

Estácio

Nível 3 - Mundo 5 | Desenvolvimento FullStack 2024

Caique Thomaz Stopiglia - 202208292471

Tratando a imensidão dos dados

Realização de limpeza de dados utilizando comandos em Python com a biblioteca Pandas e Google Colab.

1 - Carregamento do Pandas, upload do arquivo csv, leitura do arquivo csv utilizando separação por vírgula e exibindo as 5 primeiras e últimas linhas do arquivo.

```
import pandas as pd

[2] from google.colab import files
     uploaded = files.upload()

Escolher arquivos dataset.csv
• dataset.csv(text/csv) - 45571662 bytes, last modified: 21/10/2024 - 100% done
Saving dataset.csv to dataset.csv

[9] df = pd.read_csv('dataset.csv', sep=',', engine='python')

print("Primeiras linhas:")
print(df.head())

print("\nÚltimas linhas:")
print(df.tail())

Primeiras linhas:
InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country
0 536365,85123A,WHITE HANGING HEART T-LIGHT HOLD...
1 536365,71053,WHITE METAL LANTERN,6,12/1/10 8:2...
2 536365,84406B,CREAM CUPID HEARTS COAT HANGER,8...
3 536365,84029G,KNITTED UNION FLAG HOT WATER BOT...
4 536365,84029E,RED WOOLLY HOTTIE WHITE HEART.,6...

Últimas linhas:
InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country
541904 581587,22613,PACK OF 20 SPACEBOY NAPKINS,12,12...
541905 581587,22899,CHILDREN'S APRON DOLLY GIRL ,6,12...
541906 581587,23254,CHILDRENS CUTLERY DOLLY GIRL ,4,1...
541907 581587,23255,CHILDRENS CUTLERY CIRCUS PARADE,4...
541908 581587,22138,BAKING SET 9 PIECE RETROSPOT ,3,1...
```

2 - Verificando separador e dados.

```
[5] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 1 columns):
#   Column                                                                 Non-Null Count  Dtype
---  -
0   InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country  541909 non-null  object
dtypes: object(1)
memory usage: 4.1+ MB

[6] df_copy = df.copy()

[8] print(df.columns)

Index(['InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country'], dtype='object')

[10] print(df.columns)
print(df.head())

Index(['InvoiceNo', 'StockCode', 'Description', 'Quantity', 'InvoiceDate',
       'UnitPrice', 'CustomerID', 'Country'],
      dtype='object')
InvoiceNo  StockCode  Description  Quantity  \
0    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER      6
1    536365    71053              WHITE METAL LANTERN      6
2    536365    84406B  CREAM CUPID HEARTS COAT HANGER      8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE      6
4    536365    84029E  RED WOOLLY HOTTIE WHITE HEART.      6

InvoiceDate  UnitPrice  CustomerID  Country
0  12/1/10 8:26      2,55    17850.0  United Kingdom
1  12/1/10 8:26      3,39    17850.0  United Kingdom
2  12/1/10 8:26      2,75    17850.0  United Kingdom
3  12/1/10 8:26      3,39    17850.0  United Kingdom
4  12/1/10 8:26      3,39    17850.0  United Kingdom
```

3 - Substituindo valores nulos na coluna 'CustomerID' por 0 e na Coluna 'InvoiceDate' por '1900/01/01'.

```
df['CustomerID'].fillna(0, inplace=True)
print(df.head())

InvoiceNo  StockCode  Description  Quantity  \
0    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER      6
1    536365    71053              WHITE METAL LANTERN      6
2    536365    84406B  CREAM CUPID HEARTS COAT HANGER      8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE      6
4    536365    84029E  RED WOOLLY HOTTIE WHITE HEART.      6

InvoiceDate  UnitPrice  CustomerID  Country
0  12/1/10 8:26      2,55    17850.0  United Kingdom
1  12/1/10 8:26      3,39    17850.0  United Kingdom
2  12/1/10 8:26      2,75    17850.0  United Kingdom
3  12/1/10 8:26      3,39    17850.0  United Kingdom
4  12/1/10 8:26      3,39    17850.0  United Kingdom

[12] df['InvoiceDate'].fillna('1900/01/01', inplace=True)
print(df.head())

InvoiceNo  StockCode  Description  Quantity  \
0    536365    85123A  WHITE HANGING HEART T-LIGHT HOLDER      6
1    536365    71053              WHITE METAL LANTERN      6
2    536365    84406B  CREAM CUPID HEARTS COAT HANGER      8
3    536365    84029G  KNITTED UNION FLAG HOT WATER BOTTLE      6
4    536365    84029E  RED WOOLLY HOTTIE WHITE HEART.      6

InvoiceDate  UnitPrice  CustomerID  Country
0  12/1/10 8:26      2,55    17850.0  United Kingdom
1  12/1/10 8:26      3,39    17850.0  United Kingdom
2  12/1/10 8:26      2,75    17850.0  United Kingdom
3  12/1/10 8:26      3,39    17850.0  United Kingdom
4  12/1/10 8:26      3,39    17850.0  United Kingdom
```

4 - Transformando a coluna 'InvoiceDate' em Datetime, resolvendo erros de formatação na coluna 'InvoiceData' e removendo registros com valores nulos.

```
[13] df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], errors='coerce')
print(df.info())


<ipython-input-13-369128e7e46c>:1: UserWarning: Could not infer format, so each element will be parsed individually, falling back to 'dateutil'. To ensure parsing is consistent and as-expected, please specify a format.
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], errors='coerce')
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541989 entries, 0 to 541988
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   InvoiceNo    541989 non-null  object
1   StockCode   541989 non-null  object
2   Description  540455 non-null  object
3   Quantity    541989 non-null  int64
4   InvoiceDate  541989 non-null  datetime64[ns]
5   UnitPrice   541989 non-null  object
6   CustomerID  541989 non-null  float64
7   Country     541989 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.1+ MB
None

df['InvoiceDate'] = df['InvoiceDate'].replace('1900/01/01', pd.NaT)
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], errors='coerce')

[15] df.dropna(subset=['InvoiceDate'], inplace=True)
print(df.info())

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541989 entries, 0 to 541988
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   InvoiceNo    541989 non-null  object
1   StockCode   541989 non-null  object
2   Description  540455 non-null  object
3   Quantity    541989 non-null  int64
4   InvoiceDate  541989 non-null  datetime64[ns]
5   UnitPrice   541989 non-null  object
6   CustomerID  541989 non-null  float64
7   Country     541989 non-null  object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.1+ MB
None
```

5 - Verificando alterações.

 Nivel3-Mundo5.ipynb ☆

Arquivo Editar Ver Inserir Ambiente de execução Ferramentas Ajuda Todas as alterações foram salvas

Arquivos

..

sample_data

dataset.csv

+ Código + Texto

[15]

1 StockCode 541989 non-null object
2 Description 540455 non-null object
3 Quantity 541989 non-null int64
4 InvoiceDate 541989 non-null datetime64[ns]
5 UnitPrice 541989 non-null object
6 CustomerID 541989 non-null float64
7 Country 541989 non-null object
dtypes: datetime64[ns](1), float64(1), int64(1), object(5)
memory usage: 33.1+ MB
None

print(df.head())

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2,55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3,39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2,75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3,39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3,39	17850.0	United Kingdom