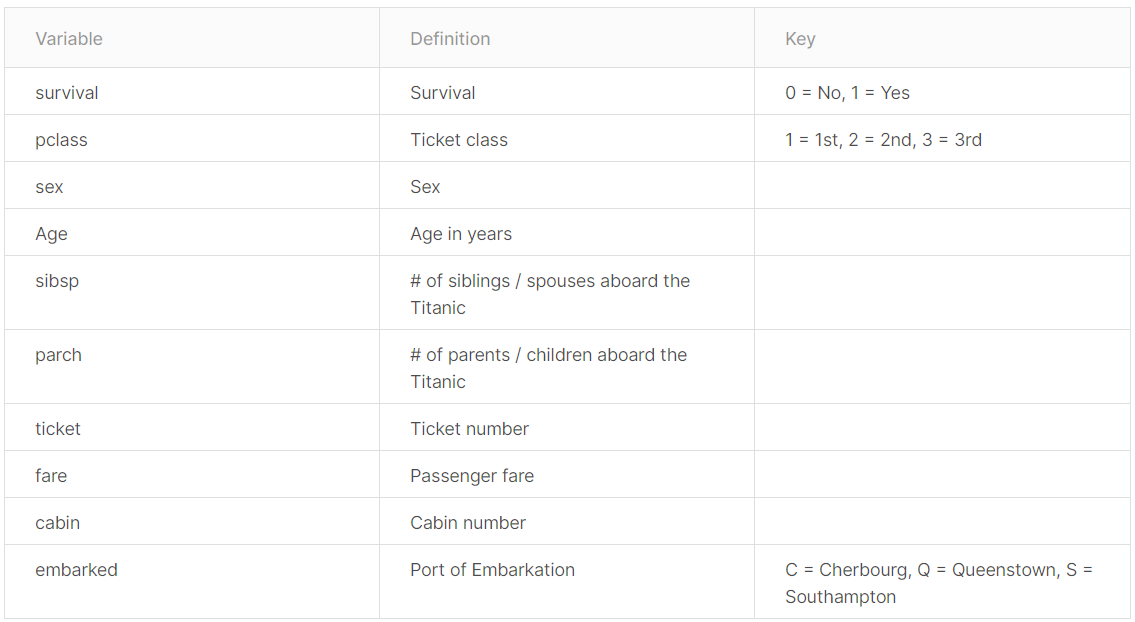
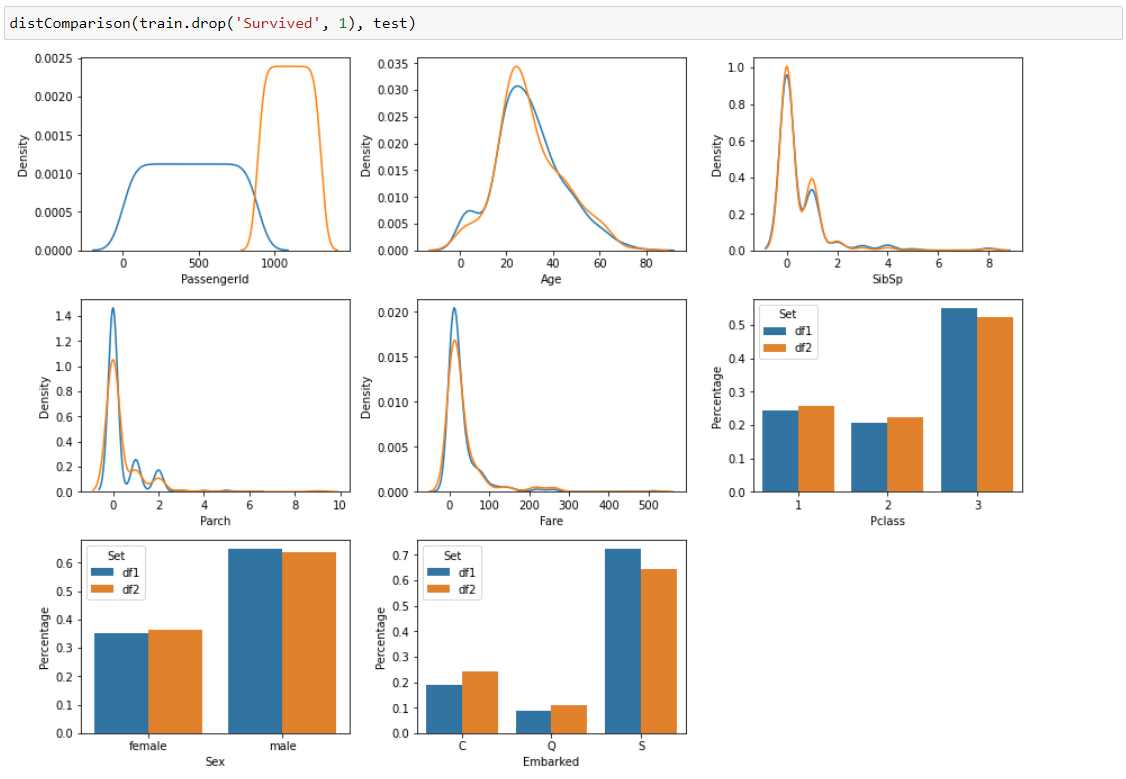
# Dataset

The dataset used for this project is a toy dataset that contains information about the titanic passengers. The dataset came pre-divided into the training and testing datasets. There are a total of 418 instances in the test dataset and 891 instances in the training set. The features in both the datasets are as follows:

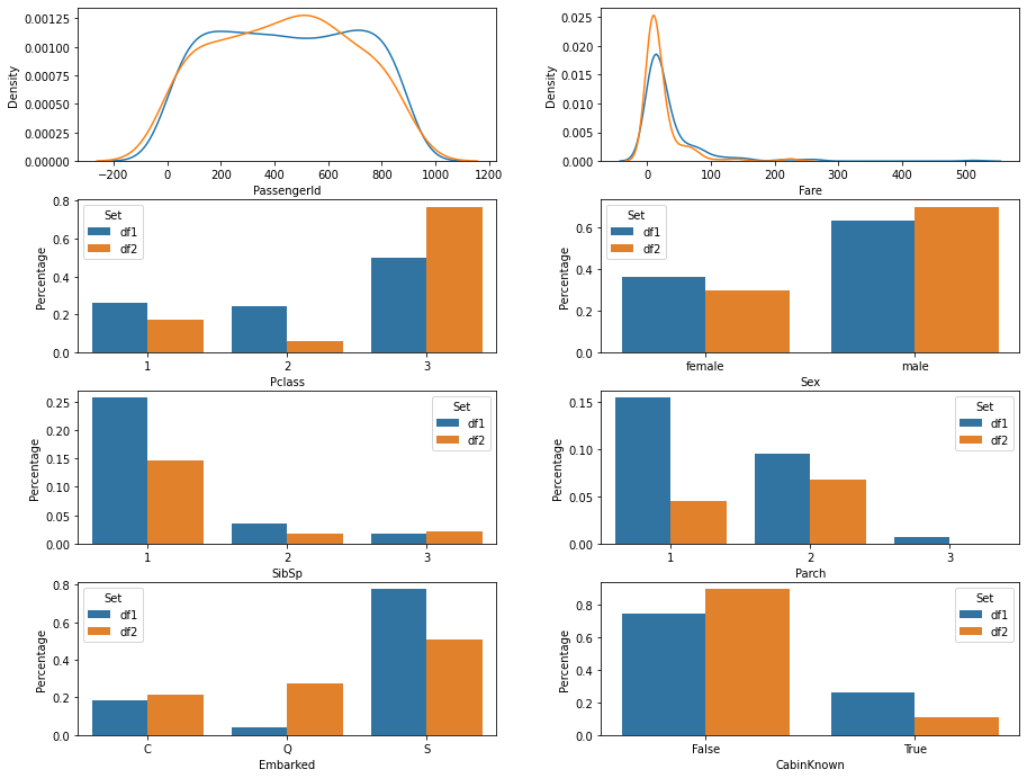


# Machine Learning

The aim of this project code is to evaluate the performance of various data imputation techniques and compare them. There are various graphs that pertain to data preprocessing and algorithm performance. I will be explaining them in this section. The algorithm used in this project is the Random Forest classifier.

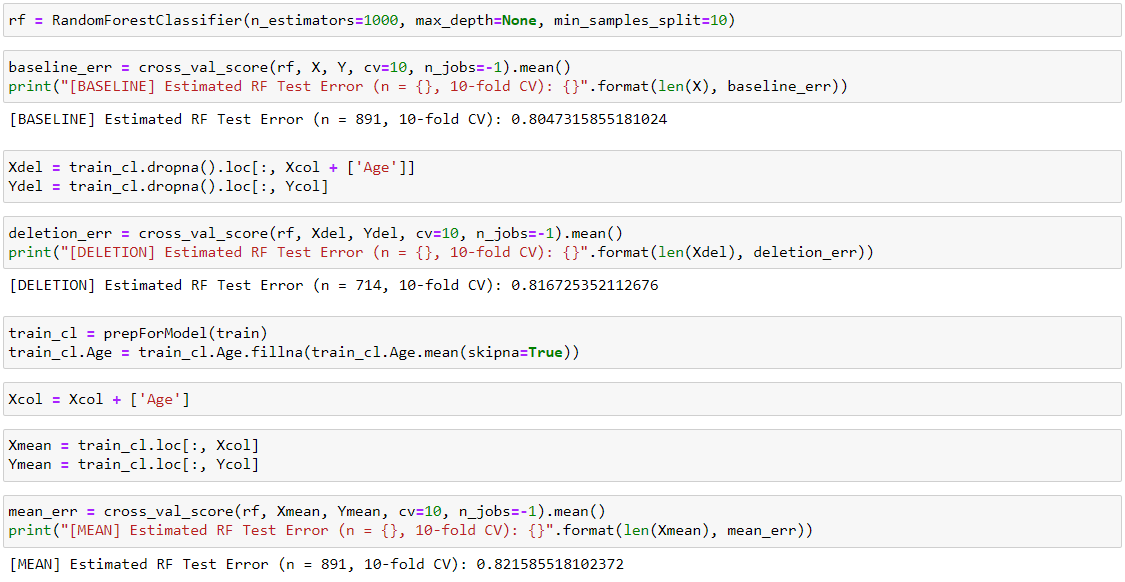


The above collection of graphs shows the comparison of the distribution of each feature in both the train and test sets. The blue represents the train set and the orange represents the test set.



The above graph also shows the comparison of the distribution of each feature in the train dataset (blue) and the test dataset (orange). But these collections of graphs have been plotted after removing the null values from the dataset and grouping the data of different classes in each of the features in both the datasets.

The algorithm used for Machine learning is the Random Forest Classifier. The three different methods used for data imputation are as shown in the code snippet below:



The results for each of the techniques are as shown in the graph below:

