Investigando um conjunto de dados - Titanic

Bruno Gonzaga Santos

06.06.2018

Fundamentos em Data Science I

INTRODUÇÃO

RMS Titanic é um lendário navio britânico que foi lançado ao mar a primeira vez no ano de 1911, projetado pelos engenheiros navais Alexander Carlisle e Thomas Andrews o navio foi pensado para ser o mais seguro e mais luxuoso de sua época, inclusive gerando lendas de que era supostamente “inquebrável”.

A embarcação partiu em sua viagem inaugural de Southampton para Nova Iorque em 10 de abril de 1912, no caminho passando em Cherbourg-Octeville na França e por Queenstown na Irlanda. Ele colidiu com um iceberg às 23h40min do dia 14 de abril e

afundou na madrugada do dia seguinte com mais de 1 500 pessoas a bordo, sendo um dos maiores desastres marítimos em tempos de paz de toda a história. Seu naufrágio destacou vários pontos fracos de seu projeto, deficiências nos procedimentos de evacuação de emergência e falhas nas regulamentações marítimas da época. Comissões de inquérito foram instauradas nos Estados Unidos e no Reino Unido, levando a mudanças nas leis internacionais de navegação que permanecem em vigor mais de um século depois.

Dataset

Para esta análise foi utilizado o dataset obtido no site Kaggle, o mesmo possuí uma amostra de 891 passageiros, portanto não contempla todos os passageiros que estavam naquela embarcação no momento da tragédia. Segue abaixo um quadro com mais informações sobre cada coluna que este dataset disponibiliza.

Dicionário de dados

Variable	Definition	Key
survival	Survival	0 = No, 1 = Yes
pclass	Ticket class	1 = 1st, 2 = 2nd, 3 = 3rd
sex	Sex	
Age	Age in years	
sibsp	# of siblings / spouses aboard the Titanic	
parch	# of parents / children aboard the Titanic	
ticket	Ticket number	
fare	Passenger fare	
cabin	Cabin number	
embarked	Port of Embarkation	C = Cherbourg, Q = Queenstown, S = Southampton

Para este trabalho, como irá ficar mais claro mais para frente. Iremos selecionar apenas algumas das propriedades do dataset.

Limitações

Observando nosso dataset verificou-se que alguns valores não possuíam valores para alguns dos passageiros. Esses atributos que identificamos são: Age, Cabin, Embarked.

Medida tomada:

- **Age:** os valores nulos foram preenchidos com base na idade média de toda população da amostra.
- **Cabin:** Como veremos mais à frente esta informação não é importante para respondermos nossos questionamentos, portanto estaremos logo mais excluído toda esta coluna.
- **Embarked:** Apresenta um número não significativo com dados faltantes, por este motivo a fins estatísticos, foi decidido excluir do dataset as linhas que continham valores nulos neste atributo.

Questionamentos

Analisando as informações que foi disponibilizado, foram elaboradas às seguintes perguntas para nos ajudar a esclarecer alguns pontos referente ao acidente. Lembrando que todas as respostas consideram apenas dados referente a amostra disponibilizada e não o número total de pessoas que estavam a bordo do Titanic.

1. Quantos foram os que sobrevieram, segundo a amostra?
2. Quantos foram os que morreram na catástrofe, segundo a amostra?
3. Dentre os que sobreviveram, qual a média da idade?
4. Dentre os que não sobreviveram, qual a média de idade?
5. Qual a relação entre os sobreviventes e a classe de ingresso?
6. Qual o sexo mais relevante entre os sobreviventes?
7. Morreram mais pessoas do sexo feminino ou masculino?
8. Em qual classe de ingresso havia mais crianças?

Wrangle (Limpeza de dados)

O primeiro à ser realizado é a leitura dos dados do dataset. Para isso iremos importar a biblioteca pandas, pois a mesma contém inúmeras funcionalidades que permite o fácil manuseio de estrutura de dados como a que está sendo estudada. Serão importados também o pyplot do matplotlib para ajudar nas plotagens de gráficos e o seaborn para tornar os gráficos mais atraentes.

```
In [59]: import pandas as pd
         from matplotlib import pyplot as plt
         import seaborn as sns
         %matplotlib inline

         titanic = pd.read_csv('titanic-data-6.csv')
```

Antes de começarmos a limpeza do nosso dataset, é importante verificar a disposição dos dados dentro do mesmo. Abaixo foram executados dois comandos: 1 para mostrar as primeiras linhas (onde já é possível entender como os dados são apresentados) e mais abaixo verificamos o tamanho total do arquivo, neste caso foi nos apresentado um dataset com **891 linhas** e **12 colunas**.

```
In [60]: print('Amostra dos dados do dataset - Titanic')
         titanic.head()
```

```
Amostra dos dados do dataset - Titanic
```

Out[60]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

```
In [61]: # Número de amostras do conjunto
         print(f'Este conjunto tem: {titanic.shape[0]} linhas e {titanic.shape[1]} colunas.')
```

```
Este conjunto tem: 891 linhas e 12 colunas.
```

Ok, será que desses 891 registros, existem valores que estão duplicados por algum motivo? Vamos verificar.

```
# Números de linhas duplicadas - Wrangle
print("Duplicadas - Titanic: {0}".format(titanic.duplicated().sum()))
```

```
Duplicadas - Titanic: 0
```

Ótimo, após a execução do comando acima é possível verificar que nosso dataset não possui linhas duplicadas.

Humn! Considerando as perguntas que iremos utilizar, será que iremos precisar de todos as colunas que existem hoje no dataset? Não, existem colunas que irão agregar muito pouco para nossa análise, vamos removê-las.

```
# Removendo colunas que não serão usadas - Wrangle
titanic.drop(['PassengerId', 'Name', 'Ticket', 'Fare', 'Cabin'], axis=1, inplace=True)
```

Pronto! Removemos as colunas: PassengerId, Name, Ticket, Fase, Cabin. Pois as mesmas pouco irão agregar para responder nossas perguntas.

Logo mais iremos gerar estatísticas de alguns dos dados, será que dados como, por exemplo idade estão sendo tratados corretamente pelo python? Vamos verificar.

```
# Checando tipos das colunas - Wrangle
def check_types(df):
    for column in df.columns:
        print("coluna: {0}, tipo: {1}".format(column, type(df[column][0])))
check_types(titanic)
```

```
coluna: Survived, tipo: <class 'numpy.int64'>
coluna: Pclass, tipo: <class 'numpy.int64'>
coluna: Sex, tipo: <class 'str'>
coluna: Age, tipo: <class 'numpy.float64'>
coluna: SibSp, tipo: <class 'numpy.int64'>
coluna: Parch, tipo: <class 'numpy.int64'>
coluna: Embarked, tipo: <class 'str'>
```

Ok, o campo idade que usaremos logo mais para gerar médias já está sendo tratado como float, isso garante que provavelmente não teremos problemas para fazer este cálculo. Porém ainda temos o risco de existirem valores nulos dentro deste dataset, iremos fazer a verificação logo mais.

```
# Verificando a ocorrência de elementos únicos em cada coluna - Wrangle
titanic.nunique()
```

```
Survived     2
Pclass       3
Sex          2
Age         88
SibSp        7
Parch        7
Embarked     3
dtype: int64
```

```
print(f'Valores -> Survived: {titanic["Survived"].unique()}')
print(f'Valores -> Pclass: {titanic["Pclass"].unique()}')
print(f'Valores -> Sex: {titanic["Sex"].unique()}')
print(f'Valores -> Age: {titanic["Age"].unique()}')
print(f'Valores -> SibSp: {titanic["SibSp"].unique()}')
print(f'Valores -> Parch: {titanic["Parch"].unique()}')
print(f'Valores -> Embarked: {titanic["Embarked"].unique()}')
```

```
Valores -> Survived: [0 1]
Valores -> Pclass: [3 1 2]
Valores -> Sex: ['male' 'female']
Valores -> Age: [22.   38.   26.   35.     nan 54.    2.   27.   14.    4.   58.   20.
 39.   55.   31.   34.   15.   28.    8.   19.   40.   66.   42.   21.
 18.    3.    7.   49.   29.   65.   28.5   5.   11.   45.   17.   32.
 16.   25.    0.83 30.   33.   23.   24.   46.   59.   71.   37.   47.
 14.5  70.5  32.5  12.    9.   36.5  51.   55.5  40.5  44.    1.   61.
 56.   50.   36.   45.5  20.5  62.   41.   52.   63.   23.5   0.92 43.
 60.   10.   64.   13.   48.    0.75 53.   57.   80.   70.   24.5   6.
  0.67 30.5   0.42 34.5  74.  ]
Valores -> SibSp: [1 0 3 4 2 5 8]
Valores -> Parch: [0 1 2 5 3 4 6]
Valores -> Embarked: ['S' 'C' 'Q' nan]
```

Na foto acima, podemos verificar os valores únicos de cada coluna, primeiramente a quantidade de valores únicos tem e na sequência seus valores.

É possível identificar que existem valores nulos dentro da coluna que representa a idade e a cidade em que o passageiro embarcou.

```
# exiba a contagem de valores faltantes para cada característica - Wrangle
titanic.isnull().sum()
```

```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Embarked      2
dtype: int64
```

Para as 2 linhas que possuem valores nulos no campo de *cidade em que o passageiro embarcou*, iremos remover essas linhas, já que não representa uma quantia significativa para nossa amostra, assim seguiremos.

Para a coluna idade existe 177 pessoas com valores nulos, como este número representa um grande percentual da nossa amostra, iremos seguir a estratégia de substituir esses valores nulos, pela média calculada da idade, como representado abaixo.

```
# descarte linhas com quaisquer valores nulos na coluna Embarked - Wrangle
titanic.dropna(subset=['Embarked'], inplace=True)
titanic.isnull().sum()
```

```
Survived      0
Pclass        0
Sex           0
Age         177
SibSp         0
Parch         0
Embarked      0
dtype: int64
```

```
# Substituindo valores nulos da idade pela média - Wrangle
media_idade = titanic['Age'].mean()
titanic['Age'].fillna(media_idade, inplace=True)
titanic.isnull().sum()
```

```
Survived    0
Pclass      0
Sex         0
Age         0
SibSp       0
Parch       0
Embarked    0
dtype: int64
```

Pronto, sem valores nulos em nosso dataset. Ainda tem problema que algumas colunas não possuem nomes tão intuitivas, além disso eles estão em inglês, como esta análise está sendo realizada para um público majoritariamente brasileira, vamos renomear colunas e seus valores de domínios para o português.

```
# Renomeando colunas para melhor entendimento - Wrangle
colunas = ['Sobreviveu', 'Classe', 'Sexo', 'Idade', 'Irmãos/Cônjugue', 'Pais/Crianças', 'Embarcou']
titanic.columns = colunas
titanic.head(1)
```

	Sobreviveu	Classe	Sexo	Idade	Irmãos/Cônjugue	Pais/Crianças	Embarcou
0	0	3	male	22.0	1	0	S

```
titanic['Sobreviveu'] = titanic['Sobreviveu'].replace([0, 1], ['Não', 'Sim'])
titanic['Classe'] = titanic['Classe'].replace([1, 2, 3], ['1st', '2st', '3st'])
titanic['Sexo'] = titanic['Sexo'].replace(['male', 'female'], ['masculino', 'feminino'])
titanic['Embarcou'] = titanic['Embarcou'].replace(['S', 'C', 'Q'], ['Cherbourg', 'Southampton', 'Queenstown'])
titanic.head()
```

	Sobreviveu	Classe	Sexo	Idade	Irmãos/Cônjugue	Pais/Crianças	Embarcou
0	Não	3st	masculino	22.0	1	0	Cherbourg
1	Sim	1st	feminino	38.0	1	0	Southampton
2	Sim	3st	feminino	26.0	0	0	Cherbourg
3	Sim	1st	feminino	35.0	1	0	Cherbourg
4	Não	3st	masculino	35.0	0	0	Cherbourg

Dataset limpo. Vamos agora verificar como nosso dataset aparenta depois da nossa limpeza.

	Sobreviveu	Classe	Sexo	Idade	Irmãos/Cônjugue	Pais/Crianças	Embarcou
0	Não	3st	masculino	22.0	1	0	Cherbourg
1	Sim	1st	feminino	38.0	1	0	Southampton
2	Sim	3st	feminino	26.0	0	0	Cherbourg
3	Sim	1st	feminino	35.0	1	0	Cherbourg
4	Não	3st	masculino	35.0	0	0	Cherbourg

```
check_types(titanic)
```

```
coluna: Sobreviveu, tipo: <class 'str'>
coluna: Classe, tipo: <class 'str'>
coluna: Sexo, tipo: <class 'str'>
coluna: Idade, tipo: <class 'numpy.float64'>
coluna: Irmãos/Cônjugue, tipo: <class 'numpy.int64'>
coluna: Pais/Crianças, tipo: <class 'numpy.int64'>
coluna: Embarcou, tipo: <class 'str'>
```

```
titanic.describe()
```

	Idade	Irmãos/Cônjugue	Pais/Crianças
count	889.000000	889.000000	889.000000
mean	29.642093	0.524184	0.382452
std	12.968346	1.103705	0.806761
min	0.420000	0.000000	0.000000
25%	22.000000	0.000000	0.000000
50%	29.642093	0.000000	0.000000
75%	35.000000	1.000000	0.000000
max	80.000000	8.000000	6.000000

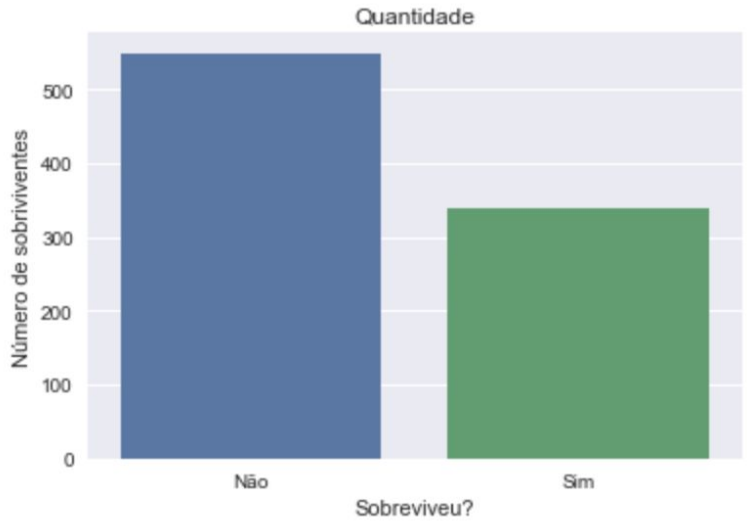
Perfeito, parece tudo em ordem. Vamos agora começar a responder nossas perguntas.

Explore (explorando os dados)

Sobreviventes vs não sobreviventes

Existia a crença que o Titanic era um dos navios mais seguros da época, como dito no início dessa análise ele foi projeto com o intuito de ser o mais seguro da época. Sabemos o destino do navio, mas será que os aparatos de emergências disponíveis garantiram a sobrevivência da maioria? Ou será que o número de óbitos acabou superando o de sobreviventes?

Infelizmente como demonstra o quadro abaixo, segundo a amostra analisada, a maioria dos ocupantes do navio não saíram vivos da catástrofe. Um número de aproximadamente 60% dos passageiros acabaram morrendo. Muito provavelmente devido às condições adversas do local e o clima onde o navio afundou.




```
Sobreviveram 340 pessoas que estavam no titanic.
Morreram 549 pessoas que estavam no titanic.
```

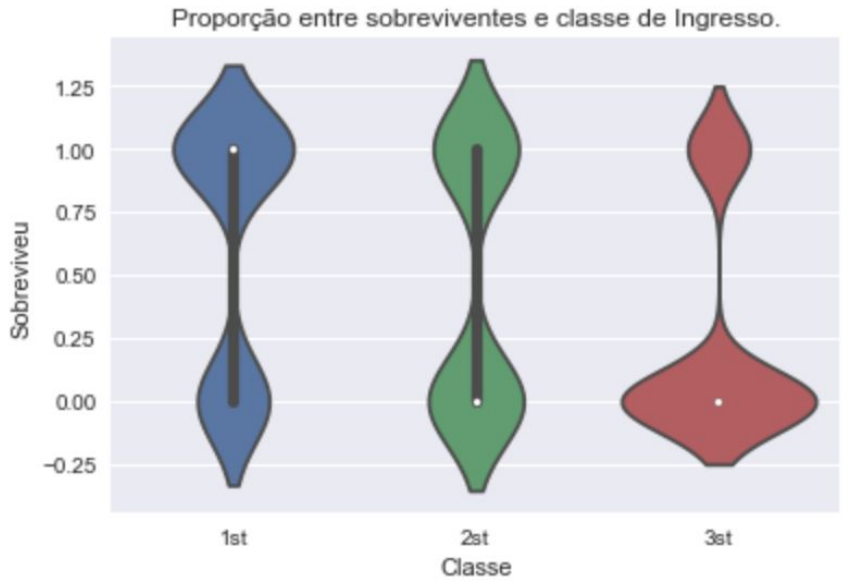
Porém, olhando as pessoas que sobreviveram? Quais características foram mais determinantes na hora em que o navio começou a ser evacuado?

Idade foi determinante?

Vamos começar pela idade. A média de idade das pessoas que sobreviveram foi de aproximadamente 28 anos, já as que não sobreviveram foi de 30 anos. Aparentemente, apesar de crianças terem preferências no uso do bote salva vidas, a idade não foi muito relevante na luta pela vida naquele dia.

Rico versus Pobre

Será que a classe em que o passageiro estava influenciou? É sabido que os passageiros das classes mais altas ficavam na parte superior do navio, portanto mais próximo aos acessos dos botes salva-vidas, vamos visualizar um gráfico violino para verificar a proporção de sobreviventes de cada classe.

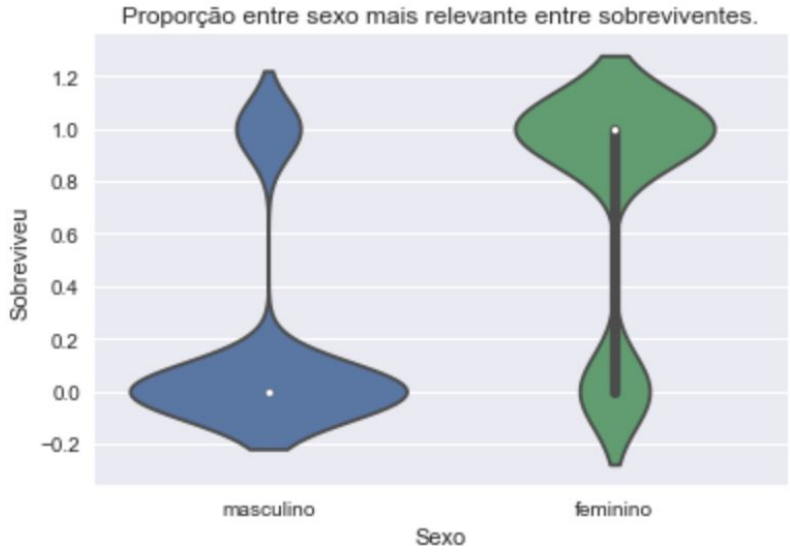


Explicando o gráfico, foi definido o valor 1 para passageiros que sobreviveram em nosso dataset, e o valor 0 para os que não sobreviveram. Analisando a proporção das três classes é possível identificar que pessoas que estavam na **3 classe**, tiveram desvantagem na evacuação do navio, possivelmente pelo fato de estarem em uma área mais afastada dos botes de segurança e mais próxima da área afetado pela entrada da água do mar.

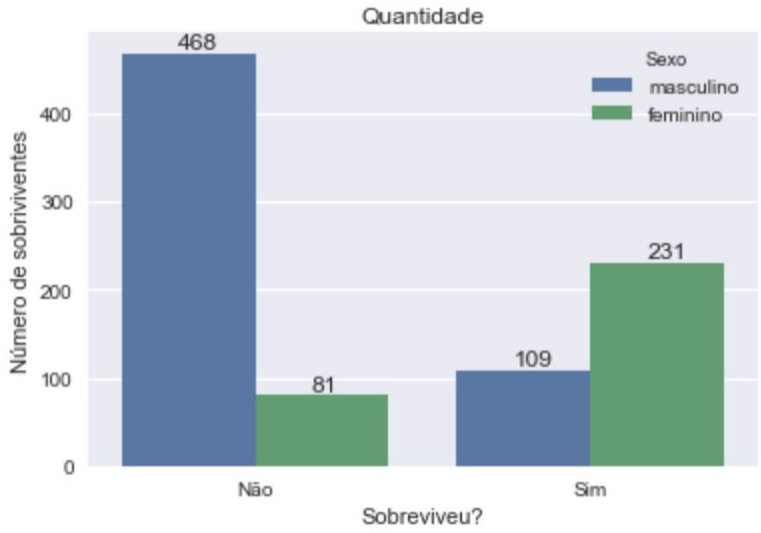
Já na primeira classe, verificamos de forma não tão acintosa, mas que a maioria das pessoas desta classe, conseguiram escapar com vida.

Guerra dos sexos

Geralmente mulheres e crianças têm prioridade para o desembarque, será que isso de alguma forma tornou mais favorável para mulheres saírem vida do episódio. Vamos descobrir no gráfico abaixo.

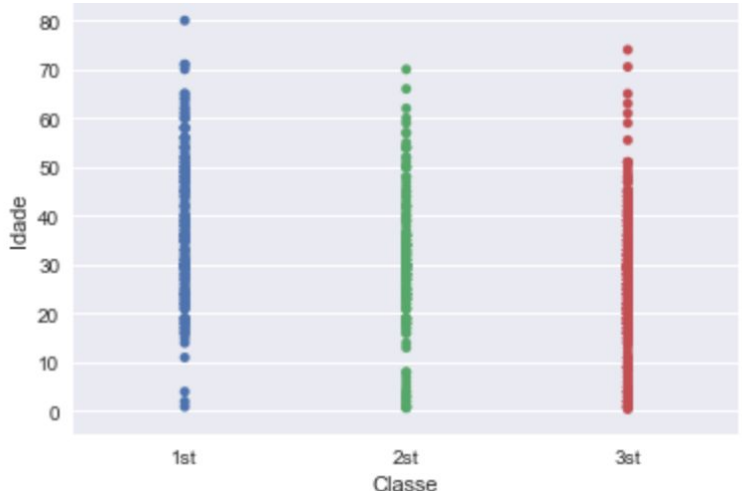


Opa, é o que parece, o gráfico violino deixa claro que a proporção de mulheres sobreviventes foi bem maior que os dos homens. Para finalizar, vamos verificar o gráfico em um formato de barras?



Classes mais jovens

Finalmente, para concluirmos nossa análise e respondermos todas às perguntas. Qual será a classe que apresenta maior distribuição de crianças e adolescentes na sua composição? Pelo gráfico abaixo podemos verificar que a 3ª classe era a que tinha uma maior ocorrência de pessoas entre 0 e 20 anos, já nas primeira classe podemos verificar uma ocorrência maior de adultos maiores de 50 anos.



CONCLUSÃO

Analisando os dados da tragédia do Titanic, ficam evidenciados os seguintes pontos:

- que tanto a classe onde o passageiro estava hospedado no navio, quanto seu sexo foram determinantes para sobrevivência;
    - Passageiros da primeira classe tiveram um índice de sobrevivência muito maior do que os das classes mais baixos;
    - Aproximadamente 62% dos passageiros do Titanic não sobreviveram;
    - O Titanic tinha uma idade total média de 29 anos;
    - Assim como nas classes de ingressos, é possível verificar que as mulheres tiveram proporcionalmente um índice de mortalidade visivelmente menor do que dos homens.

Anexo

O Titanic é um ícone mundial. A tragédia provavelmente nunca será esquecida. Sua derrocada gerou normas mais rígidas de seguranças e até mudanças na arquitetura de navios, para que o dia de 14 de abril de 1912 nunca mais volte a se repetir na mesma proporção.

Após a colisão com o iceberg no navio demorou aproximadamente 2 horas para submergir no oceano totalmente. O que seria tempo suficiente para que um número de pessoas próximo a totalidade fosse retirado com vida se os controles de seguranças fossem mais rígidos.

Limitando à conclusão apenas aos dados que analisamos, conseguimos verificar que tanto a classe onde o passageiro estava hospedado no navio, quanto seu sexo foram determinantes para sobrevivência.

Pessoas da terceira classe, tiveram menos sorte provavelmente por estarem mais longe dos botes e por suas áreas serem as primeiras afetadas pela batida, já que elas se encontravam no fundo do navio.

Por outro lado, mulheres, muito provavelmente por terem preferência na saída dos botes salva vidas tiveram um índice de mortalidade bem menor do que dos homens.

Observando que tanto as pessoas que estavam mais próximas dos botes salva vidas, quanto as que tinha preferências por embarcar nesses botes sobreviveram em maior número, fica claro que se houvesse uma quantidade maior de botes, provavelmente mais pessoas iremos conseguir prolongar suas vidas.

Encerro desejando minhas condolências à todas às vítimas do Titanic, e que o episódio sirva de lição para que novos não se repitam.

REFERÊNCIAS

1. https://www.kaggle.com/c/titanic/data
2. http://pandas.pydata.org/pandas-docs/stable/index.html
3. https://seaborn.pydata.org/index.html
4. http://udacity.com.br
5. https://92760526.weebly.com/conclusion.html