Relatorio

December 2, 2017

1 Programa Nanodegree Fundamentos de Data Science II

1.0.1 Marlesson R. O. de Santana

Public: https://nbviewer.jupyter.org/github/marlesson/Titanic-Data-Visualizer/blob/master/Data%20Visualization.ipynb

1.1 Resumo

O naufrágio do RMS Titanic é um dos naufrágios mais famosos da história. Em 15 de abril de 1912, durante sua viagem inaugural, o Titanic afundou depois de colidir com um iceberg, matando 1502 de 2224 passageiros e tripulantes.

Uma das razões pelas quais o naufrágio levou a uma grande perda de vidas era que não havia botes salva-vidas suficientes para os passageiros e tripulantes. Embora houvesse algum elemento de sorte envolvido na sobrevivência do naufrágio, alguns grupos de pessoas eram mais propensos a sobreviver do que outros, como mulheres, crianças e a classe alta. Vamos analisar o quanto esses fatores propiciaram para a sobrevivência do passageiro.

1.2 Design

O designer utilizado para representar os dados foram os gráficos de Barra, Histogramas e gráficos de tendência.

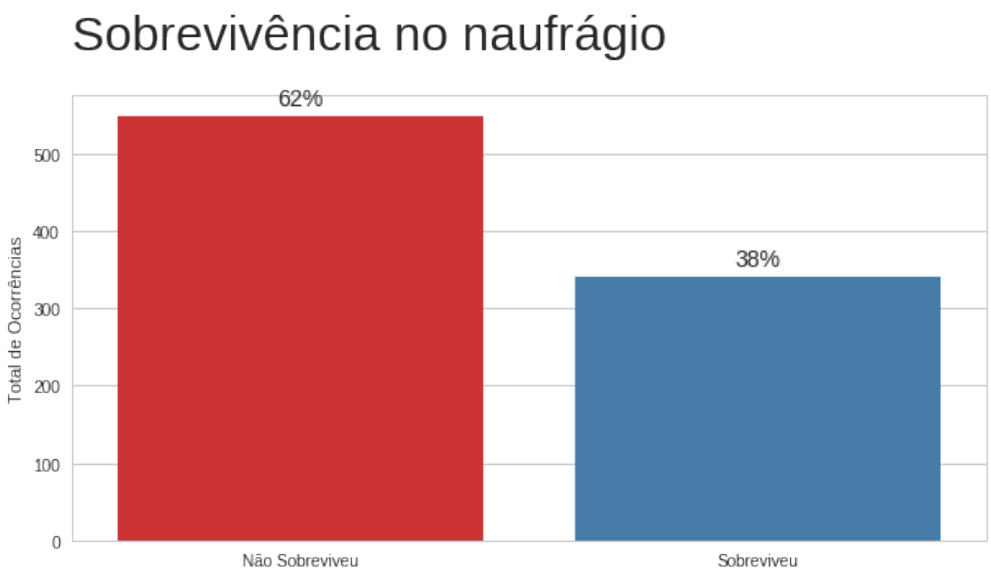
1.2.1 Paleta de cores

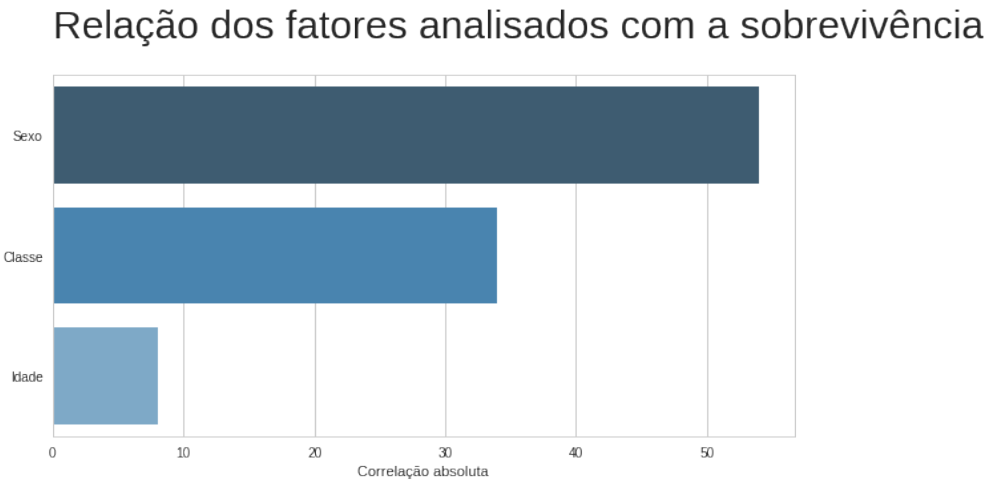
Foi dado uma preferência a **paleta de cores qualitativas** na maioria dos gráficos, por se tratar de comparações entre dois grupos principais (sobreviventes e não sobreviventes) foi dado uma preferência nas cores **vermelha** para indicar os não sobreviventes e a cor **azul** para indicar os sobreviventes. Exemplo:

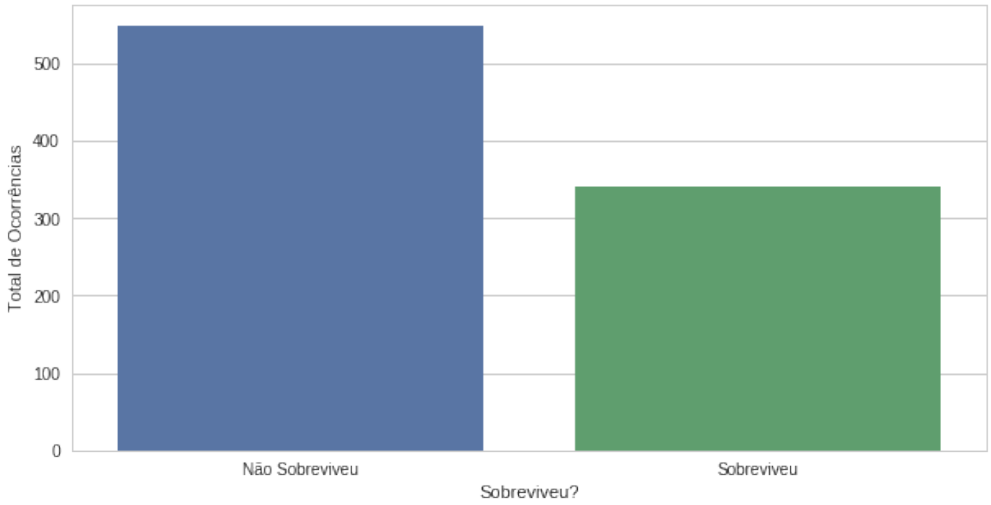
A **paleta de cores sequencial** foi utilizada apenas quando a comparação não é em sentido opostom, utilizando a a mesma cor em tonalidades diferentes para indicar a relevância de alguma informação. Exemplo:

1.2.2 Melhorias de Designer

Alguns elementos nos gráficos foram melhorando desde a primeira versão. Como exemplo, o gráfico de "Sobrevivência no naufráfio" em sua primeira versão não apresentava nem título nem a porcentagem, veja abaixo:


Sobrevivência no naufrágio
62%
38%
Total de Ocorrências
0
100
200
300
400
500
Não Sobreviveu
Sobreviveu


Relação dos fatores analisados com a sobrevivência
Sexo
Classe
Idade
0
10
20
30
40
50
Correlação absoluta



Primeira Versão

Na última versão do gráfico de “Sobrevivência no naufráfio” foram adicionaos os elementos para facilitar o entendimento.

Ao adicionar a porcentagem como número em cima das barras, facilitou o entendimento de porcentagem que antes era implícipo pelo tamanho da barra. Também foi retirada a legenda do eixo X, visto que já tem informações suficiêntes para compreender o contexto. Com relação ao título do gráfico, essa mudança foi em todos os gráficos, que na versão apresenta sempre o título.

Um outro exemplo das mesmas modificações, onde é perceptível a melhora no entendimento do gráfico, estão nos gráficos abaixo.

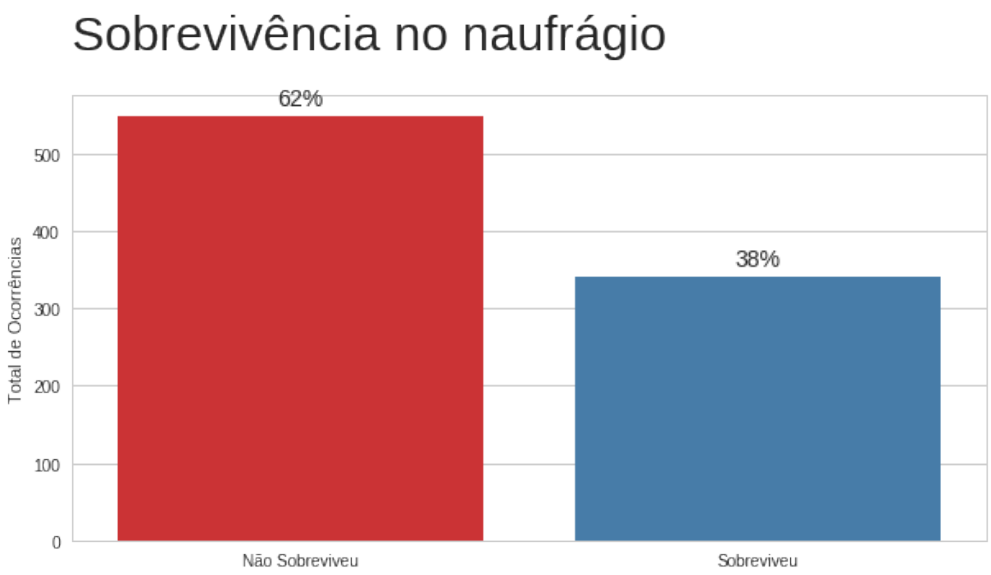
Antes:

Depois:

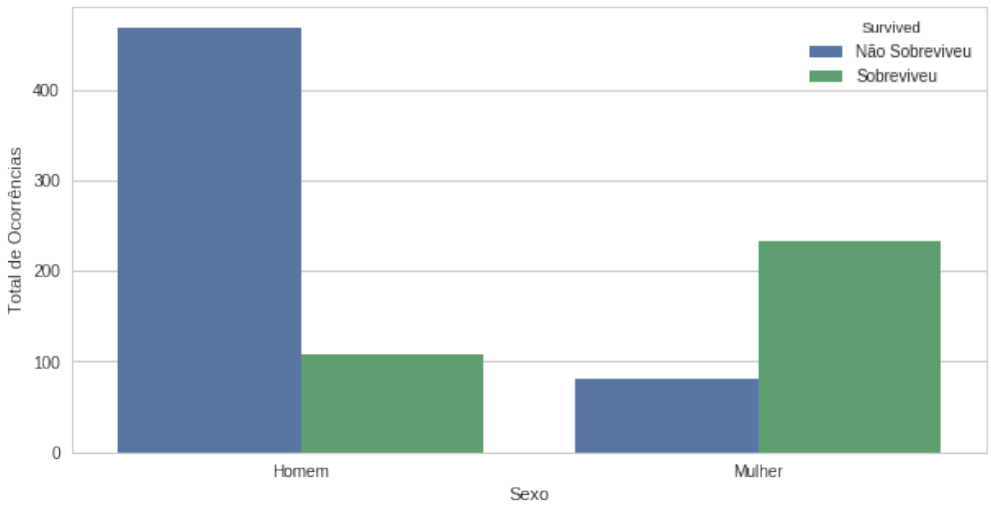
Um outro designer interessante é o de histograma, onde é possível analisar diferentes distribuições de dados para chegar a uma conclusão. A imagem abaixo é a versão final do comparativo de distribuição de idades para o grupo de Sobreviventes e Não sobreviventes.

As vezes é necessário modificar uma visão ou simplesmente adicionar um outro gráfico para evidenciar uma informação. Na escolha abaixo, a primeira versão conta com apenas o gráfico de barras, deixando implícito a tendência entre as classes.

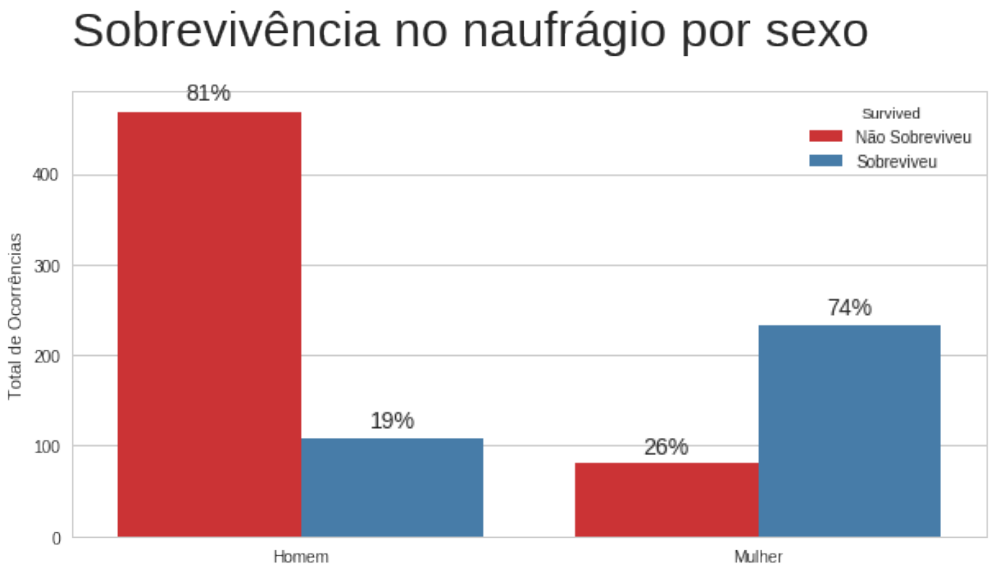
Na evolução do gráfico, além das informações de título e porcentagem, foi adicionada também um novo gráfico para evidênciar a tendência da queda na sobrevivência pela classe.



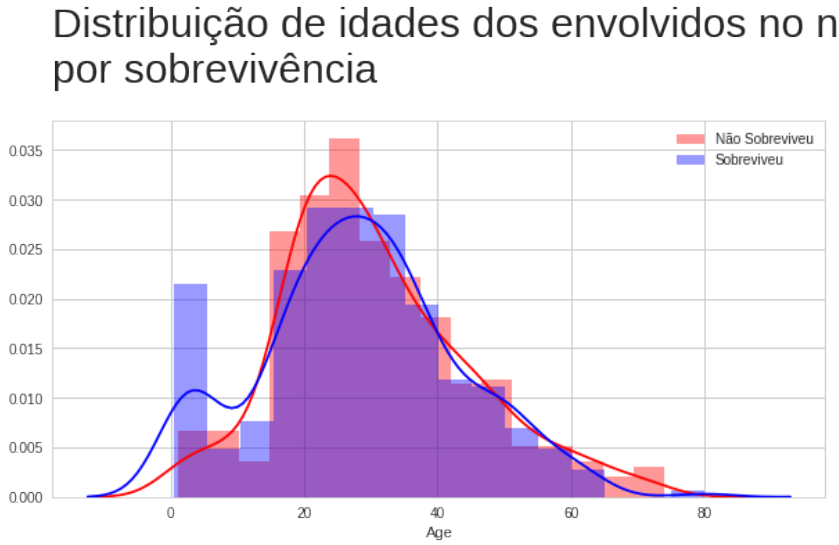
Última Versão



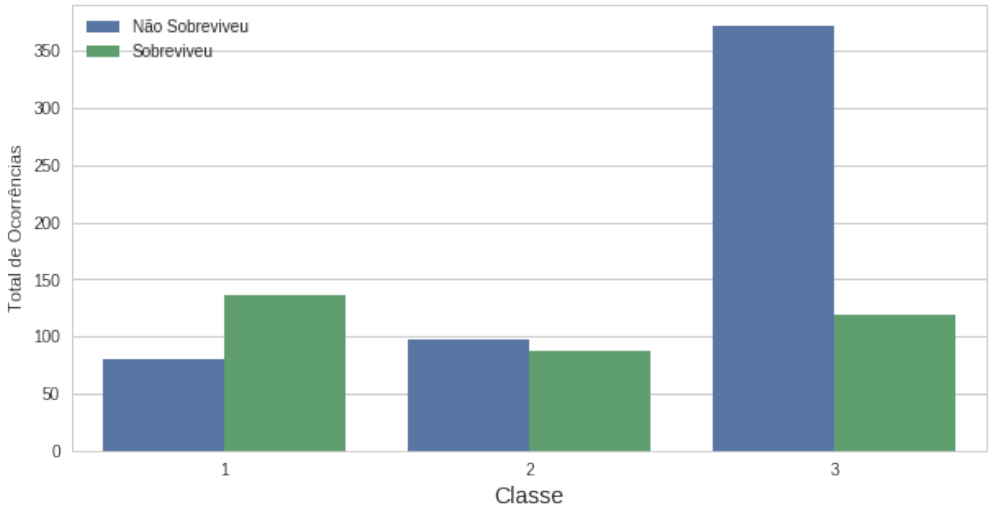
Primeira Versão



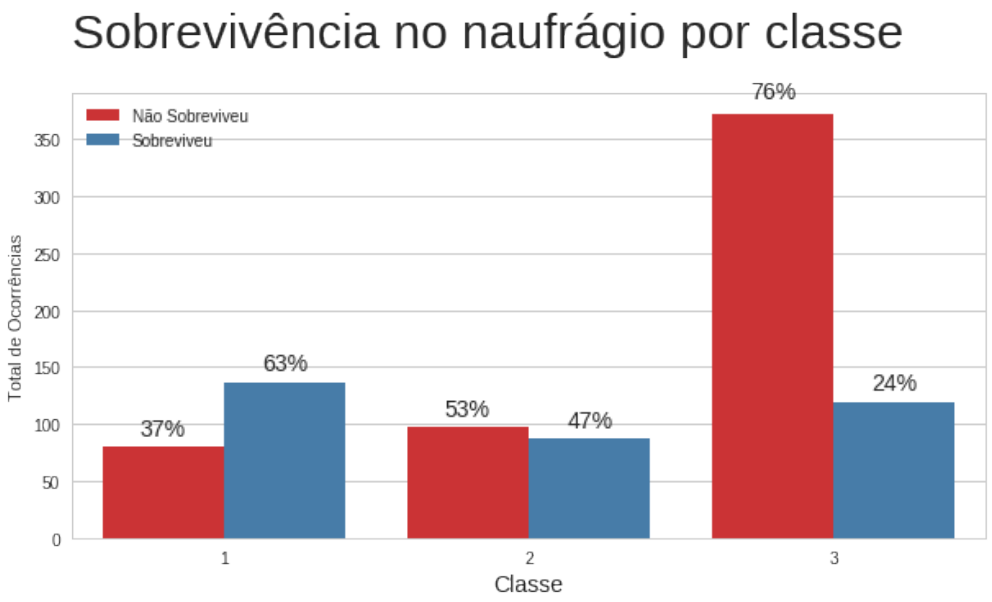
Última Versão

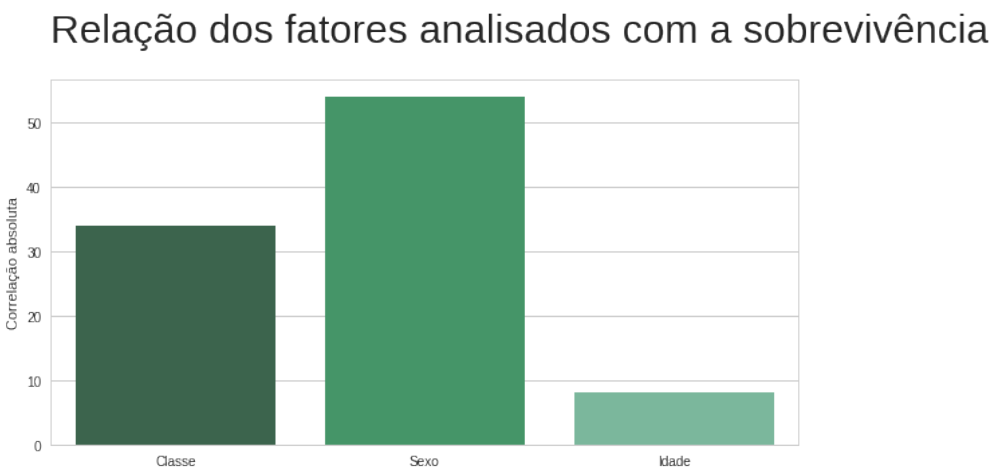


Primeira Versão

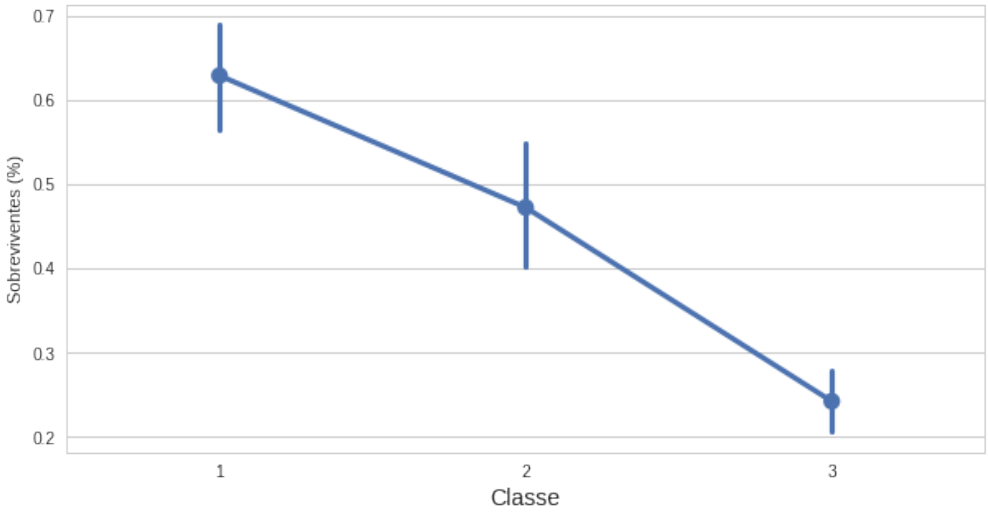


Primeira Versão





Original



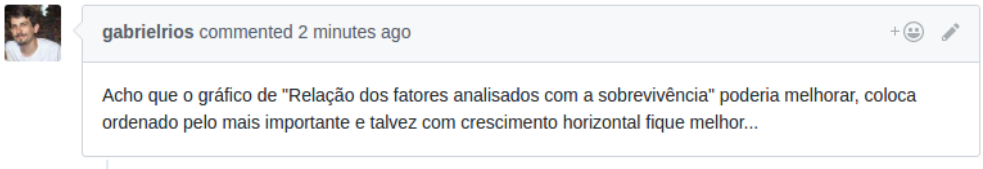
Outros exemplos podem ser encontrados na visualização

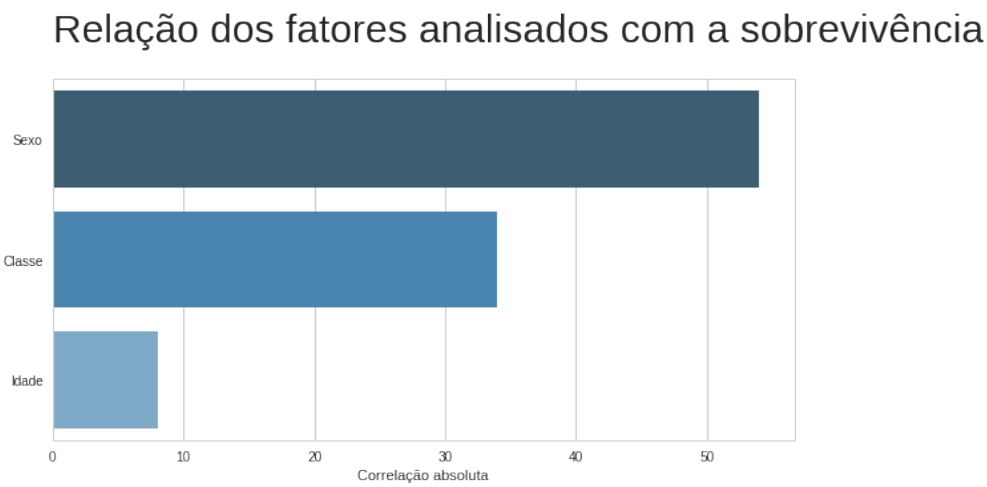
1.3 Feedback

A visualização está pública em https://nbviewer.jupyter.org/github/marlesson/Titanic-Data-Visualizer/blob/master/Data%20Visualization.ipynb para feedbacks pelo GitHub.

1.3.1 1. Comentário

A partir do comentário abaixo, o gráfico foi aterado para melhoarar o entendimento.





Alterada

Original

Alterado

1.4 Recursos

Os recursos utilizados foram:

- **Python**: Linguagem de programação amplamente utilizada para Data Science e Machine Learning
  - https://www.python.org/
- **Kaggle**: Comunidade mundialmente conhecida que centraliza análises de dados e compeonatos de Data Science e Machine Learning. O dataset utilizado nesse trabalho foi baixado do Kaggle https://www.kaggle.com/c/titanic/data
  - http://kaggle.com
- **Seaborn** Biblioteca python de visualização de dados
  - https://seaborn.pydata.org/
- **Matplotlib** Biblioteca python de visualização de dados
  - https://matplotlib.org/examples/index.html
- **Jupyter** Software open-source com utilização interativa de Python, amplamente utilizada para análise de dados.
  - http://jupyter.org/

```
In [ ]:
```