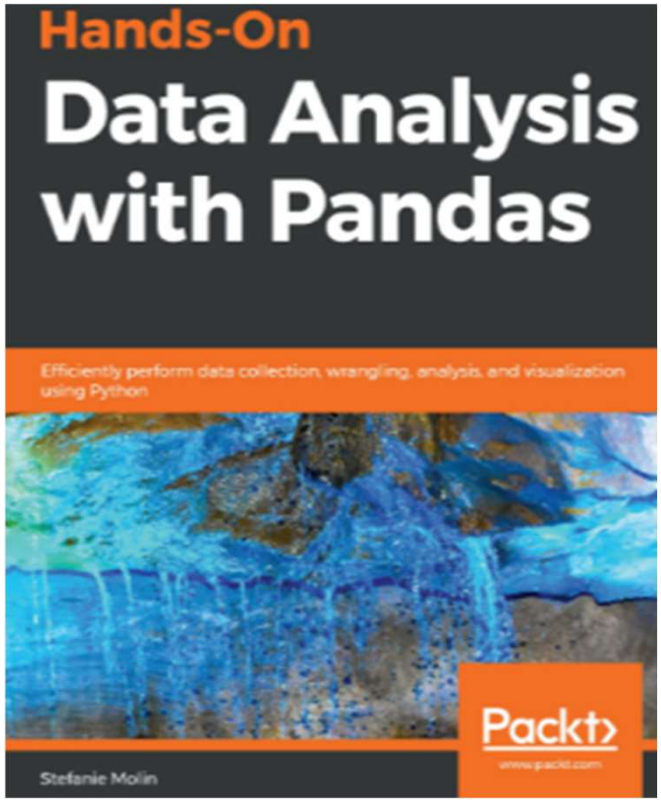
Visualização gráfica dos dados

- Distribuição:
  - Histograma
  - ramo-e-folhas
- Relação entre as variáveis:
  - Diagrama de dispersão
- Diferenças entre grupos:
  - Box-plot (observações atípicas podem aparecer somente após agrupamento)

Referências







Hands on...

NOTEBOOK:

EDA, EDA1, EDA2