



Relatório Sobre Análise de Dados do Desastre do Titanic



Atec - TPSICAS1123 - 5417

Relatório realizado por: João Pedro Fernandes Silva, Ricardo Lopes da Conceição e Diogo Baptista Louro

Formador: Nelson Alexandre Santos

Índice

Introdução.....	3
Estrutura dos dados	3
Leitura e Exploração dos Dados	4
Limpeza e Pré-Processamento.....	5
Visualização de dados	10
Análise Adicional – Relação entre Família e Sobrevivência	16
Análise Adicional – Comparação por Porto de Embarque	17
Exportação dos resultados.....	18
Utilização do Git	19
Conclusão.....	20

Índice de Figuras

Figura 1 - Utilização de .head, .tail e .describe	4
Figura 2 - Resumo estatístico	4
Figura 3 - Info geral sobre o DataFrame	4
Figura 4 - Valores nulos existentes, por coluna	5
Figura 5 - Valores nulos após tratamento.....	5
Figura 6 - Tipo de dados.....	6
Figura 7 - Idades em milissegundos desde o Epoch.....	6
Figura 8 - Taxa de sobrevivência por classe e género.....	7
Figura 9 - Media de idade dos sobreviventes	7
Figura 10 - Idade média por classe e sobrevivência	8
Figura 11 - Distribuição da Idade por Sobrevivência	8
Figura 12 - Gráfico de Linha: Representa a distribuição dos sobreviventes por classe e sexo	10
Figura 13 - Gráfico de Dispersão: Representa a correlação entre as variáveis Age, Fare e Survived	11
Figura 14 - Gráfico de Dispersão: Representa a correlação entre Idade e Sobrevivência	11
Figura 15 - Gráfico de Dispersão: Representa a correlação entre a Tarifa vs Sobrevivência	12
Figura 16 - Gráfico de Barras: Distribuição das Idades	13
Figura 17 - Gráfico de Barras: Distribuição de Tarifas.....	14
Figura 18 - Gráfico de Barras: Distribuição da Sobrevivência	15
Figura 19 - Prova de criação da nova coluna Tamanho_Familia.....	16
Figura 20 - Gráfico de Barras: Relação entre o Tamanho das Famílias e a Sobrevivência.....	16
Figura 21 - Coluna Embarked.....	17
Figura 22 - Gráfico de Barras: Taxa de sobrevivência por Porto de Embarque	17
Figura 23 - Ficheiro .csv com dados tratados.....	18
Figura 24 - Ficheiro final .csv criado.....	18
Figura 25 - Amostra dos Ficheiros.CSV	18
Figura 26 - Utilização do Git.....	19

Introdução

Neste relatório analisámos um conjunto de dados utilizando Python e bibliotecas como Pandas para manipulação e análise de dados e Matplotlib para visualizações gráficas. O conjunto de dados fornecido contém informações sobre os passageiros como a idade, sexo, classe de embarque e se sobreviveram ou não.

O objetivo deste projeto é realizar uma análise com uma exploração inicial dos dados e apresentação de insights relevantes. O relatório inclui etapas de limpeza e pré-processamento dos dados, criação de métricas adicionais e visualizações gráficas.

Estrutura dos dados

Os dados fornecidos contêm 12 colunas sobre as características dos passageiros e a sobrevivência dos mesmos. Analisando os dados temos o seguinte:

- PassengerId: identificador único;
- Survived: 1 (sobreviveu) ou 0 (não sobreviveu);
- Pclass: classe do passageiro (1ª, 2ª ou 3ª classe);
- Name: nome do passageiro;
- Sex: género do passageiro;
- Age: idade dos passageiros;
- SibSp: número de irmãos e esposo/a do passageiro;
- Parch: números de pais e filhos do passageiro;
- Ticket: número do bilhete do passageiro;
- Fare: taxa cobrada por passageiro;
- Cabin: número da cabine do passageiro;
- Embarked: porto de embarque C, Q ou S.

O dataset foi carregado com `pd.read_csv`. Os métodos `.head()`, `.tail()`, `.describe()` e `.info()` foram usados para analisar a estrutura de dados, identificar valores nulos e compreender o formato das colunas.

Leitura e Exploração dos Dados

Os dados foram carregados com `pd.read_csv()` e inspecionados utilizando `.head`, `.tail`, `.describe` e `.info` respetivamente, obtendo os seguintes resultados:

```
Atividade 3.1 - Leitura e exploração dos dados

Primeiros 3 registos:
  PassengerId  Survived  Pclass                    Name  Sex  Age  SibSp  Parch  Ticket   Fare Cabin Embarked
0          892         0       3            Kelly, Mr. James  male  34.5    0    0   330911  7.8292   NaN      Q
1          893         1       3  Wilkes, Mrs. James (Ellen Needs)  female  47.0    1    0   363272  7.0000   NaN      S
2          894         0       2         Myles, Mr. Thomas Francis  male  62.0    0    0   240276  9.6875   NaN      Q

Últimos 3 registos:
  PassengerId  Survived  Pclass                    Name  Sex  Age  SibSp  Parch  Ticket   Fare Cabin Embarked
415         1307         0       3   Saether, Mr. Simon Sivertsen  male  38.5    0    0   SOTON/O.Q. 3101262  7.2500   NaN      S
416         1308         0       3         Ware, Mr. Frederick  male   NaN    0    0   359309  8.0500   NaN      S
417         1309         0       3   Peter, Master. Michael J  male   NaN    1    1         2668 22.3583   NaN      C
```

Figura 1 - Utilização de `.head`, `.tail` e `.describe`

```
Resumo estatístico só de colunas numéricas:
  PassengerId  Survived  Pclass      Age      SibSp      Parch      Fare
count  418.000000  418.000000  418.000000  332.000000  418.000000  418.000000  417.000000
mean    1100.500000    0.363636    2.265550    30.272590    0.447368    0.392344    35.627188
std     120.810458    0.481622    0.841838    14.181209    0.896760    0.981429    55.907576
min      892.000000    0.000000    1.000000    0.170000    0.000000    0.000000    0.000000
25%     996.250000    0.000000    1.000000    21.000000    0.000000    0.000000    7.895800
50%    1100.500000    0.000000    3.000000    27.000000    0.000000    0.000000    14.454200
75%    1204.750000    1.000000    3.000000    39.000000    1.000000    0.000000    31.500000
max    1309.000000    1.000000    3.000000    76.000000    8.000000    9.000000    512.329200

Resumo estatístico só de colunas não-numéricas:
  PassengerId  Survived  Pclass      Name  Sex      Age      SibSp      Parch  Ticket   Fare      Cabin Embarked
count  418.000000  418.000000  418.000000      418    418  332.000000  418.000000  418.000000      418  417.000000      91    418
unique      NaN      NaN      NaN      418      2      NaN      NaN      NaN      NaN      363      NaN      76      3
top      NaN      NaN      NaN  Kelly, Mr. James  male      NaN      NaN      NaN  PC 17608      NaN  B57 B59 B63 B66      S
freq      NaN      NaN      NaN      1    266      NaN      NaN      NaN      5      NaN      3    270
mean    1100.500000    0.363636    2.265550      NaN      NaN    30.272590    0.447368    0.392344      NaN    35.627188      NaN      NaN
std     120.810458    0.481622    0.841838      NaN      NaN    14.181209    0.896760    0.981429      NaN    55.907576      NaN      NaN
min      892.000000    0.000000    1.000000      NaN      NaN    0.170000    0.000000    0.000000      NaN    0.000000      NaN      NaN
25%     996.250000    0.000000    1.000000      NaN      NaN    21.000000    0.000000    0.000000      NaN    7.895800      NaN      NaN
50%    1100.500000    0.000000    3.000000      NaN      NaN    27.000000    0.000000    0.000000      NaN    14.454200      NaN      NaN
75%    1204.750000    1.000000    3.000000      NaN      NaN    39.000000    1.000000    0.000000      NaN    31.500000      NaN      NaN
max    1309.000000    1.000000    3.000000      NaN      NaN    76.000000    8.000000    9.000000      NaN    512.329200      NaN      NaN
```

Figura 2 - Resumo estatístico

```
Informação geral sobre o DataFrame:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 418 entries, 0 to 417
Data columns (total 12 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   PassengerId  418 non-null    int64
1   Survived     418 non-null    int64
2   Pclass       418 non-null    int64
3   Name         418 non-null    object
4   Sex          418 non-null    object
5   Age         332 non-null    float64
6   SibSp       418 non-null    int64
7   Parch       418 non-null    int64
8   Ticket      418 non-null    object
9   Fare        417 non-null    float64
10  Cabin       91 non-null     object
11  Embarked    418 non-null    object
dtypes: float64(2), int64(5), object(5)
memory usage: 39.3+ KB
```

Figura 3 - Info geral sobre o DataFrame

Limpeza e Pré-Processamento

Para tratar dos valores nulos e ajustar os dados na coluna Age os valores ausentes foram preenchidos com a mediana das idades, a coluna Cabin foi descartada na análise inicial devido à alta quantidade de valores ausentes e na coluna Embarked os valores ausentes foram preenchidos com o valor mais frequente.

```
Atividade 3.2 - Limpeza e pré-processamento de dados
Valores nulos por coluna:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             86
SibSp            0
Parch           0
Ticket           0
Fare             1
Cabin           327
Embarked         0
dtype: int64
```

Figura 4 - Valores nulos existentes, por coluna

```
Valores nulos após tratamento:
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age             0
SibSp            0
Parch           0
Ticket           0
Fare             0
Cabin            0
Embarked         0
dtype: int64
Novo ficheiro CSV criado: titanic_tratado.csv
```

Figura 5 - Valores nulos após tratamento

Relativamente ao tipo de dados, consideramos não haver a necessidade de converter qualquer tipo.

```
Novo ficheiro CSV criado: titanic_tratado.csv

Tipos de dados após conversão:
PassengerId      int64
Survived         int64
Pclass           int64
Name             object
Sex              object
Age             float64
SibSp            int64
Parch            int64
Ticket           object
Fare            float64
Cabin           object
Embarked         object
```

Figura 6 - Tipo de dados

Nota: PassengerId, Survived, Pclass: Está como int64, o que está correto. Como é um identificador único, não precisa de conversão. Age, Flare: Estão corretamente como float64. Name, Ticket, Cabin, Embarked, Sex: Mantêm o tipo object (string).

Depois criamos uma coluna chamada *Idade_Milissegundos*, onde as idades foram convertidas para milissegundos.

	Age	Idade_Milissegundos
0	34.5	631152000000
1	47.0	220924800000
2	62.0	62
3	27.0	852076800000
4	22.0	1009843200000
5	14.0	1262304000000
6	30.0	757382400000
7	26.0	883612800000
8	18.0	1136073600000
9	21.0	1041379200000

Figura 7 - Idades em milissegundos desde o Epoch

Nota: Para idades inferiores à data do Epoch (1970-01-01) decidimos colocar o número inteiro pois apresentava número negativo em milissegundos.

Análise e Manipulação dos Dados

Realizamos algumas análises com as funções do Pandas, como *groupby* para calcular a taxa de sobrevivência por classe e género. Foi possível observar que a classe não interferiu na taxa de sobrevivência. Relativamente ao género, as mulheres apresentaram uma taxa de sobrevivência absoluta com nenhum homem a sobreviver.

```
Taxa de Mortalidade por sexo:
Sex
female    0.0
male      1.0
Name: Survived, dtype: float64
Taxa de Sobrevivência por Sexo (%):
Sex
female   100.0
male      0.0
Name: Survived, dtype: float64
Taxa de Sobrevivência por Classe e Sexo (%):
Pclass  Sex
1      female   100.0
       male      0.0
2      female   100.0
       male      0.0
3      female   100.0
       male      0.0
Name: Survived, dtype: float64
```

Figura 8 - Taxa de sobrevivência por classe e género

De seguida analisámos qual a idade média de um sobrevivente o que nos fez concluir que a probabilidade de sobreviver é mais favorável para pessoas mais jovens.

```
Name: Survived, dtype: float64
Média de Idade por Sobrevivência:
Survived
0    29.522218
1    29.734145
```

Figura 9 - Média de idade dos sobreviventes

Na próxima análise, vamos interpretar os dados da classe vs sobrevivência e perceber se o status e/ou a idade prevaleceu sobre a decisão dos que sobreviviam.

Idade Média por Classe e Sobrevivência:

Pclass	Survived	
1	0	38.859649
	1	40.760000
2	0	30.690476
	1	24.464000
3	0	25.372671
	1	24.273194

Name: Age, dtype: float64

Figura 10 - Idade média por classe e sobrevivência

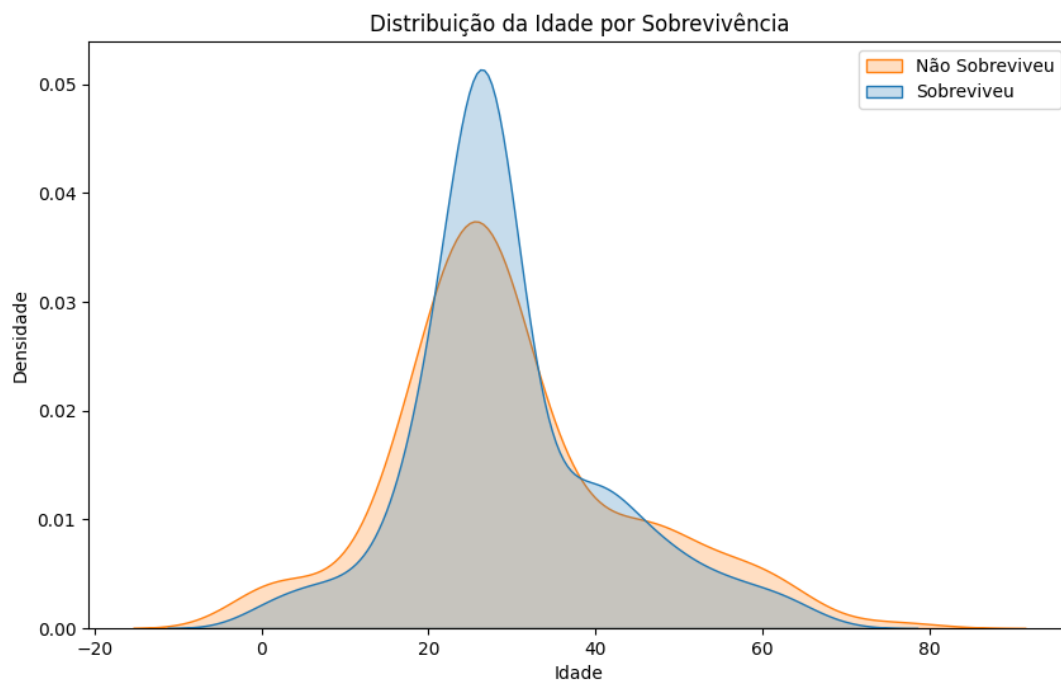


Figura 11 - Distribuição da Idade por Sobrevivência

Interpretação dos dados:

Classe 1 (Pclass = 1):

- Não sobreviveram (Survived = 0): A idade média dos passageiros que não sobreviveram nesta classe foi de aproximadamente 38,86 anos.
- Sobreviveram (Survived = 1): A idade média dos passageiros que sobreviveram nesta classe foi de aproximadamente 40,76 anos.

Conclusão: Pode indicar que passageiros mais velhos tinham maior probabilidade de sobrevivência nessa classe, ou que havia mais passageiros mais velhos sobreviventes na 1ª classe.

Classe 2 (Pclass = 2):

- Não sobreviveram (Survived = 0): A idade média foi de 30,69 anos.
- Sobreviveram (Survived = 1): A idade média foi de 24,46 anos.

Conclusão: Pode indicar que passageiros mais jovens tiveram maior probabilidade de sobreviver nesta classe.

Classe 3 (Pclass = 3):

- Não sobreviveram (Survived = 0): A idade média foi de 25,37 anos.
- Sobreviveram (Survived = 1): A idade média foi de 24,27 anos.

Conclusão: Os sobreviventes tendem a ser ligeiramente mais jovens.

Portanto, estes padrões podem ser influenciados por fatores como:

- Acesso a botes salva-vidas (maior nas classes mais altas).
- Prioridades dadas durante o resgate (potencialmente favorecendo crianças e mulheres).

Visualização de dados

No gráfico seguinte podemos ter uma visão mais clara sobre a classe e sexo dos sobreviventes onde imediatamente podemos concluir que relativamente às mulheres houve um número muito baixos de sobreviventes de segunda classe e um número mais alto para mulheres de terceira classe.

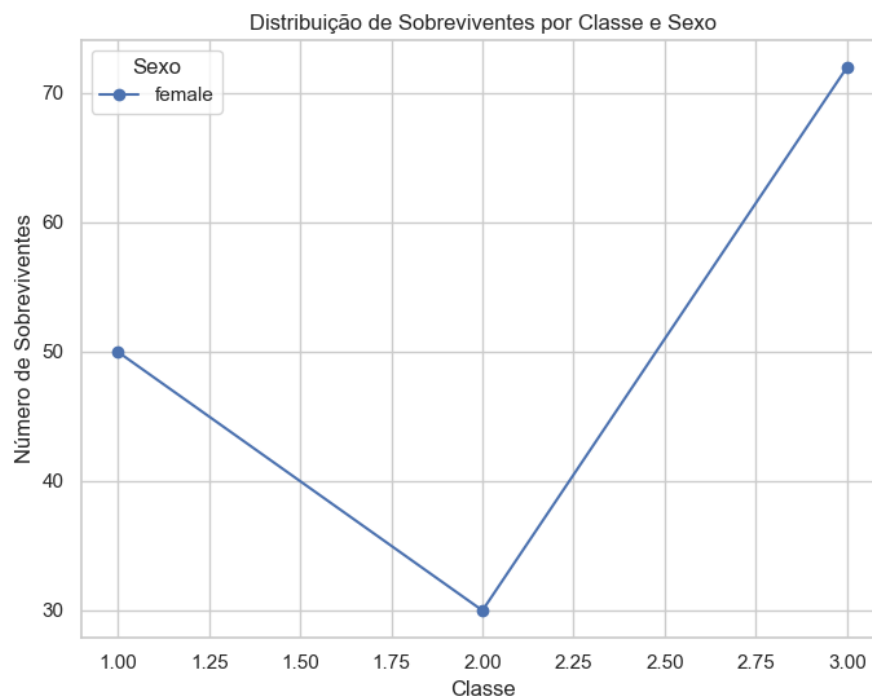


Figura 12 - Gráfico de Linha: Representa a distribuição dos sobreviventes por classe e sexo

Analisando as idades com a tarifa e sobrevivência podemos observar que a tarifa não teve impacto na taxa de sobrevivência pois maior parte dos sobreviventes pagaram uma tarifa Fare menor que 100. O número 0 representa os não-sobreviventes enquanto que o número 1 representa os sobreviventes.

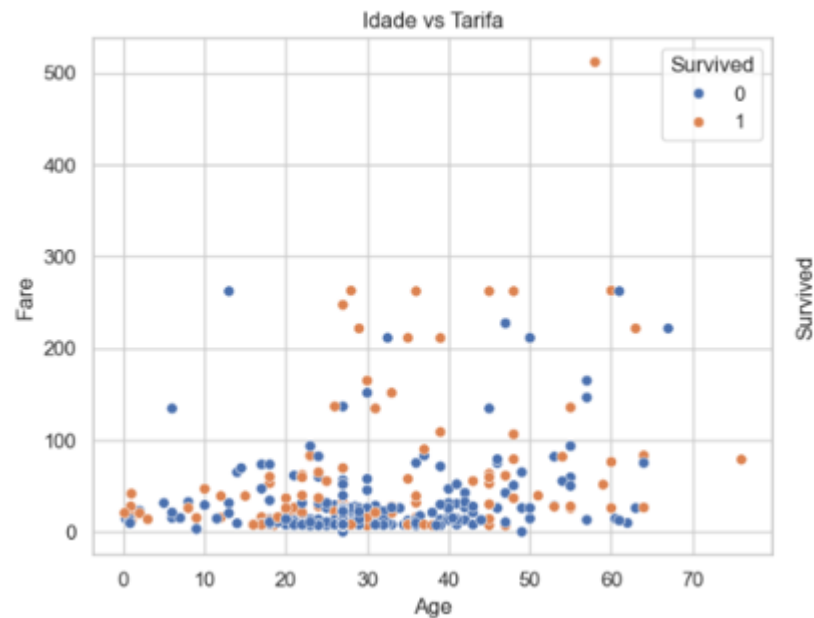


Figura 13 - Gráfico de Dispersão: Representa a correlação entre as variáveis Age, Fare e Survived

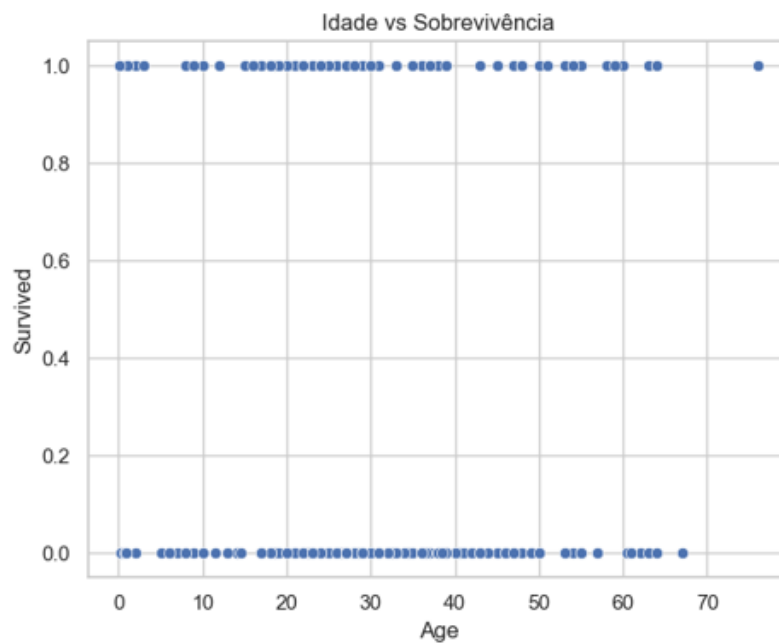


Figura 14 - Gráfico de Dispersão: Representa a correlação entre Idade e Sobrevivência

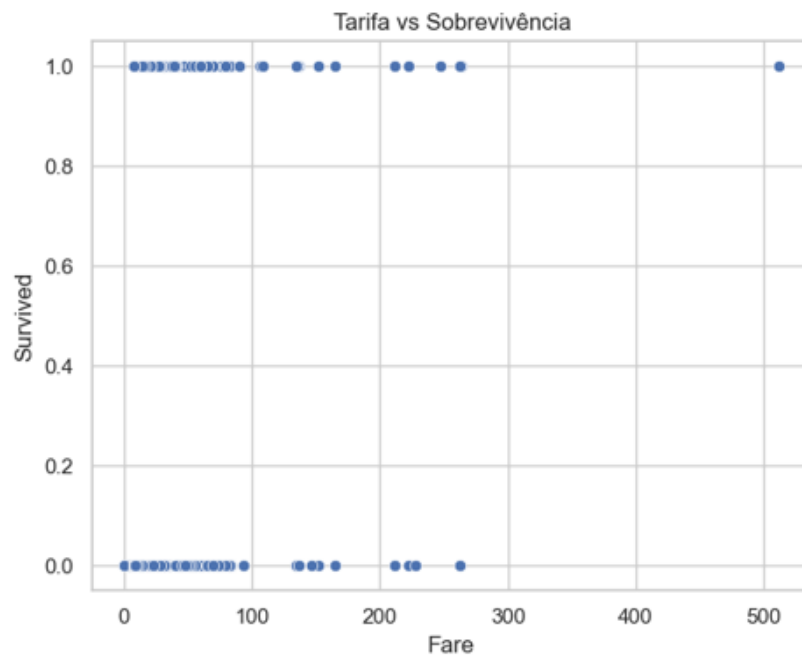


Figura 15 - Gráfico de Dispersão: Representa a correlação entre a Tarifa vs Sobrevivência

No histograma seguinte podemos visualizar melhor a distribuição de idades sobre a taxa de sobrevivência. O número 0 representa os não-sobreviventes enquanto que o número 1 representa os sobreviventes.

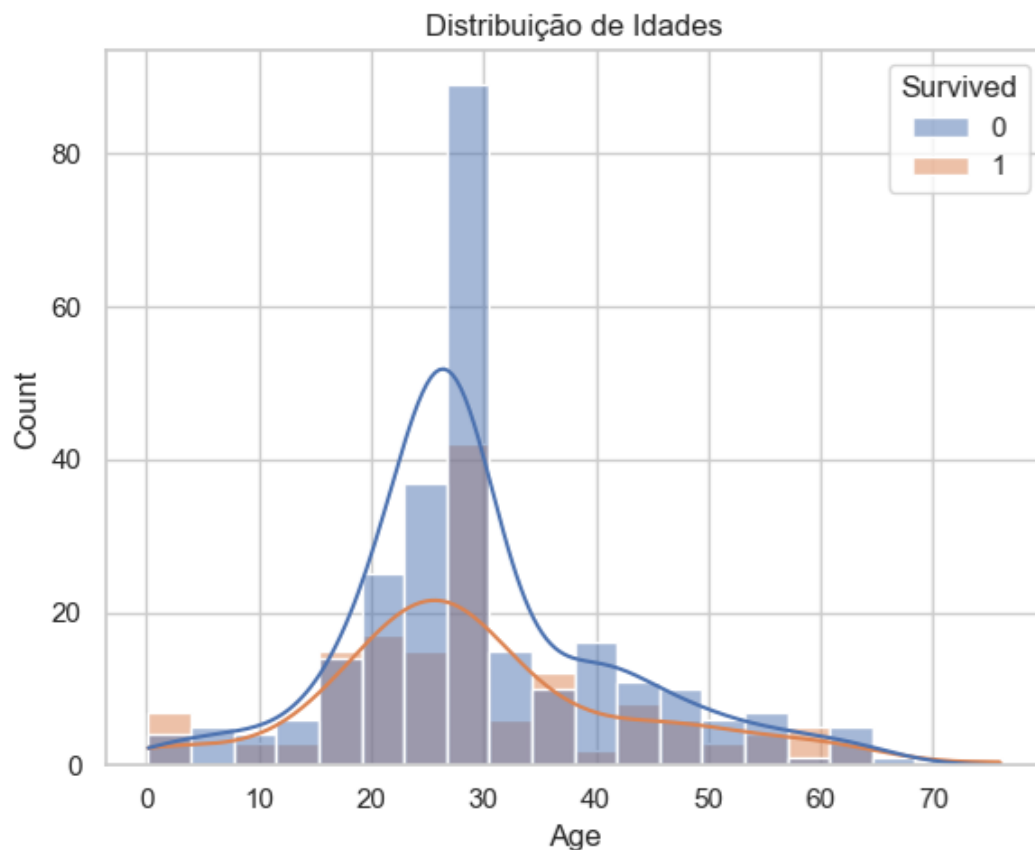


Figura 16 - Gráfico de Barras: Distribuição das Idades

Análise: Pode-se concluir que o grosso das pessoas com idade entre 20-30 anos foi o que não sobreviveu, e ainda com 40 anos. Por outro lado, foi entre os 0-30 anos que mais sobreviventes se pode observar, e ainda na casa dos 45-50 anos.

Neste histograma podemos visualizar a distribuição das tarifas sobre a taxa de sobreviventes.

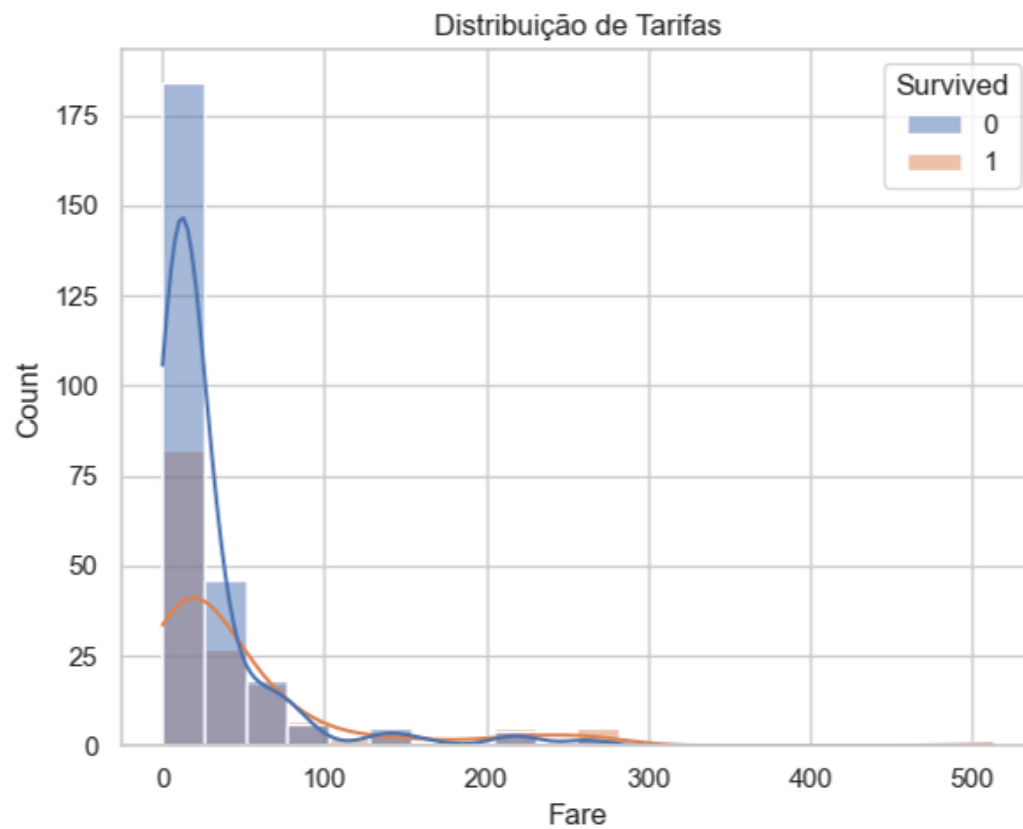


Figura 17 - Gráfico de Barras: Distribuição de Tarifas

Análise: O número de não sobreviventes que adquiriram uma tarifa mais barata é bastante visível, enquanto quem pagou uma tarifa mais cara, teve mais chances de sobreviver.

E neste gráfico podemos observar melhor a diferença do número de passageiros que sobreviveram e o número de passageiros que não sobreviveram em que claramente houve muitos mais passageiros que não sobreviveram.

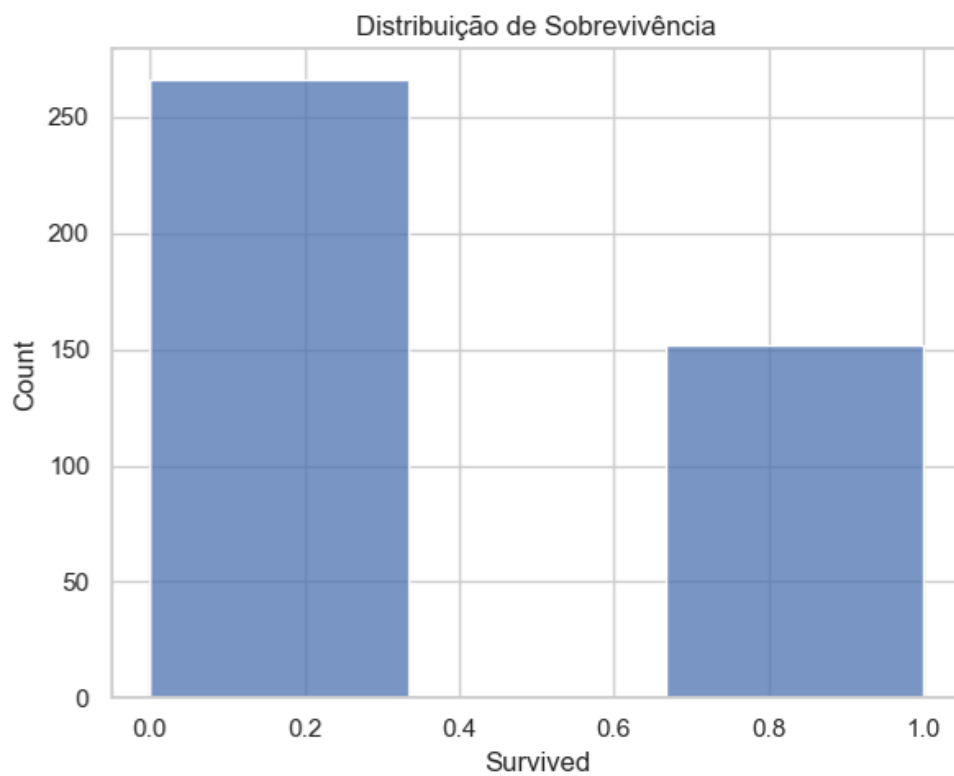


Figura 18 - Gráfico de Barras: Distribuição da Sobrevivência

Análise Adicional – Relação entre Família e Sobrevivência

Uma análise adicional que fizemos, a qual necessitou de se criar uma nova coluna, foi explorar qual a influência do tamanho da família na taxa de sobrevivência e obtemos o seguinte gráfico que mostra que um número de membros familiares de 4 e 6 apresentam uma taxa de sobrevivência relativamente mais elevada que passageiros com 1, 5 e 7 membros familiares com os restantes valores por volta de 50% de probabilidade de sobrevivência.

```
Atividade 3.7 Análise Adicional - Relação entre Família e Sobrevivencia
Tamanho_Familia
1      0.268775
2      0.486486
3      0.526316
4      0.714286
5      0.285714
6      0.666667
7      0.250000
8      0.500000
11     0.500000
Name: Survived, dtype: float64
```

Figura 19 - Prova de criação da nova coluna Tamanho_Familia

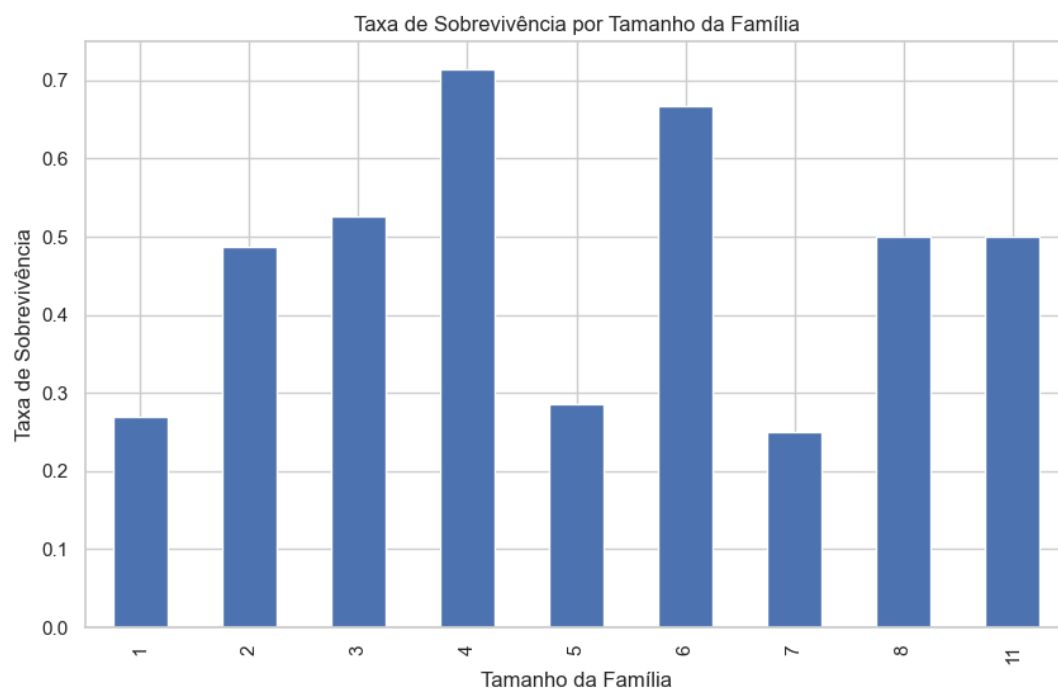


Figura 20 - Gráfico de Barras: Relação entre o Tamanho das Famílias e a Sobrevivência

Análise Adicional – Comparação por Porto de Embarque

Ao analisarmos o gráfico obtido da taxa de sobrevivência por porto de embarque concluímos que os passageiros que embarcaram em Queenstown tiveram uma maior probabilidade de sobreviver seguido do porto de Cherbourg e por fim os passageiros de Southampton tiveram a menor probabilidade de sobrevivência.

```
Atividade 3.7 Análise Adicional - Comparação por Porto de Embarque
Embarked
C    0.392157
Q    0.521739
S    0.325926
Name: Survived, dtype: float64
```

Figura 21 - Coluna Embarked

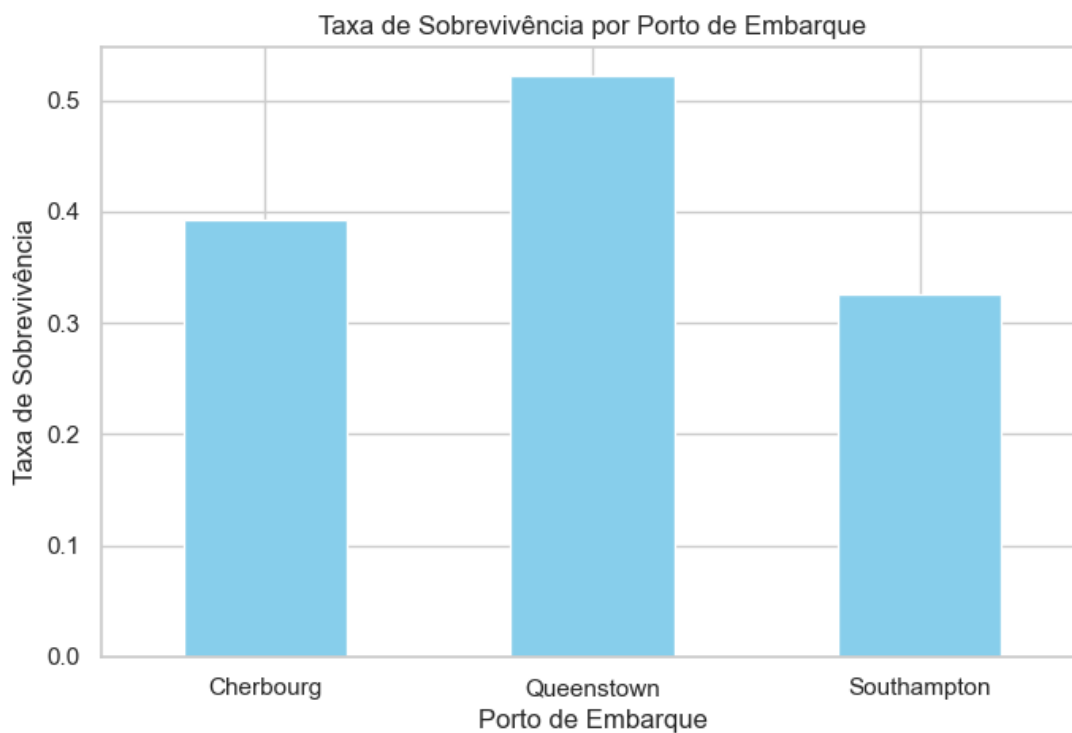


Figura 22 - Gráfico de Barras: Taxa de sobrevivência por Porto de Embarque

Exportação dos resultados

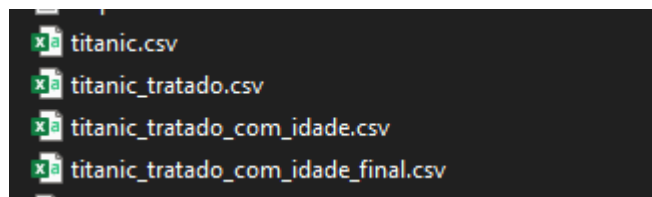
Ao longo do projecto, foi necessário exportar para ficheiro Excel, aquando da conclusão das atividades que resultou na seguinte criação de ficheiros .csv:

```
dtype: int64  
Novo ficheiro CSV criado: titanic_tratado.csv  
  
Tipos de dados após conversão:
```

Figura 23 - Ficheiro .csv com dados tratados

```
Name: Survived, dtype: float64  
Novo ficheiro CSV criado: titanic_tratado_com_idade_final.csv  
Traceback (most recent call last):
```

Figura 24 - Ficheiro final .csv criado



```
titanic.csv  
titanic_tratado.csv  
titanic_tratado_com_idade.csv  
titanic_tratado_com_idade_final.csv
```

Figura 25 - Amostra dos Ficheiros.CSV

Utilização do Git

Neste trabalho foi utilizada a plataforma Git, para que todos os elementos do grupo pudessem trabalhar em simultâneo, e atualizar o projeto conforme as funcionalidades criadas:

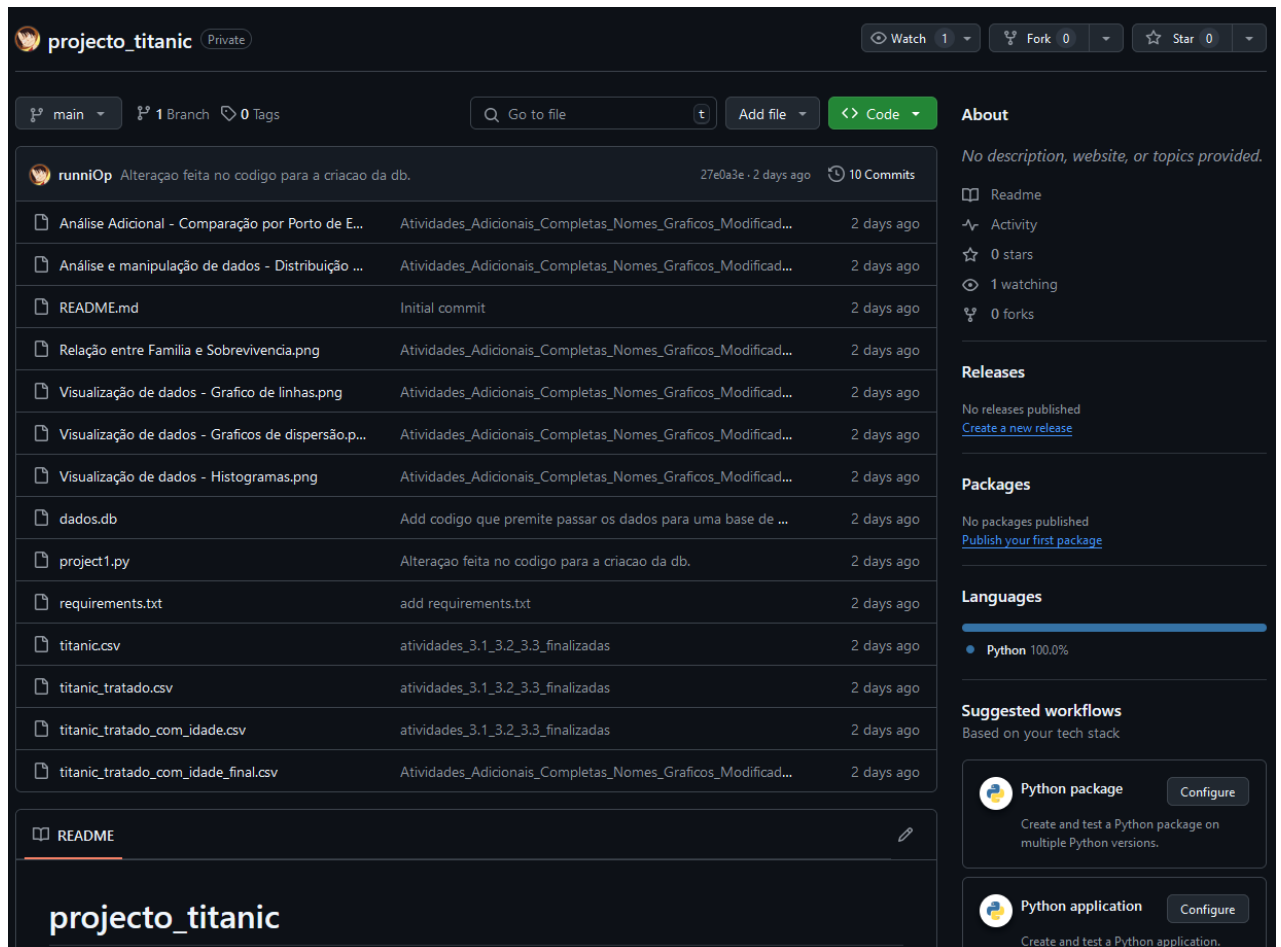


Figura 26 - Utilização do Git

Conclusão

Neste relatório analisámos os dados sobre os passageiros do Titanic e desta forma conseguimos identificar padrões que influenciaram a sobrevivência ao naufrágio de 1912.

Nos dados analisados encontramos claras evidências da política "mulheres e crianças primeiro" com taxas de sobrevivência mais altas para mulheres e passageiros mais jovens. Os passageiros da 3ª e 1ª classe tiveram mais sobreviventes que 2ª classe, onde destacamos a posição e acessibilidade aos botes salva-vidas.

Relativamente a tarifas não houve um impacto o sendo que os que tiveram tarifas inferiores a 100 tinham maior taxa de sobrevivência.

Com esta análise, foi possível perceber que os fatores sociais e económicos tiveram a sua influência na sobrevivência deste desastre.

Por fim, achamos que este trabalho foi uma boa experiência, que nos permitiu consolidar os conceitos da Programação em Python, e como é uma mais valia na Analise de Dados.