Data Science I - Projeto final ¶

Conclusão

Ao analisar a amostra de dados de passageiros do Titanic, foi possível chegar as seguintes conclusões.

- A primeira classe apresentou o maior percentual de sobreviventes, com 62,6%, contra 47,2% da segunda e 24,2% da terceira. Tal diferença pode apontar que houve alguma facilidade ou favorecimento no acesso aos botes salva vidas para os passageiros da primeira classe.
- Dentre os sobreviventes, 68% são mulheres e 32% homens
- O grupo de classe e sexo que apresentou maior taxa de sobreviventes foi o feminino da primeira classe, com a relação de 96,7%
- 59,0% das crianças a bordo foram salvas
- A idade média entre os sobreviventes era de 28,4 anos.

Limitações

Inicialmente ao avaliar o conjunto de dados foi detectado que haviam campos como idade, embarcou e cabine, não preenchidos e que potencialmente prejudicariam a análise. Assim, foram feitos alguns ajustes como:

- Campos de idade com valor nulo foram preenchidos com a média
- Foi efetuada uma breve análise sobre os passageiros que apresentavam o campo 'embarcou' como nulo e se haviam sobrevivido, assim, dado que o retorno foi positivo, estes foram removidos do conjunto de dados considerando que estes não embarcaram.

Valores abreviados em local de embarque (embarcou) foram substituídos pelo nome do local sem abreviação.

Foram aplicadas traduções nos nomes das colunas e nos valores dos campos sobreviveu e sexo para auxiliar na criação de legendas dos gráficos.

Foi criado um campo de categorização da idade dos passageiros com o objetivo de mapea-los facilmente e levantar informações a respeito de cada grupo.

Apresentação

Mesmo após um século de seu naufrágio, que ocorreu em 1912, o Titanic é considerado um dos maiores desastres marítimos em tempos de paz. Sua história rendeu livros, filmes e diversos documentarios que buscam explorar e levar informações e curiosidades a seu respeito ao público. Nesse mesmo sentido, esse projeto tem como objetivo explorar o conjunto de dados do Titanic e tentar responder uma série de perguntas pertinentes. O arquivo está disponível no formato CSV através do link (https://d17h27t6h515a5.cloudfront.net/topher/2017/October/59e4fe3d_titanic-data-6/titanic-data-6.csv).

No primeiro instante, ao carregar o conjunto de dados, foi necessario avaliar suas características como, seu esquema de organização, consistência dos dados e a necessidade possíveis correções e adaptações que pudessem contribuir na manipulação e pesquisa. Neste passo foram aplicadas as seguintes mudanças:

- Nomes de colunas e valores foram traduzidos para o português de modo que ficassem no mesmo idioma da análise;
- Valores nulos na coluna 'idade' foram preenchidos com o valor médio;
- Passageiros foram classificados por idade como Criança, menores de 15 anos, Jovem, entre 15 e 25, Adulto, entre 25 e 65, e Idoso para maiores de 65 anos;
- Valores abreviados na coluna 'embarcou' foram substituidos pelo nome correspondente sem abreviação;
- Passageiros que sobreviveram e não possuiam local de embarque foram removidos do conjunto, considerando que não embarcaram;

Feito isso, foram levantadas perguntas das quais poderiam ser respondidas com o conjunto de dados disponível, são elas:

1. Quantos sobreviveram?
2. Quantos morreram?
3. Quantas crianças sobreviveram?
4. Quantas crianças morreram?
5. Qual a idade média entre os sobreviventes?
6. Qual a idade média entre os que morreram?
7. Qual a relação dos sobreviventes com a classe de ingresso?
8. Qual o sexo mais relevante entre os sobreviventes?
9. Qual a relação dos sobreviventes com a classe e sexo?
10. Qual ponto de embarque recebeu mais passageiros?
11. Do sexo masculino, qual é a idade do sobrevivente mais velho?
12. Do sexo masculino, qual é a idade do sobrevivente mais novo?
13. Do sexo feminino, qual é a idade da sobrevivente mais velha?
14. Do sexo feminino, qual é a idade da sobrevivente mais nova?

Perguntas

1 e 2, Quantos sobreviveram e quantos morreram?

Complementares, a primeira e a segunda pergunta foram respondidas ao agrupar os dados por tipo de valor na coluna 'sobreviveu' e apontam que, conforme a tabela e gráfico abaixo, 340 (38,25%) pessoas sobreviveram e outras 549 (61,75%) morreram.

```
In [328]: plot_survivors()
```

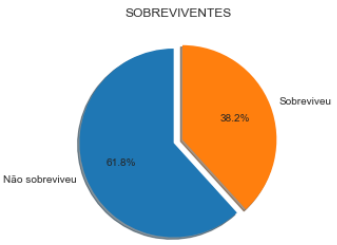


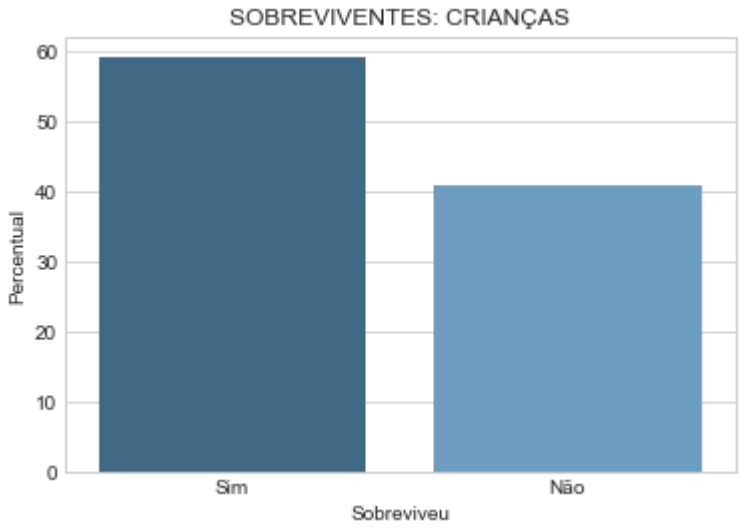
TABELA DE SOBREVIVENTES

SOBREVIVEU	TOTAL	PROPORÇÃO
Não	549	61.75
Sim	340	38.25

3 e 4, Quantas crianças sobreviveram e quantas morreram?

Para as perguntas três e quatro, que também são complementares, foram filtrados os registros que apresentavam o valor 'Criança' na coluna 'categoria_idade' para posteriormente contabilizar os valores de sobreviventes.

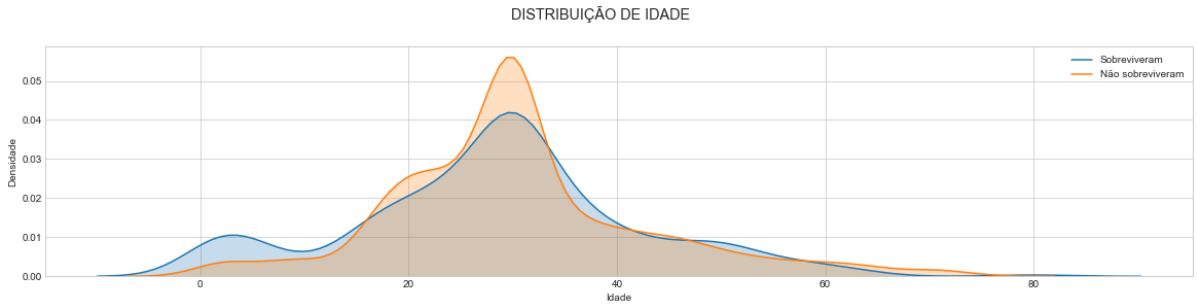
```
In [329]: plot_children_balance()
```



5 e 6, Qual a idade média entre os sobreviventes? E entre os que não sobreviveram?

Os sobreviventes aprensentavam idade média de 28,47 anos e os que não sobreviveram com 30,48. Ambos aprensetaram um pico de densidade de passageiros entre 20 e 40 anos, sendo de sobreviventes levemente superior a 4% e dos não sobreviventes superior a 5%.

```
In [330]: plot_age_dist()
```



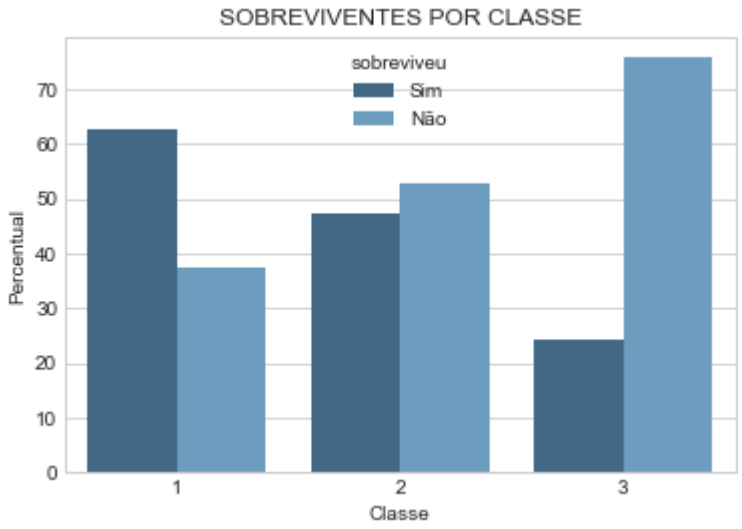
7. Qual a relação dos sobreviventes com a classe de ingresso?

A classe que apresentou maior número de sobreviventes foi a primeira com 62,62%, seguida da segunda classe com 52,72% e a terceira classe com 24,24%. Podemos concluir que algum fator, possivelmente, favoreceu os passageiros da primeira classe no acesso aos botes salva vidas.

In [331]:
```
plot_surv_by_class()
```

Out[331]:

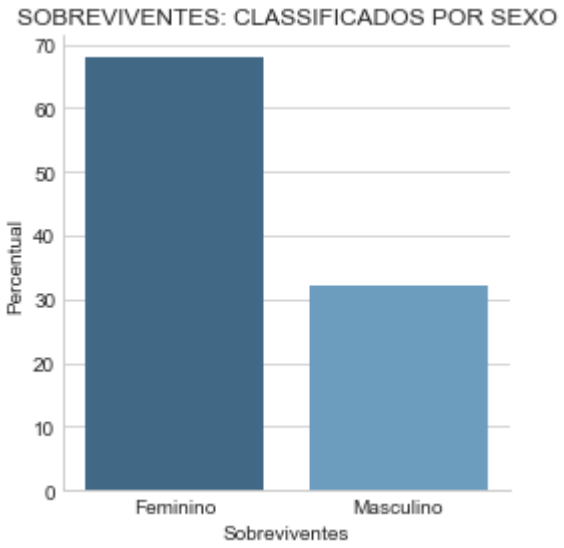
	classe	sobreviveu	percentual
0	1	Sim	62.62
1	1	Não	37.38
2	2	Não	52.72
3	2	Sim	47.28
4	3	Não	75.76
5	3	Sim	24.24



8. Qual o sexo mais relevante entre os sobreviventes?

O sexo mais relevante entre os sobreviventes é o feminino com 67,94%.

In [332]:
```
sex_survivors_compare()
```



In [333]:
```
df.query('sobreviveu == "Sim"')['sexo'].value_counts(normalize=True)
```

Out[333]:
```
Feminino     0.679412
Masculino    0.320588
Name: sexo, dtype: float64
```

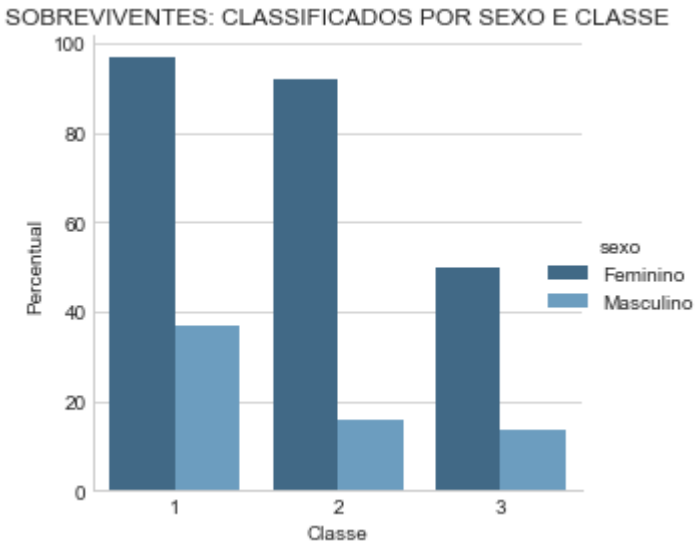
9. Qual a relação dos sobreviventes com a classe e sexo?

Ao analisarmos a proporção de sobreviventes por gênero e classe é possível notar como as mulheres da primeira e segunda se destacam em relação a seu grupo. Sobreviveram, 96,74% das mulheres da primeira classe e 92,11% das mulheres da segunda classe, ao passo que a proporção de homens, respectivamente, foi de 36,89% e 15,74%. O gênero que apresentou pior proporção foi o masculino na terceira classe com apenas 13,54% de sobreviventes.

In [334]:
```
plot_survivors_by_class_sex()
```

Out[334]:

	classe	sexo	sobreviveu	percentual
0	1	Feminino	Sim	96.74
3	1	Masculino	Sim	36.89
4	2	Feminino	Sim	92.11
7	2	Masculino	Sim	15.74
9	3	Feminino	Sim	50.00
11	3	Masculino	Sim	13.54



10. Qual ponto de embarque recebeu mais passageiros?

In [335]:
```
df['embarcou'].value_counts().to_frame()
```

Out[335]:

	embarcou
Southampton	644
Cherbourg	168
Queenstown	77

O ponto de embarque que recebeu mais passageiros foi de Southampton, ao sul do Reino Unido e também ponto de partida do navio.

11. Do sexo masculino, qual é a idade do sobrevivente mais velho?

```
In [336]: df.query('sobreviveu == "Sim" & sexo == "Masculino"').idade.max()

Out[336]: 80.0
```

O sobrevivente mais velho do sexo masculino tinha 80 anos.

12. Do sexo masculino, qual é a idade do sobrevivente mais novo?

```
In [337]: df.query('sobreviveu == "Sim" & sexo == "Masculino"').idade.min()

Out[337]: 0.42
```

O sobrevivente mais novo do sexo masculino tinha 0.42 ano (5 meses);

13. Do sexo feminino, qual é a idade da sobrevivente mais velha?

```
In [338]: df.query('sobreviveu == "Sim" & sexo == "Feminino"').idade.max()

Out[338]: 63.0
```

A sobrevivente mais velha do sexo feminino tinha 63 anos.

14. Do sexo feminino, qual é a idade da sobrevivente mais nova?

```
In [339]: df.query('sobreviveu == "Sim" & sexo == "Feminino"').idade.min()

Out[339]: 0.75
```

A sobrevivente mais nova do sexo feminino tinha 0.75 ano (9 meses).

Fim

Este é um projeto aberto, não tem como objetivo encontrar respostas definitivas. Há várias perguntas a serem feitas, campos e outros conjuntos de dados a serem explorados que podem esclarecer mais sobre este desastre marítimo que levou consigo tantas vidas.

Consultas e referências

- Seaborn API documentation (https://seaborn.pydata.org/api.html)
- Stack Overflow: Seaborn (https://stackoverflow.com/questions/33524694/plotting-with-seaborn)
- Pandas documentation (https://pandas.pydata.org/pandas-docs/stable/)
- Stack Overflow: Pandas (https://stackoverflow.com/questions/tagged/pandas)
- Matplotlib documentation (https://matplotlib.org/contents.html)