Tarefa 3.6 - Análise Exploratória e Estatística Descritiva - Correlação

Esta tarefa consiste em realizar a análise de correlação de variáveis (atributos) por meio de diagramas de dispersão (scatterplot) e coeficiente de correlação de Pearson utilizando as bibliotecas e funções da linguagem Python.

Você deve implementar em Python e responder os itens solicitados abaixo utilizando a biblioteca Matplotlib e/ou Seaborn e outras que desejar ou forem necessárias.

Clique aqui para acessar um exemplo de notebook que contém as funções necessárias para a realização desta tarefa.

PARTE I: *dataset Advertising*

Este dataset mostra uma série de valores investidos em anúncios de TV, Rádio e jornais e os respectivos resultados de venda.

a-) Ler o dataset "*Advertising/Advertising.csv*" como dataframe utilizando a biblioteca Pandas e mostrar as 5 primeiras e 5 últimas linhas.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

!git clone "https://github.com/malegopc/DSBD"

df = pd.read_csv("DSBD/Datasets/Advertising/Advertising.csv")

print(df.head(5))
print(df.tail(5))
```

```
fatal: destination path 'DSBD' already exists and is not an empty directory.
   Unnamed: 0     TV  radio  newspaper  sales
0           1  230.1   37.8       69.2   22.1
1           2   44.5   39.3       45.1   10.4
2           3   17.2   45.9       69.3    9.3
3           4  151.5   41.3       58.5   18.5
4           5  180.8   10.8       58.4   12.9
     Unnamed: 0     TV  radio  newspaper  sales
195         196   38.2    3.7       13.8    7.6
196         197   94.2    4.9        8.1    9.7
197         198  177.0    9.3        6.4   12.8
198         199  283.6   42.0       66.2   25.5
199         200  232.1    8.6        8.7   13.4
```

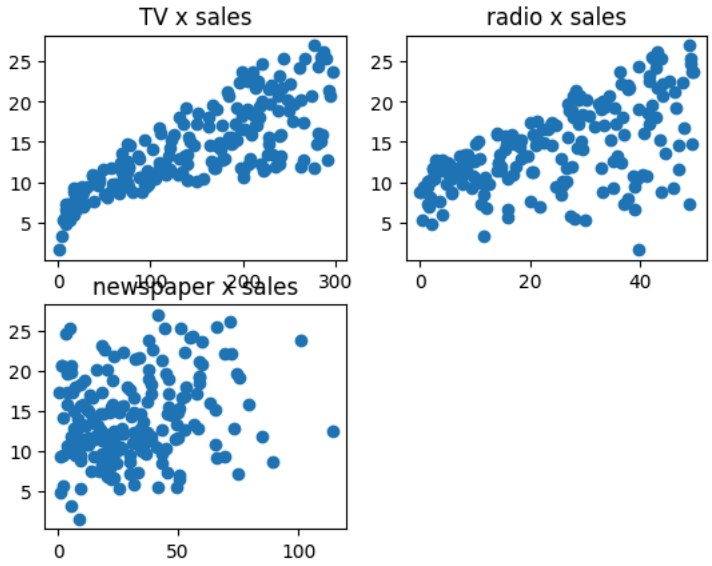
b-) Mostre num mesmo quadro (usando subplots) os três diagramas de dispersão para os pares de variáveis: TV x sales, radio x sales e newspaper x sales.

```
plt.subplot(2,2,1)
plt.title("TV x sales")
plt.scatter(df['TV'], df['sales'])

plt.subplot(2,2,2)
plt.title("radio x sales")
plt.scatter(df['radio'], df['sales'])

plt.subplot(2,2,3)
plt.title("newspaper x sales")
plt.scatter(df['newspaper'], df['sales'])

plt.show()
```



c-) Calcule o coeficiente de correlação de Pearson para os três pares de variáveis: TV x sales, radio x sales e newspaper x sales.

```
dfTV = pd.DataFrame({'TV': df['TV'], 'sales': df['sales']})
print(dfTV.corr())
print("\n")

dfRadio = pd.DataFrame({'Radio': df['radio'], 'sales': df['sales']})
print(dfRadio.corr())
print("\n")

dfNews = pd.DataFrame({'Newspaper': df['newspaper'], 'sales': df['sales']})
print(dfNews.corr())
print("\n")
```

```
              TV     sales
TV     1.000000  0.782224
sales  0.782224  1.000000


           Radio     sales
Radio  1.000000  0.576223
sales  0.576223  1.000000


           Newspaper     sales
Newspaper   1.000000  0.228299
sales       0.228299  1.000000
```

d-) Analisando os diagramas de dispersão acima e os valores dos coeficientes de correlação responda qual dos tipos de anúncios influencia mais e qual deles influencia menos nos resultados das vendas?

Respostas:

Analisando os dados acima, é possível afirmar que a TV é o tipo de anúncio que mais influencia com 0.78 de correlação com as vendas.

Em 2° lugar fica o rádio, com 0.57 de correlação com as vendas.

Em 3° lugar fica o jornal, com 0.22 de correlação com as vendas.

Também é possível observar os mesmos resultados nos gráficos de dispersão.

PARTE II: *dataset Tips*

Este dataset mostra diferentes atributos (valor da conta, gorjeta, número de pessoas, etc.) dos clientes de um restaurante.

e-) Faça a leitura do dataset "tips" disponível na biblioteca seaborn.

```
!git clone "https://github.com/mwaskom/seaborn-data"

tips = pd.read_csv("seaborn-data/tips.csv")
print(tips)
```

```
fatal: destination path 'seaborn-data' already exists and is not an empty directory.
     total_bill   tip     sex smoker   day    time  size
0         16.99  1.01  Female     No   Sun  Dinner     2
1         10.34  1.66    Male     No   Sun  Dinner     3
2         21.01  3.50    Male     No   Sun  Dinner     3
3         23.68  3.31    Male     No   Sun  Dinner     2
4         24.59  3.61  Female     No   Sun  Dinner     4
..          ...   ...     ...    ...   ...     ...   ...
239       29.03  5.92    Male     No   Sat  Dinner     3
240       27.18  2.00  Female    Yes   Sat  Dinner     2
241       22.67  2.00    Male    Yes   Sat  Dinner     2
242       17.82  1.75    Male     No   Sat  Dinner     2
243       18.78  3.00  Female     No  Thur  Dinner     2

[244 rows x 7 columns]
```

f-) Mostre o diagrama de dispersão entre as variáveis "total_bill" e "tip" com a reta de regressão utilizando a função *polyfit* da biblioteca numpy.

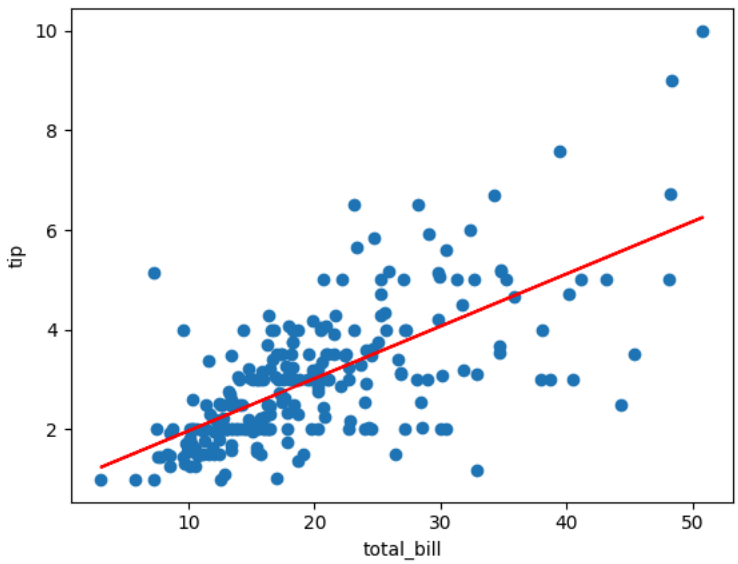
```
import numpy as np

a, b = np.polyfit(tips['total_bill'], tips['tip'], 1)

plt.plot(tips['total_bill'], a * tips['total_bill'] + b, color='red')
plt.scatter(tips['total_bill'], tips['tip'])

plt.xlabel('total_bill')
plt.ylabel('tip')

plt.show()
```



g-) Utilizando os coeficientes a e b da reta de regressão, calcule qual será a gorjeta (tip) estimada para uma conta total (total_bill) igual a um valor fornecido pelos usuário.

```
total = float(input())
gorjeta = a * total + b
print(f"A gorjeta estimada é: {gorjeta}")
```

```
21.01
A gorjeta estimada é: 3.126834723799938
```

h-) Faça o mesmo para estimar o valor total da conta (total_bill) para um valor de gorjeta (tip) fornecido pelo usuário?

```
gorjeta = float(input())
total = (gorjeta - b) / a
print(f"O total estimado é: {total}")
```

```
3.12
O total estimado é: 20.9449225878856
```

i-) Calcule o coeficiente de correlação de Pearson para total_bill e tip.

```
df_total_bill = pd.DataFrame(tips, columns=['total_bill', 'tip'])
df_total_bill.corr()
```

	total_bill	tip
total_bill	1.000000	0.675734
tip	0.675734	1.000000

j-) Responda:

i-) Existe correlação entre 'total_bill' e 'tip'? Resp:

Sim, existe correlação.

ii-) Se sim, a correlação é positiva ou negativa? Resp:

A correlação entre 'total_bill' e 'tip' é positiva.

iii-) Se sim, qual o grau de correlação? Forte, moderada ou fraca? Resp:

A correlação entre 'total_bill' e 'tip' é forte, pois é acima de 0.6.

PARTE III: *dataset Atlas Brasil*

Este dataset mostra diferentes atributos valores de variáveis sociais para cada um dos estados brasileiros:

- ANOSEST = Média de anos de estudo
- T_ANALF25M = Taxa de analfabetismo - 25 anos ou mais
- MORT1 = Mortalidade infantil
- RDPC = Renda per capita
- POPTOT = População total
- IDHM = IDHM

k-) Ler o dataset "*Atlas/Atlas_Brasil_2014.csv*" como dataframe utilizando a biblioteca Pandas

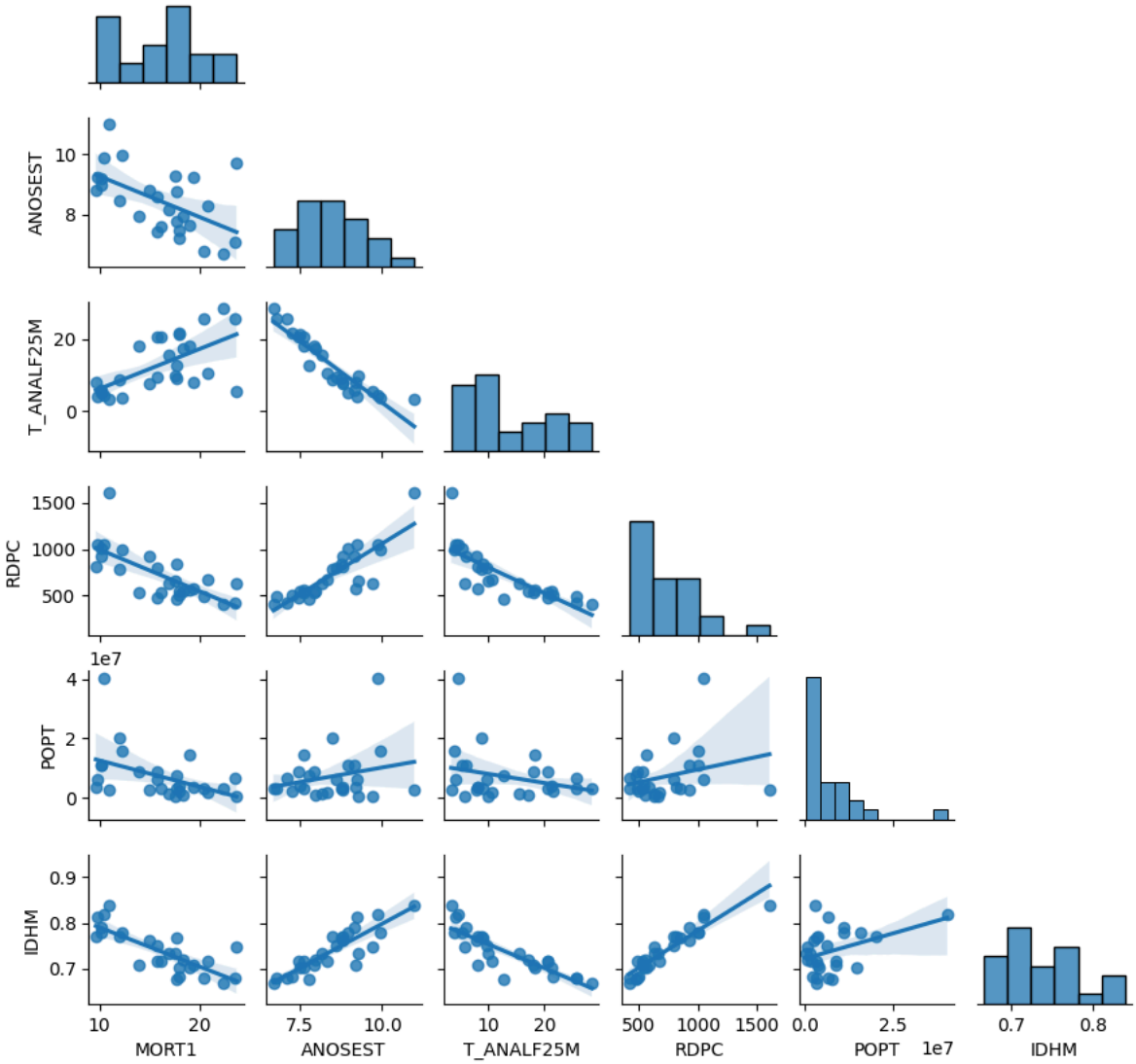
```
atlas = pd.read_csv("DSBD/Datasets/Atlas/Atlas_Brasil_2014.csv")
print(atlas)
```

```
   ESTADO  MORT1  ANOSEST  T_ANALF25M     RDPC      POPT   IDHM
0      RO  20.82     8.31       10.66   667.41   1641072  0.715
1      AC  18.37     7.95       17.19   548.24    762502  0.719
2      AM  19.38     9.21        8.06   579.45   3769444  0.709
3      RR  17.57     9.29        9.95   665.19    472758  0.732
4      PA  17.65     7.78       12.81   469.47   7615746  0.675
5      AP  23.67     9.69        5.67   630.72    741289  0.747
6      TO  16.86     8.16       15.40   628.99   1456641  0.732
7      MA  23.52     7.09       25.63   424.12   6525585  0.678
8      PI  20.37     6.80       25.62   487.90   3178322  0.678
9      CE  15.81     7.44       20.65   481.88   8624448  0.716
10     RN  16.14     7.60       20.67   531.69   3368229  0.717
11     PB  17.97     7.48       21.39   545.34   3844026  0.701
12     PE  13.99     7.94       18.10   529.47   8750240  0.709
13     AL  22.36     6.69       28.46   414.51   3265569  0.667
14     SE  17.94     7.24       21.56   500.83   2156005  0.681
15     BA  18.95     7.63       18.20   558.53  14673318  0.703
16     MG  11.97     8.47        8.67   790.11  20242670  0.769
17     ES   9.64     8.80        8.00   816.08   3678560  0.771
18     RJ  12.29     9.94        3.81   997.62  15635042  0.778
19     SP  10.49     9.87        4.53  1047.38  40350800  0.819
20     PR  10.13     9.17        6.04   926.12  10817027  0.790
21     SC   9.80     9.25        4.06  1042.82   6408973  0.813
22     RS  10.16     8.97        5.30  1000.62  10851074  0.779
23     MS  14.94     8.80        7.77   930.50   2599715  0.762
24     MT  17.71     8.74        9.09   844.21   3141541  0.767
25     GO  15.79     8.60        9.59   796.18   6373595  0.750
26     DF  10.99    10.98        3.35  1606.40   2756412  0.839
```

l-) Utilizando a função "pairplot" da biblioteca seaborn mostre os diagramas de dispersão com suas respectivas retas de regressão entre todos os pares de variáveis.

```
sns.pairplot(atlas, kind ="reg", height=1.5, corner= True)
```

```
<seaborn.axisgrid.PairGrid at 0x7d469f7ccfa0>
```



m-) Obtenha a matriz de correlação entre todos os pares de variáveis.

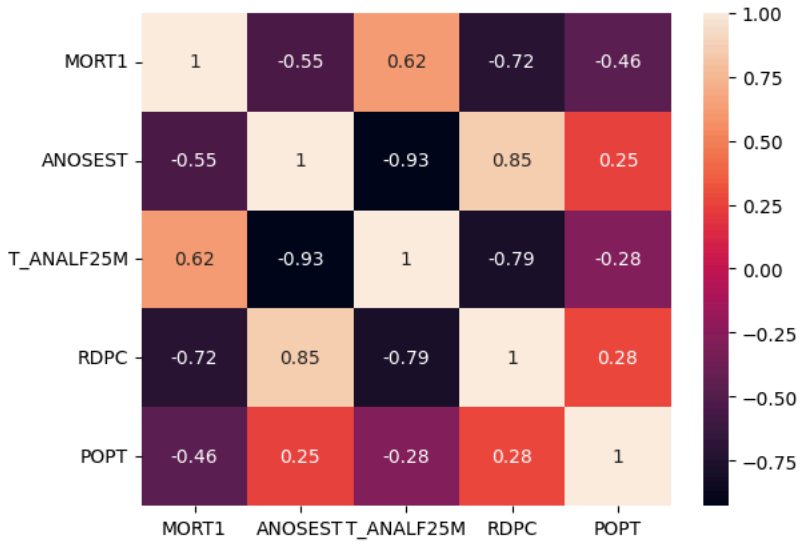
```
atlasCorr = pd.DataFrame(atlas, columns=['MORT1', 'ANOSEST', 'T_ANALF25M', 'RDPC', 'POPT'])
atlasCorr.corr()
```

	MORT1	ANOSEST	T_ANALF25M	RDPC	POPT
MORT1	1.000000	-0.549384	0.619011	-0.716724	-0.460198
ANOSEST	-0.549384	1.000000	-0.927319	0.849124	0.248012
T_ANALF25M	0.619011	-0.927319	1.000000	-0.786432	-0.281324
RDPC	-0.716724	0.849124	-0.786432	1.000000	0.275724
POPT	-0.460198	0.248012	-0.281324	0.275724	1.000000

n-) Mostre um "mapa de calor" (heatmap) baseado na matriz de correlação entre todos os pares de variáveis.

```
corrMatrix = atlasCorr.corr()

sns.heatmap(corrMatrix, annot=True)
plt.show()
```



o-) Analisando os resultados acima, responda:

i-) Quais pares de variáveis são correlacionadas positivamente? Resp:

- Taxa de mortalidade x Analfabetismo;
- Anos de Estudo x Renda perCapita;
- Anos de Estudo x População Total;
- Renda perCapita x População Total;

ii-) Quais pares de variáveis são correlacionadas negativamente? Resp:

- Taxa de Mortalidade x Anos de Estudo;
- Taxa de Mortalidade x Renda perCapita;
- Taxa de Mortalidade x População Total;
- Anos de Estudo x Taxa de Analfabetismo;
- Taxa de Analfabetismo x Renda perCapita;
- Taxa de Analfabetismo x População Total;

iii-) Qual par de variáveis apresenta maior correlação (positiva e negativa)(desconsidere pares com as mesmas variáveis)? Resp:

- Das variáveis correlacionadas negativamente, 'Taxa de Analfabetismo' e 'Anos de Estudo' é a mais correlacionada.
- Das variáveis correlacionadas positivamente, 'Anos de Estudo' e 'Renda perCapita' é a mais correlacionada.