IST 678: INTRODUCTION TO DATA SCIENCE MANAGEMENT GROUP 3

FLIGHT SATISFACTION ANALYSIS



Project by:

Rachit Kaushik

Mudit Bisht

Maaz Kamal

Scott Dillon

**TABLE OF CONTENTS**

* Background 3
* Business Questions 4
* Data Cleaning, transformation and munging 5
* Descriptive Analysis and Visualization 5
* Modeling Techniques 8
* Actionable Insights 10

**Background**

Our main client, SouthWest Airlines has seen low satisfactions scores from customers (44%) for their air travel. They want to identify the factors leading to their customers being dissatisfied from their air travel and have tasked our team to do the necessary analysis and recommend the best possible solution.

We were given a dataset which consisted approximately 10,300 entries of customers, along with more than 28 characteristics, which included, their age, gender, flight details, number of flights in a year, satisfaction level (derived from “Likelihood.to.recommend

” attribute ranging from 1-10, with 8-10 rated as the high satisfaction level and 7 or below rated as the low satisfaction level), etc. The aforementioned factors are related to each other and help us in determining what affects a customer’s satisfaction score. Why is satisfaction score important? Because high satisfaction score suggests that the customer was happy with his/her experience and is highly likely to travel again with the same airlines, whereas, low satisfaction score suggests otherwise. Not only that, fact of the matter is that ‘word of mouth’ is the best form of marketing for any brand. Thus, it is in the interest of our client to make sure their customers are satisfied with their flight experience and likely to recommend the same to others.

To make sure we identify the problem and provide with the best possible solution, we had to ask the right business questions. Then, once we had the idea of what we are trying to do and what we want to achieve, we began with data acquisition, data cleaning and munging. We then moved forward to getting an idea of what kind of data we were dealing with, so we ran descriptive statistics for each variable. Once we were done with that, we ran various visualizations and modeling techniques in order to figure out any trends among highly or satisfied or unsatisfied customers. Finally, we summarized our data insights to the client and within this report. Our main objective is to advise the client as to how to raise customer satisfaction - which we will describe and validate throughout the remainder of this analysis.

**Business Questions**

* What are the factors leading to low satisfaction scores?
* What steps need to be taken to identify those factors?
* How much weightage each factor has on the satisfaction score?
* What are the things which customers care about the most?
* How to make sure the satisfaction scores of some characteristics increases?
* Are enough incentives being provided which would increase customer retention?
* Should the flight start In-Flight Meals?
* Is there a connection between flight dates and special events or festivals?
* Identify relationship between flight duration and customer satisfactions
* Identify top partner companies and whether there should be an incentive for them
* Is there any meaningful pattern between shoppers, foodies?
* Identify destinations/departures that make up the bulk and whether there be more booths to better handle the customers
* Identify destinations/departures where most flights were cancelled and whether services should be discontinued (since it may cause bad Public relations and loss of revenue)
* Identify relationship between flight duration and customer satisfactions
* Identify which flights are cancelled/delayed due to bad weather based on the combination of dates, city and states
* Predict need of new/better infrastructure or automation and self-service solutions in increasingly favorite cities

**DATA CLEANING, TRANSFORMATION, MUNGING**

Before we began with our analysis, we had to clean our data and make sure there weren’t any NA values or redundant information present in the data. First, we started by importing the JSON file into R and converted it into data frame which is readable in R using *jsonlite::fromJSON* function.

Once we had the dataset ready, we first ran a *for loop* which traverses through the all attributes in the entire dataset to identify all the NA. There were a total of 4 numerical attributes that had NA values. We replaced NA values with the mean values instead of 0 or the max value because we did not want to skew the data. We assumed that a canceled flight can have either: 1) a longer delay (in minutes) before it is canceled, or 2) no delay several days before the scheduled flight time. Therefore, we used the mean value to account for both situations. And there was only 1 character variable with NA values, which upon inspection was basically the comments from customers about their experience. Since there were about a thousand NA values (basically those values where the customer chose not to leave a comment) in this attribute and no apparent way to ‘fix’ them. We separated this attribute from rest of the data, leaving us with 31 variables instead of the original 32 variables. After this step, we ran various analysis techniques to achieve the goal we set out in the beginning.

**Descriptive Analysis**

In order to understand the data, we need to understand each numerical and factor variable. We did that by creating histogram for each numerical variable to observe a general trend (positively skewed, negatively skewed or symmetric) and to get a sense each of those variables may be affecting the satisfaction score and to analyze how these variables may be used to drive up the satisfaction score.

We classified these variables into three categories, which are as follows:

Customer attributes

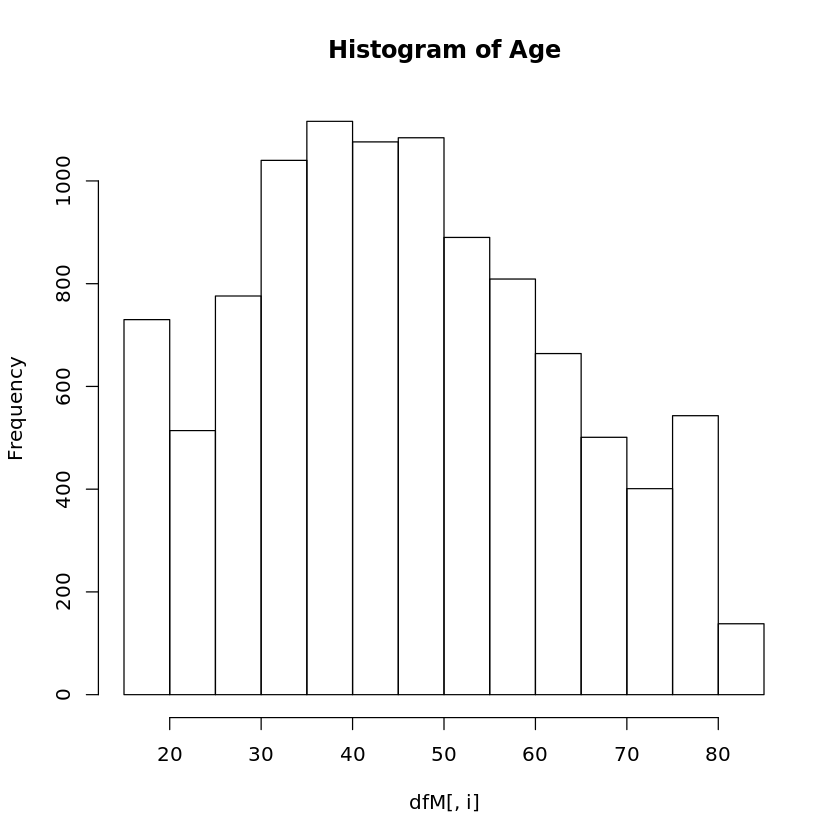
Flight attributes

Flight quality

Customer Attributes

For customer attributes, we chose Gender, Age, Travel Type, Class, and Status variables to display because these were the most interesting. Following is a description of these attributes.

**Age:** We created a histogram for the age variable to understand the distribution of customers using airlines which gave us interesting insights into their demographic. Below is the histogram of the age variable. As is clear from the histogram, we get a symmetric curve concentrated towards the center, that is, middle aged people (30 to 55 age) are the most frequent fliers compared to the other age groups. This can help the airlines to direct their advertisement campaigns towards the aforementioned age group.



**Gender:** For this categorical variable, we decided to use *table() command* to get the insights of the distribution of each gender type and the frequency associated with each of them. Using this command, we can infer that female passengers (5708) are moderately higher than male passengers (4574). This gives us another insight into which gender type to focus on.

**Type of Travel:** This categorical variable has three sub-categories, namely, Business Travel, Personal Travel, and Mileage Tickets. With our analysis, we were able to derive that the frequency of business travelers (6282) is considerably higher than personal travel (3176) and mileage tickets (824). This is another important demographic which gives us insights into what is the purpose of travel of the customers.

**Class:** Using the *table()* command, we could infer that this categorical variable is subdivided into three categories, these are, Economy, Economy Plus, and Business with majority of customers preferring Economic class compared to other categories.

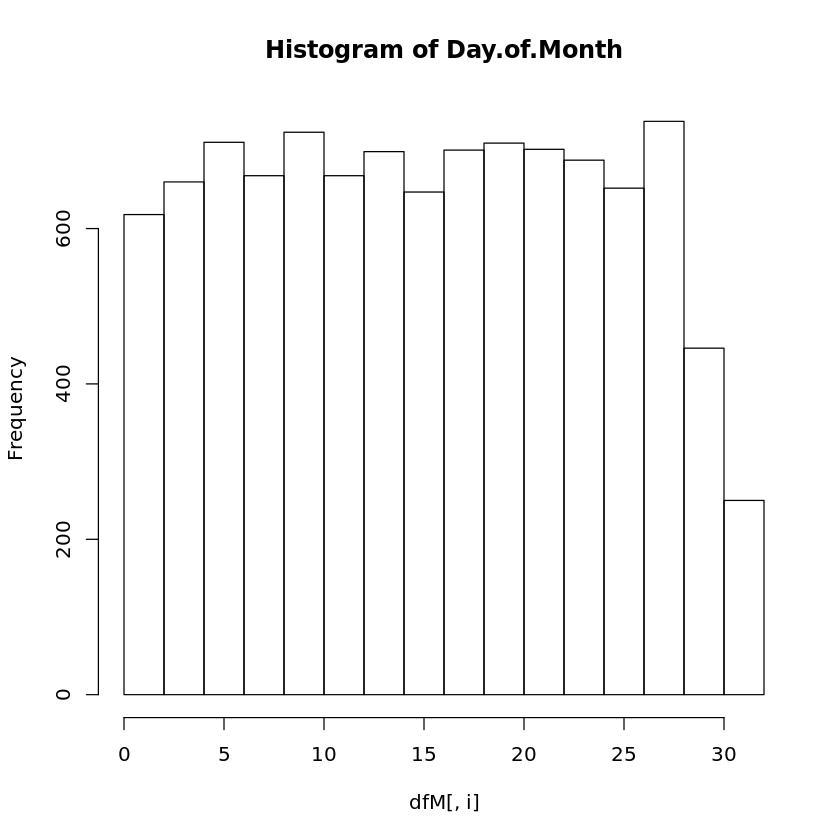
**Airline Status:** We were able to derive that most of the customers had a Blue status (7003) compared to Silver (2051), Gold (850) and Platinum (378). Thus, we can say that airlines should focus on the Blue status travelers to get the maximum profit.

Flight Attributes

For customer attributes, we chose Airline Name, Flight Day of the Month, Origin and Destination State to understand our data. Following is the descriptive analysis of these variables.

**Partner Name:** We created an analysis to understand the demographic of all the partners of our airlines. We were able to conclude that Cheapseats Airlines and Sigma Airlines were the most prominent partners.

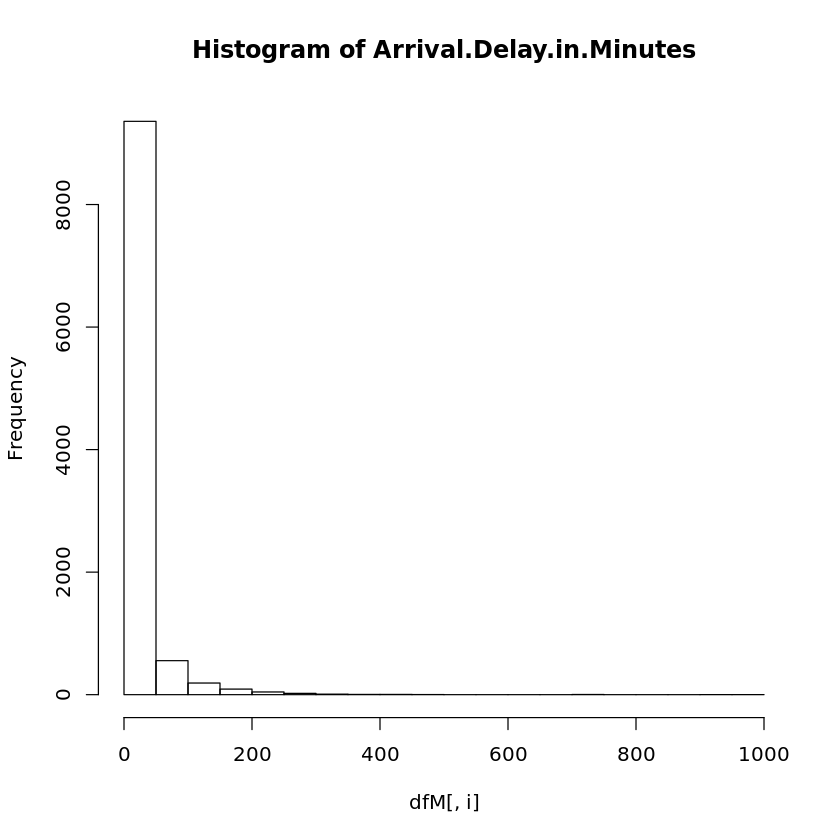
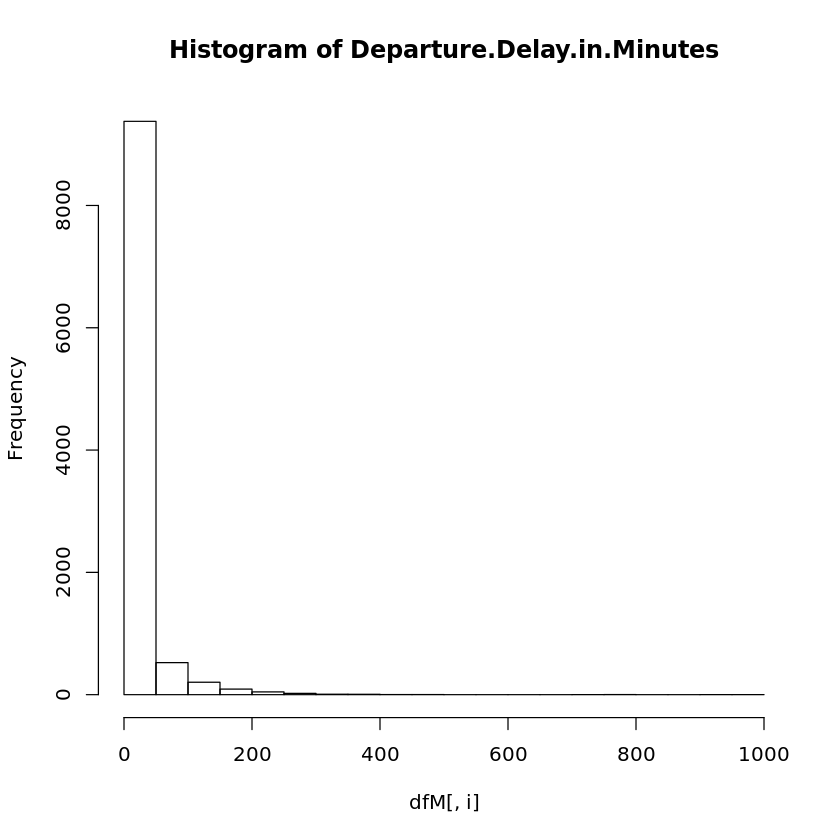
**Flight Day of The Month:** We created a histogram for this variable. We can see that histogram is symmetric, that is, most of the data is accumulated in the center. But another interesting insight we have is that just before the end of the month, there is a spike in the frequency of travelers, which shows that we also need to focus on those days as well.



**Origin State and Destination State:** We can infer that California is the busiest airport for both origin and destination.

Flight Quality

**Arrival Delay and Departure Delay:** The data for arrival delay and departure delay generate a right skewed histogram, suggesting that flights are generally not delayed at all, but if there is a departure or arrival delay, it is more likely that the flight is delayed by more than five minutes.

Free Text attribute

Lastly, we decided to use that 32nd attribute (of the comments from the customer ) which was separated from the rest of the data to perform sentiment analysis and derive as much insight as we could (Results shared on page 11).

**Modeling Techniques**

We utilized three different modeling techniques to address our business questions:

1) Linear Regression, 2) Association Rules and 3) Support Vector Machines (“SVM”)

**Linear Regression:** We used multivariate linear regression models to try and establish a causal link between different variables and customer satisfaction. And to highlight all significant attributes that were influencing satisfaction score of the customer.

First, we ran linear regression models using all traveler information. This produced regression results with an adjusted R-squared value of approximately 0.4146 .

We then removed the variables that were not statistically significant, and ran the regression again which reduced adjusted R-squared value to 0.4057. The reduced adjusted R-squared value may or may not be a cause of alarm depending, how we look at it. But the reduced adjusted R-squared may simply be, since we have lesser dependent variables for our model. Regardless, at the end of this exercise we have filtered and statistically significant attributes, which is perhaps the most important challenge when dealing with large amounts of data, that is, narrowing our focus down to only significant and actionable data.

Hence our client should pay most attention to the following attributes whenever making any decisions, since these attributes directly affect customers satisfaction:

* Airline Status
* Age
* Gender
* Price Sensitivity
* Flights Per Year
* Type of Travel
* Eating and Drinking at Airport
* Class
* Day of Month
* Flight date
* Partner Code
* Scheduled Departure Hour
* Departure Delay in Minutes
* Arrival Delay in Minutes

**Association Rules:** We used the association rules model to: 1) create profiles for happy and unhappy customers, 2) find out factors that make a flight satisfied or unsatisfied, and 3) relate our findings to specific airlines in order to understand why some airlines have higher satisfaction rate over others.

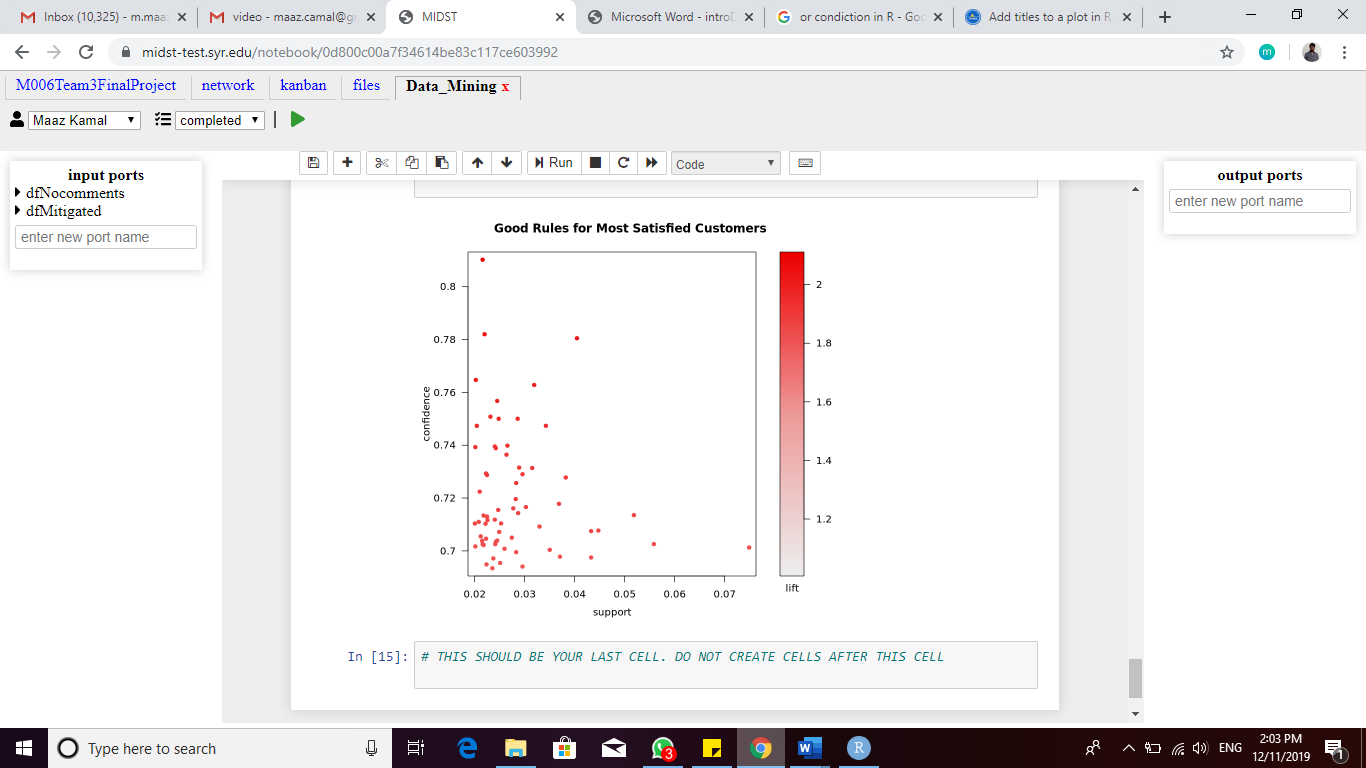
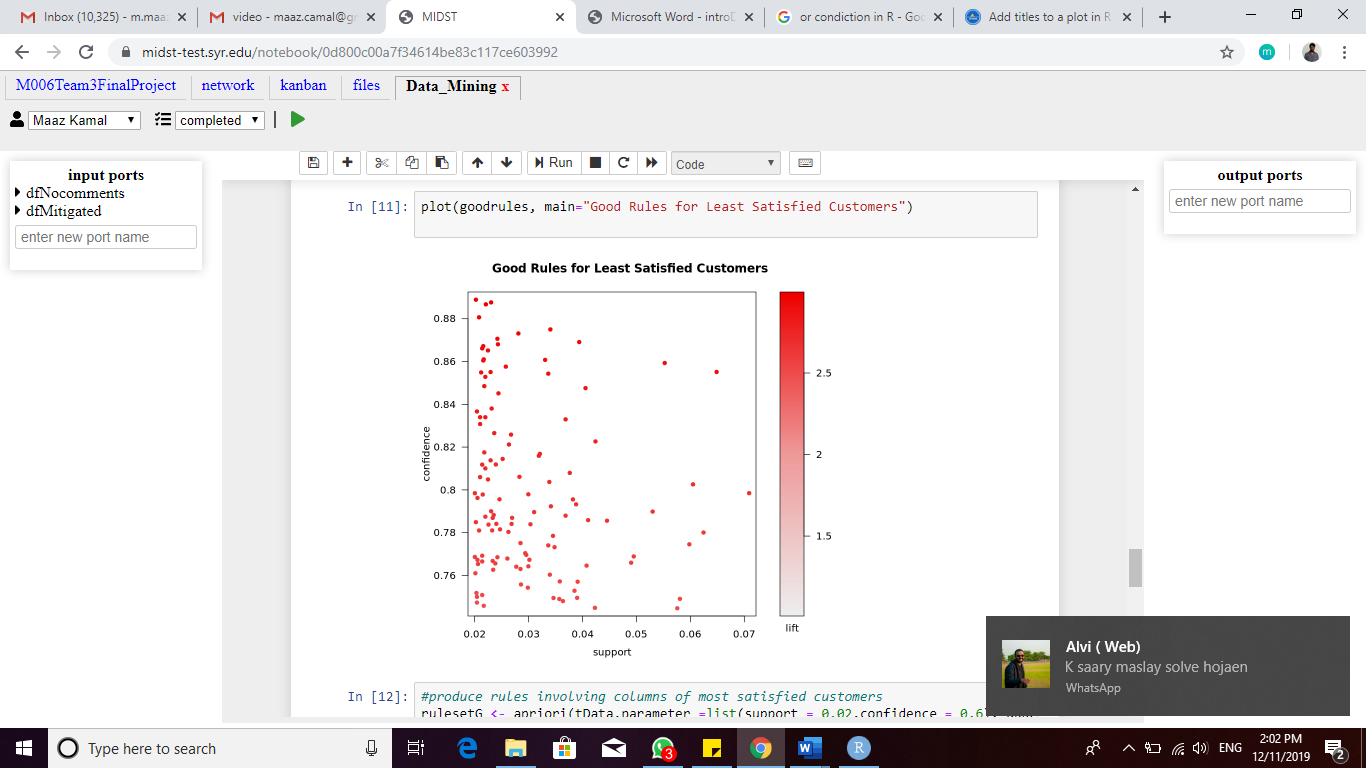
#Association Rules were used on 2 different (fixed) RHS. Firstly, on the Lowest Satisfied Customers (score between 1-7) and secondly most satisfied customers (score between 9-10). The gap (of 7-8) between the two was intentionally left out, so that the differences can be clearly marked out between the satisfied and not unsatisfied customers. Following were the consultations of these 2 exercises:

**RHS= Least satisfied customers (of score between 1-7)**

* "Personal Travel" type of travel, appears to be closely related to the low satisfactory score.
* Especially those "Personal travel" who are using Client's Partner, FlyFast Airways Inc.
* People who spend under $40 on eating and drinking at the airport are particularly unsatisfied with our Client's services.

**RHS= most satisfied customers (of score between 9-10)**

* "Business Travel" type of travel, appears to be closely related to high satisfactory score.
* Especially those " Business Travel" who are using Client's Partner, Northwest Business Airlines Inc. and Sigma Airlines Inc.
* People who spend more than $40 on eating and drinking at the airport are particularly satisfied with Client's services.

**Support Vector Machines (SVM):** We ran SVM model to try to generate an accurate model that can predict whether a customer is going to be happy or unhappy, based on significant customer attributes. As we have found from the association rule analysis and linear regression model~~s~~, personal attributes such as gender, traveler type, and airline status can influence the group’s satisfaction rate, we decided to use these (and other )variables in our SVM model to predict the satisfaction level of potential customers.

After much tweaking and experimenting with parameters values, we came up with a promising model with Cross validation error of 0.244614. Once we had our model, we used the Test Data (that was separated and not used in the formation of the model) to test the accuracy and validation of our model. The resulting error rate of this (unseen) test data was found out to be 0.2429127, which was very much comparable to original cross validation error of our model (.244614).

Hence, we successfully came up with a model that can potentially predict the satisfaction level of a group of travelers based on their personal attributes at almost 75.9% precision rate to our client. This may help our client to take preventative measures beforehand if it is predicted that travelers might be unhappy about certain flights.

**Actionable Insights**

We have seen from our analysis above that most of the customers who are unhappy were either “Personal” traveling, traveling with client’s partner airline “FlyFast Airways Inc.” or were the ones spending less than $40 on eating and drinking at the airport.

We also saw that most satisfied customers were the ones doing “Business” travelling, travelling with client’s airline partners Northwest Business Airlines Inc. or Sigma Airlines Inc, or were the ones spending more than $40 on eating and drinking at the airport.

Based on these insights, following recommendations may be proposed:

* + Particular attention be given to the customers doing “Personal” travel via better, quality of service and communication.
  + Introduction of incentive and disincentive for the partner airlines based on their performance. Moreover, based on the disapproval of customers of client’s partner airline “FlyFast Airways Inc.”, discontinuation of the partnership may be considered.
  + Based on the patterns between satisfied and unsatisfied customers and their spending on food and drinking at the airport, perhaps the policy of not offering inflight meals, could be reviewed.

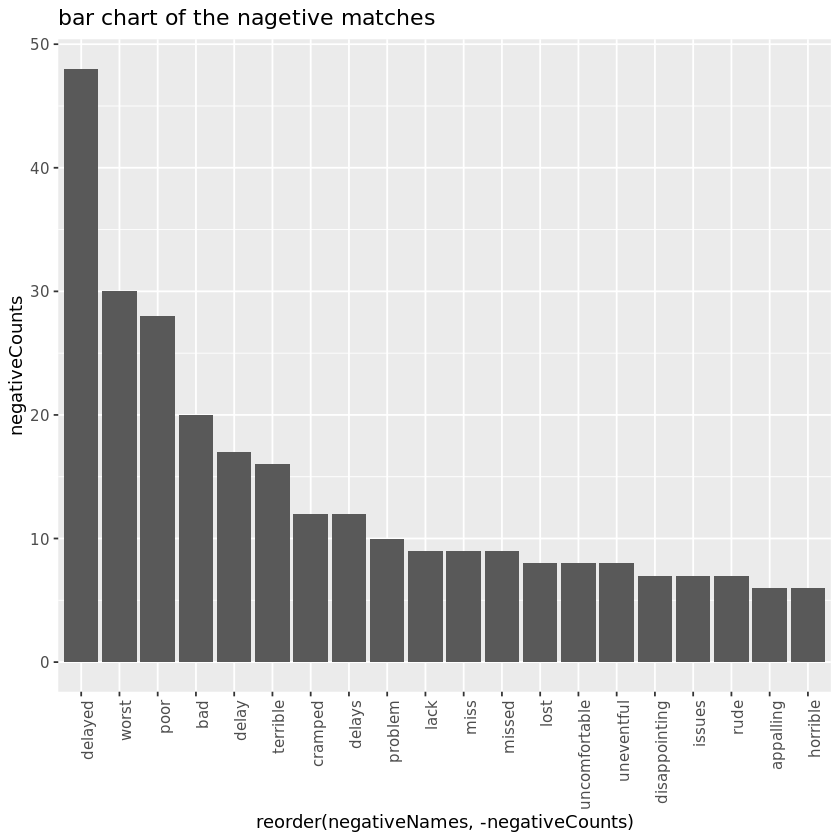
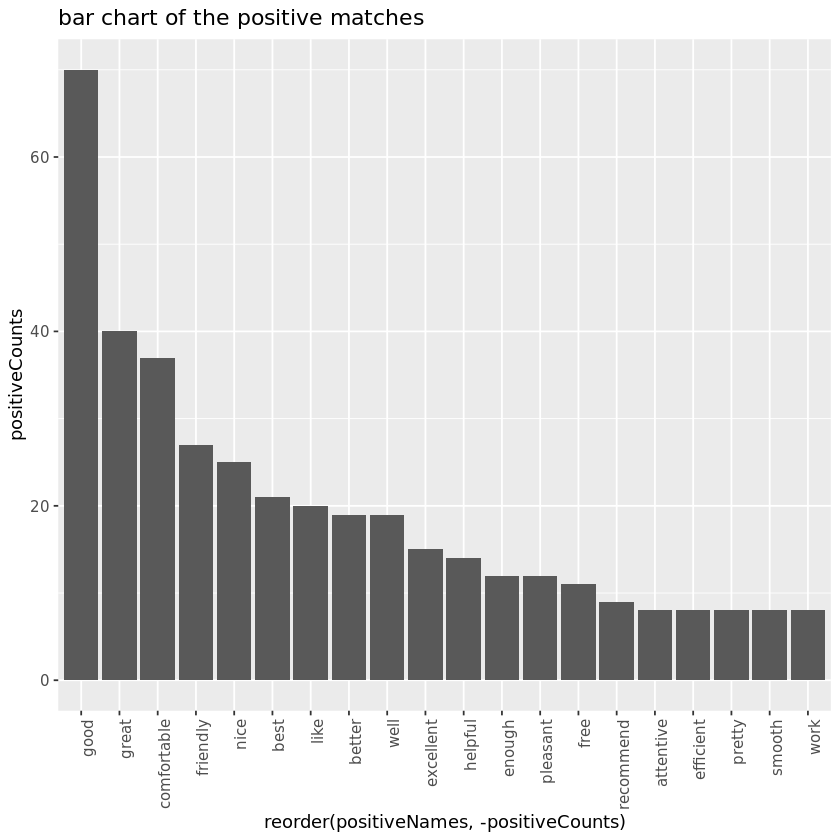
Since Atlanta, Chicago and Los Angeles make the bulk of the clients’ customers travelled cities, increasing (trained) staff should be able to better handle flight delays, customer bumping etc.

Furthermore, utilization of the prediction model to work on increasing the general customer satisfaction rate and to come up with methods to raise satisfaction rate before they board a flight is highly recommended.

Performing sentiment analysis on the comments of the customers provides us additional insights.



Extracting the positive and negative words from the reviews of customers and analyzing the number of times they are used, it can be detected where the flight is going wrong (as seen in the negative word bar chart) and what things are right (as seen in the positive word bar chart). This can help the airlines to improve the things that people don’t like about the flights and to advertise and promote the things that are already going for the company. A higher ratio of positive to negative words would signify a general happiness and lead to higher satisfaction score which in this case is true with the negative rating being equal to 6.76 and the positive ratings equal to 8.77.

 Judging from the Positive words’ graph, flight delays and inflight inconveniences bother the customers the most. Perhaps improved and near-real time communication with the customers could be considered.

Whereas, judging from the graph, most customers seem to be satisfied with the staff of the airline, which should be appropriately awarded, and similar attention be paid to the other (above mentioned) areas