
Tarefa 3.5 - Análise Exploratória e Estatística Descritiva - Visualização Gráfica

Esta tarefa consiste em explorar por meio de métodos de descrição gráfica (visualização de dados) o conjunto de dados *Hourly Wages*.

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.

Exemplos dos gráficos solicitados estão nos links das bibliotecas acima e também no notebook Visualização_gráfica.ipynb.

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.

Obs: Você pode ignorar este item caso realize a tarefa no Jupyter Notebook. O conjunto de dados Hourly Wages pode ser baixado clicando aqui.

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

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

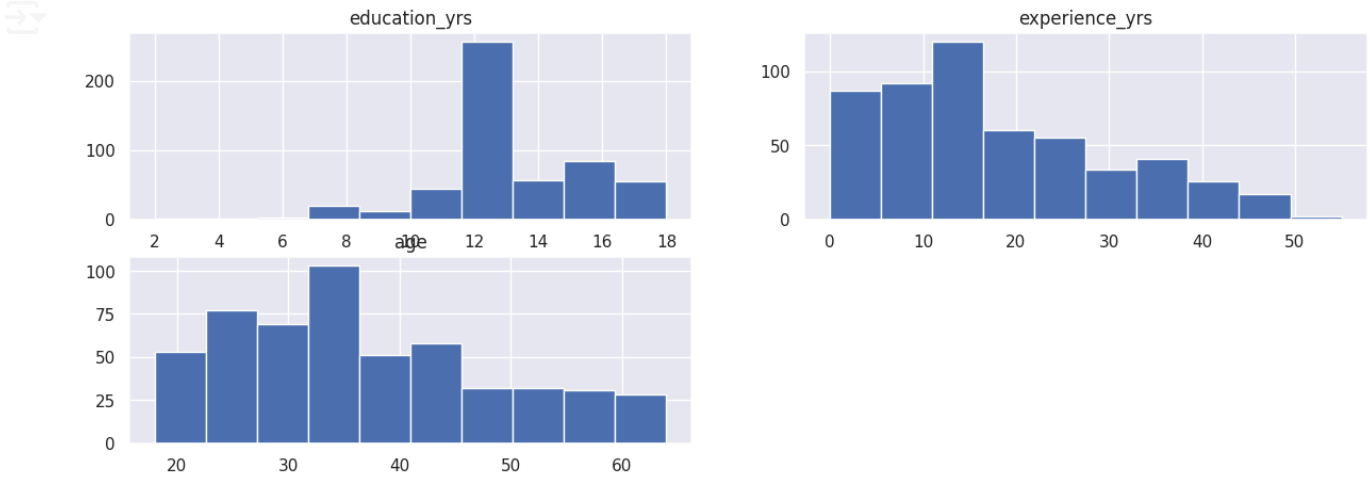
b-) Ler o dataset "*Hourly_wages/hourly_wages.csv*" como dataframe utilizando a biblioteca Pandas e mostrar as 5 primeiras e 5 últimas linhas.

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


df = pd.read_csv("DSBD/Datasets/Hourly_wages/hourly_wages.csv")
df = df.dropna()

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
```

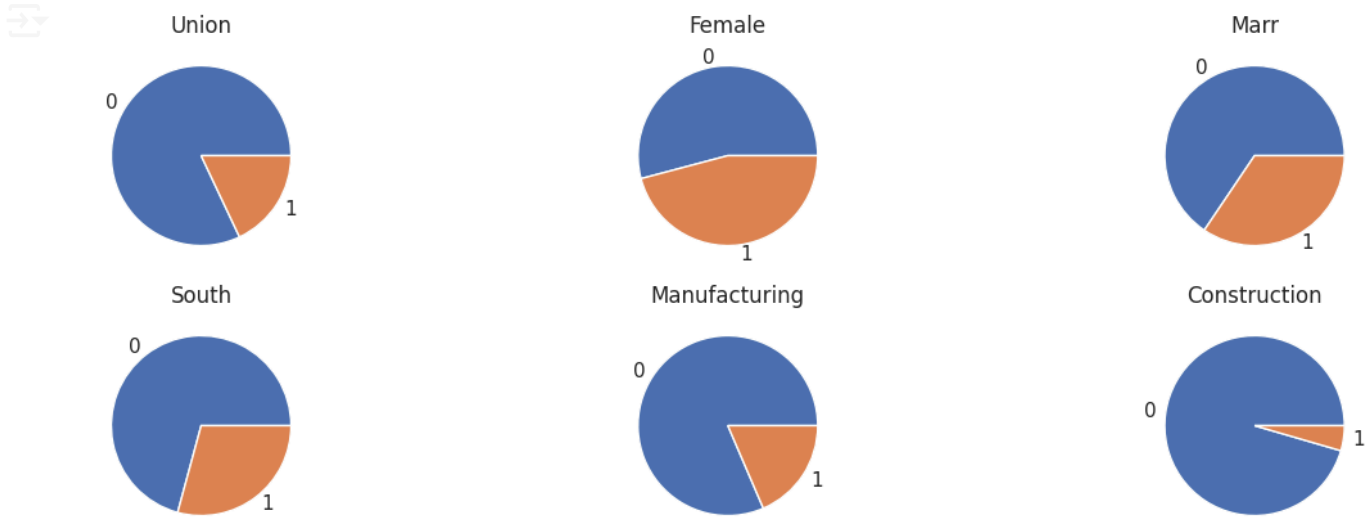
c-) Mostrar os histogramas dos atributos: *education_yrs*, *experience_yrs* e *age* num mesmo quadro de plotagem (usando *subplots*).

```
plt.subplot(2,2,1)
plt.title('education_yrs')
plt.hist(df['education_yrs'])

plt.subplot(2,2,2)
plt.title('experience_yrs')
plt.hist(df['experience_yrs'])

plt.subplot(2,2,3)
plt.title('age')
plt.hist(df['age'])

plt.show()
```



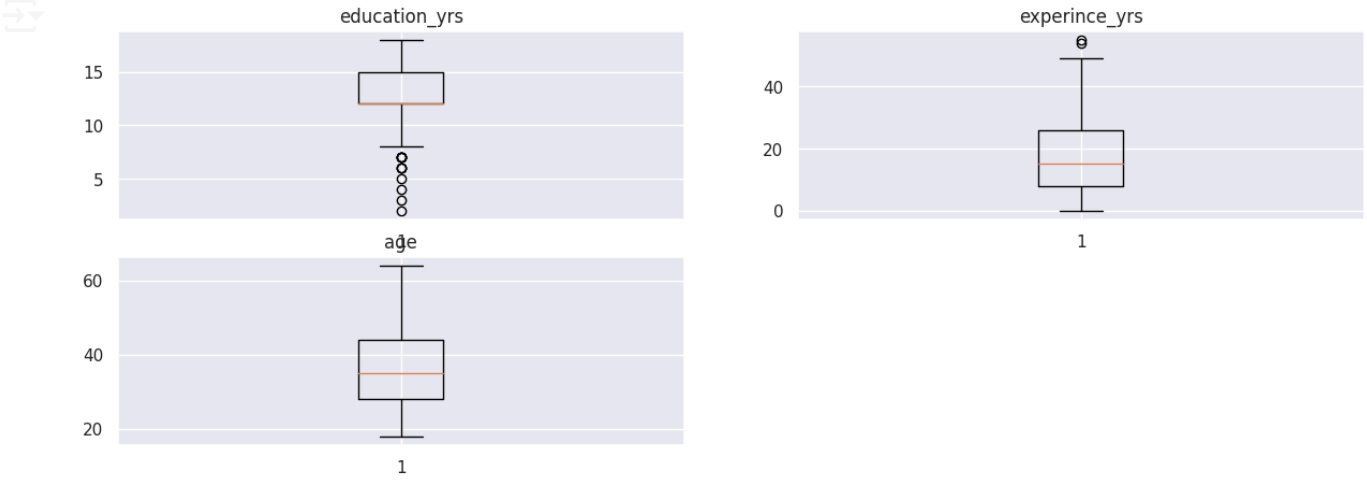
d-) Mostrar os gráficos de setores (*pie chart*) de cada um dos atributos: *union*, *female*, *marr*, *south*, *manufacturing* e *construction* num mesmo quadro de plotagem (usando subplots).

```
plt.subplot(2,3,1)

label = ['0', '1']

plt.pie(df['union'].value_counts(), labels = label)
plt.title('Union')

plt.subplot(2,3,2)
plt.pie(df['female'].value_counts(), labels = label)
plt.title('Female')

plt.subplot(2,3,3)
plt.pie(df['marr'].value_counts(), labels = label)
plt.title('Marr')

plt.subplot(2,3,4)
plt.pie(df['south'].value_counts(), labels = label)
plt.title('South')

plt.subplot(2,3,5)
plt.pie(df['manufacturing'].value_counts(), labels = label)
plt.title('Manufacturing')

plt.subplot(2,3,6)
plt.pie(df['construction'].value_counts(), labels = label)
plt.title('Construction')

plt.show()
```



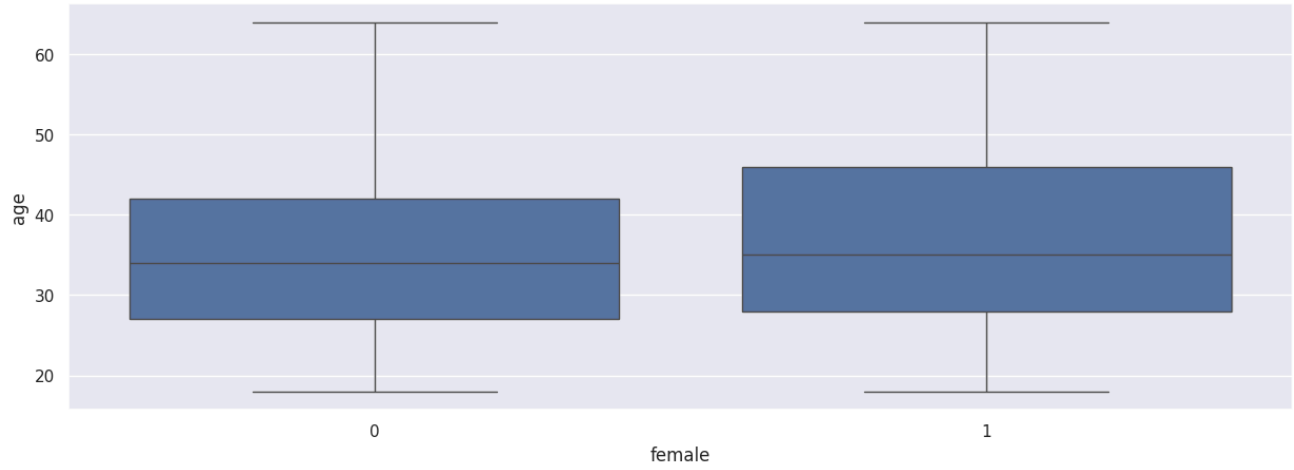
e-) Mostrar os boxplots das variáveis *education_yrs*, *experience_yrs* e *age* num mesmo quadro de plotagem (usando *subplots*).

```
plt.subplot(2,2,1)
plt.title('education_yrs')
plt.boxplot(df['education_yrs'])

plt.subplot(2,2,2)
plt.title('experince_yrs')
plt.boxplot(df['experience_yrs'])

plt.subplot(2,2,3)
plt.title('age')
plt.boxplot(df['age'])

plt.show()
```



f-) Mostre lado a lado os boxplots da variável *age* discriminados por sexo (female).

```
sns.boxplot(x = 'female', y = 'age', data=df)
plt.show()
```



g-) Responda observando os boxplots construídos no item anterior:

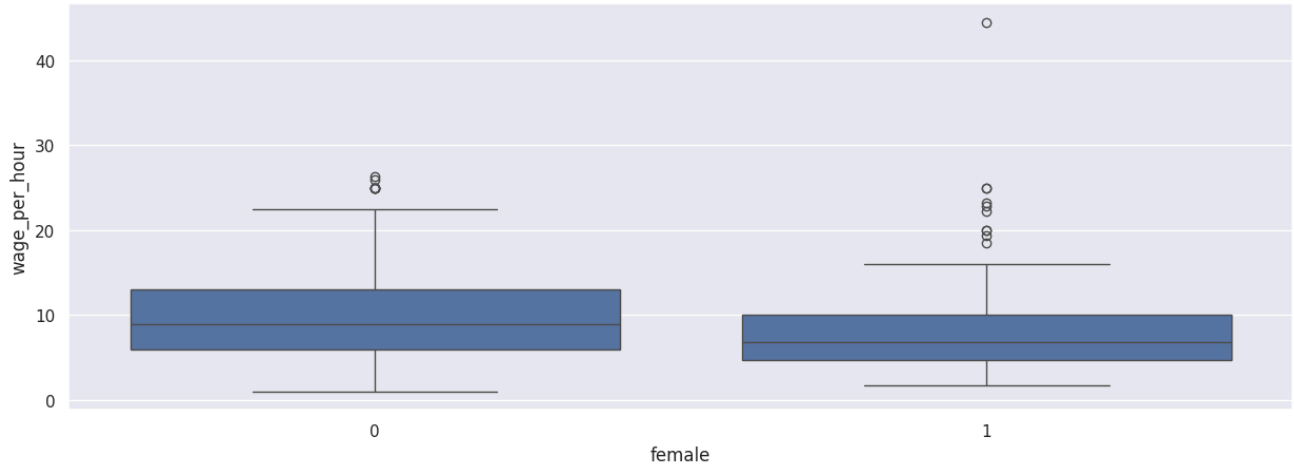
i-) Qual sexo apresenta maior variabilidade (dispersão) de idade? Resposta: **A mulher apresenta maior variabilidade de idade.**

ii-) Qual é aproximadamente o 2o. quartil das idades das mulheres? Resposta: **Aproximamente 35 anos.**

iii-) Qual é aproximadamente o 1o. quartil das idades dos homens? Resposta: **Aproximadamente 28 anos.**

h-) Mostre lado a lado os boxplots da variável *wage_per_hour* discriminados por sexo (female).

```
sns.boxplot(x = 'female', y = 'wage_per_hour', data = df)
plt.show()
```

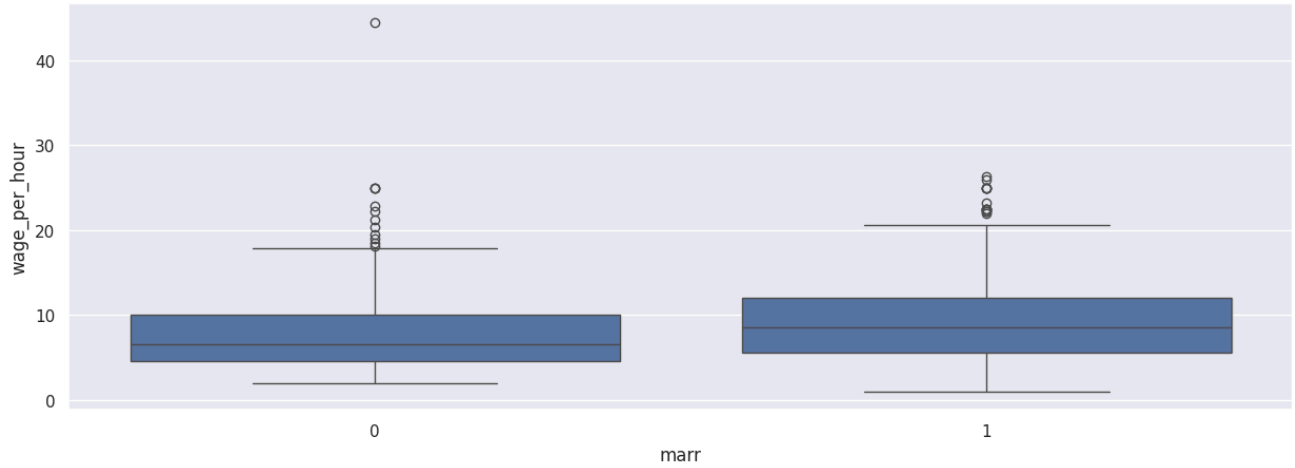


i-) Análise comparativamente os boxplots do item anterior e comente as informações que podem ser exploradas a respeito dos salários por hora de homens e mulhores. Quais conclusões podemos obter?

Resposta:**Analisando o boxplot anterior, é possível afirmar que os homens possuem uma remuneração por hora maiores que a das mulheres, onde a média nos homens é de aproximadamente 9, enquanto o das mulheres é aproximadamente 7. Também é possível observar que as mulheres possuem mais outliers que os homens.**

j-) Mostre lado a lado os boxplots da variável wage_per_hour discriminados por estado civil (marr).

```
sns.boxplot(x = 'marr', y = 'wage_per_hour', data = df)
plt.show()
```



k-) Análise comparativamente os boxplots do item anterior e comente as informações que podem ser exploradas a respeito dos salários por hora de casados e não casados. Quais conclusões podemos obter?

Resposta: **Analisando o boxplot anterior, é possível observar que em média, as pessoas casadas recebem mais e possuem maior variabilidade no salário por hora, quando comparados com os solteiros.**