
Tarefa 3.4 - Análise Exploratória e Estatística Descritiva - Medidas estatísticas utilizando Python

Esta tarefa consiste analisar por meio de métodos de descrição paramétrica (medidas estatísticas) o conjunto de dados *Hourly Wages*.

Você deve implementar em Python os itens solicitados abaixo utilizando a biblioteca Pandas e outras que forem necessárias.

Conjunto de dados *Hourly Wages*

Esse conjunto de dados é popularmente empregado para a tarefa de regressão em Machine Learning. Constrói-se e ajusta-se um modelo de aprendizado de máquina para predizer/estimar o salário de um funcionário em função de suas características (anos de estudos, experiência de trabalho, filiação sindical, região, ocupação e sexo).

- Número de Instâncias: 534
- Número de Atributos: 9 atributos numéricos e o target (*wage_per_hour*)
- Informações dos Atributos:
 - union (filiação sindical)
 - education_yrs (anos de instrução)
 - experience_yrs (anos de experiência)
 - age (idade)
 - female (sexo)
 - marr (estado civil - casado)
 - south (região)
 - manufacturing (indústria)
 - construction (construção)
- Variável de destino (target): *wage_per_hour*

✓ Implementar os itens abaixo:

- ✓ a-) Clonar o repositório de dados da disciplina (DSBD) hospedado no GitHub.

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

```
fatal: destination path 'DSBD' already exists and is not an empty directory.
```

- ✓ b-) Ler o dataset "*Hourly_wages/hourly_wages.csv*" como dataframe utilizando a biblioteca Pandas e mostrar as 5 primeiras e as 5 últimas linhas do dataset.

```
import pandas as pd
import numpy as np
```

```
df = pd.read_csv("DSBD/Datasets/Hourly_wages/hourly_wages.csv", na_values=['?'])
```

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

```

wage_per_hour  union  education_yrs  experience_yrs  age  female  marr  \
0             5.10     0              8              21    35      1      1
1             4.95     0              9              42    57      1      1
2             6.67     0             12               1    19      0      0
3             4.00     0             12               4    22      0      0
4             7.50     0             12              17    35      0      1

south  manufacturing  construction
0      0              1              0
1      0              1              0
2      0              1              0
3      0              0              0
4      0              0              0

wage_per_hour  union  education_yrs  experience_yrs  age  female  marr  \
529          11.36     0             18               5    29      0      0
530           6.10     0             12              33    51      1      1
531          23.25     1             17              25    48      1      1
```

532	19.88	1	12	13	31	0	1
533	15.38	0	16	33	55	0	1

	south	manufacturing	construction
529	0	0	0
530	0	0	0
531	0	0	0
532	1	0	0
533	0	1	0

- ✓ c-) Obter um resumo da estatística descritiva dos dados utilizando um único comando (função).

```
df.describe()
```

```

count    wage_per_hour    union    education_yrs    experience_yrs    age    femal
mean      9.024064    0.179775    13.018727    17.822097    36.833333    0.45880
std       5.139097    0.384360    2.615373    12.379710    11.726573    0.49876
min       1.000000    0.000000    2.000000    0.000000    18.000000    0.00000
25%       5.250000    0.000000    12.000000    8.000000    28.000000    0.00000
50%       7.780000    0.000000    12.000000    15.000000    35.000000    0.00000
75%      11.250000    0.000000    15.000000    26.000000    44.000000    1.00000
max      44.500000    1.000000    18.000000    55.000000    64.000000    1.00000

```

- ✓ d-) Calcule a proporção de funcionários de sexo masculino e feminino.

```
df.female.value_counts()
```

```

count
female
0      289
1      245

dtype: int64

```

- ✓ e-) Calcule a proporção de casados e solteiros.

```
df.marr.value_counts()
```

```

count
marr
1      350
0      184

dtype: int64

```

- ✓ f-) Calcule as médias de cada uma das variáveis: anos de educação, anos de experiência e idade

```

print('Média da variável anos de educação: ',df['education_yrs'].mean())
print('Média da variável anos de experiência: ',df['experience_yrs'].mean())
print('Média da variável idade: ', df['age'].mean())

```

```

Média da variável anos de educação: 13.0187265917603
Média da variável anos de experiência: 17.822097378277153
Média da variável idade: 36.833333333333336

```

- ✓ g-) Calcule as medianas de cada uma das variáveis: anos de educação, anos de experiência e idade.

```

print('Mediana da variável anos de educação: ',df['education_yrs'].median())
print('Mediana da variável anos de experiência: ',df['experience_yrs'].median())
print('Mediana da variável idade: ', df['age'].median())

```

```
➦ Mediana da variável anos de educação: 12.0
Mediana da variável anos de experiência: 15.0
Mediana da variável idade: 35.0
```

- ✓ h-) Calcule as modas de cada uma das variáveis: anos de educação, anos de experiência e idade.

```
print('Moda da variável anos de educação: ', df['education_yrs'].mode())
print('Moda da variável anos de experiência: ', df['experience_yrs'].mode())
print('Moda da variável idade: ', df['age'].mode())
```

```
➦ Moda da variável anos de educação: 0 12
Name: education_yrs, dtype: int64
Moda da variável anos de experiência: 0 14
Name: experience_yrs, dtype: int64
Moda da variável idade: 0 32
Name: age, dtype: int64
```

- ✓ i-) Analisando especificamente a variável **idade**, responda: qual é a idade do conjunto de dados em que **70%** dos funcionários apresentam idade inferior?

```
print('70% dos funcionários tem idade igual ou inferior a:', np.percentile(df['age'], 70))
```

```
➦ 70% dos funcionários tem idade igual ou inferior a: 42.0
```

- ✓ j-) Analisando especificamente a variável **anos de experiência**, responda: qual é a quantidade de anos de experiência no conjunto de dados em que 10% dos funcionários ainda não possuem?

```
yrs_exp = np.percentile(df['experience_yrs'], 10)
print(f'10% dos funcionários ainda não alcançaram {yrs_exp} anos de experiência!')
```

```
➦ 10% dos funcionários ainda não alcançaram 3.0 anos de experiência!
```

- ✓ k-) Calcule o desvio-padrão de cada um dos atributos: anos de educação, anos de experiência e idade.

```
print('Desvio Padrão da variável anos de educação:', df['education_yrs'].std())
print('Desvio Padrão da variável anos de experiência:', df['experience_yrs'].std())
print('Desvio Padrão da variável idade:', df['age'].std())
```

```
➦ Desvio Padrão da variável anos de educação: 2.6153726283543635
Desvio Padrão da variável anos de experiência: 12.379710087848084
Desvio Padrão da variável idade: 11.726572722555636
```

- ✓ l-) Determine qual dos três atributos apresenta maior variabilidade: anos de educação, anos de experiência ou idade? Justifique a sua resposta.

```
DP_edu_yrs = df['education_yrs'].std()
DP_exp_yrs = df['experience_yrs'].std()
DP_age = df['age'].std()
```

```
if DP_edu_yrs > DP_exp_yrs and DP_edu_yrs > DP_age:
    print("A variável education_yrs possui maior variabilidade, com ", DP_edu_yrs)
```

```
if DP_exp_yrs > DP_edu_yrs and DP_exp_yrs > DP_age:
    print("a variável experience_yrs possui maior variabilidade, com ", DP_exp_yrs)
```

```
if DP_age > DP_edu_yrs and DP_age > DP_exp_yrs:
    print("A variável age possui maior variabilidade, com ", DP_age)
```

```
➦ a variável experience_yrs possui maior variabilidade, com 12.379710087848084
```

- ✓ m-) Leia o conjunto de dados [Atlas_Brasil_2014.csv](#)

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

- n-) Calcule a média do IDHM.

```
print("A média do IDHM é: ", data['IDHM'].mean())
```

 A média do IDHM é: 0.7376296296296296

o-) Qual é a taxa de mortalidade em que 50% dos estados apresentam taxa inferior?

```
print("50% dos estados apresentam taxa igual ou inferior de :", data['MORT1'].quantile(0.5))
```

 50% dos estados apresentam taxa igual ou inferior de : 16.86

p-) Calcule qual o valor de RDPC (renda per capita) em que 25% dos estados apresentam valor superior.

```
print("25% dos estados apresentam valor de RDPC acima de :", data['RDPC'].quantile(0.25))
```

 25% dos estados apresentam valor de RDPC acima de : 530.58

q-) Determine dentre as duas variáveis: T_ANALF25M (Taxa de analfabetismo - 25 anos ou mais) ou RDPC (Renda per capita média), qual apresenta maior variabilidade. Justifique a sua resposta.

```
DP_TA = data["T_ANALF25M"].std()
```

```
DP_RDPC = data["RDPC"].std()
```

```
if DP_TA > DP_RDPC:
```

```
    print("A variável T_ANALF25M possui maior variabilidade, com ", DP_TA)
```

```
else:
```

```
    print("A variável RDPC possui maior variabilidade, com ", DP_RDPC)
```

 A variável RDPC possui maior variabilidade, com 268.68154446939883