Análise Exploratória de Dados.

Unidade I

INSTITUTO DE COMPUTAÇÃO

UFMT

O que é AED?

- A análise exploratória de dados (AED) refere-se ao processo de realizar investigações inicias nos dados, utilizando um conjunto de técnicas estatísticas e representações gráficas, cujo objetivo é detectar anomalias, testar hipóteses, compreender as relações entre as variáveis e verificar suposições, maximizando a compreensão do conjunto de dados.

Benefícios da AED

- Entender melhor os padrões nos dados.
- Detectar outliers ou eventos anômalos.
- Testar hipóteses.
- Identificar informações importantes nos dados.
- Revelar relações entre as variáveis do conjunto de dados.
- Determinar a melhor forma de manipular fontes de dados.
- Determinar as técnicas mais apropriadas aos dados.

Ferramentas para AED

- Descrição estatística dos dados.
- Visualização univariada de cada atributo do conjunto de dados.
- Visualização bivariada, permitindo avaliar a relação entre cada atributo do conjunto de dados.
- Visualizações multivariadas, permitindo mapear e entender as interações entre diferentes atributos.
- Técnicas de agrupamento e redução de dimensionalidade.

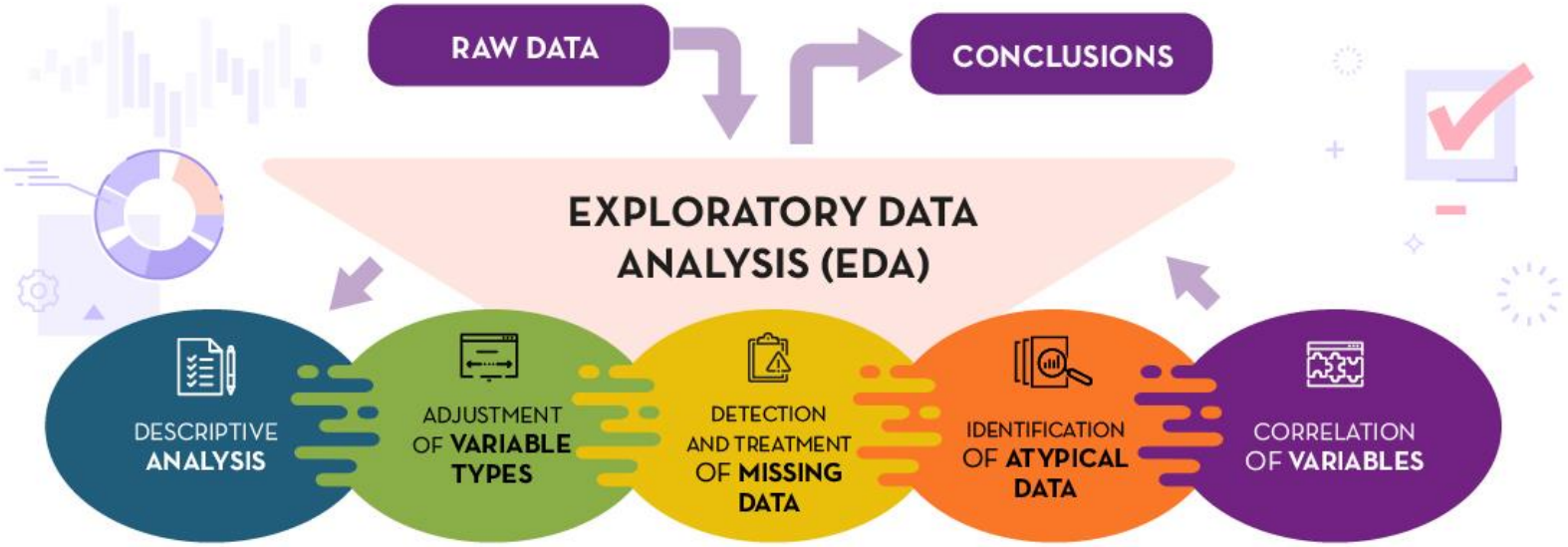
Ferramentas da AED

https://devopedia.org/exploratory-data-analysis

Ferramentas de AED

- Softwares utilizados
  - Python: Uma linguagem de programação interpretada e orientada a objetos com semântica dinâmica.
  - R: Uma linguagem de programação de código aberto e ambiente de software livre para computação estatística e gráficos suportados pela R Foundation for Statistical Computing.

Etapas da AED

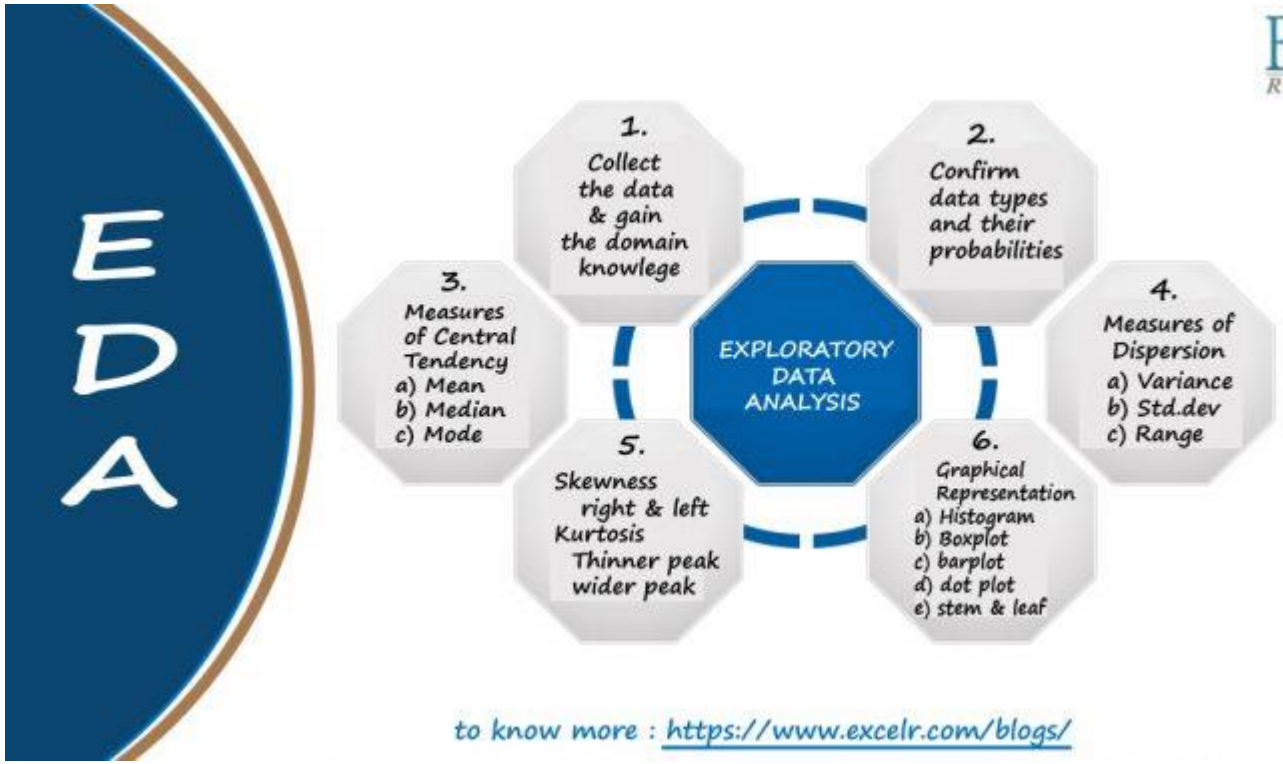


https://datos.gob.es/en/documentacion/practical-introductory-guide-exploratory-data-analysis

Etapas da AED

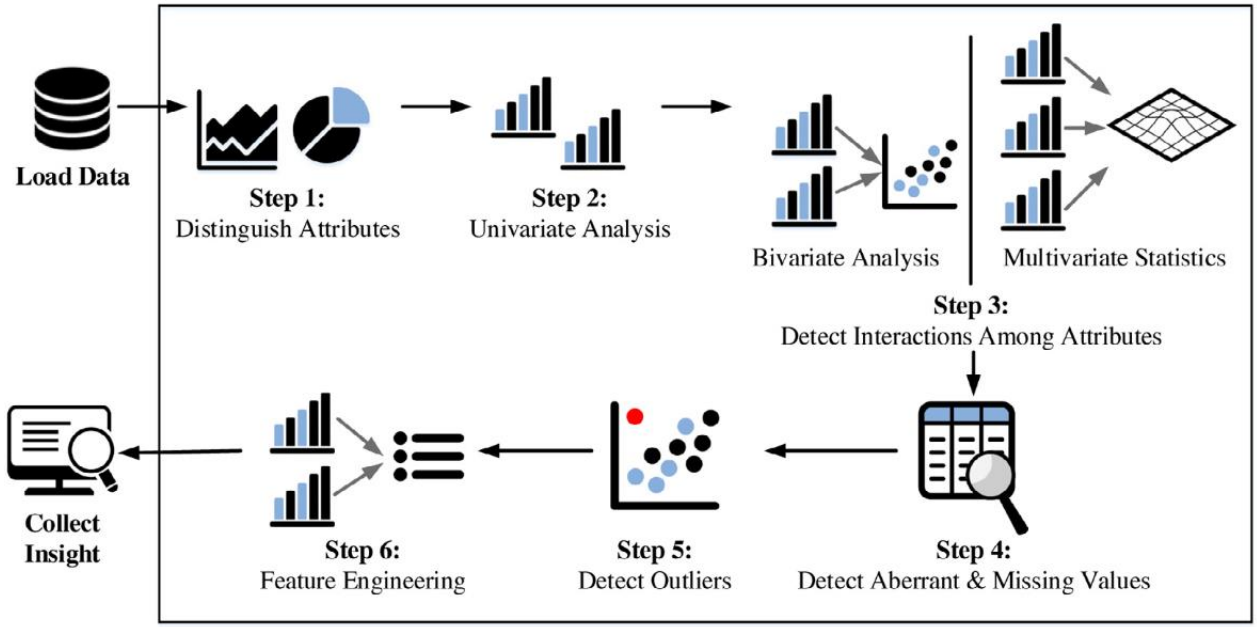
1. Análise descritiva das variáveis.
2. Ajuste dos tipos das variáveis para que sejam consistentes.
3. Detecção e tratamento de dados ausentes.
4. Identificação e tratamento de outliers.
5. Análise numérica e gráfica das relações entre as variáveis, identificando o grau de correlação entre elas.

Etapas da AED



https://www.excelr.com/exploratory-data-analysis-in-data-science

Etapas da AED



https://devopedia.org/exploratory-data-analysis

Tipos de Dados

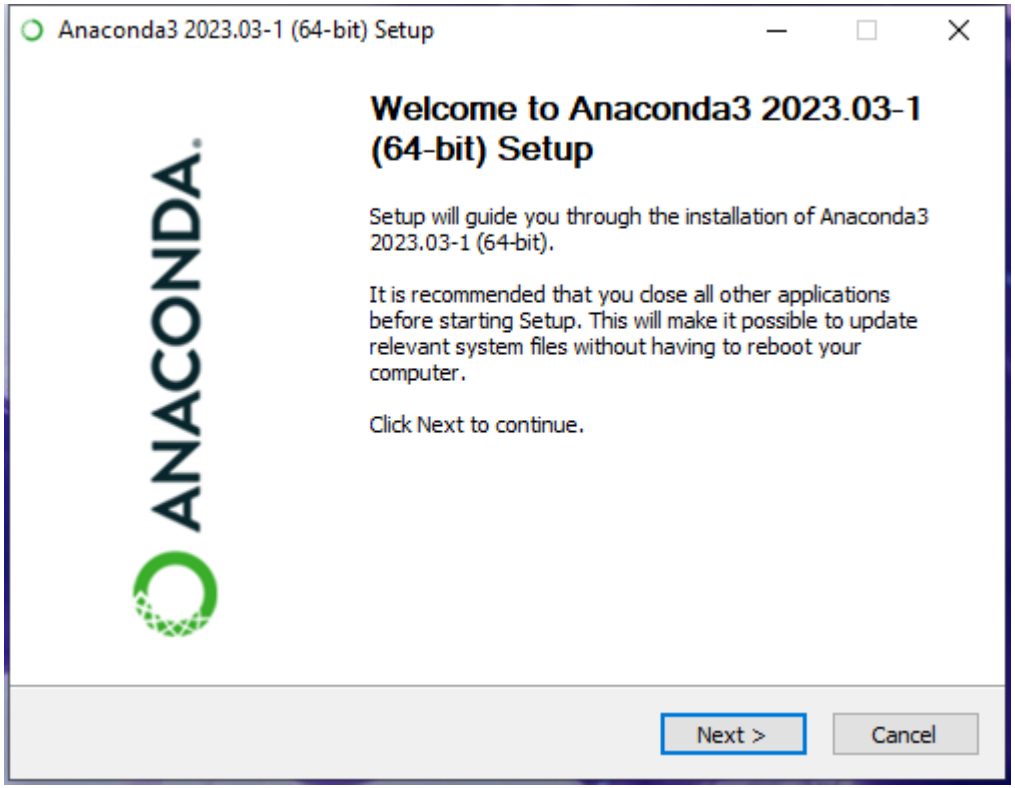
- Variável categórica:
  - Possuem o formato de string ou texto.
  - Ex.:Datas, Gênero, etc.
  - Podem ser divididas em:
    - Variáveis nominais:
      - São valores que não têm ordem entre si, ou seja, nenhum valor é maior que o outro.
      - Ex.: Gêneros, clima, país, etc.
    - Variáveis ordinarias:
      - Podem ser organizadas em alguma ordem em relação umas às outras.
      - Ex.: Classificação em uma competição, nível de escolaridade, etc.

Tipos de Dados

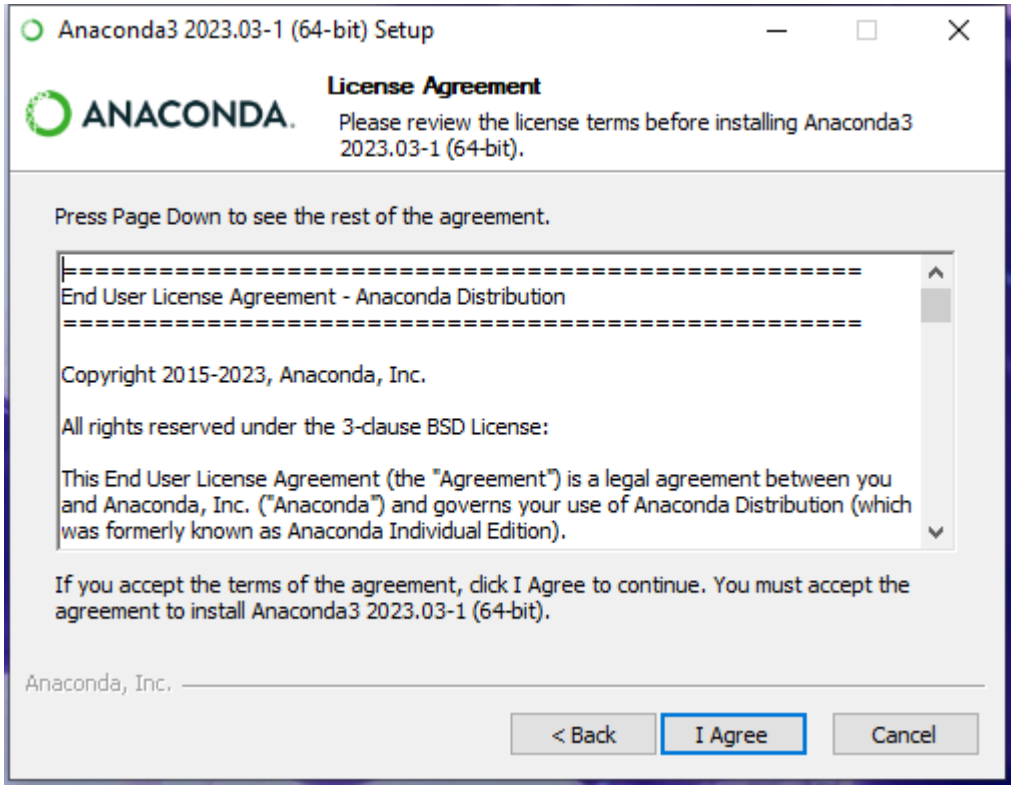
- Variáveis numéricas:
  - Possuem valores numéricos que podem ser medidos.
  - Ex.: Idade, valor do estoque, peso, altura, etc.
  - Pode ser divididos em:
    - Variáveis contínuas:
      - Possuem valores infinitos.
      - Ex.: Altura, peso, comprimento, distância, etc.
    - Varáveis discretas:
      - Possuem valores finitos.
      - Ex.: Número de medalhas, número de diplomas, etc.

Instalação Anaconda

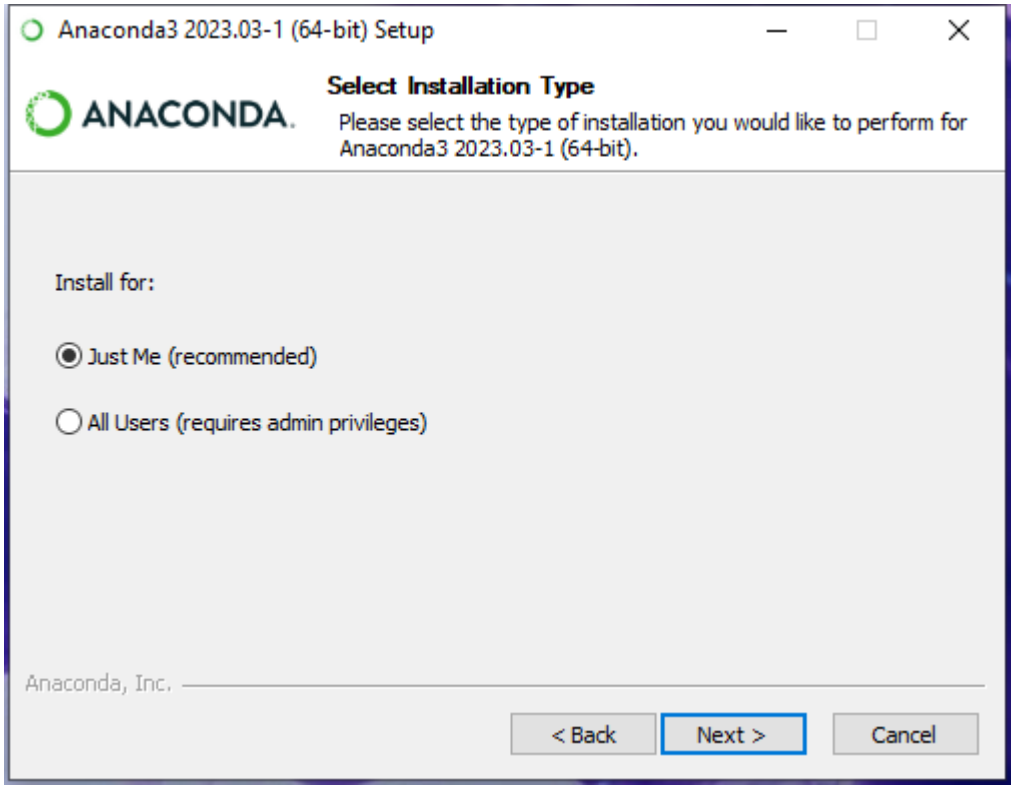
- https://www.anaconda.com/



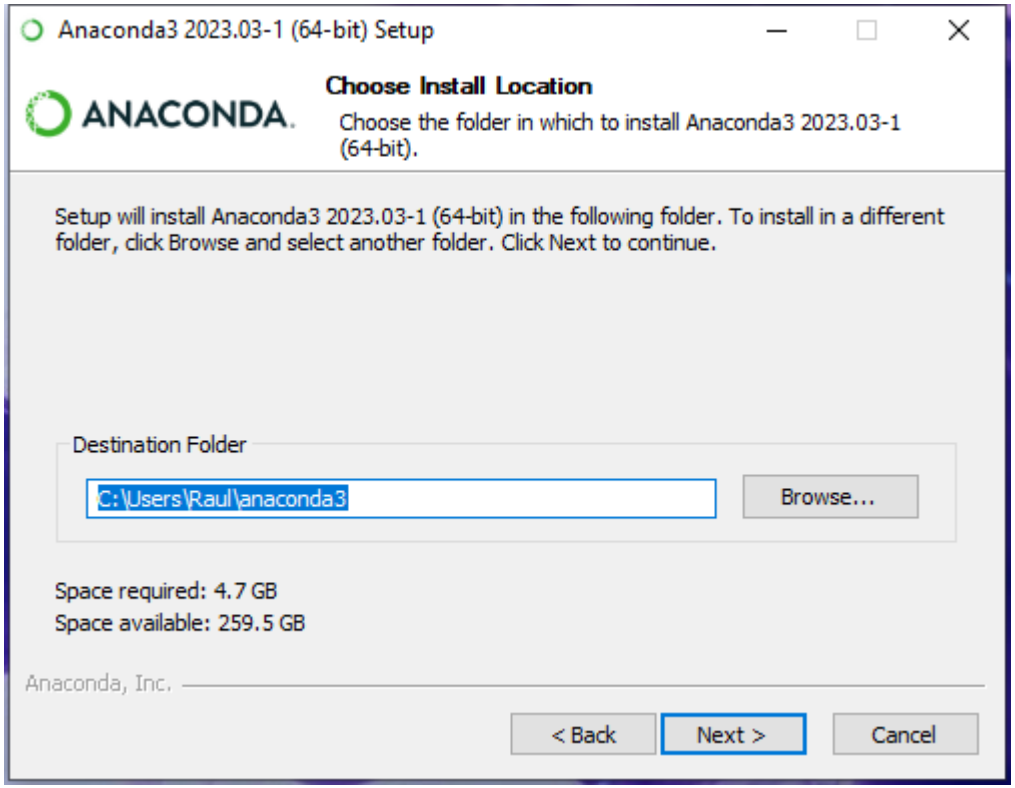
Instalação Anaconda



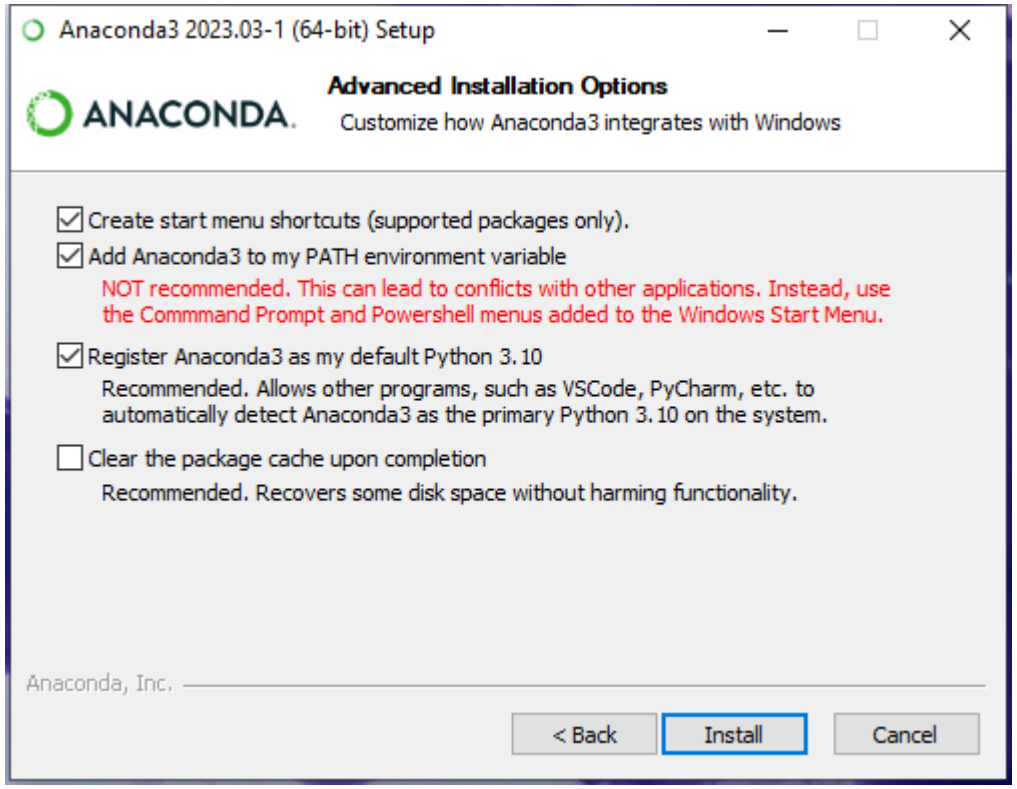
Instalação Anaconda



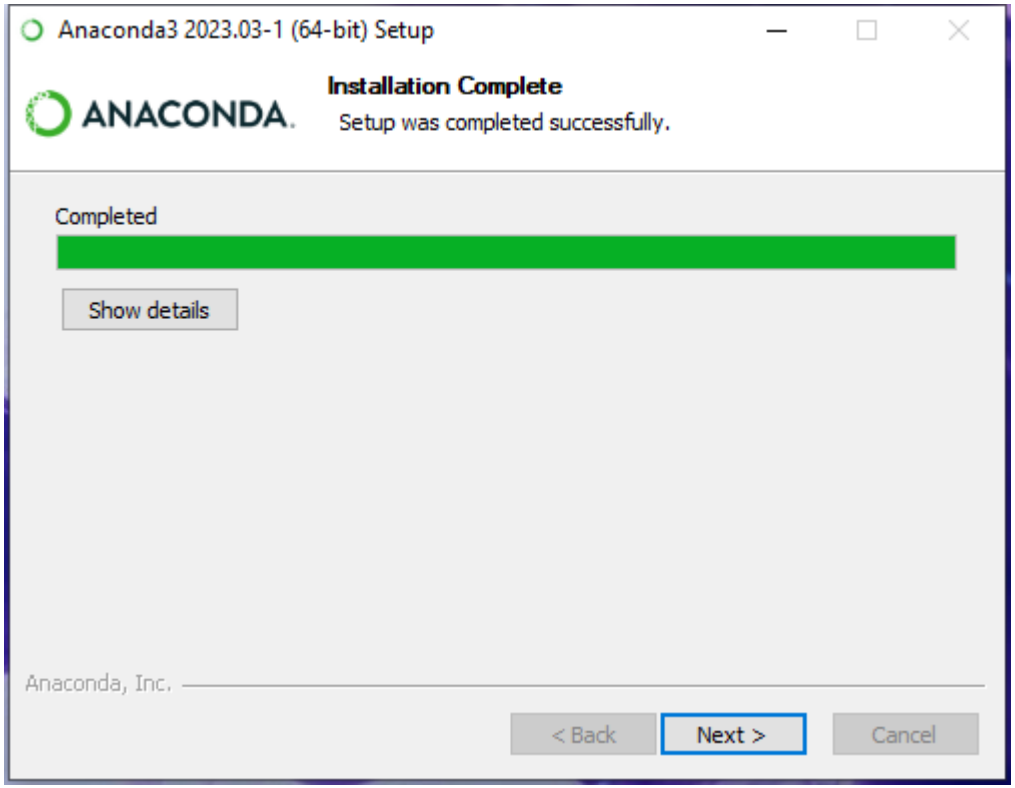
Instalação Anaconda



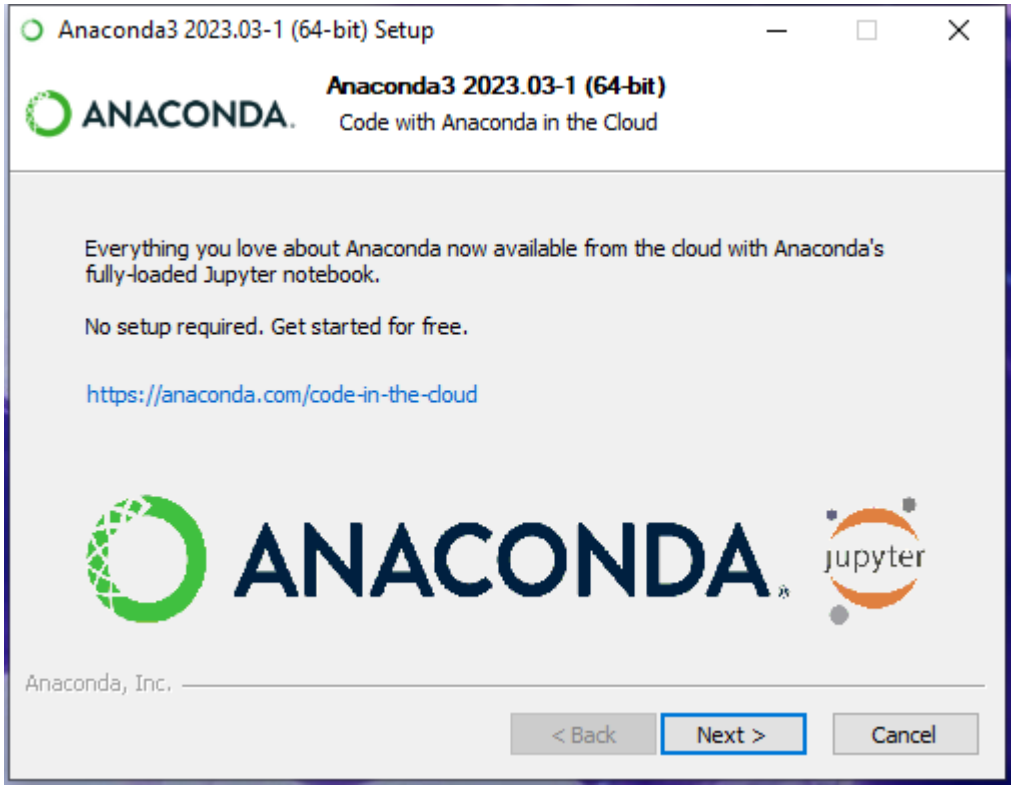
Instalação Anaconda



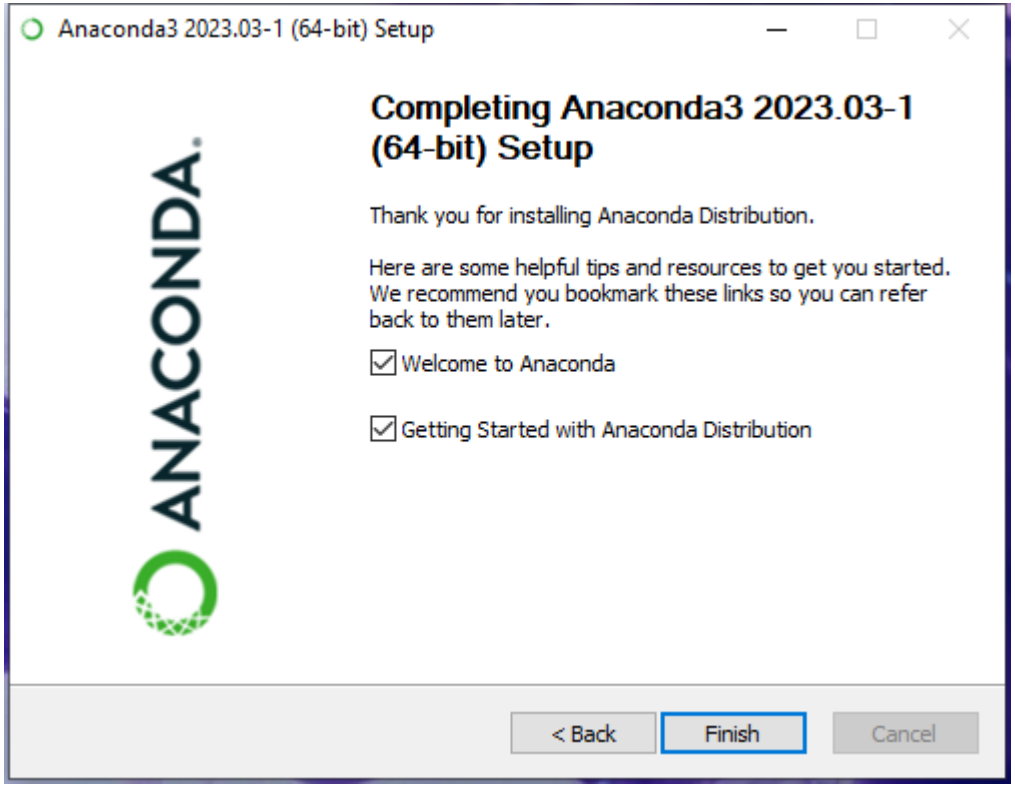
Instalação Anaconda



Instalação Anaconda



Instalação Anaconda



Bibliotecas

- Numpy
- Matplotlib
- Seaborn
- Pandas

Importando dados

- CSV
  - *import pandas as pd*
  - *emprego_df = pd.read_csv('emprego.csv', delimiter=',')*
  - *emprego_df*
  - *emprego_df = pd.read_csv('emprego.csv', delimiter=',', index_col=0)*
- XLS
  - *import pandas as pd*
  - *professores_df = pd.read_excel('..|disciplina_AED|dados.xlsx')*
  - *professores_df = pd.read_excel('..|disciplina_AED|dados.xlsx', index_col=0)*

Visualizando dados

- É importante verificar as colunas e os tipos de dados.
- *professores_df.columns*
- *professores_df.info()*
- *professores_df.head()*
- *professores_df.tail()*
- *professores_df['name']*
- *professores_df['genero'].unique()*

Dados ausentes

- Algumas causas para dados ausentes.
  - Informações não preenchidas devido a questões de privacidade.
  - Falhas no processo de coleta de dados.
  - Perdas de dados no processo de transferência.
- Podem existir muitas outras razões para dados ausentes.

Identificando dados ausentes

- *professores_df*
- *len(professores_df)*
- *professores_df.info()*
- *professores_df.isna()*
- *professores_df.notna()*
- *professores_df.isnull()*
- *professores_df.isnull().sum()*
- *professores_df.describe()*
- Valores NaN são considerados float pelo Pandas.
- Dados ausentes no formato datetime são denominados NaT.

Tratando dados ausentes

- Diferentes formas para tratar dados ausentes
  - Ignorar linhas com valores ausentes
    - Algumas vezes a quantidade de linhas com dados ausentes pode ser menor do que 1-5%.
    - Um método é remover linhas contendo a maioria das colunas sem preenchimento.
    - Devemos ter cuidado para não remover a maior parte do conjunto de dados.
    - Remover muitas linhas pode reduzir a qualidade para modelos de Aprendizado de Máquina.

Tratando dados ausentes

- Diferentes formas para tratar dados ausentes
  - Preenchimento de dados ausentes.
    - Um forma é o preenchimento com um valor genérico ou inferir a partir do conjunto de dados.
    - Não é sempre possível inferir o valor a ser preenchido.

Idade	Localização do assento
65	Inferior
70	Inferior
15	Superior
24	Meio
72	?????

Tratando dados ausentes

- Diferentes formas para tratar dados ausentes
  - Preenchimento de dados ausentes.
    - Outra forma de preenchimento é utilizar medidas de tendência central (média, mediana, moda, etc.).
    - É importante tomar cuidado para que eles não alterem os padrões gerais nos dados (a média é afetada por outliers).
    - O valor a ser utilizado no preenchimento pode ser escolhido com base nos valores de outros atributos.

Tratando dados ausentes

Gênero	Peso (kg)
M	70
F	55
M	65
F	??
F	60
M	??
F	52
F	53
M	85
M	75
M	??
F	68

Média de Peso por Gênero:
M: 73.75
F: 57.6

Preenchendo valores ausentes

Gênero	Peso (kg)
M	70
F	55
M	65
F	57.6
F	60
M	73.75
F	52
F	53
M	85
M	75
M	73.75
F	68

Tratando dados ausentes

- Removendo linhas com dados ausentes
  - *import pandas as pd*
  - *professores2_df = pd.read_excel('dados_2.xlsx', index_col=0)*
  - *professores2_df*
  - *professores2_df_no_missing = professores2_df.dropna()*
  - *professores2_df_no_missing*
  - Não devemos remover a maioria das linhas.

Tratando dados ausentes

- Removendo linhas com a maioria das colunas com dados ausentes
  - *import pandas as pd*
  - *professores2_df = pd.read_excel('dados_2.xlsx', index_col=0)*
  - *professores2_df*
  - *professores2_df_limiar_missing = professores2_df.dropna(thresh=5)*
  - *professores2_df_limiar_missing*
  - O parâmetro *thresh* indica mínimo de colunas preenchidas para manter o registro.

Tratando dados ausentes

- Removendo colunas com um percentual elevado de dados ausentes
  - *professores2_df_percentual_columns_missing = professores2_df.dropna(axis=1, thresh=int(0.6*len(professores2_df)))*
  - *professores2_df_percentual_columns_missing*
  - Remove colunas com mais de 60% (0.6) de dados ausentes
  - O parâmetro axis de deve ter valor 0 para remover linhas ou 1 para remover colunas.

Tratando dados ausentes

- Preenchimento de dados ausentes com valores genéricos.
  - *professores2_df_generic_no_missing = professores2_df.fillna(-1)*
  - *professores2_df_generic_no_missing*

Tratando dados ausentes

- Preenchimento de dados ausentes com valores das linhas adjacentes.
  - Utiliza o próximo valor não nulo.
    - *professores2_df_next_value = professores2_df.bfill()*
  - Utiliza um valor anterior não nulo.
    - *professores2_df_forward_value = professores2_df.ffill()*

Tratando dados ausentes

- Preenchimento de dados ausentes com valores de tendências centrais.
  - *idade_media = professores2_df['idade'].mean()*
  - *altura_media = professores2_df['altura'].mean()*
  - *print("Idade Média:", idade_media, " - Altura Média:", altura_media)*
  - *dados_preenchimento = {'idade': idade_media, 'altura': altura_media} # Dicionário*
  - *dados_preenchimento*
  - *professores2_df_central_tendency=professores2_df .fillna(value=dados_preenchimento)*

Tratando dados ausentes

- Preenchimento de dados ausentes baseado em condições.
  - *professores2_df_peso_genero = professores2_df[['genero', 'peso']]*
  - professores2_df_peso_genero.groupby("genero").transform(lambda x: x.fillna(x.mean()))
  - professores2_df[['genero','peso']] = professores2_df_peso_genero

Tratando dados ausentes

- https://pandas.pydata.org/pandas-docs/stable/user_guide/missing_data.html
- Atividade: Efetuar o tratamento de dados ausentes de um conjunto de dados.