The following is an explanation on the methodology used in my analysis of the Titanic dataset.

1. The csv is loaded into a Pandas data-frame. A screen shot of a computer

   Description automatically generated

The count refers to the number of non null values in the dataset, from this we can deduce that all features except from ‘Age’ and ‘Cabin’ have no null values. Removal of the null values in age will be necessary for analysis. Due to the high null rate in the ‘Cabin’ feature it will be excluded from analyis.

The min and max values will also be usefull later to normalize the data for the logistic model.

1. Using a Phik matrix shows that the most important features for predicting survival are Sex, Passenger class, Fare, and Embarked so these will be used to fit the model

|  |  |
| --- | --- |
| Feature | Corelation coefficient |
| PassengerId 0.156077 | |
| Age 0.204231 | |
| SibSp 0.218441 | |
| Parch 0.167872 | |
| Fare 0.274939 | |
| Sex\_female 0.742633 | |
| Sex\_male 0.742633 | |
| Pclass\_1 0.443398 | |
| Pclass\_2 0.117118 | |
| Pclass\_3 0.496300 | |
| Embarked\_C 0.291650 | |
| Embarked\_Q 0.028397 | |
| Embarked\_S 0.235120 | |

1. The histograms of ages of people who survived with the differences from the initial distribution of ages multiplied out show the probability of surviving to be non linear. Because of this I will onehot age ranges.A graph of blue bars

   Description automatically generated