



LADE

LABORATÓRIO DE ANÁLISES DE DADOS
EDUCACIONAIS E ESTATÍSTICA APLICADA

IFCE - CAMPUS FORTALEZA

Estatística Descritiva com Python

Análise de dados : Teoria e Tecnologias

1. Importação de Dados
2. Limpeza e Preparação dos Dados
3. Análise Exploratória de Dados (EDA)
4. Modelagem de Dados
5. Visualização de Resultados

Fonte: <https://www.cursospm3.com.br/blog/python-para-analise-de-dados/>

1. Planejamento
2. Coleta de Dados
3. Crítica dos dados
4. Apuração dos dados
5. Análises de Dados
6. Emissão do Relatório Final
7. Comunicação dos Resultados

Fonte: Costa, Giovani. Estatística Aplicada A Educacao Com Abordagem Além da análise descritiva

1. Olhar para o quadro geral;
2. Obter os dados;
3. Descobrir e visualizar os dados para obter informações;
4. Preparar os dados para os algoritmos do Aprendizado de Máquina;
5. Selecionar e treinar um modelo;
6. Ajustar o seu modelo;
7. Apresentar sua solução;
8. Lançar, monitorar e manter seu sistema.

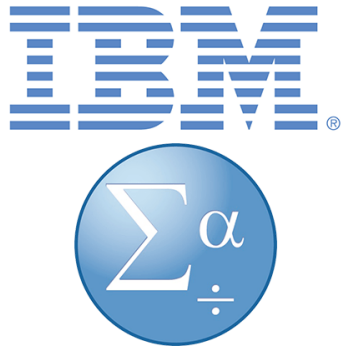
Fonte :Géron, Aurélien. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn & TensorFlow

- 1. Defina seus objetivos:** Antes de começar a análise, é crucial entender o que você quer alcançar. Qual é a pergunta ou problema que você está tentando resolver?
- 2. Coleta de dados:** Reúna os dados relevantes para sua análise. Isso pode envolver a obtenção de dados de fontes diferentes, como bancos de dados, arquivos CSV, APIs da web, etc.
- 3. Limpeza de dados:** Os dados geralmente precisam ser limpos e preparados para análise. Isso pode incluir a remoção de dados duplicados, tratamento de valores ausentes, padronização de formatos de dados, entre outras tarefas.
- 4. Exploração de dados:** Nesta etapa, você explora seus dados para entender melhor sua estrutura e identificar padrões, tendências e relações. Isso pode envolver a criação de visualizações de dados, como gráficos e tabelas.
- 5. Análise estatística:** Use técnicas estatísticas para analisar seus dados de maneira mais aprofundada. Isso pode incluir medidas de tendência central, dispersão, correlação, regressão, entre outras.
- 6. Modelagem de dados (se aplicável):** Se você estiver realizando análises preditivas ou inferenciais, pode ser necessário criar modelos estatísticos ou de aprendizado de máquina para fazer previsões ou extrair insights dos dados.
- 7. Interpretação dos resultados:** Analise os resultados de sua análise e interprete o que eles significam em relação aos seus objetivos iniciais. Isso pode envolver a elaboração de conclusões e recomendações com base nos insights obtidos.
- 8. Comunicação dos resultados:** Comunique seus resultados de forma clara e compreensível para as partes interessadas relevantes. Isso pode incluir a criação de relatórios, apresentações ou visualizações de dados para transmitir suas descobertas.
- 9. Iteração e refinamento:** Às vezes, é necessário revisitar etapas anteriores, refinar suas análises ou coletar mais dados para obter uma compreensão mais completa do problema em questão.

(ChatGPT 3.5)



Análise de dados : Teoria e Tecnologias



Usaremos :





Olá, este é o Colaboratory

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda



Índice



+ Código

+ Texto

Copiar para o Drive



Vamos começar



Ciência de dados



Machine learning



Mais recursos

Exemplos em destaque


+ Seção

Conheça o Colab

(Novidade) Teste a API Gemini

- [Generate a Gemini API key](#)
- [Talk to Gemini with the Speech-to-Text API](#)
- [Gemini API: Quickstart with Python](#)
- [Gemini API code sample](#)
- [Compare Gemini with ChatGPT](#)
- [More notebooks](#)

Importação de Dados

 AulaCursoEstatisticaDecritiva.ipynb ☆
Arquivo Editar Ver Inserir Ambiente de execução Ferramentas /

Índice



+ Seção

{x}



Clica em arquivos

+ Código + Texto



Comece a progr



AulaCursoEstatisticaDecritiva.ipynb ☆

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda [Todas as alterações foram salvas](#)

Arquivos



{x}

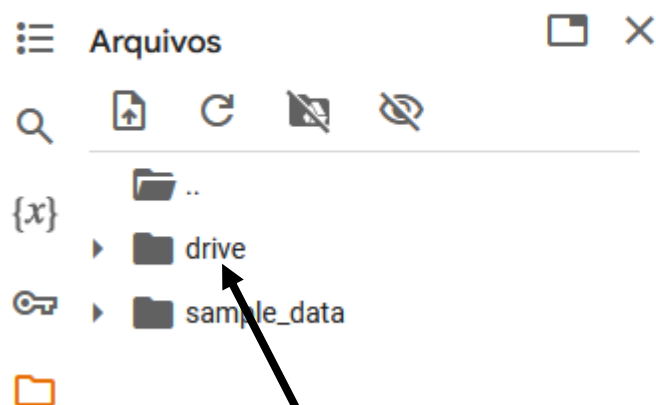


..

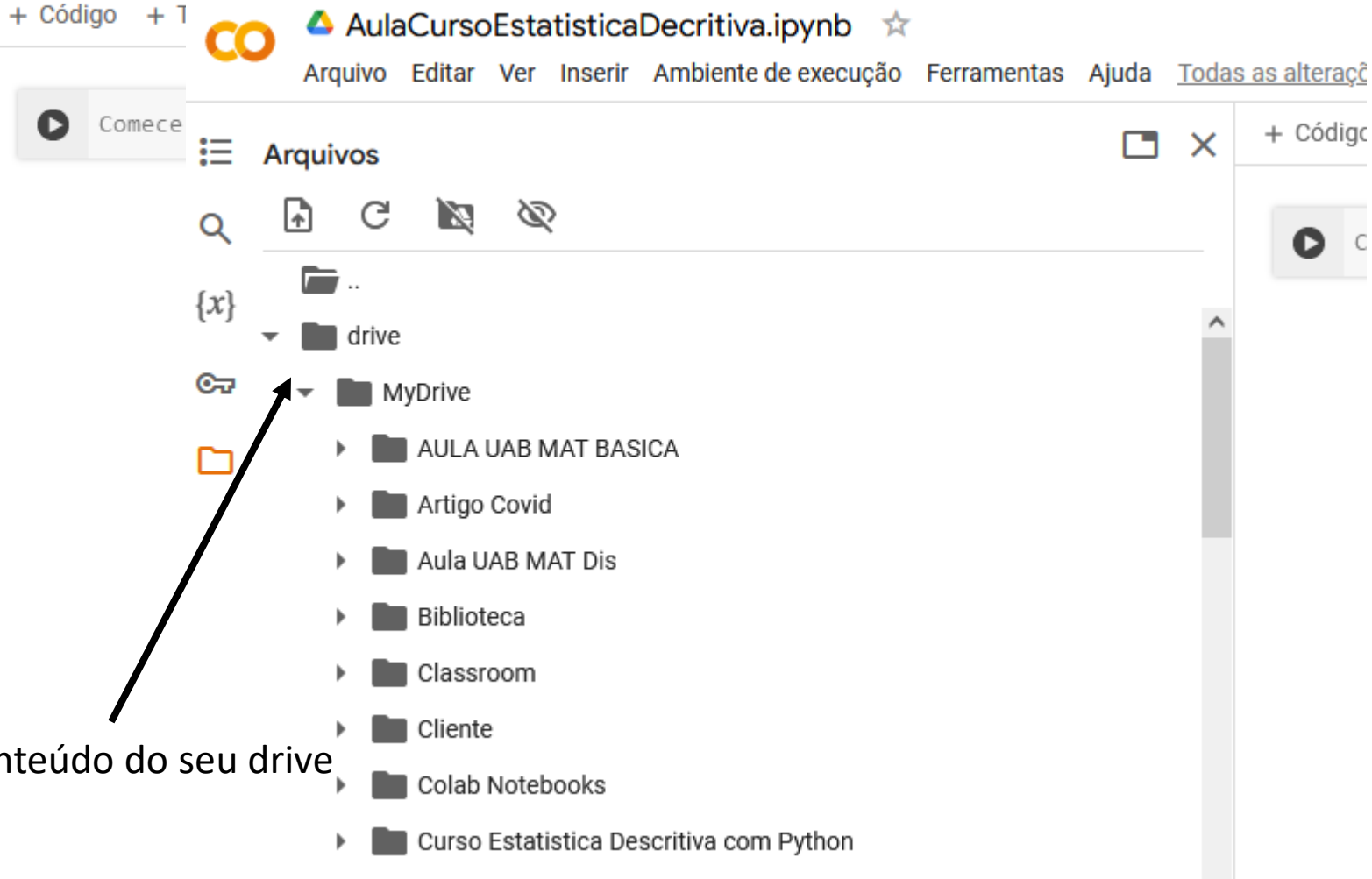
sample_data



Montar seu Drive, para ter acesso aos seus arquivos.



Drive Montado



Acessando ao conteúdo do seu drive



AulaCursoEstatisticaDecritiva.ipynb ☆

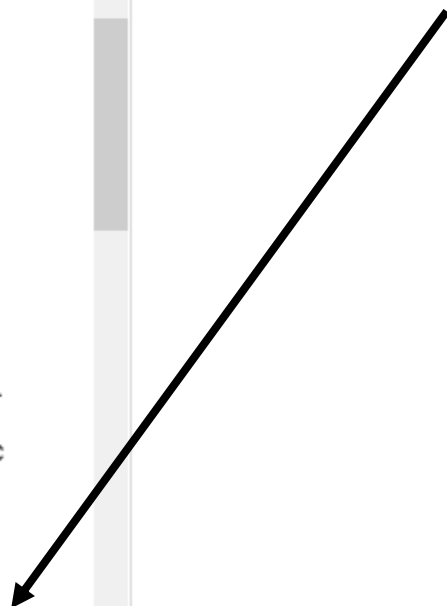
Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alteraç

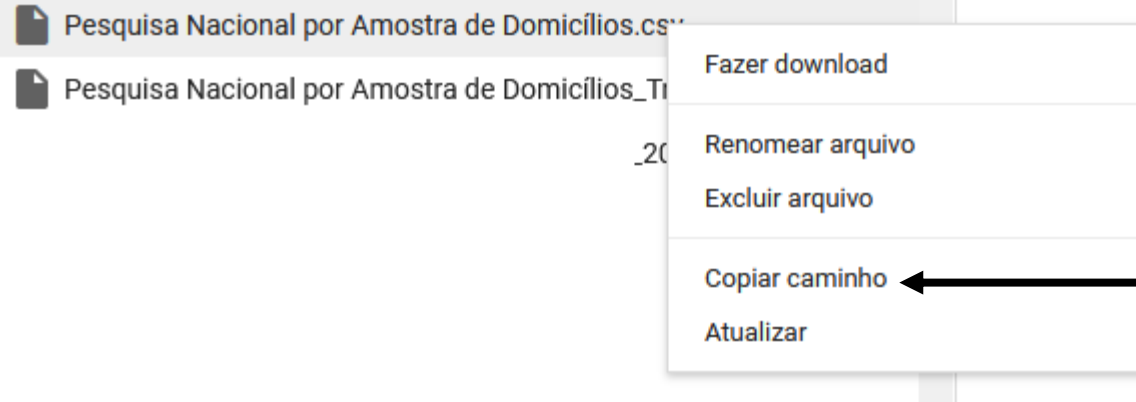
Arquivos



- Curso Estatística Descritiva com Python
- Curso de Introdução a Python
- Curso de programação
- Cursos
- Despesas
- Disciplina de Probabilidade e Estatística
 - PROBABILIDADE E ESTATISTICA
 - Aula02Exercicio_Pro_Est_PesquisaAmostragem.ipynb
 - Chamada dos Alunos 2024_1_Probabilidade_Computação.d...
 - Chamada dos Alunos 2024_1_Probabilidade_Matemática.doc
 - Disciplina de Probabilidade e Estatística.docx
 - Horário das Aulas_trabalho_LADE_estudos_2024_1.docx
 - Pesquisa Nacional por Amostra de Domicílios.csv

Procurando arquivo





Copiar o caminho do arquivo

/content/drive/MyDrive/Disciplina de Probabilidade e Estatística/Pesquisa Nacional por Amostra de Domicílios.csv

```
[2] import pandas as pd
import numpy as np
```

```
[ ] # Outra forma de montar seu drive

# from google.colab import drive
# drive.mount('/content/drive')
```

```
[3] PesquisaNacional = pd.read_csv('/content/drive/MyDrive/Disciplina de Probabilidade e Estatística/Pesquisa Nacional por Amostra de Domicílios.csv')
```

▶ PesquisaNacional



	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	0	23	8	12	800	1.603808
1	11	1	23	2	12	1150	1.739790
2	11	1	35	8	15	880	1.760444
3	11	0	46	2	6	3500	1.783158
4	11	1	47	8	9	150	1.690631
...
76835	53	1	46	2	11	812	1.687030
76836	53	0	30	4	7	1500	1.792934
76837	53	0	32	8	12	1300	1.830587



Temos a nossa base de dados, o quê fazer?
Só um pouco de calma, vamos ao Pandas e Numpy.

Pandas - <https://pandas.pydata.org/>

Uma das bibliotecas mais conhecidas e mais usadas por profissionais de dados, ela permite que o usuário **manipule, transforme e analise dados** de maneira muito otimizada.

A **Pandas** possibilita a leitura em vários formatos, como SQL, CSV, Excel, etc., além de funcionar, principalmente, com dois tipos de estrutura de dados: *Series* e *DataFrames*.

DataFrames seguem uma estrutura semelhante a uma planilha de **Excel**.

Series, se refere a um **array unidimensional**, que pode ser entendido como uma lista simples de valores.

NumPy - <https://numpy.org/>

A biblioteca **Numpy compila funções** relacionadas à álgebra linear e computação numérica, trabalhando com *arrays* multidimensionais, cálculos rápidos, entre outras funcionalidades.

Além disso, a biblioteca **NumPy está no núcleo de basicamente todos** os programas e bibliotecas que lidam com operações matemáticas e usam a linguagem de programação Python.

Como por exemplo, a própria biblioteca Pandas baseia sua estrutura de dados (*DataFrames* e *Series*) em *arrays* de NumPy.

Fonte: <https://www.cursospm3.com.br/blog/python-para-analise-de-dados/>



```
[2] import pandas as pd
import numpy as np
```

```
[ ] # Outra forma de montar seu drive

# from google.colab import drive
# drive.mount('/content/drive')
```

```
[3] PesquisaNacional = pd.read_csv('/content/drive/MyDrive/Disciplina de Probabilidade e Estatística/Pesquisa Nacional por Amostra de Domicílios.csv')
```

▶ PesquisaNacional



	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	0	23	8	12	800	1.603808
1	11	1	23	2	12	1150	1.739790
2	11	1	35	8	15	880	1.760444
3	11	0	46	2	6	3500	1.783158
4	11	1	47	8	9	150	1.690631
...
76835	53	1	46	2	11	812	1.687030
76836	53	0	30	4	7	1500	1.792934
76837	53	0	32	8	12	1300	1.830587



Vamos conhecer no base de dados, para isso usaremos

`PesquisaNacional.info()`



PesquisaNacional.info()



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 76840 entries, 0 to 76839
Data columns (total 7 columns):
#   Column                Non-Null Count  Dtype
---  -
0   UF                    76840 non-null  int64
1   Sexo                 76840 non-null  int64
2   Idade                76840 non-null  int64
3   Cor                  76840 non-null  int64
4   Anos de Estudo       76840 non-null  int64
5   Renda                76840 non-null  int64
6   Altura               76840 non-null  float64
dtypes: float64(1), int64(6)
memory usage: 4.1 MB
```

Tipos de Variáveis

Qualitativas

Nominal

- Profissão
- Sexo
- Religião

Ordinal

- Escolaridade
- Estágio da doença
- Classe social

Quantitativas

Discreta

- Nº de filhos
- Nº de acessos à plataforma

Contínua

- Altura
- Peso
- Salário

Iremos a documentação/dicionário da nossa base de dados

Pesquisa Nacional por Amostra de Domicílios - PNAD

Dicionário – Códigos.

Variáveis utilizadas

Renda

Documento na página do curso/Google sala de aula

Rendimento mensal do trabalho principal para pessoas de 10 anos ou mais de idade.

Idade

Idade do morador na data de referência em anos.

Altura (elaboração própria)

Altura do morador em metros.



Trocaremos os códigos pelos significados

```
PesquisaNacional['Sexo'].replace({0: 'Masculino', 1: 'Feminino'}, inplace=True)
```

```
[ ] PesquisaNacional['Sexo'].replace({0: 'Masculino', 1: 'Feminino'}, inplace=True)
```



PesquisaNacional



	UF	Sexo	Idade	Cor	Anos de Estudo	Renda	Altura
0	11	Masculino	23	8	12	800	1.603808
1	11	Feminino	23	2	12	1150	1.739790
2	11	Feminino	35	8	15	880	1.760444
3	11	Masculino	46	2	6	3500	1.783158
4	11	Feminino	47	8	9	150	1.690631
...
76835	53	Feminino	46	2	11	812	1.687030
76836	53	Masculino	30	4	7	1500	1.792934
76837	53	Masculino	32	8	12	1300	1.830587
76838	53	Masculino	57	8	4	1500	1.726344
76839	53	Masculino	38	8	4	900	1.658305

76840 rows × 7 columns

Exercício :

Rotular as variáveis qualitativas , usando o dicionário, do arquivo :

Pesquisa Nacional por Amostra de Domicílios.csv

Exercício para Casa :

Rotular as variáveis qualitativas , usando o dicionário, do arquivo :

Enem_Fortaleza_2021_DisciplinaProbabilidadeEstatística.xlsx

Agora temos um arquivo xlsx, como abri-lo usando o pandas?



LADE

LABORATÓRIO DE ANÁLISES DE DADOS
EDUCACIONAIS E ESTATÍSTICA APLICADA
— IFCE - CAMPUS FORTALEZA —

OBRIGADO!!!