Professor: Bruno Pimentel

Aluno: Kamila de Almeida Benevides

1) Qual a diferença entre Big Data e Ciência de Dados? (0,5 ponto)

Big Data é um dos recursos usados na área de ciência de dados, especificamente lidando com as tecnologias para coletar, armazenar e pré-processar esses dados. Já a ciência de dados irá usar trabalhar na modelagem desses dados para tomada de decisão.

2) De que forma Estatística, Mineração de Dados e Aprendizagem de Máquina interagem com Ciência de Dados? (1 ponto)

Estatística, mineração de dados e aprendizagem de máquina são uma ferramenta da ciência de dados, para analisar, modelar e extrair dados de uma determinada base de dados a fim de obter Insights para possíveis tomadas de decisões.

3) Mostre a importância do conhecimento de domínio para o cientista de dados. (0,5 ponto)

É importante ter o domínio da área onde se vai aplicar os conhecimentos de ciência de dados, para que as soluções possam melhor aplicadas no contexto do problema. Ter um domínio da área onde será aplicada é importante para não encontrar soluções não viáveis.

4) Crie um conjunto de dados com duas variáveis V1 e V2, tal que:

a) Mediana de V1 < Média de V1 (0,5 ponto)

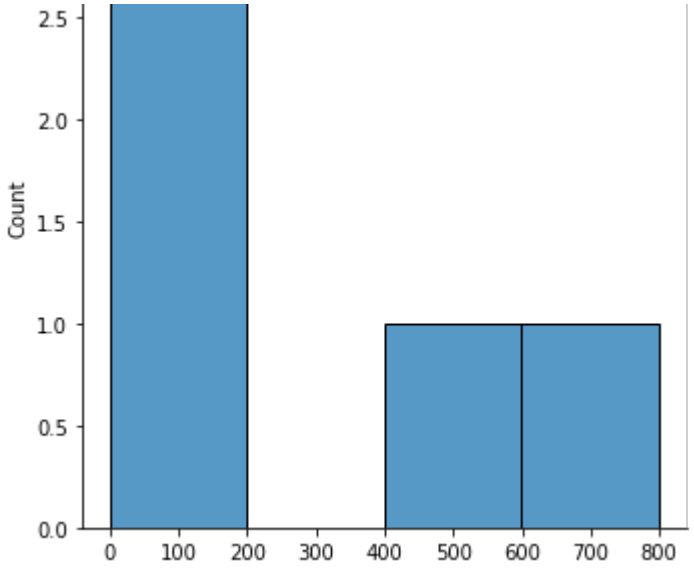
b)Mediana de V2 > Média de V2 (0,5 ponto)

```
#A
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

def mediana(x):
    x.sort()
    n = len(x)
    if n % 2 == 1:
        return x[n // 2]
    else:
        return (x[n // 2 - 1] + x[n // 2]) / 2

V1 = [1, 15, 20, 578, 799]
print("Media: ",np.mean(s),"\nMediana: ", mediana(s))

sns.displot(V1)
plt.show()
```



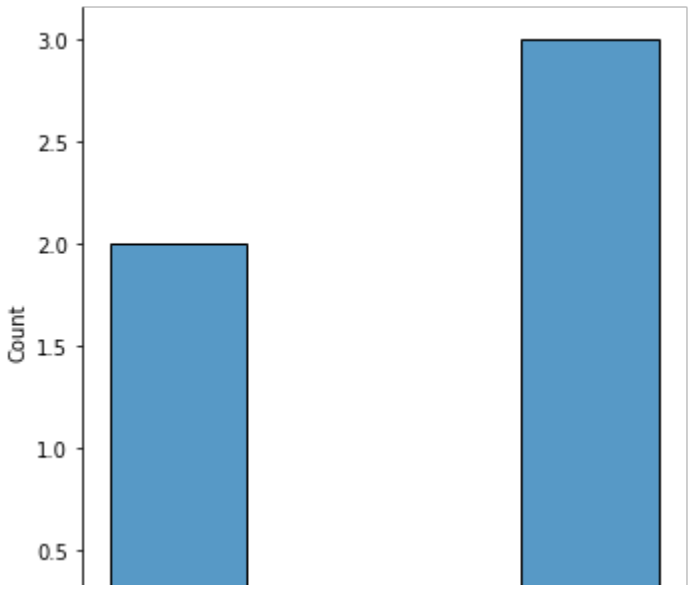

```
#B
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

def mediana(x):
    x.sort()
    n = len(x)
    if n % 2 == 1:
        return x[n // 2]
    else:
        return (x[n // 2 - 1] + x[n // 2]) / 2

V2 = [10, 9, 17, 16, 15]
print("Media: ",np.mean(V2),"\nMediana: ", mediana(V2))


sns.displot(V2)
plt.show()
```

```
Media:  13.4
Mediana:  15
```



```
sat = NormalDist(1060, 195)
```

```
---------------------------------------------------------------------------
ImportError                               Traceback (most recent call
last)
<ipython-input-56-da4c4903f76e> in <module>()
----> 1 from statistics import NormalDist
      2
      3 sat = NormalDist(1060, 195)

ImportError: cannot import name 'NormalDist' from 'statistics' (/usr/lib
/python3.7/statistics.py)

---------------------------------------------------------------------------
NOTE: If your import is failing due to a missing package, you can
manually install dependencies using either !pip or !apt.

To view examples of installing some common dependencies, click the
"Open Examples" button below.
---------------------------------------------------------------------------
```

5) Baseando-se no conjunto de dados criado na questão 4, crie uma função em Python que:

a)Mostra o histograma de cada variável; (1 ponto)

b) Verifica se as variáveis seguem uma distribuição Normal (use teste de hipótese) (1 ponto)

```
#B
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import math

V1 = [1, 15, 20, 578, 799]

mediaV1 = np.mean(V1)
sigma = 0.1

s = np.random.normal(mediaV1, sigma, 1000)
sns.histplot(s)
```

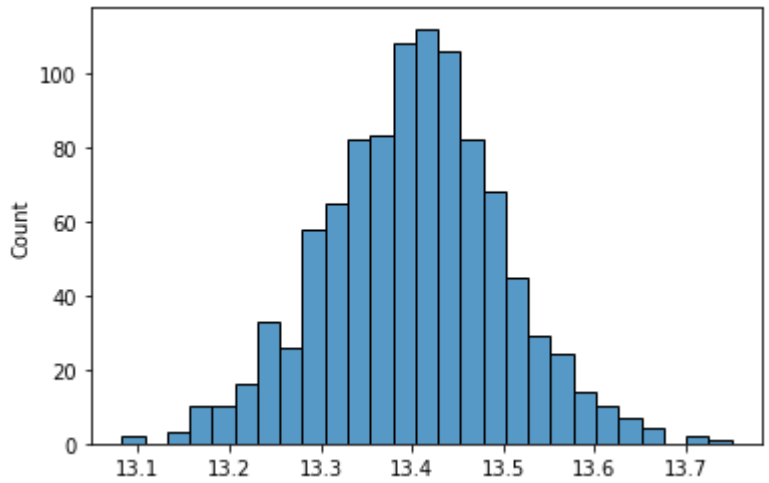
```
import seaborn as sns
import math

V2 = [10, 9, 17, 16, 15]

mediaV2 = np.mean(V2)
sigma = 0.1

s = np.random.normal(mediaV2, sigma, 1000)
sns.histplot(s)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fbcf8f7c8d0>
```



6) Cite 2 técnicas para remoção de ruídos e, para cada uma, mostre uma vantagem e uma desvantagem. (1 ponto)

```
Regressão, estabelecendo uma função de r
podemos eliminar os dados que estiverem
discrepantes dessa reta, a desvantagens
```

Regressão, estabelecendo uma função de regressão, podemos eliminar os dados que