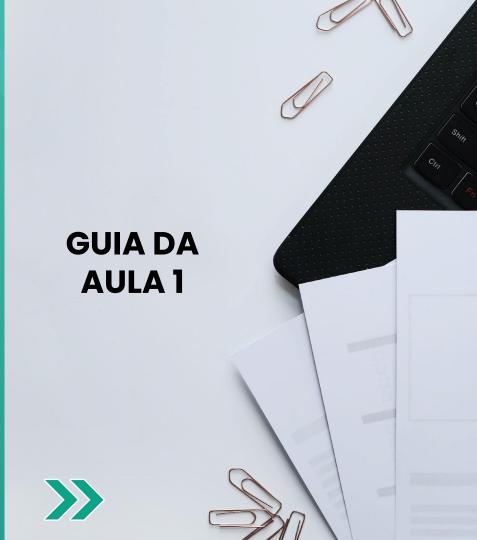


Profissão: Analista de dados





COLETA DE DADOS I







Manuseie arquivos CSV

- Estrutura de dados
- Pacote CSV



Acompanhe aqui os temas que serão tratados na videoaula



1. Estruturas de dados



Não estruturado: texto, imagem, áudio, etc.

Semi estruturado: html, json, etc.

Estruturado: tabelas, planilhas, etc.

2. Arquivos CSV

1.1. Formato

Um arquivo **csv** é um tipo de arquivo de **texto** com uma estrutura específica (**estruturado**) para organizar os dados num formato tabular:

- **Linhas** são separadas pelo caractere de nova linha $\cdot \setminus_n \cdot$, normalmente a primeira coluna é o cabeçalho (header);
- **Colunas** por um separador: ',' (mais comum), ';' , etc. É um tipo de arquivo muito utilizado (talvez o mais utilizado) para armazenar dados no mundo analítico.





Arquivo CSV: banco.csv

```
In []:

%%writefile banco.csv
age,job,marital,education,default,balance,housing,loan
30,unemployed,married,primary,no,1787,no,no
33,services,married,secondary,no,4789,yes,yes
35,management,single,tertiary,no,1350,yes,no
30,management,married,tertiary,no,1476,yes,yes
59,blue-collar,married,secondary,no,0,yes,no
35,management,single,tertiary,no,747,no,no
36,self-employed,married,tertiary,no,307,yes,no
39,technician,married,secondary,no,147,yes,no
41,entrepreneur,married,tertiary,no,221,yes,no
43,services,married,primary,no,-88,yes,yes
```





Exemplo: Extraindo os valores da primeira coluna (idade).

```
idades = []

with open(file='./banco.csv', mode='r', encoding='utf8') as arquivo:
    cabecalho = arquivo.readline().split(sep=',')
    indice_idade = cabecalho.index('age') linha =
    arquivo.readline()

while linha:
    idade = linha.split(sep=',')[indice_idade]
    idades.append(idade)
    linha = arquivo.readline()

print(idades)
```





Exemplo: Tipo dos dados.

```
In [ ]:
    tipos_idades = set(map(lambda idade: type(idade), idades))
    print(tipos_idades)
```

Exemplo: Média das idades.





1.2. Pacote CSV

Pacote nativo do Python que facilita a leitura de arquivos no formato CSV.

```
In []: import csv
saldos = None
with open(file='./banco.csv', mode='r', encoding='utf8') as arquivo:
    leitor_csv_iter = csv.reader(arquivo,
    delimiter=',') cabecalho = next(leitor_csv_iter)
    indice_saldo = cabecalho.index('balance')
    saldos = [linha[indice_saldo] for linha in leitor_csv_iter]
    print(saldos)
```



Exemplo: Média dos saldos.



