PROJECT REPORT

Title: Airline Passenger Satisfaction

Authors: Akhil Chitreddy, Yashwanth Pothireddy, Chamanthi Aki,
Aishwarya Kurnutala.

Summary:

- Due to COVID-19 one of the hardest hit industries is the airline industry, but as soon as we get past this pandemic, there will be a surge in the demand for the airline industry.
- Our project helps to give airlines a competitive edge when the expected surge arrives by building a predictive model based on the previous surveys data.
- Our aim of the project is to predict whether the airline passengers are satisfied or not with the airline services.
- The dataset contains satisfaction levels of customers in various departments such as cleanliness, legroom, seat comfort, etc. The dataset also consists of continuous variables such as departure delay, arrival delay, and flight distance.
- We will be pre-processing, visualizing our data to identify any relation between predictor and response variables and choose the predictors accordingly. We then build a prediction model, where the above attributes are given as input and the output would be “satisfied” or “dissatisfied” or “neutral”.
- Since the data is categorical and most of the variables are not continuous, we can use random forest classification which is an ensemble model built on decision trees. We will finally use classification metrics such as sensitivity, specificity, accuracy to judge our model performance and tweak the model’s parameters as necessary to improve the model’s performance.

Methods:

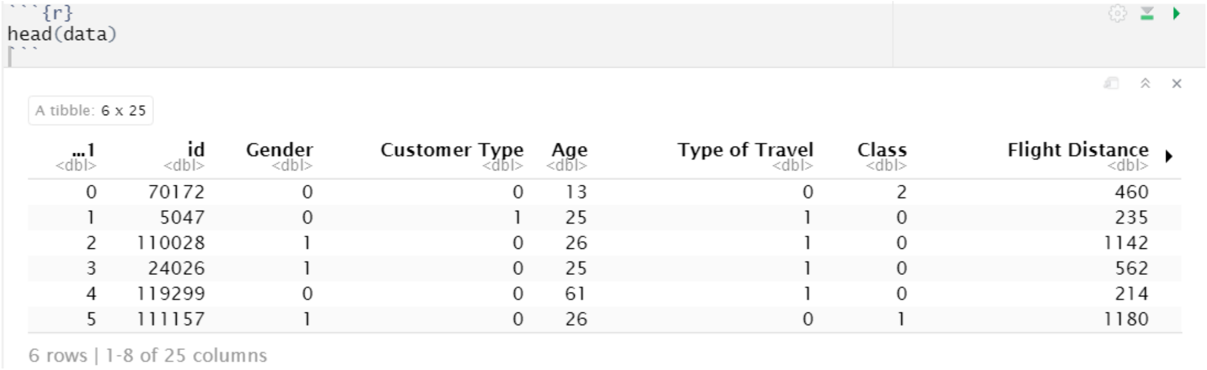
Exploratory Data Analysis:

- Our dataset has 100 thousand rows and 25 columns. The attributes are collected from a satisfaction survey which give us details about the flight and satisfaction levels of customers in various departments.

- We would use visualizations such as box plots, scatter plots to explore the data and identify any potential relationship between the variables.

Data pre-processing:

- We start by dropping the irrelevant attributes such as ID.
- We check the data for the null values. In our data we find no missing data.
- As few attributes are of type character, we need encode them into numeric values by mapping the numeric values to the string values.
- We label encode the response variable also as it needs to be passed into the select-k-best method using chi square test to check the correlation between the response variable and the other attributes.
- If all the data lies on the same scale, then we could proceed for feature selection. The data after label encoding does not lie on the same scale. So, we use standard scaling to get the data on the same scale.



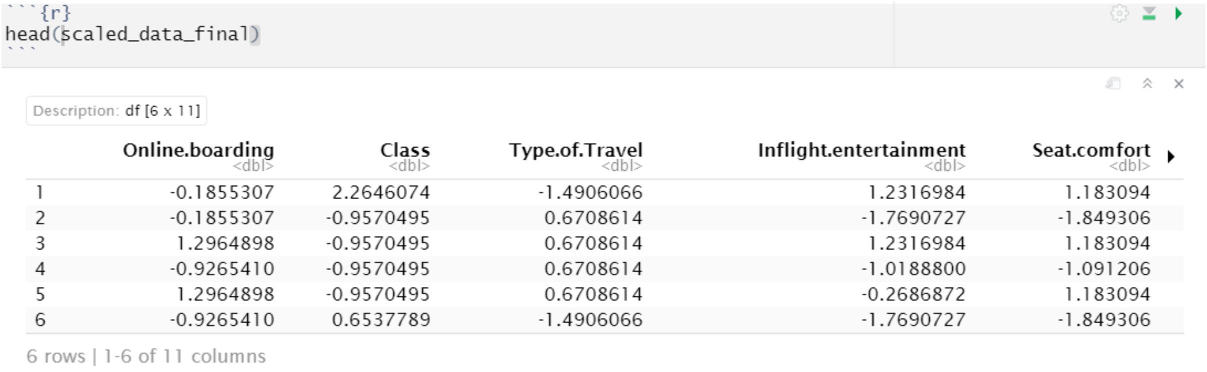
~~~
```{r}
head(data)
```
~~~

A tibble: 6 x 25

...1 <dbl>	id <dbl>	Gender <dbl>	Customer Type <dbl>	Age <dbl>	Type of Travel <dbl>	Class <dbl>	Flight Distance <dbl>
0	70172	0	0	13	0	2	460
1	5047	0	1	25	1	0	235
2	110028	1	0	26	1	0	1142
3	24026	1	0	25	1	0	562
4	119299	0	0	61	1	0	214
5	111157	1	0	26	0	1	1180

6 rows | 1-8 of 25 columns

Data before pre-processing



~~~
```{r}
head(scaled_data_final)
```
~~~

Description: df [6 x 11]

	Online.boarding <dbl>	Class <dbl>	Type.of.Travel <dbl>	Inflight.entertainment <dbl>	Seat.comfort <dbl>
1	-0.1855307	2.2646074	-1.4906066	1.2316984	1.183094
2	-0.1855307	-0.9570495	0.6708614	-1.7690727	-1.849306
3	1.2964898	-0.9570495	0.6708614	1.2316984	1.183094
4	-0.9265410	-0.9570495	0.6708614	-1.0188800	-1.091206
5	1.2964898	-0.9570495	0.6708614	-0.2686872	1.183094
6	-0.9265410	0.6537789	-1.4906066	-1.7690727	-1.849306

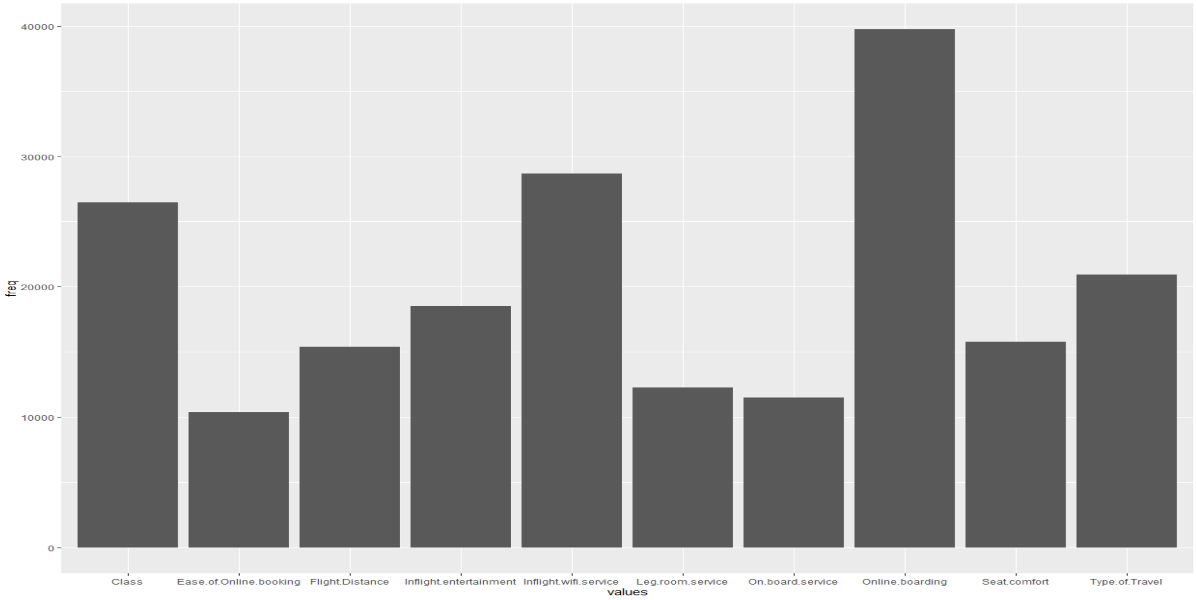
6 rows | 1-6 of 11 columns

Data after pre-processing

Feature Selection:

- Feature selection is used to eliminate the attributes which bring down the model's performance or which do not contribute to the model's performance.

- Training the model on too many attributes would also bring down the model's performance. So, to optimize, we use feature selection methods such as select-k-best using chi square test.
- We choose the top 10 attributes in the data for passing into the model. We chose only 10 attributes as they had a comparatively higher feature importance score among the 24 attributes.

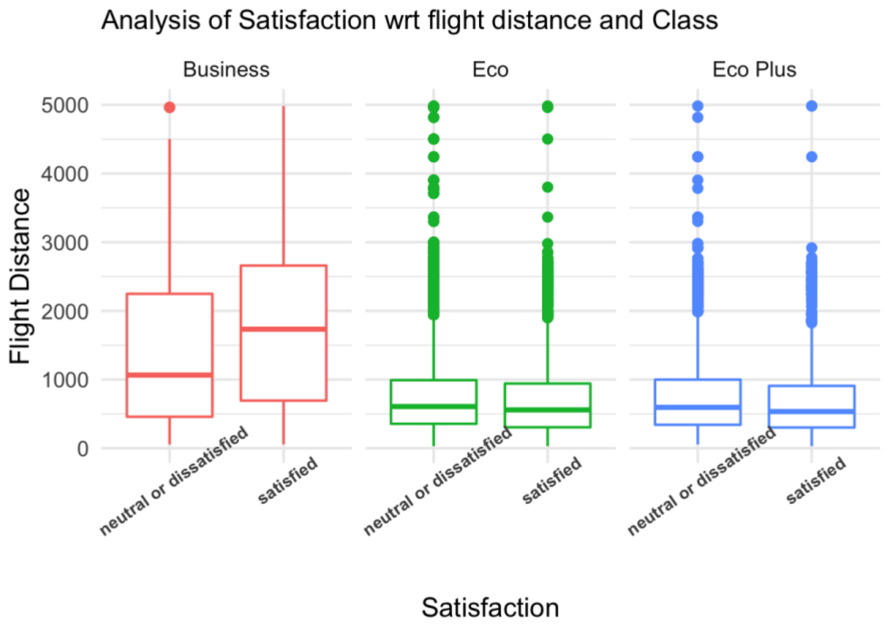


Building Predictive Model:

- In our data we need to predict the satisfaction level response variable which is a categorical variable.
- As there are only two levels in the satisfaction attribute which are "satisfied" and "neutral or dissatisfied", we could use models such as logistic regression, support vector machine and random forest.
- We divided the data into train and test datasets in the ratio of 75% to 25%.
- We use the train data for training the machine learning models. We use the test dataset to predict the satisfaction response variable.
- We then use the metrics accuracy, sensitivity, specificity etc. for each model and choose one of the models as final model.

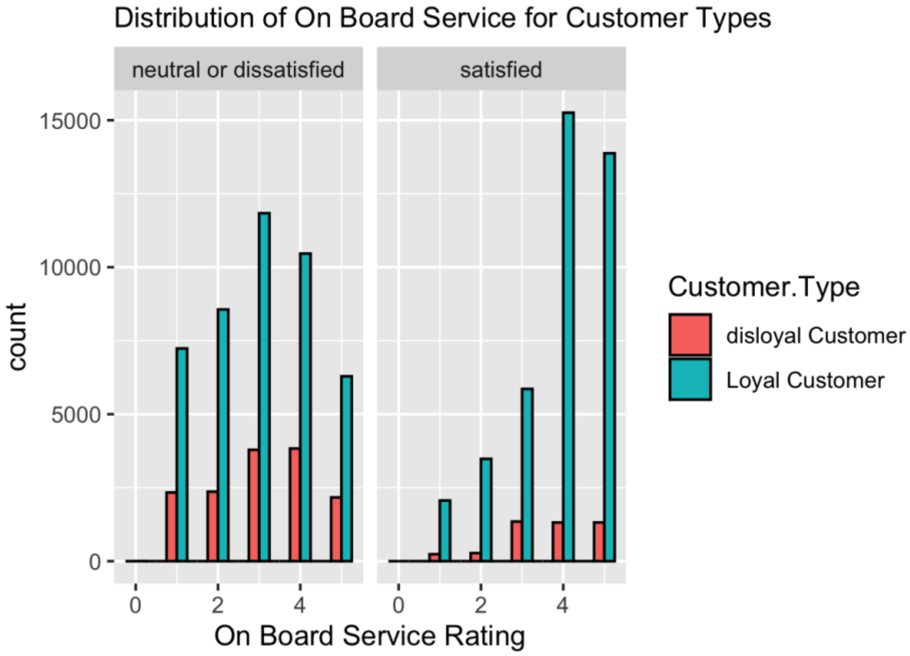
Results:

Some of the graphs that we got from our EDA analysis are:



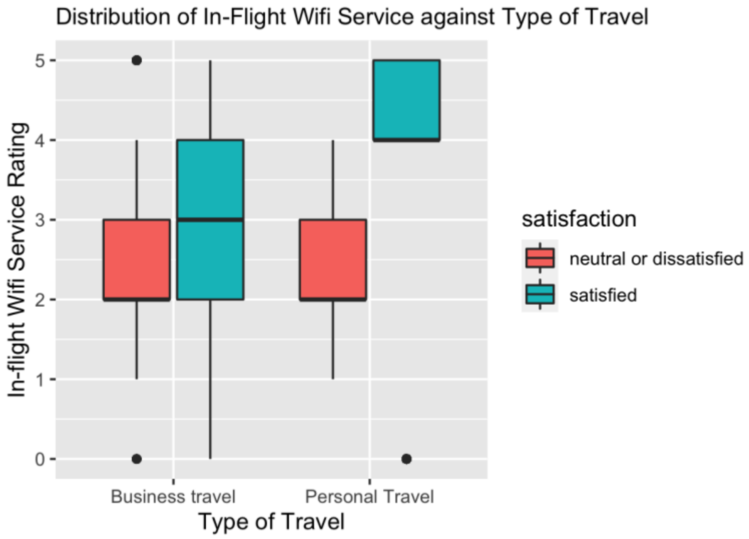
Observations:

- The average flight distance travelled by satisfied business class passengers is the longest.
- The average flight distance travelled by satisfied passengers in the Eco and Eco plus classes is nearly same.
- As a result, more individuals are satisfied after flying long distances in business class.



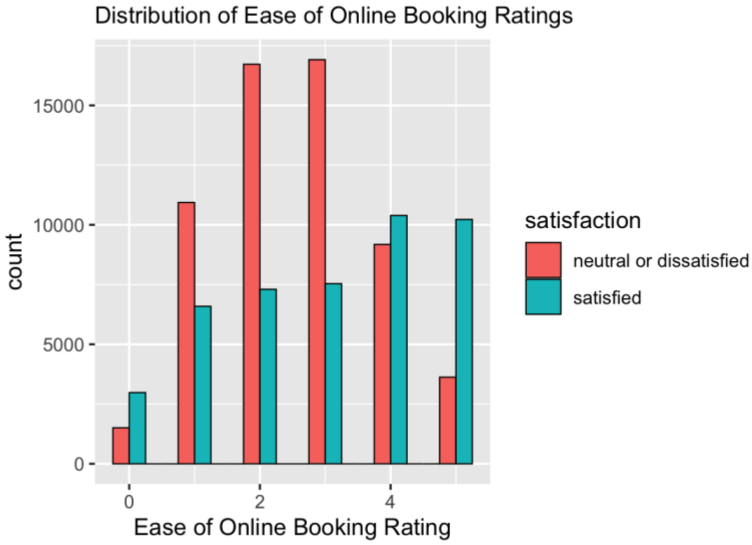
Observations:

- Customers who are loyal have provided more ratings than those who are disloyal.
- On Board Service has received higher ratings from loyal customers who are satisfied.
- The ratings of neutral or dissatisfied passengers form a bell curve.
- As a result, more individuals are satisfied with the On-Board Service.



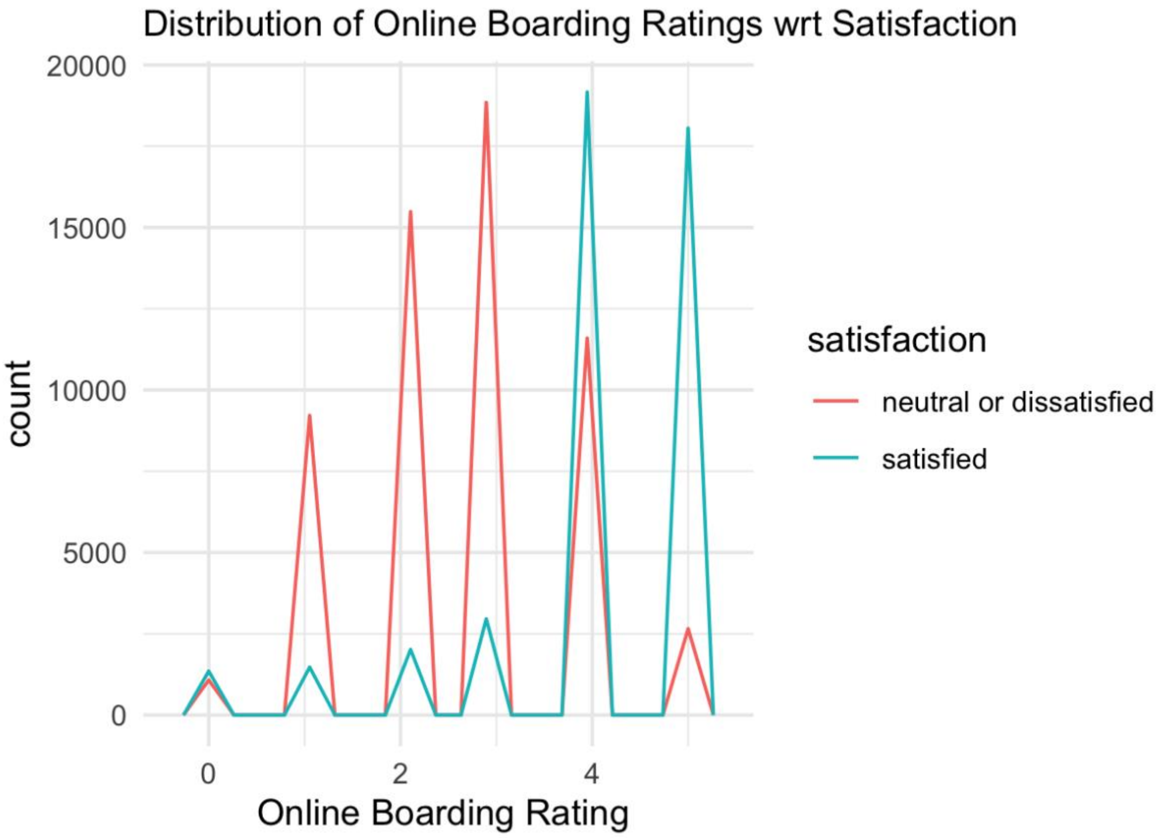
Observations:

- In-Flight Wi-Fi Service has received higher ratings from satisfied passengers for both business and personal travel.
- As a result, more people are happy with the in-flight Wi-Fi service.



Observations:

- More passengers rated that booking online was difficult for them.
- This implies that more customers who were dissatisfied with the service gave it a low rating.
- As a result, more people are unhappy with Online Booking.



Observations:

- The ratings of dissatisfied or neutral passengers form a bell curve, indicating that majority of dissatisfied passengers gave average ratings for Online Boarding.
- Passengers who were dissatisfied gave a greater number of ratings than passengers who were satisfied.
- As a result, more people are unhappy with Online Boarding.

Model Metrics:

- Logistic Regression:
  After training the model, the metrics we received are: -

Confusion Matrix:

	Actual dissatisfied	**Actual satisfied**
Predicted Dissatisfied	12850	2066
Predicted satisfied	1723	9337

Accuracy	85.4
Sensitivity	88.1
Specificity	81.8
F1-Score	83.1

- Support Vector Machine:
  After training the model, the metrics we received are: -

Confusion Matrix:

	Actual dissatisfied	**Actual satisfied**
Predicted Dissatisfied	9209	1541
Predicted satisfied	2194	13032

Accuracy	85.6
Sensitivity	80.7
Specificity	89.4
F1-Score	87.4

- Random Forest:
  After training the model, the metrics we received are: -

Confusion Matrix:

	Actual dissatisfied	**Actual satisfied**
Predicted Dissatisfied	14011	861
Predicted satisfied	562	10542

Accuracy	94.5
Sensitivity	96.1
Specificity	92.4
F1-Score	95.1

Among the three models, the best performance in terms of accuracy, sensitivity, specificity and F1-score was given by the random forest classifier with an accuracy of 94.5%.

Discussion:

- By predicting the satisfaction level of the passenger, airline service provider could find the departments where they need to focus in improving the passenger satisfaction levels.
- By looking at the feature importance scores airline service provider can focus on developing a software for the online boarding process of the passengers.
- They could also improve their inflight Wi-Fi service which has a good correlation with the response variable.
- After improving their service in a specific department, the prediction system will be of a great use to find out the satisfaction level by inputting the previous data from the passengers.
- For the future work, we can consider collecting data by introducing additional categories in the response variable such as extremely dissatisfied, extremely satisfied, slightly dissatisfied, slightly satisfied etc.
- By having these additional categories, we can categorize customers further and use advanced machine learning models such as XGboost, KNN (k nearest neighbors) in predicting the satisfaction levels.

Statement of Contributions:

Akhil Chitreddy – Performed Feature Selection, worked on logistic regression, random forest models.

Yashwanth Pothireddy – Performed Feature Selection, worked on support vector machine model.

Chamanthi Aki – Performed EDA and preprocessing of the data.

Aishwarya Kurnutala – Performed EDA and preprocessing of the data

References:

- https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction
- https://www.rdocumentation.org/packages/e1071/versions/1.7-9/topics/svm
- https://www.rdocumentation.org/packages/randomForest/versions/4.6-14/topics/randomForest
- https://www.rdocumentation.org/packages/FSinR/versions/1.0.8/topics/selectKBest
- https://www.rdocumentation.org/packages/spatialEco/versions/1.3-7/topics/logistic.regression

Appendix:

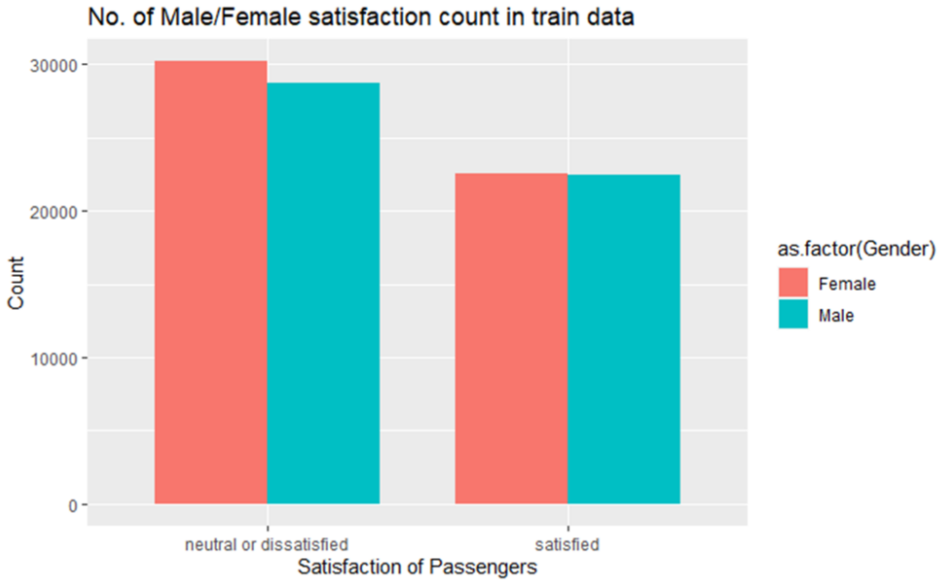
Code-

```
#Import libraries
library(tidyverse)
library(readr)
library(ggplot2)

#Import Dataset
train<-read.csv("C:/Users/Chams/Downloads/idmp/train.csv")
test<-read.csv("C:/Users/Chams/Downloads/idmp/test.csv")
#Printing the first few lines of the dataset
columns<c("Online.boarding","Class","Type.of.Travel","Inflight.entertainment","Seat.comfort","On.board.service","Leg.room.service","Cleanliness","Flight.Distance","Inflight.wifi.service","satisfaction")
traindf = train[columns]
testdf=test[columns]
library(dplyr)

count_data_train<- train %>% count(Gender,satisfaction)
count_data_train

#Plot-1
ggplot(count_data_train,
mapping=aes(x=satisfaction,y=n,fill=as.factor(Gender)))+
geom_bar(stat="identity",position=position_dodge(0.75),width=0.75)+
ggtitle("No. of Male/Female satisfaction count in train data")+
xlab("Satisfaction of Passengers")+ ylab("Count")
```

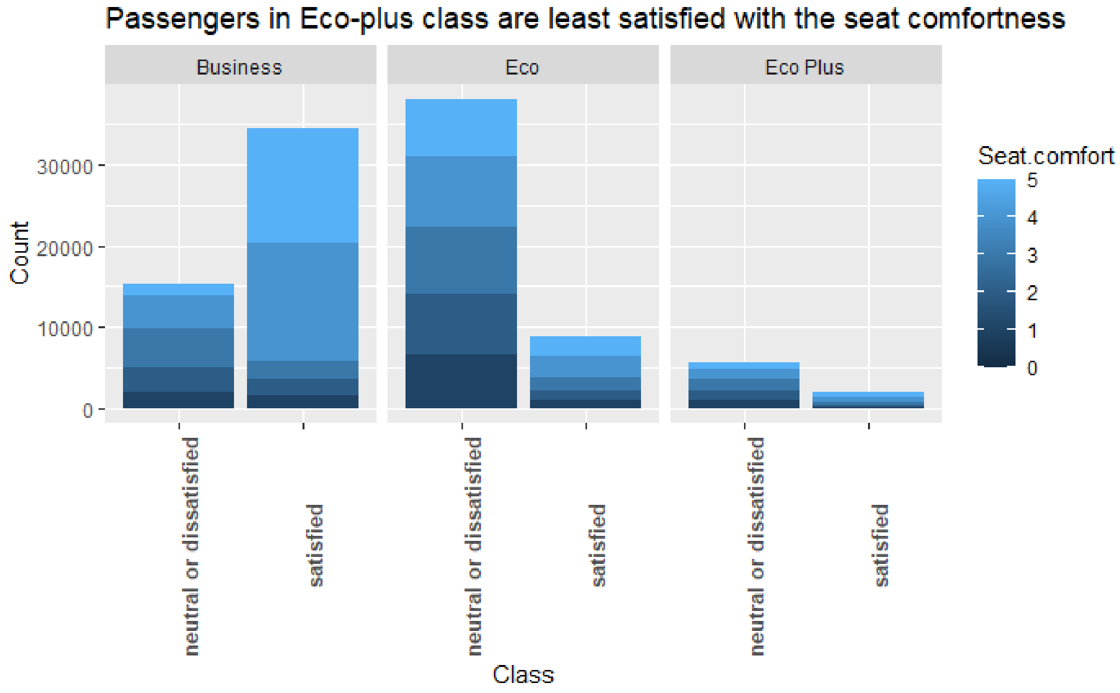


Observations:

- Women are more dissatisfied with the overall travel than men.
- Number of satisfied women and men is same.

#Plot-2

```
seat_count<- traindf %>% count(Class,Seat.comfort,satisfaction)
p1 <- ggplot(seat_count) + geom_col(aes(x = satisfaction, y = n, fill = Seat.comfort))
p1+ facet_wrap(~Class)+ ggtitle("Passengers in Eco-plus class are least satisfied with the seat comfortness")+ xlab("Class")+ ylab("Count")
```

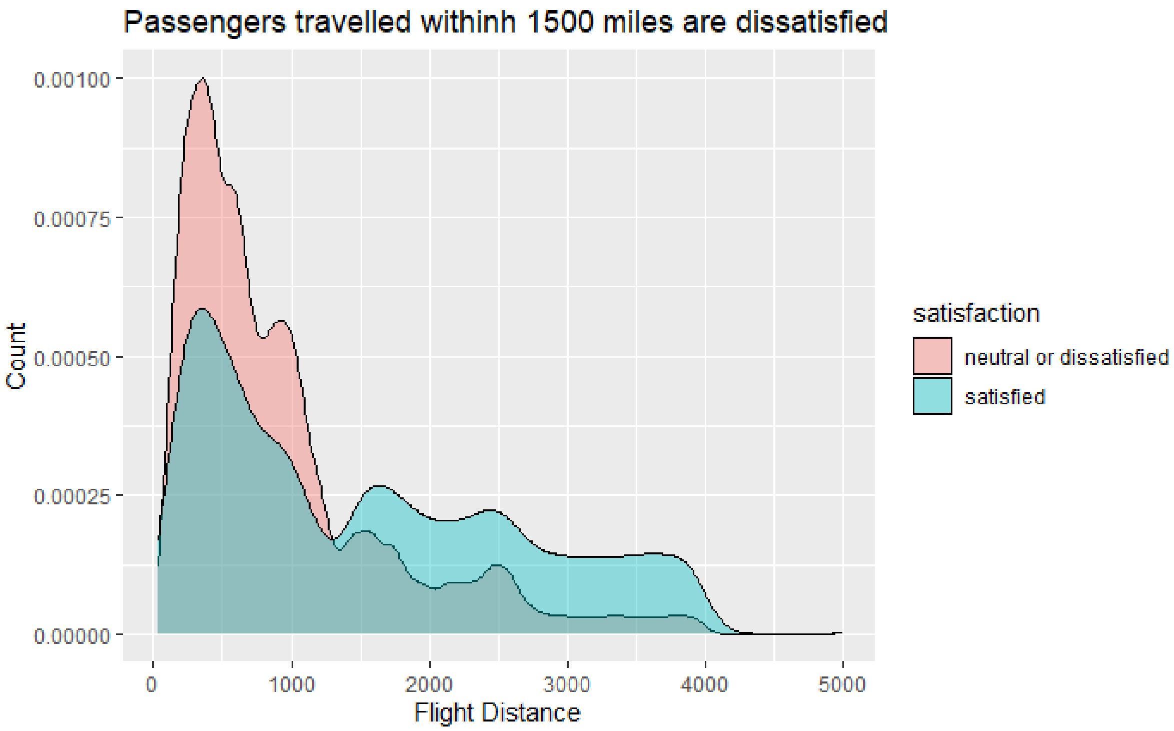


Observations:

- More number of people in business class are satisfied and have rated seat comfort high.
- More number of economy class passengers are dissatisfied and their rating with respect to seat comfort is spread across the range.

#Plot-3

```
ggplot(data=traindf, aes(x=Flight.Distance, fill=satisfaction))+
 geom_density()+ ggtitle("Passengers travelled within 1500 miles are dissatisfied")+ xlab("Flight Distance")+ ylab("Count")
```

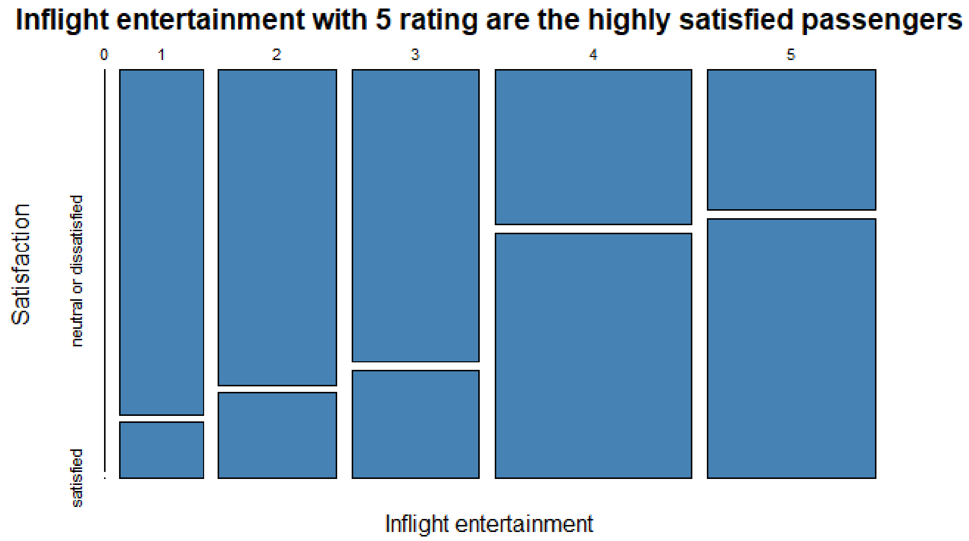


Observations:

- On average, people travelling shorter distance tend to be more dissatisfied.

#Plot-4

```
#create table of counts
counts <- table(traindf$Inflight.entertainment, traindf$satisfaction)
mosaicplot(counts, xlab='Inflight entertainment', ylab='Satisfaction',
        main='Inflight entertainment with 5 rating are the highly satisfied passengers',col='steelblue')
```

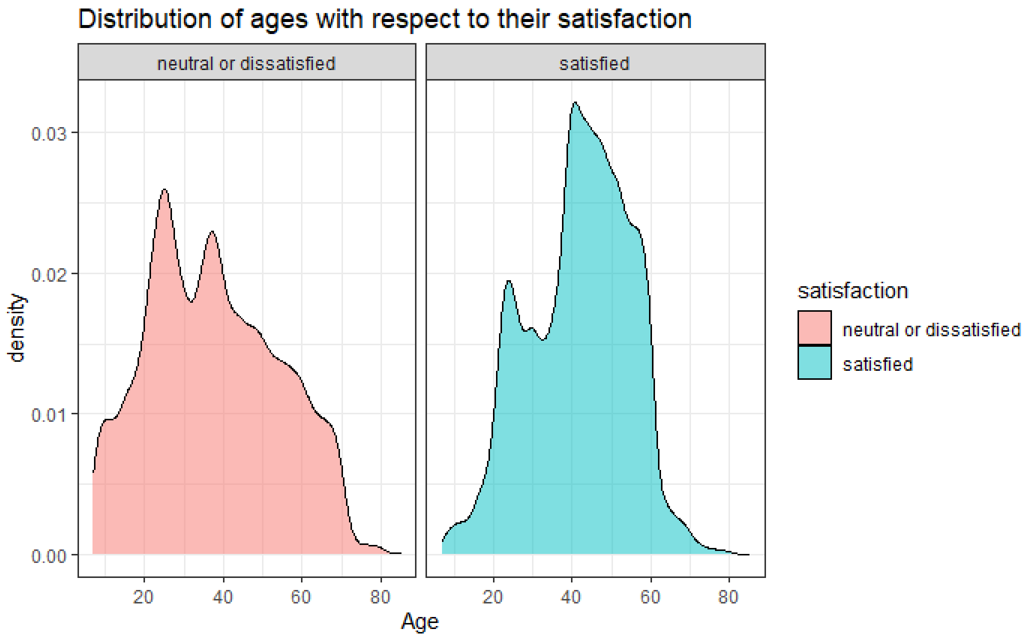


Observations:

- Inflight Entertainment is directly proportional to overall satisfaction
- Passengers who have rated ‘5’ for the Inflight entertainment are the majority ones who are satisfied with the travel.

<u>#Plot-5</u>

```
train %>% mutate(satisfaction = recode(satisfaction, `0` = "neutral or dissatisfied", `1` = "satisfied")) %>%
 ggplot(aes(x = Age, fill = satisfaction)) + geom_density(alpha = 0.5) +
 labs(x = NULL) +theme_bw() + facet_wrap(~satisfaction)
```

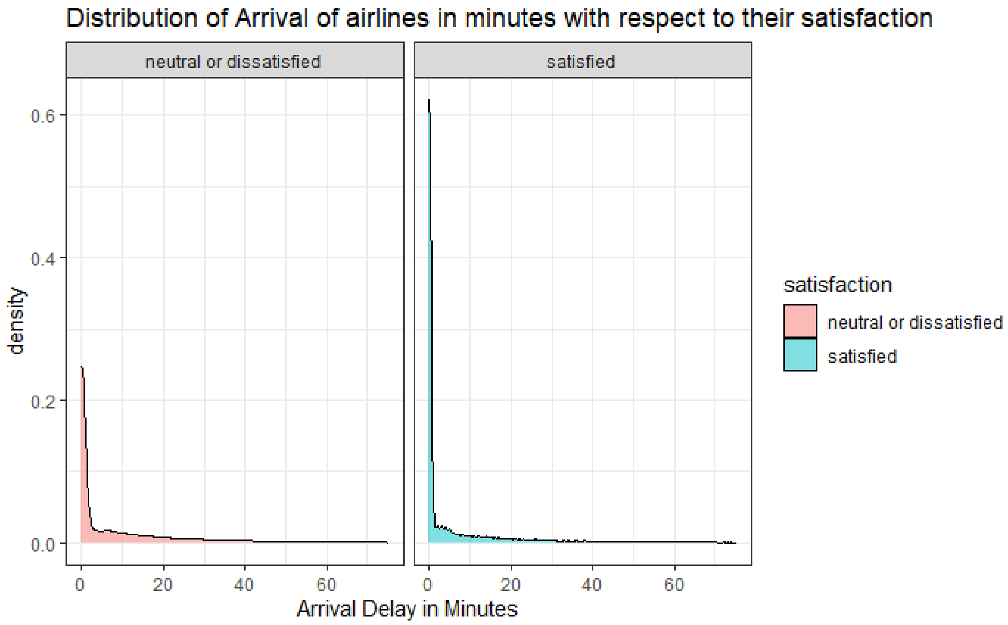


Observations:

- The passengers within the age range of 40-50 are the majority ones with overall satisfaction.
- The passengers within the age range of 20-40 are the majority ones with overall dissatisfaction.

#Plot-6

```
train %>% mutate(satisfaction = recode(satisfaction, `0` = "neutral or dissatisfied", `1` = "satisfied")) %>%
 ggplot(aes(x = Arrival.Delay.in.Minutes, fill = satisfaction)) +
 geom_density(alpha = 0.5) + xlim(0, 75) + labs(x = NULL) + theme_bw() +
 facet_wrap(~satisfaction)
```



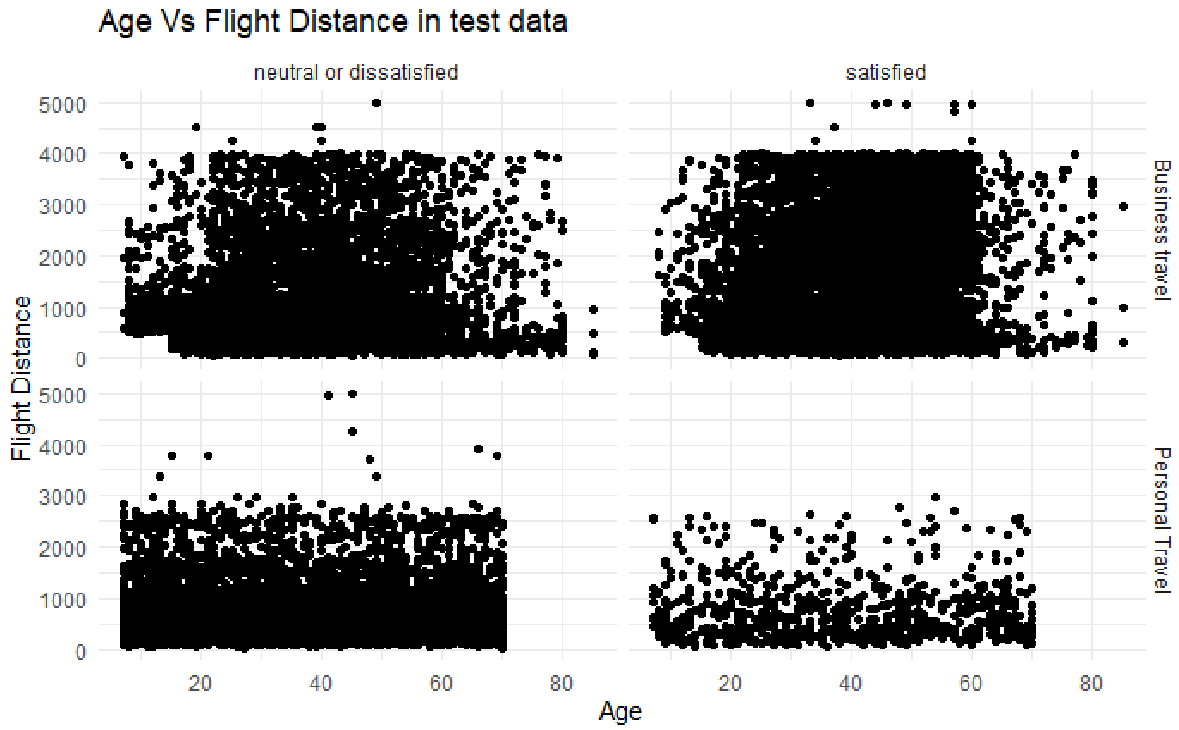
Observations:

- The number of satisfied passengers is **inversely proportional** to Arrival delay in minutes.
- The number of satisfied passengers with the arrival delay are high when compared with the number of dissatisfied passengers with the arrival delay.
- This shows that the overall satisfaction had no much impact with the consideration of arrival delay in minutes .

#Plot-7

test %>% ggplot(aes(x=Age, y=Flight.Distance)) + geom_point() +

```
  facet_grid(Type.of.Travel~satisfaction) +  labs(title="Age Vs Flight Distance in test data", x="Age", y="Flight Distance ") + theme_minimal()
```

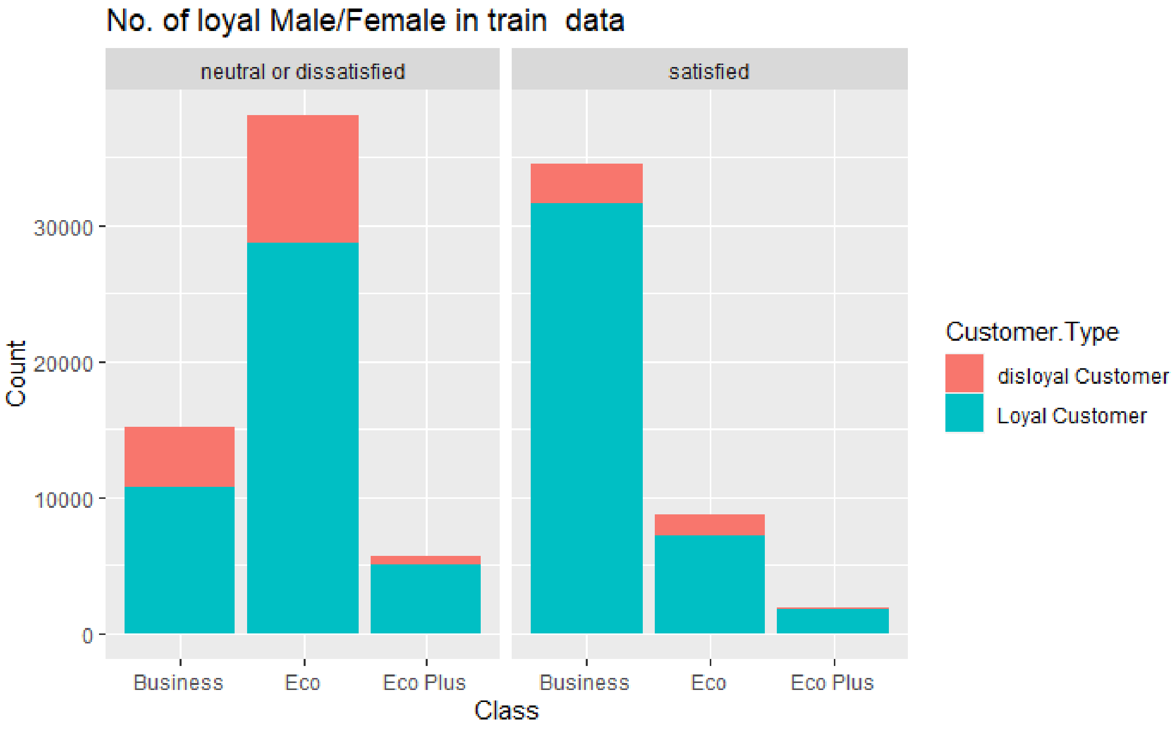


Observations:

- Passengers under Business travel type are majorly overall satisfied when their age range is within 20-40 and they are independent of the distance travelled.
- Passengers under Personal travel type are more dissatisfied than satisfied.
- When only satisfied passengers under Personal travel type are considered, the overall satisfaction is reducing with increase in travel distance.

<u>#Plot-8</u>

```
loyal_data_train<- train %>% count(Class,Customer.Type,satisfaction)
loyal_data_train
p1 <- ggplot(loyal_data_train) +
 geom_col(aes(x = Class, y = n, fill = Customer.Type))
p1+ facet_wrap(~satisfaction)+
  ggtitle("No. of loyal Male/Female in train  data")+
  xlab("Class")+ ylab("Count")
```



Observations:

- Maximum number of disloyal customers are under Eco-class and maximum loyal customers are under Business class.
- Most of the loyal Eco and Eco-plus customers are dissatisfied.

```
#Import libraries
library(tidyverse)
library(readr)
library(ggplot2)

#Import Dataset
data<-read.csv("/Users/aishwaryakurnutala/Downloads/train.csv")
#Printing the first few lines of the dataset
head(data)
```

<u>#Plot-9</u>

```
ggplot(data, aes( x=satisfaction,y=Flight.Distance, color=Class)) +
  geom_boxplot() + labs(title="Analysis of Satisfaction wrt flight distance and Class", x="Satisfaction", y="Flight Distance") +
  facet_wrap(~Class)+  theme_minimal()+
theme(legend.position='none',axis.text.x=element_text(face='bold',   angle=35,
size=7))+ theme(plot.title = element_text(size=11))
```

#Plot-10

```
ggplot(data,aes(x=On.board.service,color=Customer.Type, fill=Customer.Type))
+geom_histogram( position="dodge", size = 0.5, color="black", binwidth = 0.5)+
theme(legend.position="top")+ labs(title="Distribution of On Board Service for
Customer Types", x="On Board Service Rating") +
 theme_grey() +facet_grid(~satisfaction)+
 theme(plot.title = element_text(size=11)) + scale_x_continuous()
```

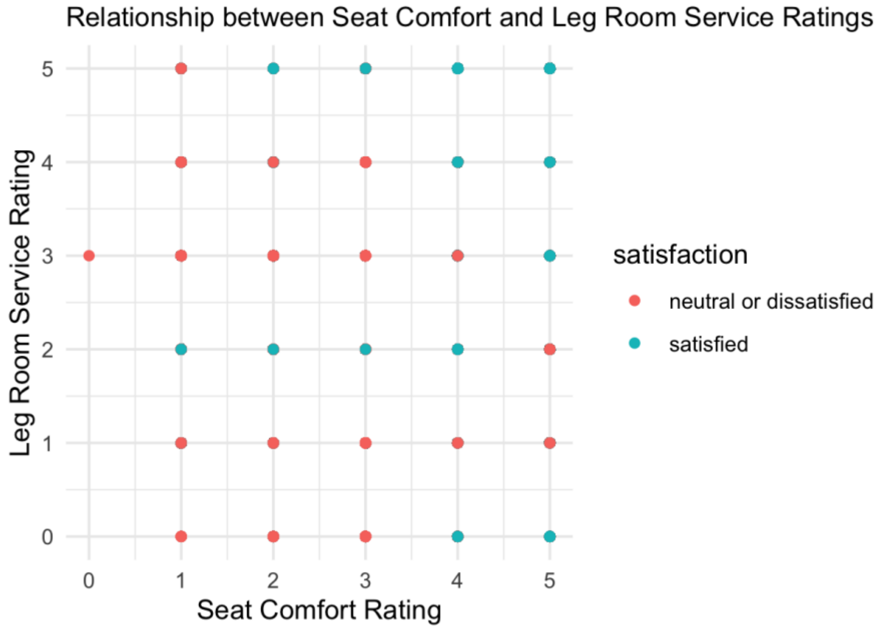
#Plot-11

```
ggplot(data, aes(x=Online.boarding, color=satisfaction, fill=satisfaction)) +
geom_freqpoly(bins=20) + labs(title="Distribution of Online Boarding Ratings
wrt Satisfaction", x="Online Boarding Rating ") + scale_x_continuous()+
theme_minimal()+theme(plot.title=element_text(size=11))+
scale_x_continuous()
```

#Plot-12

```
ggplot(data, aes(x=Seat.comfort, y=Leg.room.service,color=satisfaction)) +
  geom_point() + labs(title="Relationship between Seat Comfort and Leg Room
Service Ratings", x="Seat Comfort Rating", y="Leg Room Service Rating")+
  theme_minimal()+theme(plot.title=element_text(size=11))+
scale_x_continuous()
```

Output-



Observations:

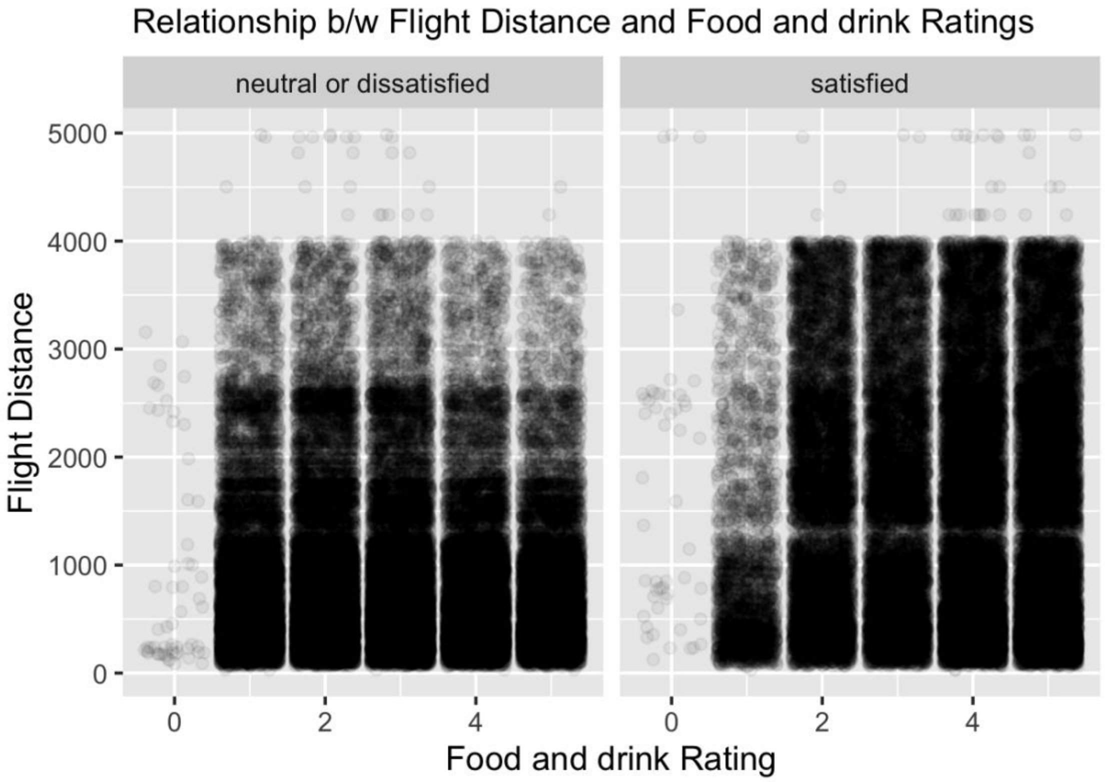
- Seat Comfort and Leg Room Service have received high ratings from majority of satisfied travelers.
- Some customers who rated Seat Comfort highly are dissatisfied with Leg Room Service, and vice - versa.
- As a result, a significant number of passengers are dissatisfied with Seat Comfort and Leg Room service.

```
#Plot-13
ggplot(data, aes( x=Type.of.Travel,y=Inflight.wifi.service)) +
  geom_boxplot(aes(fill = satisfaction)) +
  labs(title="Distribution of In-Flight Wifi Service against Type of Travel", x="Type
of Travel", y="In-flight Wifi Service Rating" ) +theme_gray()+
  theme(plot.title = element_text(size=11))
```

```
#Plot-14
ggplot(data, aes(x=Ease.of.Online.booking, color=satisfaction, fill=satisfaction))
+ geom_histogram( position="dodge", size = 0.3, color="black", binwidth=0.5)+
  theme(legend.position="top")+ labs(title="Distribution of Ease of Online
Booking Ratings",  x="Ease of Online Booking Rating") + theme_grey() +
 theme(plot.title = element_text(size=11)) + scale_x_continuous()
```

```
#Plot-15
ggplot(data,aes(x=Food.and.drink, y=Flight.Distance))+
geom_jitter(alpha=1/20)+ geom_smooth()+
  labs(x="Food and drink Rating", y="Flight Distance",
       title = " Relationship b/w Flight Distance and Food and drink Ratings")+
  facet_wrap(~satisfaction)+
  theme(plot.title = element_text(size=11)) + scale_x_continuous()
```

Output-



Observations:

- Passengers who flew a shorter distance in the flight are more likely to be neutral or dissatisfied.
- Satisfied passengers gave food and drink mixed ratings, even on longer flights.
- As a result, both satisfied and unsatisfied passengers gave mixed ratings for food and drinks.

Preprocessing of data:

```
library(tidyverse)
library(modelr)
library(ggplot2)
library(readr)
library(plyr)
library(dplyr)
library(e1071)
library(corrplot)
library(tidyr)
library(mlbench)
library("MLmetrics")
library(caret)
library(reshape2)
library("FSinR")
```

~~~
library(randomForest)
```


```{r}
data <- read_csv("datasets/airline/archive/train.csv")
data_test <- read_csv("datasets/airline/archive/test.csv")
```


```{r}
data_copy <- data
data_test_copy <- data_test
```
~~~

Label Encoding:

~~~
```{r}
data["Customer Type"] <- mapvalues(unlist(data["Customer Type"]),
        from=c("Loyal Customer","disloyal Customer"),
        to=c("0","1"))
data["Customer Type"] = as.numeric(unlist(data["Customer Type"]))

data["Type of Travel"] <- mapvalues(unlist(data["Type of Travel"]),
        from=c("Personal Travel","Business travel"),
        to=c("0","1"))
data["Type of Travel"] = as.numeric(unlist(data["Type of Travel"]))

data["Class"] <- mapvalues(unlist(data["Class"]),
        from=c("Business","Eco", "Eco Plus"),
        to=c("0","1","2"))
data["Class"] = as.numeric(unlist(data["Class"]))

data["Gender"] <- mapvalues(unlist(data["Gender"]),
        from=c("Male","Female"),
        to=c("0","1"))
data["Gender"] = as.numeric(unlist(data["Gender"]))

data["satisfaction"] <- mapvalues(unlist(data["satisfaction"]),
        from=c("satisfied","neutral or dissatisfied"),
        to=c("0","1"))
data["satisfaction"] = as.numeric(unlist(data["satisfaction"]))
```
~~~

~~~
```{r}
data_test["Customer Type"] <- mapvalues(unlist(data_test["Customer Type"]),
        from=c("Loyal Customer","disloyal Customer"),
        to=c("0","1"))
data_test["Customer Type"] = as.numeric(unlist(data_test["Customer Type"]))

data_test["Type of Travel"] <- mapvalues(unlist(data_test["Type of Travel"]),
        from=c("Personal Travel","Business travel"),
        to=c("0","1"))
data_test["Type of Travel"] = as.numeric(unlist(data_test["Type of Travel"]))

data_test["Class"] <- mapvalues(unlist(data_test["Class"]),
        from=c("Business","Eco", "Eco Plus"),
        to=c("0","1","2"))
data_test["Class"] = as.numeric(unlist(data_test["Class"]))

data_test["Gender"] <- mapvalues(unlist(data_test["Gender"]),
        from=c("Male","Female"),
        to=c("0","1"))
data_test["Gender"] = as.numeric(unlist(data_test["Gender"]))

data_test["satisfaction"] <- mapvalues(unlist(data_test["satisfaction"]),
        from=c("satisfied","neutral or dissatisfied"),
        to=c("0","1"))
data_test["satisfaction"] = as.numeric(unlist(data_test["satisfaction"]))
```


```{r}
sapply(data,class)
```


```{r}
scaled_data = scale(data[,3:24])
scaled_data=data.frame(scaled_data)
scaled_data["satisfaction"] = data["satisfaction"]
```


```{r}
scaled_data_test = scale(data_test[,3:24])
scaled_data_test=data.frame(scaled_data_test)
~~~

~~~
scaled_data_test["satisfaction"] = data_test["satisfaction"]
```


```{r}
sum(is.na(scaled_data$satisfaction))
```
```{r}
sum(is.na(scaled_data_test$satisfaction))
```



```{r}
filter_evaluator <- filterEvaluator('chiSquared')
skb_direct_search <- selectKBest(k=20)
skb_direct_search(scaled_data, 'satisfaction', filter_evaluator)
```


```{r}
scaled_data_final <- scaled_data %>%
select("Online.boarding","Class","Type.of.Travel","Inflight.entertainment","Seat.comfort"
,"On.board.service","Leg.room.service","Ease.of.Online.booking","Flight.Distance","Inflight.wifi.service","satisfaction")
scaled_data_final["satisfaction"]= as.factor(data_copy$satisfaction)
```
~~~

Data Scaling:

~~~
```{r}
scaled_data_test_final <- scaled_data_test %>%
select("Online.boarding","Class","Type.of.Travel","Inflight.entertainment","Seat.comfort"
,"On.board.service","Leg.room.service","Ease.of.Online.booking","Flight.Distance","Inflight.wifi.service","satisfaction")
scaled_data_test_final["satisfaction"]= as.factor(data_test_copy$satisfaction)
```
~~~

Model Building:

~~~
```{r}
lr <- glm(satisfaction~.,family=binomial,data=scaled_data_final)
lr
```
~~~

~~~
```{r}
svm <- svm(satisfaction~., data = scaled_data_final, kernel = "linear", cost = 10,
scale = FALSE)
svm
```

```{r}
rf <- randomForest(satisfaction~.,data = scaled_data_final)
rf
```
~~~

<u>Prediction and Model Metrics:</u>

~~~
```{r}
p1 <- predict(lr,scaled_data_test_final,type = "response")
p1 <- ifelse(p1 > 0.5, "satisfied", "neutral or dissatisfied")
confusionMatrix(factor(p1), scaled_data_test_final$satisfaction)
```

```{r}
p2 <- predict(rf,scaled_data_test_final)
confusionMatrix(p2, scaled_data_test_final$satisfaction)
```



```{r}
p1 <- mapvalues(unlist(p1),
        from=c("satisfied","neutral or dissatisfied"),
        to=c("0","1"))
p1 = as.numeric(p1)
RMSE(p1, data_test$satisfaction)
MAPE(as.numeric(p1), as.numeric(scaled_data_test_final$satisfaction))
MAE(p1, data_test$satisfaction)
F1_Score(p1, data_test$satisfaction)
```

```{r}
p2 <- mapvalues(unlist(p2),
        from=c("satisfied","neutral or dissatisfied"),
        to=c("0","1"))
p2 = as.numeric(levels(p2))
p2_f1 = as.numeric(p2)
~~~

~~~
RMSE(p2, data_test$satisfaction)
MAPE(as.numeric(p2), as.numeric(scaled_data_test_final$satisfaction))
MAE(p2, data_test$satisfaction)
F1_Score(p2_f1, data_test$satisfaction)
```


```{r fig.width=14, fig.height=10}
M = cor(data[,3:25][,!(names(data[,3:25]) %in% c("Arrival Delay in Minutes"))])
corrplot(M, method = 'color', order = 'alphabet')
```


```{r}
actual <- c(1.1, 1.9, 3.0, 4.4, 5.0, 5.6)
predicted <- c(0.9, 1.8, 2.5, 4.5, 5.0, 6.2)
MAPE(actual, predicted)
```
```{r}
head(scaled_data_final)
```
~~~

Feature Importance Score Chart:

~~~
```{r fig.width=13, fig.height=10}
freq <- c(39751.00, 28696.41, 26471.86, 20947.19,18508.07, 15756.13,
15406.58, 12271.37, 11508.56,10407.61 )
values <-
c("Online.boarding","Inflight.wifi.service","Class","Type.of.Travel","Inflight.ent
ertainment","Seat.comfort","Flight.Distance","Leg.room.service"
,"On.board.service", "Ease.of.Online.booking")
df <- data.frame(values,freq)
ggplot(df, aes(x = values, y = freq)) +
     geom_bar(stat="identity")
~~~