Akshay Gade
G01025094

AIT 580 – Project Report

Milestone 1 :

The Data Acquisition and Conversion has been done with python using pandas package. The data has been downloaded programmatically and converted into a dataframe from a JSON file. The output file looks like:

	CUSTOMERID	DESCRIPTION	FARE	GUESTS	SEATCLASS	SUCCESS
1	1	Braund, Mr. Owen Harris;22	7.25	1	3	0
2	2	Cumings, Mrs. John Bradley (Florence Briggs Th...	71.2833	1	1	1
3	3	Heikkinen, Miss. Laina;26	7.925	0	3	1
4	4	Futrelle, Mrs. Jacques Heath (Lily May Peel);35	53.1	1	1	1
5	5	Allen, Mr. William Henry;35	8.05	0	3	0

Milestone 2 :

The DESCRIPTION in the dataset has been divided into five different columns (First Name, Last Name, Salutation, Age, Alternate Name) using split, extract and replace functions. The additional metadata that has been added are has_alternate_name and Ethnicity based on the first and last names. All of the columns have been converted to either numerical or nominal data. Each ethnicity is set with a number to convert everything into nominal data (like SEATCLASS).

Null values in the Age column has been replaced based on the mean data of each salutation. For example, if a null value has a salutation of Master, the null value is replaced by the mean value of Master salutation. The output file after milestone 2 looks like:

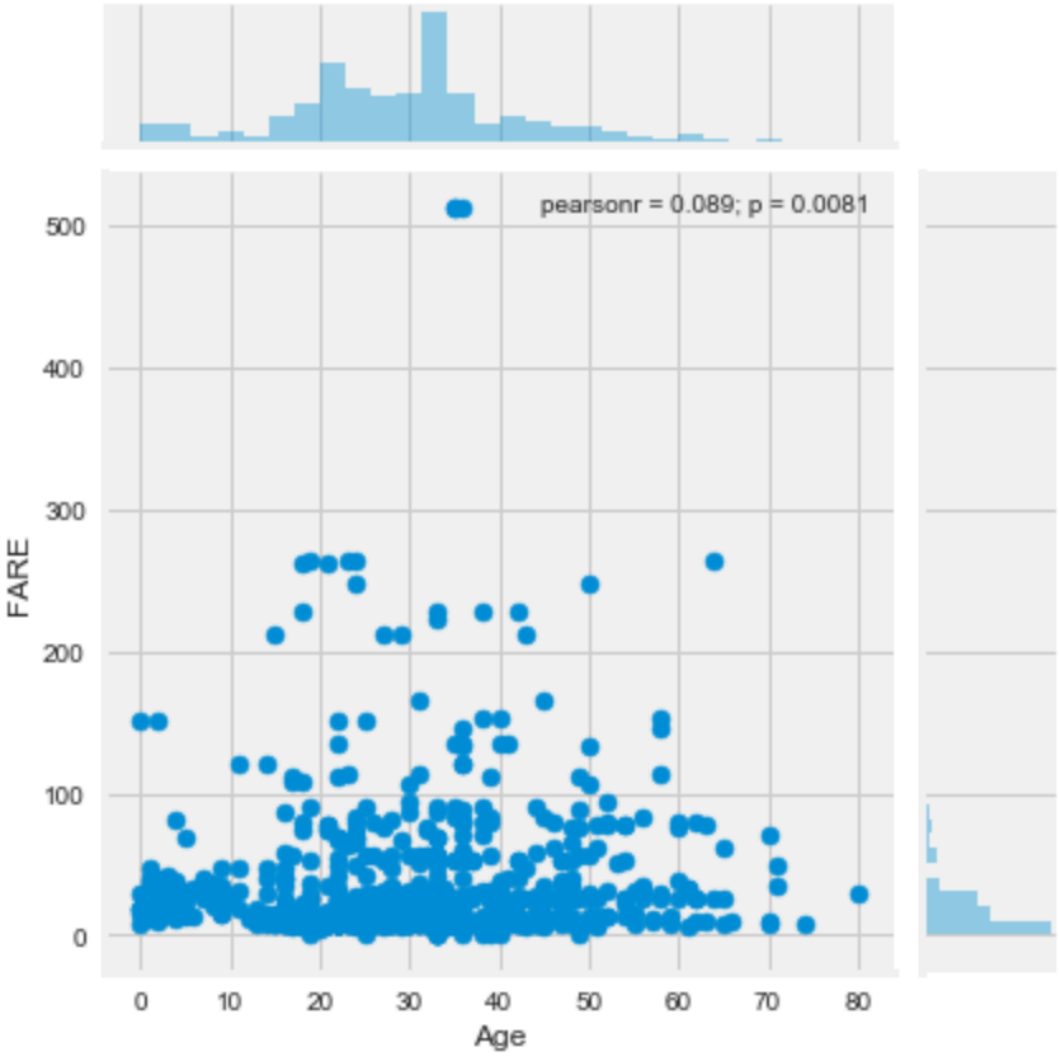
	CUSTOMERID	FARE	GUESTS	SEATCLASS	SUCCESS	Last_Name	Title	First_Name	Age	Alternate_Name	ethnicity	ethni
1	1	7.2500	1	3	0	Braund	Mr	Owen Harris	22	nan	english	5.0
2	2	71.2833	1	1	1	Cumings	Mrs	John Bradley	38	Florence Briggs Thayer	nordic	17.0
3	3	7.9250	0	3	1	Heikkinen	Miss	Laina	26	nan	french	6.0
4	4	53.1000	1	1	1	Futrelle	Mrs	Jacques Heath	35	Lily May Peel	english	5.0
5	5	8.0500	0	3	0	Allen	Mr	William Henry	35	nan	english	5.0

Milestone 3:

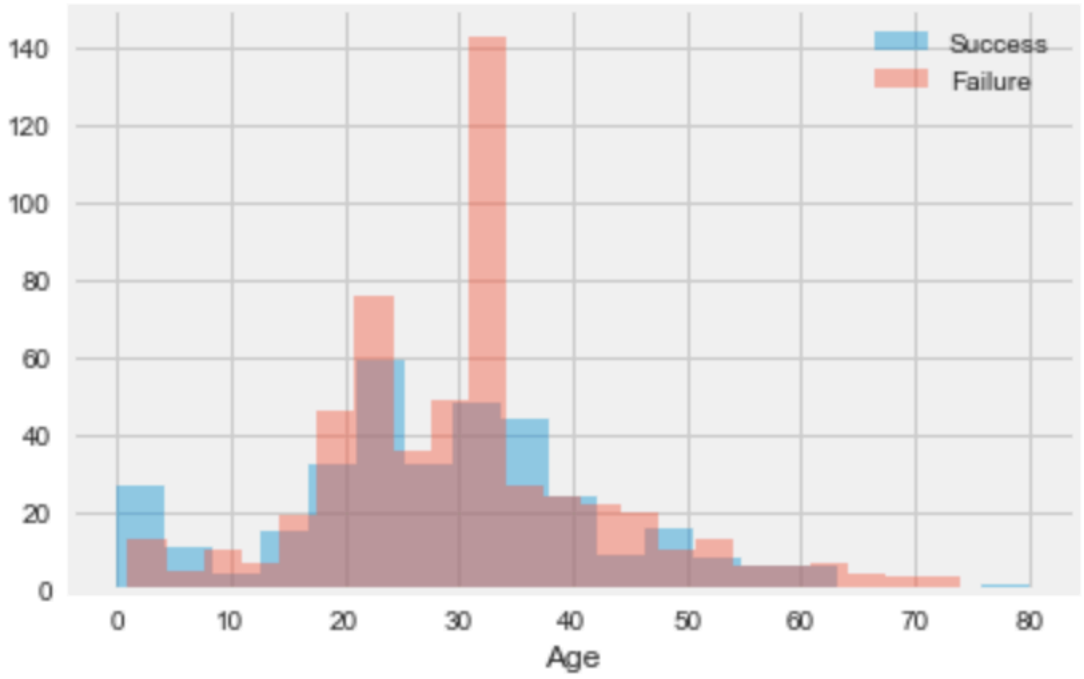
Metadata exploration has been done using seaborn package in python. The visuals are given below:

Akshay Gade
G01025094

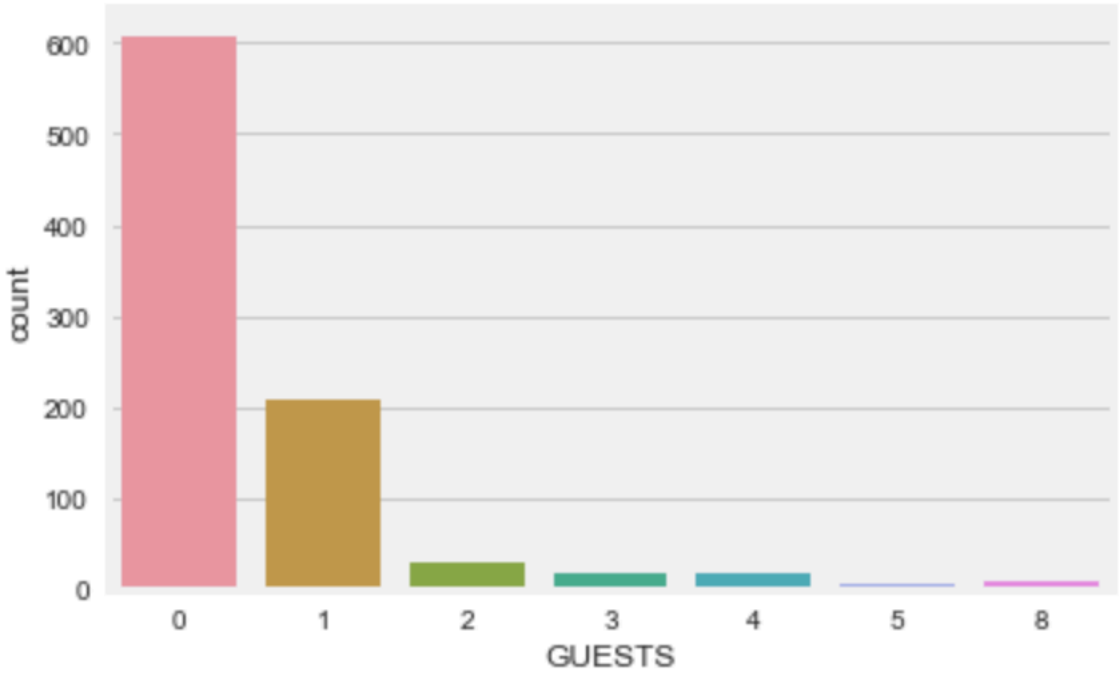
Distribution of Fare with respect to Age



Akshay Gade
G01025094

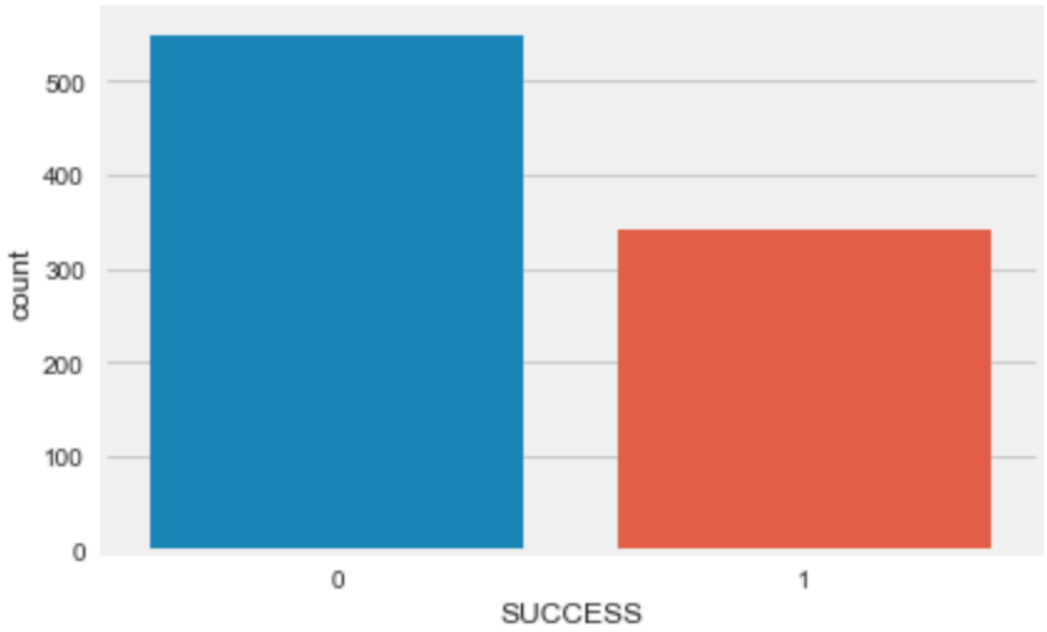


Distribution of Fare with respect to Age, Grouped by SUCCESS
It indicates that most of the un successful customers fall in the age of 30 - 40



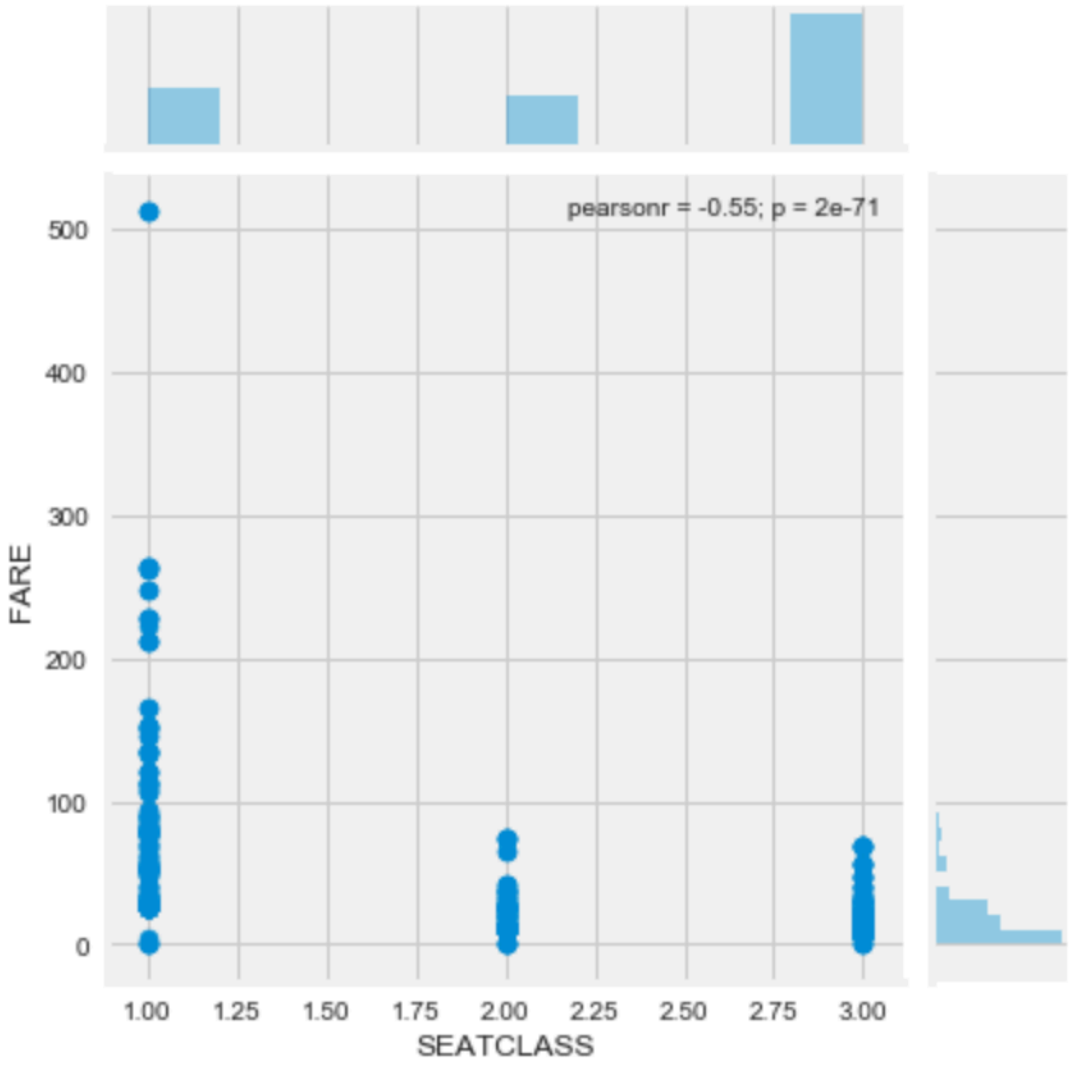
Countplot of customers with respect to number of GUESTS
This indicates that most of the customers don't have any guests.

Akshay Gade
G01025094



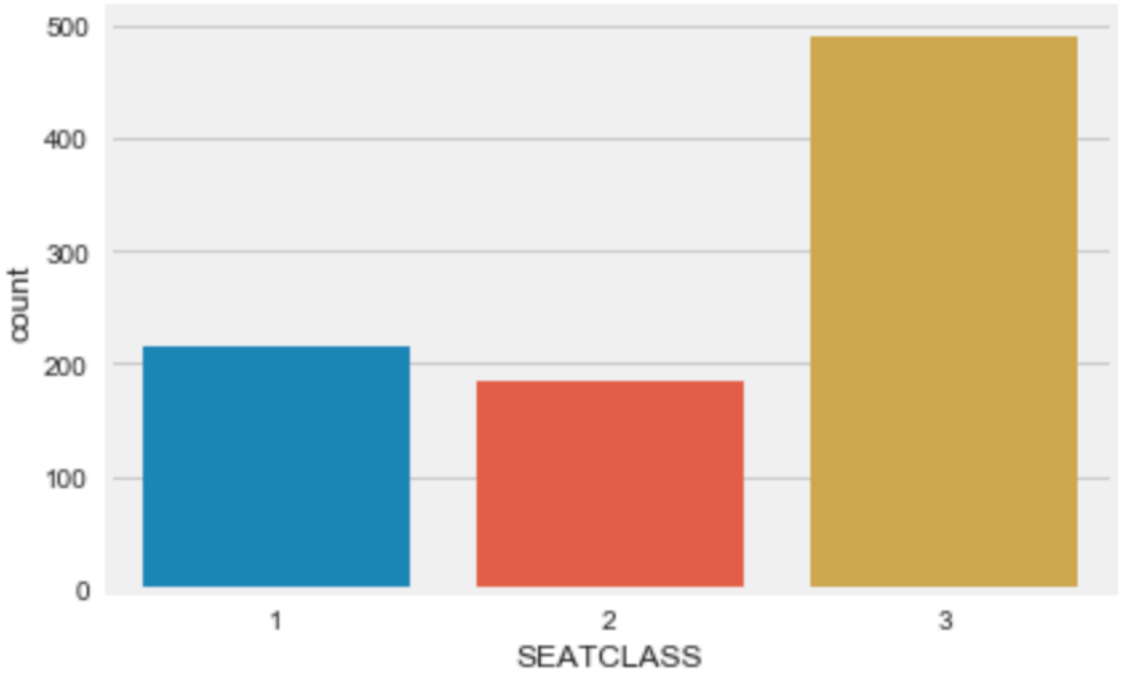
Countplot of Customers with respect to SUCCESS
The number of customers with failure is greater than success.

Akshay Gade
G01025094

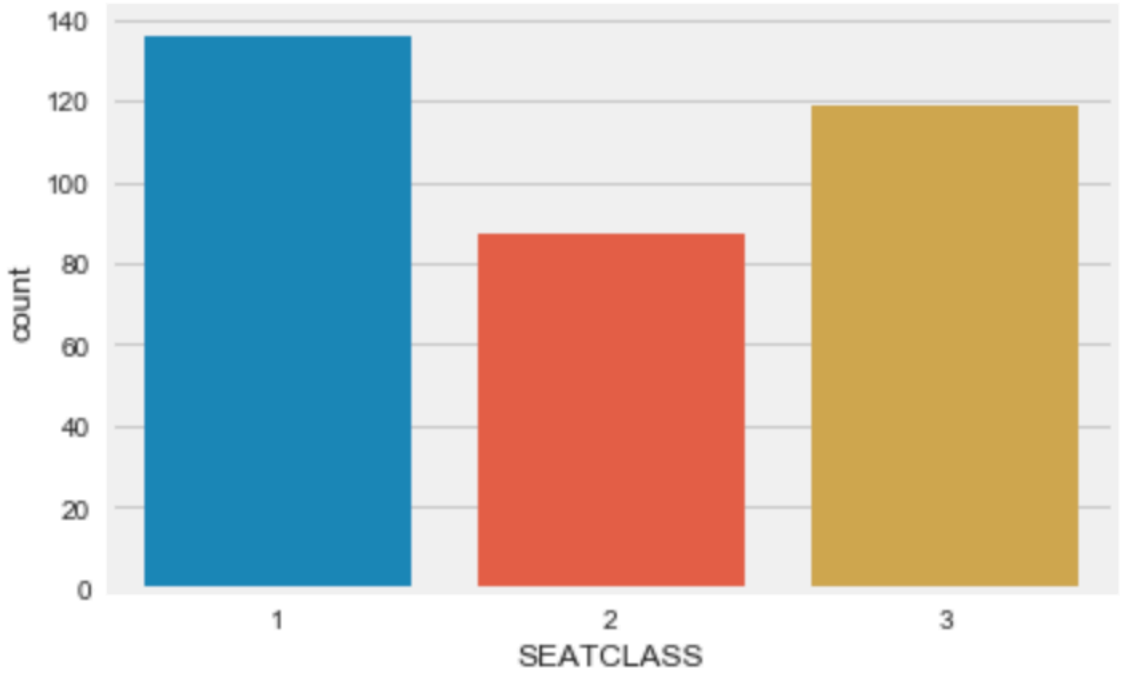


Jointplot of customers with respect to FARE, grouped by SEATCLASS
This indicates that seatclass 1 is the costliest among all.

Akshay Gade
G01025094

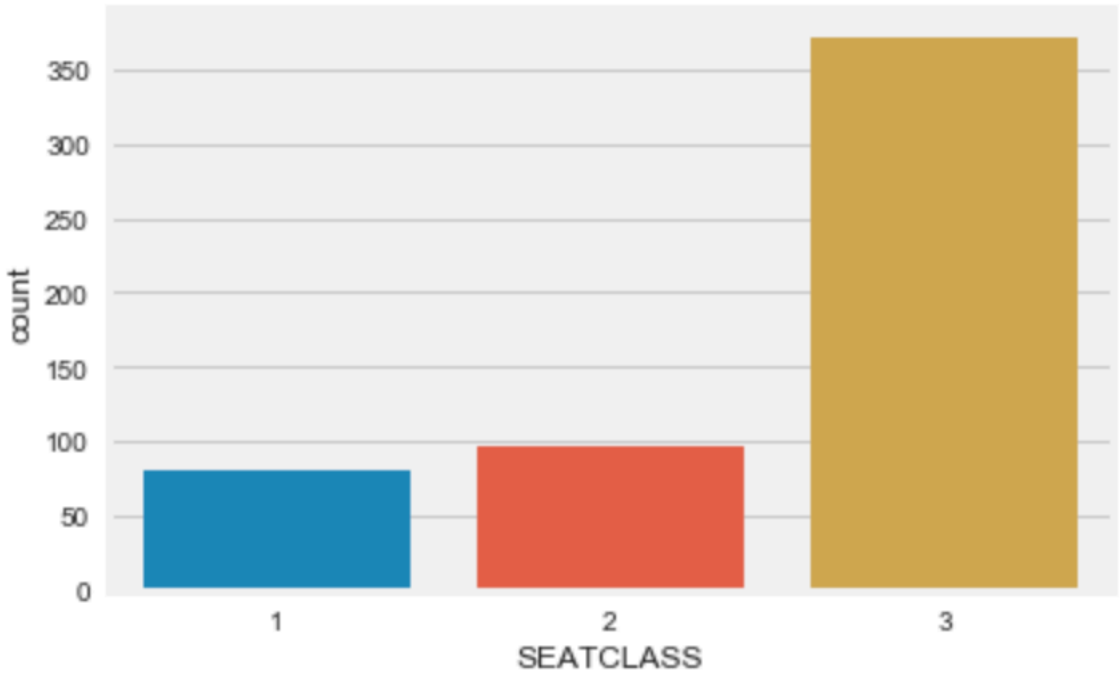


Number of customers grouped by seatclass

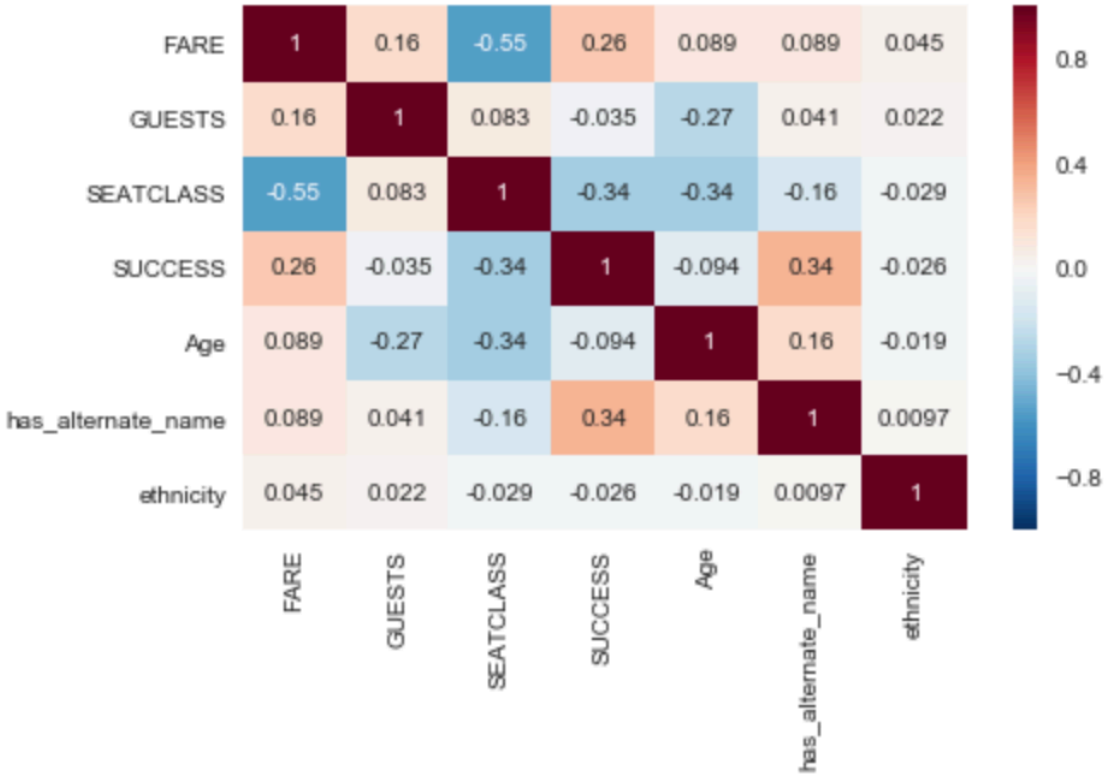


Seatclass grouped by number of customers with success.
Seatclass 1 has highest number of success rate.

Akshay Gade
G01025094



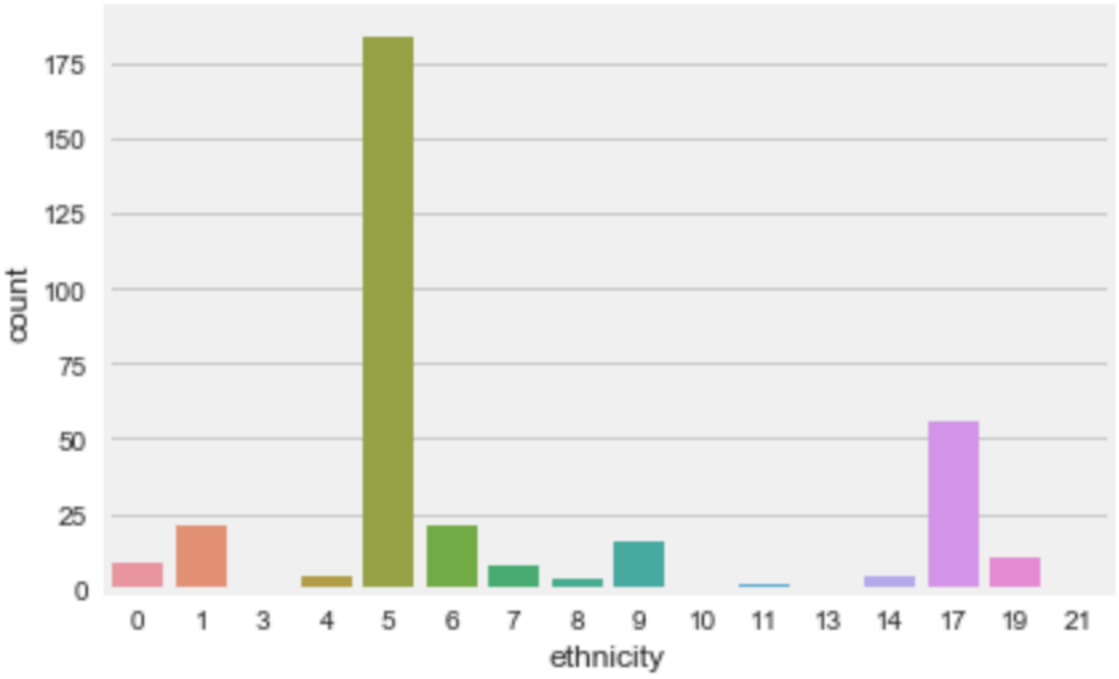
Seatclass grouped by number of customers with out success.
Seatclass 3 has highest number of failures.



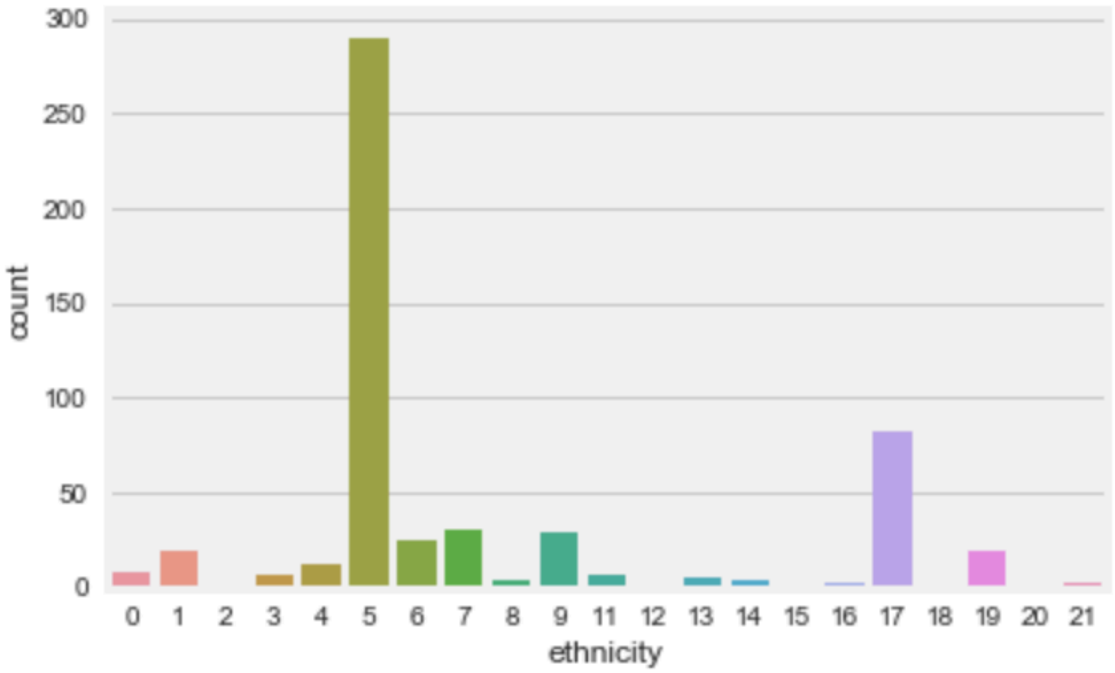
Correlation plot for all the variables

Akshay Gade
G01025094

This indicates that success is positively correlated with Fare and Seatclass.

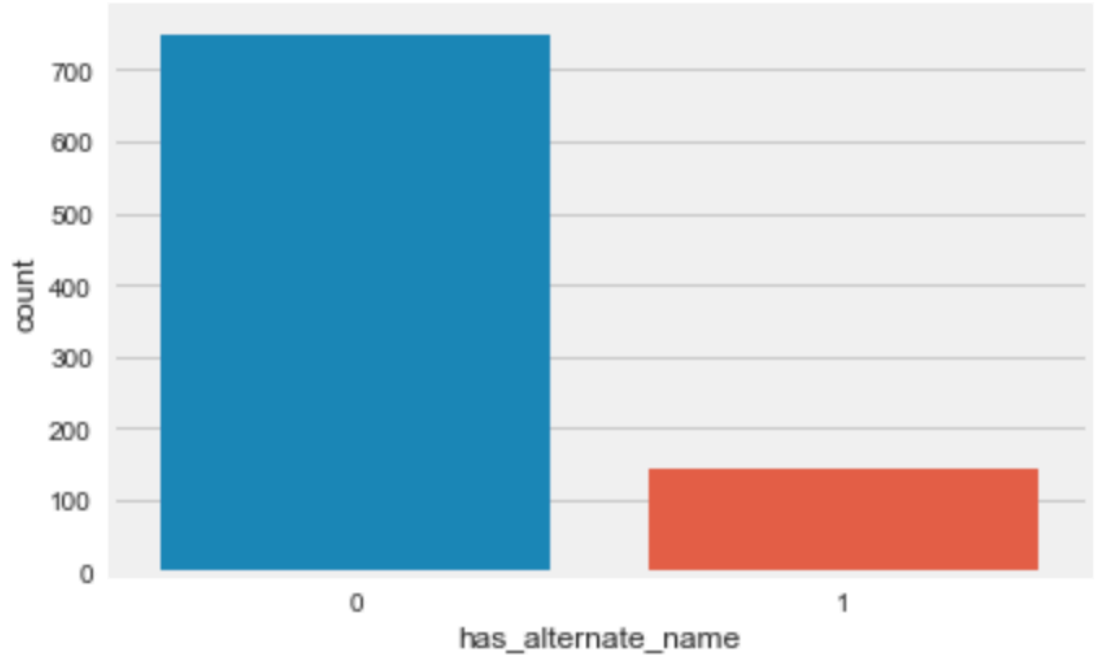


Ethnicity countplot with success (English is the highest)



Ethnicity countplot without success (English is the highest)

Akshay Gade
G01025094



Number of customers with alternate name (0 for No, 1 for Yes)

Milestone 4:

The data is loaded into Weka and the numerical data is converted into nominal by using Discretize filter in Weka. Using the attribute selector, the best attributes from the data that we have obtained are **Fare, Seat class and has_alternate_name**.

```
=== Attribute selection 10 fold cross-validation (stratified), seed: 1

number of folds (%)  attribute
          10(100 %)   1 FARE
           0(  0 %)   2 GUESTS
          10(100 %)   3 SEATCLASS
           0(  0 %)   5 Age
          10(100 %)   6 has_alternate_name
           0(  0 %)   7 ethnicity
```

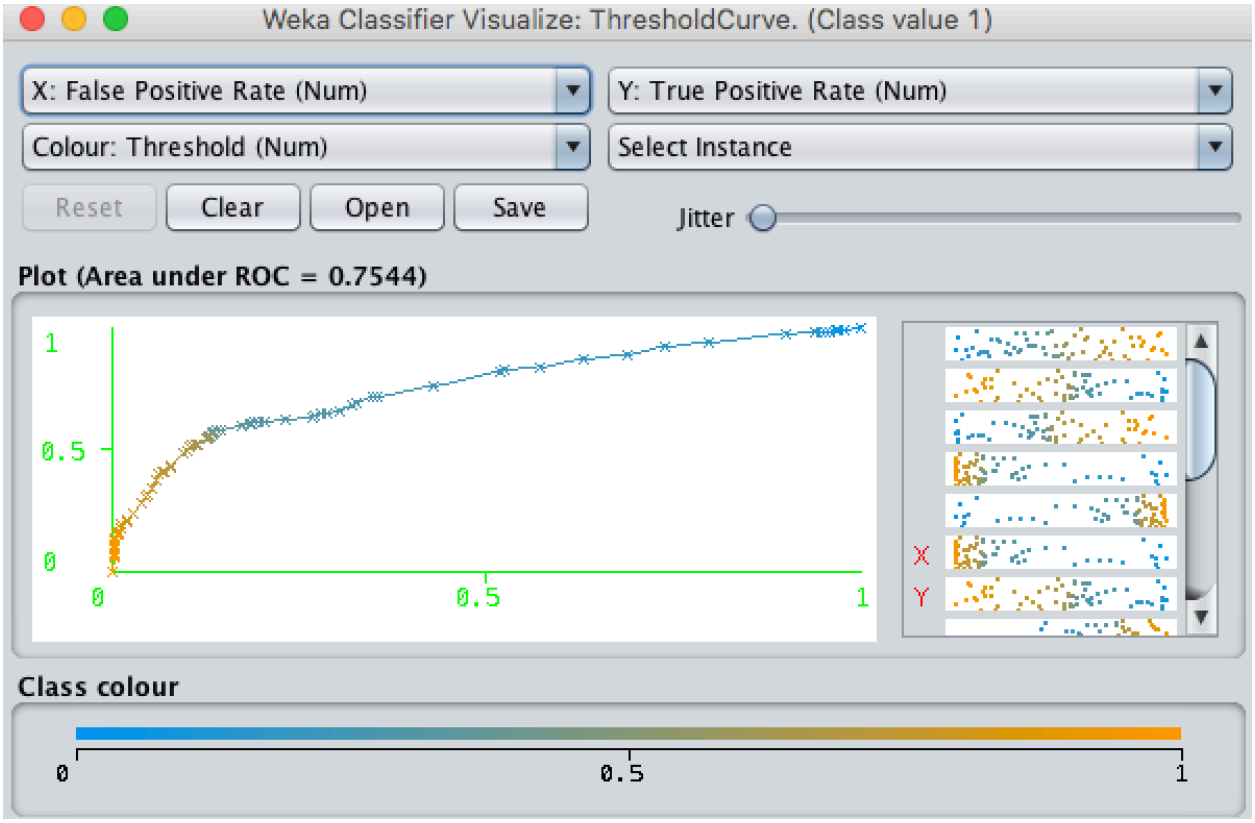
Milestone 5:

Akshay Gade
G01025094

J48 decision tree classifier is being run on the dataset, considering success as the output variable. The accuracy of the classifier is 75.44%. For Random forest, the accuracy is 77.2% The ROC curve and the precision recall curve is shown below. The closer the curve is to 1, the better the model is.

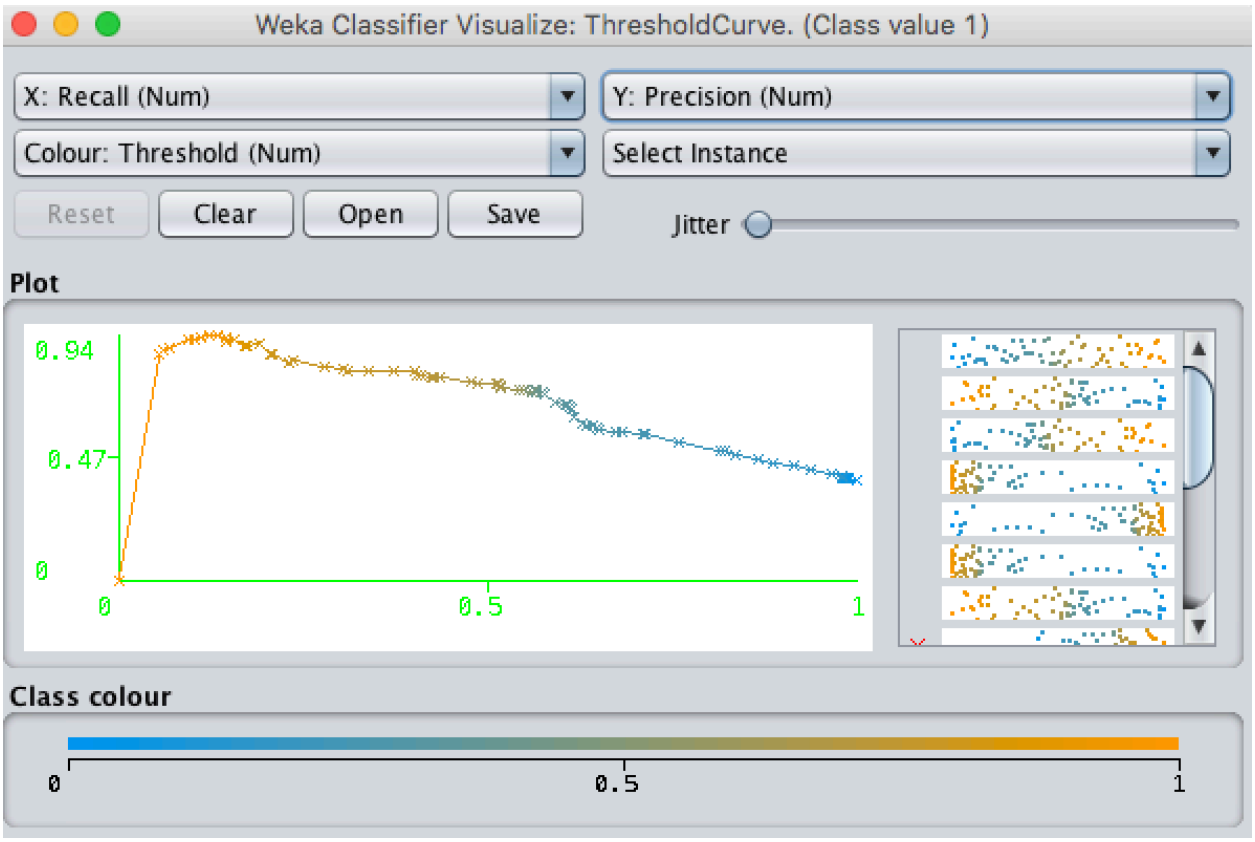
J48:

ROC:



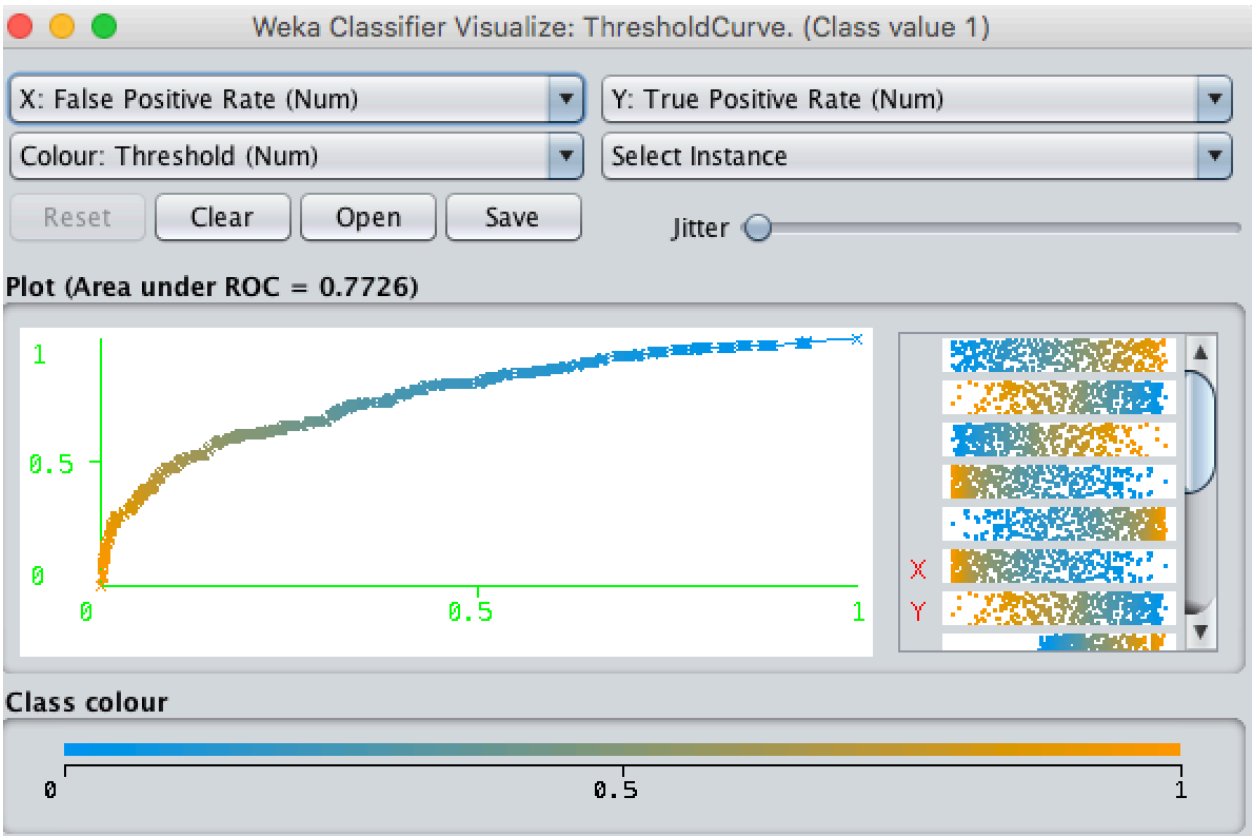
Precision and Recall:

Akshay Gade
G01025094



Random Forest:
ROC

Akshay Gade
G01025094



Precision and Recall

Akshay Gade
G01025094