seminario

October 3, 2022

1 Refazendo Seminario em Python

[ ]:
```
#Importando Bibliotecas
import warnings
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.stats.proportion import proportions_ztest
from statsmodels.stats.proportion import proportion_confint
```

2 Introdução

No dia 10 de abril de 1912, o maior transatlântico já construído até então zarpava do porto de Southampton, na Inglaterra, com destino a Nova York, em sua viagem inaugural. Na madrugada do dia 14 para o dia 15 de abril, o luxuoso transatlântico colidiu com um iceberg no Atlântico Norte por volta da meia-noite, afundando em menos de três horas.

Foram 1.517 pessoas mortas e 706 sobreviventes, dos 2.223 passageiros e tripulantes, de acordo com o relatório do Senado dos Estados Unidos sobre o desastre.

[ ]:
```
Base = pd.read_csv('train.csv')              #Lendo a Base de dados e atribuindo␣
 ↪ao objeto Base
Base['Age'] = Base['Age'].fillna(Base['Age'].median())          #Preenchendo␣
 ↪NAs com a mediana na coluna Age
Base['Sex'] = np.where(Base.Sex == 'male',0,1)
```

Para esse trabalho vamos utilizar a base de dados do titanic do kaggle.

[ ]:
```
Base.head(10)       #Mostrando 20 linhas da base
```

[ ]:

	PassengerId	Survived	Pclass	\
0	1	0	3	
1	2	1	1	
2	3	1	3	
3	4	1	1	
4	5	0	3	
5	6	0	3	

```
6            7         0       1
7            8         0       3
8            9         1       3
9           10         1       2
```

```
                                                 Name  Sex   Age  SibSp  Parch  \
0                             Braund, Mr. Owen Harris    0  22.0      1      0
1  Cumings, Mrs. John Bradley (Florence Briggs Th…    1  38.0      1      0
2                              Heikkinen, Miss. Laina    1  26.0      0      0
3        Futrelle, Mrs. Jacques Heath (Lily May Peel)    1  35.0      1      0
4                            Allen, Mr. William Henry    0  35.0      0      0
5                                    Moran, Mr. James    0  28.0      0      0
6                             McCarthy, Mr. Timothy J    0  54.0      0      0
7                      Palsson, Master. Gosta Leonard    0   2.0      3      1
8  Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)    1  27.0      0      2
9                Nasser, Mrs. Nicholas (Adele Achem)    1  14.0      1      0

             Ticket     Fare Cabin Embarked
0         A/5 21171   7.2500   NaN        S
1          PC 17599  71.2833   C85        C
2  STON/O2. 3101282   7.9250   NaN        S
3            113803  53.1000  C123        S
4            373450   8.0500   NaN        S
5            330877   8.4583   NaN        Q
6             17463  51.8625   E46        S
7            349909  21.0750   NaN        S
8            347742  11.1333   NaN        S
9            237736  30.0708   NaN        C
```

2.0.1 Informações sobre a base de dados

Homens = 0 e Mulheres = 1

Sobrevivente = 1 e Morto = 0

[ ]:
```
Base.describe()          #Descrição da base de dados com algumas informações.
```

[ ]:
```
       PassengerId    Survived      Pclass         Sex         Age  \
count   891.000000  891.000000  891.000000  891.000000  891.000000
mean    446.000000    0.383838    2.308642    0.352413   29.361582
std     257.353842    0.486592    0.836071    0.477990   13.019697
min       1.000000    0.000000    1.000000    0.000000    0.420000
25%     223.500000    0.000000    2.000000    0.000000   22.000000
50%     446.000000    0.000000    3.000000    0.000000   28.000000
75%     668.500000    1.000000    3.000000    1.000000   35.000000
max     891.000000    1.000000    3.000000    1.000000   80.000000

            SibSp       Parch        Fare
```

```
count  891.000000  891.000000  891.000000
mean     0.523008    0.381594   32.204208
std      1.102743    0.806057   49.693429
min      0.000000    0.000000    0.000000
25%      0.000000    0.000000    7.910400
50%      0.000000    0.000000   14.454200
75%      1.000000    0.000000   31.000000
max      8.000000    6.000000  512.329200
```

3 Histograma de Sobreviventes, Classes e Sexo

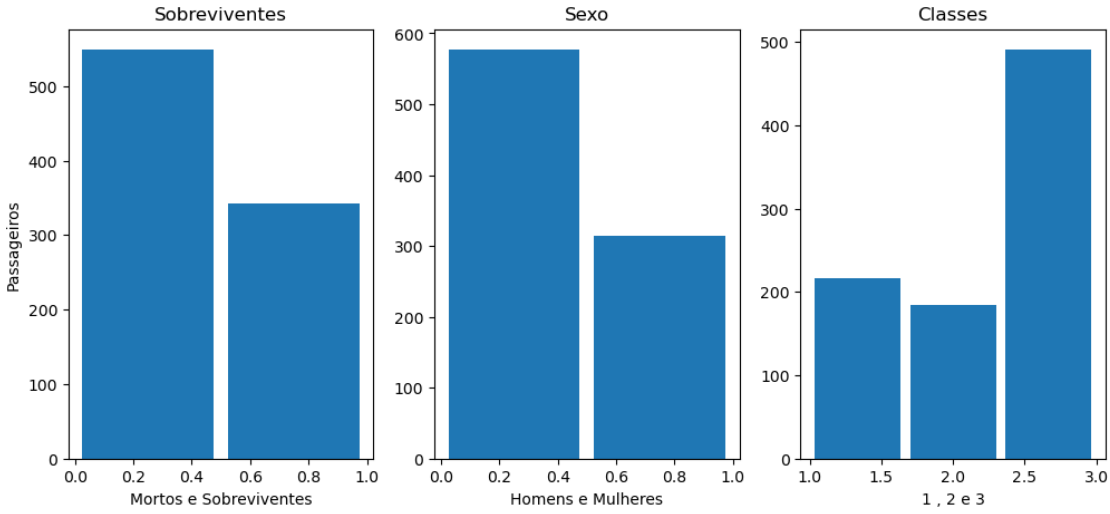
[ ]:
```
fig, ax = plt.subplots(1,3,figsize=(12,5))          #Criando figura com␣
 ↪subplots(1 linha e 3 colunas)

ax[0].set_title('Sobreviventes')                    #Título do subplot
ax[0].set_ylabel('Passageiros')                     #Nome do eixo Y
ax[0].set_xlabel('Mortos e Sobreviventes')          #Nome do eixo X
ax[0].hist(Base.Survived,2,rwidth=0.9)              #Histograma de␣
 ↪Sobreviventes com 2 colunas

ax[1].set_title('Sexo')                             #Título do subplot
ax[1].set_xlabel('Homens e Mulheres')               #Nome do eixo X
ax[1].hist(Base.Sex,2,rwidth=0.9)                   #Histograma de Sexo com 2␣
 ↪colunas

ax[2].set_title('Classes')                          #Título do subplot
ax[2].set_xlabel('1 , 2 e 3')                       #Nome do eixo X
ax[2].hist(Base.Pclass,3,rwidth=0.9)                #Histograma de Classes com␣
 ↪3 colunas
```

[ ]:
```
(array([216., 184., 491.]),
 array([1.        , 1.66666667, 2.33333333, 3.        ]),
 <BarContainer object of 3 artists>)
```



No primeiro histograma é possível observar a distribuição da quantidade de pessoas que sobreviveram e que morreram. Como foi utilizado o sistema binário para definir sobreviventes como 1 e mortos como 0 (eixo x), é possível perceber através do gráfico que a quantidade de pessoas mortas foi maior do a de sobreviventes (eixo y).

No segundo histograma é possível observar a quantidade de pessoas que estavam a bordo (eixo y) separadas por classes (1, 2 e 3 no eixo x). Percebe-se através do gráfico que a maior parte das pessoas que estavam no Titanic eram da Terceira, Primeira e Segunda classe, respectivamente. Sendo a Terceira classe predominantemente maior com aproximadamente 500 pessoas, sendo pelo menos o dobro de alguma das demais classes.

O terceiro histograma mede a quantidade de pessoas segmentada pelos sexos feminino e masculino. Foi utilizado o sistema binário para definir Homens como 0 e Mulheres como 1. Foi possível interpreter que a quantidade de homens a bordo era superior do que a de mulheres.

4 Boxplot de Idade e Fare

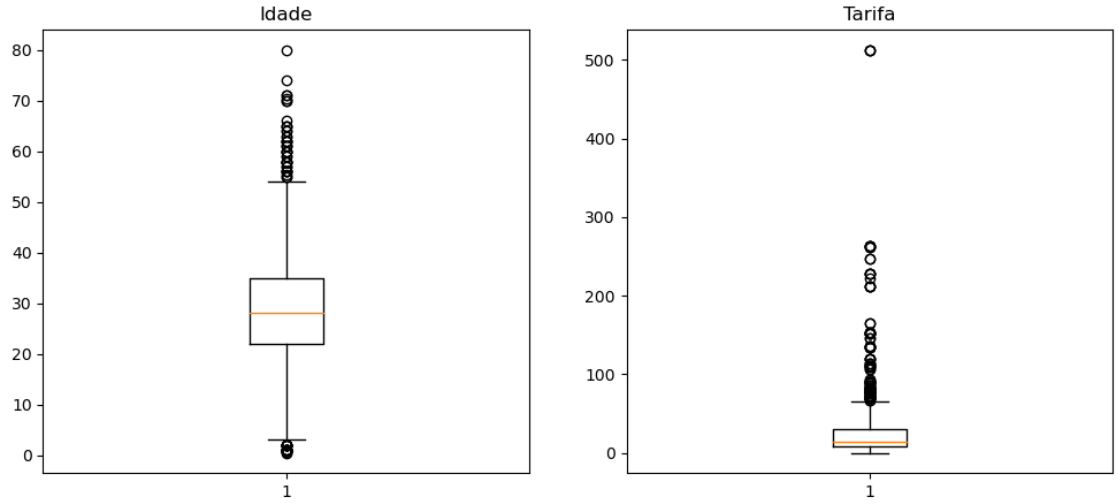
[ ]:
```
fig1, ax1 = plt.subplots(1,2,figsize=(12,5))          #Criando figura com␣
 ↪subplots(1 linha e 2 colunas)

ax1[0].set_title('Idade')                             #Título
ax1[0].boxplot(Base.Age)                              #Criando boxplot da␣
 ↪coluna Age

ax1[1].set_title('Tarifa')                            #Título
ax1[1].boxplot(Base.Fare)                             #Criando boxplot da␣
 ↪coluna Fare
```

[ ]:
```
{'whiskers': [<matplotlib.lines.Line2D at 0x1d152d6f6d0>,
  <matplotlib.lines.Line2D at 0x1d152d6f970>],
```

```
'caps': [<matplotlib.lines.Line2D at 0x1d152d6fc10>,
 <matplotlib.lines.Line2D at 0x1d152d6feb0>],
'boxes': [<matplotlib.lines.Line2D at 0x1d152d6f430>],
'medians': [<matplotlib.lines.Line2D at 0x1d152d7d190>],
'fliers': [<matplotlib.lines.Line2D at 0x1d152d7d430>],
'means': []}
```

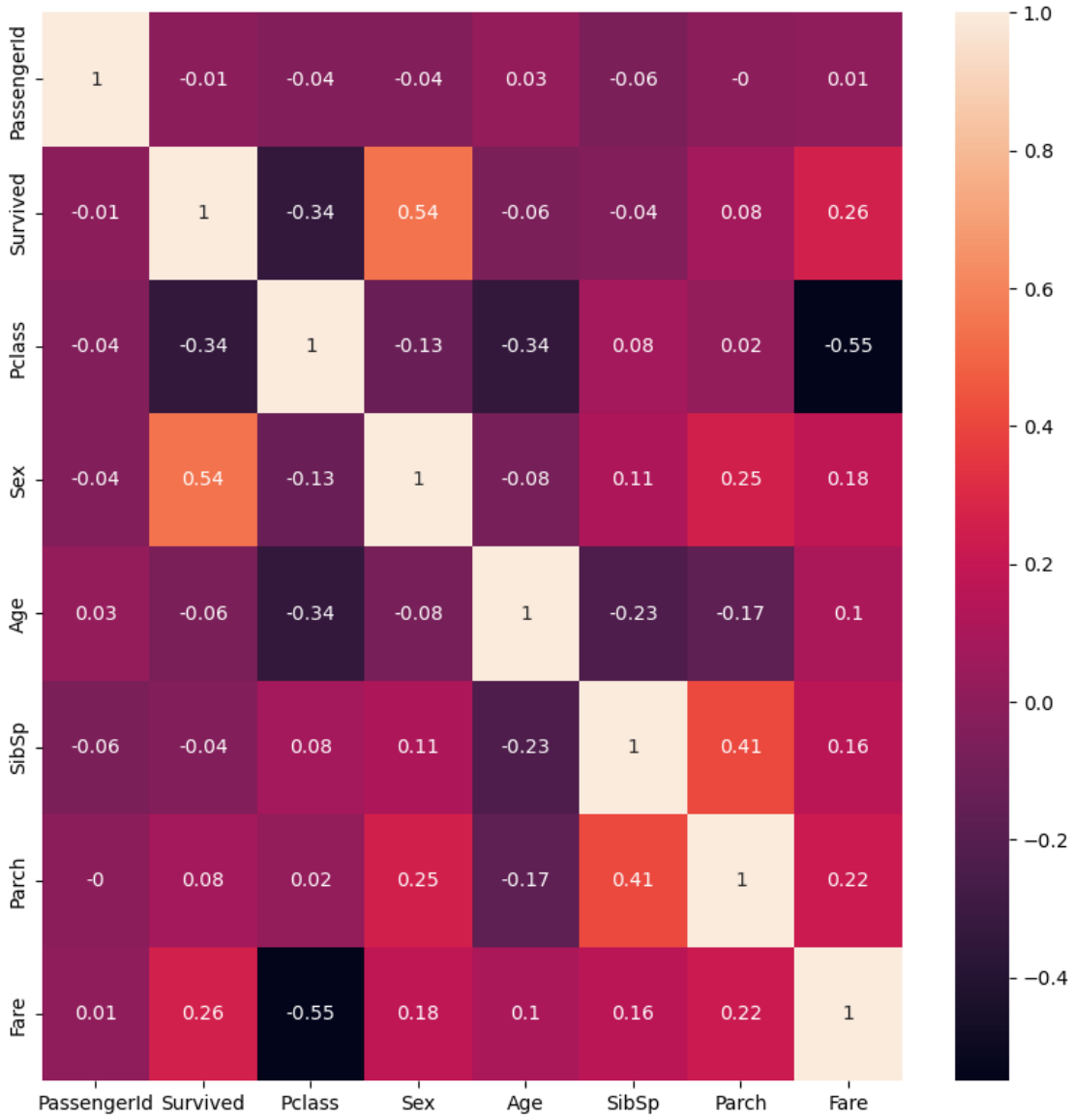


No boxplot acima podemos estimar as idades e fare dos passageiros no geral.

5 Gráfico de correlação da base de dados

[ ]:
```
plt.figure(figsize=(10,10))              #Criando imagem e definindo tamanho
sns.heatmap(Base.corr().round(2), annot = True)          #gráfico de correlação␣
 ↪usando biblioteca seaborn(sns) e arredondando para 2 casas decimais
```

[ ]:
```
<AxesSubplot: >
```



Foi realizado o teste de correlação entre as variáveis para procurar entender se suas respctivas variações apresentariam alguma relação. Apesar de não ter notado nenhuma correlação forte, isto é, próxima de 1 ou -1, a variável “Sex” apresentou a maior relação com o número de sobreviventes, sendo esta uma correlação positiva, ou seja, quando uma aumenta, a outra aumenta também. No sistema binário tanto mulheres quanto sobreviventes correspondiam ao número 1. Desta forma, a correlação sugere que mulheres e sobreviventes estavam correlacionados.

5.0.1 Obtendo total de mulheres sobreviventes

[ ]:
```
ms = Base[(Base.Sex == 1 ) & (Base.Survived == 1)]      #Filtrando mulheres␣
↪sobreviventes
ms = len(ms)                #Obtendo quantidade através do tamanho do dataframe
ms
```

[ ]:
```
233
```

5.0.2 Obtendo total de sobreviventes

[ ]:
```
ts = Base[Base.Survived == 1]
ts = len(ts)
ts
```

[ ]:
```
342
```

5.0.3 Teste de proporção de mulheres sobreviventes

[ ]:
```
print(proportion_confint(ms,ts))            #Intervalo de confiança da␣
↪proporção
print(proportions_ztest(ms,ts,0.5))
print(round(ms/ts,2))
```

```
(0.6319009716644249, 0.7306721277507798)
(7.1947139709042505, 6.259183826006731e-13)
0.68
```

Ao perceber a correlação entre mulheres e sobreviventes, foi feito um teste de proporção, onde foi concluído com 95% de confiança, que embora a quantidade de homens a bordo fosse maior, dentre o número absoluto de sobreviventes, 68% eram mulheres.

6 Criando base de passageiros por classe

[ ]:
```
warnings.filterwarnings("ignore")
Base3 = Base[['Survived','Pclass']]           #Base por classe
Base3['C1'] = np.where(Base3.Pclass == 1,1,0)
Base3['C2'] = np.where(Base3.Pclass == 2,1,0)
Base3['C3'] = np.where(Base3.Pclass == 3,1,0)
Base3.head(5)
```

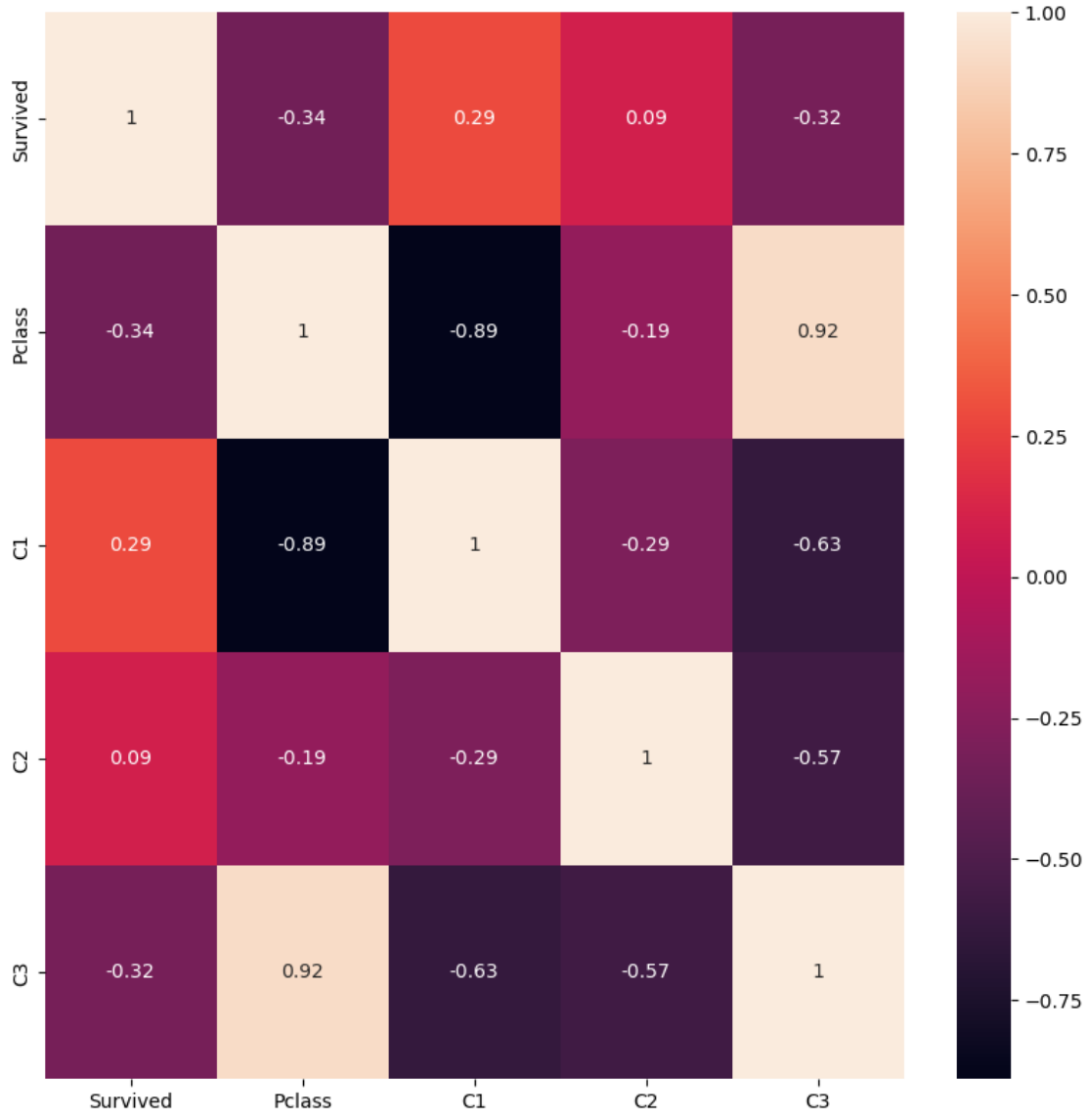
[ ]:

	Survived	Pclass	C1	C2	C3
0	0	3	0	0	1
1	1	1	1	0	0
2	1	3	0	0	1
3	1	1	1	0	0
4	0	3	0	0	1

7 Gráfico de correlação por classes

```
[ ]: plt.figure(figsize=(10,10))              #Criando imagem e definindo tamanho
     sns.heatmap(Base3.corr().round(2), annot = True)
```

```
[ ]: <AxesSubplot: >
```



Ao analisar o gráfico acima, foi percebida uma correlação positiva entre sobreviventes e a primeira classe e uma correlação negativa entre sobreviventes e a terceira classe. Isto é, o número de sobreviventes e pessoas da primeira classe crescem na mesma direção, enquanto o número de pessoas da terceira classe crescem na mesma direção que o número de mortos. Apesar de entender que as classes poderiam, mesmo que de forma sutil (baixa correlação) ter influenciado no "poder de

sobrevivência", foi feita a estratificação das classes por sexo, a fim de entender melhor, se dentro de cada sexo as classes foram um fator determinante.

8 Comparações por Classe

8.0.1 Obtendo mulheres sobreviventes da classe 1 e total de de mulheres da classe 1

[ ]:
```
# Obtendo mulheres sobreviventes da classe 1
msc1 = Base[(Base.Pclass == 1 ) & (Base.Survived == 1) & (Base.Sex == 1)]
msc1 = len(msc1)
print(msc1)
# Obtendo total de mulheres da classe 1
tmc1 = Base[(Base.Pclass == 1) & (Base.Sex == 1)]
tmc1 = len(tmc1)
print(tmc1)
```

```
91
94
```

8.0.2 Teste de proporção das mulheres sobreviventes da classe 1

[ ]:
```
print(proportion_confint(msc1,tmc1))              #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(msc1,tmc1,0.5))
print(round(msc1/tmc1,2))
```

```
(0.9325516352530691, 1.0)
(25.818754009596567, 5.460783834601893e-147)
0.97
```

8.0.3 Obtendo mulheres sobreviventes da classe 3 e total de de mulheres da classe 3

[ ]:
```
# Obtendo mulheres sobreviventes da classe 3
msc3 = Base[(Base.Pclass == 3 ) & (Base.Survived == 1) & (Base.Sex == 1)]
msc3 = len(msc3)
print(msc3)
# Obtendo total de mulheres da classe 3
tmc3 = Base[(Base.Pclass == 3) & (Base.Sex == 1)]
tmc3 = len(tmc3)
print(tmc3)
```

```
72
144
```

8.0.4 Teste de proporção das mulheres sobreviventes da classe 3

[ ]:
```
print(proportion_confint(msc3,tmc3))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(msc3,tmc3,0.5))        #Descarta a hipótese alternativa␣
 ↪pois na classe 3 exatamente a metade das mulheres sobreviveram
print(round(msc3/tmc3,2))
```

```
(0.4183348339774977, 0.5816651660225023)
(0.0, 1.0)
0.5
```

A partir dos testes realizados, foi possível perceber que dentro do sexo feminino, as classes tiveram comportamento de sobreviventes diferentes, onde a primeira classe aparentou ter maior chance de sobreviver do que a terceira: 97% das mulheres da primeira classe sobreviveram enquanto 50% das mulheres da terceira classe sobreviveram.

8.0.5 Obtendo homens sobreviventes da classe 1 e total de de homens da classe 1

[ ]:
```
# Obtendo homens sobreviventes da classe 1
hsc1 = Base[(Base.Sex == 0) & (Base.Survived == 1) & (Base.Pclass == 1)]
hsc1 = len(hsc1)
print(hsc1)
# Obtendo total de homens da classe 1
thc1 = Base[(Base.Sex == 0) & (Base.Pclass == 1)]
thc1 = len(thc1)
print(thc1)
```

```
45
122
```

8.0.6 Teste de proporção de homens sobreviventes da classe 1

[ ]:
```
print(proportion_confint(hsc1,thc1))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(hsc1,thc1,0.5))            #Não podemos descartar a␣
 ↪hipótese nula. Menos de 50% dos homens da classe 1 sobreviveram.
print(round(hsc1/thc1,2))
```

```
(0.2832354672379669, 0.45446945079482004)
(-3.0022598511061984, 0.0026798331913469088)
0.37
```

8.0.7 Obtendo homens sobreviventes da classe 3 e total de de homens da classe 3

[ ]:
```
# Obtendo homens sobreviventes da classe 3
hsc3 = Base[(Base.Sex == 0) & (Base.Survived == 1) & (Base.Pclass == 3)]
hsc3 = len(hsc3)
```

```
print(hsc3)
# Obtendo total de homens da classe 3
thc3 = Base[(Base.Sex == 0) & (Base.Pclass == 3)]
thc3 = len(thc3)
print(thc3)
```

```
47
347
```

8.0.8 Teste de proporção de homens sobreviventes da classe 3

[ ]:
```
print(proportion_confint(hsc3,thc3))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(hsc3,thc3,0.5))         #Não podemos descartar a␣
 ↪hipótese nula. Menos de 50% dos homens da classe 3 sobreviveram.
print(round(hsc3/thc3,2))
```

```
(0.09944163448599802, 0.17145173727192703)
(-19.844753402062032, 1.2232728571920844e-87)
0.14
```

A partir dos testes realizados, foi possível perceber que dentro do sexo masculino, as classes tiveram comportamento de sobreviventes diferentes, onde a primeira classe aparentou ter maior chance de sobreviver do que a terceira: 37% dos homens da primeira classe sobreviveram enquanto 14% dos homens da terceira classe sobreviveram.

9 Comparações por Tarifa

9.0.1 Obtendo quantidade de passageiros por valor pago

[ ]:
```
# Obtendo quantidade de passageiros por valor pago 31, 14 , 7
maior31 = Base[(Base.Fare >= 31)]
maior31 = len(maior31)
print(maior31)
maior14 = Base[(Base.Fare < 31) & (Base.Fare >= 14)]
maior14 = len(maior14)
print(maior14)
menor14 = Base[(Base.Fare < 14)]
menor14 = len((menor14))
print(menor14)
```

```
225
230
436
```

9.0.2 Obtendo quantidade de passageiros sobreviventes pelo valor pago

[ ]:
```
surv31 = Base[(Base.Fare >= 31) & (Base.Survived == 1)]
surv31 = len(surv31)
print(surv31)
surv14 = Base[(Base.Fare < 31) & (Base.Fare >= 14) & (Base.Survived == 1)]
surv14 = len(surv14)
print(surv14)
surv0 = Base[(Base.Fare < 14) & (Base.Survived == 1)]
surv0 = len((surv0))
print(surv0)
```

```
131
100
111
```

9.0.3 Teste de proporção de quem pagou mais que 31

[ ]:
```
print(proportion_confint(surv31,maior31))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(surv31,maior31,0.5))
print(round(surv31/maior31,2))
```

```
(0.517779498418524, 0.6466649460259204)
(2.500710472376967, 0.012394446013367241)
0.58
```

9.0.4 Teste de proporção de quem pagou mais que 14 e menos que 27

[ ]:
```
print(proportion_confint(surv14,maior14))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(surv14,maior14,0.5))
print(round(surv14/maior14,2))
```

```
(0.3707165491352943, 0.49884866825601004)
(-1.995186515283529, 0.04602254120887567)
0.43
```

9.0.5 Teste de proporção de quem pagou menos que 14

[ ]:
```
print(proportion_confint(surv0,menor14))            #Intervalo de confiança da␣
 ↪proporção
print(proportions_ztest(surv0,menor14,0.5))
print(round(surv0/menor14,2))
```

```
(0.2136967401408829, 0.29547757178572254)
(-11.763155888274051, 6.043240470251041e-32)
0.25
```

A partir dos testes realizados, foi possível entender que dos 342 sobreviventes, 25% pagaram mais de 31, 26% pagaram entre 14 e 31 e 49% pagaram menos de 14. Embora a maior parte seja de pagantes dos tickets com menores valores, ao analisar a proporção de sobreviventes de cada grupo, é possível perceber que existiu uma chance maior de sobrevivência para os grupos que pagaram por tickets com valores maiores:

- 58% dos que pagaram + de $31 sobreviveram;
- 43% dos que pagaram + de $14 sobreviveram;
- 25% dos que pagaram - de $14 sobreviveram.

10 Conclusão

Através das análises deste trabalho foi possível obter algumas perceções curiosas. Embora tivessem muito mais homens do que mulheres a bordo, as mulheres sobreviveram mais do que os homens. Apesar da terceira classe ser a maioria, a primeira classe se mostrou mais propensa a sobreviver tanto entre homens como entre mulheres. O poder aquisitivo demonstrado pelo valor dos tíckets (variável fare) também mostraram que os grupos que pagavam mais, tinham uma proporção maior de sobreviventes.