# Data Analysis for Southeast Airlines

SOUTHEAST AIRLINES

The Spirit of Liberty™

Ryan Ondocin | Conor Dugan | Zeyi Luo | Jian Shi | Yaakob Vaisman

# Table of Contents

# Introduction

## Description

This project focuses on using US airline data from over 10,000 customers in order to provide recommendations and actionable insights to Southeast Airline Co. Our insights ultimately serve the purpose of increasing customer satisfaction for Southeast Airline Co.

## Project Overview

This project aimed at reducing the customer defection or 'churn'-rate for Southeast Airlines. The survey data contained information from Southeast and 13 other partner airlines. Each row provided customer information, flight statistics and the customer's likelihood to recommend the airline to another individual. This work was ostensibly done in response to the observed devaluation of loyalty programs for frequent flyers according to the International Air Transport Association. Rewarding frequent flyers with discounted miles has become an antiquated practice in the industry that is merely putting Southeast Airlines Co. deeper in debt. This notion has prompted our team to analyze the dataset and figure out what areas SouthEast airlines needs to improve upon. Following rigorous statistical and modeling techniques, our team has proposed a few suggestions that can help our client reduce customer attrition.

# Deliverables

- Clean the data to ensure that there are no invalid fields/null values for each attribute.
- Isolate and visualize attributes that have the largest effect on customer satisfaction ratings (positive and negative).
- Predict likelihood to recommend using a support vector machine and linear/logistic regression models.
- Derive analysis-based insights from the aforementioned models and visualizations.
- Provide suggestions to client based on these insights.

# Data Acquisition

The data set was made available to us by the professor. It is a survey dataset that contains thousands of observations of flight segment data. Each row represents one flight segment, by one airline (either southeast or one of its partner airlines), for a specific customer. Each column represents an attribute of that particular flight segment. Each row captures 26 characteristics of the flight (ex. day of month, date, airline, origin and destination city, if the flight was delayed), the customer (ex. age, gender, price sensitivity, the person's frequent flyer status). The row also contains a simple survey-based rating of each customer's likelihood to recommend the airline that they just flew on as well as a field for open-ended text comments. The data was previously used mostly to examine Net Promoter Score, but we hope to gain a better understanding of our clientele's behavior by increasing the scope of attributes we examine.

## Data Cleaning and Preprocessing

We imported the data from a JSON url and converted into a dataframe. Firstly,  we manually detected NAs in each column. We found there were NAs in the following columns:

- Flights per year
- Departure.Delay.in.Minutes
- Arrival.Delay.in.Minutes
- Flight.time.in.minute
- Likelihood.to.recommend

In order to replace those missing values, we created several histograms of each column containing NAs to visualize spread and figure out appropriate NA substitution values. Based on approximate right skewed distribution of those histograms in all but one instance, we decided to use **median** to replace missing values. For the Likelihood.to.Recommend column, we replaced NAs with mean since it is left skewed. Finally, since our evaluation is based on the Net Promoter Score (NPS) method, we sorted each observation into three categories: detractors, passive, and promoters.

Because our partner airlines operate independently, the majority of our analyses are done specifically with data from Southeast Airlines, unless otherwise stated. Because we cannot change the behavior or experience for those flying on our partner airlines, we have focused our analyses on Southeast Airlines, except when appropriate.

### Code:

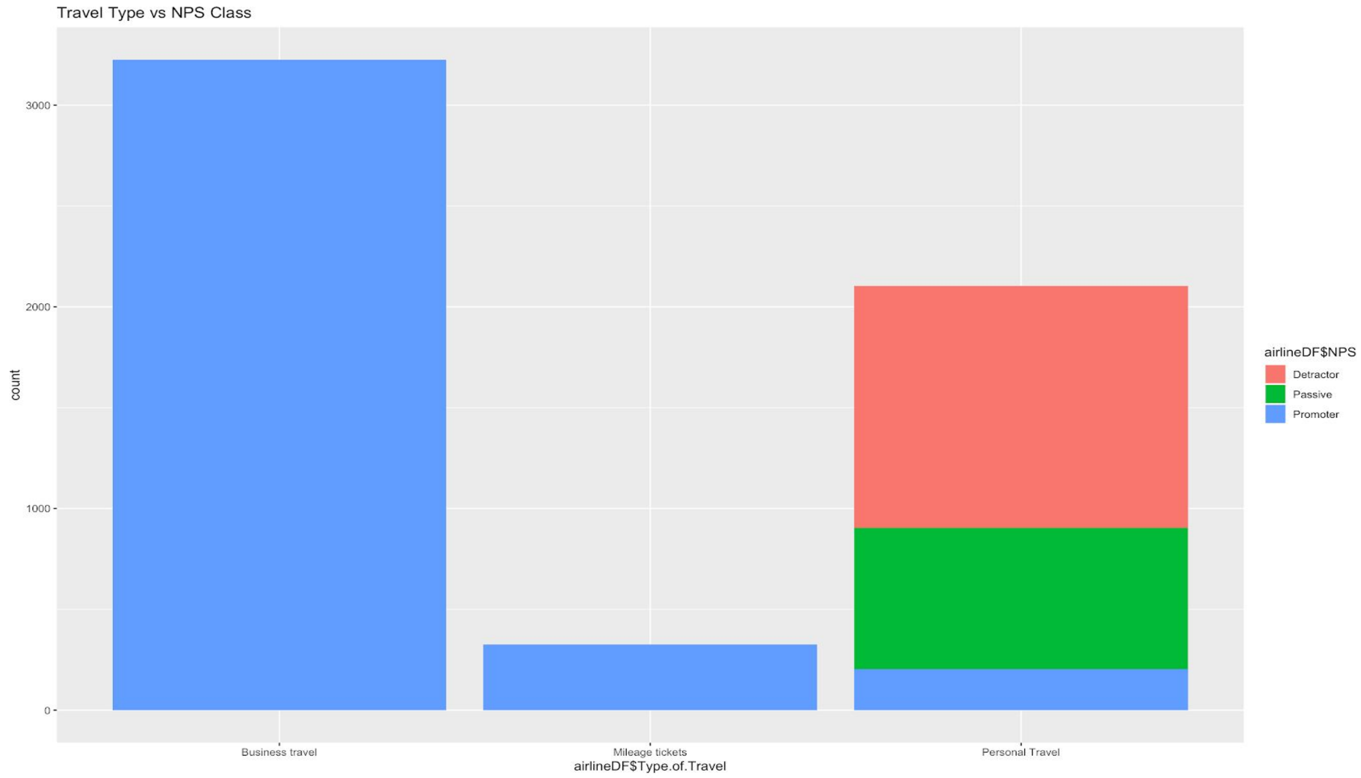
```
#checking for missing values in each column

sum(is.na(airlineDF$Destination.City))

sum(is.na(airlineDF$Origin.City))
sum(is.na(airlineDF$Airline.Status))
sum(is.na(airlineDF$Age))
sum(is.na(airlineDF$Gender))
sum(is.na(airlineDF$Price.Sensitivity))
sum(is.na(airlineDF$Year.of.First.Flight))
sum(is.na(airlineDF$Flights.Per.Year))
sum(is.na(airlineDF$Loyalty))
sum(is.na(airlineDF$Type.of.Travel))
sum(is.na(airlineDF$Total.Freq.Flyer.Accts))
sum(is.na(airlineDF$Shopping.Amount.at.Airport))
sum(is.na(airlineDF$Eating.and.Drinking.at.Airport))
sum(is.na(airlineDF$Class))
sum(is.na(airlineDF$Day.of.Month))
sum(is.na(airlineDF$Flight.date))
sum(is.na(airlineDF$Partner.Code))
sum(is.na(airlineDF$Partner.Name))
sum(is.na(airlineDF$Origin.State))
sum(is.na(airlineDF$Destination.State))
sum(is.na(airlineDF$Scheduled.Departure.Hour))
sum(is.na(airlineDF$Departure.Delay.in.Minutes))#208 missing
sum(is.na(airlineDF$Flight.cancelled))
sum(is.na(airlineDF$Flight.time.in.minutes)) #238 missing
sum(is.na(airlineDF$Flight.Distance))
sum(is.na(airlineDF$Likelihood.to.recommend)) #1 missing
sum(is.na(airlineDF$olong))
sum(is.na(airlineDF$olat))
sum(is.na(airlineDF$dlong))
sum(is.na(airlineDF$dlat))
sum(is.na(airlineDF$freeText)) #10000 missing
#creating a histogram of flights per year column to visualize spread and figure out appropriate NA substitution values
hist(airlineDF$Flights.Per.Year)
hist(airlineDF$Departure.Delay.in.Minutes[airlineDF$Departure.Delay.in.Minutes<120],)
#right skewed distribution
#therefore median is more appropriate replacement value for na
hist(airlineDF$Arrival.Delay.in.Minutes[airlineDF$Arrival.Delay.in.Minutes<120],)#right skewed distribution
hist(airlineDF$Flight.time.in.minutes[airlineDF$Departure.Delay.in.Minutes],)
#imputing missing values with median due to distribution being rightly-skewed
med_flight<-median(airlineDF$Flight.time.in.minutes, na.rm=TRUE)
med_dep<-median(airlineDF$Departure.Delay.in.Minutes, na.rm=TRUE) #median is equal to 0

#imputing the values
airlineDF$Departure.Delay.in.Minutes[is.na(airlineDF$Departure.Delay.in.Minutes)]<-med_dep
airlineDF$Flight.time.in.minutes[is.na(airlineDF$Flight.time.in.minutes)]<-med_flight

sum(is.na(airlineDF$Departure.Delay.in.Minutes)) #0 remain
sum(is.na(airlineDF$Flight.time.in.minutes)) #0 remain

#Imputing likelihood to recommend NAs with mean due to negative distribution
mean_like<-mean(airlineDF$Likelihood.to.recommend, na.rm=TRUE)
airlineDF$Likelihood.to.recommend[is.na(airlineDF$Likelihood.to.recommend)]<-mean_like
sum(is.na(airlineDF$Likelihood.to.recommend)) #0 remain
```

# Data Visualization

The following visualizations were created using ggplot2. They helped our team understand the given data and figure out which attributes are crucial when it comes to determining customer satisfaction.

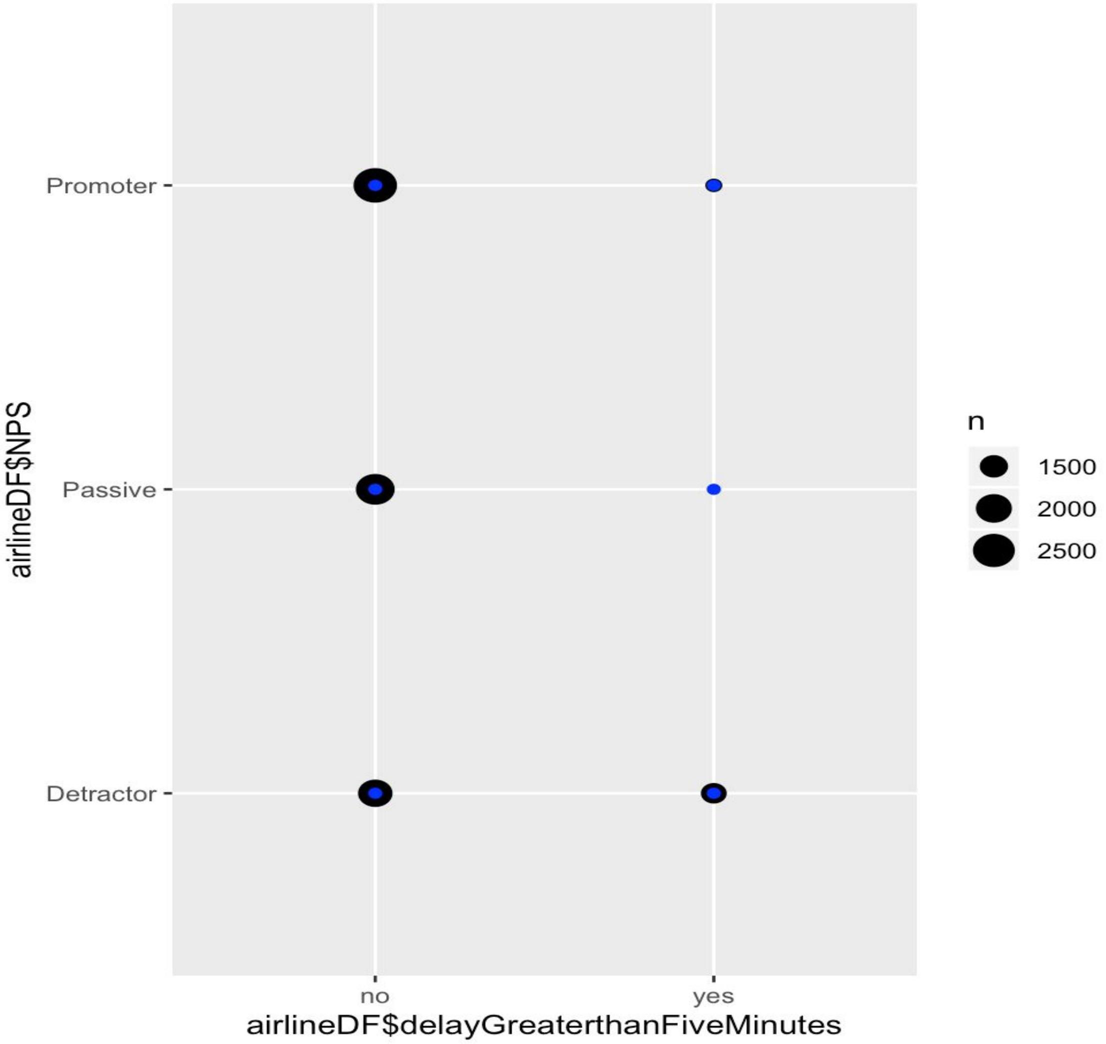
## Primary Visualizations:

### Figure 1: Net Promoter Class Versus Type of Travel



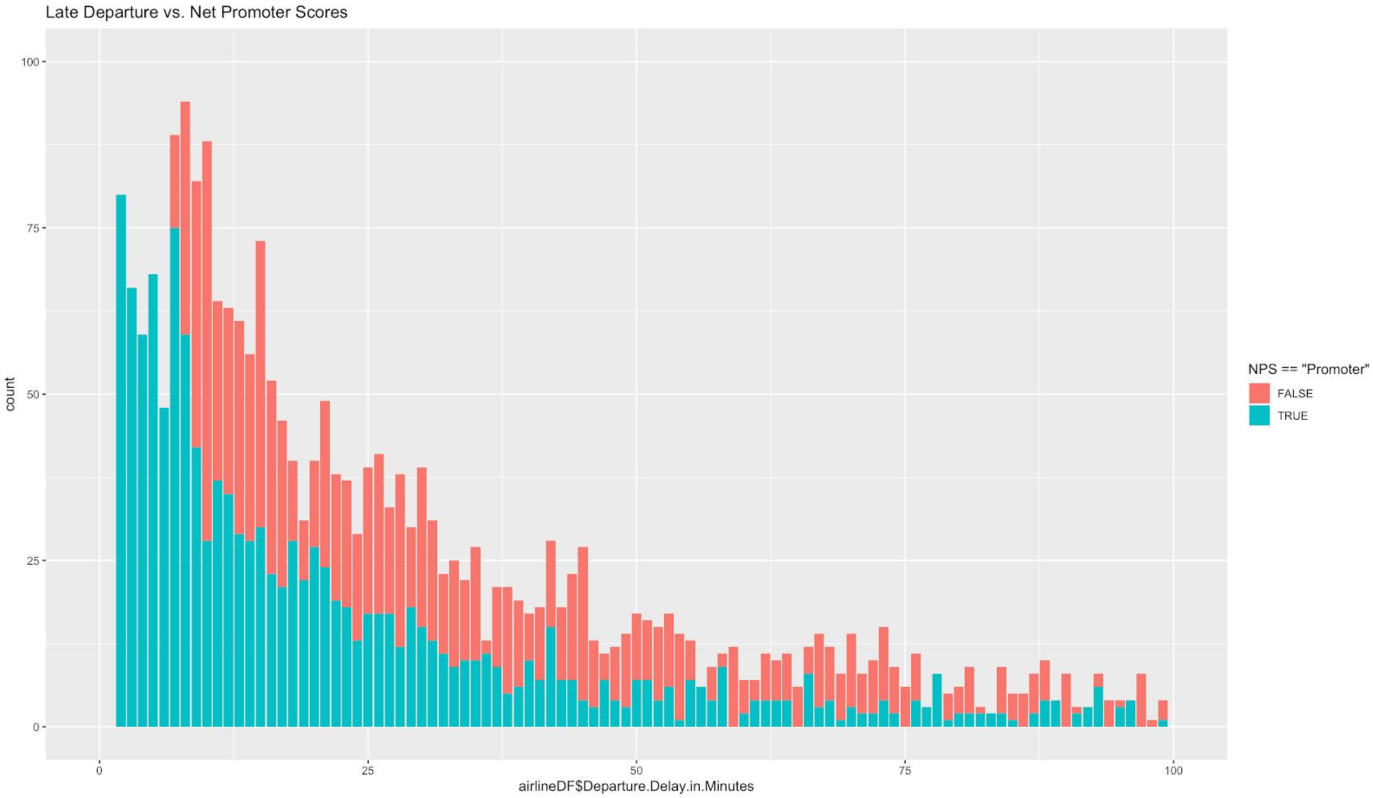
This graph shows that the largest portion of detractors and passives for Southeast Airlines are Personal travelers. This is a group that should be focused on and investigated in more detail if Southeast Airlines would like to reduce their customer defection rate. Why do business travelers enjoy the airline, but personal travelers do not?

Figure 2: NPS relation to Flight Delay



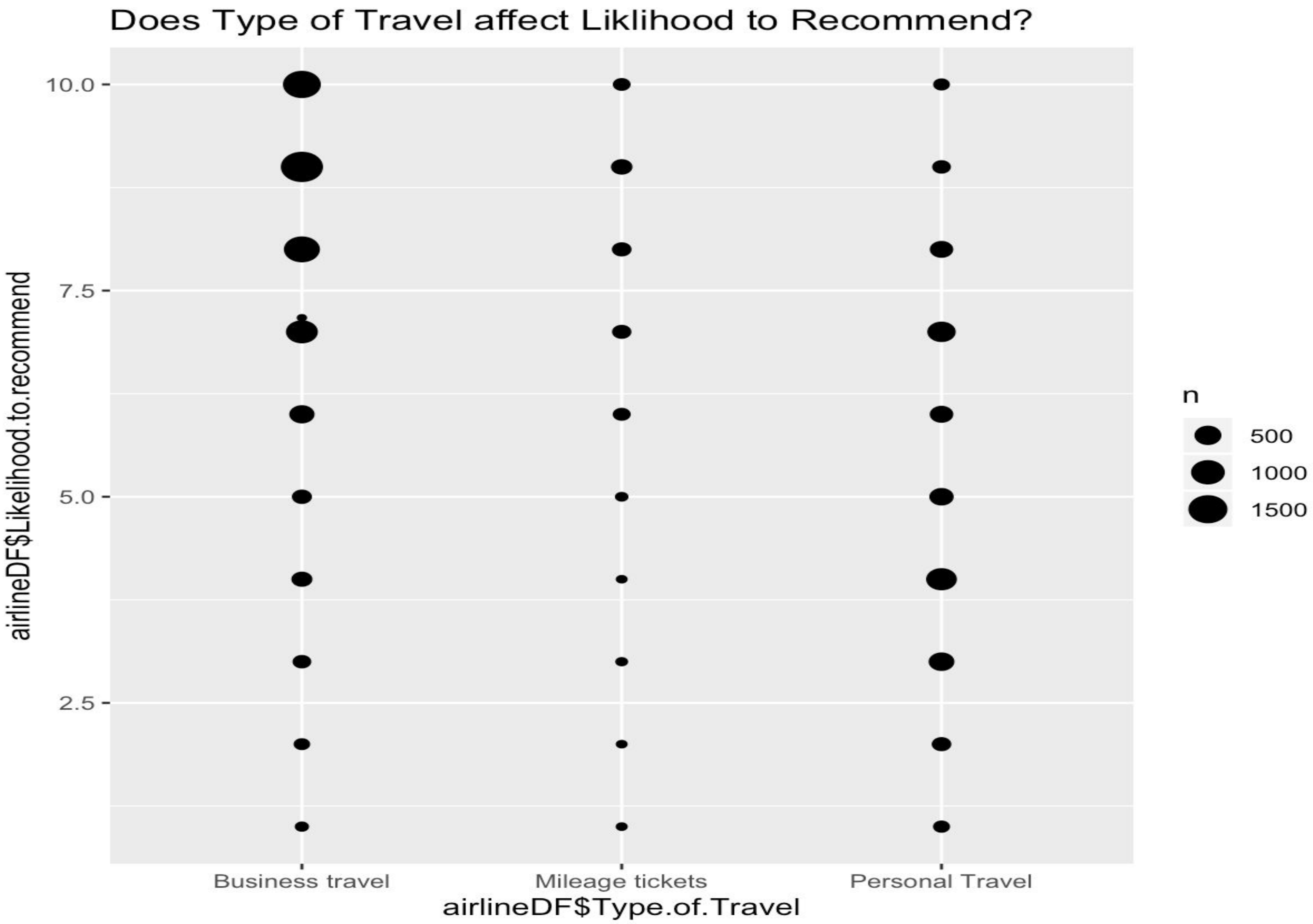
This figure was created following the creation of a new attribute: delayGreaterthanFiveMinutes, which takes a value of 1 for true and 0 for false. This plot suggests that flight delays greater than five minutes can contribute to customer becoming a detractor. This is because we see a noticeable increase in detractors when there is a delay greater than five minutes.

## Figure 3: Late Departure as compared to NPS



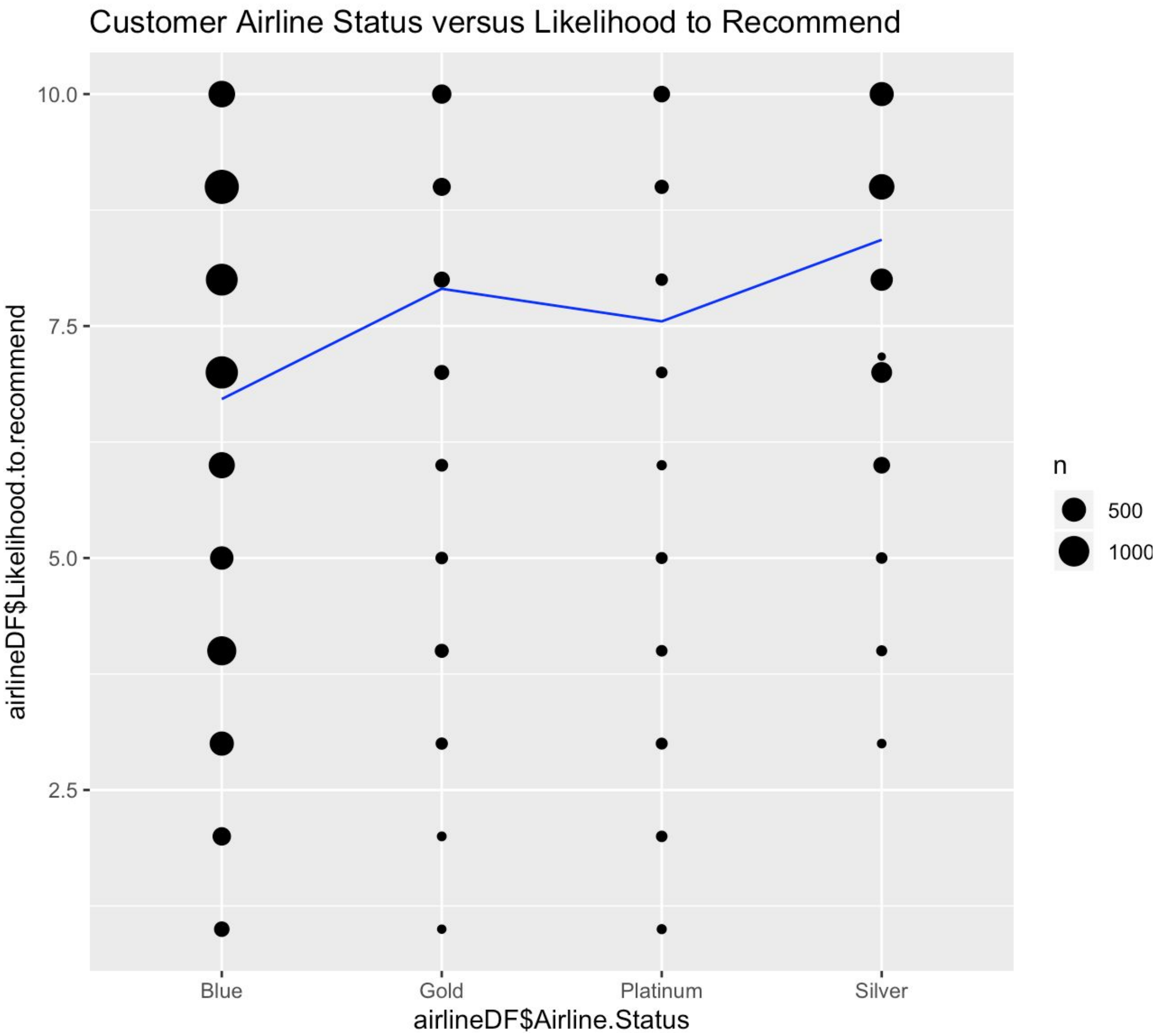
This figure is a histogram that displays departure delay in minutes as well as whether or not the customer was a promoter. This chart makes it clear that as flight delay time becomes greater than five minutes, customers tend to fall into the Passive and Detractors categories. Note the lack of any detractors or passives early on.

Figure 4: Travel Type's affect on NPS



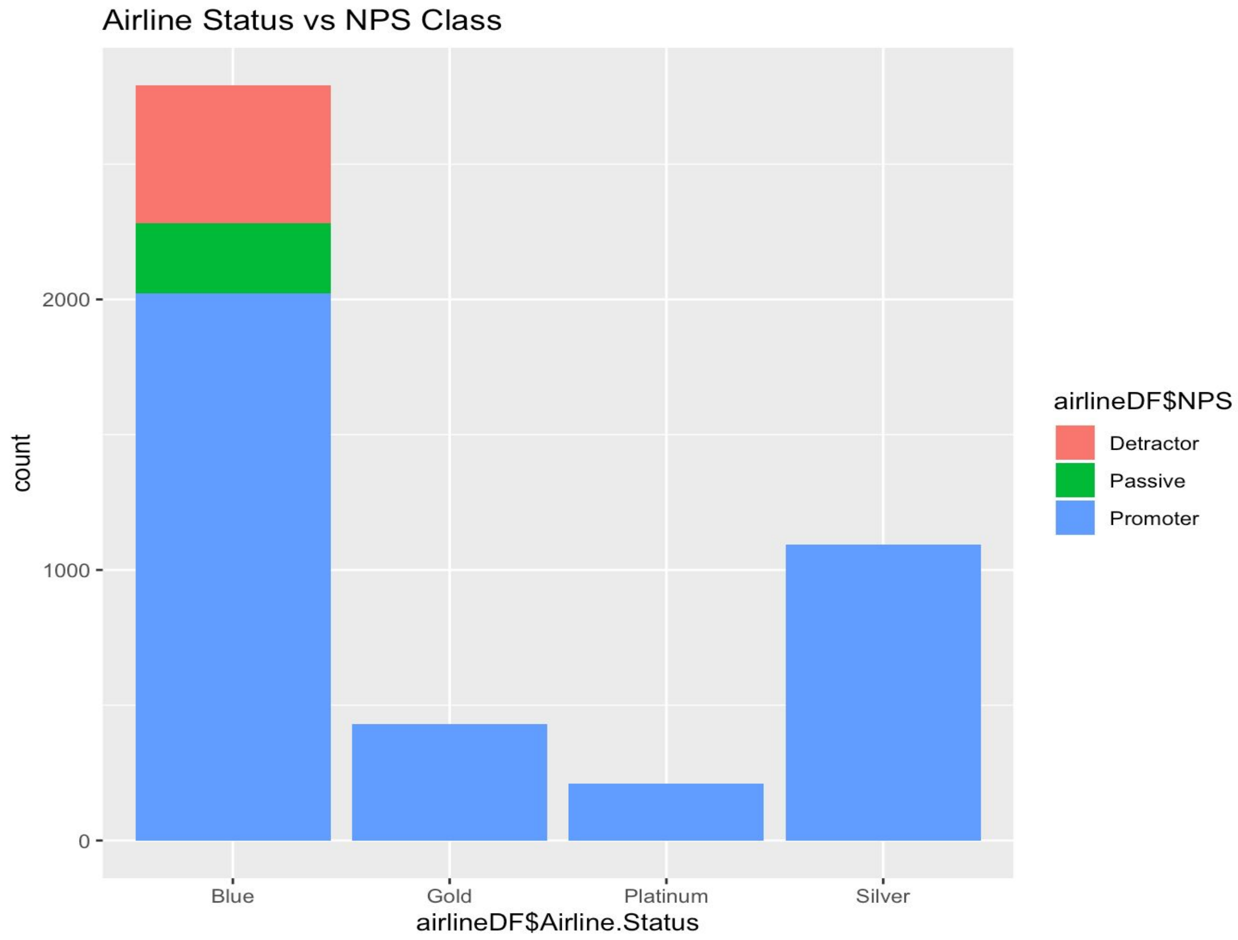
This chart demonstrates that Business travelers were most likely to give a higher promoter score, while personal travelers were more likely to give a lower promoter score. Free Mileage travels tended towards higher scores, but did not represent a significant portion of the data.

Figure 5: Airline Flyer Status affect on NPS



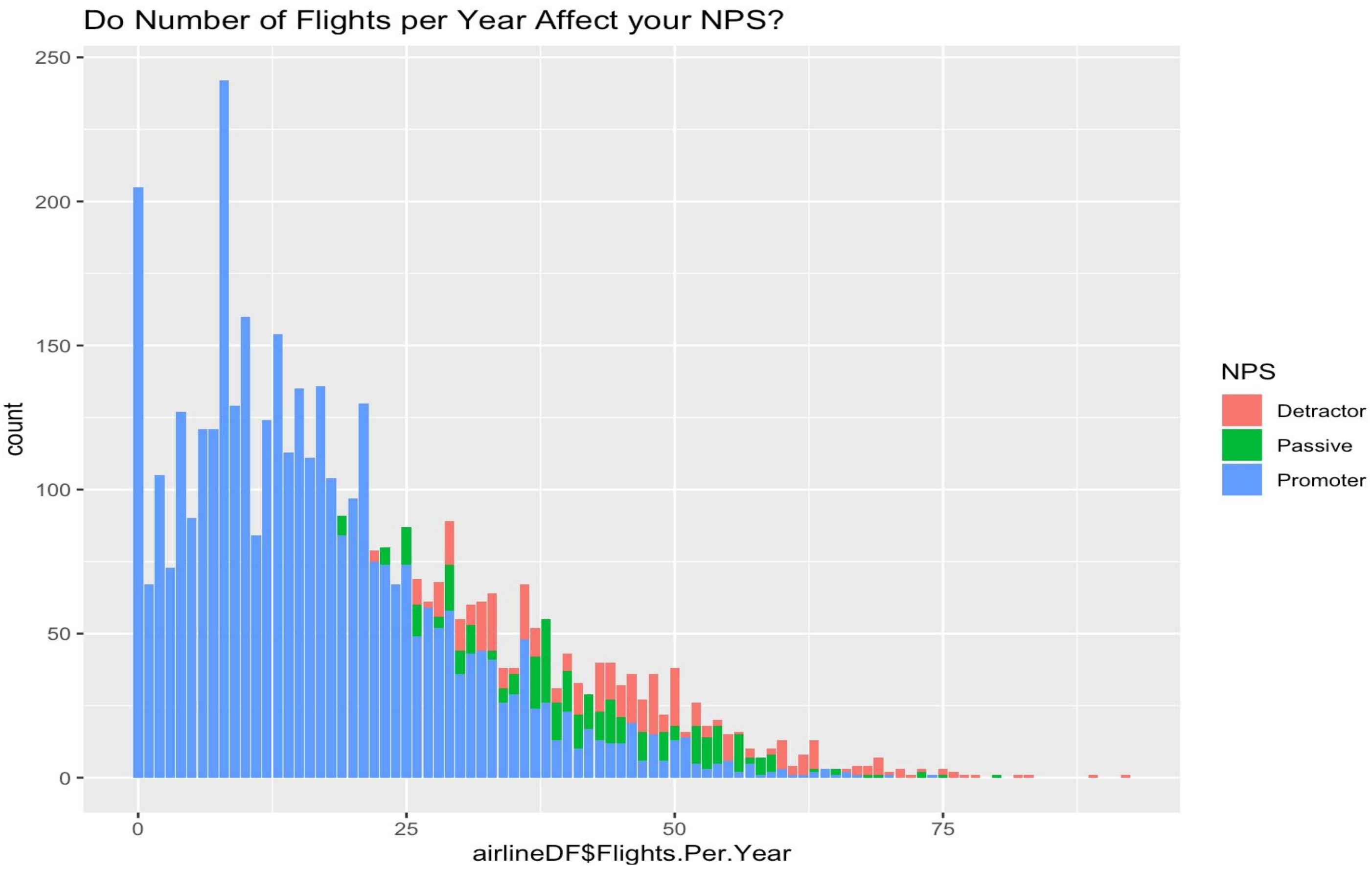
This figure demonstrates how Silver status flyers have the greatest likelihood to recommend across all flyer statuses and with all airline partners. Note that there are still more promoters in Blue status, but also more detractors.

Figure 6: Southeast Airlines Status Flyers Categories



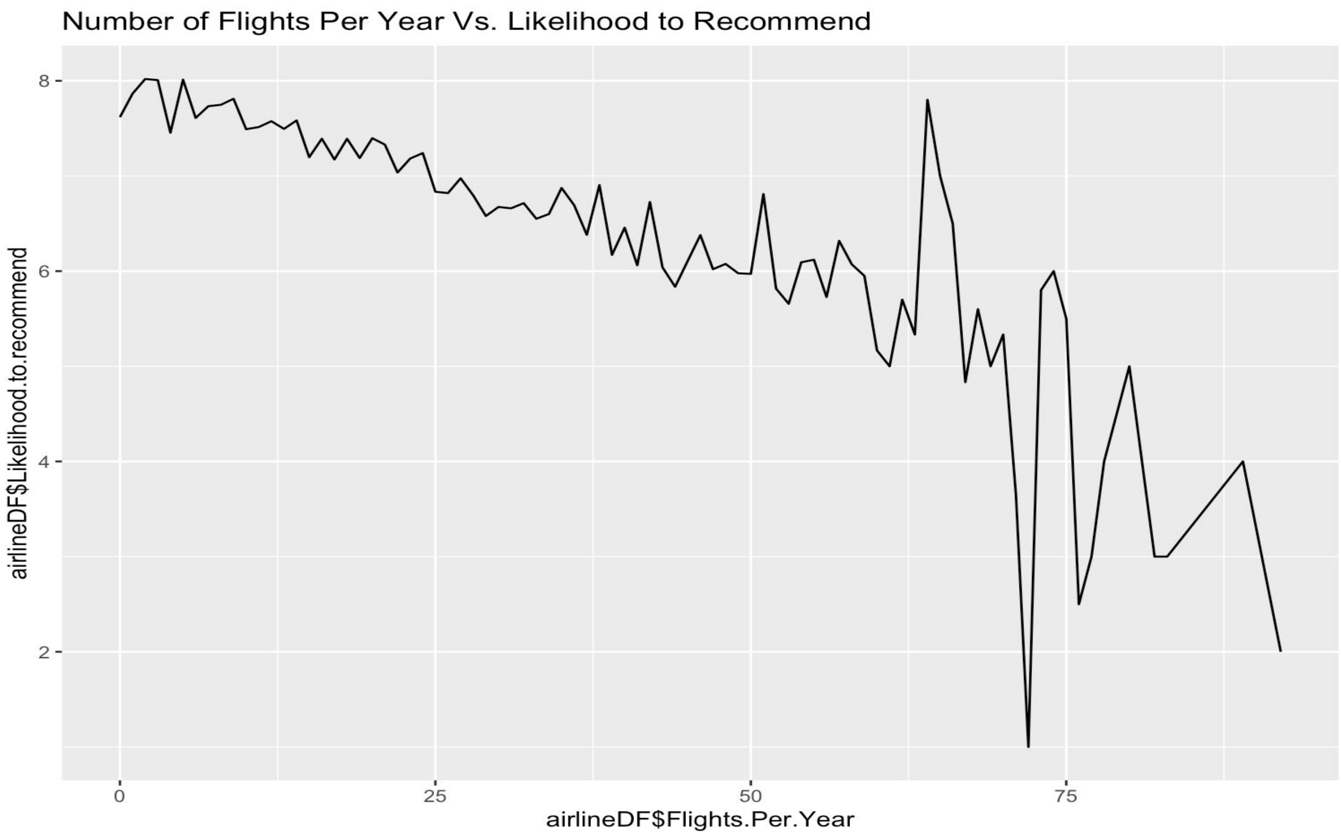
We can see here that for those passengers flying with Southeast Airlines, any upgraded status identified the flyer as being a likely Promoter. Detractors and Passive clients only existed in the Blue level, or those who are new to our flyer program.

Figure 7: Flights per Year effect on Promoter status



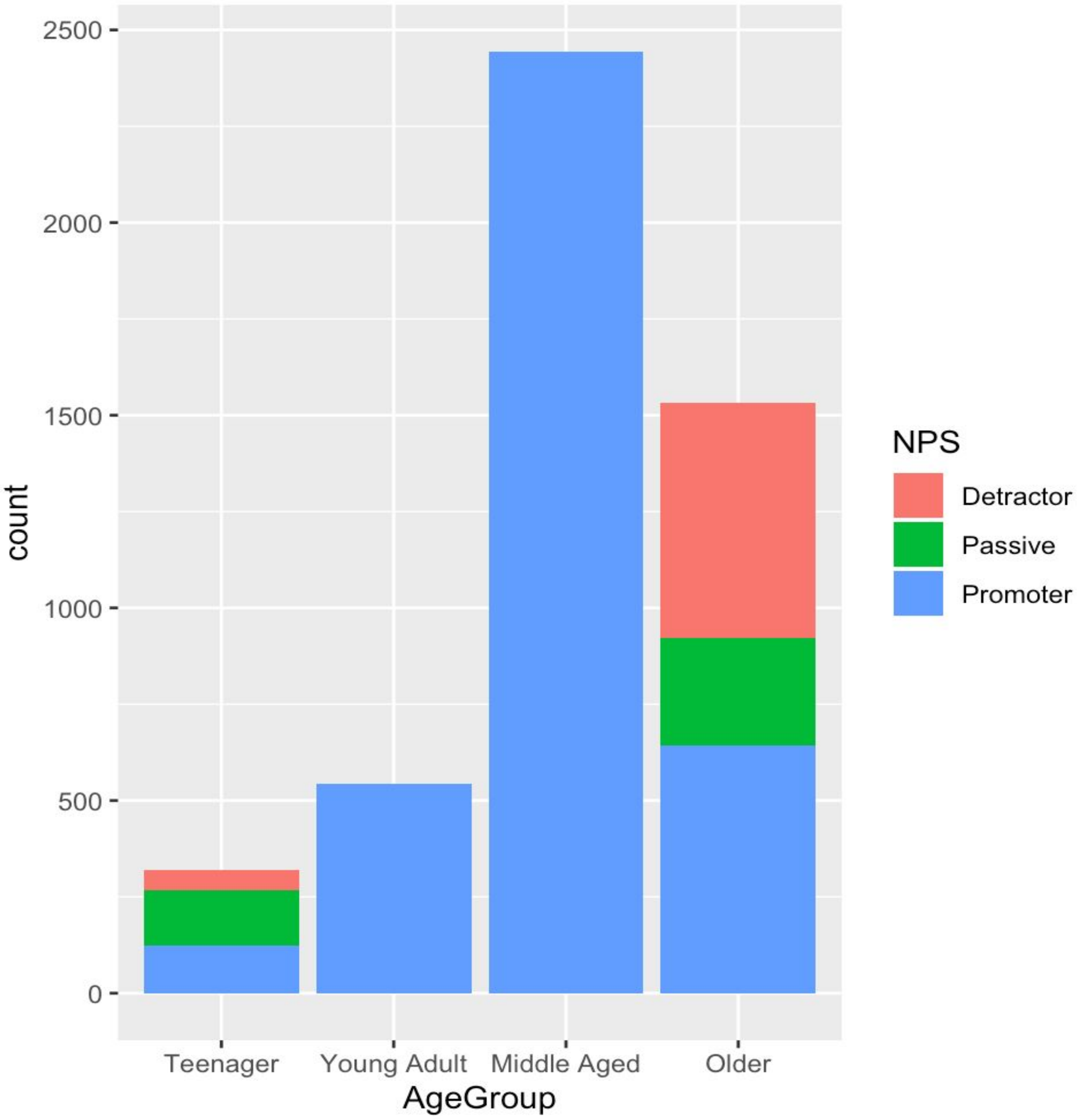
This chart demonstrates how the Detractors, Passives, and Promoters are spread depending on the number of flights they take per year.

Figure 8: Flights Per Year and Correlation to Likelihood to Recommend



We can see here that the mean score of a customer’s likelihood to recommend score falls as their number of flights increases. The large jumps in the graph are to be expected as the number of data points decreases as the flights per year increases. Still this graph is able to confidently show us a significant decrease in customer satisfaction for frequent flyers.

Figure 9: Promoter Category by Age



This chart shows that passives and detractors occur mostly at either end of the age spectrum, with teenagers, and mature customers most often acting as detractors.

# Code:

```
########################################################
Data Visualization with ggplot2
library('ggplot2')
pl<-ggplot(airlineDF, aes(x = airlineDF$Type.of.Travel, y = ..count..)) + geom_bar(aes(fill=airlineDF$NPS),position =
"identity")
pl+ggtitle("Travel Type vs NPS Class")
pl<-ggplot(southEast, aes(x = southEast$NPS, y = ..count..)) + geom_bar(aes(fill=airlineDF$Type.of.Travel),position =
"identity")
pl+ggtitle("NPS Class vs. Travel Type")

pl<-ggplot(airlineDF, aes(x = airlineDF$Airline.Status, y = ..count..)) + geom_bar(aes(fill=airlineDF$NPS),position = "identity")
pl+ggtitle("Airline Status vs NPS Class")



pll<-ggplot(airlineDF, aes(x = airlineDF$NPS, y = ..count..,fill=airlineDF$Age)) + geom_bar()
pll+ggtitle("Southeast Airlines Co. NPS Scores")
ggplot(southEast, aes(x = southEast$NPS, y = ..count..)) + geom_bar()



countp<-ggplot(airlineDF,aes(x=airlineDF$Arrival.Delay.in.Minutes,y=airlineDF$Likelihood.to.recommend))+geom_count()+
  stat_summary(aes(y=airlineDF$Likelihood.to.recommend ,group=1), fun.y=mean, colour="red", geom="point",group=1)
countp<-countp+ggtitle("Arrival Delay in Minutes versus Likelihood to recommend")
countp<-countp+xlim(0,100)
countp


ddd<-ggplot(airlineDF,aes(x=airlineDF$Departure.Delay.in.Minutes,y=airlineDF$Likelihood.to.recommend))+geom_count()+
  stat_summary(aes(y=airlineDF$Likelihood.to.recommend ,group=1), fun.y=mean, colour="green", geom="point",group=1)
ddd<-ddd+ggtitle("Departure Delay in Minutes versus Likelihood to recommend")
ddd<-ddd+xlim(0,100)
ddd




countp<-ggplot(airlineDF,aes(x=airlineDF$Arrival.Delay.in.Minutes,y=airlineDF$Likelihood.to.recommend))+geom_count()+
  stat_summary(aes(y=airlineDF$Likelihood.to.recommend ,group=1), fun.y=mean, colour="red", geom="point",group=1)



pll<-ggplot(airlineDF, aes(x = airlineDF$Likelihood.to.recommend, y = ..count..)) + geom_bar()
pll+ggtitle("Likelihood to Recommend Airline")
```

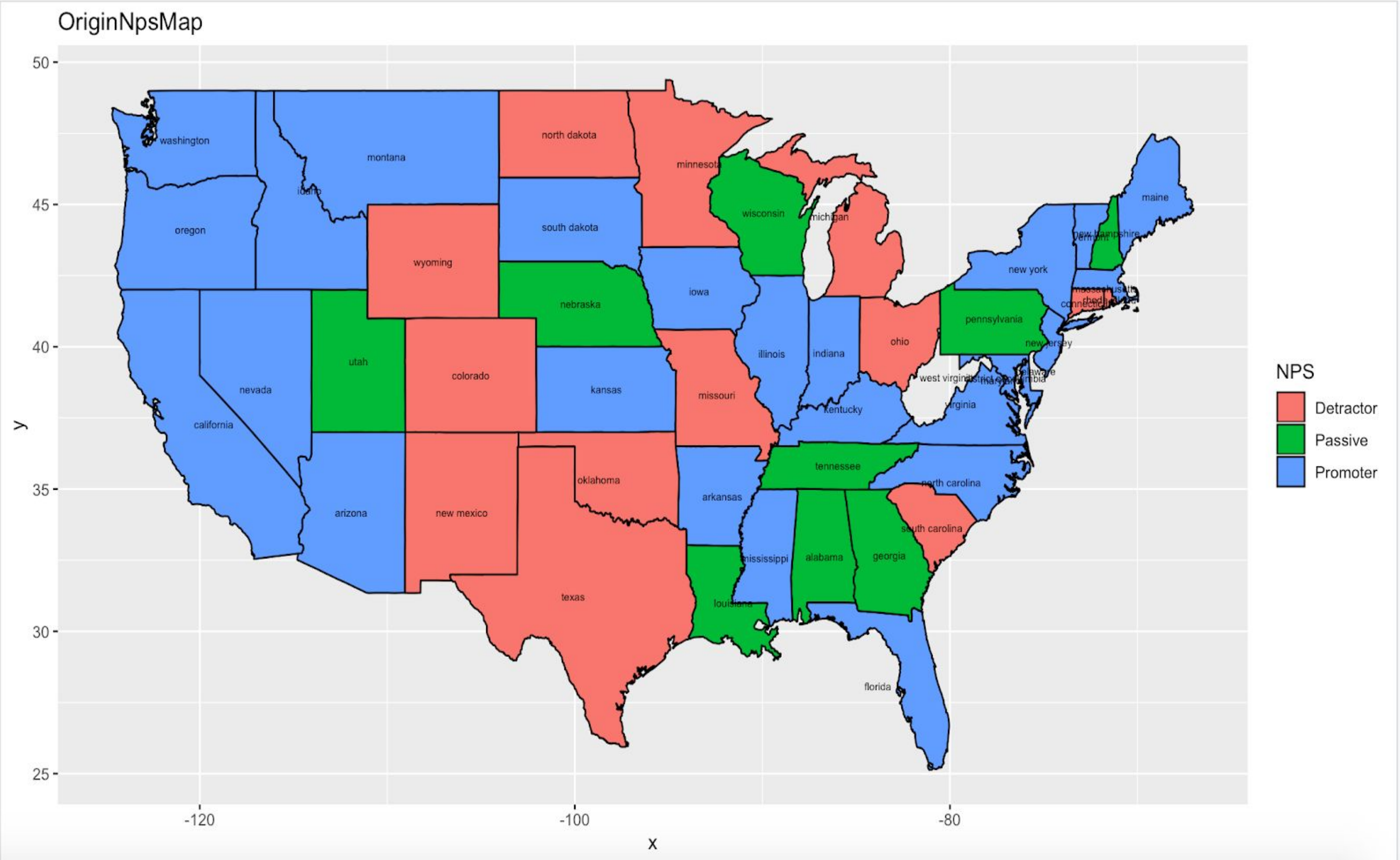
```
ggplot(southEast, aes(x = southEast$NPS, y = ..count..)) + geom_bar()


View(problemStates)
p<-ggplot(problemStates,aes(x=problemStates$delayGreaterthanFiveMinutes,
                            y=problemStates$NPS, main="does departure delay affect NPS?"))+geom_count()+
  stat_summary(aes(y=problemStates$NPS ,group=1), fun.y=mean, colour="blue", geom="point",group=1)

p + ggtitle("Affect on Delay time and NPS in Problem States")


p<-ggplot(southEast,aes(x=southEast$delayGreaterthanFiveMinutes,
                        y=southEast$NPS, main="does departure delay affect NPS?"))+geom_count()+
  stat_summary(aes(y=southEast$NPS ,group=1), fun.y=mean, colour="blue", geom="point",group=1)

p + ggtitle("Affect on Delay time and NPS in Problem States")


dd<-ggplot(airlineDF, aes(x = airlineDF$Arrival.Delay.in.Minutes, y = ..count..)) +
geom_bar(aes(fill=NPS=="Detractor"),position = "identity")
dd<-dd+xlim(0,100)
dd<-dd+ylim(0,100)
dd+ggtitle("Late Arrival vs. Net Promoter Scores")
aa<-ggplot(airlineDF, aes(x = airlineDF$Departure.Delay.in.Minutes, y = ..count..)) +
geom_bar(aes(fill=NPS=="Detractor"),position = "identity")
aa<-aa+xlim(0,100)
aa<-aa+ylim(0,100)
aa+ggtitle("Late Departure vs. Net Promoter Scores")

dd<-ggplot(southEast, aes(x = southEast$Arrival.Delay.in.Minutes, y = ..count..)) +
geom_bar(aes(fill=NPS=="Promoter"),position = "identity")
dd<-dd+xlim(0,50)
dd<-dd+ylim(0,18)
dd+ggtitle("Late Arrival vs. Net Promoter Scores for SouthEast Airlines")
aa<-ggplot(southEast, aes(x = southEast$Departure.Delay.in.Minutes, y = ..count..)) +
geom_bar(aes(fill=NPS=="Promoter"),position = "identity")
aa<-aa+xlim(0,50)
aa<-aa+ylim(0,18)
aa+ggtitle("Late Departure vs. Net Promoter Scores for SouthEast Airlines")


dd<-ggplot(airlineDF, aes(x = airlineDF$Flights.Per.Year, y = ..count..)) + geom_bar(aes(fill=NPS),position = "identity")

dd+ggtitle("Do Number of Flights per Year Affect your NPS?")

table(airlineDF$Class, airlineDF$Likelihood.to.recommend)
summary(airlineDF$Flight.Distance)
```

```
p1<-ggplot(airlineDF,aes(x=airlineDF$Type.of.Travel,
                         y=airlineDF$Likelihood.to.recommend))+geom_count()

p1 + ggtitle("Does Type of Travel affect Liklihood to Recommend?")


ok<-ggplot(airlineDF,aes(x=airlineDF$Airline.Status,y=airlineDF$Likelihood.to.recommend))+geom_count()+
  stat_summary(aes(y =airlineDF$Likelihood.to.recommend ,group=1), fun.y=mean, colour="blue", geom="line",group=1)
ok+ggtitle("Customer Airline Status versus Likelihood to Recommend")

ggplot(data_a,aes(x=airline_status,y=satisfaction))+geom_count()+
  stat_summary(aes(y =data_a$satisfaction ,group=1), fun.y=mean, colour="red", geom="line",group=1)


ff<-ggplot(airlineDF, aes(x=airlineDF$Flights.Per.Year, y=airlineDF$Likelihood.to.recommend)) +
stat_summary(fun.y="mean", geom="line")
ff+ggtitle("Number of Flights Per Year Vs. Likelihood to Recommend")
```

## Map Visualizations

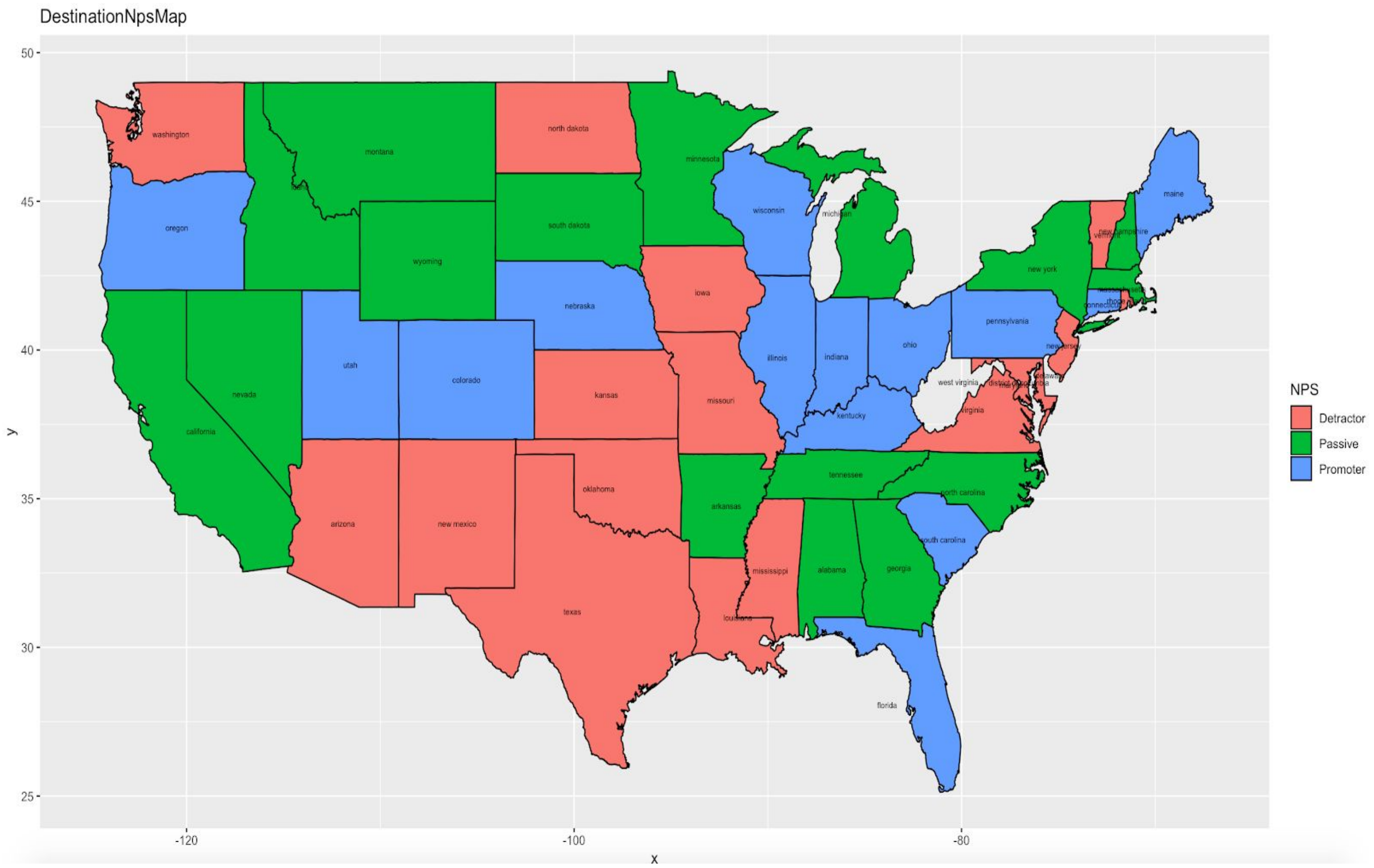
A brief study focusing on Promoter Category by state was done in order to give a greater focus to airport specific experience
.

### Figure 10: Origin State by Promoter Category



This map shows which states most often fell in a specific promoter category across all partner airlines related to the origin of their flights.

## Figure 11: Destination State by Promoter Category



This map shows how a state fell related to promoter scores according to a flight's destination across all partner airlines.

## Interpretation and Analysis: Maps and States

From our previous work, we can see that a customer's airport experience has a clear impact on their Net Promoter Score. For instance, money spent at the airport, whether shopping or dining, is a good indicator of a promoter. Therefore we needed to identify our standout airports, and our airports which need improvement.

Unfortunately we did not receive airport specific data, only state specific data. As such we are assuming that each state has only one airport (unlikely except in rural areas) or that the experience is representative across state airports. Until more granular data can be collected, we will examine this information at a state level.

We also are focusing on the best airports and the worst airports. Therefore it is important to note the following states:

States which were **majority detractors** for Departure and Arrival:

- New Mexico
- North Dakota
- Oklahoma
- Rhode Island
- Texas

States which were **majority promoters** for Departure and Arrival:

- Oregon
- Illinois
- Indiana
- Kentucky
- Maine

We believe it is in the best interest of the company to travel to those airports which produce a majority of promoters and study them. Through further collection of data, we can discover how these airports create a positive experience for our customers, and use them to create strategic plans which may improve the airports where our customers have more negative experiences. By learning from our best, we can improve our worst airports, and create a drastic improvement in our customers' experiences.

# Code:

```
################################################################################

##MAP

install.packages("maps")

library(maps)

state <- map_data("state")

airlineDF$Origin.State<-tolower(airlineDF$Origin.State)

airlineDF$Destination.State<-tolower(airlineDF$Destination.State)


StateNames <- aggregate(cbind(long, lat) ~ region, data=state,

                    FUN=function(x)mean(range(x)))

View(StateNames)


map1<-ggplot(airlineDF)+


  expand_limits(x=state$long,y=state$lat)+

  geom_map( map=state,aes(map_id=Origin.State,fill=NPS),color="black")+

#draw a map with NPS filled in each state

  geom_text(data = StateNames,aes(x=long,y=lat,label=region),size=2)+

# add state names on the map

  ggtitle("OriginNpsMap")

map1

#the distribution of NPS at each origin states

map2<-ggplot(airlineDF)+

 expand_limits(x=us$long,y=us$lat)+
  geom_map( map=us,aes(map_id=Destination.State,fill=NPS),color="black")+

  geom_text(data = StateNames,aes(x=long,y=lat,label=region),size=2)+

  ggtitle("DestinationNpsMap")

map2

#the distribution of NPS at each destination state
```

# Modeling Techniques

To explore how to reduce customer churn in Southeast Airlines, we conducted various modelling analyses to identify important indicators that influence customers' loyalty. We used a linear regression model, a logistic regression model and an SVM model, respectively.

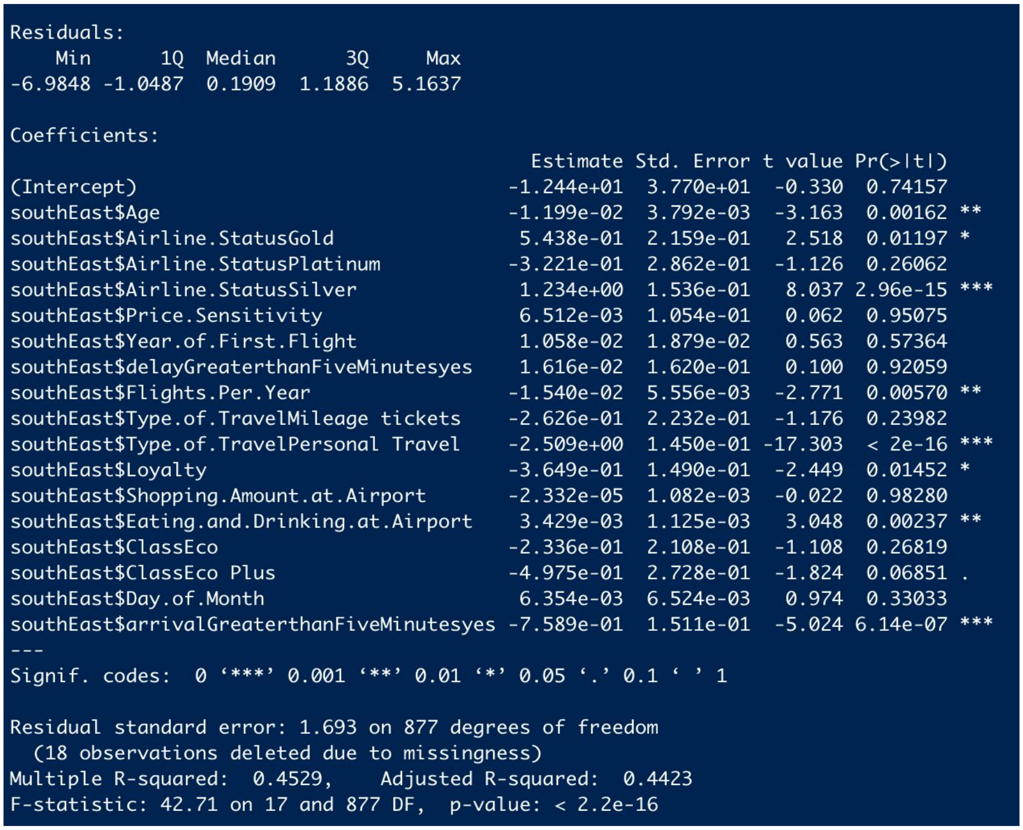
## Linear Regression and Logistic Regression

This study developed two types of regression models utilizing linear regression and logistic regression, that separately predicted the Net Promoter Score (NPS).

First, by applying linear regression model, we summarized the correlations between predictors and NPS among continuous variables. As shown in the following figure, among Southeast airlines' customers, *airline flyer status, type of travel, arrival delay time,* and *spendings on shopping at the airport* remained significantly correlated with NPS.

Our results showed that customers who have frequently taken this airline instead of other airlines would be less likely to recommend it to their friends, which was consistent with our previous finding that loyalty programs for frequent flyers have devalued this airline. Customers who spend more money on food/drink at the airport are more likely to recommend this airline. In addition, customers who have silver or gold airline status tended to have greater intention to recommend the airline to their friends. In contrast, people traveling for personal reasons (such as vacation) or who have experienced a flight arrival delay of more than five minutes are less likely to recommend the airline. It may be of note that customers who are traveling for personal reasons may have higher expectations for their experience than those who have taken the flight for business purposes.

Interestingly, we also found that arrival delay matters to NPS, whereas departure delay did not affect the overall ratings of the airline at all. In other words, people were more tolerant of flight departure delay than arrival delay. Overall, the findings suggested that those indicators jointly played important roles in predicting the likelihood of recommendation of customers in Southeast Airlines, which explained 44.23% of the variance.



```
Residuals:
    Min      1Q  Median      3Q     Max 
-6.9848 -1.0487  0.1909  1.1886  5.1637 

Coefficients:
                                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                                -1.244e+01  3.770e+01  -0.330  0.74157    
southEast$Age                              -1.199e-02  3.792e-03  -3.163  0.00162 ** 
southEast$Airline.StatusGold                5.438e-01  2.159e-01   2.518  0.01197 *  
southEast$Airline.StatusPlatinum           -3.221e-01  2.862e-01  -1.126  0.26062    
southEast$Airline.StatusSilver              1.234e+00  1.536e-01   8.037 2.96e-15 ***
southEast$Price.Sensitivity                 6.512e-03  1.054e-01   0.062  0.95075    
southEast$Year.of.First.Flight              1.058e-02  1.879e-02   0.563  0.57364    
southEast$delayGreaterthanFiveMinutesyes    1.616e-02  1.620e-01   0.100  0.92059    
southEast$Flights.Per.Year                 -1.540e-02  5.556e-03  -2.771  0.00570 ** 
southEast$Type.of.TravelMileage tickets    -2.626e-01  2.232e-01  -1.176  0.23982    
southEast$Type.of.TravelPersonal Travel    -2.509e+00  1.450e-01 -17.303  < 2e-16 ***
southEast$Loyalty                          -3.649e-01  1.490e-01  -2.449  0.01452 *  
southEast$Shopping.Amount.at.Airport       -2.332e-05  1.082e-03  -0.022  0.98280    
southEast$Eating.and.Drinking.at.Airport    3.429e-03  1.125e-03   3.048  0.00237 ** 
southEast$ClassEco                         -2.336e-01  2.108e-01  -1.108  0.26819    
southEast$ClassEco Plus                    -4.975e-01  2.728e-01  -1.824  0.06851 .  
southEast$Day.of.Month                      6.354e-03  6.524e-03   0.974  0.33033    
southEast$arrivalGreaterthanFiveMinutesyes -7.589e-01  1.511e-01  -5.024 6.14e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.693 on 877 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0.4529,	Adjusted R-squared:  0.4423 
F-statistic: 42.71 on 17 and 877 DF,  p-value: < 2.2e-16
```

Interpretation of the intercept (Non-economic):

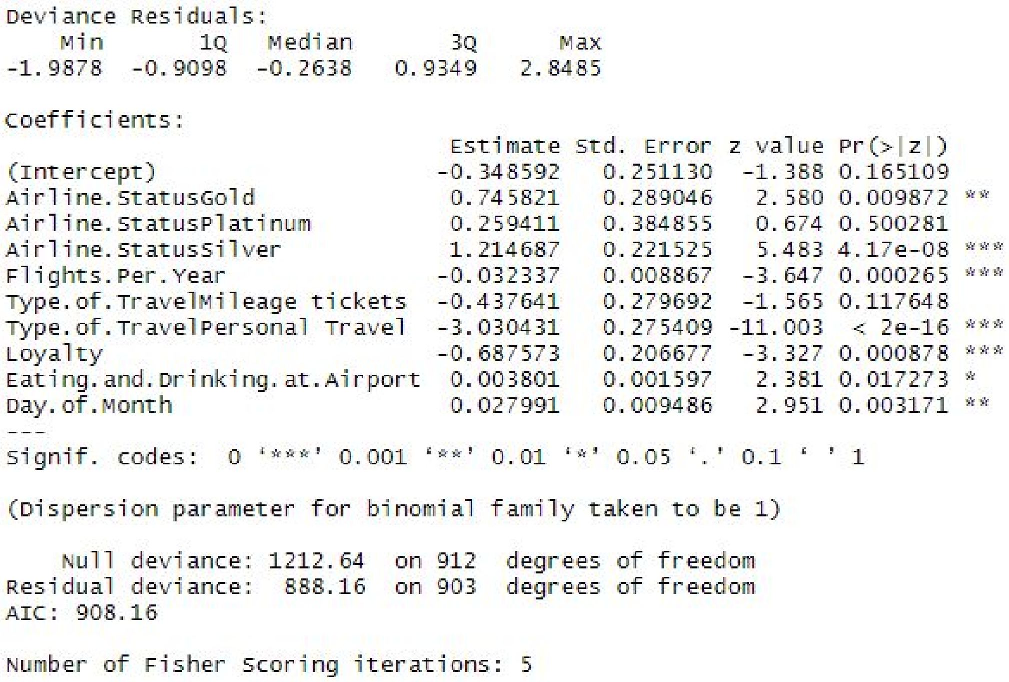
Within the theoretical and impossible scenario of age 0, Blue status, 0 flights per year, business traveller, loyalty score of 0, $0.00 of eating and drinking at the airport and arrival delay was less than 5 minutes - the NPS is predicted to be -12.44.

Economic interpretation of the significant explanatory variables:

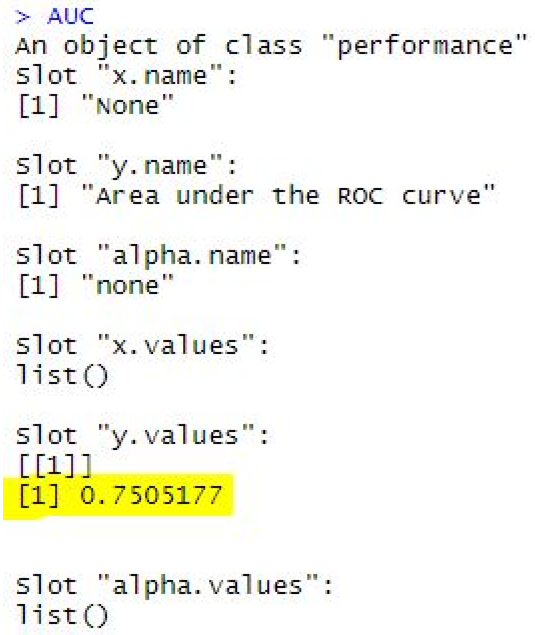
- Every additional one year in a passenger’s age is predicted to decrease the NPS by 0.01, holding all other independent variables constant.
- A Silver-status passenger is predicted to have a higher NPS by 1.23 compared to a Blue-status passenger, holding all other independent variables constant.
- A Gold-status passenger is predicted to have a higher NPS by 0.54 compared to a Blue-Status passenger, holding all other independent variables constant.
- A Platinum-status passenger is predicted to have a lower NPS by 0.32 compared to a Blue-Status passenger, holding all other independent variables constant.

- Every additional passenger's flight per year is predicted to decrease the NPS by 0.02, holding all other independent variables constant.
- A mileage traveller is predicted to have a lower NPS by 0.26 compared to a business traveller, holding all other independent variables constant.
- A personal traveller is predicted to have a lower NPS by 2.51 compared to a business traveller, holding all other independent variables constant.
- Every addition of 1 point in the loyalty score is predicted to decrease the NPS by 0.36, holding all other independent variables constant.
- Every additional $1 spent on eating and drinking on airport is predicted to increase the NPS by 0.003, holding all other independent variables constant.
- A passenger on a flight which had a delay on its arrival of more than five minutes is predicted to decrease the NPS by 0.76 compared to a passenger on a flight which had a delay on its arrival of less than five minutes, holding all other independent variables constant.

Second, by applying a logistic regression model, we are able to examine what are the strongest predictors for increasing the probability that a passenger is a promoter (Likelihood to recommend >= 9). Being particularly interested in the similar results between linear regression model and logistic regression model, we defined a new, dichotomous variable which derives from the likelihood to recommend variable, where a score which equals to 9 or 10 is converted to 1, otherwise 0. Consistent with the results we found in linear model, we also found that customers who have a silver airline status and spend more money on eating and drinking at the airport are more likely to be promoters. In contrast, people who have taken a greater number of flights in the past 12 months, who have taken flight for personal travel purposes, and who have experienced a flight arrival delay of more than five minutes are less likely to be promoters. The AUC measurement for the final logistic model is 75.05%, which is considered a fair prediction power.



```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9878  -0.9098  -0.2638   0.9349   2.8485

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.348592   0.251130  -1.388 0.165109
Airline.StatusGold            0.745821   0.289046   2.580 0.009872 **
Airline.StatusPlatinum        0.259411   0.384855   0.674 0.500281
Airline.StatusSilver          1.214687   0.221525   5.483 4.17e-08 ***
Flights.Per.Year             -0.032337   0.008867  -3.647 0.000265 ***
Type.of.TravelMileage tickets -0.437641   0.279692  -1.565 0.117648
Type.of.TravelPersonal Travel -3.030431   0.275409 -11.003  < 2e-16 ***
Loyalty                      -0.687573   0.206677  -3.327 0.000878 ***
Eating.and.Drinking.at.Airport  0.003801   0.001597   2.381 0.017273 *
Day.of.Month                  0.027991   0.009486   2.951 0.003171 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1212.64  on 912  degrees of freedom
Residual deviance:  888.16  on 903  degrees of freedom
AIC: 908.16

Number of Fisher Scoring iterations: 5
```



```
> AUC
An object of class "performance"
Slot "x.name":
[1] "None"

Slot "y.name":
[1] "Area under the ROC curve"

Slot "alpha.name":
[1] "none"

Slot "x.values":
list()

Slot "y.values":
[[1]]
[1] 0.7505177


Slot "alpha.values":
list()
```

Interpretation of the Intercept:

Within the theoretical and impossible scenario of Blue status, 0 flights per year, business traveller, loyalty score of 0, $0.00 of eating and drinking at the airport and day ‘0’ in a month - the log-odds for being a promoter is predicted to be -0.35.

Interpretation of the significant explanatory variables:

- A Silver-status passenger is predicted to have a higher log-odds to be a promoter by 1.21 compared to a Blue-status passenger, holding all other independent variables constant.
- A Gold-status passenger is predicted to have a higher log-odds to be a promoter by 0.75 compared to a Blue-status passenger, holding all other independent variables constant.
- A platinum-status passenger is predicted to have a higher log-odds to be a promoter by 0.26 compared to a Blue-status passenger, holding all other independent variables constant.
- Every passenger's additional flight per year is predicted to decrease the log-odds to be a promoter by 0.03, holding all other independent variables constant.
- A mileage traveller is predicted to have a lower log-odds to be a promoter by 0.44 compared to a business traveller, holding all other independent variables constant.
- A personal traveller is predicted to have a lower log-odds to be a promoter by 3.03 compared to a business traveller, holding all other independent variables constant.
- Every addition of 1 point in the loyalty score is predicted to decrease the log-odds to be a promoter by 0.69, holding all other independent variables constant.
- Every additional $1 spent on eating and drinking on airport is predicted to increase the log-odds to be a promoter by 0.004, holding all other independent variables constant.
- Every additional day in a month is predicted to increase the log-odds to be a promoter by 0.03, holding all other independent variables constant.

(Linear/Logistic Regression with charts and code)

Calculating accuracy of each model using testing data
Graphs of model acc/false positives

## Support Vector Machine Model

Finally, we utilized an SVM with the kernlab package to predict customer satisfaction based on statistically significant variables from our linear and logistic regression models. A new column in our dataframe needed to be created in order to determine whether or not the customer was "satisfied".

Scores greater than or equal to 9 were denoted with a 1 and others with a 0. Scores of 9 and 10 are the airline promoters which is why we used this breakpoint. Binary classification helps simplify our model to determine how well these attributes can predict a promoter.

## Code:

```
library(kernlab)
library(caret)
set.seed(2154)
airlineDF$sat<-cut(airlineDF$Likelihood.to.recommend, breaks = c(0,9,10), labels=c(0,1))

#Initially, our data was split into a training set that consisted of 70% of the airline dataframe and a test set for rest.

train<-createDataPartition(y=airlineDF$sat,p=0.7, list=FALSE)
training<-airlineDF[train,]
testing<-airlineDF[-train,]

#Training the svm:
myKsvm<-ksvm(sat ~Eating.and.Drinking.at.Airport+
             Type.of.Travel+
             Loyalty+
             Flights.Per.Year+
             Airline.Status,
             data=training, kernel = "rbfdot",kpar="automatic", C=5,cross=3, prob.model=TRUE)
myKsvm
```

Output:

```
SV type: C-svc  (classification)
 parameter : cost C = 5

Gaussian Radial Basis kernel function.
 Hyperparameter : sigma =  0.289899885072513
```

```
Number of Support Vectors : 2224

Objective Function Value : -10053.53
Training error : 0.130157
Cross validation error : 0.139463
Probability model included.
```

Creating a histogram of the SVM model:

```
hist(alpha(myKsvm)[[1]], main="support vector histogram C=5, cross=3",
    xlab="support values")
```

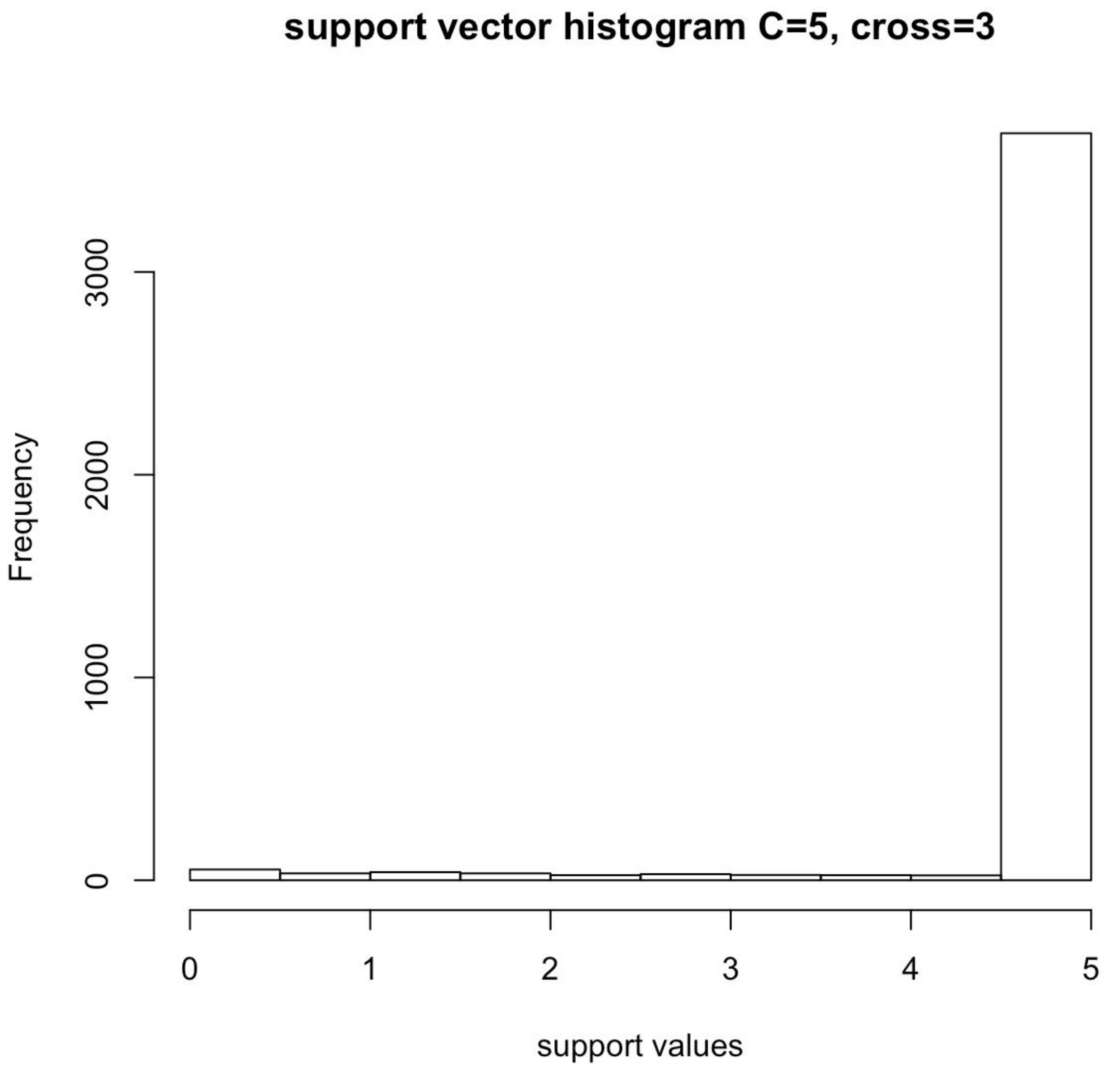


Figure 14: Support Vector Machine Histogram

SVM Prediction on testing dataset

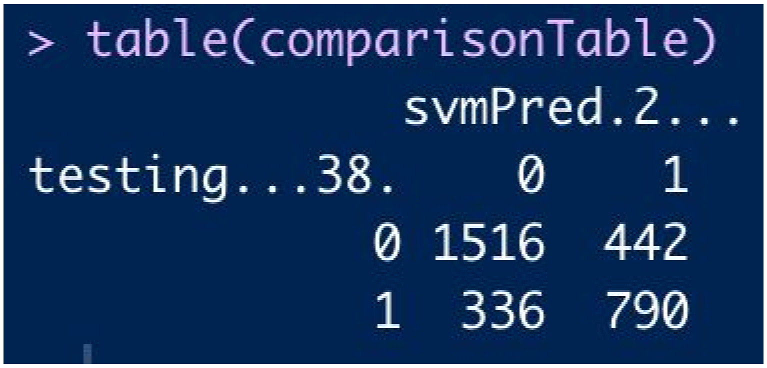
svmPred <- predict(myKsvm, testing, type = "votes")

svmPred

Creating a confusion matrix to see how well model is predicting satisfied customers.

comparisonTable <- data.frame(testing[ ,38], svmPred[2, ])

table(comparisonTable)



```
> table(comparisonTable)
               svmPred.2...
testing...38.     0    1
              0 1516  442
              1  336  790
```

Figure 15: SVM Confusion Matrix of prediction vs. ground truth predictions

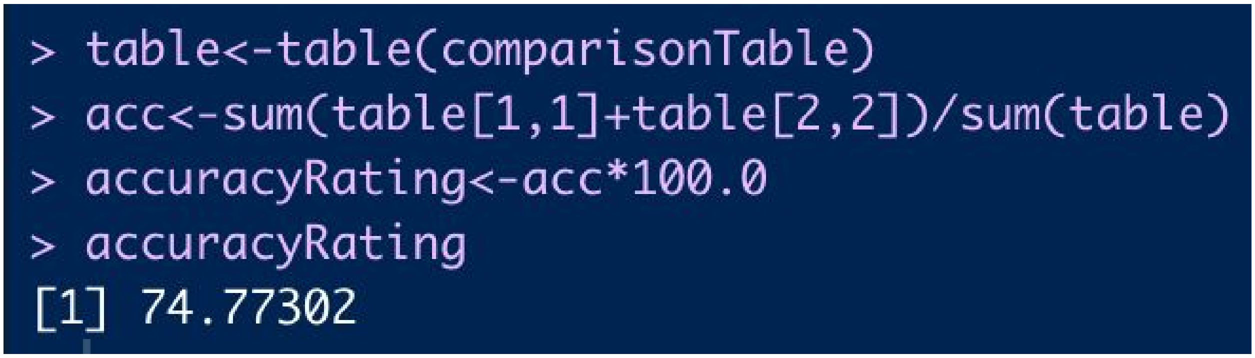
table<-table(comparisonTable)

Creating an accuracy rating based on the confusion matrix:

acc<-sum(table[1,1]+table[2,2])/sum(table)

accuracyRating<-acc*100.0

accuracyRating



```
> table<-table(comparisonTable)
> acc<-sum(table[1,1]+table[2,2])/sum(table)
> accuracyRating<-acc*100.0
> accuracyRating
[1] 74.77302
```

Figure 16: SVM Accuracy

## Business Questions and Answers

The goal of this project is to reduce customer churn by identifying the most important predictors of high and low customer ratings. The following business questions would provide us reasonable suggestions toward how to keep a customer in Southeast Airlines. Several key questions were identified and answered through the project:

- Do customers from different age groups have a different likelihood of recommending Southeast Airlines?
  - We found that emerging adults and middle aged customers were most likely to be promoters. More mature customers were most likely to be detractors.
- How are customers experiences impacted by the airports they are traveling to and from? How does this manifest on a state level?
  - We found that a customer's experience was greatly affected by their time in an airport. Some states provide a more pleasant experience, while others like Texas need drastic improvement.
- Do customers who experience flight departure/arrival delays tend to have a lower likelihood of recommendation?
  - Customers are much more likely to be promoters when there flight delay is five minutes or less.
- Does type of travel have an impact on customers' likelihood of recommendation?

    - Those customers who were traveling for business were more likely to be promoters than passengers traveling for personal reasons.
- Do customers with different airline status members show differences in the NPS category?
    - Those customers who were above blue status were more likely to be promoters for Southeast Airlines
- Do numbers of flights per year influence customers’ likelihood of recommendation?
    - As the number of flights per year increased, likelihood to recommend decreased.

# Conclusions and Takeaways

- Departure delay of 5 min or more only slightly affects satisfaction ratings whereas arrival delay greater or equal to five minutes significantly reduces the score
    - **→People want to arrive on time! Stick to the schedule**

- Certain partner airlines are hurting Southeast Airlines image. Flyfast Airlines and Cheapseats Airlines have consistently low Net Promoter Scores
    - **→Discontinue contracts with these airlines. We may lose some flight availability, but customers will think more of our airline experience**
- Age affects satisfaction ratings for all airlines. Middle aged flyers reported the highest scores, however it was much lower amongst older flyers
    - **→ Work towards providing greater commodities for mature customers. This demographic is only going to increase as the population ages.**

- When customers travel for personal reasons, their satisfaction is the lowest
    - **→ Common Folk have Higher Expectations, we need to study this group more and understand why their experiences differ from business travelers.**

- For Southeast airlines, Likelihood to recommend increases as the number of yearly flights decreases
    - **→ Make sure seasoned flyers are being rewarded. The current reward program does not work. Consider free seat upgrades for frequent flyers.**

- Blue and Platinum members report the lowest Recommendation scores, whereas Gold and Silver members tend to give much higher ratings.
    - **→ Subset Eco flyers and analyze flight notes to determine what could've been better**

- A majority of the Detractors in our study were flying around the midwest
    - **→ Midwest airports have less infrastructure than major airports, what can we do to improve our midwest experience?**