Airline Survey Analysis

Cool&Young Airlines Inc.

Sammy Pardes
IST687 M400
Final Project
12/15/19

Project Summary

Cool&Young Airlines Inc. Airlines Inc. recently participated in a multi-airline initiative to survey of over 125,000 passengers to gain insight on overall client satisfaction. The survey took place over 3 months, from January 1, 2014 through March 31, 2014.

Of the over 125,000 participants from 14 different airlines, 1,288 of the sample flew on Cool&Young Airlines Inc. Each passenger was asked to rate their experience on a scale from 1 to 5, where 1 is the least satisfied and 5 is the most satisfied. As the Cool&Young Airlines Inc. consultant, I have been tasked with analyzing the survey responses to determine the strengths and weaknesses of the Cool&Young Airlines experience.

Information gathered in the survey includes personal attributes, travel attributes, and flight attributes. To understand trends in client Satisfaction, some of the personal attributes I analyzed include Age, Gender, Percent of Flights with other Airlines, and Total Number of Flights Taken. The travel attributes I have considered include Airline Status, Type of Travel, Class, and Origin/Departure City. I also took into account flight attributes such as Arrival and Departure Delay.

Business Questions and Assumptions

- **Assumption 1:** Everyone who participated in the Airline survey filled it out as honestly and accurately as possible.
- **Assumption 2:** The passengers sampled for this survey provide a good representation of the total client base not only for Cool&Young Airlines Inc., but for all participating airlines.
- **Assumption 3:** Using the data provided, insight can be gained on how the passenger experience can be improved.
- **Assumption 4:** Enough data has been collected to confidentially make informed business decisions.

Data Questions

1. **Which cities/states have the largest impact on satisfaction? Do these cities have higher/lower delays?**
    a. Identify airports that need improvement, and whether that needs to be in the airport experience or with the flight trafficking
        i. Satisfaction Score
        ii. Origin City
        iii. Origin State
        iv. Arrival City
        v. Arrival State
        vi. Departure Delay in Minutes
        vii. Arrival Delay in Minutes
2. **How are airlines performing across different types of customers?**
    a. Compare different demographics and their respective satisfaction scores to determine who each airline is preferable with and which targets they are missing.

   i. Gender
   ii. Age
   iii. Class
   iv. Satisfaction
   v. Type of Travel

3. **Who are the most loyal customers? Who are the least loyal? Does loyalty align with satisfaction?**
   a. If an airline already has a loyal customer demographic, we could offer additional rewards for their continued support.
   b. For those who are not loyal, an airline could offer promotions to gain their patronage.
   c. We must determine what makes a customer most/least loyal. (ex: smallest % of flights with other airlines.)
      i. Airline Status
      ii. Age
      iii. Gender
      iv. Status
      v. % of Flight with other Airlines
4. **What variables have the strongest impact on Satisfaction score?**
   a. Find out what is driving customer satisfaction.
   b. Identify areas to prioritize to improve overall customer satisfaction?
      i. Satisfaction score
      ii. All other variables

Data Cleaning

The data was originally presented as a .csv file. After reading the file into R, I discovered the only columns missing data were Arrival Delay in Minutes and Flight Time in Minutes.

Since a significant number of rows were missing data from one or both fields, and to avoid removing those responses all together, I adjusted the data where NAs were present.

Where Arrival Delay in Minutes was left black, I assumed an on-time arrival occurred or the flight was cancelled. In both cases, the Arrival Delay was adjusted to be 0.

Where Flight Time in Minutes was NA, we assumed that the flight was cancelled and therefore, replaced the missing data with a 0 to indicate that the flight did not occur. Since flights that took place cannot be 0 minutes long, where the fight duration was left blank, I replaced the NA values with zeros.

To further ready to data for analysis, I ensured that all the state names were lowercase. Since I wanted to use the state.name data frame to incorporate regional information, the survey data needed to match the state.names data.

Finally, I removed all the periods from the column names. To use the SQLDF package to run SQL code on the survey data, I needed to remove all the punctuation in the column names. Unlike R, SQL uses the period to denote column names within a table.

Data Analysis

The first step in analyzing the results of the airline survey was to determine how Cool&Young Airlines Inc. ranks in relation to the 13 other participating airlines. On average, Cool&Young airlines ranked second among its competitors.

	Satisfaction	AirlineName	AirlineCode
8	3.486967	West Airways Inc.	HA
13	3.442547	Cool&Young Airlines Inc.	VX
2	3.425301	FlyToSun Airlines Inc.	AS
1	3.399167	Paul Smith Airlines Inc.	AA
4	3.397547	Sigma Airlines Inc.	DL
12	3.396888	Southeast Airlines Co.	US
7	3.395002	FlyHere Airways	FL
10	3.394798	Northwest Business Airlines Inc.	OO
11	3.386534	Oursin Airlines Inc.	OU
9	3.360199	EnjoyFlying Air Services	MQ
14	3.357318	Cheapseats Airlines Inc.	WN
5	3.352567	FlyFast Airways Inc.	EV
3	3.346803	OnlyJets Airlines Inc.	B6
6	3.297194	GoingNorth Airlines Inc.	F9

Next, I analyzed the survey results for only Cool&Young Airlines Inc. passengers. This included 1,288 participants.

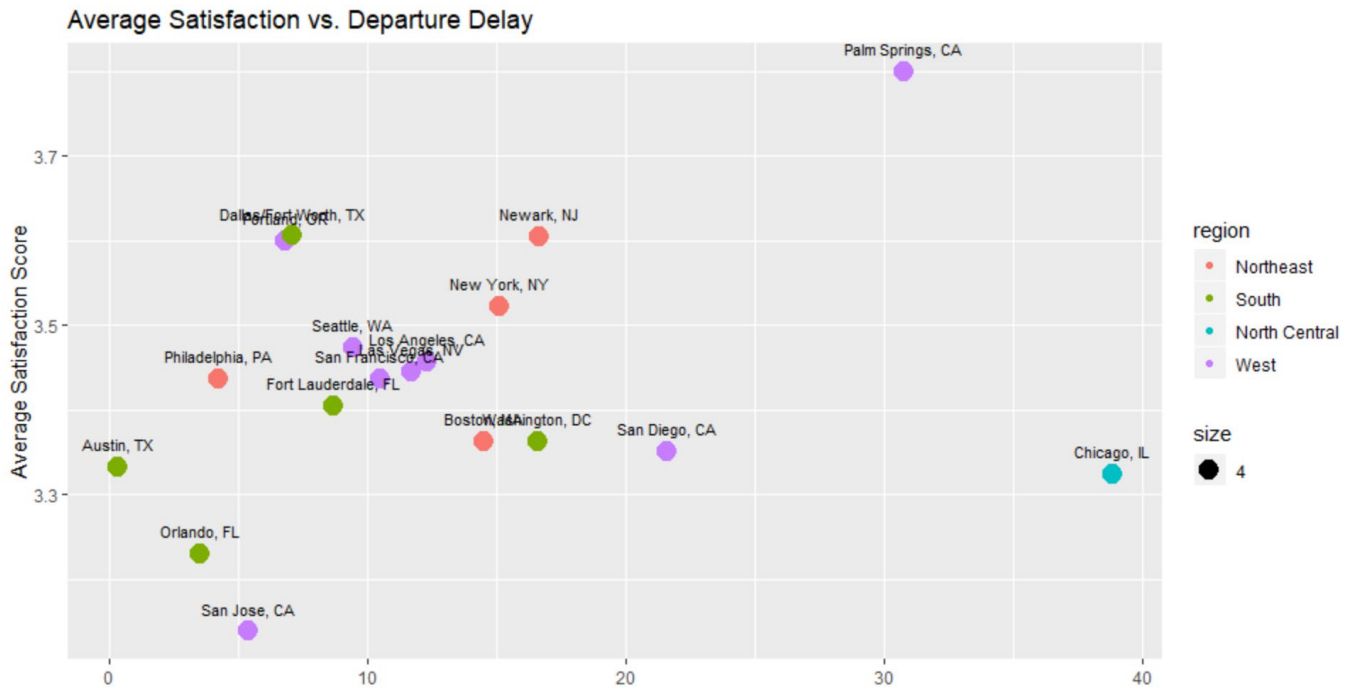
Which cities/states have the largest impact on satisfaction? Do these cities have higher/lower delays?

To determine the cities that have the largest impact on satisfaction score, I first sorted the Cool&Young Airlines Inc. dataset by average satisfaction score based on Origin City.

Next, I found the cities with the highest delay times to see if we could find any correlation between delay times and customer satisfaction. The top 5 cities with the highest Arrival Delays and highest Departure Delays are the same. Of the 18 cities represented, Chicago and San Diego appear in the top 5 for most delayed and least satisfied cities.

	ratingByCity.OriginCity.1.5.	delaysByCity.OriginCity.1.5.
1	San Jose, CA	Chicago, IL
2	Orlando, FL	Palm Springs, CA
3	Chicago, IL	San Diego, CA
4	Austin, TX	Newark, NJ
5	San Diego, CA	Washington, DC

Next, I combined the survey data with the state.names data frame to plot Average Departure Delay vs. Average Satisfaction, sorted by City and Region.



Here, you can see that the West region is the most represented within Cool&Young Airlines Inc. Cities in the West include: Los Angeles, San Jose, Palm Springs, San Diego, San Francisco, Las Vegas, Portland, and Seattle.

Unlike I initially predicted, there does not seem to be any kind of correlation between minutes delayed and customer satisfaction.

Similarly, the average number of cancellations did not seem to sway satisfaction.

How are airlines performing across different types of customers?

When comparing the Status of customers, there is a significant difference between the satisfaction of those who are Blue compared to those who are Platinum, Silver, or Gold.

```
AvgRating   Status
1  3.227624     Blue
2  3.923913     Gold
3  4.048780 Platinum
4  3.944000   Silver
```

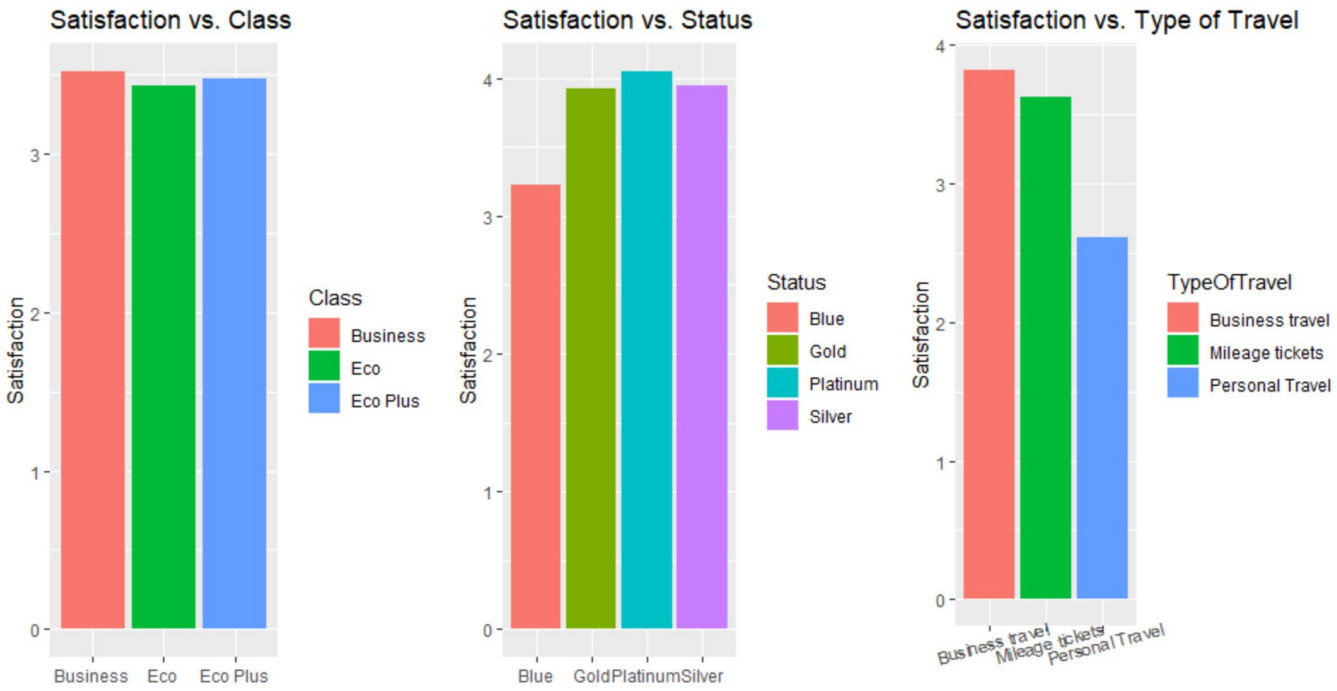
In a similar vein, although not as varying, customer in Business class are marginally more satisfied, than their Eco Plus and Eco flight-mates.

```
AvgRating    Class
1  3.521008 Business
2  3.429933      Eco
3  3.475000 Eco Plus
```

Employing the same procedure that I used to compare Class and Status, I discovered passengers who travelled for Business were much more satisfied than those who's travel type was Personal.

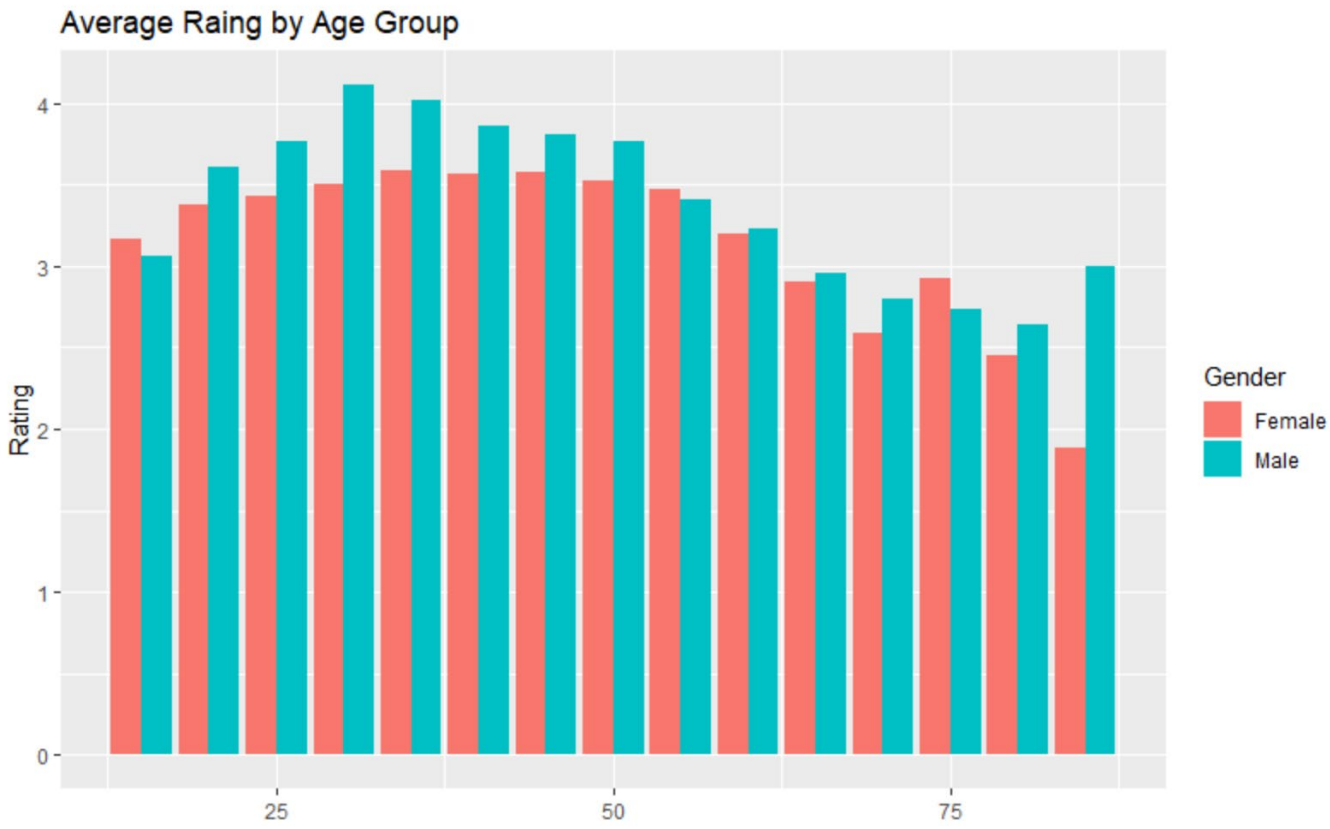
```
AvgRating    TypeOfTravel
1  3.816121 Business travel
2  3.619469 Mileage tickets
3  2.611549 Personal Travel
```

Below, I combined the Status, Class, and Travel Type plots to give an overview of how satisfied passengers are, classified by each respective attribute. While the difference in satisfaction between passengers in different classes is not very substantial, the different in Status and Travel type is significant.



To compare satisfaction by Age and Gender, I combined the information into one plot. I calculated the average satisfaction of passengers grouped by Age in intervals of 5 years and then by Gender.

In almost every age group, males are significantly more satisfied than their female contemporaries. Independent of gender, passengers who are between the ages of 30-50 are the most satisfied with Cool&Young Airlines Inc. The least satisfied passengers are women over 85.



Who are the most loyal customers? Who are the least loyal? Does loyalty align with satisfaction?

To determine which customers are the most and least loyal, we must first define loyalty. I decided to define loyalty by the percentage of flights passengers have taken on airlines other than Cool&Young compared to their total number of flights. Those with the lowest percentage are considered the most loyal.

Some of our data included passengers who took zero flights. Presumably, these passengers had their flights cancelled. I removed these passengers from consideration when calculating loyalty. This left us with 1,234 rows of data to evaluate.

Women who fly on Cool&Young Airlines are much less loyal than men.

```
AvgLoyalty Gender
1  1.1475500 Female
2  0.8516448   Male
```

As shown in the table below, Business class passengers are the most loyal followed by Eco Plus, and lastly, Eco.

```
AvgLoyalty    Class
1  0.8780523 Business
2  1.0264822      Eco
3  0.9813352 Eco Plus
```

While, class and Gender seemed to replicate the results of our Satisfaction evaluation, Interestingly, Status rendered the opposite result. Platinum passengers were far less loyal than Blue.

```
AvgLoyalty   Status
1  0.9547865     Blue
2  1.2373694     Gold
3  1.5717837 Platinum
4  1.0355013   Silver
```

What variables have the strongest impact on Satisfaction score?

I decided to use linear regression to determine which variables have the greatest impact on predicting Satisfaction. I tried several variations of the regression to generate the highest R-squared value and the lowest p-value. What yielded the best result with the fewest number of columns was the combination of Travel Type and Status. Each column is statistically significant and the R^2 value is close to 36%. Adding Gender and Age only improved the R^2 value by about 1%.

```
Call:
lm(formula = satisfactionNum ~ travelNum + statusNum, data = surveyCool)

Residuals:
    Min      1Q  Median      3Q     Max
-5.5960 -1.1264  0.4648  0.8750  6.8750

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.20996    0.13057   47.56   <2e-16 ***
travelNum   -1.20510    0.05460  -22.07   <2e-16 ***
statusNum    0.53038    0.04107   12.91   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.754 on 1285 degrees of freedom
Multiple R-squared:  0.3582,   Adjusted R-squared:  0.3572
F-statistic: 358.5 on 2 and 1285 DF,  p-value: < 2.2e-16
```

In addition to the linear model method, I also used KSVM and SVM model to anticipate passenger satisfaction.

```
TestData Prediction
1        4   5.898684
2        3   4.947411
3        3   5.656699
4        4   5.080738
5        3   3.856389
6        3   4.132905
```

The RMSE for the KSVM model I used to predict satisfaction using all other columns is 2.132004.

The RMSE for the SVM mode I used is 2.132854. Since the scale of our predicted variable is from 1-5, an error of ~2 means that these models are not great predictors of how satisfied our customers will be.

```
> rmse(surveyTest$satisfactionNum, surveyPredictSvm)
[1] 2.132854
> rmse(surveyTest$satisfactionNum, surveyPredict)
[1] 2.132004
```

Results Validation

I was able to use Microsoft SQL Server and Excel to replicate the basis of my calculations. Thus, I ensured the results of my findings are both accurate and perceptive. Additionally, I provided graphics for many of my calculations to confirm my results visually.

Recommendations

After my analysis on the Airline Survey data for Cool&Young Airlines Inc., I have gathered enough data to recommend that Cool&Young Airlines Inc. focus on their least satisfied passengers rather than the markets they have already captured.

Recommendation #1: Increase accessibility

As their name suggest, Cool&Young Airlines Inc. could significantly improve their relationship with their older passengers. While customers between the ages of 25 and 50 are the most content, there is a gradual drop-off in satisfaction with passengers over the age of 50. The least satisfied population based on age and gender is women over 80. I recommend that Cool&Young Airlines improve their accessibility and assistance available for senior travelers.

This could include having more wheelchairs available or perhaps an express lane for older passengers. Although it may be off-brand, it is important to keep all customers happy!

Recommendation #2: Incentivize Status upgrades

Considering the discrepancy in Satisfaction of Blue Status passengers compared to Gold, Silver, and Platinum passengers, Cool&Young Airlines Inc. should work to have fewer Blue passengers. To do this, Cool&Young Airlines Inc. could offer incentives for passengers to upgrade their status. These incentives could be priority seat selection, or free snacks.

Recommendation #3: Offer discounts for non-mileage flights

Since I have determined that passengers who fly for personal travel and don't use miles are less gratified, Cool&Young Airlines should cater to these travelers by offering cheaper fares for non-mileage bookings.

Recommendation #4: Reward passenger loyalty

Assuming Business Class and Platinum passengers are paying more for their tickets, we want them to keep coming back to Cool&Young Airlines Inc. To keep them coming back, Cool&Young Airlines Inc. should offer special accommodations repeat customers. This could include priority boarding, free beverages, or lounge access.

Code Used

```
#Sammy Pardes
#IST687
#Final Project


######load the data
survey <- read.csv("c:/Users/samantha.pardes/Desktop/graduate/IST687/final-project/SatisfactionSurvey2_2_2_2.csv")


######cleanse the data
head(survey)
```

```
str(survey)


any(is.na(survey)) #there are NAs in the data

sum(is.na(survey)) #there are 7,821 rows with NAs

colSums(is.na(survey)) #the only columns missing data are Arrival.Delay.in.Minutes,
Flight.time.in.minutes, and Departure.Delay.in.Minutes


#let's assume there were 0 minute delays where left blank

survey$Arrival.Delay.in.Minutes[is.na(survey$Arrival.Delay.in.Minutes)] <- 0

any(is.na(survey$Arrival.Delay.in.Minutes))


survey$Departure.Delay.in.Minutes[is.na(survey$Departure.Delay.in.Minutes)] <- 0

any(is.na(survey$Departure.Delay.in.Minutes))


#let's replace NA flight with 0 minutes

library(sqldf)


sqldf('SELECT "Flight.time.in.minutes", "Flight.cancelled" FROM survey WHERE
"Flight.cancelled"=="No" AND "Flight.time.in.minutes"==0')

#no rows where flight time is 0 and flight is not cancelled


NoFlightTime <- 0

survey$Flight.time.in.minutes[is.na(survey$Flight.time.in.minutes)] <- NoFlightTime

survey$Flight.time.in.minutes


#ensure state names are lowercase

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

survey$Origin.State <- tolower(survey$Origin.State)
```

```
#remove "." from column names to use SQL
names(survey)[names(survey) == "Origin.State"] <- "OriginState"
names(survey)[names(survey) == "Orgin.City"] <- "OriginCity" #fixes typo in col name
names(survey)[names(survey) == "Price.Sensitivity"] <- "PriceSensitivity"
names(survey)[names(survey) == "Year.of.First.Flight"] <- "YearOfFirstFlight"
names(survey)[names(survey) == "No.of.Flights.p.a."] <- "NoOfFlights"
names(survey)[names(survey) == "X..of.Flight.with.other.Airlines"] <-
"NumFlightsWithOtherAirlines"
names(survey)[names(survey) == "Type.of.Travel"] <- "TypeOfTravel"
names(survey)[names(survey) == "No..of.other.Loyalty.Cards"] <- "NumLoyaltyCards"
names(survey)[names(survey) == "Shopping.Amount.at.Airport"] <- "ShoppingAmtAtAirport"
names(survey)[names(survey) == "Eating.and.Drinking.at.Airport"] <- "EatDrinkAtAirport"
names(survey)[names(survey) == "Day.of.Month"] <- "DayOfMonth"
names(survey)[names(survey) == "Flight.date"] <- "FlightDate"
names(survey)[names(survey) == "Airline.Code"] <- "AirlineCode"
names(survey)[names(survey) == "Airline.Name"] <- "AirlineName"
names(survey)[names(survey) == "Scheduled.Departure.Hour"] <- "SchedDepartHr"
names(survey)[names(survey) == "Departure.Delay.in.Minutes"] <- "DepartDelayMin"
names(survey)[names(survey) == "Flight.cancelled"] <- "FlightCancelled"
names(survey)[names(survey) == "Flight.time.in.minutes"] <- "FlightTimeMin"
names(survey)[names(survey) == "Flight.Distance"] <- "FlightDist"
names(survey)[names(survey) == "Arrival.Delay.greater.5.Mins"] <- "ArrivalDelayOver5Min"
names(survey)[names(survey) == "Destination.City"] <- "DestinationCity"
names(survey)[names(survey) == "Destination.State "] <- "DestinationState"
names(survey)[names(survey) == "Arrival.Delay.in.Minutes"] <- "ArrDelayMin"
names(survey)[names(survey) == "Destination.State"] <- "DestState"
names(survey)[names(survey) == "Airline.Status"] <- "Status"
```

```
head(survey)

dim(survey)


#use Cool&Young Airlines Inc. responses only

any(is.na(survey$AirlineCode))


surveyCool <- sqldf('SELECT * FROM survey WHERE AirlineCode=="VX"')

head(surveyCool)

dim(surveyCool)

#1,288 responses for Cool&Young Airlines Inc.


######data is cleansed!!


#How do we compare?

ratingByAirline <- sqldf('SELECT AVG(Satisfaction) as Satisfaction, AirlineName, AirlineCOde
from survey GROUP BY AirlineCode')

ratingByAirline <- ratingByAirline[order(-ratingByAirline$Satisfaction),]

ratingByAirline

#Cool&Young Airlines Inc. ranks second among it's competitors for most satisfied clients


######business question 1: which cities/states have the largest impact on satisfaction? Do
these cities have higher/lower delays?

#find cities w/ lowest satisfaction scores

ratingByCity <- sqldf('SELECT AVG(surveyCool.Satisfaction)AS AvgScore, OriginCity, OriginState
FROM surveyCool GROUP BY OriginCity')

dim(ratingByCity) #18 cities represented
```

```
ratingByCity <- ratingByCity[order(ratingByCity$AvgScore),] #order from lowest to highest satisfaction

ratingByCity[1:5,]

#the 5 lowest ranking cities are:

#1) San Jose, CA 2.) Orlando, FL 3.) Chicago, IL 4.) Austin, TX 5.) San Diego, CA


#find max delay times

delaysByCity <- sqldf('SELECT AVG(surveyCool.DepartDelayMin)AS AvgDelay, OriginCity, OriginState FROM surveyCool GROUP BY OriginCity')

dim(delaysByCity)


delaysByCity <- delaysByCity[order(-delaysByCity$AvgDelay),]

delaysByCity[1:5,]

#the cities w/ the most departure delays are:

#1) Chicago, IL 2.) Palm Springs, CA 3.) San Diego, CA 4.) Newark, NJ 5.) Washington, DC


#compare low ranking cities to delayed cities

data.frame(ratingByCity$OriginCity[1:5],delaysByCity$OriginCity[1:5])

#Chicago and San Diego appear in the 5 lowest rated and most departure delays


#let's try arrival delays

arrDelaysByCity <- sqldf('SELECT AVG(surveyCool.ArrDelayMin)AS AvgArrDelay, OriginCity, OriginState FROM surveyCool GROUP BY OriginCity')

dim(arrDelaysByCity)


arrDelaysByCity <- arrDelaysByCity[order(-arrDelaysByCity$AvgArrDelay),]

arrDelaysByCity[1:5,]

#the cities w/ the most arrivial delays are:

#1.) Chicago, IL 2.) Palm Springs, CA 3.) San Diego, CA 4.) Washington, DC 5.) Newark, NJ
```

```
#same cities as most departure delays


#avg rating vs. avg dept. delays

RatingVsDelay <- sqldf('SELECT AVG(DepartDelayMin)AS AvgDelay, AVG(Satisfaction)AS
AvgRating, OriginCity, OriginState AS state FROM surveyCool GROUP BY OriginCity')

head(RatingVsDelay)


#plot findings

library(ggplot2)

library(ggmap)

myStates <- data.frame(tolower(state.name), state.region)

colnames(myStates) <- c('state', 'region')


RatingVsDelay <- merge(RatingVsDelay, myStates, by="state") #leaves out U.S. Pacific Trust
Territories


ggRatingDelay <- ggplot(RatingVsDelay, aes(x=AvgDelay, y=AvgRating)) + geom_point()

ggRatingDelay <- ggRatingDelay + xlab("Average Delay in Minutes") + ylab("Average
Satisfaction Score")

ggRatingDelay <- ggRatingDelay + aes(color=region, size=4) + ggtitle("Average Satisfaction vs.
Departure Delay")

ggRatingDelay <- ggRatingDelay + geom_text(aes(label=OriginCity),hjust=0.5, vjust=-1.3,
size=3, color="black")

ggRatingDelay


#cities in West region

RatingVsDelay[(RatingVsDelay$region=="West"),]


#cancellations

surveyCool$cancelledNum <- as.numeric(surveyCool$FlightCancelled)
```

```
RatingVsCancel <- sqldf('SELECT AVG(cancelledNum)AS AvgCancelled, AVG(Satisfaction)AS AvgRating, OriginCity, OriginState AS state FROM surveyCool GROUP BY OriginCity')

RatingVsCancel <- RatingVsCancel[order(-RatingVsCancel$AvgCancelled),]

head(RatingVsCancel)


######business question 2: how are airlines performing across different types of customers?

#status

ratingByStatus <- sqldf('SELECT AVG(Satisfaction)AS AvgRating, Status FROM surveyCool GROUP BY Status')

ratingByStatus

#most to least satisfied by status: platinum silver, gold, blue

#big difference between blue and gold


ratingbyStatusPlot <- ggplot(ratingByStatus, aes(x=Status, y=AvgRating)) + geom_histogram(stat="identity")

ratingbyStatusPlot <- ratingbyStatusPlot + aes(fill=Status) + xlab("Status") + ylab("Satisfaction") + ggtitle("Satisfaction vs. Status")

ratingbyStatusPlot


#class

ratingByClass <- sqldf('SELECT AVG(Satisfaction)AS AvgRating, Class FROM surveyCool GROUP BY Class')

ratingByClass

#Business class most satisfied, then Eco Plus, then Eco


ratingbyClassPlot <- ggplot(ratingByClass, aes(x=Class, y=AvgRating)) + geom_histogram(stat="identity")

ratingbyClassPlot <- ratingbyClassPlot + aes(fill=Class) + xlab("Class") + ylab("Satisfaction") + ggtitle("Satisfaction vs. Class")
```

```
ratingbyClassPlot


#age

ratingByAge <- sqldf('SELECT AVG(Satisfaction)AS AvgRating, Age FROM surveyCool GROUP BY Age')

ratingByAge <- ratingByAge[order(-ratingByAge$AvgRating),]

ratingByAge[1:10,]

#most satisfied ages: 43, 35, 40, 52, 51


ratingByAge <- ratingByAge[order(ratingByAge$AvgRating),]

ratingByAge[1:10,]

#least satisfied ages: 70, 85, 71, 80, 69


#gender

ratingByGender <- sqldf('SELECT AVG(Satisfaction)AS AvgRating, Gender FROM surveyCool GROUP BY Gender')

ratingByGender

#males are on average, more satisfied than females


#gender + age plot

agesGenders <- sqldf('SELECT AVG(Satisfaction) AS AvgRating, Gender, FLOOR((Age)/5)*5 AS AgeGroup FROM surveyCool GROUP BY AgeGroup, Gender')

agesGenders


ageGenderPlot <- ggplot(agesGenders, aes(AgeGroup, AvgRating, fill=Gender))

ageGenderPlot <- ageGenderPlot + geom_bar(stat = "identity", position = 'dodge')

ageGenderPlot <- ageGenderPlot + xlab("Age Group") + ylab("Rating") + ggtitle("Average Raing by Age Group")

ageGenderPlot
```

```
#type of travel

ratingByTravel <- sqldf('SELECT AVG(Satisfaction)AS AvgRating, TypeOfTravel FROM surveyCool
GROUP BY TypeOfTravel')

ratingByTravel

#business travelers most satisfied, then mileage travelers, personal travelers least satisfied


ratingbyTravelPlot <- ggplot(ratingByTravel, aes(x=TypeOfTravel, y=AvgRating)) +
geom_histogram(stat="identity")

ratingbyTravelPlot <- ratingbyTravelPlot + aes(fill=TypeOfTravel) + xlab("Type of Travel") +
ylab("Satisfaction") + ggtitle("Satisfaction vs. Type of Travel")

ratingbyTravelPlot <- ratingbyTravelPlot + theme(axis.text.x = element_text(angle=15))

ratingbyTravelPlot


#multiple plots

library(gridExtra)

ratingPlots <- grid.arrange(ratingbyClassPlot, ratingbyStatusPlot, ratingbyTravelPlot, nrow = 1)

ratingPlots


######business question 3: who are the most loyal customers within each airline? Does
loyalty align with satisfaction? Who are the least loyal?


#define loyalty: NumFlightsWithOtherAirlines/NoOfFlights (low % is more loyal)


#can't divide by 0. must exclude passengers w/ 0 # of flights

sqldf('SELECT COUNT(*) FROM surveyCool WHERE NoOfFlights ==0') #54 participants have taken
0 flights

sqldf('SELECT COUNT(*) FROM surveyCool WHERE NoOfFlights !=0') #1234 participants have
taken at least 1 flight
```

```
surveyCoolWFlights <- sqldf('SELECT * FROM surveyCool WHERE NoOfFlights !=0')

dim(surveyCoolWFlights)

dim(surveyCool)


#create loyal % column, 0% would be most loyal

head(sqldf('SELECT NumFlightsWithOtherAirlines, NoOfFlights, loyalty FROM
surveyCoolWFlights'))


surveyCoolWFlights$loyalty <-
(surveyCoolWFlights$NumFlightsWithOtherAirlines/surveyCoolWFlights$NoOfFlights)

head(surveyCoolWFlights$loyalty)


#gender

mostLoyalGender <- sqldf('SELECT AVG(loyalty) AS AvgLoyalty, Gender FROM
surveyCoolWFlights GROUP BY Gender')

mostLoyalGender

#males are more loyal than females


#class

mostLoyalClass <- sqldf('SELECT AVG(loyalty) AS AvgLoyalty, Class FROM surveyCoolWFlights
GROUP BY Class')

mostLoyalClass

#most loyal is Business, then Eco Plus, then Eco


#status

mostLoyalStatus <- sqldf('SELECT AVG(loyalty) AS AvgLoyalty, Status FROM surveyCoolWFlights
GROUP BY Status')

mostLoyalStatus

#most loyal is Blue, then Silver, then Gold. Least loyal is platinum
```

```
#opposite from rating by Status. Blue is most loyal but least satisfied. Platinum is most satisfied, least loyal.


#travel type

mostLoyalTravel <- sqldf('SELECT AVG(loyalty) AS AvgLoyalty, Class FROM surveyCoolWFlights GROUP BY TypeOfTravel')

mostLoyalTravel

#most loyal is Personal, then Mileage, then Business

#also the opposite from rating by Type of Travel. Business travelers are most satisfied, least loyal.

#Personal travelers are least satisfied but most loyal


######business question 4: what variables have the strongest impact on Satisfaction score?


#linear models

str(surveyCool$Satisfaction) #need all variables to be the same type (numeric)


surveyCool$travelNum <- as.numeric(surveyCool$TypeOfTravel)

surveyCool$classNum <- as.numeric(surveyCool$Class)

surveyCool$statusNum <- as.numeric(surveyCool$Status)

surveyCool$satisfactionNum <- as.numeric(surveyCool$Satisfaction)

surveyCool$genderNum <- as.numeric(surveyCool$Gender)


#shopping

shoppingRating <- sqldf('SELECT Avg(Satisfaction) as avgSatisfaction, ShoppingAmtAtAirport FROM surveyCool WHERE ShoppingAmtAtAirport > 0 GROUP BY ShoppingAmtAtAirport')


shoppingLm <- lm(formula=avgSatisfaction ~ ShoppingAmtAtAirport, data=shoppingRating)

summary(shoppingLm)
```

```
#very low R^2 value, not a good predictor


#eat drink

eatingRating <- sqldf('SELECT Avg(Satisfaction) as avgSatisfaction, EatDrinkAtAirport FROM surveyCool GROUP BY EatDrinkAtAirport')

eatingRating


eatingLm <- lm(formula=avgSatisfaction ~ EatDrinkAtAirport, data=eatingRating)

summary(eatingLm)

#very low R^2 value, not a good predictor


#multiple variable models

#travel type, status

surveyCoolLm <- lm(formula=satisfactionNum ~ travelNum + statusNum, data=surveyCool)

summary(surveyCoolLm)

#R^2 value ~36%, p-value very low


#travel type, status, class, gender, age

surveyCool2 <- lm(formula=satisfactionNum ~ travelNum + statusNum + classNum + genderNum + Age, data=surveyCool)

summary(surveyCool2)

#R^2 value ~37%, p-value very low


#travel type, status, class, gender, age, shopping, eating

surveyCoolLm3 <- lm(formula=satisfactionNum ~ travelNum + statusNum + classNum + genderNum + Age + ShoppingAmtAtAirport + EatDrinkAtAirport, data=surveyCool)

summary(surveyCoolLm3)

#R^2 value ~37%, p-value very low
```

```
#KSVM and SVM Models

#creat training and testing data sets

surveyRows <- nrow(surveyCool)

surveyCutPoint <- floor((surveyRows*2)/3)

surveyRandom <- sample(1:surveyRows)


surveyTrain <- surveyCool[surveyRandom[1:surveyCutPoint],] #create training data set w/ first 2/3 of data

surveyTest <- surveyCool[surveyRandom[(surveyCutPoint+1):surveyRows],] #create test data set w/ remaining 1/3


dim(surveyTrain) #check # of rows

dim(surveyTest)


#ksvm model

library(kernlab)

surveyKsvm <- ksvm(surveyTrain$satisfactionNum ~., data=surveyTrain, kernel="rbfdot", kpar="automatic", prob.model=TRUE, cross=10, C=10) #create model training data

surveyKsvm


surveyPredict <- predict(surveyKsvm, surveyTest, type = "votes")


surveyCompare <- data.frame(surveyTest[,1], surveyPredict[,1])

colnames(surveyCompare) <- c("TestData", "Prediction")

head(surveyCompare) #compare predicted score to actual satisisfaction


#calculate RSME

library(Metrics)
```

```
rmse(surveyTest$satisfactionNum, surveyPredict)
#Root Mean Squared Error for KSVM = 2.132004

#SVM
library(e1071)
library(caret)
surveySvm <- svm(surveyTrain$satisfactionNum ~., data=surveyTrain) #create model
surveyPredictSvm <- predict(surveySvm, surveyTest, type = "votes") #predict values

rmse(surveyTest$satisfactionNum, surveyPredictSvm)
#Root Mean Squared Error for SVM = 2.132854
```