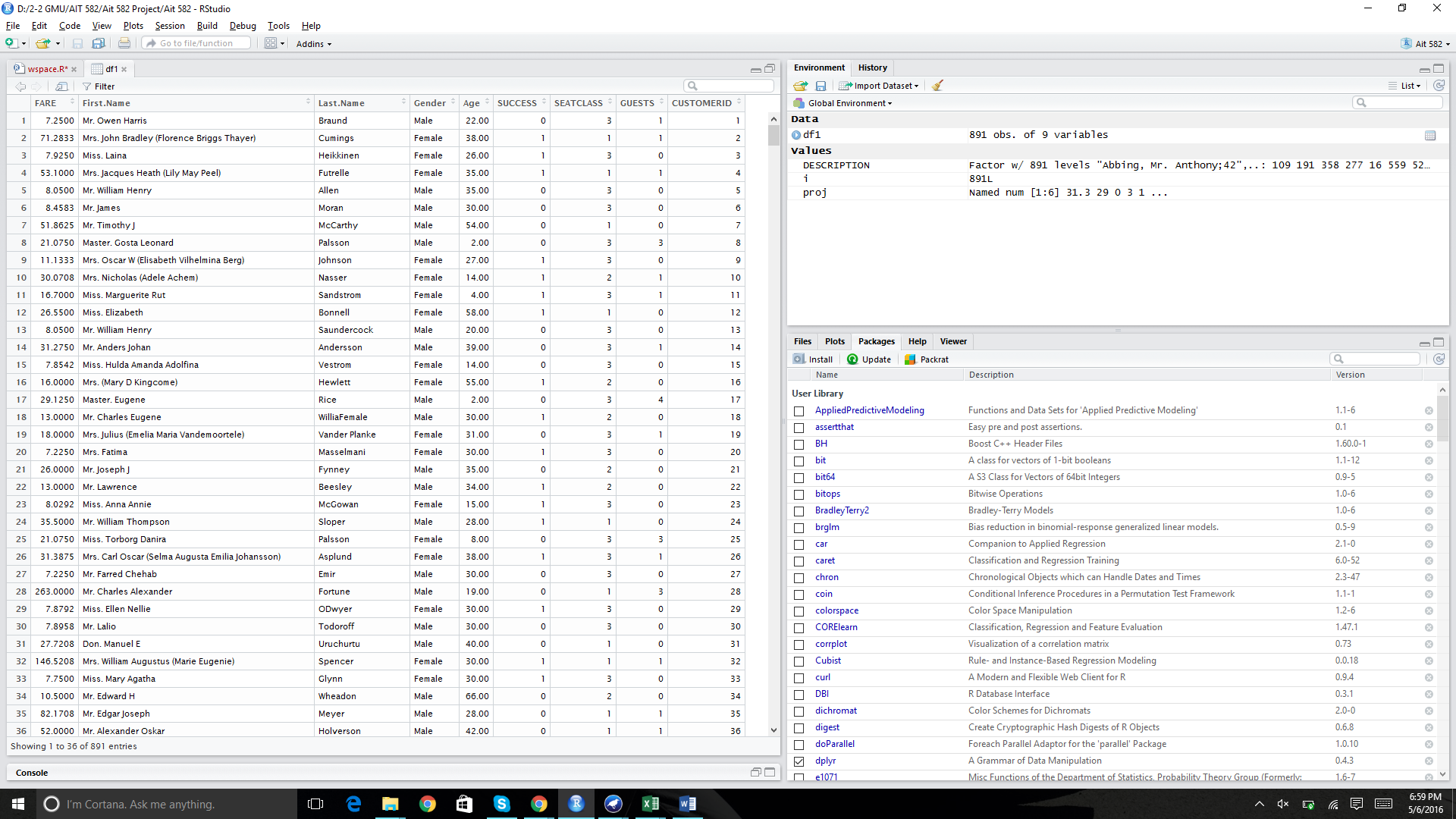
AIRFARE METADATA

848-667-4511 | sai9100091@gmail.com

Ranga Sai

2016

1. Data Acquisition and Conversion



1. Metadata Extraction and Imputation
   1. Which metadata types do you observe in the given data field of DESCRIPTION and which implicit semantic metadata can be derived?

Age, Gender

* 1. Can you extract such metadata and append them as additional fields for each of the data record?

Yes

R Code: for(i in 1:nrow(df1)){

if (is.na(df1$Age[i])== T){

df1$Age[i]=round(mean(df1$Age, na.rm =T))

}

}

mean(df1$Age) : [1] 29.75889

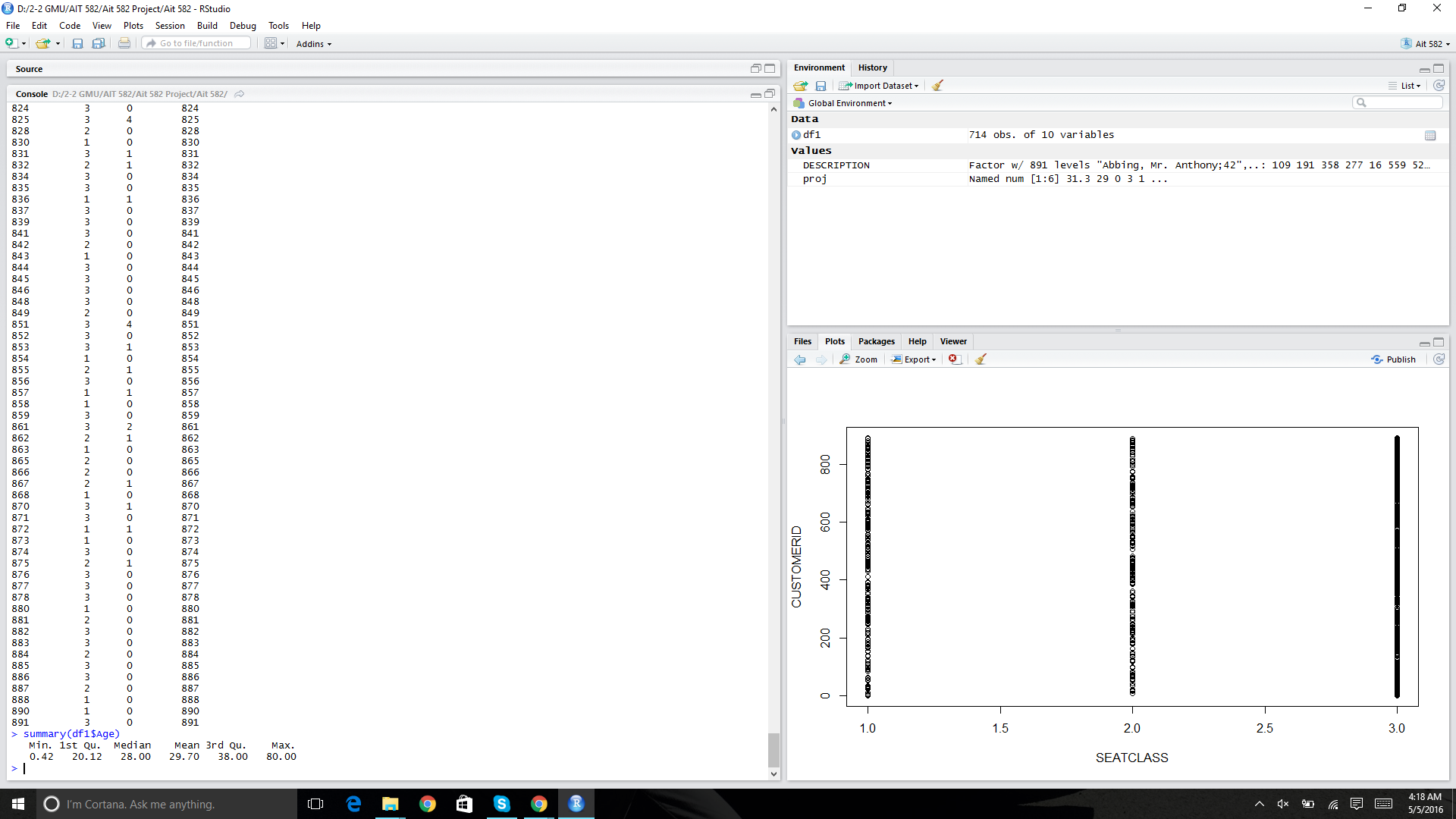
* 1. How can you impute the missing data values? Impute with one of your choice of imputing method.

#Missing Values : df1[is.na(df1$Age),]

summary(df1$Age)

# Removing Missing Values: df1<- df1[!is.na(df1$Age),]

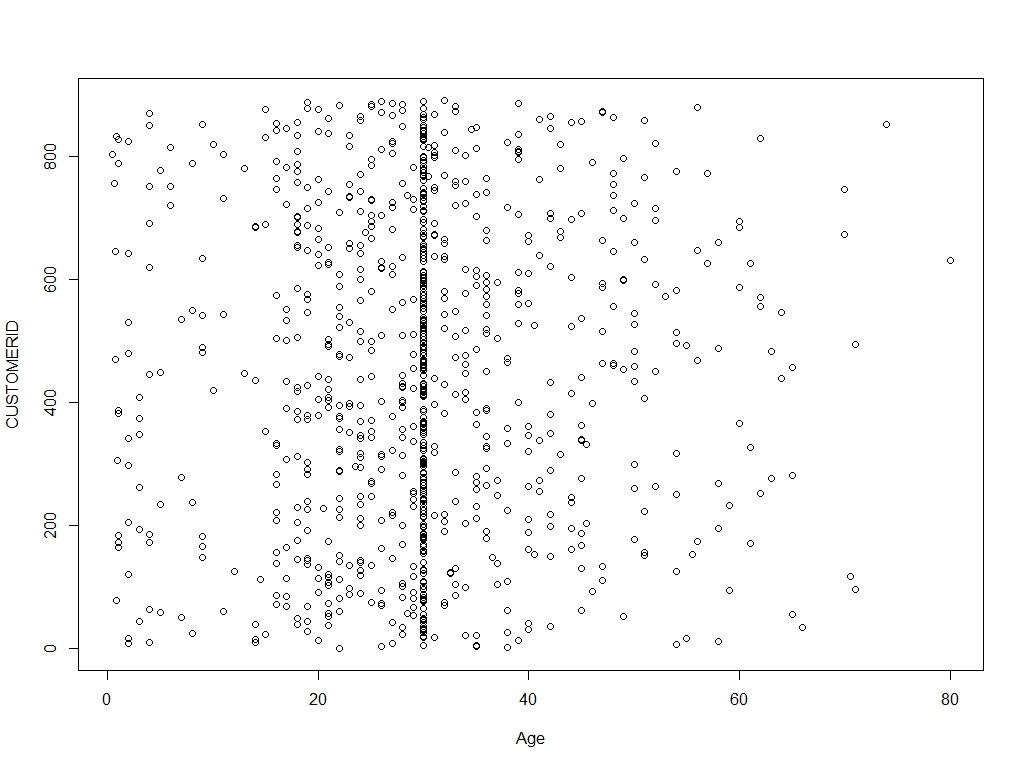
summary(df1$Age)



1. Metadata Feature Exploration
   1. Explore the data records along each of the metadata and the CUSTOMERID, and output distribution chart of data along each of the metadata?
      1. e.g., SEATCLASS has specific classes of 1, 2 or 3. So, how many customers under each of them.

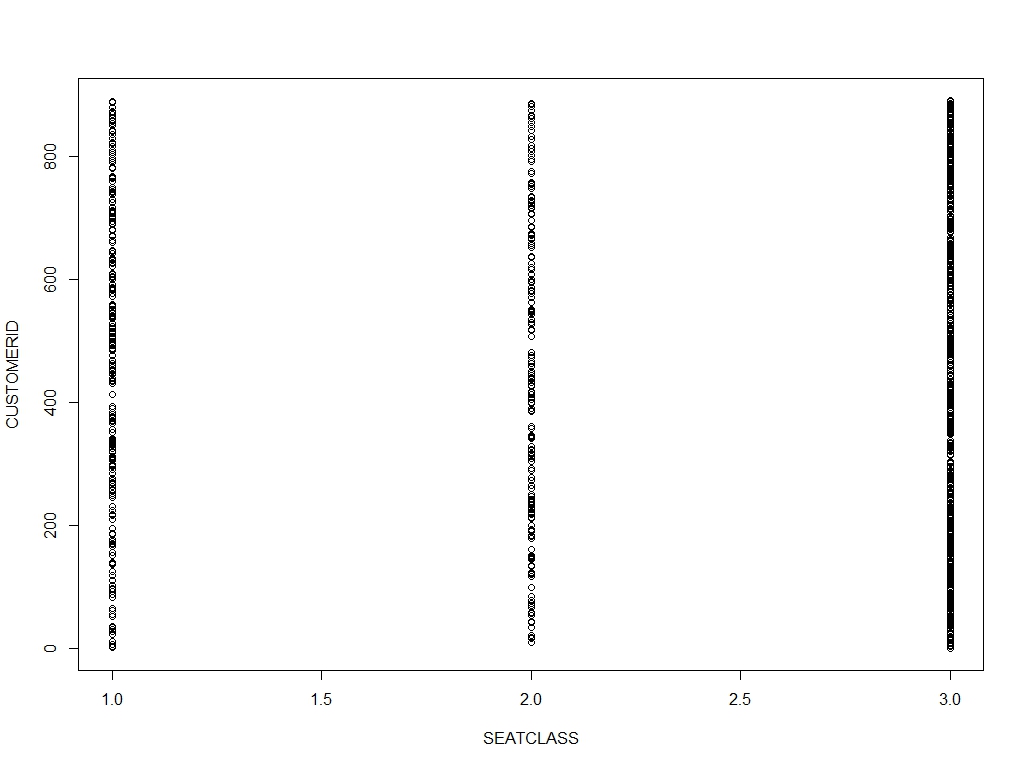
plot(CUSTOMERID ~ Age + Gender + SUCCESS + GUESTS + FARE + SEATCLASS, data=df1)

1. CustomerID vs Age



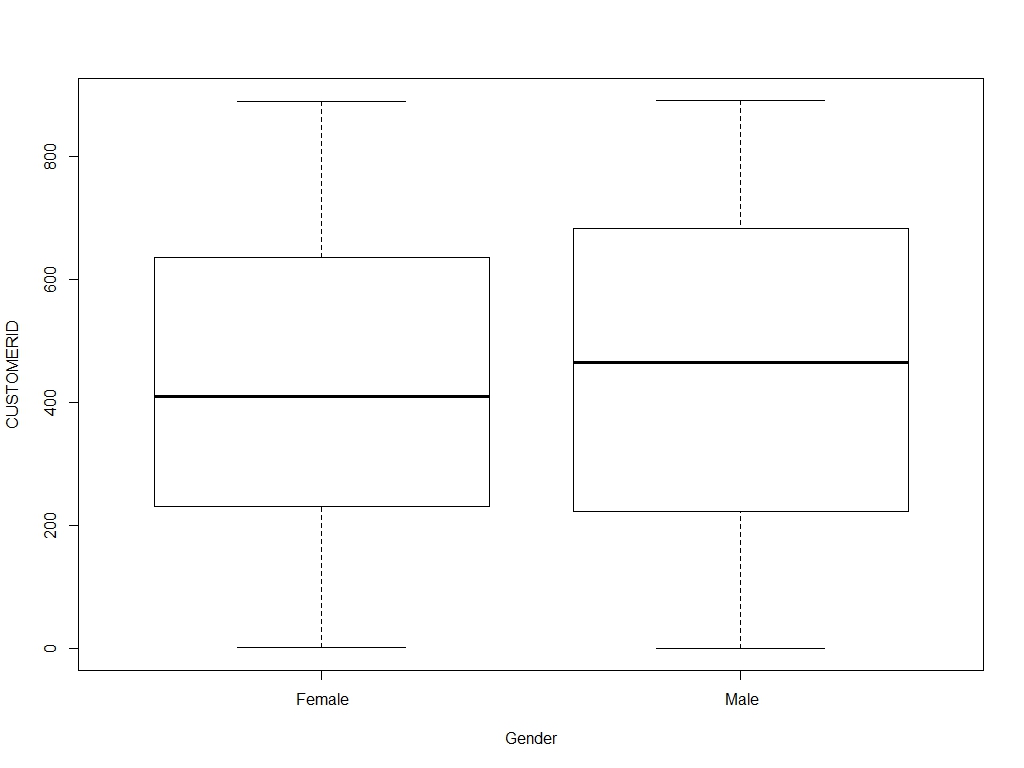
The above plot indicates customers of Age around 30 fly frequently. This might also be a reason because of the NA’s rounded to a mean value of 29.75.

1. CustomerID vs SeatClass

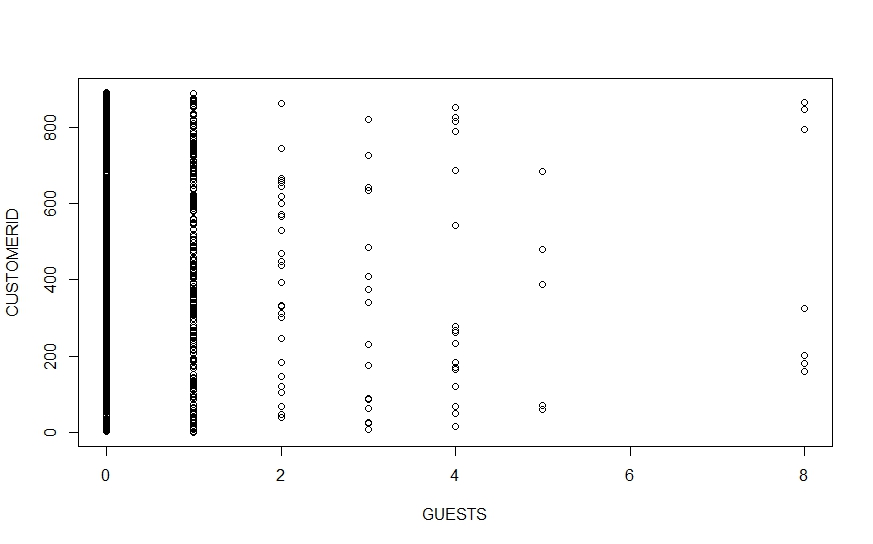


The above plot indicates Class 3 to be used by customers.

1. CustomerID vs Gender

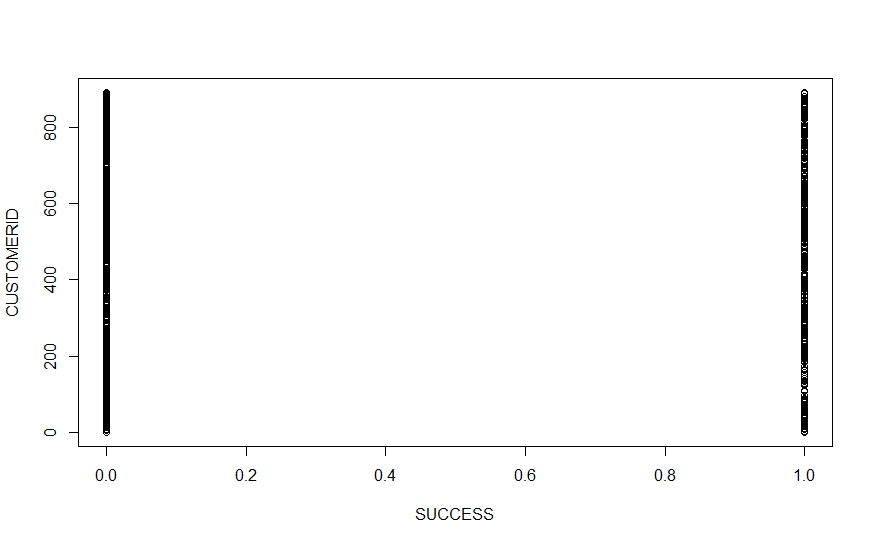


1. CustomerID vs Guests

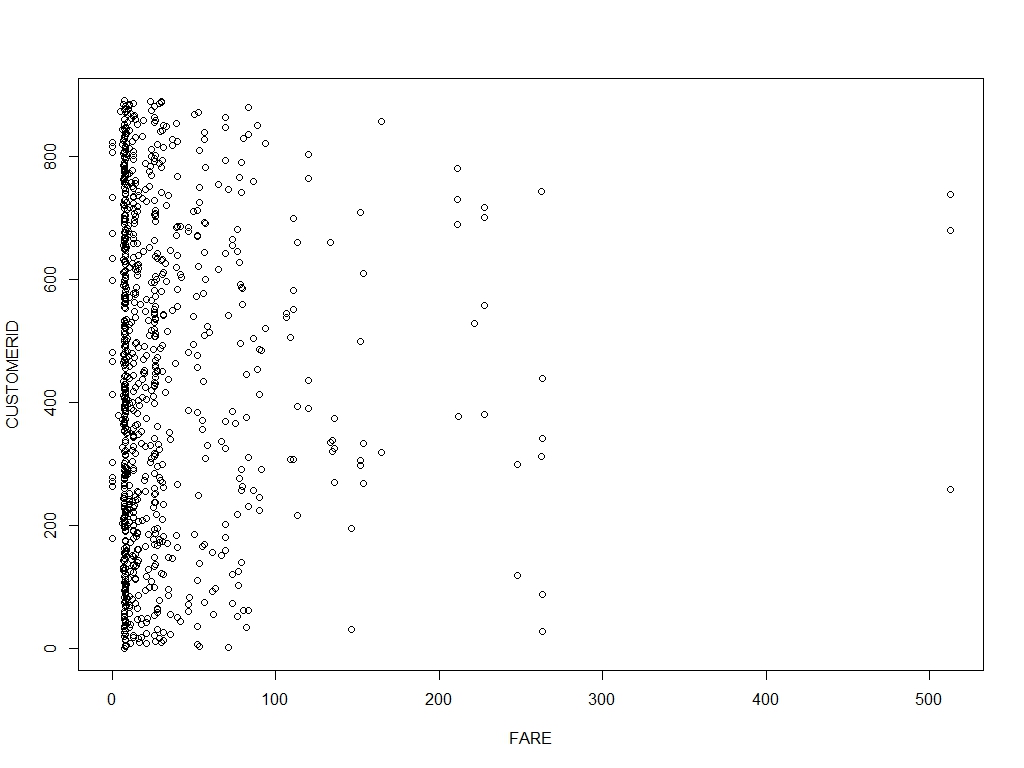


The plot describes that most customer’s prefer traveling with 1 guest at most.

1. CustomerID vs Success

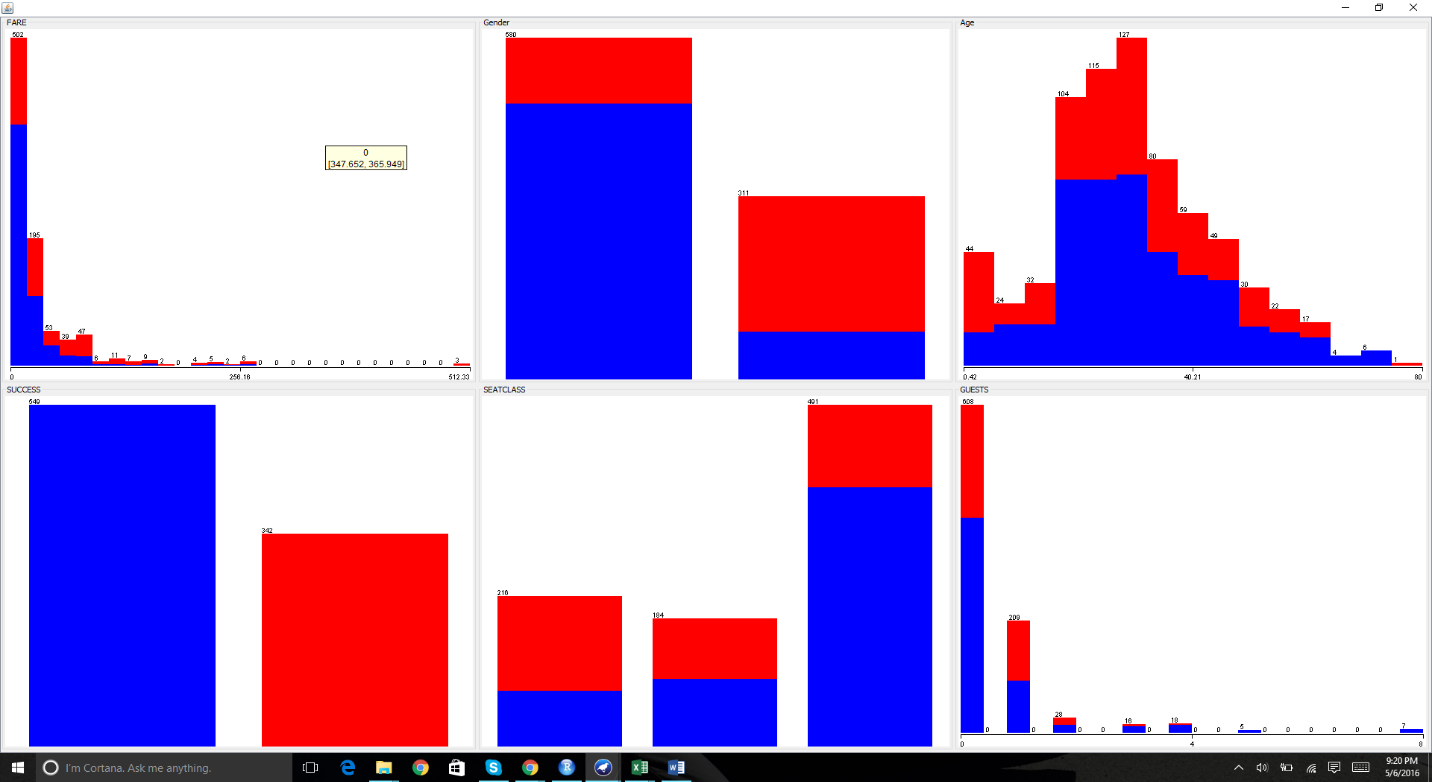


1. CustomerID vs Fare

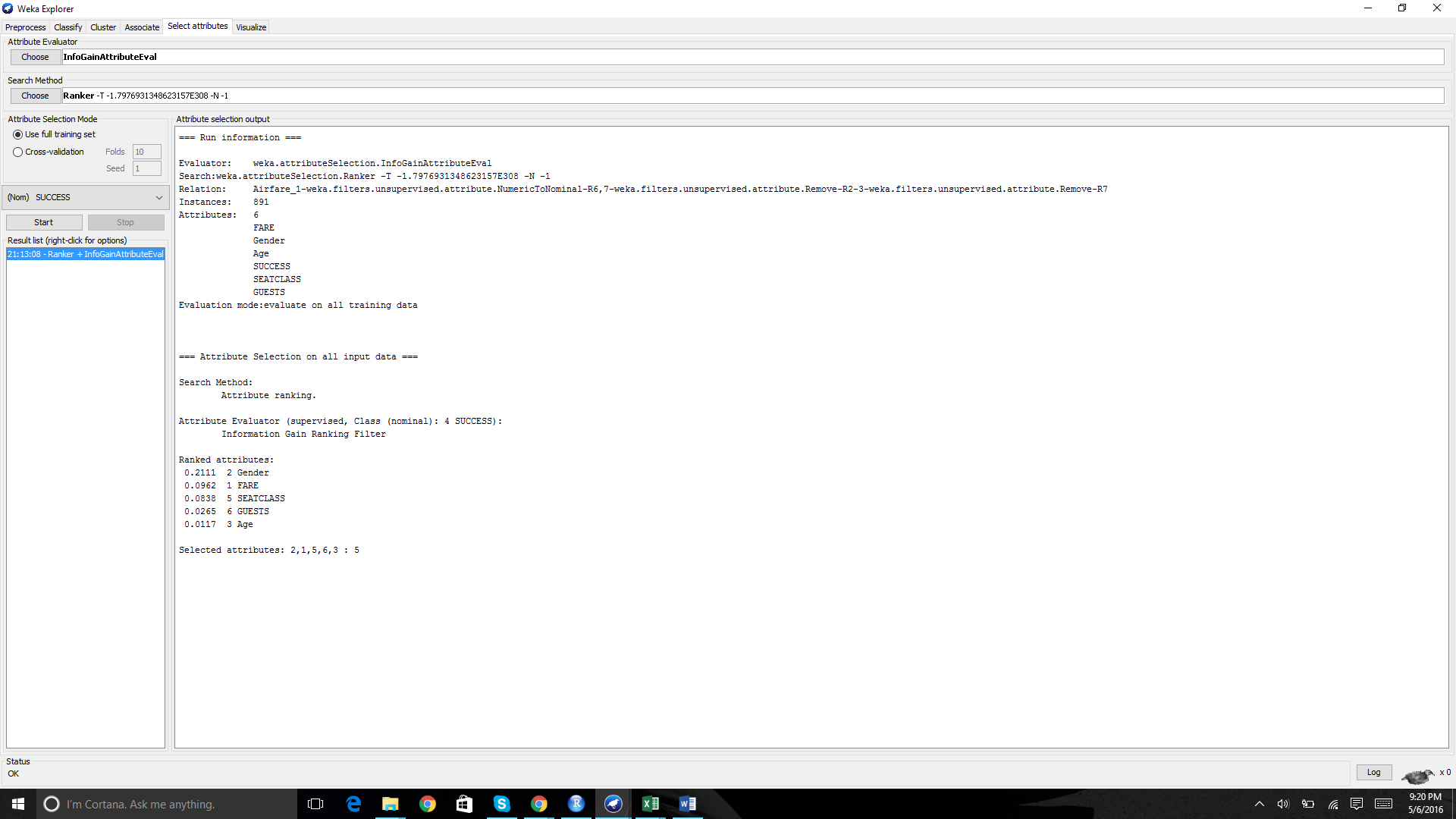


Customer’s travelled mostly when the Fare’s of flight were $10.

1. Feature Preparation and Engineering



Attribute Selection

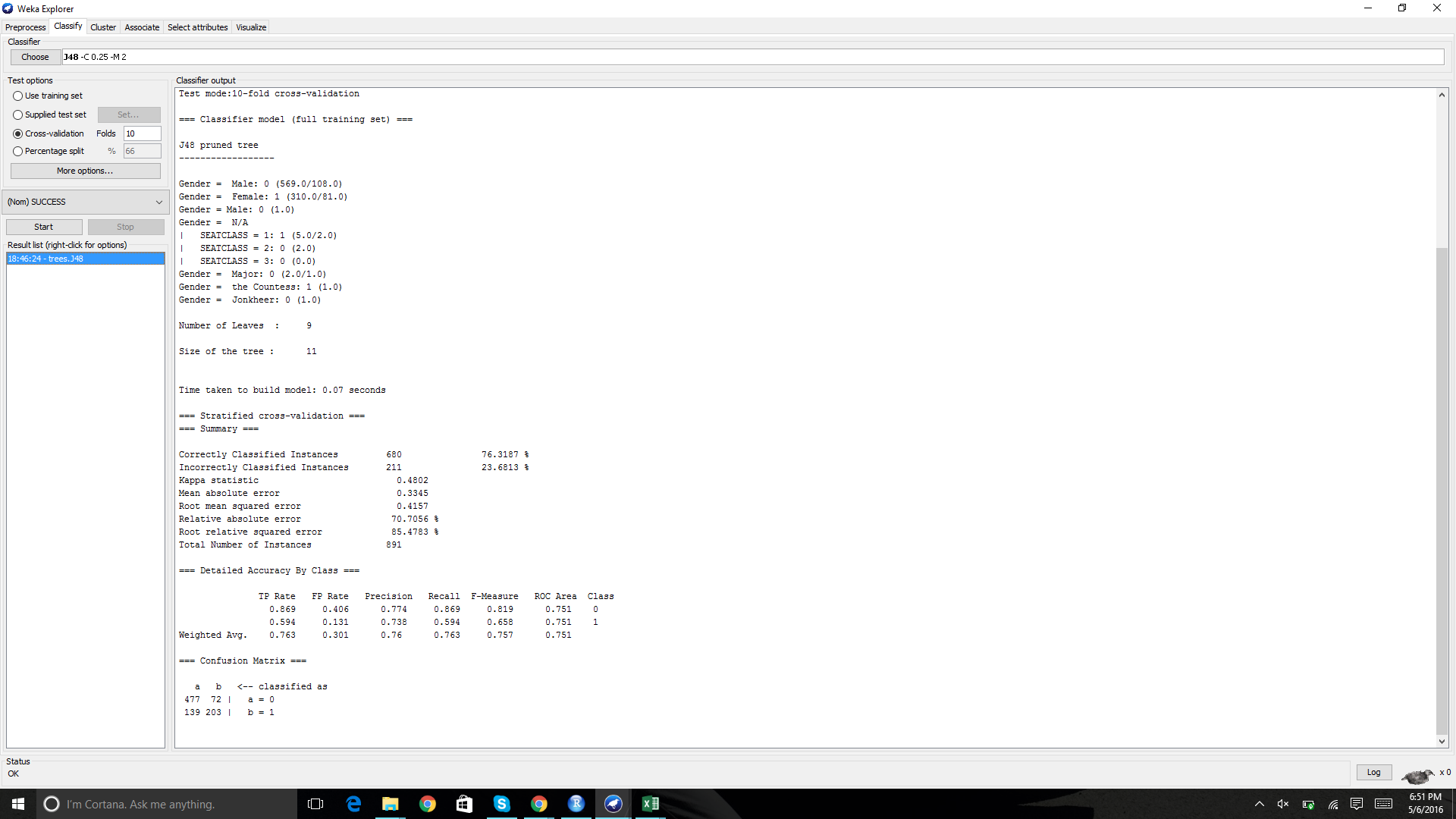


Attribute Selection: For this model we have used the InfoGain evaluator for attribute evaluation and for ranking we have used the Ranker method. InfoGain initially considers the complete set of attributes, and after continuous evaluation, it keeps on removing attributes till the time there is a decrease in the evaluation value. Once it reaches this point the attributes which are still standing will be used to generate the classifier model.

Result: For the SUCCESS class variable Gender variable ranks first and the other variables follow.

* 1. Prediction Modeling and Visualization
  2. Learn a classification model using Decision Tree and Random Forest algorithms.

**CLASSIFICATION using DECISION TREE**

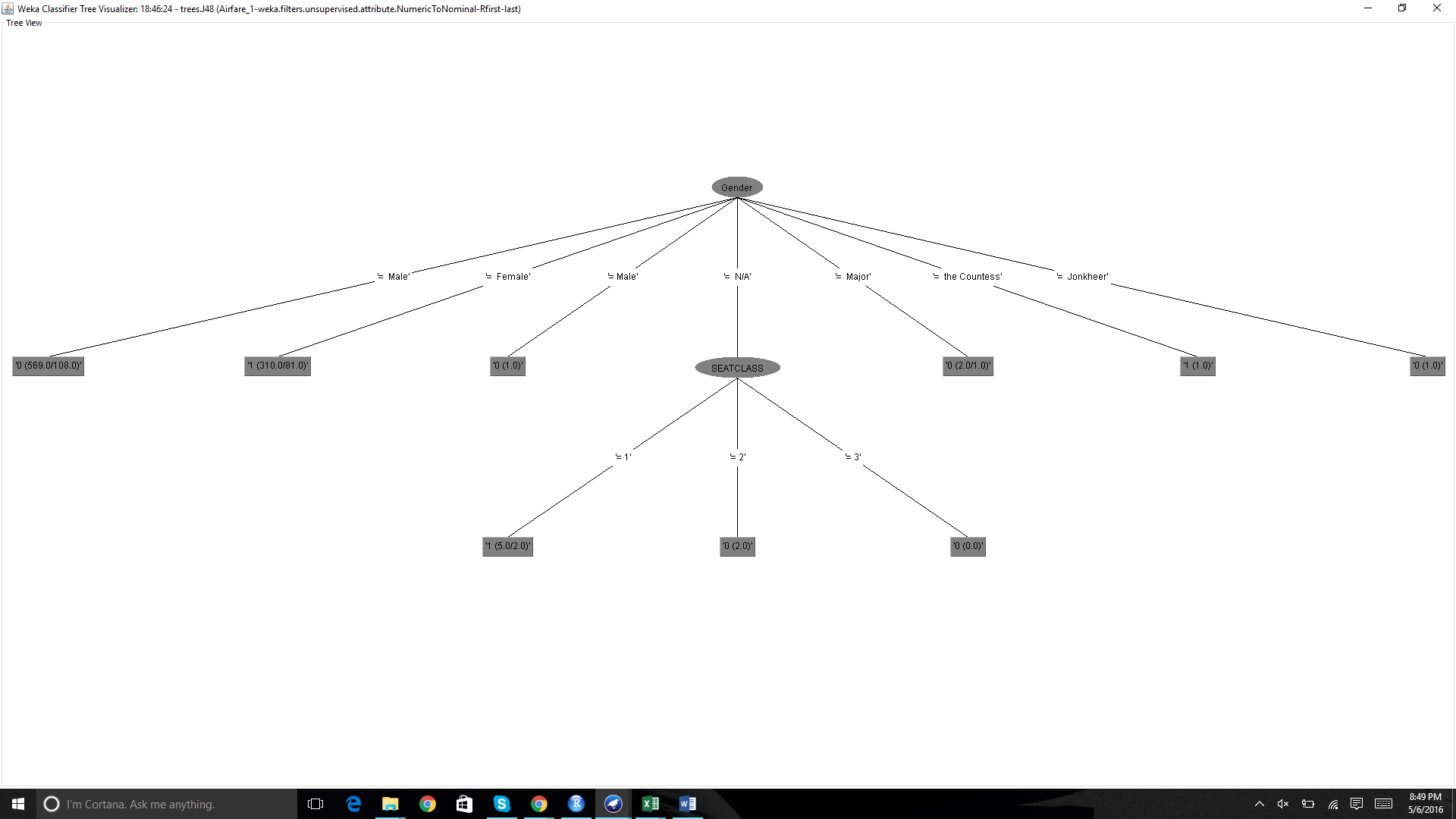


Classification requires the discretization of the Numeric data i.e we have to convert it to the nominal data. The decision tree algorithm makes use of information gain and Entropy for the construction of the decision tree.

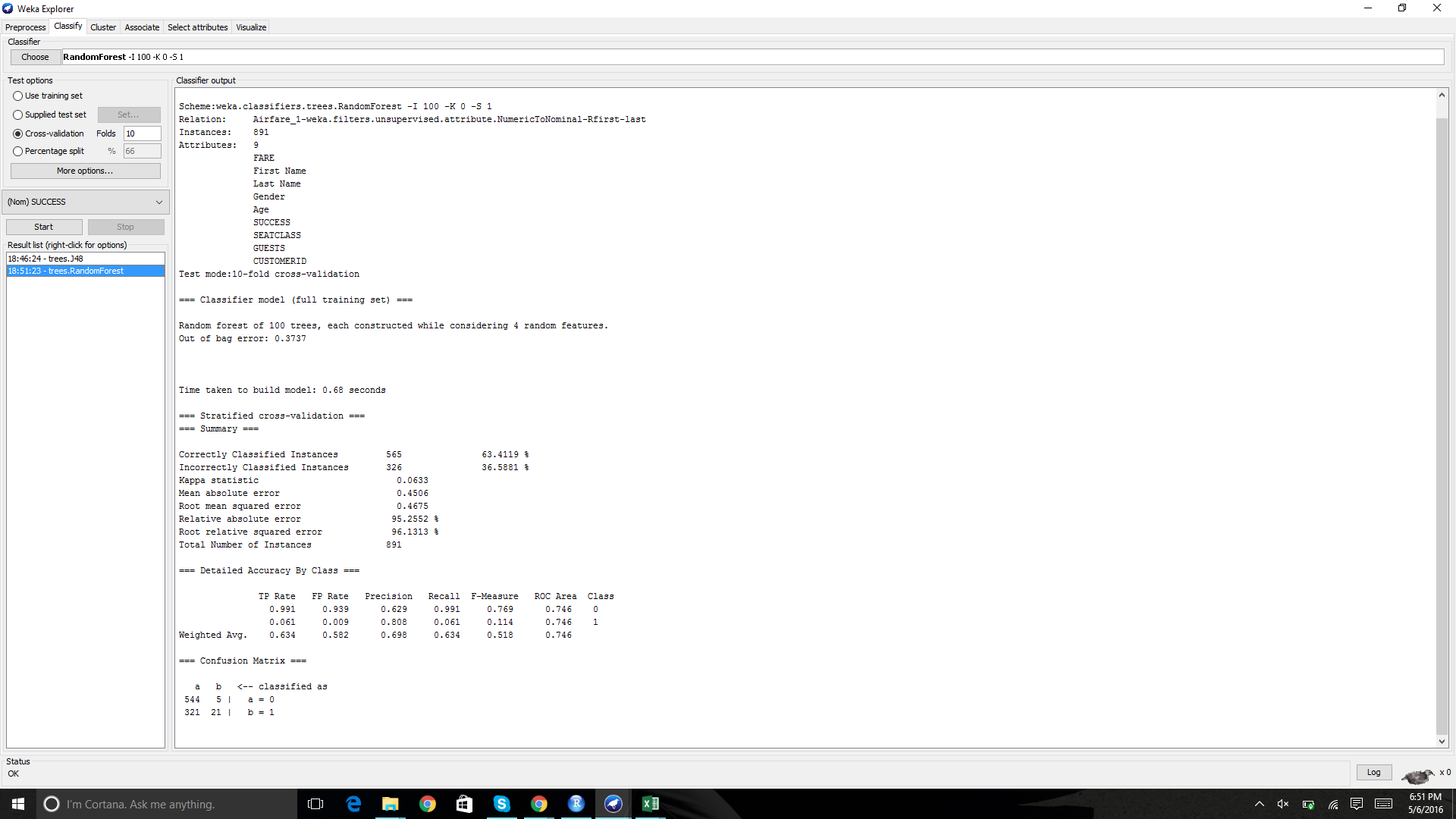
Entropy: J48 uses it to calculate the similarity between the dataset & if the dataset is equally distributed then the entropy is equal to 1.  
  
Information Gain: In this we calculate the entropy before and after splitting the node, this is called the information gain or decrease in entropy. The attribute with the largest information gain is selected as the decision node.

The topmost decision node in a tree which corresponds to the best predictor called root node. In this case, it is gender. What the algorithm does is that it partitions the dataset into subsets that contain instances with similar values and based on the decision nodes the tree will be constructed.

**DECISION TREE**



**RANDOM FOREST**



Random Forest: This is also similar to the working of J48 algorithm but the difference is that in J48 the tree is constructed with the best predictor as root node whereas in the decision tree algorithm few of the best attributes are chosen and then randomly picked amongst them. That gives us different trees every time. Generally, if we bag decision trees, by initially randomizing them and bag the result, we get better performance.

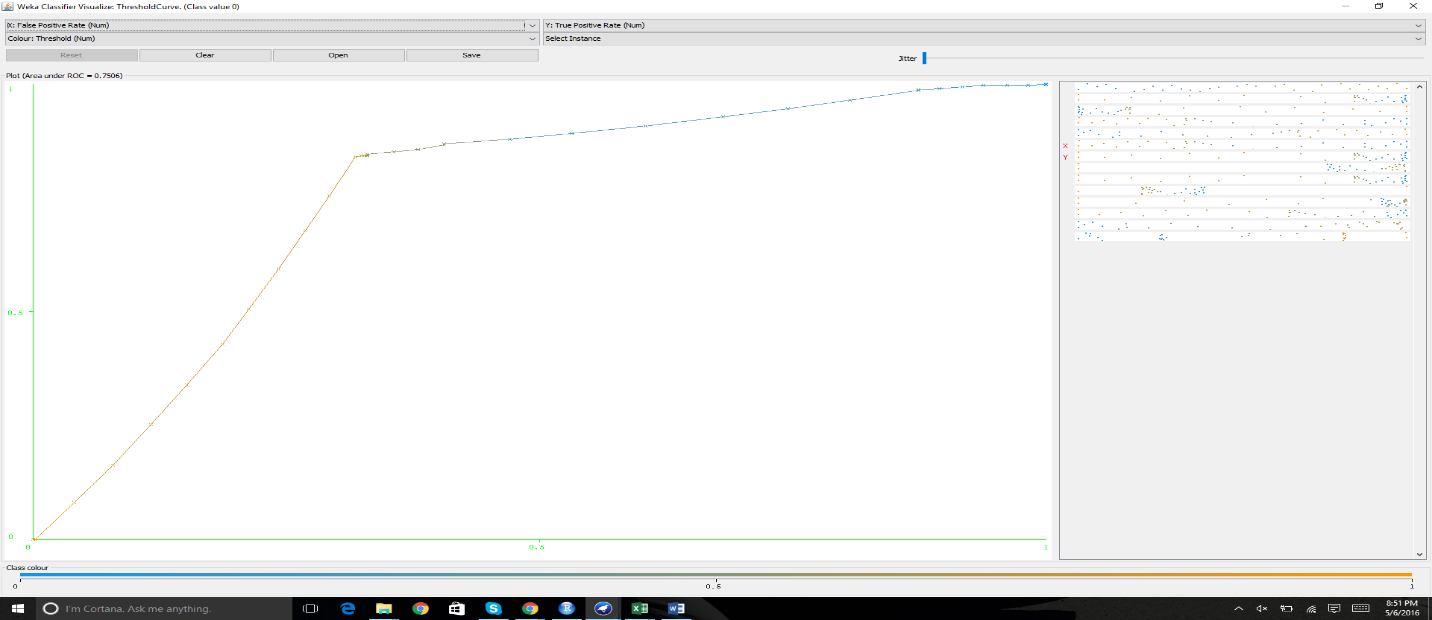
* 1. Visualize a ROC curve, and also the classification performance for 10-fold cross validation (CV).

ROC Curve: ROC stands for Receiver Operating Characteristics Curve. This curve is created by plotting the true positive vs false positive at various threshold values. When comparing the two models i.e.

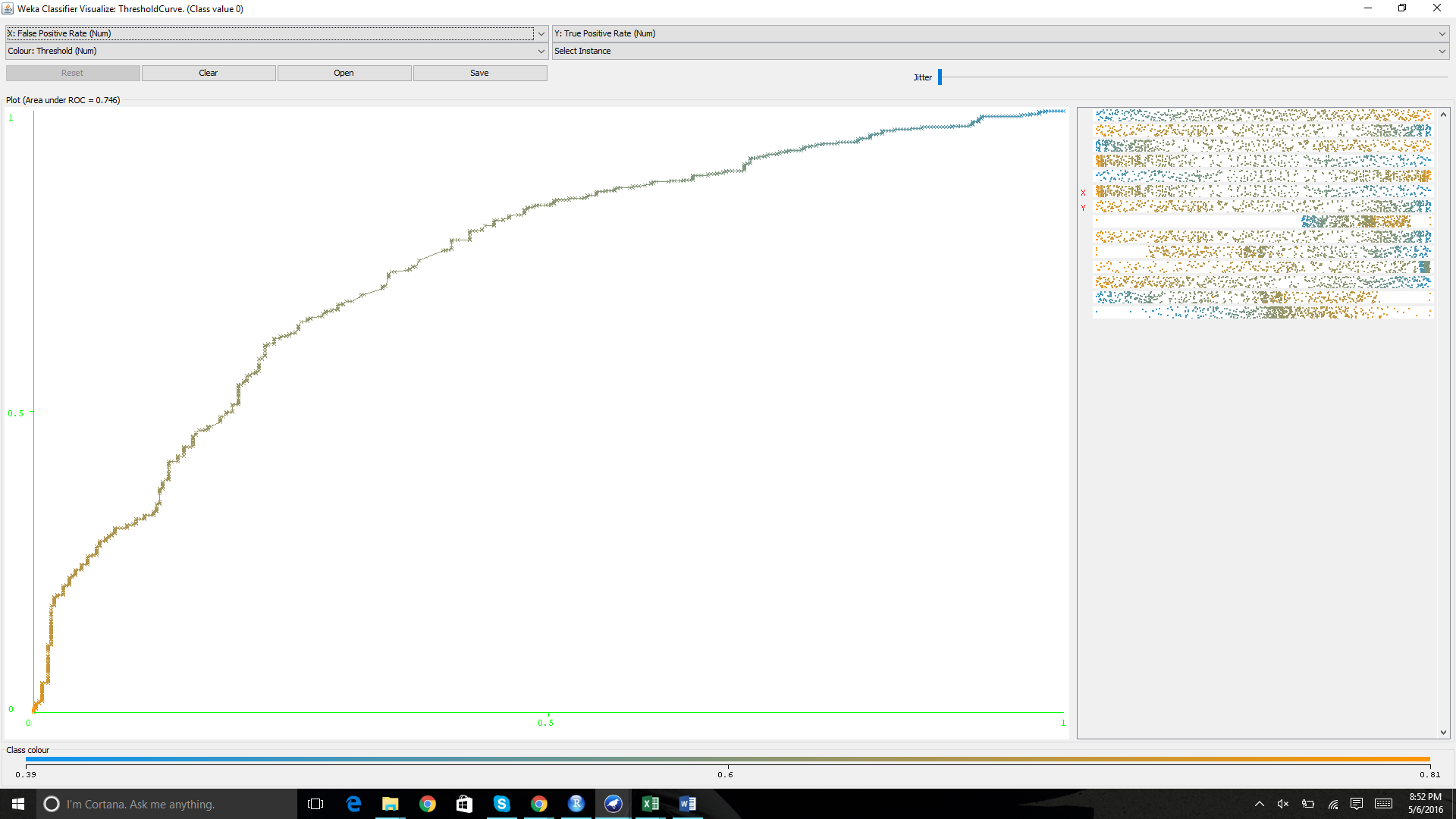
……………………………………………………………………………………………

J48 and decision tree, the model whose curve is closer to the Y axis or the curve which has higher ROC value will be the better choice

**ROC for Decision Tree**



**ROC for Random Forest**



Result: The Random Forest ROC curve performs better than the Decision Tree ROC curve.

**R codes:**

# 1.Data Acquistition and Conversion

df1<- read.csv(file.choose(), header=T)

View(df1)

str(df1)

head(df1)

tail(df1)

class(df1$Age)

class(df1)

#Removing DESCRIPTION

df1<- c[c(2)]

#2.Metadat Extraction & Imputation

## Imputing values for Age

for(i in 1:nrow(df1)){

if (is.na(df1$Age[i])== T){

df1$Age[i]=round(mean(df1$Age, na.rm =T))

}

}

mean(d1f$Age)

library(dplyr)

glimpse(df1)

summary(df1)

summary(Age)

names(df1)

attach(df1)

summary(Age)

#Missing VAlues

df1[is.na(df1$Age),]

summary(df1$Age)

# Removing Missing Values

df1<- df1[!is.na(df1$Age),]

df1

summary(df1$Age)

attach(df1)

# 3. Metadata Feature Exploration

pairs(CUSTOMERID ~ Age + Gender + SUCCESS + GUESTS + FARE + SEATCLASS, data = df1, main="Simple Scatterplot Matrix")

## OR

plot(CUSTOMERID ~ Age + Gender + SUCCESS + GUESTS + FARE + SEATCLASS, data=df1)