Table of contents:

9.	**Support Vector Machines**	
10.	**Actionable Insights**	
11.	**Trello dashboard**	
12.	**Recommendations**	

Problem Statement:

The project is aimed at analyzing the data from the dataset of customers flying within Southeast airlines and to generate actionable insights by predicting customers with low satisfaction. The real goal is to reduce churn by getting ahead of the loss (of the customer) by identifying some leading indicators, or metrics, that might help keep a customer and identify factors affecting their likelihood to recommend these airlines. We also have to suggest feedback or suggestions to improve the business and help lower the customer churn for the airlines

Scope of the Project:

We created correlational trends between various variables of the data set and customer's likelihood to recommend using the analysis models and this helped us answer the business questions formulated for the data set. On answering the business questions we could infer better insights and solutions for increasing the customer satisfaction and airline services.

In this report, we use the Association Rule Mining and Linear Modeling to analyse our data, and another model, which is Support Vector Machine to validate our result.

Business Questions:

1. •What are the attributes/factors that affect customer satisfaction ?
2. •What causes customers to give high satisfaction ?
3. •What causes customers to give low satisfaction ?
4. •What can be recommended to the airline to reduce their customer churn ?

Data munging:

- ## Importing the dataset

Importing the file into a Dataframe using JSON

```
#df <-
getURL("https://s3.us-east-1.amazonaws.com/blackboard.learn.xythos.prod/595662
1d575cd/8606160?response-content-disposition=inline%3B%20filename%2A%3D
UTF-8%27%27fall2019-survey-M09.json&response-content-type=application%2Fjs
on&X-Amz-Algorithm=AWS4-HMAC-SHA256&X-Amz-Date=20191205T023408Z&X-
Amz-SignedHeaders=host&X-Amz-Expires=21600&X-Amz-Credential=AKIAIL7WQ
YDOOHAZJGWQ%2F20191205%2Fus-east-1%2Fs3%2Faws4_request&X-Amz-Sign
ature=080af8a024ff530299823a713fcf4a4e8a827d5990bbf6048bd64945fbdc2629")
Dataset <- jsonlite::fromJSON('dataset.json')
View(Dataset)
```

· Cleaning the dataset

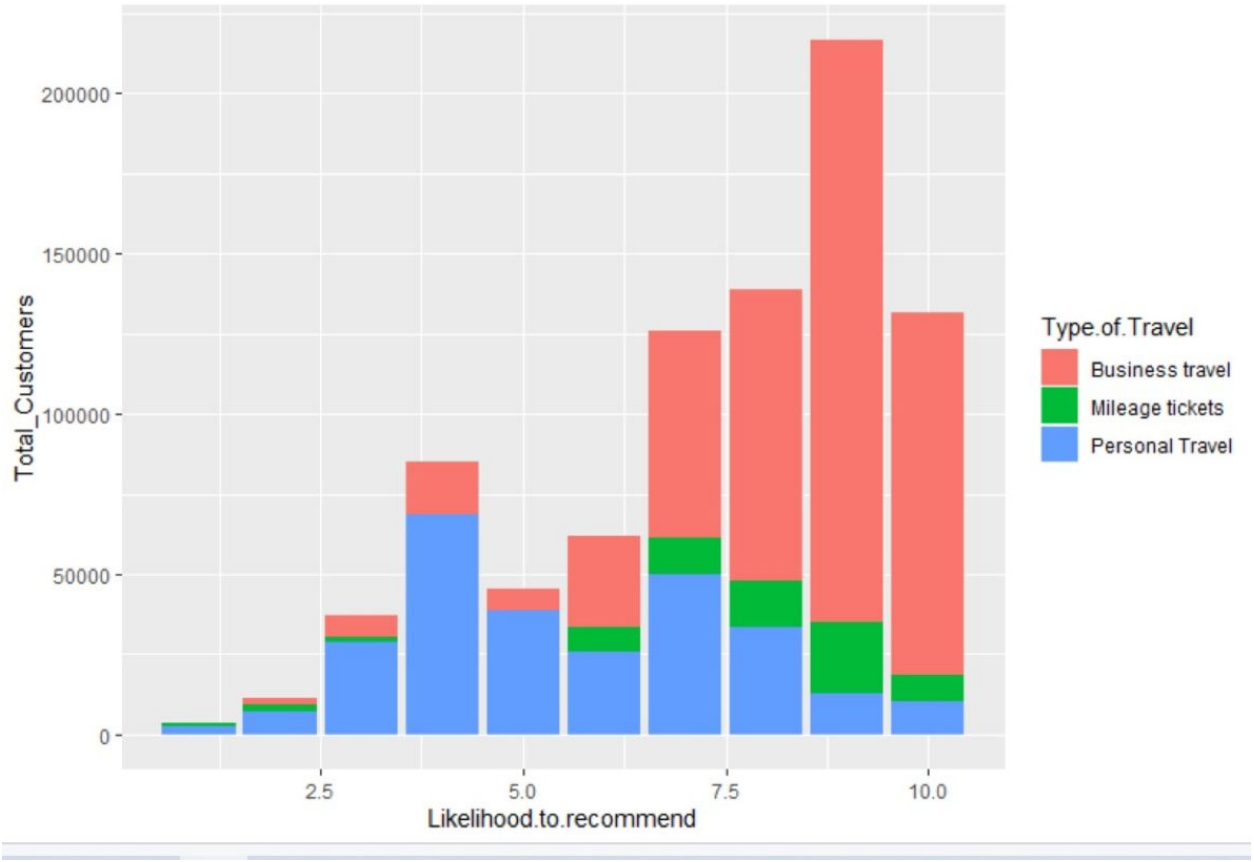
```
#Checking for NA's
is.na(Dataset)

#checking for any rows which are not complete
sum(!complete.cases(Dataset))
ncol(Dataset)
nrow(Dataset)

#Converting columns into number
Dataset$Age <- as.numeric(Dataset$Age)
Dataset$Flight.time.in.minutes <- as.numeric(Dataset$Flight.time.in.minutes)
Dataset$Day.of.Month <- as.numeric(Dataset$Day.of.Month)
Dataset$Flight.Distance <- as.numeric(Dataset$Flight.Distance)

#checking for any rows which are not complete for the new dataset
sum(!complete.cases(Dataset))
sum(is.na(Dataset$Arrival.Delay.in.Minutes))
sum(is.na(Dataset$Flight.time.in.minutes))
sum(is.na(Dataset$Departure.Delay.in.Minutes))


#replacing na values with mean values of adjacent cells
Dataset <- na.interpolation(as.numeric(Dataset),option = "linear",maxgap = Inf)

#Validating that the data is cleaned
sum(!complete.cases(Dataset))
is.na(Dataset)
is.null(Dataset)
```

View(Dataset)
nrow(Dataset)
str(Dataset)

Descriptive statistics using Data visualization:

We did visualization of attributes which we thought might have an affect a customer's likelihood to recommend SouthEast Airlines.

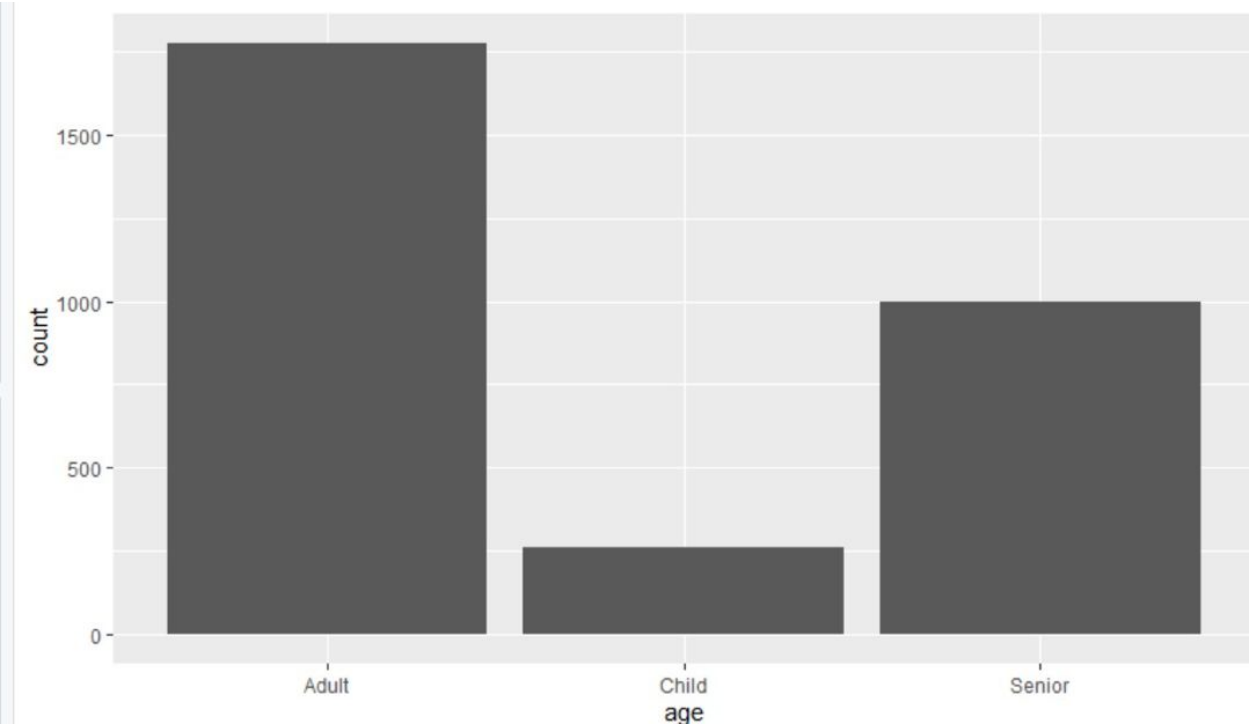
a) Customers' likelihood to recommend with type of travel.



As you can tell from the above graph, people who's type of travel is for personal reasons are less likely to recommend SouthEast airline.

```
Total_Customers <- nrow(CleanData)
travel <- ggplot(CleanData, aes(x=Likelihood.to.recommend, y=Total_Customers
,fill=Type.of.Travel)) + geom_col()
travel
```

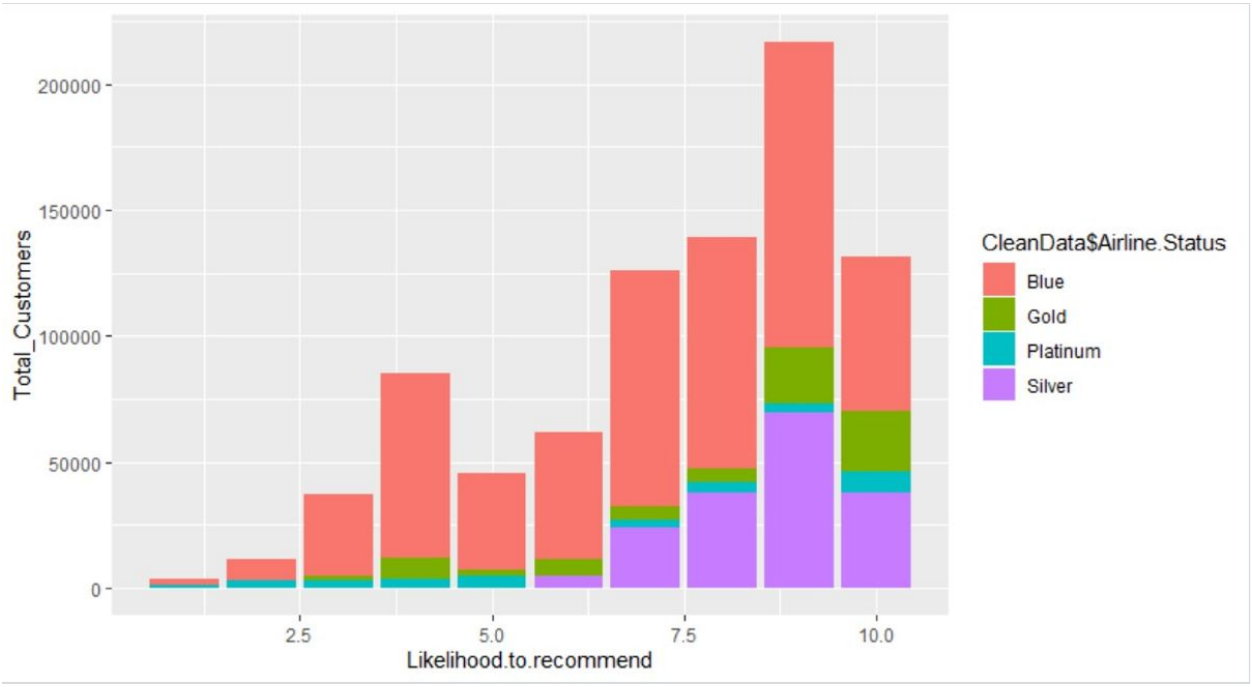
b) Customers' likelihood to recommend with Age group



The above graph clearly shows that the customers who are children or senior citizens are unhappy with the services provided by the airlines, thereby the chances of recommending SouthEast airlines is significantly low.

```
age<-agefunction(dataLowSatisfaction$Age)
ggplot(dataLowSatisfaction,aes(x=age,fill=Likelihood.to.recommend))+geom_bar
(position='dodge')
```

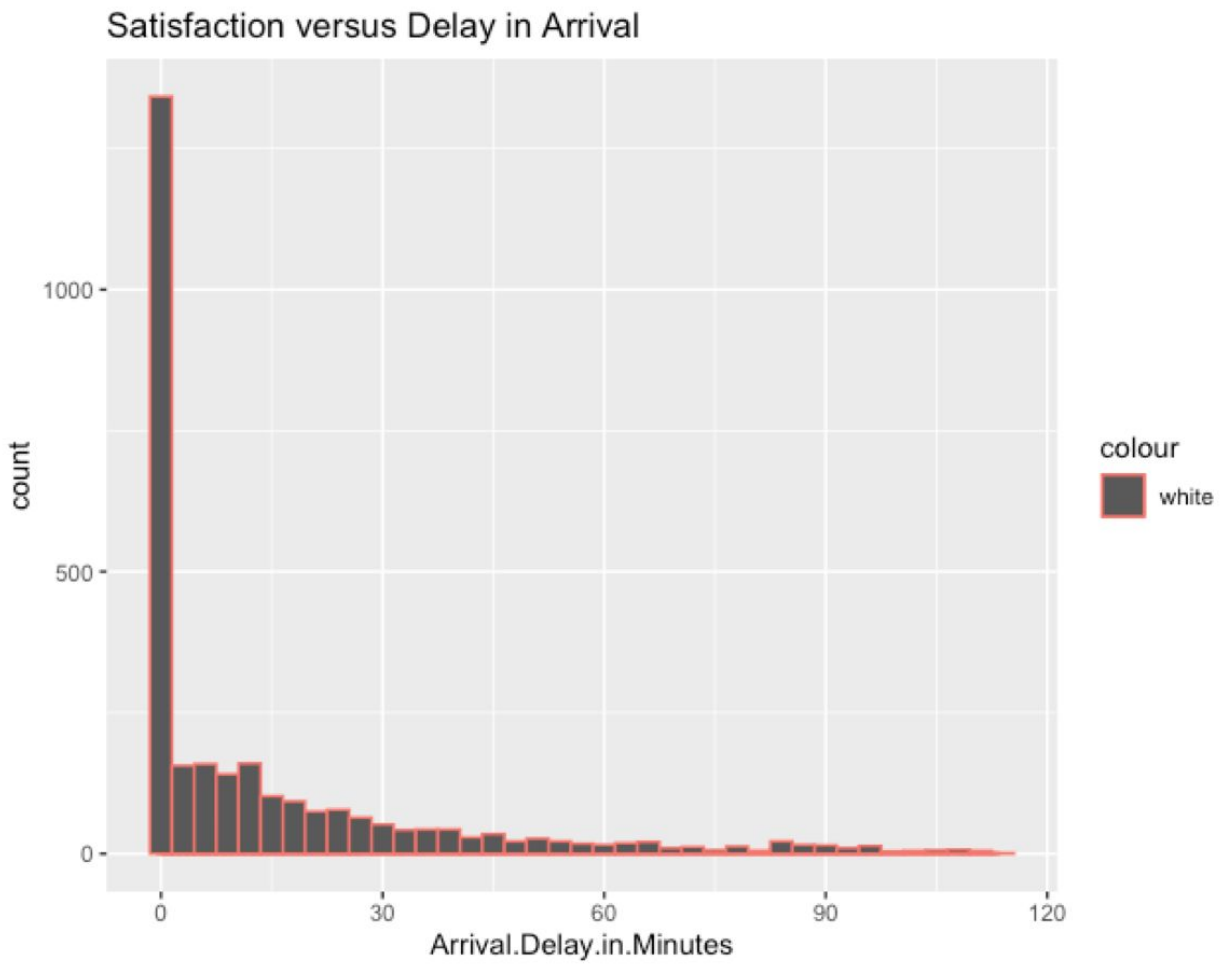
c) Customers' Likelihood to recommend with Airline Status.



From the above graph, we deduced that customers with Blue airline status are probably unhappy with the benefits provided and are hence unlikely to recommend.

```
Airline_status_satis <- ggplot(CleanData, aes(x=CleanData$Airline.Status, y=Total_Customers ,fill=Likelihood.to.recommend)) + geom_col()
Airline_status_satis
```

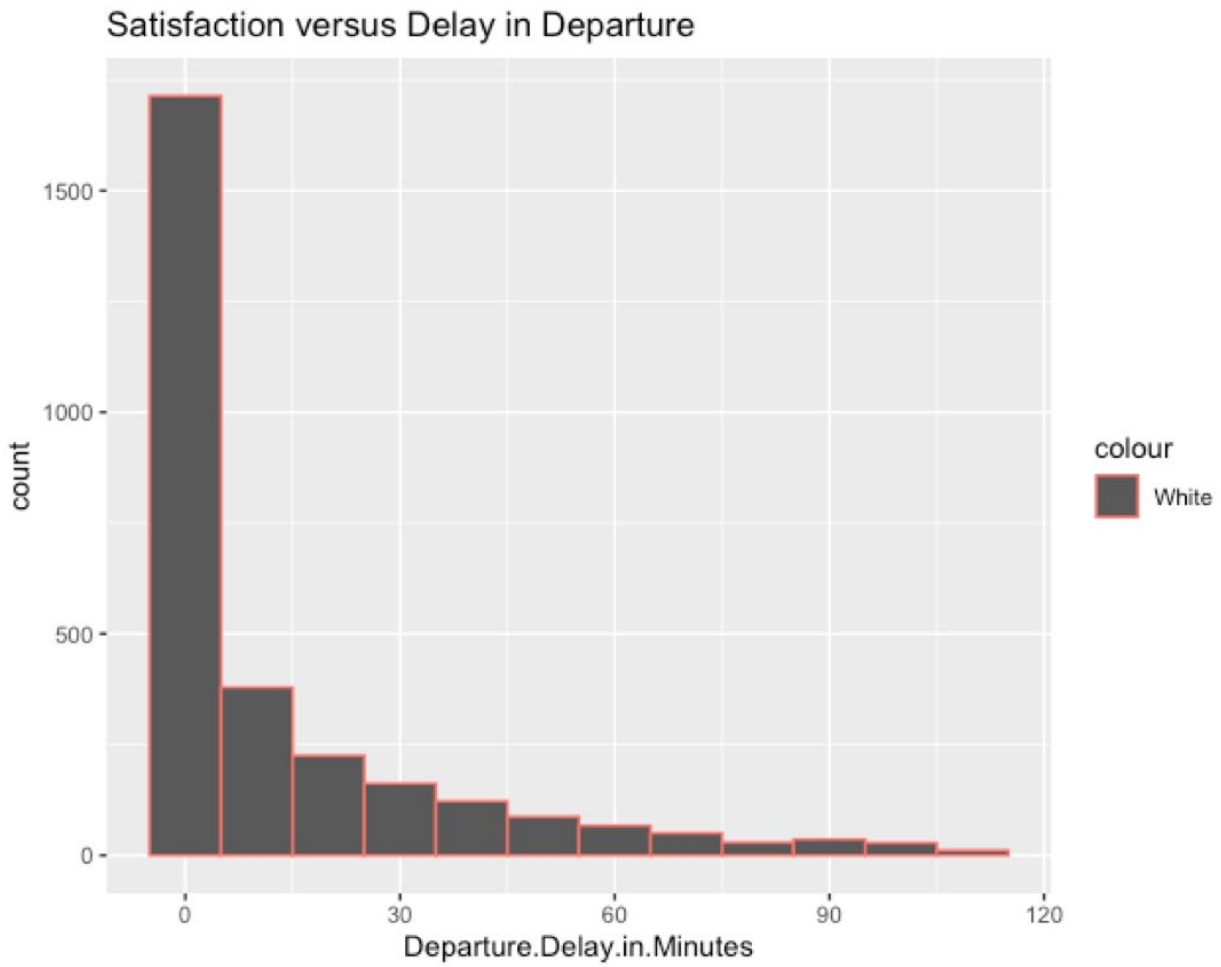
d) Customers likelihood to recommend with Delay in Arrival



The above graph shows Satisfaction vs Delay in Arrival

```
ADD <- quantile(dataLowSatisfaction$Arrival.Delay.in.Minutes,0.96)
SatisfactionDelayinArr
<-ggplot(dataLowSatisfaction[dataLowSatisfaction$Arrival.Delay.in.Minutes<ADD
,],aes(x=Arrival.Delay.in.Minutes))
SatisfactionDelayinArr <-
SatisfactionDelayinArr+geom_histogram(aes(fill=Likelihood.to.recommend,color=
"white"),binwidth = 1,position = "dodge")
SatisfactionDelayinArr <- SatisfactionDelayinArr+ggtitle("Satisfaction versus
Delay in Arrival")
SatisfactionDelayinArr
```

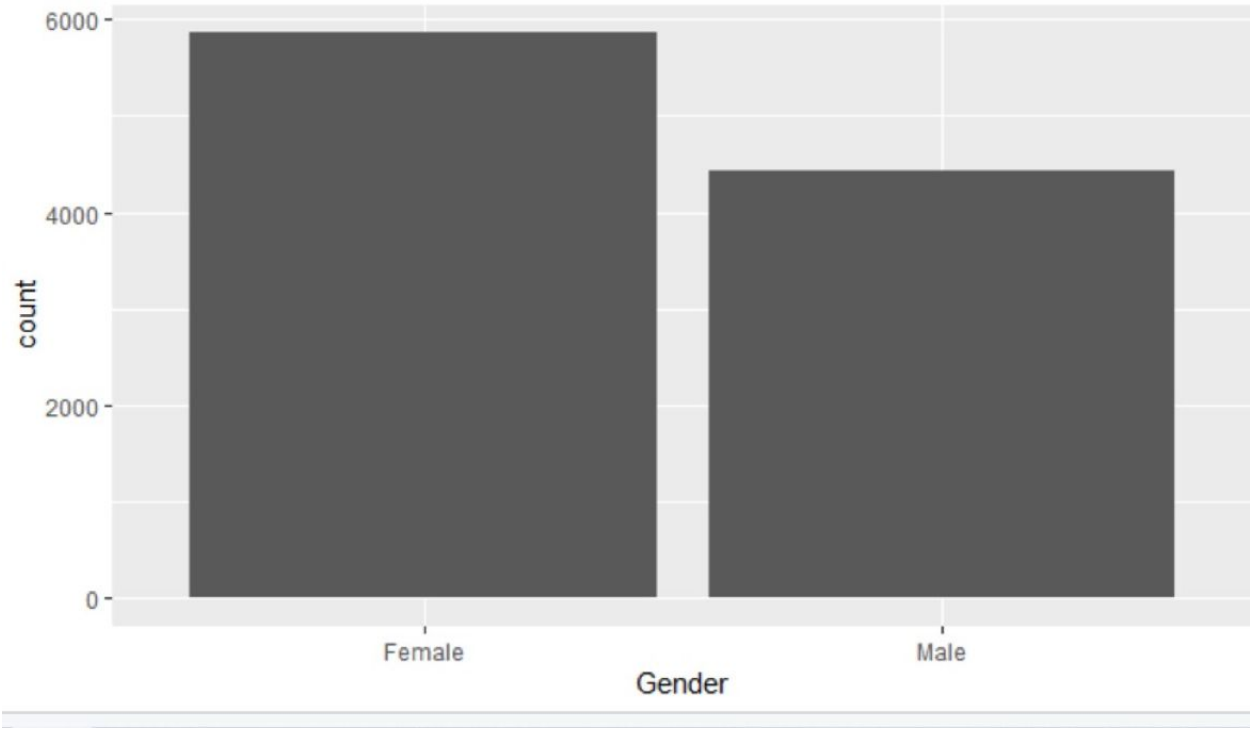
e) Customer's likelihood to recommend vs Delay in Departure



The above graph shows customers likelihood to recommend with Delay in departure.

```
SDD <- quantile(dataLowSatisfaction$Departure.Delay.in.Minutes,0.96)
SatisfactionDelayInDept <-
ggplot(dataLowSatisfaction[dataLowSatisfaction$Departure.Delay.in.Minutes<SDD,],aes(x=Departure.Delay.in.Minutes))
SatisfactionDelayInDept <-
SatisfactionDelayInDept+geom_histogram(aes(fill=Likelihood.to.recommend,color="white"),binwidth = 10,position = "dodge")
SatisfactionDelayInDept <- SatisfactionDelayInDept+ggtitle("Satisfaction versus Delay in Departure")
SatisfactionDelayInDept
```

f) Customer's likelihood to recommend based on Gender



The above graph shows that females are less likely to recommend SouthEast airlines over Male.

ggplot(Dataset,aes(x=Gender,fill=Likelihood.to.recommend))+geom_bar(position ='dodge')

Linear Modelling:

Linear models are widely used in statistical data analysis when the dependent or the response variable is quantitative, whereas the independent variables may be quantitative, qualitative, or both. It can also be used for some types of nonlinear modeling as an example given below will show.
Linear models are a way of describing a response variable in terms of a linear combination of predictor variables. The response should be a continuous variable and be at least approximately normally distributed.

Code:

```
#Linear Modelling


str(CleanData)

#Removing columns with lot of factors as they are not useful in lm. Took all the other columns as lm's job is to remove all the not important columns
lm_data <- CleanData[,c(3,4,5,6,8,9,10,11,12,13,14,15,21,22,23,25,26,27)]

str(lm_data)

# chnaging chr to factors as lm only runs on int, factors>2 and num
lm_data$Airline.Status <- as.factor(lm_data$Airline.Status)
lm_data$Class <- as.factor(lm_data$Class)
lm_data$Gender <- as.factor(lm_data$Gender)
lm_data$Type.of.Travel <- as.factor(lm_data$Type.of.Travel)
str(lm_data)

#removing rows NA - If there are no NAs in our dataset then please remove the following code
colSums(is.na(lm_data)) #check NA with this code
lm_data <- filter(lm_data, !is.na(Flight.time.in.minutes))
lm_data <- filter(lm_data, !is.na(Departure.Delay.in.Minutes))
lm_data <- filter(lm_data, !is.na(Arrival.Delay.in.Minutes))


#running 1st lm with no direction wrt likelihood to recommend
Linear.model <- lm(formula = Likelihood.to.recommend ~., data = lm_data)
library(MASS)

stepAIC(Linear.model) #this tells us the best columns to pick
summary(Linear.model)
```

```
Model_1 <- lm(formula = Likelihood.to.recommend ~ Airline.Status + Age +
                                    Type.of.Travel + Total.Freq.Flyer.Accts +
Shopping.Amount.at.Airport +
                                  Eating.and.Drinking.at.Airport + Day.of.Month +
Departure.Delay.in.Minutes +
                                Arrival.Delay.in.Minutes + Flight.time.in.minutes +
Flight.Distance,
           data = lm_data)

summary(Model_1)



#running lm with backwards
null<-lm(Likelihood.to.recommend~1,lm_data)
stepAIC(Linear.model, direction='backward')

back <-lm(formula = Likelihood.to.recommend ~ Airline.Status + Age +
                                      Type.of.Travel + Total.Freq.Flyer.Accts +
Shopping.Amount.at.Airport +
                                  Eating.and.Drinking.at.Airport + Day.of.Month +
Departure.Delay.in.Minutes +
         Arrival.Delay.in.Minutes + Flight.time.in.minutes + Flight.Distance,
        data = lm_data)

summary(back)

#running lm forward model
stepAIC(null,direction='forward',scope=list(upper=Linear.model,lower=null))

forward <- lm(formula = Likelihood.to.recommend ~ Type.of.Travel +
Airline.Status +
            Arrival.Delay.in.Minutes + Departure.Delay.in.Minutes + Age +
```

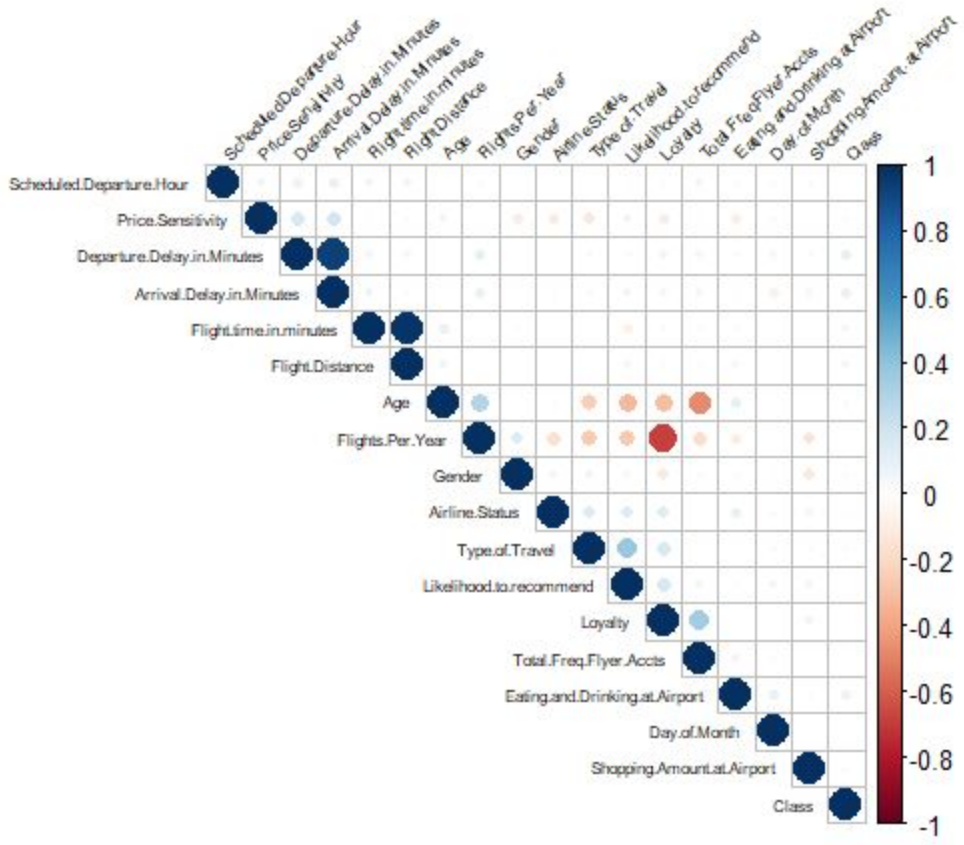
```
                                    Day.of.Month + Total.Freq.Flyer.Accts +
Eating.and.Drinking.at.Airport,
          data = lm_data)

summary(forward)


Model_common <- lm(formula = Likelihood.to.recommend ~  Airline.Status
+ Age + Type.of.Travel +
                     Total.Freq.Flyer.Accts +  Eating.and.Drinking.at.Airport +
Day.of.Month +
                Departure.Delay.in.Minutes + Arrival.Delay.in.Minutes, data =
lm_data)




summary(Model_common)

#Corelatin Model

lm_data <- CleanData[,c(3,4,5,6,8,9,10,11,12,13,14,15,21,22,23,25,26,27)]

colSums(is.na(lm_data)) #check NA with this code
lm_data <- filter(lm_data, !is.na(Flight.time.in.minutes))
lm_data <- filter(lm_data, !is.na(Departure.Delay.in.Minutes))
lm_data <- filter(lm_data, !is.na(Arrival.Delay.in.Minutes))

str(CleanData)

str(lm_data)

cor_df <- lm_data
```

```
# converting categorical variables into dummy variables
cor_df$Airline.Status[cor_df$Airline.Status == 'Blue'] <- 1
cor_df$Airline.Status[cor_df$Airline.Status == 'Silver'] <- 2
cor_df$Airline.Status[cor_df$Airline.Status == 'Gold'] <- 3
cor_df$Airline.Status[cor_df$Airline.Status == 'Platinum'] <- 4

cor_df$Gender[cor_df$Gender == "Female"] <- 0
cor_df$Gender[cor_df$Gender == "Male"] <- 1

cor_df$Type.of.Travel[cor_df$Type.of.Travel == 'Mileage tickets'] <- 1
cor_df$Type.of.Travel[cor_df$Type.of.Travel == 'Personal Travel'] <- 2
cor_df$Type.of.Travel[cor_df$Type.of.Travel == 'Business travel'] <- 3

cor_df$Class[cor_df$Class == 'Eco'] <- 1
cor_df$Class[cor_df$Class == 'Eco Plus'] <- 2
cor_df$Class[cor_df$Class == 'Business'] <- 3
#cor_df <- as.matrix(cor_df)

#Converting dummy variables into numeric format
cor_df$Airline.Status <- as.numeric(cor_df$Airline.Status)
cor_df$Gender <- as.numeric(cor_df$Gender)
cor_df$Type.of.Travel <- as.numeric(cor_df$Type.of.Travel)
cor_df$Class <- as.numeric(cor_df$Class)

#Creating a correlation matrix
str(cor_df)
cor_model <- cor(cor_df)
cor_model

install.packages("corrplot")
library(corrplot)

#Visualizing correlation matrix
```

corrplot(cor_model, type = "upper", order = "hclust", tl.col = "black", tl.srt = 45, tl.cex = 0.5)

str(cor_df)

Co relation Matrix:



Association Rules:

Association Rules Mining:

Association rule mining is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. After Linear Modeling, we applied the association rules to the dataframe which had clean data of Southeast Airlines. The variables used in linear modeling were also used for Association Rules Mining.

We created functions to use in the mining. The association rules cannot have the numeric or integer variables as an input and thus by creating the functions we could run the mining rules. The variables we chose are: likelihood.to.recommend, Price.Sensitivity, Loyalty, Age, Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Flights.Per.Year, Shopping.Amount.at.Airport, Eating.and.Drinking.at.Airport, Scheduled.Departure.Hour, Flight.Distance,

Airline.Status, Gender, Type.of.Travel, Flight.cancelled, Class

The functions were then created independently depending on the scale of measure. The likelihood to recommend which is our independent variable, has three ranges. The scale is from 1-10. If the value is greater than 8, the customer is a "Promoter". If the value is less than 7, then the customer is a "Detractor" which means they are less likely to promote the airline. And "passive" for customer with other range of likelihood to recommend.

For the price sensitivity, the range is from 1-5. The more the range, greater is the customer's sensitivity towards the price of the ticket. The function for that was:

The loyalty index was given. The range was from -1 to +1. We created a function for the same.

If the customer had the value 1, we deemed them as old and loyal customer. If the value was between 0 to 1, their loyalty was lesser and were named as Recent customer. New customers or less loyal customers were with loyalty index less than 1.

Then we used for Airline Delays including Departure delay and Arrival delay, to understand the impact of the variables on the satisfaction or likelihood to recommend of a customer. The function we created was AirlinesDelay. The range used was high low and average, depending on the value v(delay).

The age of the customers was divided into three levels: teen, adults, senior citizens.

For other numeric variables the quartiles were considered. Values less than 40% quartiles were defined as “Low”, values more than 60% quartiles were defined as “High”, others were defined as “Average”. The function named was “Airlinesother”

Then we created the variables using the above function for the numeric variables. The screenshot for them is below.

The mining rules now needed to be applied to these newly created variables. Before that, we put the variables into a new dataframe named “data_flight_categorized”

Now applying the association mining rules, we decided to use minimum level of “support” to 0.5

This provided with 11 results greater than 0.5 value of item frequency. These were the main variables that we will be getting further.

Then we proceeded to divide the results into two sections. We focused on customers with high satisfaction and the important variables affecting it and customers with lower satisfaction (likelihood to recommend) and the important variables affecting it.

```
ruleset_satisfied <-
apriori(data_flight_categorizedX, parameter =
list(support =0.3, confidence = 0.2), appearance
=
list(default="lhs",rhs=("custsatis=Promoters")))
```

From these the top 10 rules for Promoters were inspected using the inspect and head function.

The insights from the association mining rules for promoters are:

- **Adults are likely to be the promoters**
- **The customer's type of travel for Business reasons are likely to be promoters**
- **Also, Economy class travelers are likely to be promoters**

Similarly, the association rules were applied to customer who were detractors, or lesser likelihood to recommend.

```
ruleset_unsatisfied <- apriori(data_flight_categorizedX, parameter = list(support =0.1, confidence = 0.4), appearance =
list(default="lhs",rhs=("custsatis=Detractors")))
```

The scatter plot for 47 rules

The lift was set to get lower number of rules as to get the most accurate variables to use.

We inspected 10 results using the head and inspect function for detractors.

The insights we have from the Detractors are:

- **The number of people who shop less are likely to be detractors**
- **The customers with loyalty less than 0 are likely to be detractors**
- **Gender as female comes out as one of the variables for detractors**
- **The Airline Status of the customers most likely to be detractors is Blue**
- **The type of travel of the customers likely to be detractors is for Personal Reason**

Association Rules Mining:

Association rule mining is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. After Linear Modeling, we applied the association rules to the dataframe which had clean data of Southeast Airlines. The variables used in linear modeling were also used for Association Rules Mining.

We created functions to use in the mining. The association rules cannot have the numeric or integer variables as an input and thus by creating the functions we could run the mining rules. The variables we chose are: likelihood.to.recommend, Price.Sensitivity, Loyalty, Age, Departure.Delay.in.Minutes, Arrival.Delay.in.Minutes, Flights.Per.Year, Shopping.Amount.at.Airport,

Eating.and.Drinking.at.Airport, Scheduled.Departure.Hour, Flight.Distance, Airline.Status, Gender, Type.of.Travel, Flight.cancelled, Class

The functions were then created independently depending on the scale of measure. The likelihood to recommend which is our independent variable, has three ranges. The scale is from 1-10. If the value is greater than 8, the customer is a "Promoter". If the value is less than 7, then the customer is a "Detractor" which means they are less likely to promote the airline. And "passive" for customer with other range of likelihood to recommend.

For the price sensitivity, the range is from 1-5. The more the range, greater is the customer's sensitivity towards the price of the ticket. The function for that was:

The loyalty index was given. The range was from -1 to +1. We created a function for the same.

If the customer had the value 1, we deemed them as old and loyal customer. If the value was between 0 to 1, their loyalty was lesser and were named as Recent customer. New customers or less loyal customers were with loyalty index less than 1.

Then we used for Airline Delays including Departure delay and Arrival delay, to understand the impact of the variables on the satisfaction or likelihood to recommend of a customer. The function we created was AirlinesDelay. The range used was high low and average, depending on the value v(delay).

The age of the customers was divided into three levels: teen, adults, senior citizens.

For other numeric variables the quartiles were considered. Values less than 40% quartiles were defined as “Low”, values more than 60% quartiles were defined as “High”, others were defined as “Average”. The function named was “Airlinesother”

Then we created the variables using the above function for the numeric variables. The screenshot for them is below.

The mining rules now needed to be applied to these newly created variables. Before that, we put the variables into a new dataframe named “data_flight_categorized”

Now applying the association mining rules, we decided to use minimum level of “support” to 0.5

This provided with 11 results greater than 0.5 value of item frequency. These were the main variables that we will be getting further.

Then we proceeded to divide the results into two sections. We focused on customers with high satisfaction and the important variables affecting it and customers with lower satisfaction (likelihood to recommend) and the important variables affecting it.

```
ruleset_satisfied <-
apriori(data_flight_categorizedX, parameter =
list(support =0.3, confidence = 0.2), appearance
=
list(default="lhs",rhs=("custsatis=Promoters")))
```

From these the top 10 rules for Promoters were inspected using the inspect and head function.

The insights from the association mining rules for promoters are:

- **Adults are likely to be the promoters**
- **The customer's type of travel for Business reasons are likely to be promoters**
- **Also, Economy class travelers are likely to be promoters**

Similarly, the association rules were applied to customer who were detractors, or lesser likelihood to recommend.

```
ruleset_unsatisfied <- apriori(data_flight_categorizedX, parameter = list(support =0.1, confidence = 0.4), appearance = list(default="lhs",rhs=("custsatis=Detractors")))
```

The scatter plot for 47 rules

The lift was set to get lower number of rules as to get the most accurate variables to use.

We inspected 10 results using the head and inspect function for detractors.

Support Vector Machine:

Actionable insights:

Insights from Linear Model:

The linear model portrayed the dependency of the customer satisfaction on variables:

Airline Status

Age

Type of travel

Total frequency flyer accounts

Eating and drinking at airport

Day of month

Departure delay in minutes

Arrival delay in minutes

The strength of the relation between likelihood to recommend and the above factors is 41%

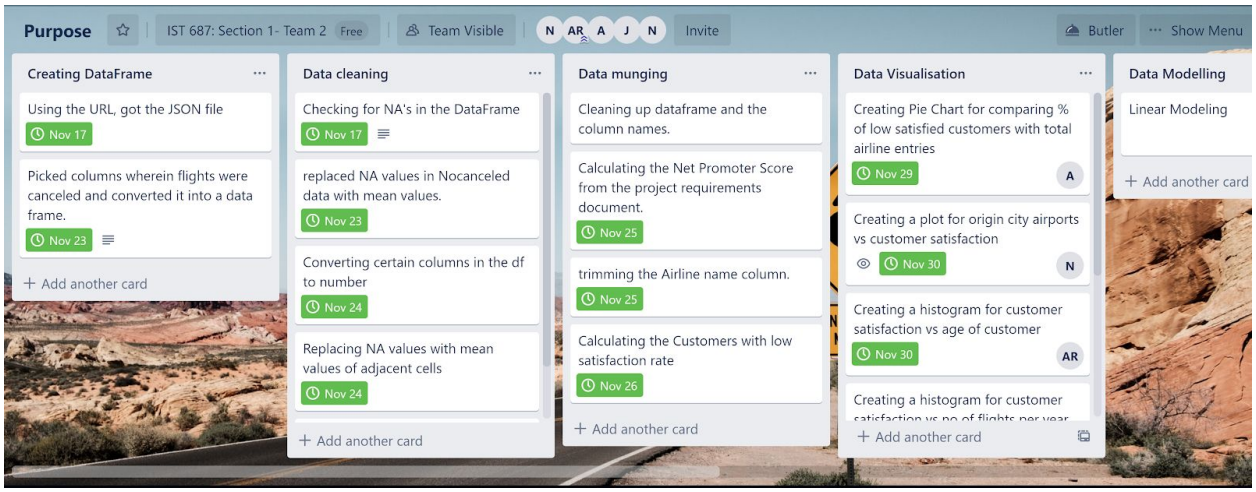
Insights from Association Rules:

The insights we have from the Detractors are:

- **The number of people who shop less are likely to be detractors**
- **The customers with loyalty less than 0 are likely to be detractors**
- **Gender as female comes out as one of the variables for detractors**
- **The Airline Status of the customers most likely to be detractors is Blue**
- **The type of travel of the customers likely to be detractors is for Personal Reasons**

Insights from Support Vector Machine:

Trello:



Recommendations:

- The customers in blue airline status should be given more benefits and opportunities.
- Better in-flight services for kids and senior citizens.
- Provide free lounge access or stay over facilities to the customers who have delayed flights.
- Co-passenger preference to female customers travelling alone.
- Sending alerts or text messages regarding any delay in arrival to the customer's emergency contacts or family members.