Projeto II

Análise Exploratória de Dados Utilizando o RStudio

FATEC Rubens Lara | Professor Jobel Corrêa. MBA, MsC. | 02/03/2023

I.1 Base de dados

Utilizaremos a base de dados "*Titanic - Machine Learning from Disaster*" disponibilizada no serviço Kaggle no endereço https://www.kaggle.com/competitions/titanic. Daqui em diante essa base será chamada de Titanic para facilitar a compreensão.

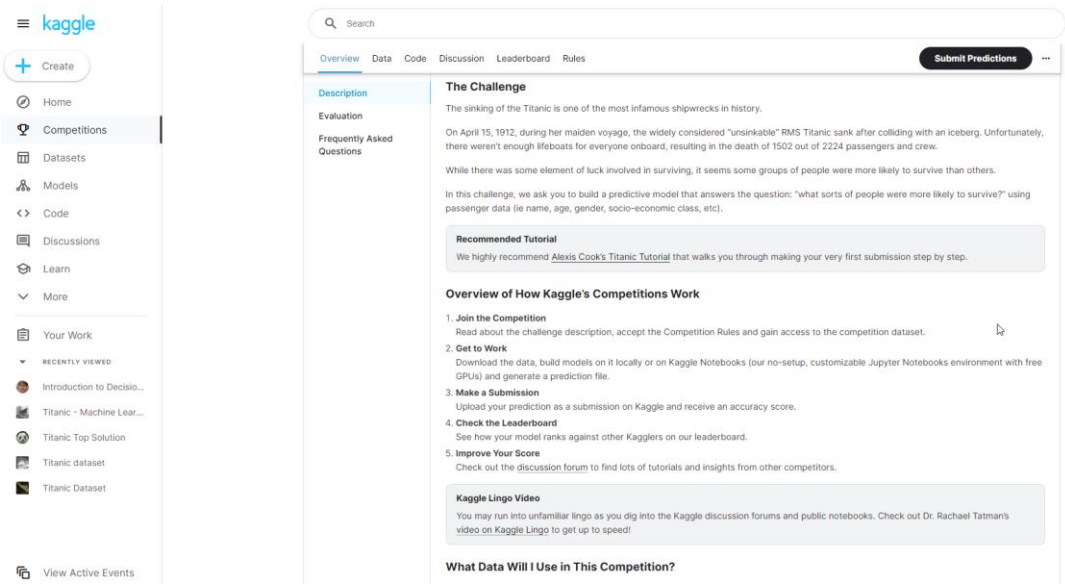

Figura I.1: Base de dados Titanic | Fonte: Kaggle

I.1.1 Descrição da base de dados

O naufrágio do Titanic é um dos naufrágios mais infames da história. Em 15 de abril de 1912, durante sua viagem inaugural, o amplamente considerado "inafundável" RMS Titanic afundou após colidir com um iceberg. Infelizmente, não havia botes salva-vidas suficientes para todos a bordo, resultando na morte de 1.502 dos 2.224 passageiros e tripulantes. Embora houvesse algum elemento de sorte envolvido na sobrevivência, parece que alguns grupos de pessoas eram mais propensos a sobreviver do que outros. Nesta Análise Exploratória de Dados (AED), vamos estudar as variáveis disponíveis e verificar "que pessoas tem maior probabilidade de sobreviver?" usando os dados fornecidos.

I.1.2 Lista de campos existentes na base de dados

A Figura I.2 mostra um resumo da base de dados HDD, com a descrição de cada campo a seguir.

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	725		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	712833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	531	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	805		S
6	0	3	Moran, Mr. James	male		0	0	330877	84583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	518625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	111333		S

Figura I.2: Visão compacta da base de dados Titanic.csv | Fonte: Autor, adaptado da base de dados

1. PassengerId: Número de identificação do passageiro;
2. Survived: Passageiro sobreviveu ao naufrágio? (0 = Não; 1 = Sim);
3. Pclass: Classe onde o passageiro estava no navio (1 = 1ª classe; 2 = 2ª classe; 3 = 3ª classe);
4. Name: Nome do passageiro;
5. Sex: Sexo biológico;
6. Age: Idade;
7. SibSp: Número de irmãos ou cônjuges a bordo;
8. Parch: Número de pais ou filhos a bordo;
9. Ticket: Número do ticket de embarque (passagem);
10. Fare: Valor da tarifa (preço da passagem) em Libras Esterlinas;
11. Cabin: Identificação da cabine;
12. Embarked: Porto de embarque (C = Cherbourg; Q = Queenstown; S = Southampton).

I.1.3 Análise Exploratória de Dados (AED)

Utilizando o ambiente RStudio com auxílio dos e-books em R disponibilizados e da ferramenta ChatGPT você deve realizar:

a) Análise Exploratória de Dados (AED) dos dados da base Titanic. Realize análises de correlação linear de Pearson, verifique a necessidade de nomear variáveis categóricas, determine números absolutos percentuais de sobreviventes ou mortos baseados na classe de embarque, sexo biológico, idade, etc. Monte histogramas. Monte boxplots. Monte gráficos de dispersão. Pesquise sobre esta base na Internet.
b) Apresentação em PowerPoint ou Google Presentation com o storytelling adequado.

I.1.3 Dicas

a) Remova as colunas PassengerID, Name, Ticket e Cabin pois não contribuem para a análise;

b) Remova as linhas que possuem dados faltantes;

c) A variável de interesse ou variável de resultado é Survived.