**AIT 582: Application of Metadata in Complex Big Data Problems**

**Individual Project Report, Fall 2017**

**Meena Rapaka**

**G01101759**

**Project Overview:**  
To gain hands-on experience to extract and utilize metadata in the data mining process.

**Scenario of the Project:**

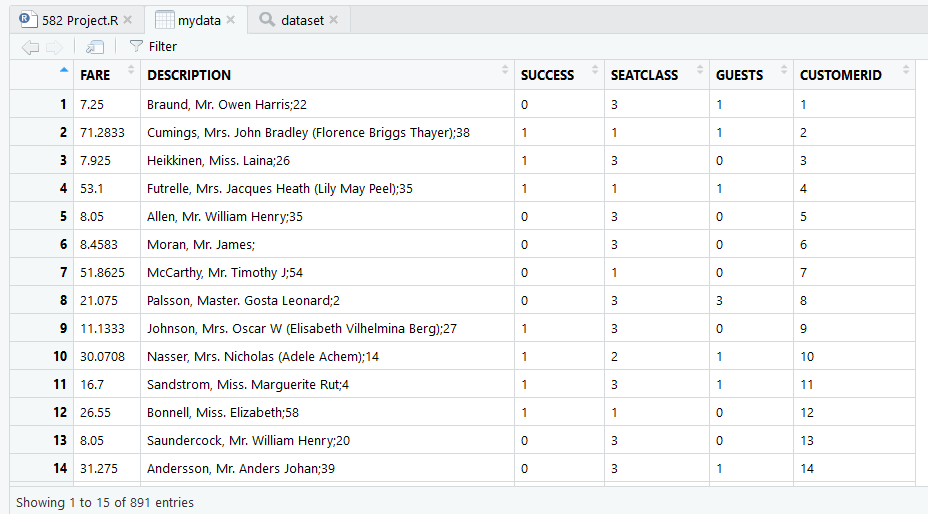
You are a data scientist for an airline A, and you analyze a customer database. You want to identify the factors that are helpful to understand why some

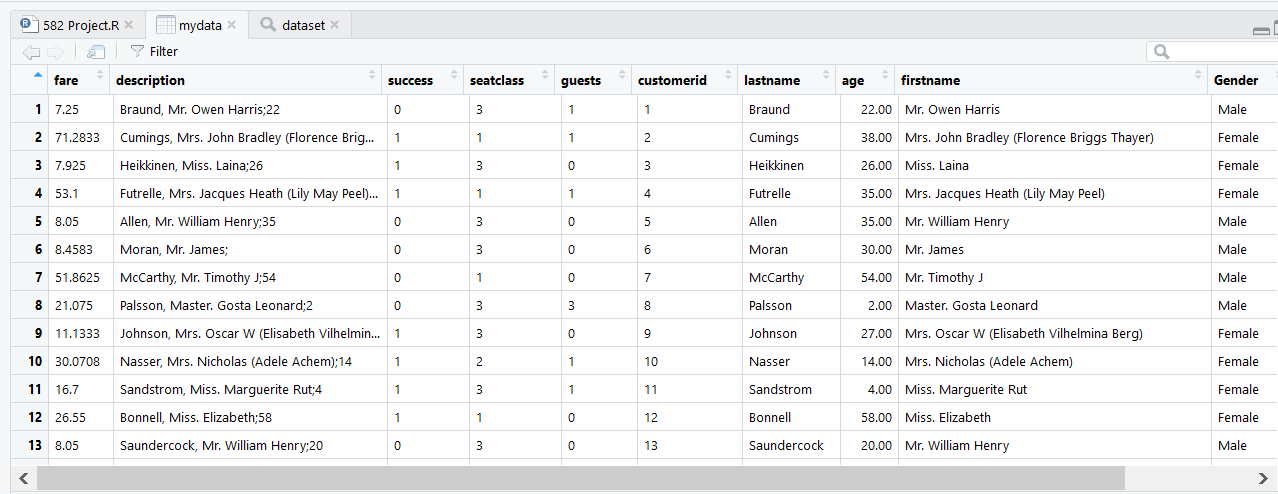
customers are flying your airline, and why others are canceling. Your data science team

wants to recommend these factors to advertising team, such as demographic-specific

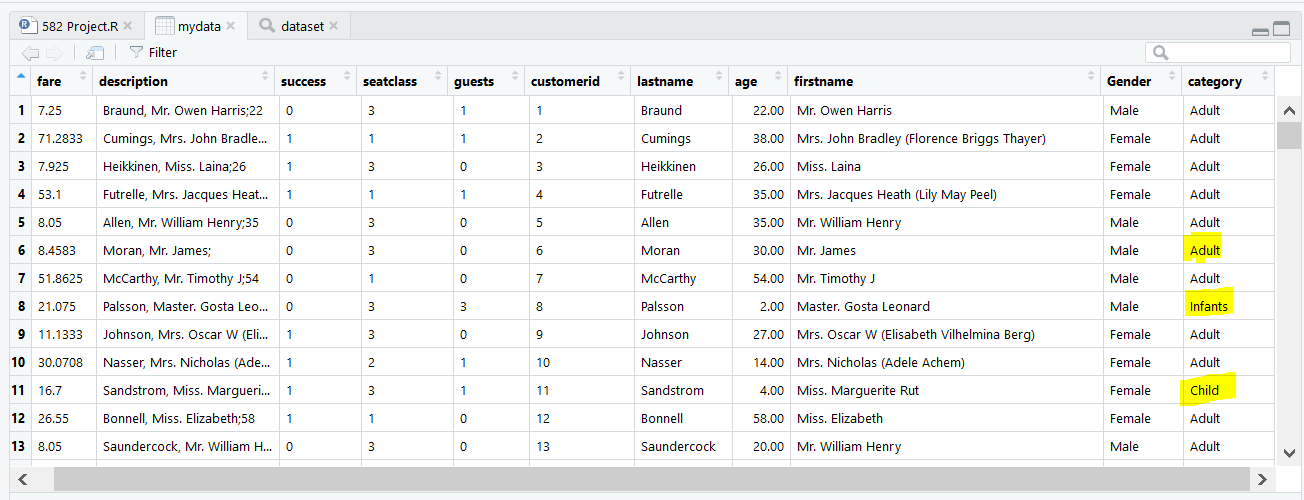
packages to attract more customers.  
  
**Milestones:**We have five milestones to fulfill for the key task completion.  
  
1. **Data Acquisition and Conversion:** The target of this step is to download and then convert the given **(**<http://ist.gmu.edu/~hpurohit/courses/ait582-proj-data-spring16.json>**)** JSON file into Comma separated values format by using R or Python Programming languages.   
Hence, I have used R programming language to convert the downloaded JSON file into CSV format. After the conversion, we find that there is an extra header which is to be deleted. The converted data is as below:





2. **Metadata Extraction and Imputation:** The intent of this step is to extract the metadata and identify the types of metadata based on the given fields (Description) and add additional metadata records.  
The data I have in Description field is further split into Firstname, Lastname, Age and Gender using strsplit function. Based on the title for the Gender I categorized them into Male, Female and Other using grepl function.  


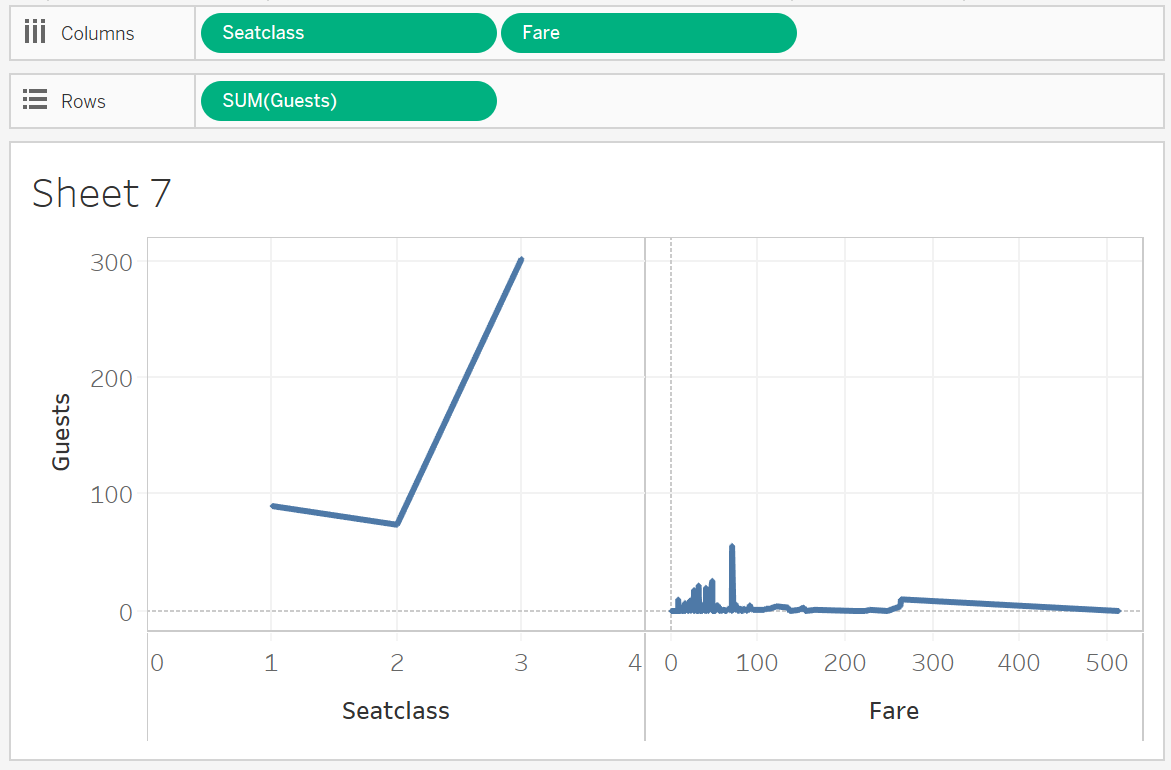
For the Age column, we can see there are some null values. For the missing age column, I have used Mean Imputation method to replace the NA values with the mean of the values in Age column which is equated to be “29.75889” which cannot be assigned as a person’s age. So, the value is rounded off to the nearest number i.e., 30.

I further created a metadata record named ‘Category’ which is categorized based on the value in the Age column as Infant, Child, Adult and Elder.  


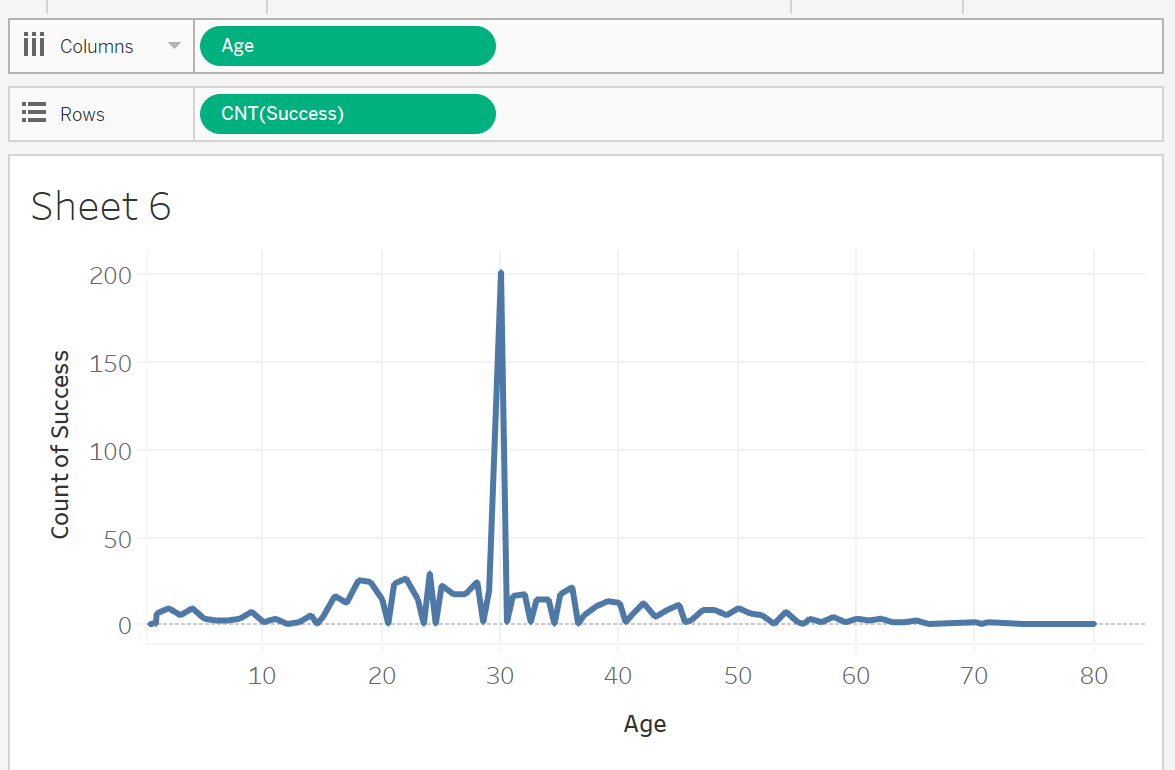
3. **Metadata Exploration:**  
 In this step, I explored the metadata records we got in the previous step and observe patterns of distributions between each of metadata fields and generate visualization using Tableau application.

Below are the visualizations for different metrics:

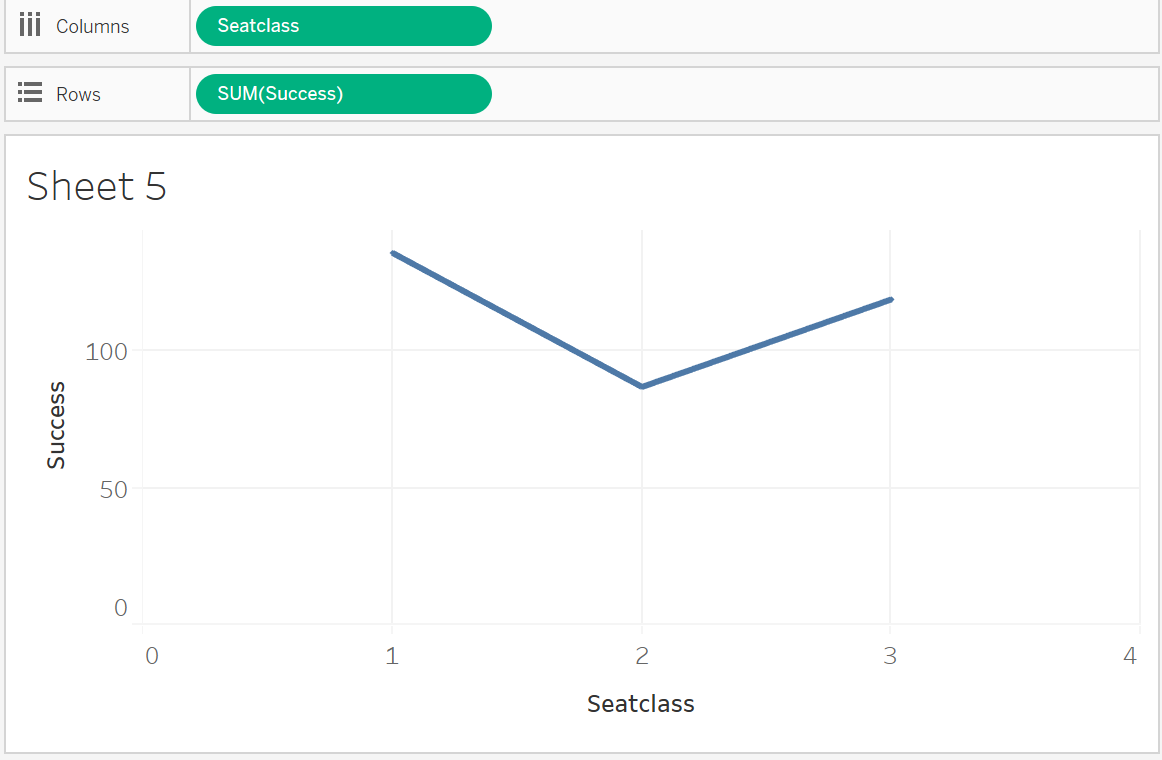
1. **Guests Vs Seatclass, Fare:** In this comparison, we can observe that there are more number of bookings along with guests for seatclass-3 which is followed by seatclass-1 and then seatclass-2.  
   And, we can also observer that for a fair price of around $69.6 there are high number of guests.



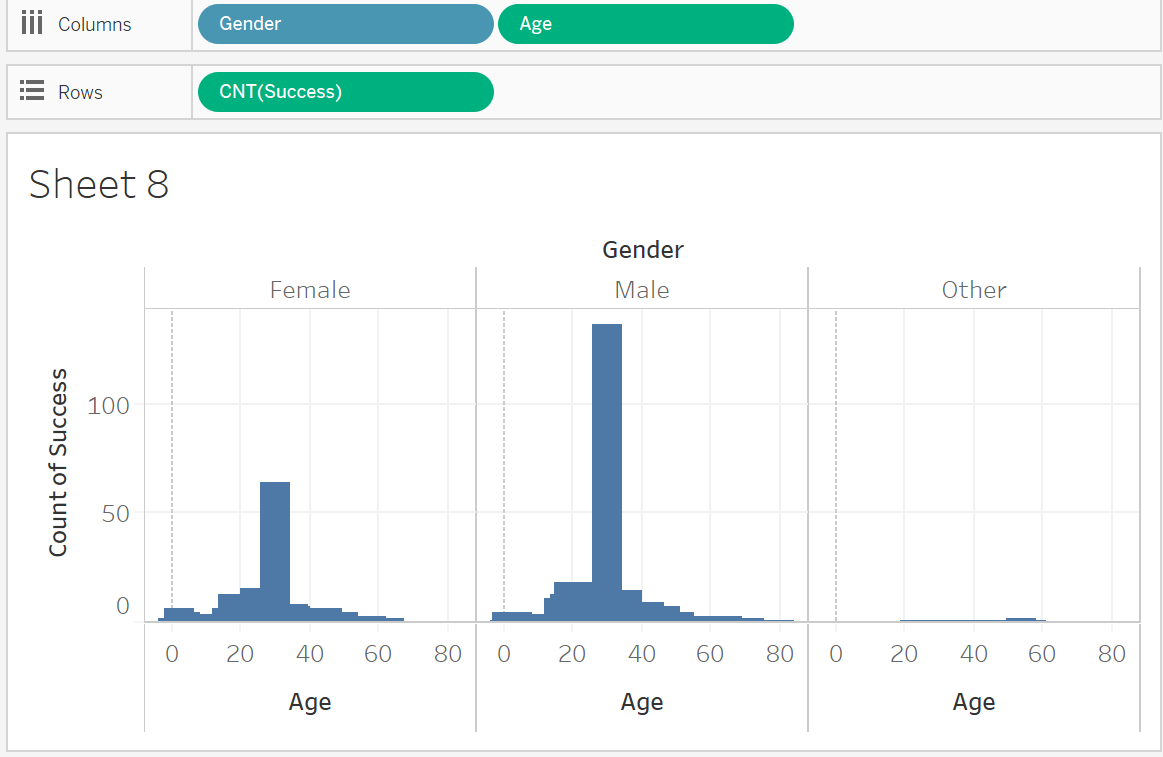
1. **Age Vs Count of Success:** In this comparison between Age and Count of success we can observe that for an Age of 30 there is highly likely for the success count as 1 which implies that for that age range, it is highly likely that the passenger would fly.



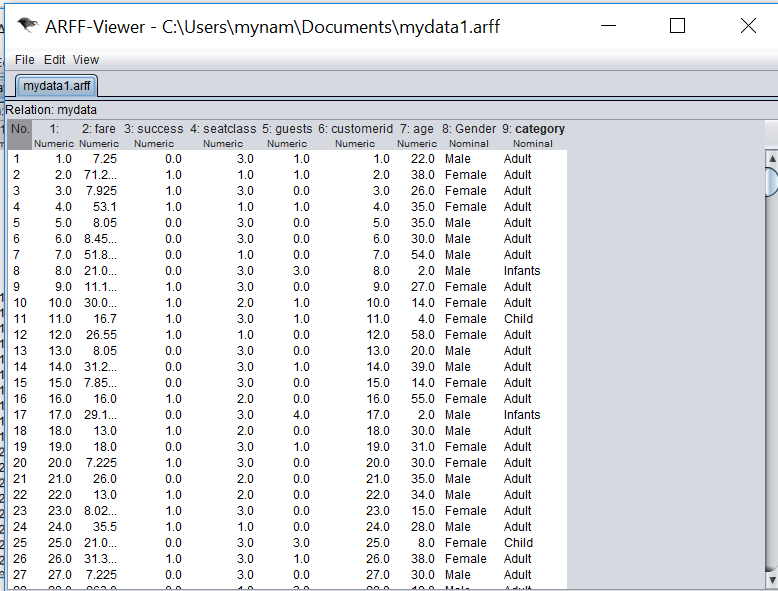
1. **Seatclass Vs Sum(Success):** From this comparison between Seatclass and Sum of Success we could observe that, the success is least for seatclass-2 which is preceded by seatclass-3 and then by seatclass-1 which implies that for the seatclass-1, the success rate is high i.e., more number having seatclass-1 would fly.



1. **Gender, Age Vs Count of Success:** From this, we can observe that for age group of 30 in female have as success count of 64 and male with a age of 30 have a success count of 137 while for the category Other they have a count of success as 2 for an age of 54, which implies that the there are less number of females flying when compared to males.

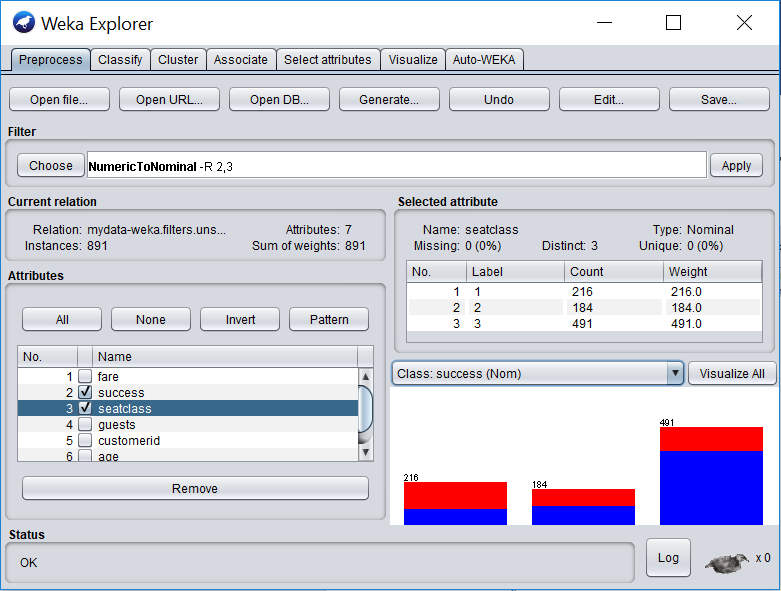


4. **Attribute Preparation and Engineering to prepare data for a supervised learning**

**algorithm using Weka tool:**In this step, we need to load the converted CSV file into attribute representation file i.e., ARFF format using WEKA tool. And then convert the attributes from Numeric to Nominal in ARFF file and identify top two attributes with 10-fold cross validation and report them accordingly.  
  
a) **Importing CSV file into WEKA:** The CSV file we obtained after programming through R is now converted into ARFF format using the ARFF viewer available under Tools tab of Weka.  
We load the CSV file into the viewer and convert it into ARFF format.  
  
 

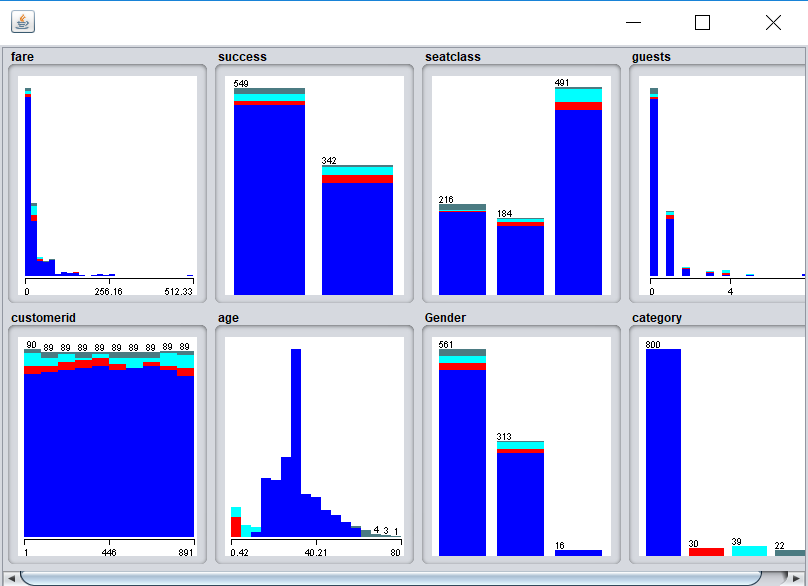
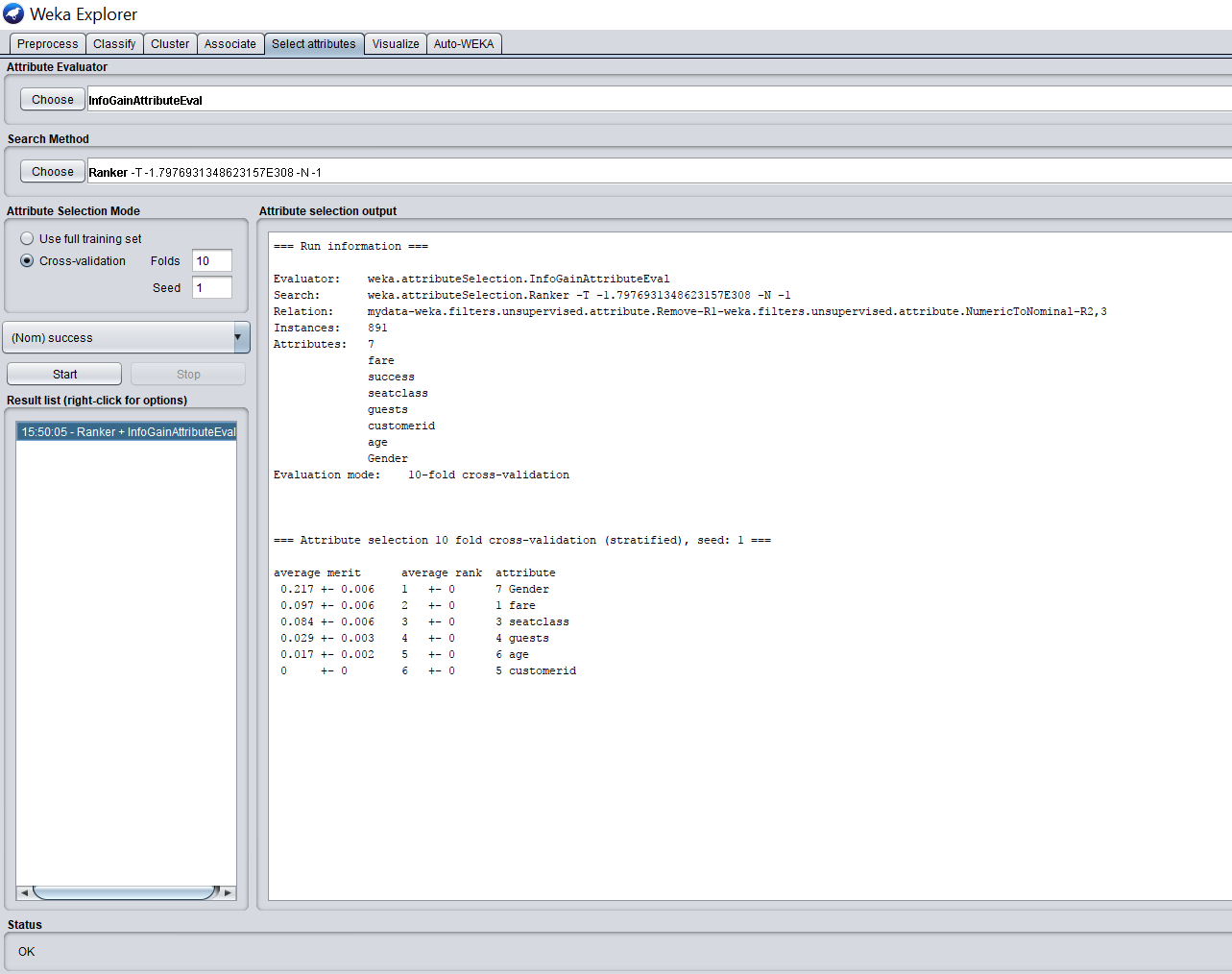
Now, using the Explorer application in Weka, we preprocess the data by loading the file and choosing the NumericToNominal filter available under Choose → Filter → Unsupervised → NumericTo Nominal.  
Here, we select the seatclass and success to convert them to Nominals and apply the changes henceforth. After the conversion, we could see that success, seatclass and gender have nominal values as they have either 1,2 or 3 values for seatclass and values 0 or 1 for success and Male, female or other for gender. They can only have either of the value for each record.

Below, we can see how the data is pre-processed and the NumericToNominal filters are applied and also visualizations of all fields after the data pre-processing.



Now, to identify the top two attributes, we go to the Select Attributes tab under Weka Explorer and set the search method as Ranker and choose the Attribute Evaluator as InfoGainAttributeEval and set the Attribute Selection Mode to Cross-validation, and set Folds value as 10. And the select the (Nom)success and run the selection list.

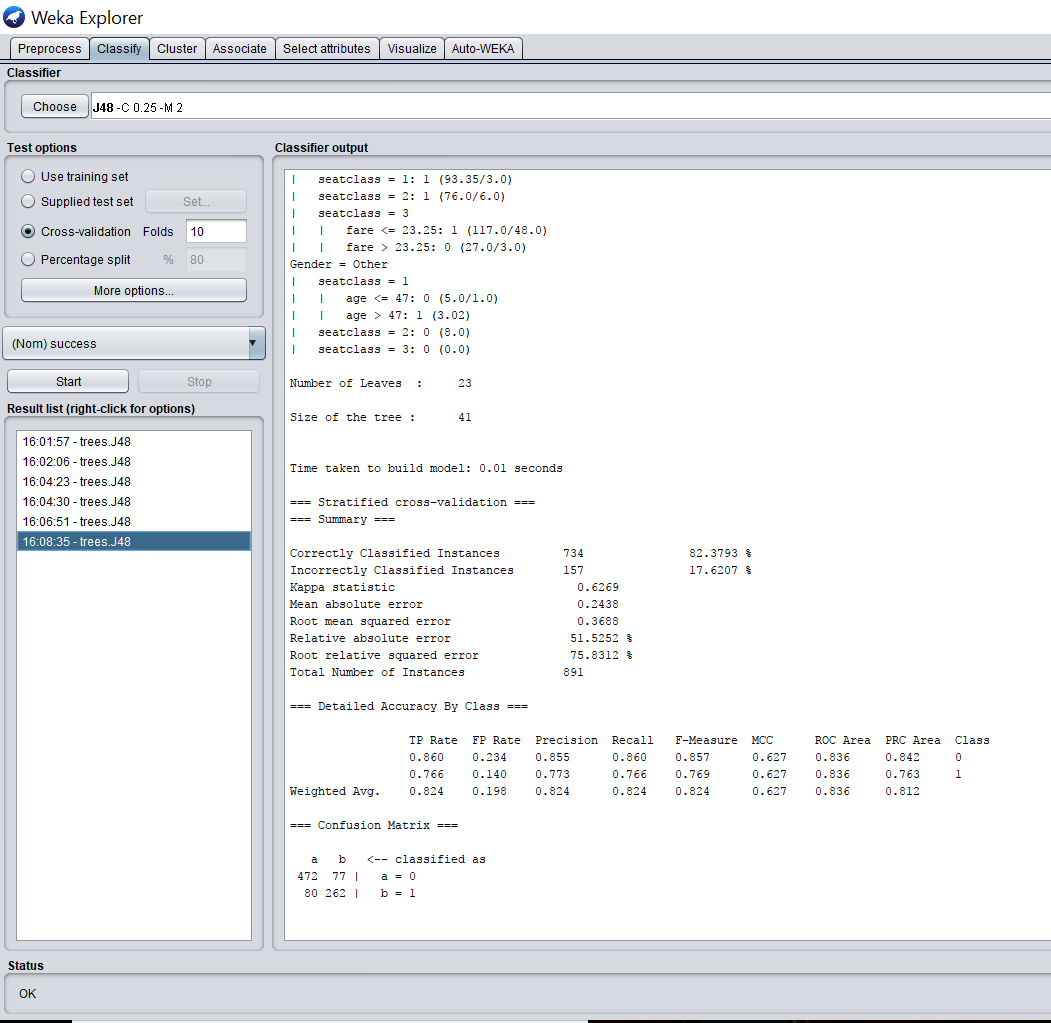
We could see the top two identified attributes as Gender and Fare based on 10-fold cross validation technique.

5**. Prediction Modeling and Visualization:**  
 The intent of this step is for us to learn a classification model such as Decision tree-J48 algorithm and Random forest algorithms and then visualize the performance using Precision-Recall curve for 10-fold cross validation setup.

1. **Classification Model:** For the classification, we select the Classify tab under Weka explorer and choose a classifier accordingly.

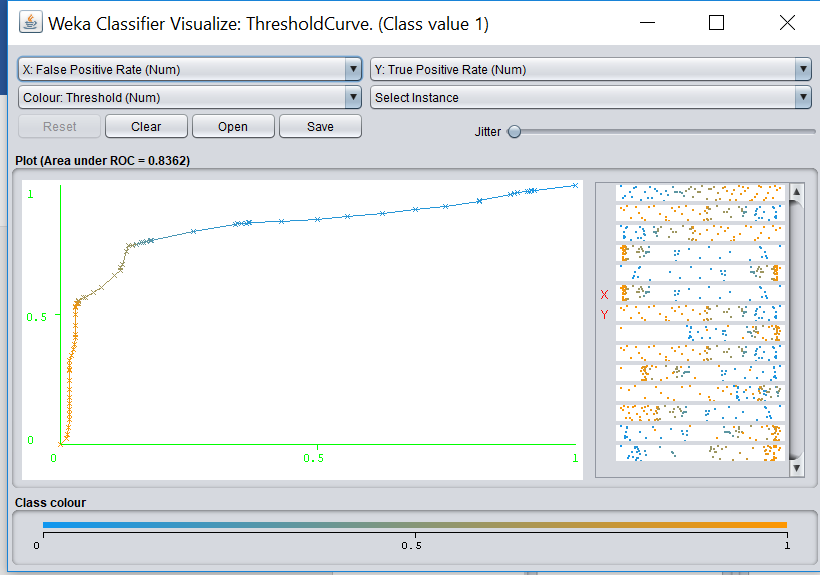
**a. Decision Tree-J48:** We set the Classifier to be J48 and select the Test option as Cross Validation with 10 folds and (Nom)Success and run the algorithm.  
We will get the result as below with 82.3793% of correctly classified instances.

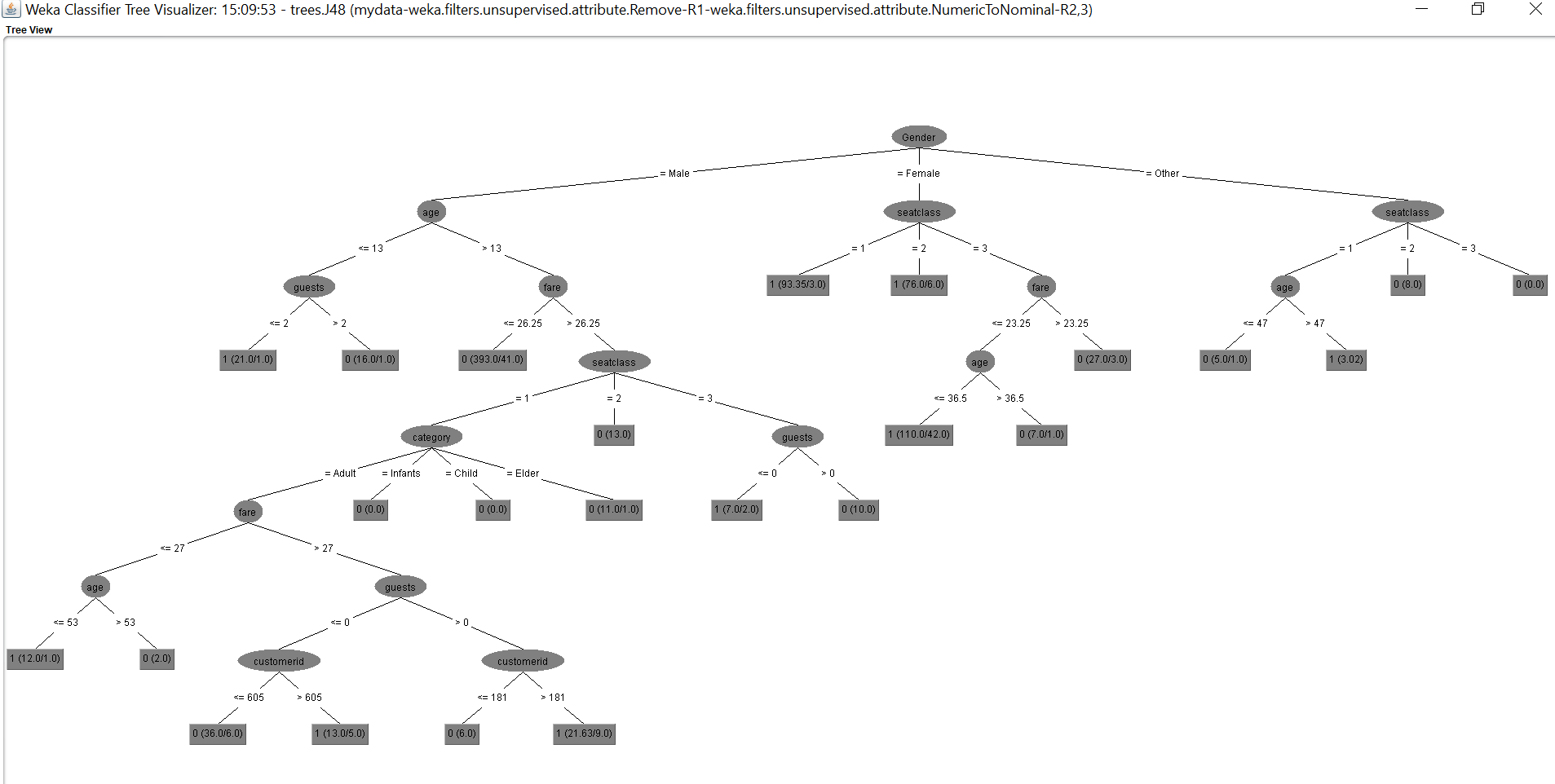


Now, by right-clicking on the result set and selecting the visualize threshold curve for success value 1, we can view the ROC curve as below.

From the curve, we can make out that there are less no of false positives and more of true positives and the curve is inclined towards the true positives with an area under ROC = 0.8362 (which is almost closure to 1) indicating this is one of the good solutions.

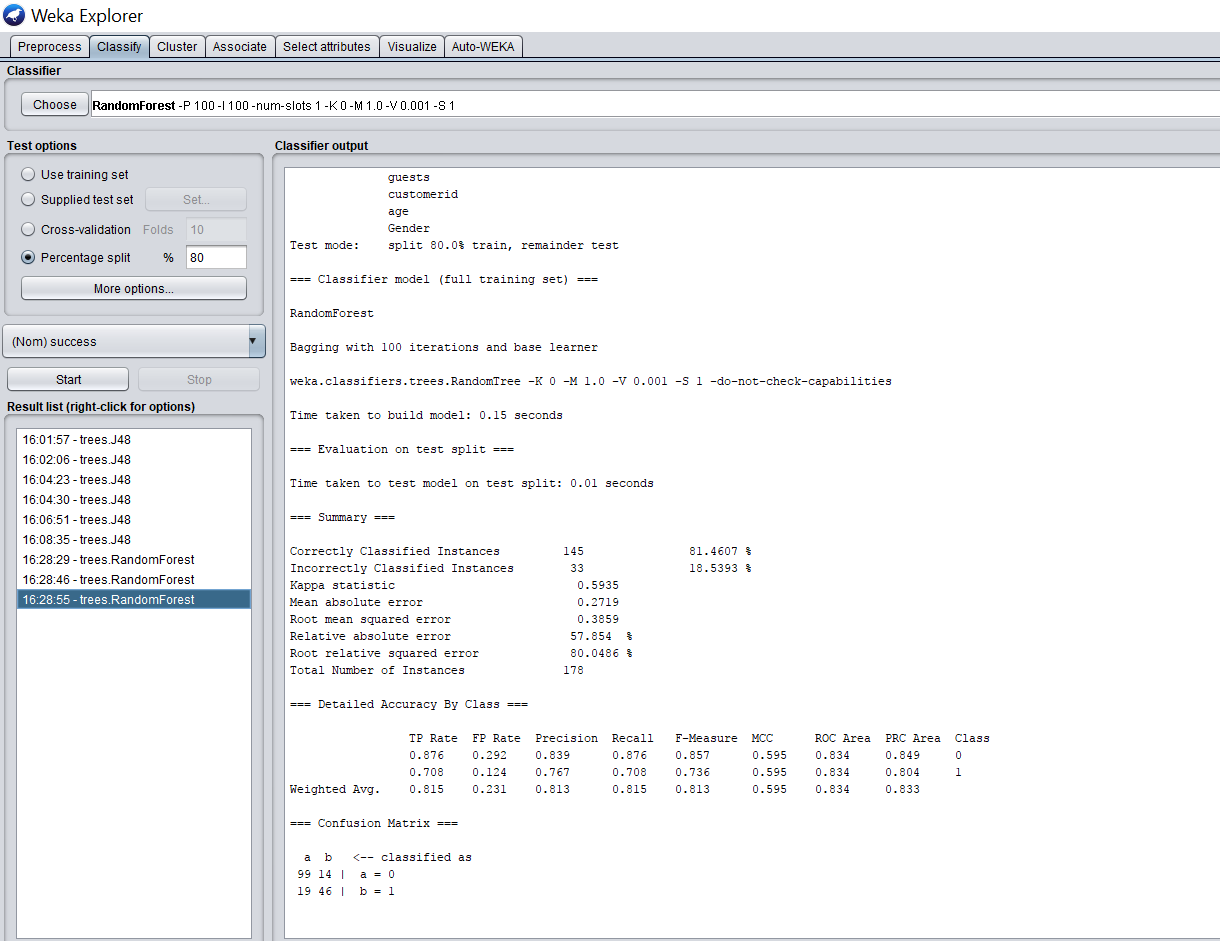
We can also view the Decision tree model by selecting the visualize tree option available for the result set. We can see that Gender is considered as a parent node for the tree.





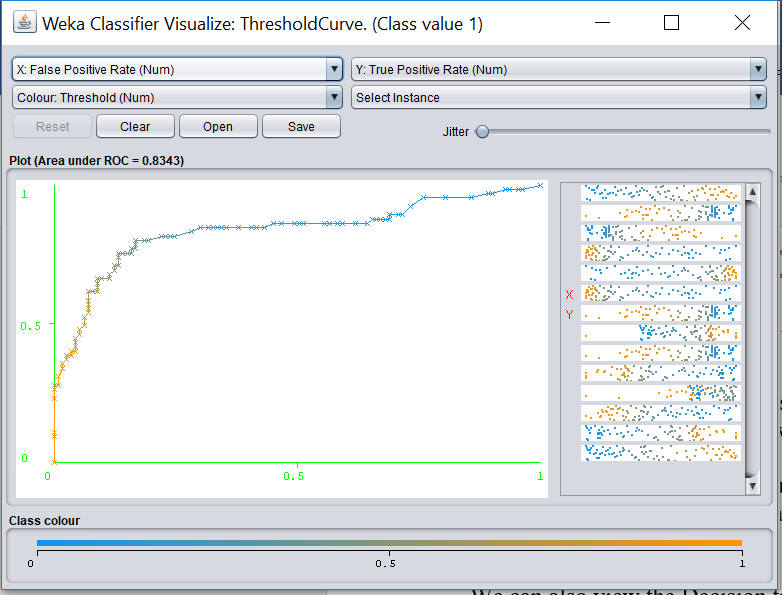
**b. Random Forest:** To apply the Random Forest algorithm, the procedure is same as J48 algorithm but instead we choose the classifier as Random Forest and run the algorithm.

We can see a result of 81.4607 as the correctly classified instances.



Now, by right-clicking on the result set and selecting the visualize threshold curve for success value 1, we can view the ROC curve as below.

From the curve, we can make out that there are less no of false positives and more of true positives and the curve is inclined towards the true positives with an area under ROC = 0.8343 indicating this is also one of the good solutions.



Based on the observations on the results from the results of algorithms, we can make out that Decision tree is a best solution as in robust and accurate one.

1. **Receiver operating characteristic(ROC) curve:**An ROC curve is created by plotting the true positives against the false positives at various threshold settings that illustrates the ability of a classifier system as its discrimination threshold varies.

The more the curve is inclined towards true positive, the more accurate is the solution.

For the J48 algorithm, we can see that there are 734 correctly classified instances with an accuracy of 82.4%. Hence it is considered as a best solution.

From all the pre-processing and analysis of metadata, we can analyze that   
Customers who are above 24 are more likely to travel and the success rate is more for male when compared to females where the top two attributes predicted to be are Gender and Fare.