# Titanic Survial Prediction

Pratikshya Parajuli

Ministry of Finance

Governmnet of Nepal

August 13, 2022

# Overview



## Introduction

## Exploratory Data Analysis(EDA)

## Feature Engineering and Data Cleaning

## Predictive Modeling

## Conclusion



# Introduction

# Overview



The sinking of the Titanic is one of the most infamous shipwrecks in history. This project aims to create a model that predicts which passengers survived the disaster.

- Useful features are Pclass, Name, Sex, Age, SibSp, Parch, Ticket, Fare, Cabin, Embarked
- Target feature is Survived



# Exploratory Data Analysis(EDA)

# Dataset



| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22.0 | 1 | 0 | A/5 21171 | 7.2500 | NaN | S |
| **1** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Th... | female | 38.0 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **2** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26.0 | 0 | 0 | STON/O2. 3101282 | 7.9250 | NaN | S |
| **3** | 4 | 1 | 1 | Futrelle, Mrs. Jacques Heath (Lily May Peel) | female | 35.0 | 1 | 0 | 113803 | 53.1000 | C123 | S |
| **4** | 5 | 0 | 3 | Allen, Mr. William Henry | male | 35.0 | 0 | 0 | 373450 | 8.0500 | NaN | S |

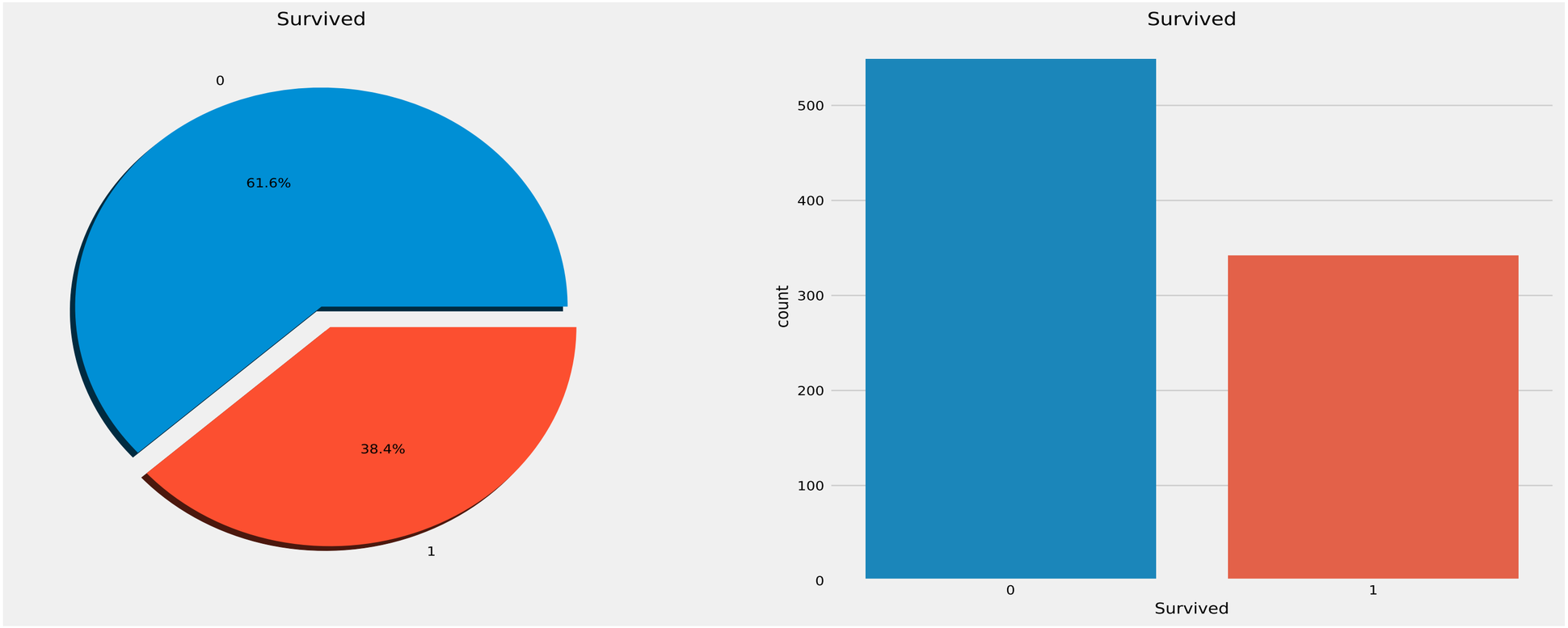
```
data.isnull().sum() #checking for total null values
```

```
PassengerId      0
Survived         0
Pclass           0
Name             0
Sex              0
Age            177
SibSp            0
Parch            0
Ticket           0
Fare             0
Cabin          687
Embarked         2
dtype: int64
```

- Age, Sex, Embarked have null values.

# How many Survived?



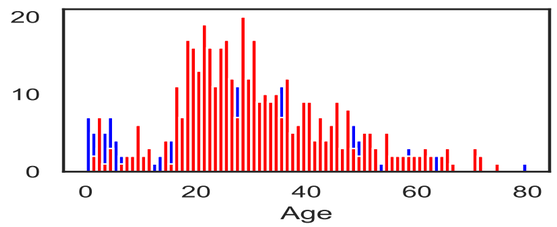


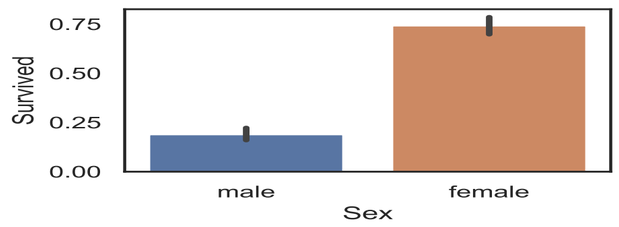

- We will try to check the survival rate by using the different features of the dataset. Some of the features being Sex, Port Of Embarcation, Age,etc.

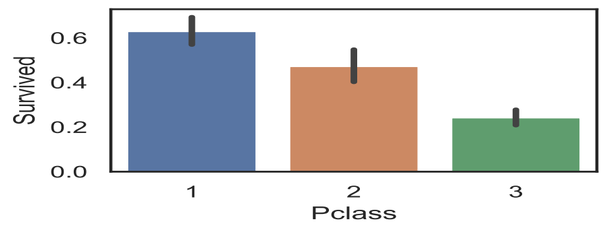
# Analysis of the Features

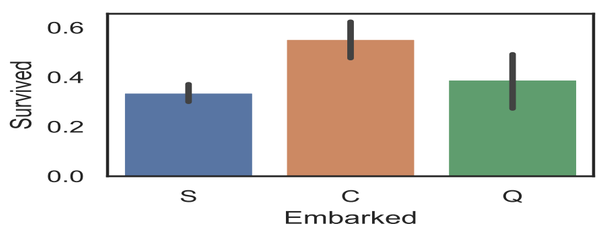


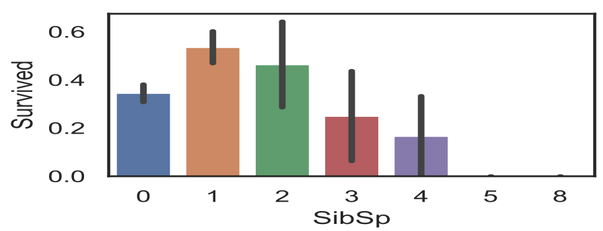
- Categorical Features in the dataset - Sex, Embarked
- Ordinal Features in the dataset - Pclass
- Continuous Features in the dataset - Age

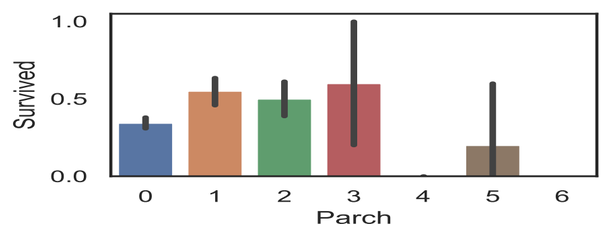


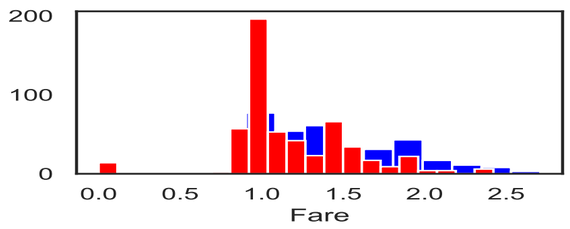



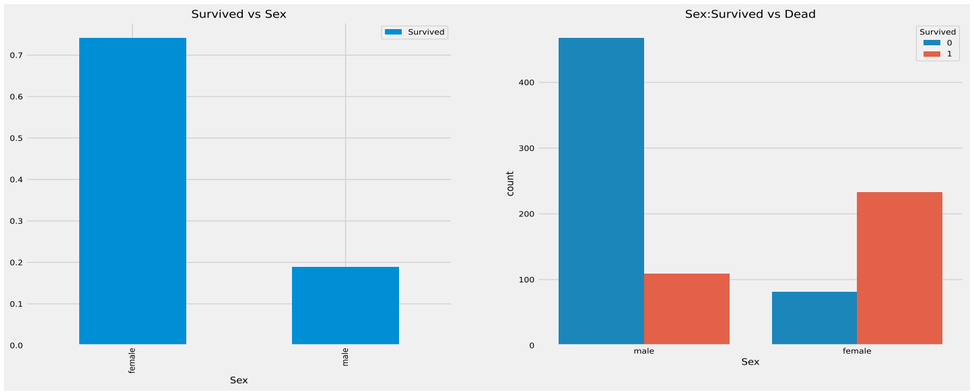



# Sex - Categorical Feature



Table 1: Survived vs. Sex

| Sex | Survived | Numbers |
|---|---|---|
| Female | 0 | 81 |
| | 1 | 233 |
| Male | 0 | 468 |
| | 1 | 109 |




- Survival rates for a women: 75 percent and men: 18-19 percent.

# Pclass - Ordianal Feature



Table 2: Numbers of Passengers by Pclass

| Survived | 0 | 1 | All |
|---|---|---|---|
| Pclass | | | |
| 1 | 80 | 136 | 216 |
| 2 | 97 | 87 | 184 |
| 3 | 372 | 119 | 491 |
| All | 549 | 342 | 891 |

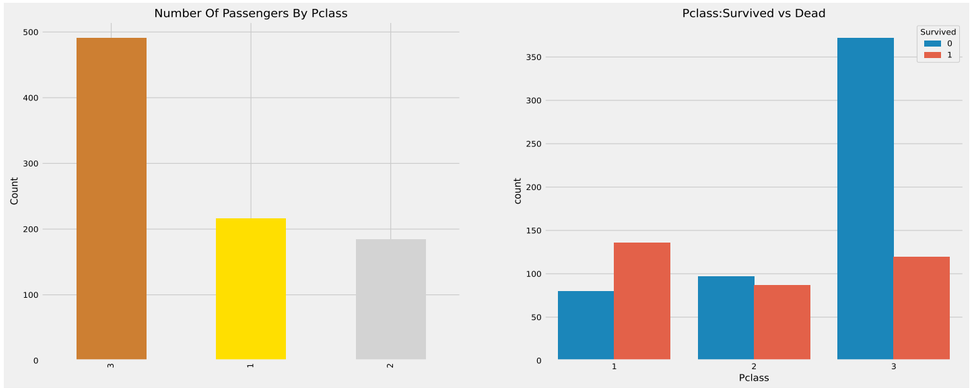



Figure 1: Pclass:Survived vs Dead

# Survival rate with Sex and Pclass Together



Table 3: Survival rate with Sex and Pclass Together

| Sex | Pclass | Survived | 2 | 3 | All |
|---|---|---|---|---|---|
| Female | 0 | 3 | 6 | 72 | 81 |
| | 1 | 91 | 70 | 72 | 233 |
| Male | 0 | 77 | 91 | 300 | 468 |
| | 1 | 45 | 17 | 47 | 109 |
| All | | 216 | 184 | 491 | 891 |

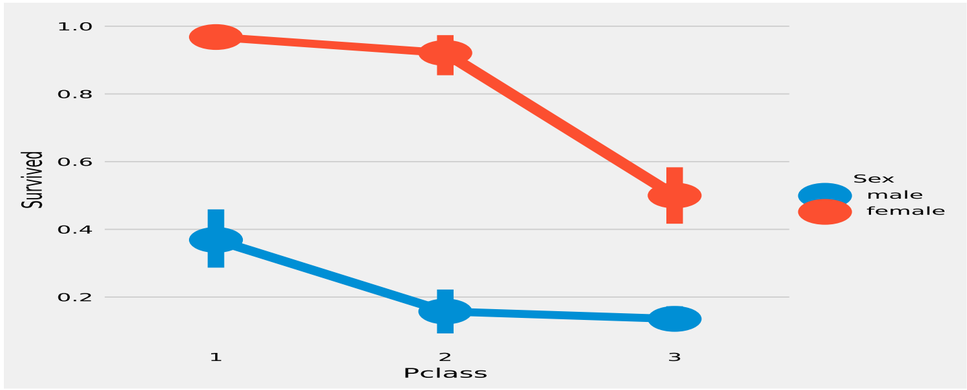



Figure 2: Survival rate with Sex and Pclass Together

# Age - Continuous Feature



Oldest Passenger was of: 80.0 Years

Youngest Passenger was of: 0.42 Years

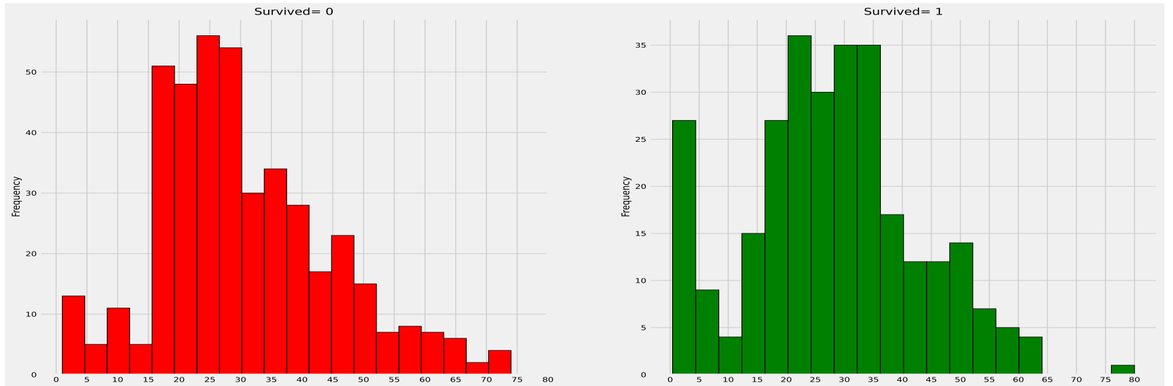



Figure 3: Survival rate with Age

Observations:

1)The Toddlers(age<5) were saved in large numbers.

2)The oldest Passenger was saved(80 years).

3)Maximum number of deaths were in the age group of 30-40.

# Embarked - Categorical Value



1)Maximum passenegers boarded from S. Majority of them being from Pclass3.

2)The Passengers from C survived.

3)The Embark S looks to the port from where majority of the rich people boarded. Still the chances for survival is low here.

4)Port Q had almost 95 percent of the passengers were from Pclass3.

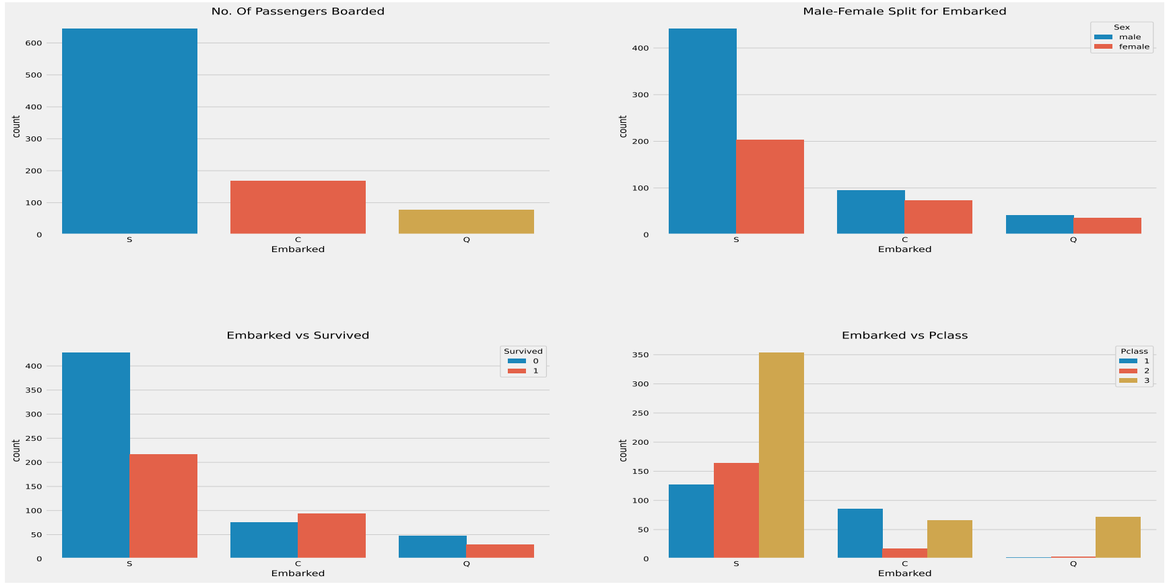



Figure 4: Survival rate with Port of Embarkation

# Correlatoin Matrix



The highest correlation is between SibSp and Parch i.e 0.41.

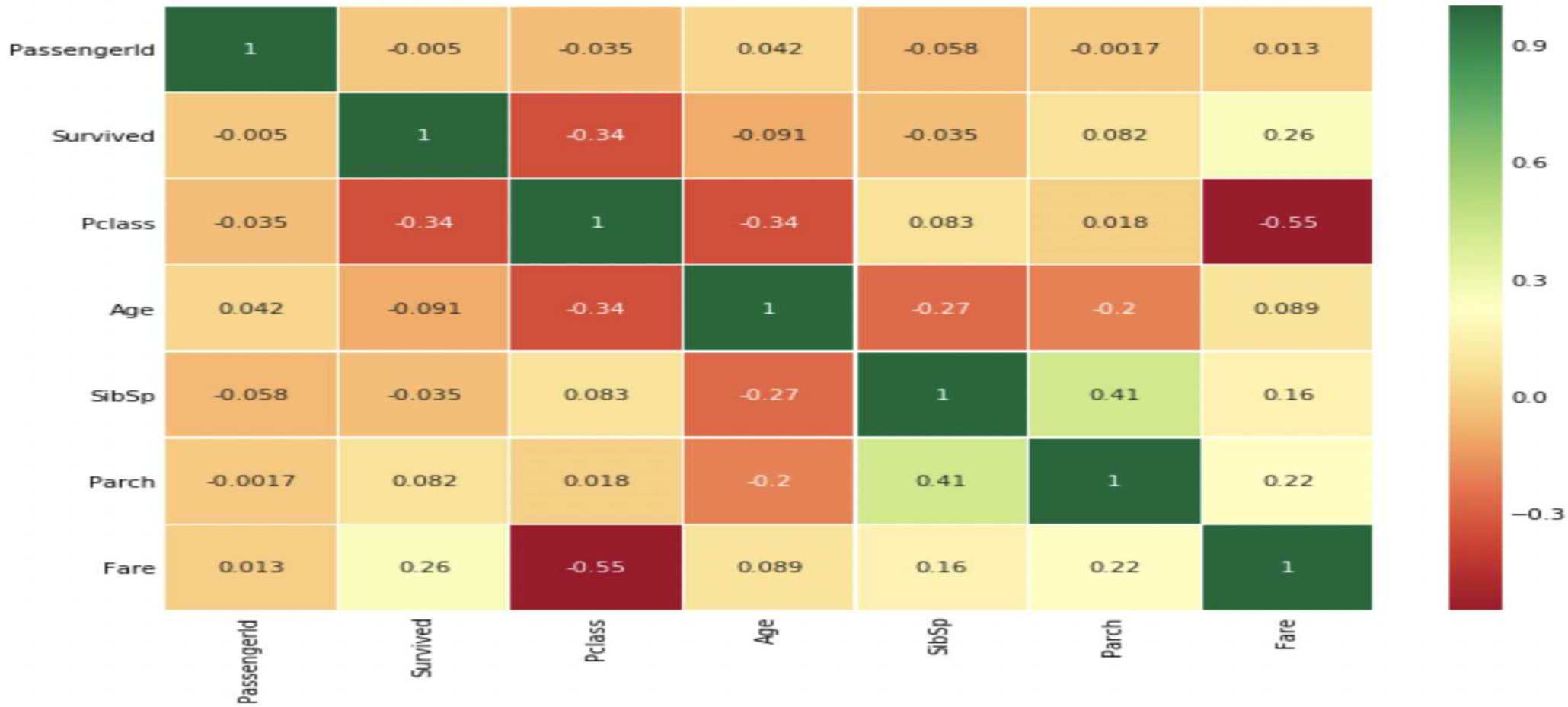



Figure 5: Interpreting the heatmap



# Feature Engineering and Data Cleaning

# Converting features into suitable form for modeling



- Age: Age_band
- Family_size and Alone: Summation of Parch and SibSp
- Fare: Fare_cat

Table 4: Age_Band

| Age_band | Numbers |
|---|---|
| 1 | 382 |
| 2 | 325 |
| 0 | 104 |
| 3 | 69 |
| 4 | 11 |

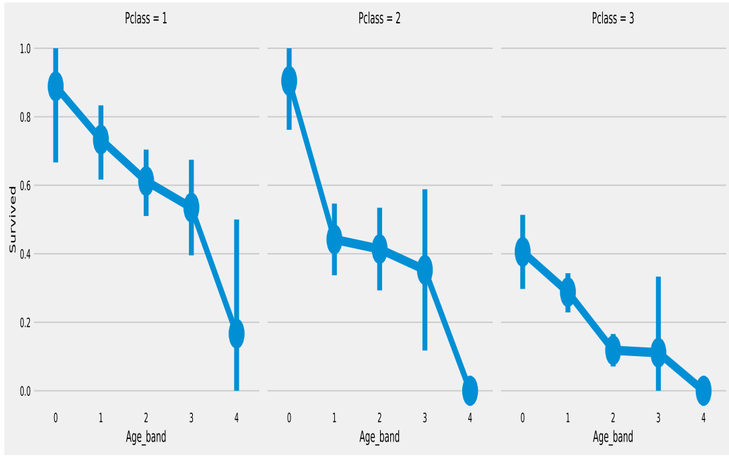



Figure 6: Age_Band

# Removing Redundant features



- Name–> We don't need name feature as it cannot be converted into any categorical value.
- Ticket–> It is any random string that cannot be categorised.
- Fare–> We have the Fare_cat feature, so unneeded
- Cabin–> A lot of NaN values and also many passengers have multiple cabins. So this is a useless feature.
- Fare_Range–> We have the fare_cat feature.
- PassengerId–> Cannot be categorised.

# Correlation Matrix after Data Cleaning



Positive correlation: SibSp andd Family_Size and Parch and Family_Size and Negative correlation: Alone and Family_Size

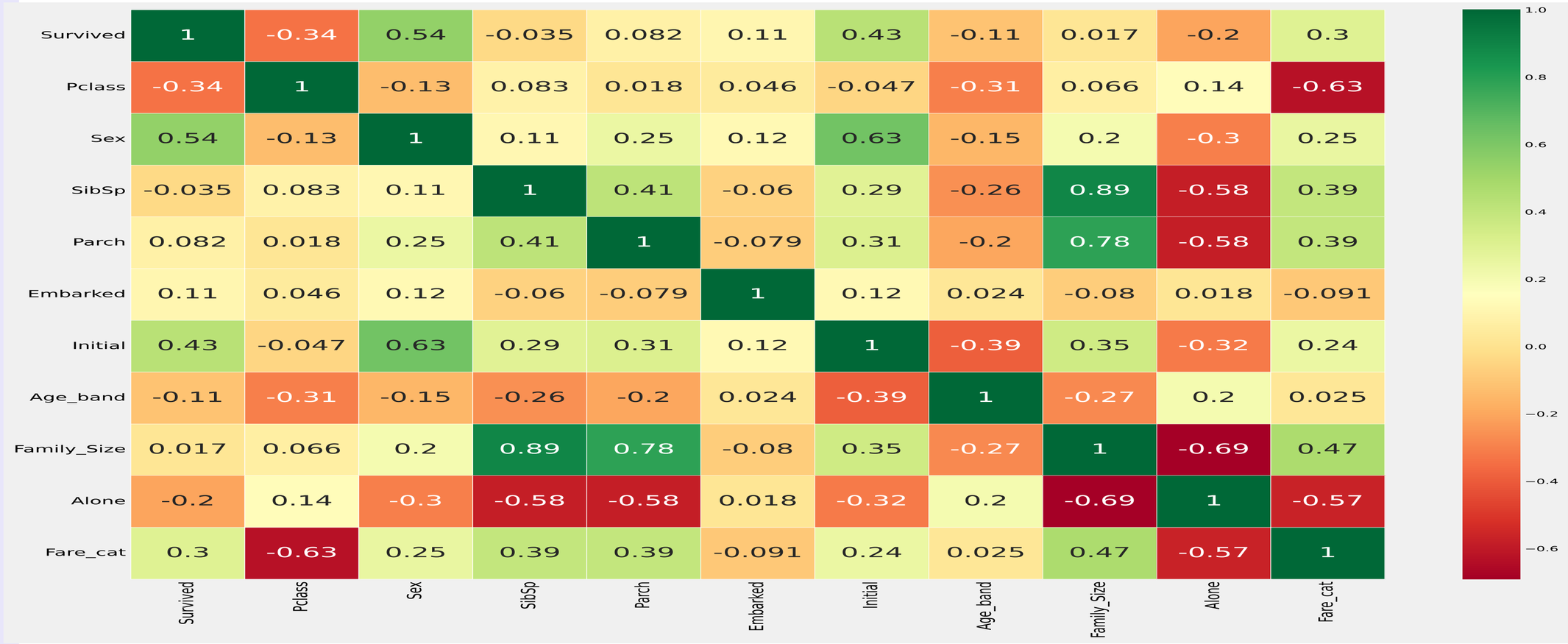



Figure 7: Correlation Matrix after Data Cleaning



# Predictive Modeling

# Evaluation Classification Algorithms



- Logistic Regression
- Support Vector Machines (Linear and radial)
- Random Forest
- K-Nearest Neighbours
- Naive Bayes
- Decision Tree

# Prediction Accuracy



- Split the train sample into train and test dataset
- Train Data_size : 0.7 and Test Data_size : 0.3
- Total sample size = 891; training sample size = 623, testing sample size = 268

Table 5: Accuracy Comparison of different Classifier Algorithms

| | Acuracy |
|---|---|
| Radial Support Vector Machines(rbf-SVM) | 0.835820895522388 |
| Linear Support Vector Machine(linear-SVM) | 0.817164179104476 |
| Logistic Regression | 0.8134328358208955 |
| Decision Tree | 0.805970149253731 3 |
| K-Nearest Neighbours(KNN) | 0.832089552238806 |
| Gaussian Naive Bayes | 0.8134328358208955 |
| Random Forests | 0.820895522388059 7 |



# Conclusion

# Conclusion



- Basic modeling of the data
- To overcome the model variance, and get a generalized model,we can use Cross Validation
- Results can be further enhanced

# Contact Information


Pratikshya Parajuli

Ministry of Finance

Government of Nepal

PPARAJULI@MOF.GOV.NP