Flight Satisfaction and Travel Time

Applications of Machine Learning

Contents

Executive Summary

The COVID-19 pandemic has widely disrupted many industries, with travel and leisure taking the brunt. This report outlines a scenario where a passenger-airliner strives to implement a re-engagement strategy by utilizing previous customer surveys and flight data. Recommendations would be tailored to the results, which would be derived through data exploration, supervised and unsupervised machine learning methods including classification, linear regression, and clustering.

<u>Overview of the Case</u>

At this point in time, the restrictions from the pandemic continue to be lifted but the travellers the airlines are servicing remain compressed compared to previous years. Airliner X believes that they can diminish the reluctance to fly through improving the most influential services when predicting for satisfaction. The reasoning for improving services rather than improving value at decreasing price is outlined by the competitive nature of the industry. The purpose of the analysis will be to provide relevant and feasible solutions when considering the results from the learning methods.

<u>Data Exploration</u>

The section dedicated to data exploration attempts to fully describe the datasets provided to the analysts. The contents include general descriptions of the variables, describe distributions found throughout the data, any variables with significant outliers discovered, correlations between variables, and other areas of interest that have been noted.

<u>Classification</u>

The first learning method discussed was classification, which included logistic regression, decision trees, random forest, and an ensemble model. Key findings were the flight arrival delay, online boarding services, Wi-Fi services and inflight entertainment were dominant features when considering flight satisfaction.

Linear Regression

To gain insight about the flight delay feature, the arrival delay data was used to possibly outline any processes that could be enhanced. When predicting for flight delay, notable features included the taxi-time and departure delay.

Clustering

The clustering section outlines the use of the unsupervised method on the customer survey dataset. The findings not only outline the largest normalized differences between features, but also depict two distinct groups of customers for their demographic features which would be used for targeting the optimal segment.

Alternatives, Recommendations and Conclusion

The last of the report would be dedicated to aggregating the results and interpreted with the situational background to provide the framework for creating sound selection for a course of action. The recommendations will be based upon the most prevalent features of every learning method and applied to Airliner X’s growth strategy.

Appendix

The appendix is where the data exploration and learning methods are visualized, with short explanations if necessary.

Overview of the Case

After recent systemic changes to travel due to pandemic-based restrictions, there has been a significant decline in airline passengers. Though the pandemic has certainly disrupted many industries, the ticket purchase quantities for airlines have fallen to the levels not seen since the 9/11 terrorist attack, which is a sixty-percent decrease from 2020 highs (Olaganathan, 2021). Considering that many global travel restrictions continue to be lifted and the number of passengers carried remain compressed, Airliner X wants to understand what their customers previously valued services in an attempt to regain the proportion of airline travellers and market share.

Airline X has provided two datasets for understanding their customer experience and thus make informed decisions for retaining their most potential customers.

The Business Problem

Why is this customer satisfaction so important to the passenger-airline industry? The North American airline industry is currently renowned for its uncompetitive nature. Though players like Southwest have somewhat disrupted the industry, the industry has significant barriers to entry such as the cost of a Boeing 737 costs between $80-$130 million (Wolla Backus, 2018). The airline would be operating within something close to an oligopoly, where there are not necessarily pressures to decrease pricing with limited competition but the pressure to increase differentiation would primarily be where airlines create their value and ultimately capture a larger market share. Airliner X wants to undergo market share expansion and re-engagement from previous travelers through further understanding the variables that customers value and which variables are being overlooked.

Objective

The primary objective is to utilize three different predictive models to gain insight into the airliner and its previous customers. Customer satisfaction will be used as a proxy of the perceived value offered to customers with a range of services, which the analysts are attempting to maximize as the output of the project. The end goal will be to provide a solution that would increase customer satisfaction. An attempt will be made to successfully analyze the survey dataset to understand if there are any forces to customer satisfaction that have gone unnoticed which can be uncovered. Additionally, it is speculated that one of the factors reducing customer satisfaction are the arrival delays experienced by the passengers. Therefore, the project also tries to explore the factors which can reduce these arrival delays using the arrival delay dataset. Finally, passengers will also be clustered in groups to identify which customer type is the least satisfied. The key drivers uncovered from the models combined with external research about the industry will assist decision makers in maximizing the value offered to customers, especially the unsatisfied group through a final implementation strategy that would follow the recommendations.

Exploring the Datasets

Two different datasets were used for this analysis, the first set of data was the survey entries of customers which capture over 90,000 observation and 24 variables including: id, Gender, Customer Type, Age, Type of Travel, Class, Flight Distance, Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, Cleanliness, Departure Delay in Minutes, Arrival Delay in Minutes, satisfaction, which was sourced from Kaggle. This dataset will be used to explore which factors impact customer satisfaction level, and therefore the dependent variable here is satisfaction. The second dataset was sourced from U.S. department of transportation's (DOT) bureau of transportation statistics, this contains over 100,000 rows and about 30 variables which captures the flight details such as: flight date, origin, destination, taxi out, departure delay, distance, delay, cancellations etc.

The dataset will be useful for the above dataset as here the factors which can reduce arrival delays are determined.

Exploring the Survey Dataset

To explore the distribution of scores on services impacting the satisfaction level (from survey dataset), three groups were created which signifies the online experience, airport experience, and inflight experience of the passengers. This way the available 13 variables can be explored effectively. Three different plots (Appendix 1, a) were created representing each category, which revealed that distribution of scores for inflight experience and airport experience were more normally distributed as compared to online experience. Furthermore, a box plot (Appendix 1, b) was created for the 13 variables denoting services, to identify which service has an average rating above the mean rating. The results from the boxplot showed that inflight-Wi-Fi services, ease of online booking, gate location, and departure arrival time convenience fall below the mean score thus Airline-X should focus on improving these. Additionally, the highest scores were given for baggage handling. The scores on services by satisfied and dissatisfied customers were also checked, and it was found out that the satisfied customers had given higher scores on the services (Appendix 1, c). From this it can be inferred that the higher scores on services are desirable to get more satisfied customers.

Exploring the Arrival Delay Dataset

The distribution of the arrival delays (Appendix 2, a) suggest that the maximum number of flights were not delayed, and if they were delayed then the delay time was less than 250 minutes for most flights. Additionally, the correlation between the arrival delays was checked with departure delay and taxi time phase, and the result showed that all these variables lead to higher arrival delays (Appendix 2, b).

Based on the exploration results, it is expected that the satisfaction level will depend on the scores a passenger gives on the services. It is also expected that taxi-time and departure delays will play a significant role in determining the arrival delays. To get the concrete results, logistic and linear regressions are trained and tested in the subsequent sections.

Classification of Customers

Various features of the customer satisfaction survey dataset have been used to not only classify the customers as satisfied and dissatisfied but also to identify factors which contribute to customer satisfaction. Here, four different algorithms namely logistic regression, decision tree, random forest, and ensemble models are trained and tested to accurately classify the passengers in two groups. Passengers classified as dissatisfied can then be targeted by the airline to improve their satisfaction and consequently increase the profits. Before analysing the data, it was pre-processed where variables were converted to the desired structure, outliers in delays were removed, and missing values (0.3% of the data) were removed.

Logistic Regression

Fitting the Model

The dependent variable in the survey dataset is categorical in nature due to which a logistic regression model is fitted rather than multi-linear regression. A range of independent variables which signifies the demographic of the passenger (gender, age, etc.), travel information (type of travel, flight distance, arrival delays, etc.), and the scores the passenger gave on different inflight, onboard, and airport services were used. Additionally, the dataset contained information on both the arrival delays and departure delays, from which only arrival delays are incorporated in the model. This is because both the variables are dependent on each other which might lead to the problem of multicollinearity in the model. After fitting the model using ‘glm’, the most important features were selected using ‘step’ where both the forward and backward directions were selected.

Results and Error

The logistic regression results (Appendix 3, a) show that the demographic factors, travel information, and the scores of services indeed have an impact on determining whether the customer will be satisfied or not as the p-value of these variables is less than 5% level of significance. According to the beta coefficients, it can be inferred that the probability of a customer being satisfied decreases when the customer is female, passengers are older, they are non-loyal customers, the type of travel is personal travel, and when the passengers are travelling in economy and economy plus. This implies males, younger passengers, loyal customers, business travellers, and travelling in business class increases the probability of being satisfied.

The results also suggest that a few services play a significant role in classifying customers as satisfied. A higher score on services like inflight Wi-Fi service, gate location, online boarding, seat comfort, inflight entertainment, on board services, leg room services, baggage handling, check-in services, inflight services, and cleanliness increases the probability of satisfaction. For instance, when a passenger increases their score on inflight Wi-Fi service by 1, then the probability of the customer being satisfied increases by 1.531 units (i.e., e^ (0.426)). There are a few services like time convenience, ease of online booking, and food and drinks which have a negative coefficient, however this is logically not right as no satisfied customer would want to have poor services. Finally, as speculated, arrival delays have a significant impact on customer satisfaction, and when arrival delays are high then the probability of being satisfied reduces.

Next, to classify the customers as satisfied and dissatisfied, an accuracy plot was plotted to check which cut-off value yields the highest accuracy. Based on this plot, the customers for whom the probability of satisfaction was higher than 0.5 were classified as ‘satisfied’, while the rest were classified as ‘dissatisfied’. The trained model was then tested on the testing dataset and a confusion matrix was obtained (Appendix 3, b). The matrix showed that the overall accuracy of the model was 85%, and the model classified 77% satisfied customers accurately and 91% dissatisfied customers accurately. An ROC curve was also plotted (Appendix 3, b) which showed that the area under the curve was 90.2% indicating good accuracy compared to a naïve model.

Decision Tree and Random Forest

Fitting the model: Decision Tree

Decision trees were also used to classify the customers as satisfied and dissatisfied. In the initial model, a very high value of complexity parameter was taken which yielded a tree with an extremely large number of leaves. Due to this, it is highly likely that the model will be overfitted, therefore pruning was done at the level where the standard error was minimum. However, despite pruning the tree obtained had 225 leaves, and the accuracy of the tree was 95% on the testing dataset. Furthermore, to derive applicable insights and to have a visual representation of the tree, the value of the complexity parameter was lowered, and the pruned version of the tree was taken as the final model (Appendix 4, a).

Results and Error

The results of the decision tree are quite similar to the logistic regression, where the customers who have higher scores on services are categorized as 'satisfied'. Among the services, higher scores on Online Boarding, Inflight Wi-Fi, Inflight Entertainment, and Checking services are important variables for classification. The tree also reveals that the customers who travel in economy and economy plus, travel for personal reasons and are disloyal customers are usually classified as 'dissatisfied'. Extracting the important variables, it was found that most of the services are important for classification, and thus it can be inferred that the Airline-X should focus on improving these services.

The final model was also tested on the testing dataset, and the confusion matrix (Appendix 4, b) showed that the model had an overall accuracy of 91%, and the model classified 94% satisfied customers accurately and 89% dissatisfied customers accurately. An ROC curve was also plotted (Appendix 4, b) which showed that the area under the curve was 95.4% indicating good accuracy compared to a naïve model.

Random Forest Model

To get better results compared to the best individual predictor, a random forest model has been fitted to aggregate the results of multiple predictive trees. The random forest model is created using the 'randomForest' function in R, and 500 different trees were used to create the random forest. According to the mean decrease in Gini (Appendix 4, c), the important variables in creating the random forest are similar to the results of other classification models. Again, Online Boarding, Inflight Wi-Fi services, type of travel, class in which the passenger is travelling, and seat comfort have an impact on satisfaction level. Additionally, when the model is tested on the testing dataset (Appendix 4, d), the overall accuracy is 96% and classified 97% satisfied customers accurately and 94% dissatisfied customers accurately. An ROC curve was also plotted (Appendix 4, d) which showed that the area under the curve was 96% indicating good accuracy compared to a naïve model.

Ensemble Model

Ensemble learning is a machine learning approach that combines the predictions from different models to improve the predictive performance. Therefore, appropriate algorithms were used to create a stacking ensemble model where a base layer was created using the logistic model, decision tree, and random forest, and the top layer was the logistic regression model. It should be noted that the dataset was reduced (5% of the original dataset) to run the algorithms as the machine had low computational power. To create the ensemble model in R, 'caretList' function was used first for creating the base layer. A cross validation set which is repeated twice and creates 5 subsets for training and testing was also applied on the base layer. These results were then used to stack the logistic regression model using the 'caretStack' function. Finally, the results of the ensemble model (Appendix 4, e) were obtained which suggested that the overall accuracy of the model is 93.79%. Although the model is a black-box model which cannot be used for deriving which factors contributed the most, the model can surely be used in predicting which customers are satisfied and dissatisfied.

Overall Results of Classification

The different classification models used provided very similar results with respect to the factors which are important in determining whether the passenger is satisfied or not. The models strongly indicate that better scores on the services is desirable to the passengers, and the passengers with high scores are usually classified as satisfied. Therefore, in terms of airline perspective, improving these services should be prioritised as better services would help improve the flying experience of the customers, and thus Airline-X would be able to retain more customers. Additionally, the results also show that lower delays are usually preferred by the passengers, therefore the airline should identify ways to lower these delays. Moreover, demographic and travel information are also important for classification, therefore a particular segment of the people who are usually classified as dissatisfied can be targeted by the Airline to improve their satisfaction level such that they choose Airline-X for their future travel plan. Building on this, the customers are also clustered to identify such segments.

Factors Affecting Arrival Delays

The results from logistic regression highlighted that arrival delays reduce the probability of satisfaction; therefore, the Airline-X must figure out ways to reduce these delays. A regression model is therefore fitted using the Arrival Delay dataset where various features are used which will help understand the causes of delays. This way, the Airline-X can focus on improving these factors to reduce the delays and thus increase customer satisfaction level.

Fitting the Model

Since the variable of interest is arrival delays, therefore arrival delays are taken as the dependent variables. After preprocessing the dataset by removing the missing values and changing the type of some variables, the most important features which determine the arrival delays were included to fit the model. Of the variables provided, day of the week, taxi-out and taxi-in time[1], total distance of the flight, and various delays including departure delay, air system delay, security delay, airline delay, and weather delays were included in the model to explain arrival delays. Although it is expected that usually delays due to various reasons will increase the delays in arrival, these were still included in the model to see which delay has the maximum impact on arrival delays. Using the 80/20 rule, the observations are split into 80% training, and 20% testing,

[1] The taxi time is the amount of time an aircraft spends in movement on the surface of an airport. This is also the phase when many complex things happen in the cockpit and cabin, and various checks and inspections are undertaken during this period.

and the model was trained on the training dataset. Additionally, a train control was set up in order to apply a cross validation process while running the model.

Results and Error

The regression results (Appendix 6, a) show that the most important factors in determining the arrival delays are taxi phase and the delays in departure, and the adjusted R-square of the model is 97%. As the taxi-out and taxi-in time increases, the arrival delays are expected to increase. Additionally, as speculated, different types of delays are also statistically significant at 5% significance level, and it is highly likely that a flight will arrive late when these delays increase. Among the different delays, departure delays have the highest beta value, indicating that departure delays have the highest impact on arrival delays. The trained model was then applied on the testing dataset, and it was found that the testing dataset error is 10.23 while the training dataset error was 10.25. Since, both training and testing errors are low, therefore the model is a good fit (Appendix 6, b).

Based on these results, it is suggested that Airline-X should focus on reducing the taxi-time phase and the delays due to departure, air-system, security, and airline to reduce the delays in departure. Although weather delays are also important in determining arrival delays, nature is not in our control, so no recommendations are made with respect to weather delays.

Clustering Customer Surveys

The classification algorithms helped understand the features which are important in classifying the customers as satisfied and dissatisfied. However, in addition to classification, the objective is also to determine the characteristics of each passenger who has claimed that they are satisfied or dissatisfied. Therefore, the passengers are clustered in different segments, from which their traits can be studied, and relevant programmes can be implemented by the Airline-X to increase satisfaction level.

Fitting the Model

To begin training the model, pre-processing was necessary when applying clustering on the customer survey dataset. Looking into the dataset there are certain columns that would not translate well in their current structure, and thus their type was changed. These variables were converted from a character to a factor structure, then to an integer. The common issue of processing the data was once again a problem, so the data had to be partitioned to analyze the figures due to the scale of the dataset. Approximately only 0.3% of the data had missing values, so omitting the observations will not diminish the impact of the findings. Before the model begins to be built, the survey number column and customer ID column had to be removed considering these variables are meaningless observations that will interfere with the results, and the data had to be normalized due to the differing measurements that were found throughout the independent variables.

The k-means clustering algorithm was chosen to cluster the passengers due to the computational power it would require running the hierarchical cluster methods plus the robustness of the k-means clusters. After calculating the distance (Appendix 6, a), the optimal size of the cluster was obtained. After reviewing the error, the best balance of r-squared and error appeared to be the cluster with a size of two. Two clusters offered the largest average silhouette width, with a diminishing total sum of square error occurring after two clusters (Appendix 6, b). Variation displayed by the r-squared is approximately 0.2 (Appendix 6, c) and visually the neatest cluster amount was certainly two (Appendix 6, d). The optimal number of clusters per the GAP stat was 10 but it appears to have an overestimation of clusters. Finally, the summary (Appendix 6, e) of the results were derived from which the traits of each cluster were identified.

In addition to k-means, PAM clustering was also used to form clusters with categorical variables. For this, 'gower' distance was calculated and based on optimal clusters, 2 clusters were created (Appendix 7, a). The summary of the clusters was then obtained to interpret the results (Appendix 7, b).

Results and Errors

The results of the k-means clustering showed that over 75% of the satisfied customers were in cluster two, and over 88% of dissatisfied customers in cluster one (Appendix 6, f). While considering the clusters and satisfaction, there are notable features from the cluster that obtained the higher satisfaction average. These results can again be used to see which attributes lead to higher satisfaction, and on this basis Airline-X could invest into these features.

The results from clutter one (containing a higher number of dissatisfied passengers) (Appendix 6, e) showed that the largest normalized average figures were for inflight entertainment rating, seat comfort rating, and cleanliness rating. The difference in mean results for the variables of interest for inflight entertainment, seat comfort and cleanliness ratings were 1.8, 1.5 and 1.5 respectively. The gaps in between the ratings appear quite large which the analysts can use when considering a recommended course of action when considering the flight satisfaction. Moreover, most of the passengers in this cluster had a lower mean on the services ranging between 2 and 3.

Also, some interesting findings from cluster one reveals that marginally more females were present, and passengers travelling in lower class categories were more dissatisfied. When compared to cluster 2, passengers who travelled lower distances were more dissatisfied, while a remarkable similarity was visible in mean results for departure and arrival delays. The average departure and arrival delay in minutes were roughly around fifteen minutes for clusters 1 and 2. When visualizing the parallel coordinate plot, it can be confirmed that there are significant outliers in departure and arrival delays data with standard deviation occurrences over ten, which is to be expected to occur when considering abnormal *distance* or *taxi-time* on a *weekend* for instance. The original parallel coordinate plot had large singular outliers displayed by this chart, which the outliers were removed in preprocessing due to those original results.

After creating the clusters, 5% of the dataset was scrapped and clusters were created again to check the stability. As similar clusters were obtained on the smaller dataset, therefore it can be concluded that the clusters are stable. Additionally, the results from PAM clustering were also very similar to the results of k-means clustering. The PAM clusters also revealed that 50% of the people travelling for personal reasons are dissatisfied, while 30% of business class travellers are in cluster one which has a higher number of dissatisfied customers. The combined results of these two clustering methods shows that females, passengers traveling in economy and economy plus, and passengers travelling for personal reasons are more unsatisfied by the Arline-X. Therefore, targeting these passengers for promotional activities will be beneficial for the Airline. Additionally, satisfied passengers have higher scores on services as compared to dissatisfied passengers.

Alternatives

Throughout the analysis, a variety of possible alternatives through clustering, including flight delay, flight class, Wi-Fi service, convenience of time, ease of online booking, gate location, food and drink services, online boarding, seat comfort, inflight services and entertainment, leg room, baggage handling or cleanliness have been highlighted. Looking further into the flight data, the flight delay variable can be improved through the taxi time and delays due to various reasons. Though investment can be made in any one of these variables, Airliner X should be optimizing their return by investing strictly in the most powerful variables at influencing customer satisfaction.

Targeting the more unsatisfied segment from clustering can assist at understanding the demographics of who Airline X can improve their services with and what services are lagging performance. If the assumption was that more satisfied customers have superior repeat purchases, then an approach at improving the satisfaction rate among the unsatisfied segment could be to invest in the lagging services, offer promotions to this segment and attempt to market the more admired services among this group of customers.

The Recommendations

After analyzing all the outputs, it is evident that the machine learning algorithms have assisted greatly in organizing the following recommended courses of action for the airliner. One important conclusion looking back at the logistic regression model, is that the arrival delay had a strong negative correlation with the satisfaction rating. Through the linear regression analysis of the flight delay dataset, the 'taxi_out' and ‘taxi_in’ attributes were considerably significant. Essentially, the company would be able to benefit from increased satisfaction and profits by reducing the taxi-time. Introduction of Taxi Time Management Tools and software, and having integrated tower, airfield, and gate support systems should be introduced at the airports. This will help predict the real-time location of the airplanes, which will make it easier for on-ground staff and air-control traffic staff to reduce the taxi out time of all the aircrafts. In addition to this, Airline-X can work on improving the automotive power of their aircrafts to further reduce the time their planes spend on moving from hangars to the runway. Furthermore, the airline can reduce time on check-in, boarding, and baggage loading by hiring efficient staff. This way the departure delays can be reduced which will further reduce the arrival delays.

The other frequent variables found to have the largest impact on flight satisfaction throughout all supervised and unsupervised models were the services provided in-flight, onboard, and on airport. So ideally, fast in-flight Wi-Fi and good entertainment, quick and easy online boarding, hassle free check-in services, higher legroom space, and careful baggage handling should be the services to invest resources into for the largest return in customer satisfaction. Multiple ways are suggested below which can be used by the Airline-X to improve these services.

Online boarding can be a service that is heavily driven for use after purchase of electronic tickets. If no online boarding currently exists for certain airports, priority for investment into developing online boarding services should be guided by the greatest number of outbound flights airports for Airliner X. Additionally, chat-bots on the website to help customers with online boarding can be introduced. The Airline-X can hire efficient ground level staff such that check-in can be quick, and baggage is handled with care. For implementing more superior Wi-Fi, Airline-X can connect with better Wi-Fi providers. Inflight entertainment can be developed through multiple avenues but offering media streaming platforms for use like Apple Music or Netflix for use should be strong considerations when considering modern-day inflight entertainment. Seat comfort would likely be the most difficult, resource-heavy, and time-consuming to implement, which the implementation would depend on the depth of the product differentiation for Airliner

X. The development of a new seat design would be needed, current and future jets would need to be re-equipped with the desired seat, along with the consideration that seats are an asset for the Airliner that deteriorates (and depreciates) which will need to be replaced over time. The first three recommendations should be given preference as software and processes will not depreciate and create infrastructure to build on further.

It is recommended that flying experience should be improved for the unsatisfied passengers. As outlined by the classification and clustering methods, it has been identified that women passengers, passengers travelling for personal reasons, and passengers travelling in Economy and Economy Plus are likely to be unsatisfied. Therefore, introduction of special discounts and promotions can be implemented to re-introduce these previously dissatisfied customers into the improved quality provided by the airline services. For instance, Airline-X can introduce a loyalty programme and marketed to travelers that have previously travelled with Airline-X and fit the segment's characteristics. Personal travel, one of the defining characteristics among the dissatisfied segment, usually is maximum around the holiday season, therefore holiday discounts on airfare can be one way to initially roll-out the loyalty service among this segment.

The segment that appeared to be more satisfied with the services can be used to derive the characteristics of a likely satisfied customer externally, though these individuals would not have flown with Airline-X previously. The targeted individual that would be male, travelling for business, in business class, which flies distances greater than 1000 miles. These characteristics can be used to target a segment that would be more likely to retain the services over time. Obtaining information on targeting this segment would be much more difficult considering there is no previous internal data on these individuals. External marketing can be conducted to suit the targeted individuals and leverage the qualities previously most admired by the satisfied customer segment including inflight entertainment, seat comfort and cleanliness.

Conclusion

The objective of the request was to improve the perceived value offered to customers, which Airline X desired for strategically timing their growth strategy. The analysis offered various recommendations that will assist in improving the value offered to customers throughout their service-encounter, while offering demographic analysis about the target segment. The solutions were framed around supervised and unsupervised learning methods, including linear regression, logistic regression, decision trees, random forest, and clustering. The solutions provided should enable Airline X to carry out their corporate-level strategy effectively through investment in customer-facing services that most significantly impact customer satisfaction, improving air traffic processes and improving customer retention among the previously unsatisfied customer segment.

References

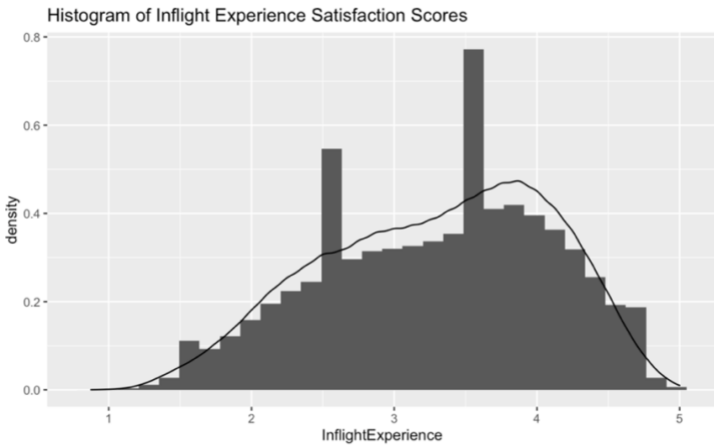
Olaganathan, R. (2021). *Impact of COVID-19 on the Airline Industry and Strategic Plan for its Recovery with Special Reference to Data Analytics Technology.* Embry-Riddle Aeronautical University. Retrieved from https://commons.erau.edu/cgi/viewcontent.cgi?article=2694&context=publication

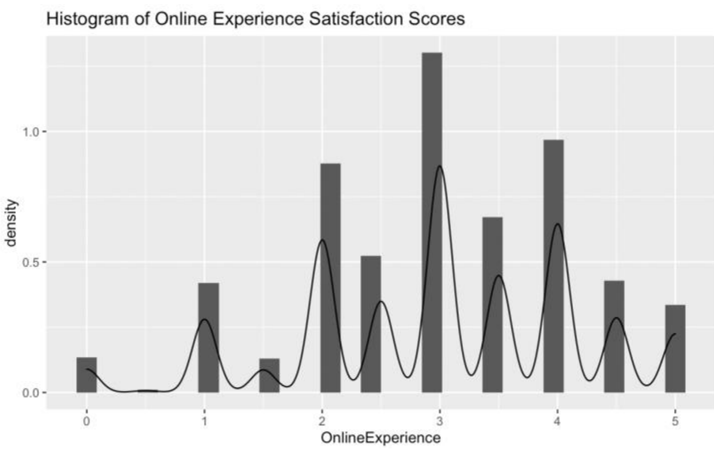
Scott A. Wolla, C. B. (2018). *The Economics of Flying: How Competitive Are the Friendly Skies?* Economic Research Federal Reserve Bank of St. Louis. Retrieved from https://research.stlouisfed.org/publications/page1-econ/2018/11/01/the-economics-of-flying-how-competitive-are-the-friendly-skies

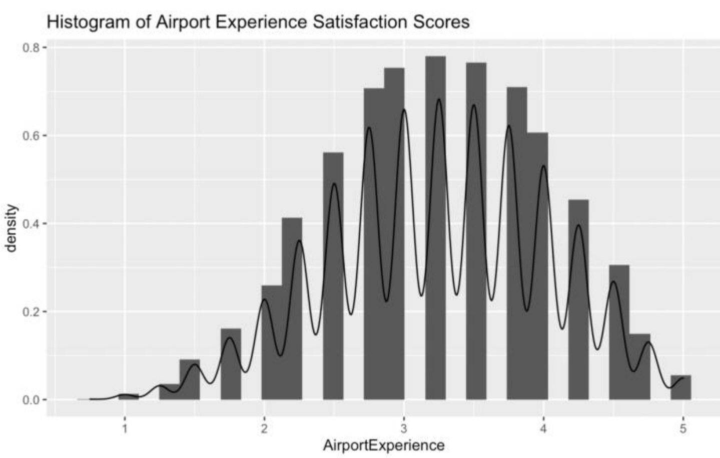
Appendix

Appendix 1: Data exploration of the survey dataset

a. Distribution of Inflight, Onboard, and Airport experience satisfaction scores





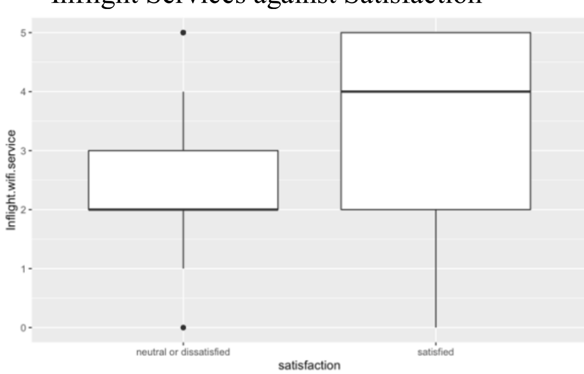


b. Boxplot of airline services to identify services above and below the mean scores

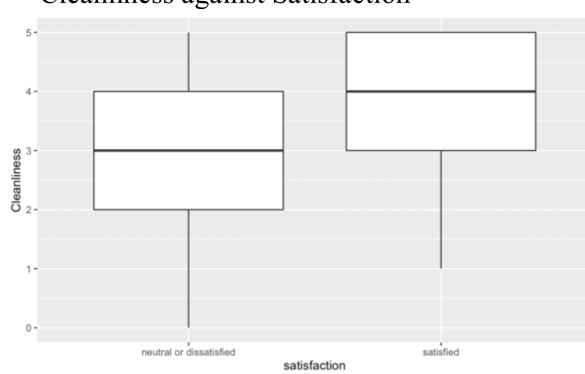
Key: A1- Seat comfort, A2- Departure arrival time convenient, A3- Food and drink, A4- Gate location, A5- Inflight Wi-Fi services, A6- inflight entertainment, A7- Ease of online booking, A8- on board service, A9- leg room services, A10- Baggage handling, A11- Checking service, A12- Cleanliness, A13- online boarding

c. Boxplot of scores of services for satisfied and dissatisfied customers

Inflight Services against Satisfaction



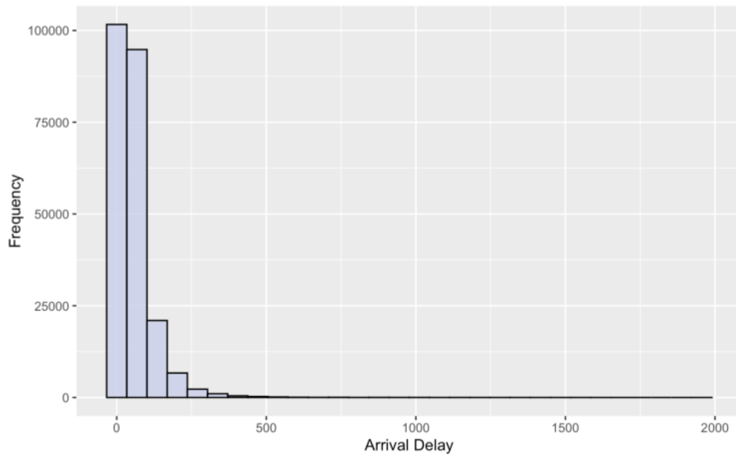
Cleanliness against Satisfaction



Note: Similar graph of other services are created in RMD

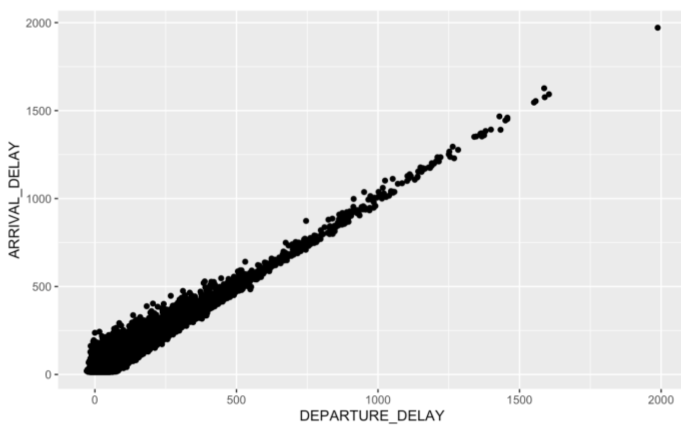
Appendix 2: Data exploration of the arrival delay dataset

a. Distribution of arrival delays

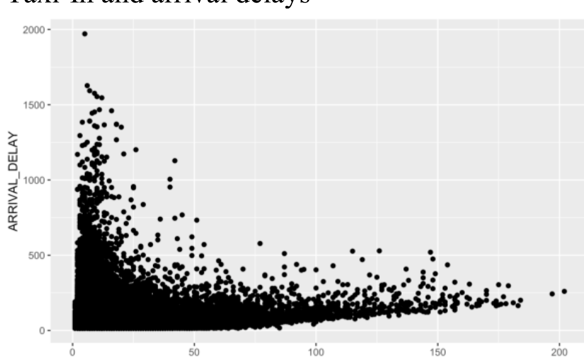


b. Scatter plot of arrival delay against departure delays, taxi-out time, and taxi-in time

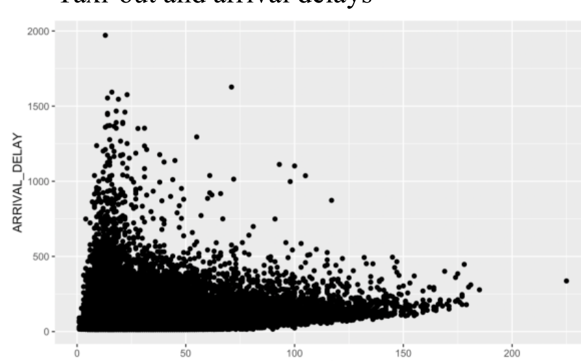
Departure delays and arrival delays



Taxi-In and arrival delays



Taxi-out and arrival delays



Appendix 3: Logistic Regression Results and Errors

a. Results of logistic regression model

```
Coefficients:
                                   Estimate Std. Error z value            Pr(>|z|)
(Intercept)                      -7.5046728  0.0858980 -87.367 < 0.0000000000000002 ***
GenderMale                        0.0486408  0.0214906   2.263              0.0236 *
Customer.TypeLoyal Customer       2.0788437  0.0319685  65.028 < 0.0000000000000002 ***
Age                              -0.0095416  0.0007766 -12.286 < 0.0000000000000002 ***
Type.of.TravelPersonal Travel    -2.6682199  0.0339176 -78.668 < 0.0000000000000002 ***
ClassEco                         -0.7077195  0.0270699 -26.144 < 0.0000000000000002 ***
ClassEco Plus                    -0.7610253  0.0443803 -17.148 < 0.0000000000000002 ***
Inflight.wifi.service             0.4257968  0.0128655  33.096 < 0.0000000000000002 ***
Departure.Arrival.time.convenient -0.1240473  0.0088546 -14.009 < 0.0000000000000002 ***
Ease.of.Online.booking           -0.1819713  0.0126371 -14.400 < 0.0000000000000002 ***
Gate.location                     0.0388148  0.0099518   3.900    0.000096082221676 ***
Food.and.drink                   -0.0580765  0.0120987  -4.800    0.000001584776660 ***
Online.boarding                   0.5993384  0.0112455  53.296 < 0.0000000000000002 ***
Seat.comfort                      0.0841054  0.0122212   6.882    0.000000000005905 ***
Inflight.entertainment            0.0350499  0.0163826   2.139              0.0324 *
On.board.service                  0.3138995  0.0113705  27.606 < 0.0000000000000002 ***
Leg.room.service                  0.2568505  0.0093286  27.534 < 0.0000000000000002 ***
Baggage.handling                  0.1239279  0.0125491   9.875 < 0.0000000000000002 ***
Checkin.service                   0.3073560  0.0093740  32.788 < 0.0000000000000002 ***
Inflight.service                  0.0980817  0.0133884   7.326    0.000000000000237 ***
Cleanliness                       0.2402937  0.0134716  17.837 < 0.0000000000000002 ***
Arrival.Delay.in.Minutes         -0.0502807  0.0022379 -22.467 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 115147  on 83619  degrees of freedom
Residual deviance:  56854  on 83598  degrees of freedom
AIC: 56898

Number of Fisher Scoring iterations: 5
```

b. Confusion Matrix and ROC curve to check the accuracy of the model

```
              Reference
Prediction     dissatisfied satisfied
  dissatisfied        13286      2586
  satisfied            1242      8779

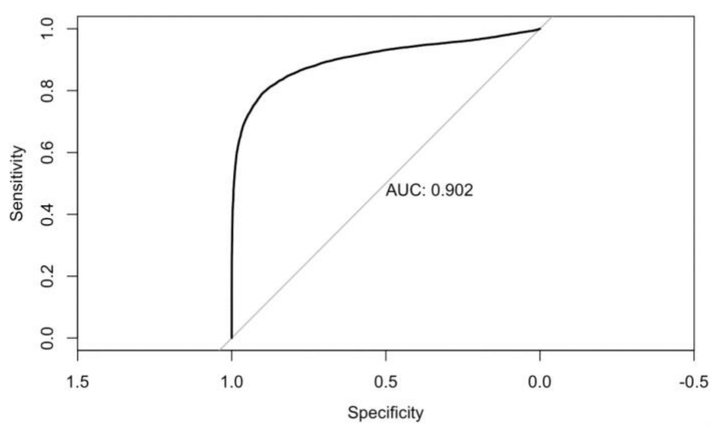
               Accuracy : 0.8522
                 95% CI : (0.8478, 0.8565)
    No Information Rate : 0.5611
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.6959

 Mcnemar's Test P-Value : < 0.00000000000000022

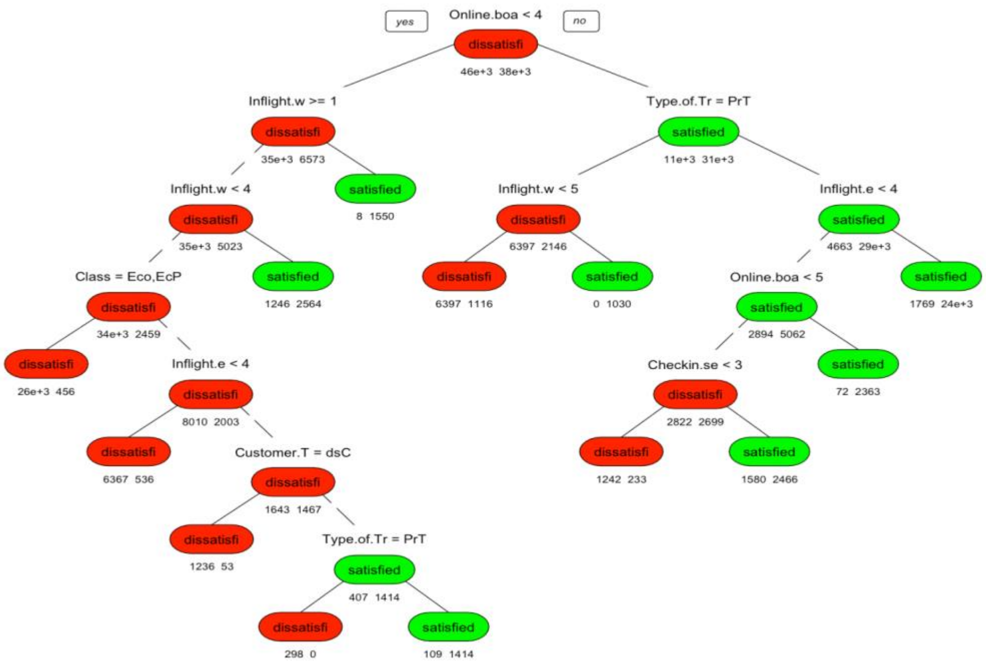
            Sensitivity : 0.7725
            Specificity : 0.9145
         Pos Pred Value : 0.8761
         Neg Pred Value : 0.8371
             Prevalence : 0.4389
         Detection Rate : 0.3390
   Detection Prevalence : 0.3870
      Balanced Accuracy : 0.8435

       'Positive' Class : satisfied
```



Appendix 4: Decision Tree Results and Errors

a. Pruned tree when cp was taken as 0.006



b. Confusion Matrix and ROC curve to check the accuracy of the model

```
Confusion Matrix and Statistics

              Reference
Prediction     dissatisfied satisfied
  dissatisfied        12930       668
  satisfied            1598     10697

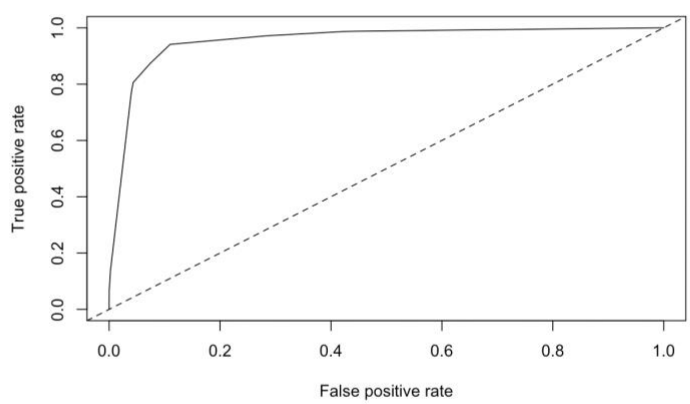
               Accuracy : 0.9125
                 95% CI : (0.909, 0.9159)
    No Information Rate : 0.5611
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.8239

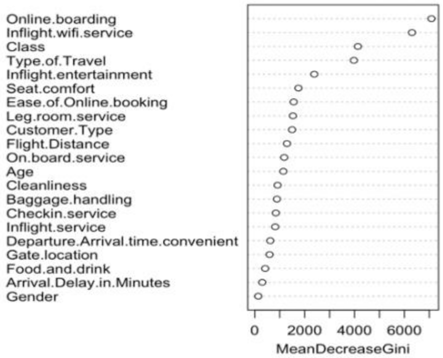
 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9412
            Specificity : 0.8900
         Pos Pred Value : 0.8700
         Neg Pred Value : 0.9509
             Prevalence : 0.4389
         Detection Rate : 0.4131
   Detection Prevalence : 0.4748
      Balanced Accuracy : 0.9156

       'Positive' Class : satisfied
```



c. Important features as described by random forest



d. Confusion Matrix and ROC curve to check the accuracy of the model

```
              Reference
Prediction     dissatisfied satisfied
  dissatisfied        14226       655
  satisfied             302     10710

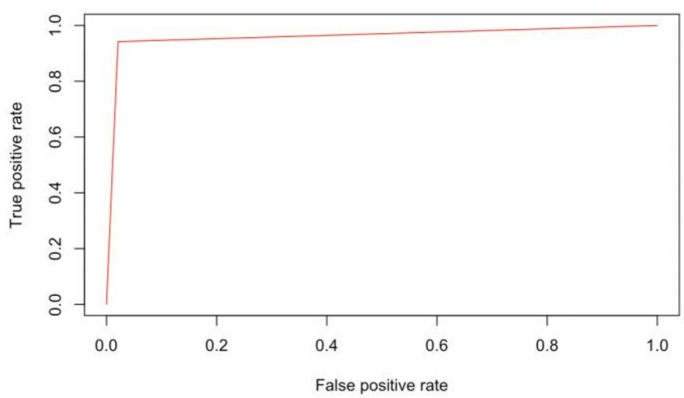
               Accuracy : 0.963
                 95% CI : (0.9607, 0.9653)
    No Information Rate : 0.5611
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.9247

 Mcnemar's Test P-Value : < 0.00000000000000022

            Sensitivity : 0.9792
            Specificity : 0.9424
         Pos Pred Value : 0.9560
         Neg Pred Value : 0.9726
             Prevalence : 0.5611
         Detection Rate : 0.5494
   Detection Prevalence : 0.5747
      Balanced Accuracy : 0.9608

       'Positive' Class : dissatisfied
```



e. Accuracy of the ensemble model

```
A glm ensemble of 3 base models: rpart, glm, rf

Ensemble results:
Generalized Linear Model

8364 samples
   3 predictor
   2 classes: 'dissatisfied', 'satisfied'

No pre-processing
Resampling: Cross-Validated (5 fold, repeated 3 times)
Summary of sample sizes: 6691, 6692, 6691, 6691, 6691, 6692, ...
Resampling results:

  Accuracy   Kappa
  0.9379883  0.8746998
```

Appendix 5: Linear regression Results and Errors

a. Results of logistic regression model

```
Residuals:
     Min       1Q   Median       3Q      Max
-110.597   -5.867    0.028    5.772  110.101

Coefficients:
                    Estimate   Std. Error  t value              Pr(>|t|)
(Intercept)      -13.92128306  0.08371675 -166.290 < 0.0000000000000002 ***
TAXI_OUT           0.67579239  0.00177733  380.229 < 0.0000000000000002 ***
TAXI_IN            0.69579181  0.00252829  275.202 < 0.0000000000000002 ***
DISTANCE          -0.00147131  0.00004061  -36.230 < 0.0000000000000002 ***
DAY_OF_WEEK2       0.18340430  0.08684220    2.112               0.0347 *
DAY_OF_WEEK3       0.05091912  0.09063409    0.562               0.5742
DAY_OF_WEEK4       0.06992588  0.08394805    0.833               0.4049
DAY_OF_WEEK5      -0.25467196  0.08423840   -3.023               0.0025 **
DAY_OF_WEEK6       0.66058868  0.09337559    7.075     0.0000000000015 ***
DAY_OF_WEEK7       1.05557750  0.08322058   12.684 < 0.0000000000000002 ***
DEPARTURE_DELAY    0.91028726  0.00054981 1655.641 < 0.0000000000000002 ***
AIR_SYSTEM_DELAY   0.20758286  0.00116709  177.863 < 0.0000000000000002 ***
SECURITY_DELAY     0.07163513  0.01319258    5.430     0.0000000564388 ***
AIRLINE_DELAY      0.06220866  0.00072552   85.743 < 0.0000000000000002 ***
WEATHER_DELAY      0.07871302  0.00114855   68.533 < 0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.25 on 182807 degrees of freedom
Multiple R-squared:  0.9739,    Adjusted R-squared:  0.9739
F-statistic: 4.879e+05 on 14 and 182807 DF,  p-value: < 0.00000000000000022
```

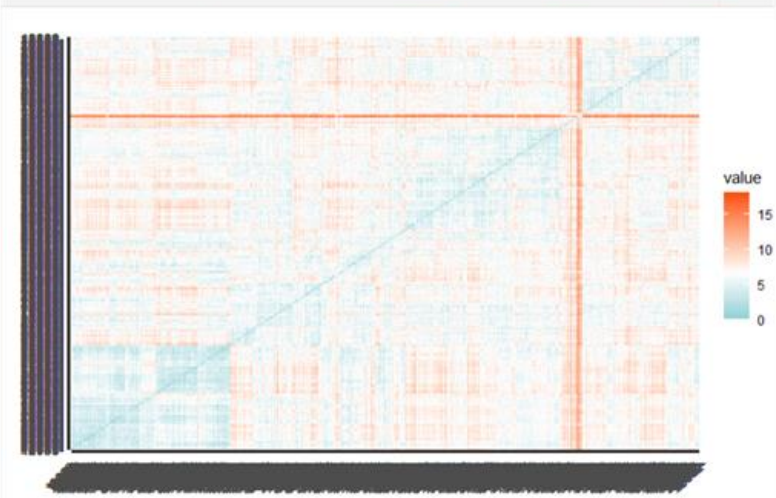
b. Training error

	ME	RMSE	MAE	MPE	MAPE
Test set	0.01640536	10.22251	7.588153	-1.80699	20.9394

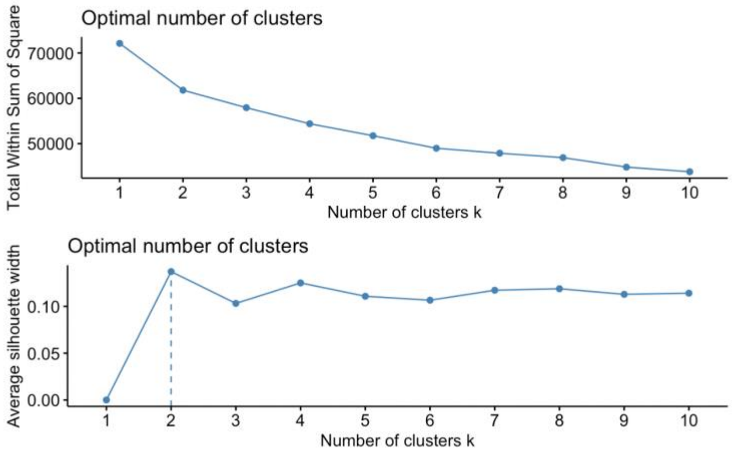
Appendix 6: K-means Clustering Results

a. Distance matrix for the normalized flight survey dataset

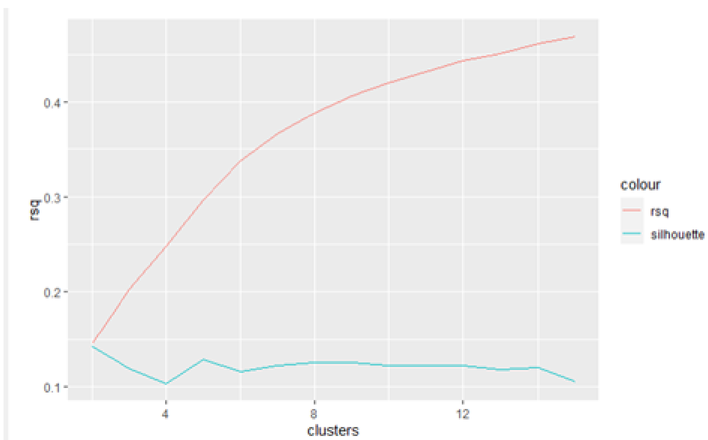
```
```{r  distance matrix}
distance <- get_dist(flight_clean_standardized, method = "euclidean")
fviz_dist(distance, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```
```



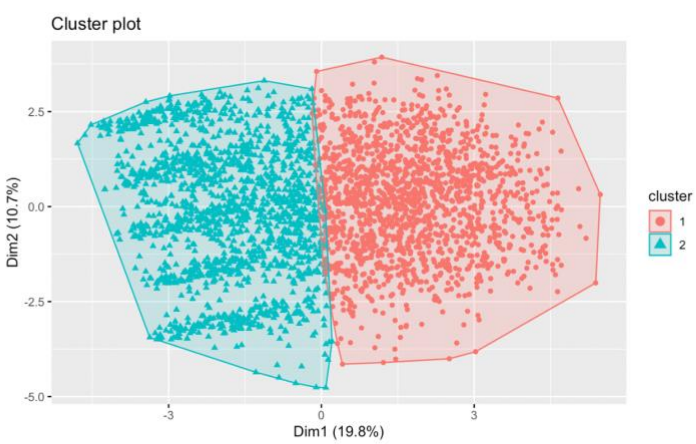
b. Optimum clusters with Silhouette and WSS Method



c. Dual Axis Chart Plotting R-Squared and Average Silhouette Width per Cluster Size



d. Clusters



e. Mean results grouped by clusters and cluster summary

Cluster <fctr>	Gender <dbl>	Customer.Type <dbl>	Age <dbl>	Type.of.Travel <dbl>	Class <dbl>	Flight.Distance <dbl>	Inflight.wifi.service <dbl>
1	1.487959	1.777567	37.51141	1.475285	1.849810	919.730	2.329531
2	1.509301	1.855035	41.09044	1.168056	1.349583	1446.327	3.186017

Departure.Arrival.time.convenient <dbl>	Ease.of.Online.booking <dbl>	Gate.location <dbl>	Food.and.drink <dbl>	Online.boarding <dbl>	Seat.comfort <dbl>
3.078580	2.503169	2.944233	2.558302	2.586819	2.693916
3.128287	3.096857	3.036562	3.874920	3.962797	4.219371

Inflight.entertainment <dbl>	On.board.service <dbl>	Leg.room.service <dbl>	Baggage.handling <dbl>	Checkin.service <dbl>	Inflight.service <dbl>	Cleanliness <dbl>
2.436629	2.890368	2.880228	3.205957	2.967047	3.266793	2.529151
4.314304	3.917255	3.831944	4.116742	3.702373	4.128287	4.091084

Cleanliness <dbl>	Departure.Delay.in.Minutes <dbl>	Arrival.Delay.in.Minutes <dbl>	satisfaction <dbl>
2.529151	3.522814	2.963245	1.114702
4.091084	2.835792	2.261065	1.758178

```
[[1]]
     Gender     Customer.Type        Age        Type.of.Travel      Class      Flight.Distance  Inflight.wifi.service
 Min.   :1.000  Min.   :1.000  Min.   : 7.00  Min.   :1.000  Min.   :1.00  Min.   :  67.0  Min.   :0.00
 1st Qu.:1.000  1st Qu.:2.000  1st Qu.:25.00  1st Qu.:1.000  1st Qu.:1.00  1st Qu.: 368.0  1st Qu.:2.00
 Median :1.000  Median :2.000  Median :37.00  Median :1.000  Median :2.00  Median : 666.0  Median :2.00
 Mean   :1.488  Mean   :1.778  Mean   :37.51  Mean   :1.475  Mean   :1.85  Mean   : 919.7  Mean   :2.33
 3rd Qu.:2.000  3rd Qu.:2.000  3rd Qu.:49.00  3rd Qu.:2.000  3rd Qu.:2.00  3rd Qu.:1107.0  3rd Qu.:3.00
 Max.   :2.000  Max.   :2.000  Max.   :80.00  Max.   :2.000  Max.   :3.00  Max.   :4243.0  Max.   :5.00
 Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding  Seat.comfort   Inflight.entertainment
 Min.   :0.000                     Min.   :0.000          Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :1.000
 1st Qu.:2.000                     1st Qu.:2.000          1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000
 Median :3.000                     Median :2.000          Median :3.000   Median :2.000   Median :3.000   Median :3.000   Median :2.000
 Mean   :3.079                     Mean   :2.503          Mean   :2.944   Mean   :2.558   Mean   :2.587   Mean   :2.694   Mean   :2.437
 3rd Qu.:4.000                     3rd Qu.:3.000          3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:3.000
 Max.   :5.000                     Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000
 On.board.service Leg.room.service Baggage.handling Checkin.service Inflight.service  Cleanliness    Departure.Delay.in.Minutes
 Min.   :1.00     Min.   :0.00     Min.   :1.000    Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000
 1st Qu.:2.00     1st Qu.:2.00     1st Qu.:2.000    1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 0.000
 Median :3.00     Median :3.00     Median :3.000    Median :3.000   Median :3.000   Median :2.000   Median : 0.000
 Mean   :2.89     Mean   :2.88     Mean   :3.206    Mean   :2.967   Mean   :3.267   Mean   :2.529   Mean   : 3.523
 3rd Qu.:4.00     3rd Qu.:4.00     3rd Qu.:4.000    3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.: 4.000
 Max.   :5.00     Max.   :5.00     Max.   :5.000    Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :41.000
 Arrival.Delay.in.Minutes  satisfaction    Cluster
 Min.   : 0.000           Min.   :1.000   1:1578
 1st Qu.: 0.000           1st Qu.:1.000   2:   0
 Median : 0.000           Median :1.000
 Mean   : 2.963           Mean   :1.115
 3rd Qu.: 4.000           3rd Qu.:1.000
 Max.   :20.000           Max.   :2.000
[[2]]
     Gender      Customer.Type        Age         Type.of.Travel       Class      Flight.Distance Inflight.wifi.service Departure.Arrival.time.convenient
 Min.   :1.000  Min.   :1.000  Min.   : 7.00  Min.   :1.000  Min.   :1.00  Min.   :  31  Min.   :0.000         Min.   :0.000
 1st Qu.:1.000  1st Qu.:2.000  1st Qu.:30.00  1st Qu.:1.000  1st Qu.:1.00  1st Qu.: 489  1st Qu.:2.000         1st Qu.:2.000
 Median :2.000  Median :2.000  Median :42.00  Median :1.000  Median :1.00  Median :1085  Median :3.000         Median :3.000
 Mean   :1.509  Mean   :1.855  Mean   :41.09  Mean   :1.168  Mean   :1.35  Mean   :1446  Mean   :3.186         Mean   :3.128
 3rd Qu.:2.000  3rd Qu.:2.000  3rd Qu.:52.00  3rd Qu.:1.000  3rd Qu.:2.00  3rd Qu.:2264  3rd Qu.:4.000         3rd Qu.:4.000
 Max.   :2.000  Max.   :2.000  Max.   :80.00  Max.   :2.000  Max.   :3.00  Max.   :4983  Max.   :5.000         Max.   :5.000
 Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding  Seat.comfort    Inflight.entertainment On.board.service Leg.room.service
 Min.   :0.000          Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :1.000          Min.   :1.000    Min.   :1.000
 1st Qu.:2.000          1st Qu.:2.000   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.:4.000          1st Qu.:3.000    1st Qu.:3.000
 Median :3.000          Median :3.000   Median :4.000   Median :4.000   Median :4.000   Median :4.000          Median :4.000    Median :4.000
 Mean   :3.097          Mean   :3.037   Mean   :3.875   Mean   :3.963   Mean   :4.219   Mean   :4.314          Mean   :3.917    Mean   :3.832
 3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000          3rd Qu.:5.000    3rd Qu.:5.000
 Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000    Max.   :5.000
 Baggage.handling Checkin.service Inflight.service  Cleanliness     Departure.Delay.in.Minutes Arrival.Delay.in.Minutes  satisfaction    Cluster
 Min.   :1.000    Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 0.000             Min.   : 0.000           Min.   :1.000   1:   0
 1st Qu.:4.000    1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000   1st Qu.: 0.000             1st Qu.: 0.000           1st Qu.:2.000   2:1559
 Median :4.000    Median :4.000   Median :4.000   Median :4.000   Median : 0.000             Median : 0.000           Median :2.000
 Mean   :4.117    Mean   :3.702   Mean   :4.128   Mean   :4.091   Mean   : 2.836             Mean   : 2.261           Mean   :1.758
 3rd Qu.:5.000    3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.: 3.000             3rd Qu.: 1.000           3rd Qu.:2.000
 Max.   :5.000    Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :38.000             Max.   :20.000           Max.   :2.000
```

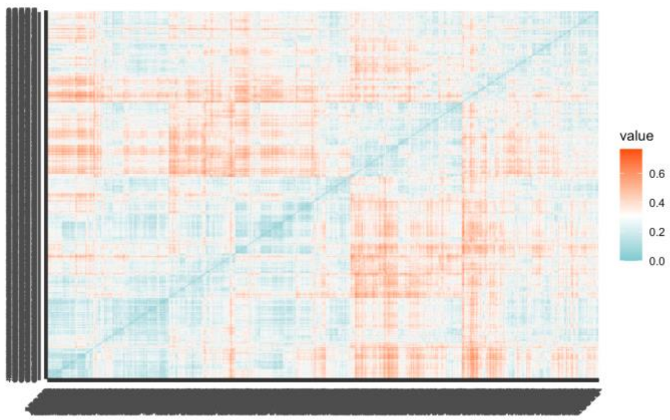
f. Sample Satisfaction Distribution by Cluster

```
                          ddataCat.cluster
flight_clean.satisfaction       1       2
                        1 21.2766 89.8827
                        2 78.7234 10.1173
```

Appendix 7: PAM Clustering Results

a. Distance matrix for the normalized flight survey dataset

```
gower_dist<-daisy(flight_clean_PAM_standardized,metric="gower")
fviz_dist(gower_dist, gradient = list(low = "#00AFBB", mid = "white", high = "#FC4E07"))
```



b. Cluster Summary

```
[[1]]
   Gender            Customer.Type      Age               Type.of.Travel      Class      Flight.Distance   Inflight.wifi.service Departure.Arrival.time.convenien
 Female:565   disloyal Customer:232   Min.   : 7.0   Business travel:646   Business:375   Min.   :  67.0   Min.   :0.000        Min.   :0.000
 Male  :556   Loyal Customer   :889   1st Qu.:25.0   Personal Travel:475   Eco     :617   1st Qu.: 384.0   1st Qu.:2.000        1st Qu.:2.000
                                      Median :38.0                         Eco Plus:129   Median : 696.0   Median :2.000        Median :3.000
                                      Mean   :38.2                                        Mean   : 969.6   Mean   :2.401        Mean   :3.056
                                      3rd Qu.:51.0                                        3rd Qu.:1235.0   3rd Qu.:3.000        3rd Qu.:4.000
                                      Max.   :85.0                                        Max.   :3998.0   Max.   :5.000        Max.   :5.000
 Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding  Seat.comfort   Inflight.entertainment On.board.service Leg.room.service Baggage.handling
 Min.   :0.000          Min.   :1.000   Min.   :0.000   Min.   :0.000   Min.   :1.000   Min.   :1.000          Min.   :1.000   Min.   :0.000   Min.   :1.00
 1st Qu.:2.000          1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000          1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00
 Median :3.000          Median :3.000   Median :2.000   Median :3.000   Median :2.000   Median :2.000          Median :3.000   Median :3.000   Median :3.00
 Mean   :2.664          Mean   :3.003   Mean   :2.235   Mean   :2.759   Mean   :2.354   Mean   :1.951          Mean   :2.722   Mean   :2.797   Mean   :3.12
 3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:4.000   3rd Qu.:3.000   3rd Qu.:2.000          3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00
 Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.00
 Checkin.service Inflight.service  Cleanliness   Departure.Delay.in.Minutes Arrival.Delay.in.Minutes               satisfaction Cluster      cluster
 Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   : 0.000             Min.   : 0.00          neutral or dissatisfied:900   1:1121   Min.   :1
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.00   1st Qu.: 0.000             1st Qu.: 0.00          satisfied              :221   2:   0   1st Qu.:1
 Median :3.000   Median :3.000   Median :2.00   Median : 0.000             Median : 0.00                                                 Median :1
 Mean   :2.968   Mean   :3.169   Mean   :2.16   Mean   : 3.277             Mean   : 2.89                                                 Mean   :1
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:3.00   3rd Qu.: 3.000             3rd Qu.: 4.00                                                 3rd Qu.:1
 Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :39.000             Max.   :20.00                                                 Max.   :1
```

```
[[2]]
   Gender             Customer.Type       Age                Type.of.Travel       Class       Flight.Distance Inflight.wifi.service
 Female:1020   disloyal Customer: 332   Min.   : 7.00   Business travel:1494   Business:1159   Min.   :  31   Min.   :0.000
 Male  : 997   Loyal Customer   :1685   1st Qu.:29.00   Personal Travel: 523   Eco     : 737   1st Qu.: 453   1st Qu.:2.000
                                        Median :41.00                          Eco Plus: 121   Median : 937   Median :3.000
                                        Mean   :40.18                                          Mean   :1312   Mean   :2.917
                                        3rd Qu.:51.00                                          3rd Qu.:2062   3rd Qu.:4.000
                                        Max.   :85.00                                          Max.   :4963   Max.   :5.000
 Departure.Arrival.time.convenient Ease.of.Online.booking Gate.location   Food.and.drink  Online.boarding Seat.comfort    Inflight.entertainment On.board.service
 Min.   :0.000                     Min.   :0.000          Min.   :1.000   Min.   :1.000   Min.   :0.000   Min.   :1.000   Min.   :1.000          Min.   :1.000
 1st Qu.:2.000                     1st Qu.:2.000          1st Qu.:2.000   1st Qu.:3.000   1st Qu.:3.000   1st Qu.:4.000   1st Qu.:4.000          1st Qu.:3.000
 Median :3.000                     Median :3.000          Median :3.000   Median :4.000   Median :4.000   Median :4.000   Median :4.000          Median :4.000
 Mean   :3.046                     Mean   :2.813          Mean   :2.963   Mean   :3.784   Mean   :3.537   Mean   :4.064   Mean   :4.206          Mean   :3.804
 3rd Qu.:4.000                     3rd Qu.:4.000          3rd Qu.:4.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:5.000          3rd Qu.:5.000
 Max.   :5.000                     Max.   :5.000          Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000   Max.   :5.000          Max.   :5.000
 Leg.room.service Baggage.handling Checkin.service Inflight.service  Cleanliness    Departure.Delay.in.Minutes Arrival.Delay.in.Minutes
 Min.   :0.000    Min.   :1.000    Min.   :1.000   Min.   :1.000    Min.   :1.000   Min.   : 0.000             Min.   : 0.000
 1st Qu.:3.000    1st Qu.:4.000    1st Qu.:3.000   1st Qu.:4.000    1st Qu.:3.000   1st Qu.: 0.000             1st Qu.: 0.000
 Median :4.000    Median :4.000    Median :4.000   Median :4.000    Median :4.000   Median : 0.000             Median : 0.000
 Mean   :3.674    Mean   :3.971    Mean   :3.517   Mean   :3.986    Mean   :3.953   Mean   : 2.798             Mean   : 2.339
 3rd Qu.:5.000    3rd Qu.:5.000    3rd Qu.:5.000   3rd Qu.:5.000    3rd Qu.:5.000   3rd Qu.: 2.000             3rd Qu.: 2.000
 Max.   :5.000    Max.   :5.000    Max.   :5.000   Max.   :5.000    Max.   :5.000   Max.   :39.000             Max.   :20.000
                  satisfaction  Cluster     cluster
 neutral or dissatisfied: 806   1:   0   Min.   :2
 satisfied              :1211   2:2017   1st Qu.:2
                                         Median :2
                                         Mean   :2
                                         3rd Qu.:2
                                         Max.   :2
```