Conclusões – PROJETO 2 - UDACITY - FUNDAMENTOS DE DATA SCIENCE I - Ricardo Aurélio de Albuquerque

- Uma nota especificando qual foi o conjunto de dados usado
  dataset (*titanic-data-6.csv*) - Kaggle.

- Uma definição de qual pergunta você fez
  Quantidade de passageiros embarcados identificados
  Quantidade de passageiros por Origem do embarque
  Quantidade de passageiros por Categoria de classe
  Quantidade de passageiros por idade
  Quantidade de sobreviventes/Não-sobreviventes
  Quantidade de sobreviventes por classe
  Quantidade no geral por sexo e classe
  Quantidade de sobreviventes/Não-sobreviventes por Categoria de classe
  Quantidade no geral de passageiros para adultos e crianças
  Total de crianças/Não-crianças que sobreviveram
  Média de sobrevivência para Criança/Não-criança

- Uma descrição do que você fez para investigar a pergunta
  Como se trata de vários questionamentos, segui a sequência conforme as perguntas eram feitas para que um encadeamento lógico fosse possível de ser realizado. O objetivo final era saber se de fato houve prioridade entre a população de passageiros no que tange ao salvamento e se isso guardava relação com as classes de embarque.
    - Estatística para análise de dados
      A estatística, segundo Mark Twain, é um elemento variável, mas os dados são teimosos. O que significa dizer, de modo mais resumido, é que analisar dados sem base estatística é apenas confiar em dados apresentados. Um modelo baseado em estatística trabalha com uma série de cruzamento de informações e isso é bem explorado na biblioteca Pandas – biblioteca desenvolvida com base na notação Pearson, que é aquela em que se utiliza da correlação linear entre variáveis. Trabalhos baseados em modelos estatísticos reforçam e fundamentam os resultados, o que ocorre nesta exploração apresentada, corolário de que se um universo numérico que representa cerca de 1.300 passageiros, como o caso do *RMS Titanic*, retirou-se uma amostra de 891, como se pode confiar nos resultados apresentados? A resposta, de forma resumida e simples, é que se trabalhou com a frequência, média e padrões para isso representar aquele universo numérico. Dentro da população, inúmeros modelos foram retirados, a exemplo, quantidade de adultos, crianças, sobreviventes, não-sobreviventes, por gênero de pessoas, padrões que se repetem para o resultado estudado.

- Descrição de qualquer limpeza de dados feita
  Foi utilizado o método *dropna* com relação às idades que possuíam valor *NaN* - df_idade.drop('Cri_Adl',axis=1, inplace=True) #Remove linhas para a feature "Age" igual a Nan. Isso na seção para apurar as crianças/Não-crianças que sobreviveram. Em tempo, o *dataset* possui *features* com valores *NaN-Not a numeric* como o caso de "*Age", "Cabin"* e *"Embarked".* Para "*Age"*, como mencionado na seção "limpeza de dados", no interior do código da análise, atribuir a média das idades para essa *feature* causaria um outro resultado que não o apresentado, já que a diferença entre os valores *NaN* e não-NaN representa quase 25%. Para a "*Cabin"*, não é um dado a ser trabalhado mediante o cenário corrente e para *"Embarked"*, dois valores Nan ela possui, o que não tem grande representatividade.

Conclusões – PROJETO 2 - UDACITY - FUNDAMENTOS DE DATA SCIENCE I - Ricardo Aurélio de Albuquerque

- Um resumo das estatísticas e gráficos comunicando seu resultado final

O trabalho ora proposta tem como objetivo expor por meio de notações gráficas e numéricas, informações sobre o *dataset* (*titanic-data-6.csv*) obtido a partir da plataforma *Kaggle.* Esse *dataset* se trata de um conjunto de dados formado por 891 passageiros dos, segundo Wikipedia, 2.435 que estiveram a bordo do navio *RMS Titanic.*

De início, tem-se a informação de que cerca de 38% dos passageiros sobreviveram. A média de idade deles era em torno de 30 anos. Como amostra de uma estatística descritiva, o valor médio pago pela tarifa foi de 30 USD (Dólar americano) e que aproximadamente 75% viajavam desacompanhados.

Seguindo a evolução deste trabalho, a distribuição da população de passageiros era de quase 65% masculina contra 35% feminina.

A título de curiosidade, os passageiros eram oriundos de 3 portos de embarque: Cherbourg-Octeville na França(168), Queenstown na Irlanda(77) e Southampton no Reino Unido(644), cujo efetivo de passageiros foi o maior e de onde o *Titanic* partiu em viagem inaugural. Vale observar que dois passageiros não obtiveram a identificação de qual porto pertenciam.

Os 891 passageiros foram distribuídos por categoria de classe, a saber:

- Primeira classe: 216 passageiros;
- Segunda classe: 184 passageiros; e
- Terceira classe: 491 passageiros.

Quanto à idade deles, tem-se que a maioria era da faixa aproximada de 18 a 35 anos. Diante do desastre ocorrido, 549 passageiros não sobreviveram, assim, cerca de 38% conseguiram sobreviver, 342 passageiros. Essa proporção, dividida por classe, chega a 136 da primeira classe, 87 da segunda e 119 da terceira classe. Também, em termos de quantidade geral por classe, na maioria delas a população masculina predominou. No entanto, pode-se observar que as mulheres foram priorizadas em relação ao salvamento, bem como as crianças, destas, quase 60% se salvaram.

Um fator que guarda certa relação com a classe embarcada é que a primeira e segunda classes obtiveram peso para sobreviverem em detrimento à terceira classe.