Data Visualization on the Titanic Dataset

Jawadul Chowdhury

February 10, 2025

1 Abstract

To be worked on.

2 Introduction

The Titanic was a British passenger ship that sank in the Atlantic Ocean on April 15, 1912. The ship had struck an iceberg on its maiden voyage from Southampton, England to New York City.

We use exploratory data analysis on the titanic dataset, using different visualizatoin tehcniques as well as determining which features are should be included in a machine learning model.

3 Dataset Description

When examining the dataset, there are a total of 16 features. Such variables are listed as follows:

- `PassengerId`: The ID of the passenger. This is a discrete numerical data type as the difference between units is constant.
- `Survived`: Whether the passenger survived or not. This is a nominal binary categorical data type as there is no order information. 0 means passenger didn't survive and and 1 means passenger did survive.
- `Pclass`: This is the passenger class. It is a ordinal categorical data type, as 1 means 1st ticket class and so on for 2 and 3, to identify passenger class.
- `Name`: The name of the passenger. This is a nominal categorical data type, as the names of the passenger have no order information.
- `Sex`: This is the sex of the passenger. This is a nominal categorical data type,as genders can't be ordered and is rather binary.
- `Age`: Age of the Passenger. This is a continuous numerical data type, as the difference between units is constant and can be counted.
- `SibSp`: Number of Siblings / Spouses of the passenger. It is a discrete numerical data type as it can be counted and has a constant difference between units.
- `Parch`: Number of Parents / Children abroad the Titanic. It is a discrete numerical data type, as it can be counted and can only take certain values
- `Ticket`: This is the ticket number. This is a nominal categorical data type as there is no ordering information.
- `Fare`: This is the passenger fare. This is a ratio numerical data type as there is a true zero, where zero means that the passenger has not paid any fare.
- `Cabin`: This is the cabin number of the passenger. It is a nominal categorical data type as there is no order information bur rather a quantitvate classifcation.
- `Embarked`: This is the port where the passenger embarked. C is Cherbourg, Q is Queenstown and S is Southampton. This is a categorical nominal data type as there is no order information and has quantitative classification.
- `Age_fill_mean`: This is a copy of the Age column but the blanks have been filled in with the mean. This is a ratio numerical data type, because there are fractional values and has a true zero point.
- `Age_fill_median`: This is a copy of the Age column but the blanks have been filled in with the median. This is a discrete numerical data type because the age is defined to be continuous numerical.
- `Age_fill_mode`: This is a copy of the Age column but the blanks have been filled in with the mode. This is a discrete numerical data type, because the age is defined to be continuous numerical.
- `Age_fill_knn`: This is a copy of the Age column but the blanks have been filled in with the mean. This is a ratio numerical data type, because there are fractional values and has a true zero point.