SDS322E Project: Predicting Flight Experience

Introduction

Flying on an airplane is a universal experience that many people encounter. More than 45,000 flights, amounting to 2.9 million people, take off and land each day throughout the globe.[2] This amounts to approximately 16.4 million flights every year for around 10.6 billion people. With potential revenue being in the billions, it is no wonder that customer satisfaction leading to retention is of the utmost importance to airline companies' bottom line. There are certain factors that contribute to the overall satisfaction of the passenger. In this project, we have data from a real, anonymous airline company where they surveyed passengers after their flight was complete.[1] Our goal is to determine (1) what variables lead to higher reports of satisfaction and to (2) predict if a customer would have reported a satisfied ride based on their parameter values.

This is a real world issue that many companies face. Corporations use data to gain insights on what to improve in the customer experience. In the context of an airline, if a decision maker knows what passengers tend to find more important and valuable, they can adjust their business model to fit. Similarly, the airline can also know what factors that customers tend to not care about and focus less on that aspect. By doing this, it will lead to a higher retention rate with their customers and better brand recognition. This problem is crucial for the current and future success of the organization.

Data

The data is collected through interviewing passengers after their ride. There is also data collected on the demographics of the rider such as their Gender, Age, Customer Type, Class, and Type of Travel.

There could be bias in this data, such as by interviewer falsification or question order bias, since the airplane riders were interviewed. It is unclear how long after their ride was complete that the company interviewed the passengers. Nonetheless, it can be speculated that the information was collected through online surveys. As depicted in Table 1, there are 23 columns/variables with 129,880 rows. 5 of the variables are character values and 18 are integers. All the variables do not have missing values except for "Arrival.Delay.in.Minutes" which has 393 NA values.

Table 1: Flight Data characterization

```
Rows: 129,880
Columns: 23
$ satisfaction                     <chr>
$ Gender                           <chr>
$ Customer.Type                    <chr>
$ Age                              <int>
$ Type.of.Travel                   <chr>
$ Class                            <chr>
$ Flight.Distance                  <int>
$ Seat.comfort                     <int>
$ Departure.Arrival.time.convenient <int>
$ Food.and.drink                   <int>
$ Gate.location                    <int>
$ Inflight.wifi.service            <int>
$ Inflight.entertainment           <int>
$ Online.support                   <int>
$ Ease.of.Online.booking           <int>
$ On.board.service                 <int>
$ Leg.room.service                 <int>
$ Baggage.handling                 <int>
$ Checkin.service                  <int>
$ Cleanliness                      <int>
$ Online.boarding                  <int>
$ Departure.Delay.in.Minutes       <int>
$ Arrival.Delay.in.Minutes         <int>
```

```
                                   [,1]
satisfaction                          0
Gender                                0
Customer.Type                         0
Age                                   0
Type.of.Travel                        0
Class                                 0
Flight.Distance                       0
Seat.comfort                          0
Departure.Arrival.time.convenient     0
Food.and.drink                        0
Gate.location                         0
Inflight.wifi.service                 0
Inflight.entertainment                0
Online.support                        0
Ease.of.Online.booking                0
On.board.service                      0
Leg.room.service                      0
Baggage.handling                      0
Checkin.service                       0
Cleanliness                           0
Online.boarding                       0
Departure.Delay.in.Minutes            0
Arrival.Delay.in.Minutes            393
```

Since there are missing values in the Arrival Delay in Minutes column, we decided to not use it in our analysis and substitute in Departure Delay in Minutes (which has a high correlation with Arrival Delay that you will see in the analysis section). Overall, the data was very structured and was stored in an Excel CSV file.

The possible values of each column are the following:
Gender: Male, Female
Customer Type: Loyal Customer, Disloyal Customer
Age: 7 - 85 years old
Type of Travel: Business Travel, Personal Travel
Class: Business, Economy Plus, Economy
Flight Distance: In miles 50 - 6951
Seat comfort - Online Boarding: On a scale of 0 - 5 where 0 means not applicable, 1 is the most negative, and 5 is the most positive.
Departure Delay in Minutes: In minutes 0 - 1592
Arrival delay in Minutes: In minutes 0 - 1584

We will call the variables in which the passengers were interviewed the rating variables. The data such as gender, customer type, Age, type of travel and class are the demographic variables.

Exploratory Analysis

Hypothesis

1. Services such as inflight entertainment, ease of online booking, online support, and technology based services in addition to legroom and seat comfort tended to be the more important factors to passenger satisfaction.
2. Females, business travel, business class, and loyal customers have higher correlation to satisfaction.
3. Females and males who are around ages 40-60 have rated their ride as satisfied whereas ages 20-40 for both females and males seemed to be more dissatisfied.
4. Economy class passengers and younger age groups tend to have the most dissatisfied passengers with long flights and delay times and females tend to care more about legroom.

Our first hypothesis was trying to determine the general variables that relate to satisfaction. Our second hypothesis was related to the demographics. Our third hypothesis was related to age and gender. Our fourth hypothesis was trying to be more specific and connect the demographics to certain rating variables.

Table 2: Figure guide

Hypothesis Number	**Figures that relate to hypothesis**
1	**3, 4, 5, 6, 7**
2	**8, 9, 10, 11, 12**
3	**18, 19, 20, 21**
4	**13, 14, 15, 16, 17, 22, 23, 24, 25, 26, 27, 28**

The following table shows computed summary statistics of all numeric variables. We can see that the youngest person who was surveyed was 7 years old and the oldest was 85 years old. It seems like the survey was trying to get as much of a representative population as possible. The rating variables ranged from 2.893 - 3.706.

Table 3: Flight Data Summary Statistics inclusive of all numeric variables

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Age	129880	39.428	15.119	7	27	51	85
Flight.Distance	129880	1981.409	1027.116	50	1359	2544	6951
Seat.comfort	129880	2.839	1.393	0	2	4	5
Departure.Arrival.time.convenient	129880	2.991	1.527	0	2	4	5
Food.and.drink	129880	2.852	1.444	0	2	4	5
Gate.location	129880	2.99	1.306	0	2	4	5
Inflight.wifi.service	129880	3.249	1.319	0	2	4	5
Inflight.entertainment	129880	3.383	1.346	0	2	4	5
Online.support	129880	3.52	1.307	0	3	5	5
Ease.of.Online.booking	129880	3.472	1.306	0	2	5	5
On.board.service	129880	3.465	1.271	0	3	4	5
Leg.room.service	129880	3.486	1.292	0	2	5	5
Baggage.handling	129880	3.696	1.156	1	3	5	5
Checkin.service	129880	3.341	1.261	0	3	4	5
Cleanliness	129880	3.706	1.152	0	3	5	5
Online.boarding	129880	3.353	1.299	0	2	4	5
Departure.Delay.in.Minutes	129880	14.714	38.071	0	0	12	1592
Arrival.Delay.in.Minutes	129487	15.091	38.466	0	0	13	1584

We have data over the ages of people who follow this distribution as shown in Figure 1.The bulk of the responses to the survey were noted to be from passengers in the 20s age group through the 50s. We also see that the mean for the ratings range from 2.839 to 3.707. It looks like the airline company is doing average in terms of their services.

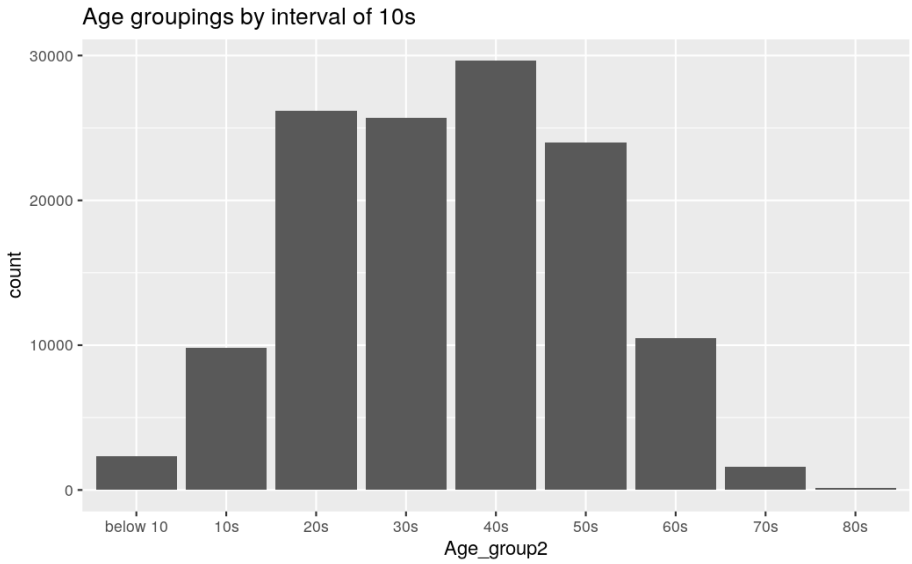

Figure 1: Age distribution of number of responses to the survey.

Figure 2 shows the general count of satisfaction. From the total of 129,880 passengers who were surveyed, there were 71,087 that were satisfied. This is more than half of the data collected which is a good sign for the airline company. People are generally rating their rides as satisfactory.

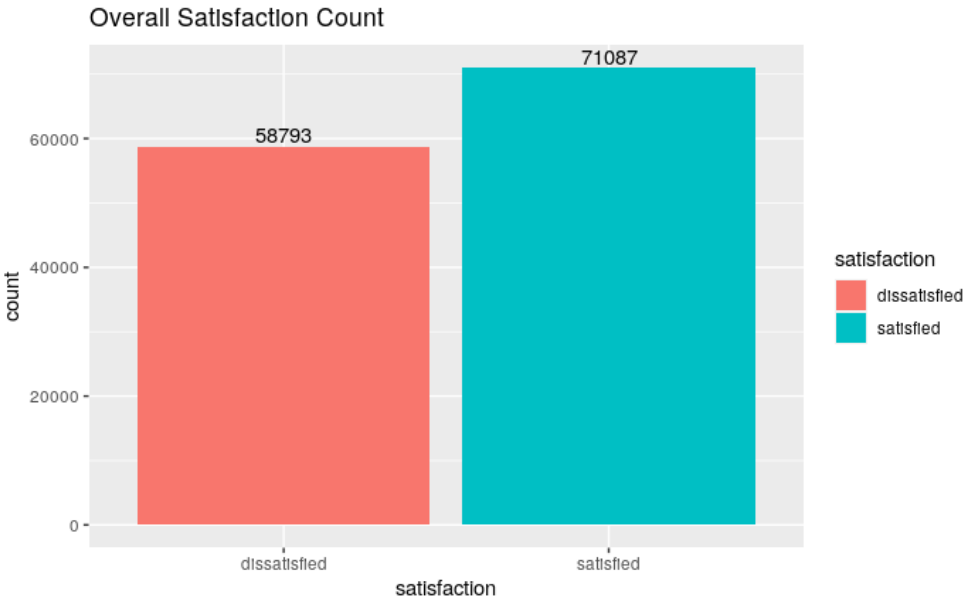

Figure 2: Count of number of passengers that overall showed to be satisfied or dissatisfied

Figure 3 depicts a correlation map between all the different rating variables. This helps visualize if there are notable connections between services. Online boarding and ease of online booking, seat comfort and food and drink, arrival delay and departure delay have a strong positive correlation. Flight distance and age have the highest negative correlation.

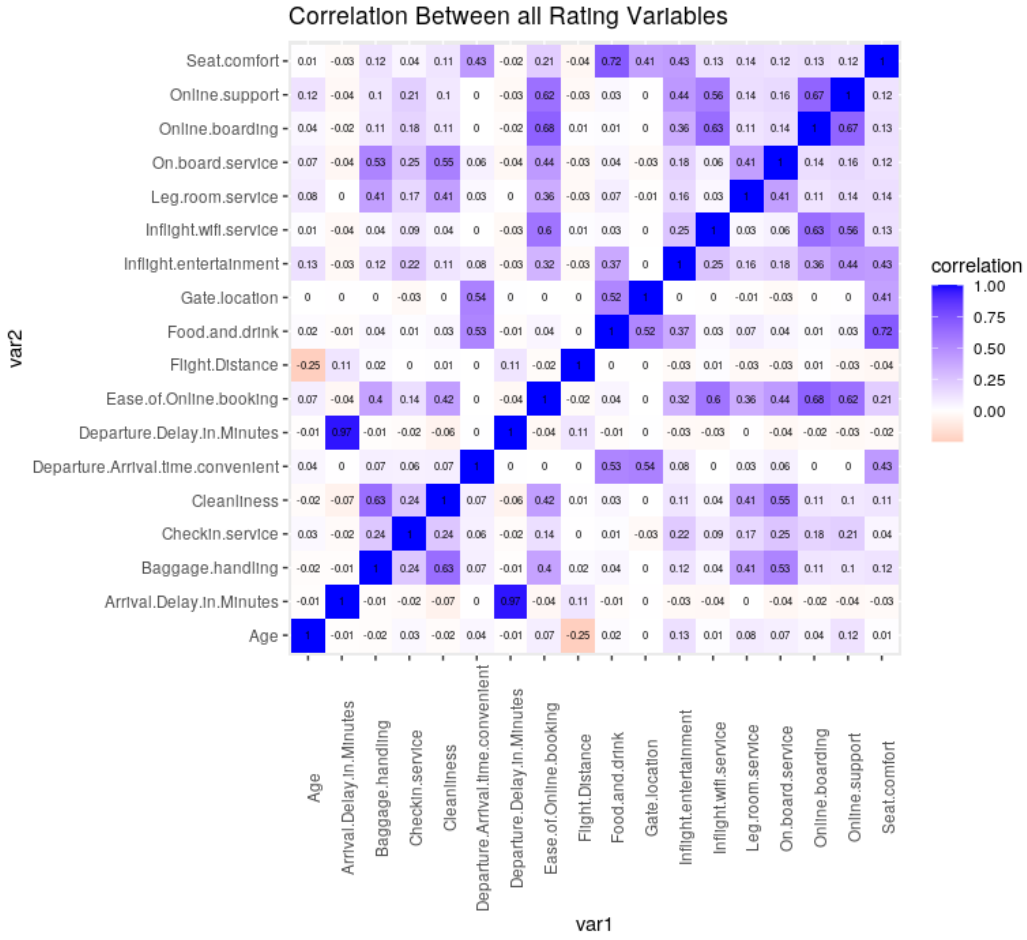

Figure 3: Correlation map of flight data across variables. Departure Delay in Minutes and Arrival Delay in Minutes show the strongest correlation.

In Figure 4, there is a visual representation of the rating variables. From this, we can start to see the variables that have higher ratings therefore leading to more satisfaction. The most distinct are ease of online booking, inflight entertainment, onboard service, and online support. These have a commonality in technology based services and were shown to be correlated in Figure 3.

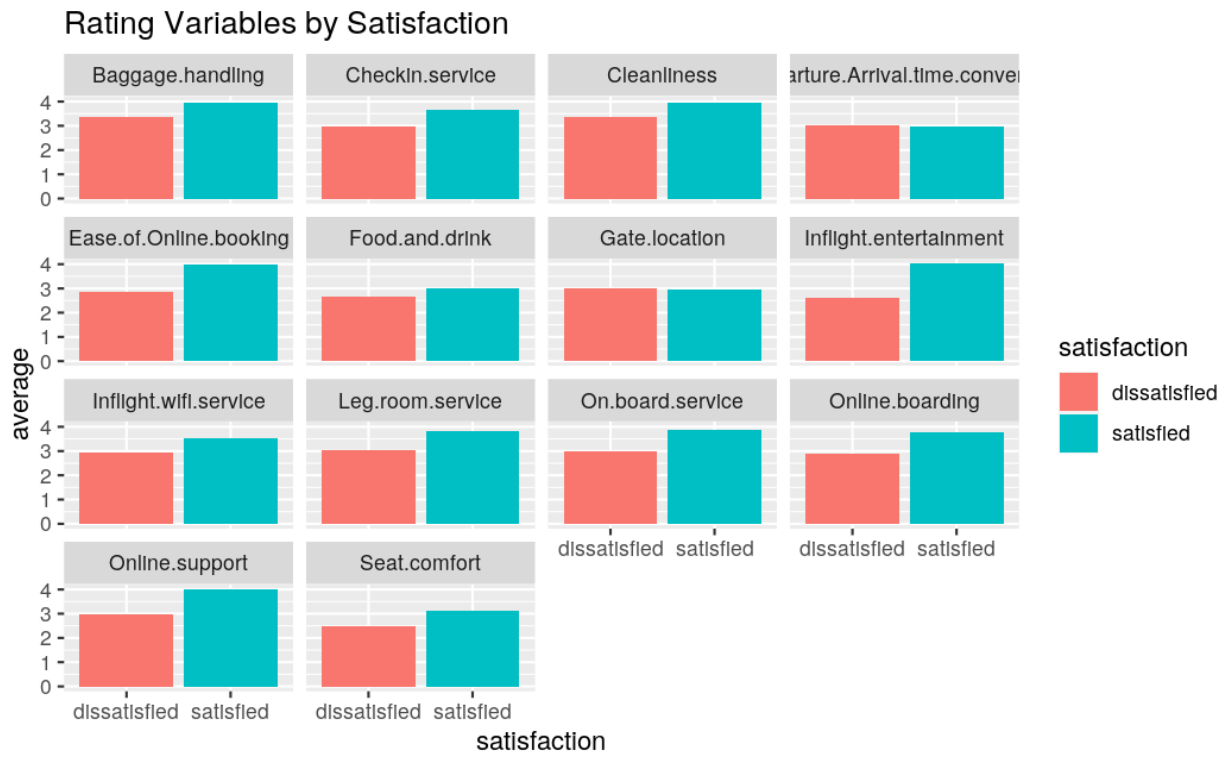

Figure 4: Average satisfaction per each surveyed variable.

Figure 5 shows the difference in the means (satisfaction mean - dissatisfaction mean) between the flight variables. This proves what was discussed previously in how the ease of online booking, inflight entertainment, and online support have the greatest difference between satisfaction levels. It also shows that the departure arrival time convenience and gate location had a higher average rating for dissatisfaction.

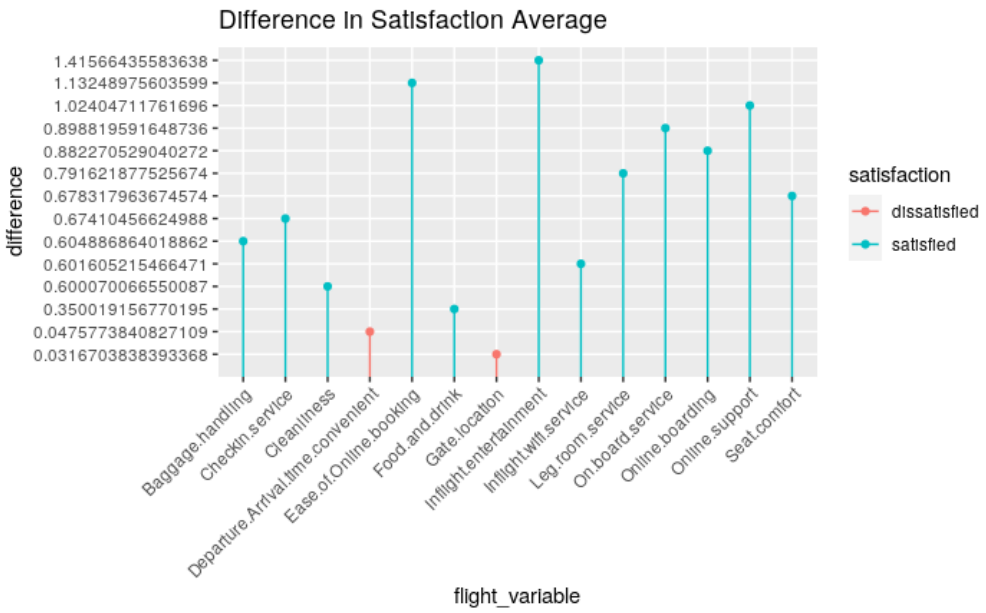

Figure 5: Calculated difference in reported satisfaction average of variables by flight survey question responses. The red means that the dissatisfied mean was higher so instead of subtraction from satisfied - dissatisfied, it was dissatisfied - satisfied.

Figure 6 is another visualization which shows the numeric values of the satisfaction and rating variables. We begin to see the pattern that inflight entertainment and ease of online booking are the two highest. The two rating variables that had a higher dissatisfaction is departure arrival time convenient and gate location.

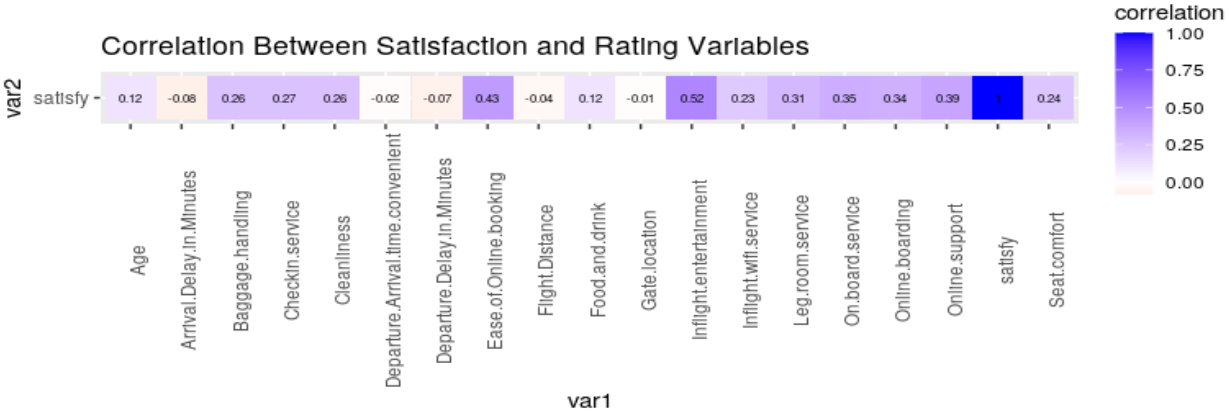

Figure 6: Numeric correlation of satisfaction across the flight data variables

.

Figure 7 depicts the count of 0's with the rating variables. 0 meaning that it was not applicable to the rider. For variables such as food and drink, inflight entertainment, and seat comfort, this can be attributed to a positive thing since there was no obvious issue for the rider to rate it or it just did not impact them at

all. Either one, it is not negative since we know their correlation to satisfaction is positive. However, for departure arrival time to be convenient, since it has the highest amount of 0s and is negatively correlated, this means the riders who rated it must have rated it poorly enough for it to be significant. This implies that the airline should take this variable into more consideration.

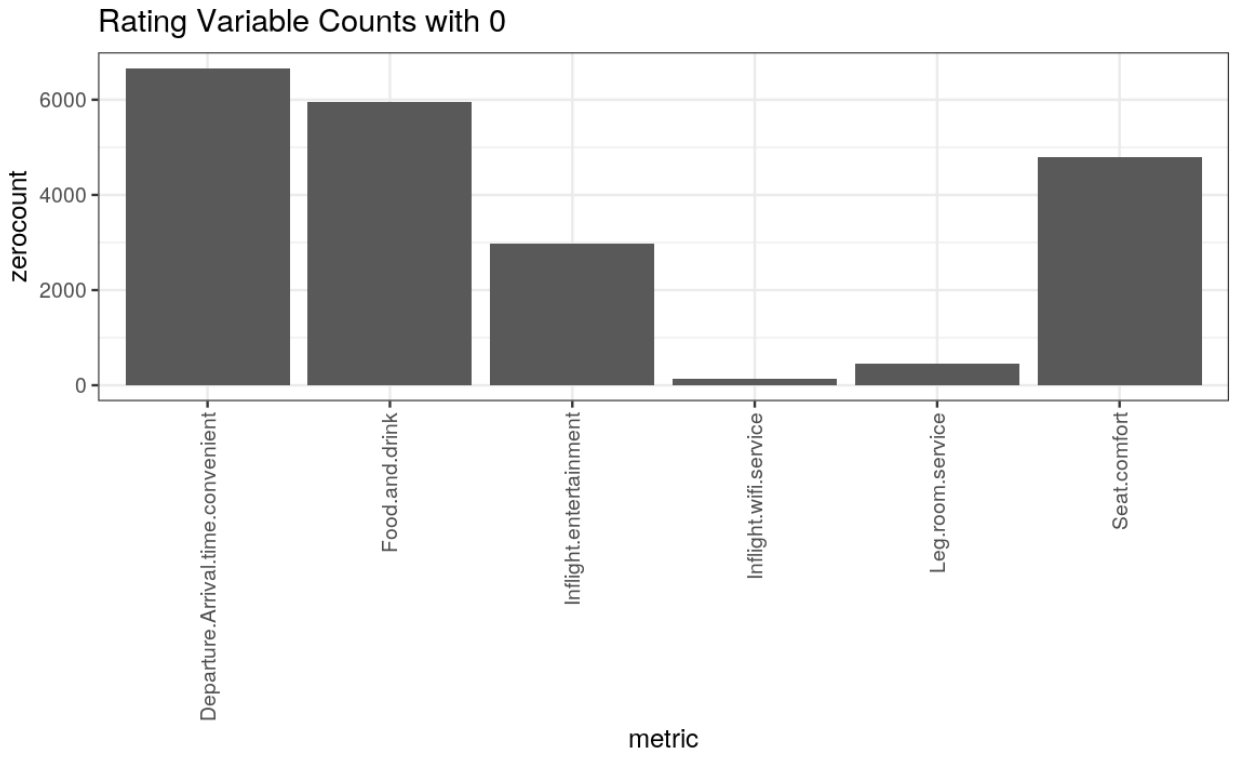

Figure 7: Number of times a survey question was inapplicable to the passengers' experience.

Gender (Figure 8), Customer Type (Figure 9), Class (figure 10), Type of Travel (Figure 11), and Age can be classified as the data relating to the passenger demographics. Females, loyal customers, business class, and business travel tend to have the highest count of passengers who are satisfied.

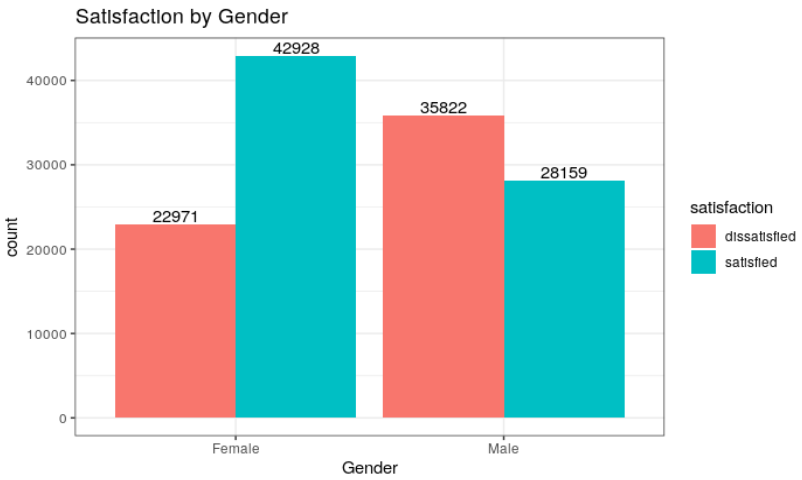

Figure 8: The count of satisfied and dissatisfied passengers based on gender.

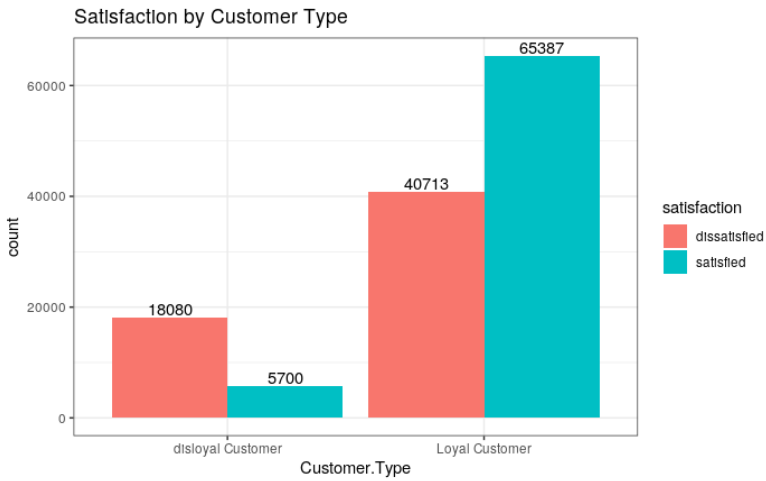

Figure 9: The ratings of satisfaction and dissatisfaction based on customer loyalty.

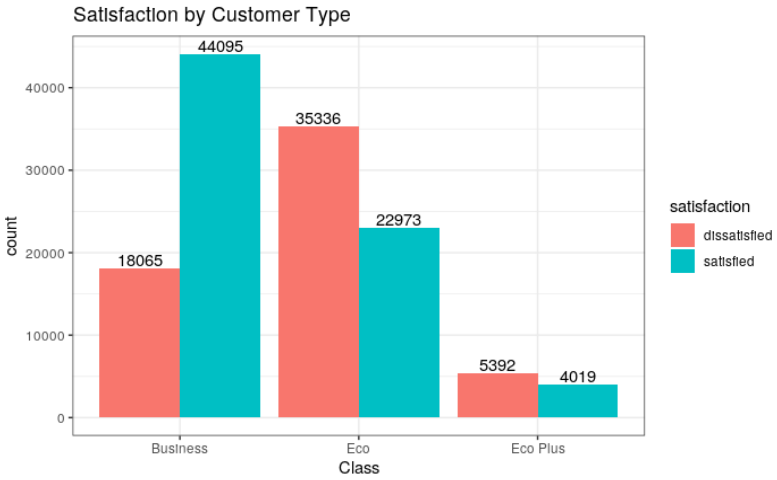

Figure 10: The count of satisfied and dissatisfied Passengers based on one of three classes, Business, Economy, and Economy Plus.

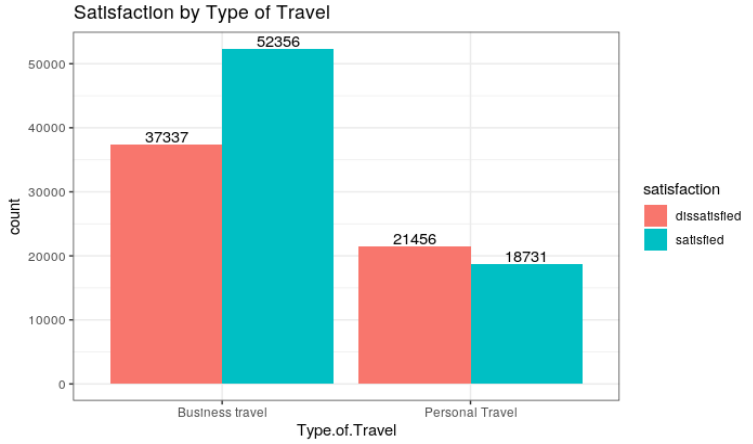

Figure 11: The ratings of satisfaction and dissatisfaction based on type of travel, either personal or business.

For our correlation matrix between satisfaction and all the demographics we have encoded them as follows:

Female 1, Male 0
Business Travel 1, Personal Travel 0
Business Class 2, Eco Plus 1, Eco 0
Loyal Customer 1, Disloyal Customer 0

Satisfaction 1, Dissatisfied 0

The variables were encoded this way based on Figure 8 - Figure 11 which depict the variables female, business travel, business class, and a loyal customer as having the highest reported overall satisfaction. For simplicity, those values were encoded as 1 with their counterpart being 0. This is true for every case except travel class, which had to have business class be encoded as 2, and descending into economy plus as 1 and economy as 0.

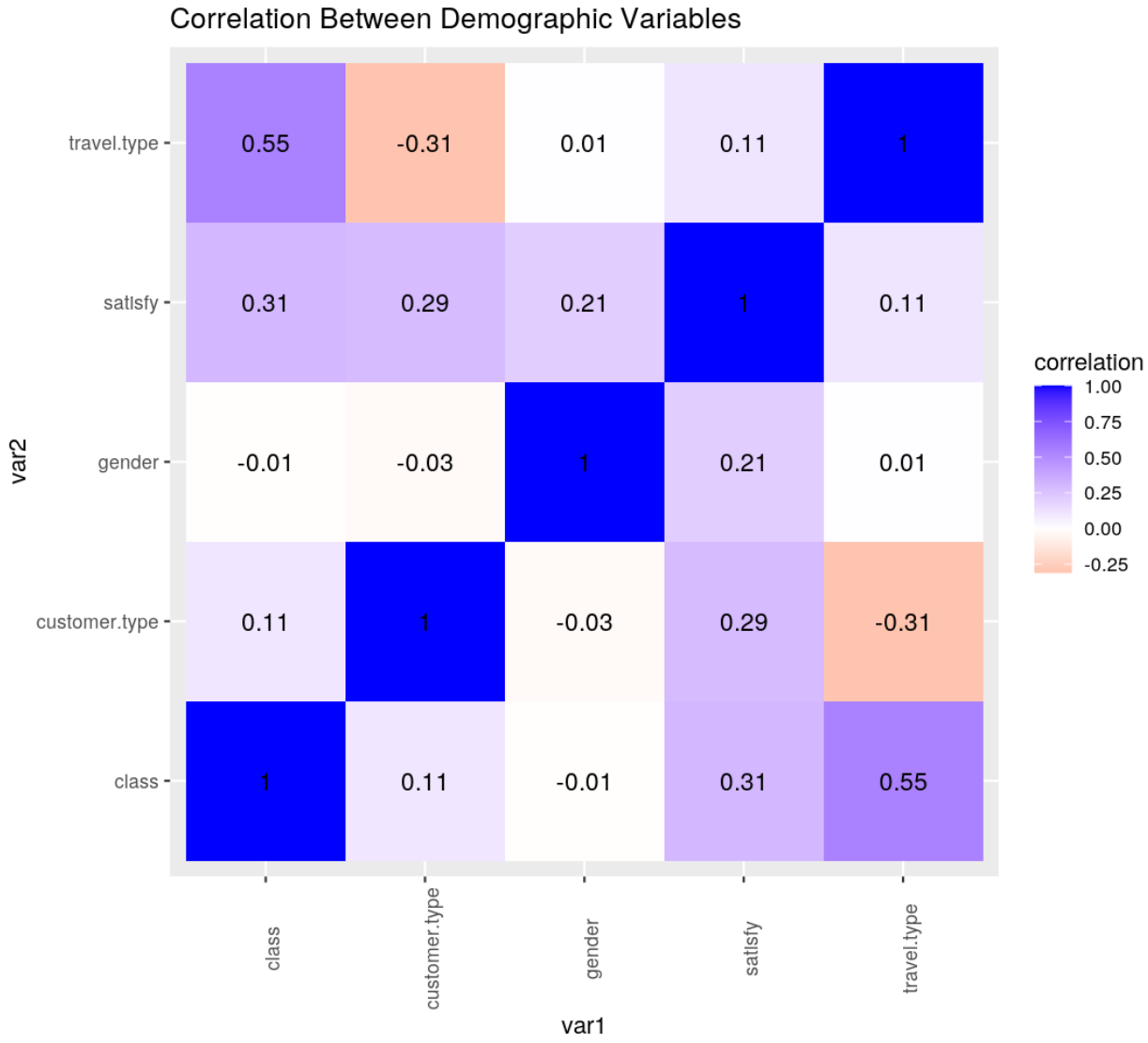

Figure 12: Correlation matrix based on variables used in Figure 8 - Figure 11.

Taking a closer look at gender distribution (Figure 14), across the surveyed variables (Figure 15) leg room has the largest correlation to gender (besides inflight entertainment which was already shown above in Figure 7). Therefore, we are looking more into this variable and how it correlates. This was depicted in Figure 13 which displayed the overall satisfaction scores given by the passenger, as well as their rating of satisfaction of leg room, grouped by gender.

Figure 13: Satisfaction rating reported for leg room based on Gender.

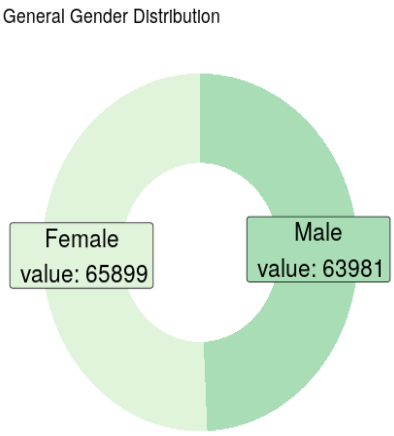

Figure 14: The overall count of Male and Female passengers who participated in the survey.

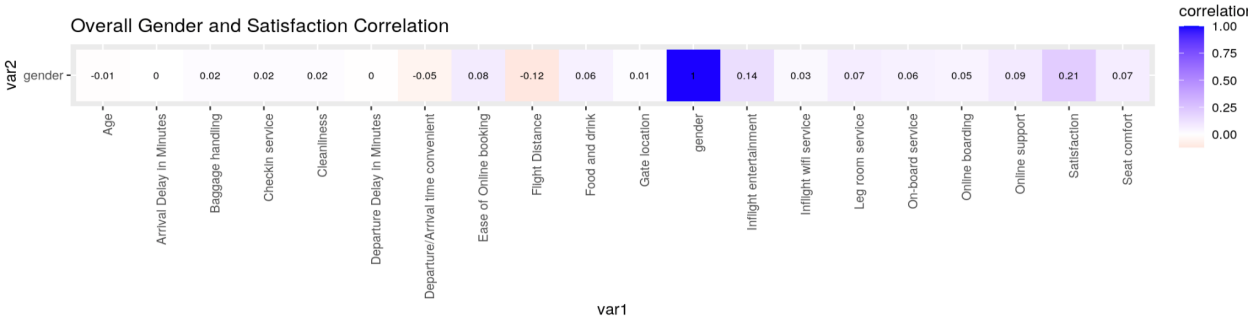

Figure 15: The overall satisfaction correlation of surveyed variables and gender.

When looking at the variables that had a high correlation to satisfaction as depicted in Figure 15, one can see that men care more about inflight entertainment and seat comfort (Figure 16) whereas women care more about leg room service and on board service (Figure 17).

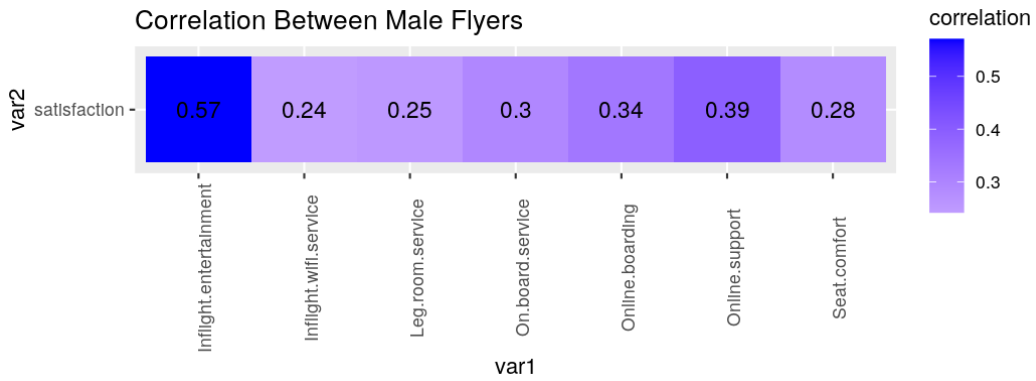

Figure 16: The overall satisfaction correlation between male passengers and surveyed variables that were notable.

While we saw the overall correlation with inflight entertainment today again, there was also a higher correlation with online support and boarding, indicating that amidst the male population there was a higher amount of preference for online mediums affecting their flight experience after inflight entertainment.

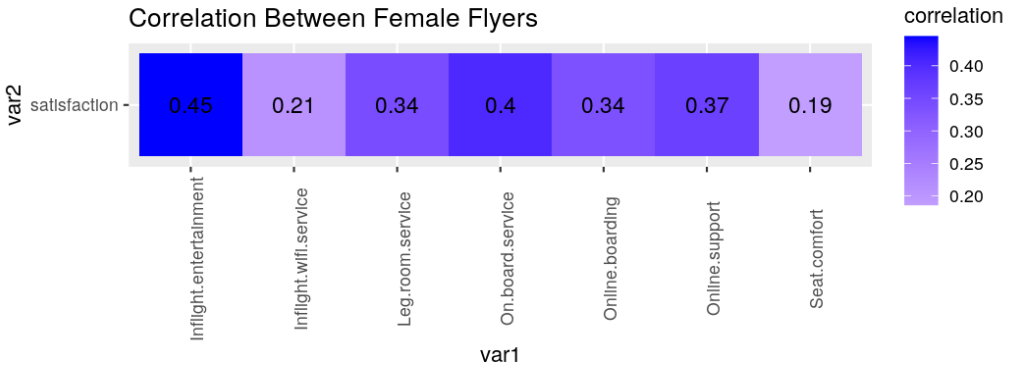

Figure 17: The overall satisfaction correlation between female passengers and surveyed variables that were notable.

Inflight entertainment was still higher with 0.45, but on board service and online support at 0.4 and 0.37 also affected the experience. For the female passengers, this indicates a sense of on board and online experience weighing equally after inflight entertainment.

Age Distribution

The ages we created are categorized as follows: Children 7 - 12, Teenager 13 - 19, Young Adult 20 - 39, Middle Age Adult 40 - 59, Senior Adult 60 - 85, depicted below in Figure 18.

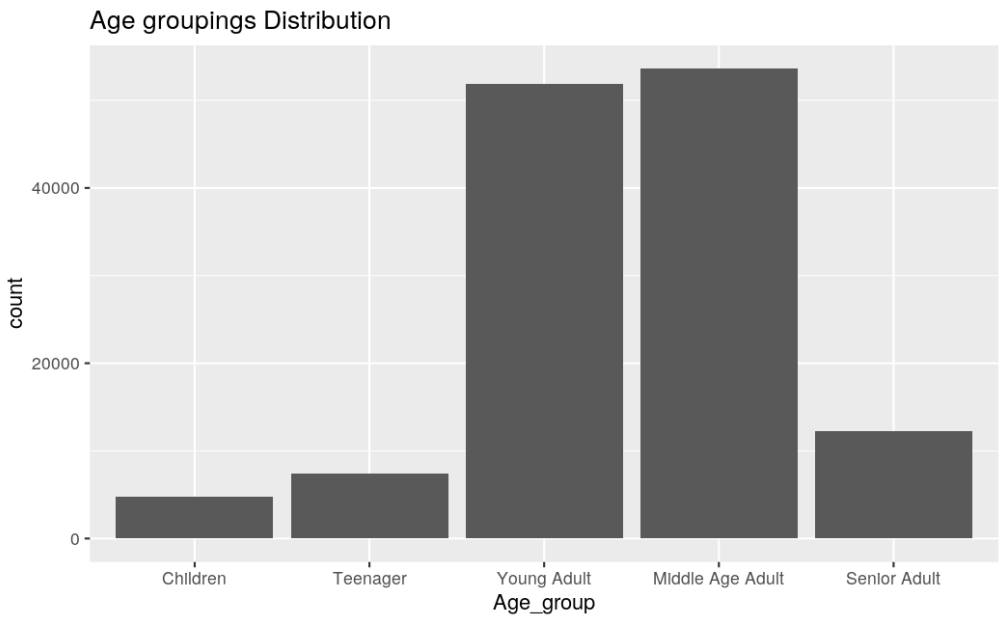

Figure 18: Overall age distribution of passengers based on predefined groupings.

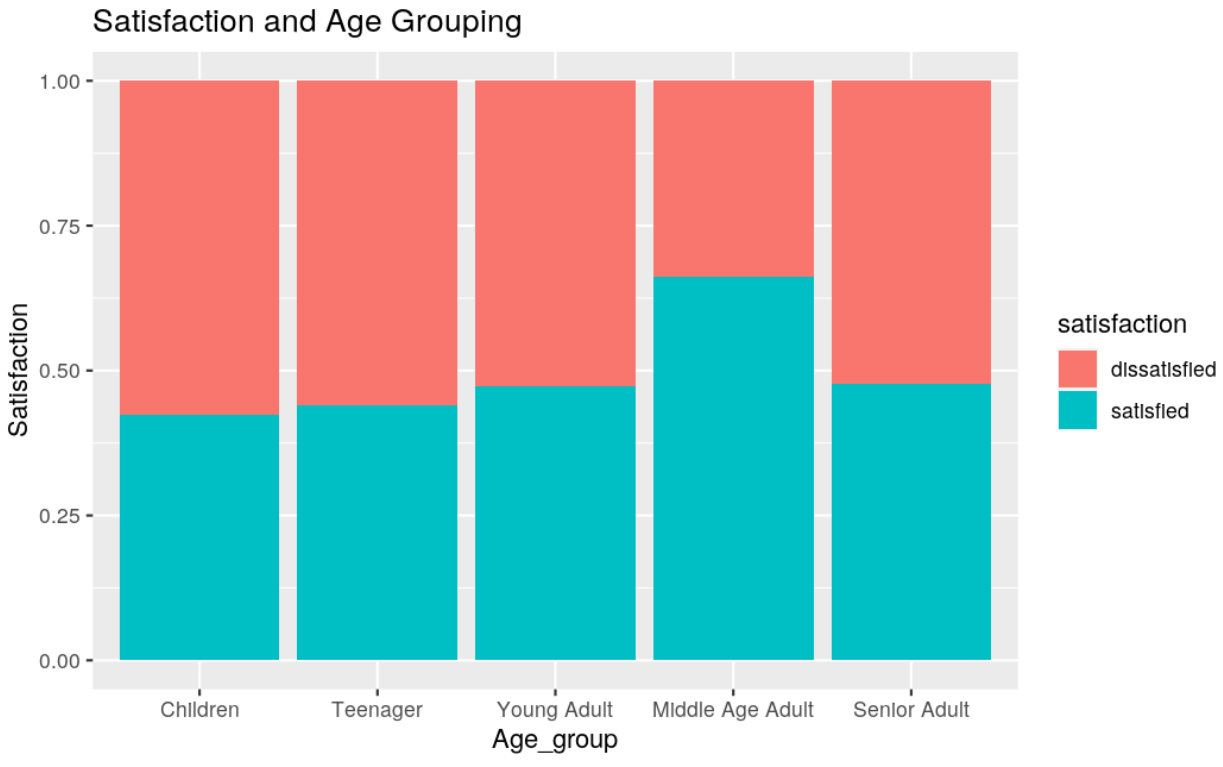

Figure 19: Another age and satisfaction plot but overall

Compared to other age groups, “Middle Age Adult” group shows the highest satisfaction as depicted in Figure 19.

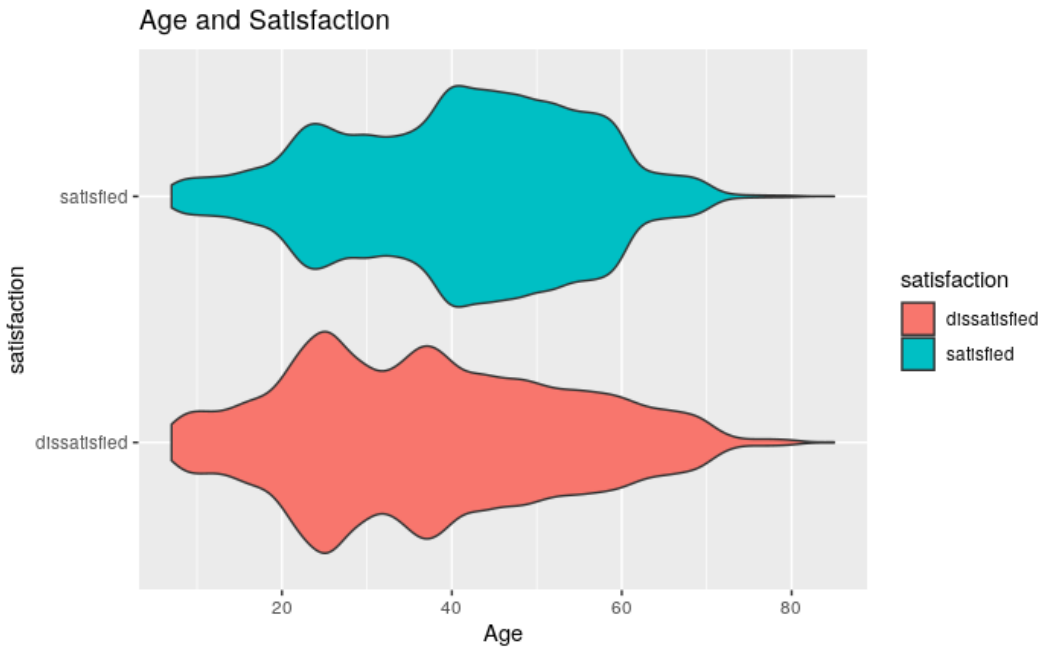

Figure 20: Graph depicting overall satisfaction rating by age.

Figure 20 also shows that ages 40-60 (Middle Aged adults) for both men and women had higher rates of satisfaction whereas 20-40 (Young Adult) for both genders had less satisfaction. The relationship further broken down below in Figure 21.

Figure 21: Graph depicting overall satisfaction rating by age and gender.

Passengers show a gradual decrease in satisfaction as delay time increases, as shown in Figure 22, below. Compared to the "Young Adult" and "Middle Age Adult" groups, "Children", "Teenager" and "Senior Adult" groups show a significant plunge in satisfaction ratio as delay time increases. Specifically, passengers in the age group of "Children" and "Teenager" tend to show a dramatic decline when delay time is over 2 hours. Also, the Middle Aged adult group also cares the most about the delay time.

We can find these out by looking at how steep the satisfaction ratio declines. In this visualization, we will look at the satisfaction ratio gap between flights that are punctual and flights that are more than 3 hours late. For "Children", "Teenager" and "Senior Adult", the gaps are 0.42%p, 0.37%p and 0.32%p, respectively. In contrast, the gaps are 0.18%p and 0.17%p for "Young Adult" and "Middle Age Adult" groups, respectively.

Not only younger age groups such as "Children" and "Teenager", but also "Senior Adult" groups who are the oldest group tend to care a lot about the extended delay time.

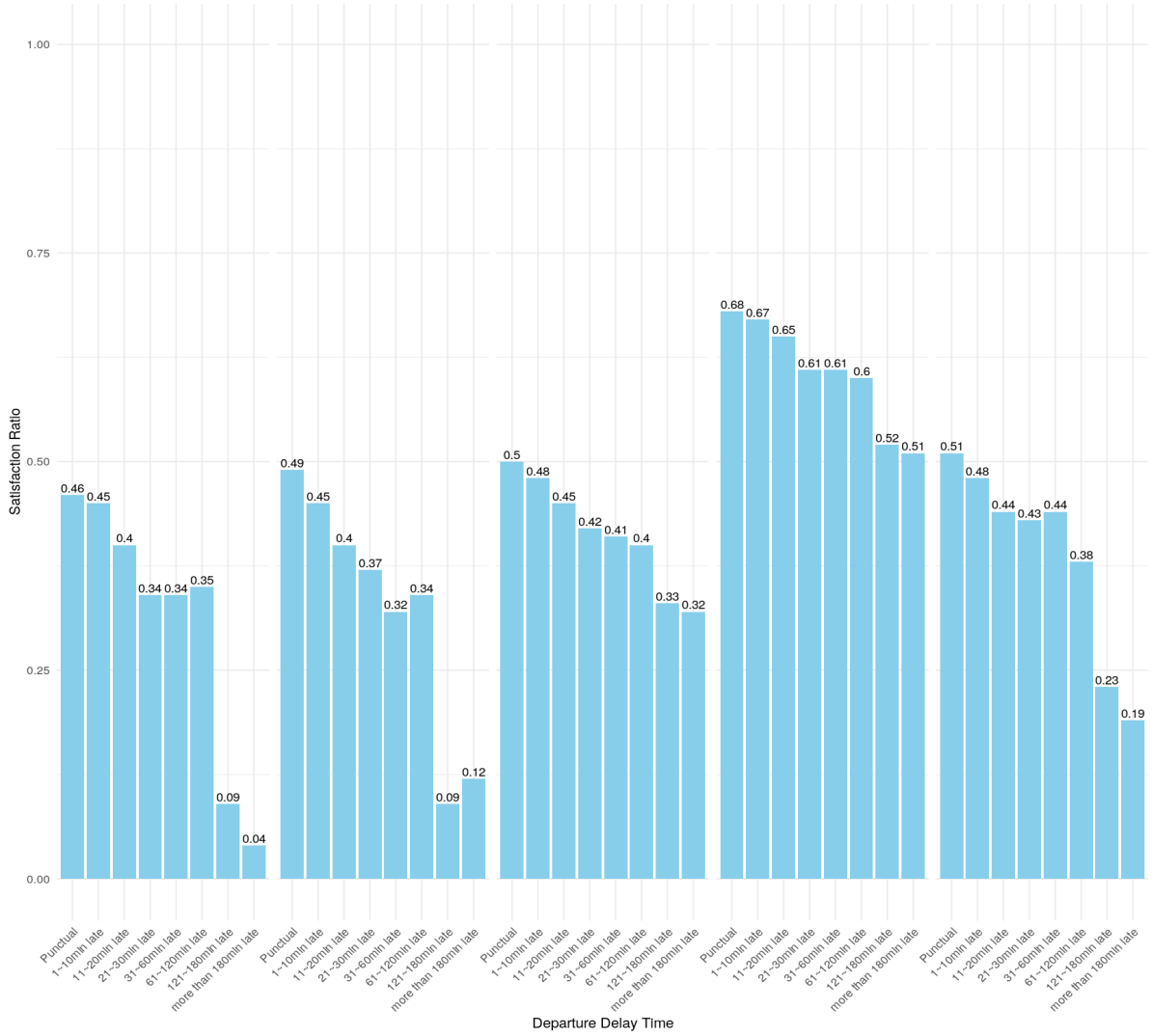

Figure 22: The satisfaction ratio across departure delay time by the predefined age groupings.

We tried to find out whether younger age groups are more likely to report ease in online booking than older age groups. However, this hypothesis has a confounding variable, airline. Whether or not it is easy to book online depends on which airline passengers take. Since we do not have the data of which certain flights passengers took, we cannot find this out with the given data. Instead, we tried to find out which age groups tend to satisfy more when they report ease in online booking.

We can see that all age groups tend to appreciate the flight experience when they report higher ease in online booking. Compared to younger age groups, older groups tend to feel satisfied when online booking is easier. Among them, “Middle Age Adult” group showed a great increase in satisfaction if they reported a high rate of ease in Online Booking, represented below in Figure 23.

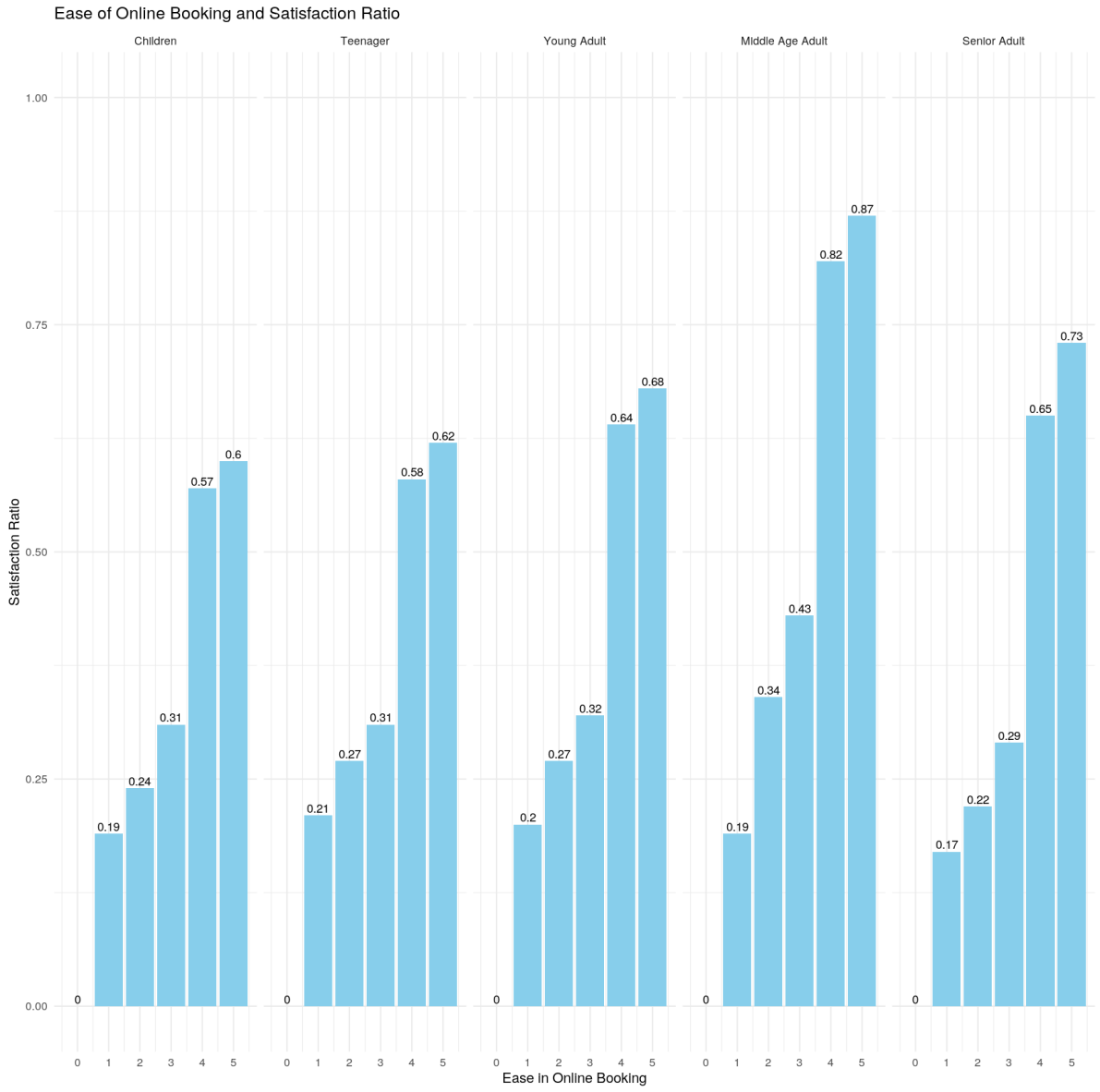

Figure 23: Calculated satisfaction ratio reported for ease in online booking as grouped by predefined age groupings.

In Figure 24, we decided to cluster between age and departure delay to see if there were any similarities. Four clusters were used since the WSS elbowed off at this value (Figure 25). Cluster 4 with Business Class had the most satisfied passengers (2154). Cluster 4 with Economy had a higher dissatisfaction reporting (2562). This means that business class is generally still satisfied no matter their age and departure delay. We calculated these numbers off to the side.

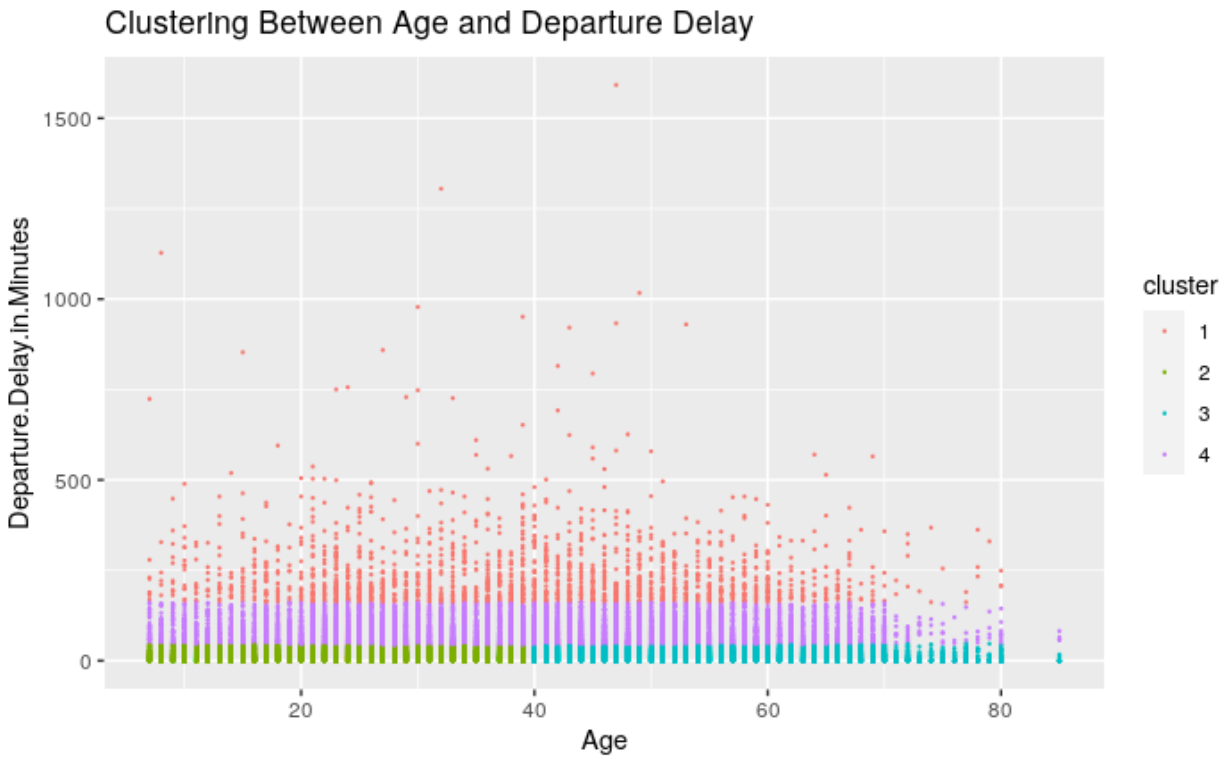

Figure 24: Clusters resulting from age and departure delay in minutes.

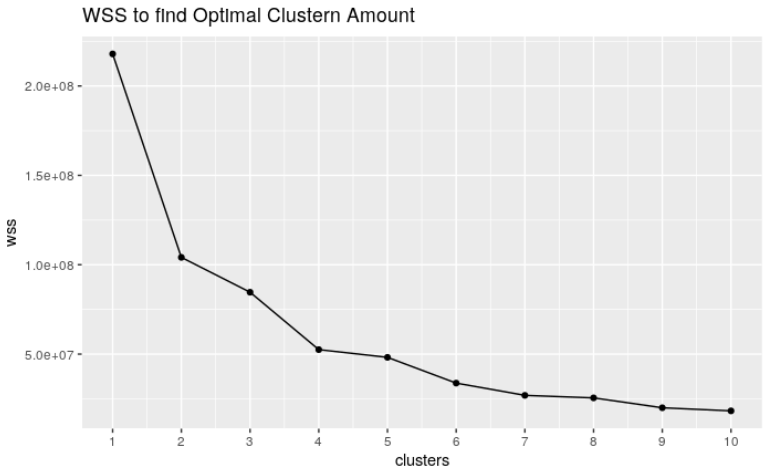

Figure 25: WSS across 10 different numbers of clusters.

Below, Figure 26 shows that business class is the most even between dissatisfied and satisfied with flight distance. However, the longer the flight is, the more dissatisfied economy and economy plus passengers become.

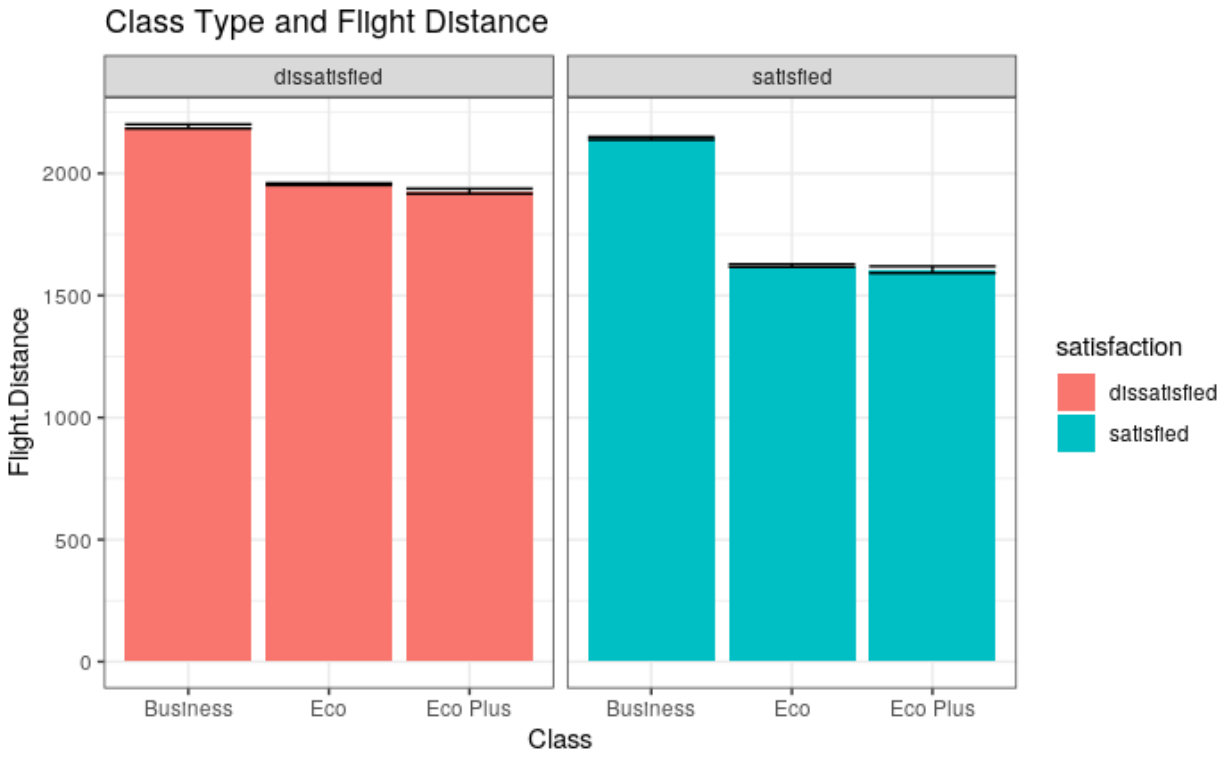

Figure 26: Overall satisfaction reported between Business, Economy Plus, and Economy as related to overall flight distance.

More business travel passengers have higher ratings for legroom than do personal travel, as shown in Figure 27 below. This could be because they are in business class since their company might have paid for them to travel better.

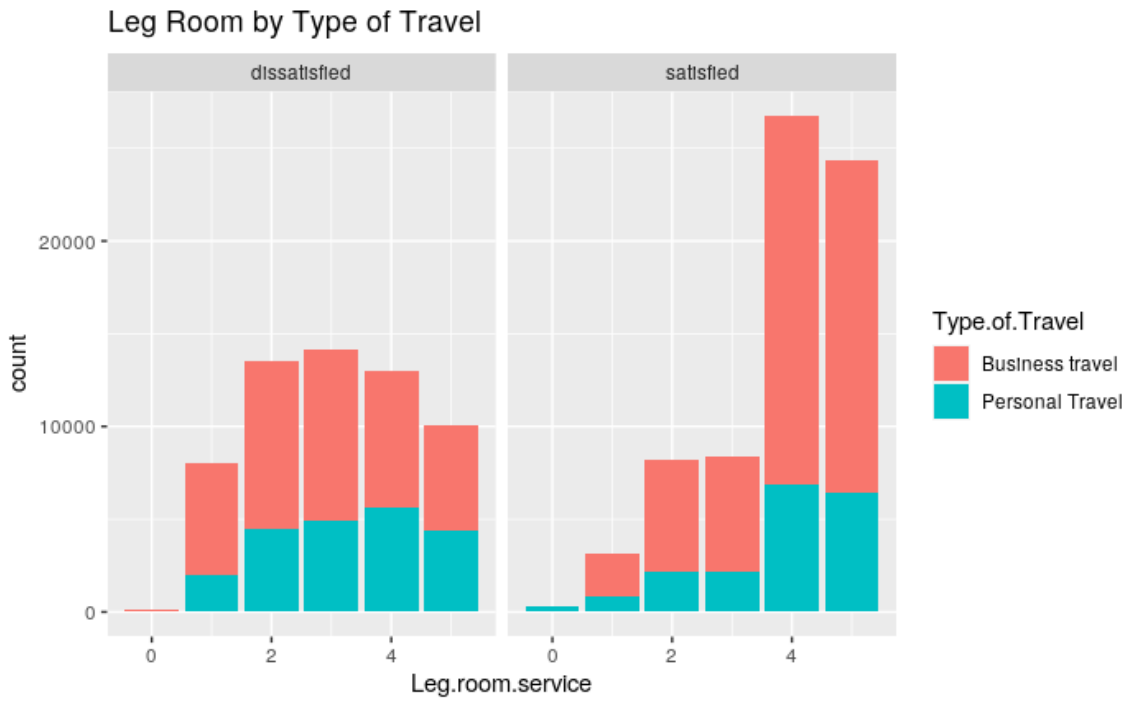

Figure 27: Overall satisfaction based on the leg room service satisfaction scale from 0-5 and type of travel being either business or personal.

Modeling

From the exploratory analysis, we have determined that the most important features are inflight entertainment, ease of online booking, online support, on board service, leg room service, seat comfort, and online boarding. In addition, we also used gender, customer type and class since they had the highest correlation to satisfaction. The label is the column satisfaction. We converted the non-numeric variables to numbers based: Female 1, Male 0, Business Travel 1, Personal Travel 0, Business Class 2, Eco Plus 1, Eco 0. We then use the scikit learn libraries to fit the model. We tuned our model to fit these features and came up with these combinations as the most important variables.

Training Features Shape: (97410, 10)
Training Labels Shape: (97410,)
Testing Features Shape: (32470, 10)
Testing Labels Shape: (32470,)

Accuracy of Logistic Regression : 0.8194333230674469
Accuracy of Decision Tree : 0.9227286726208808
Accuracy of Random Forest Classify : 0.927964274715121

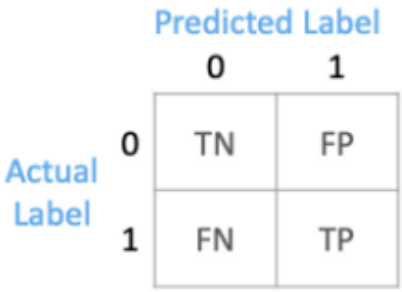

Figure 28: Label graph.

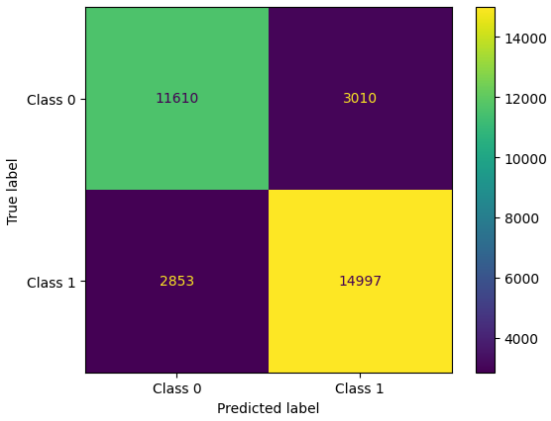

Figure 29: Logistic Regression Confusion Matrix

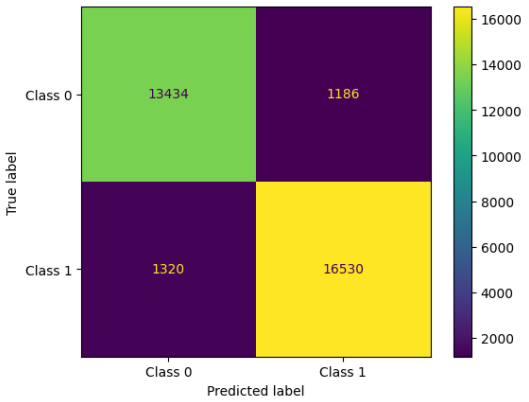

Figure 30: Decision Tree Confusion Matrix

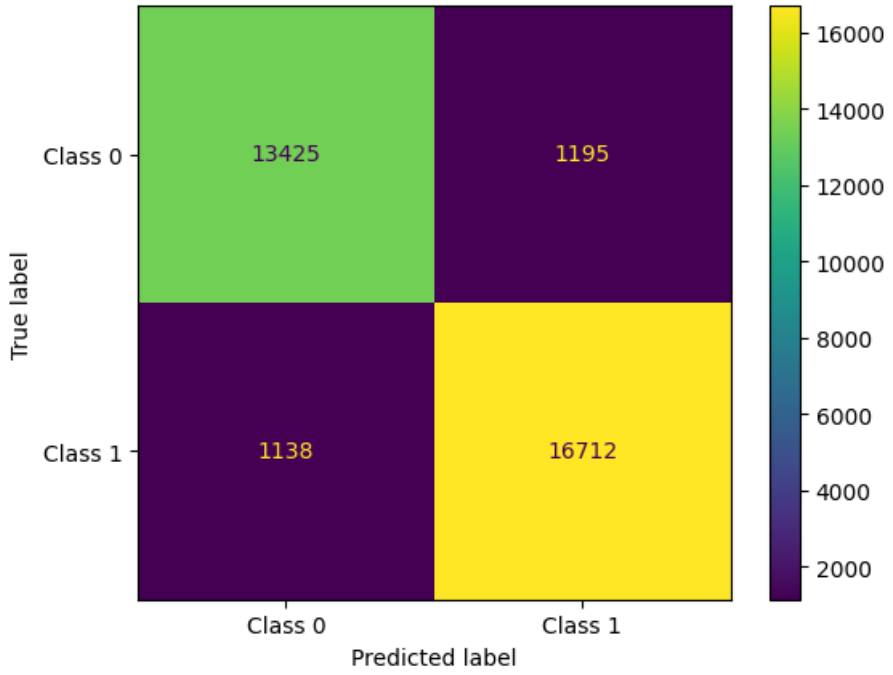

Figure 31: Random Forest Confusion Matrix

The random forest model had the highest accuracy which was pretty close to the decision tree (Figure 30). The precision of the random forest was calculated as follows from Figure 31; 16,712 / (16,712 + 1,138) = .94. Precision is the true positives divided by the true positives plus the false positives. The closer to 1, the better.

Thus, through our exploratory analysis, we have touched all the variables and have been able to understand the passengers more. In our discussion, we will look at explaining some of our hypotheses.

Discussion

From our analysis, we determined the variables that correlate the highest with passenger satisfaction. We can group these variables into two categories, pre-flight and during flight activities. These are the groupings of all the rating variables:

Pre-flight: Online support, Ease of online booking, online boarding, check in
During flight: Departure arrival time convenience, food and drink, gate location, inflight wife service and entertainment, leg room, baggage handling, cleanliness, and departure and arrival delay.

The pre-flight activities that are the most important are ease of online booking, online support, and online boarding. The during flight activities that are most important are inflight entertainment, leg room service, and seat comfort. These variables tend to have technology based items in common (which is shown in the correlation matrix of all rating variables). This makes sense since people are finding technology more convenient, especially as it can bridge the gap between airline class and mitigate boredom. The other variables that are important have physical attributes in common. The airline organization should then focus on increasing technology based services. This could, in turn, positively influence leg room service and seat comfort if the inflight entertainment is really strong since more riders could be preoccupied with technology to notice their physical discomfort. Their correlation is also relatively high at 0.43 for seat comfort and inflight entertainment and 0.16 for leg room service and inflight entertainment.

In addition, our analysis shows that females, riders who travel for business, riders who were seated in the business class, and loyal customers have high correlation with satisfaction. Females tend to care the most about inflight entertainment, on board service, and leg room amidst the result. Males tend to care the most about inflight entertainment and online services. Both females and males who are aged 40 - 60 years old are more satisfied. This could be because they are traveling for business related reasons and most of the time, their trip is paid for. They would naturally have a higher tolerance for their airplane ride. Their respective company is also likely to put them up in business class as well which can increase their experience immensely. Younger ages, around 15 - 40 years old, have a higher count for dissatisfaction. This could be because they are expecting a lot more services and have a lower threshold for tolerance. They could expect the technology services to be superior, and could be more easily let down if something goes wrong. Loyal customers make sense as they are inclined to the airline organization. Business class also makes sense as they are given more amenities. Females on the other hand make sense as females could be generally nicer and accepting of services compared to men.

Another issue is departure delay. Since there were NAs in the arrival delay and departure delay and arrival delay is very strongly correlated, any analysis we make can be translated to arrival. No matter the age, as long as the rider is in business class, they will have a more satisfying ride. Conversely, if they are in economy class, they will tend to have a more dissatisfied ride. The airline organization should implement services that help the economy class, such as technology services. The company can also take advantage of correlated services such as food and drink and seat comfort. If the company cannot directly help the seat and comfort variable, they can focus more on providing better food and drinks. Since they are correlated high, this will help improve satisfaction.

In terms of our machine learning model, the random forest classifier had the highest accuracy. We were able to predict a passenger was satisfied with their ride 92.8% of the time. It had the lowest false ratings (only 1138 false negatives and 1195 false positives) compared to the logistic regression (3010 false negatives and 2853 false positives) and decision tree (1186 false negatives and 1320 false positives). We were able to use the features that we determined were the most important in our analysis (inflight entertainment, ease of online booking, online support, on board service, online support, leg room service, and seat comfort) in our model. We also utilized gender, customer type, and class. We decided not to use SVM since we did not believe that the data was linearly separable.

Some limitations to our findings is that we don't know the full extent and meaning of all the variables. For example, for inflight entertainment, we do not know from the data given if this means that TV/movies are provided, video games, or if anything else is included. Another limitation is that there are some pieces of data missing that could be helpful to the analysis such as price tickets and destinations. Where a passenger is flying to could make a difference on how happy they are already when they fly. Additionally, the ticket price will create a difference in tolerance for flying. With different ticket prices, it affects the way that individuals go on a plane, and we did not have that variable in our report specified explicitly. As ticket price can affect many things, an analysis on that would present a greater accuracy on our study. One more confounding variable to think about would be the date, as more prominent dates and holidays such as Christmas and Thanksgiving which encourage more flyers regardless; with the date in mind, the correlations to satisfaction overall could be re-evaluated. Another piece of information that we do not know is if the riders are grouped together in any way. Are they flying alone, with friends, or family? Having access to this data could improve the analysis by providing more context to the available data. There are a lot of external factors that could have contributed to a riders experience that we cannot fully quantify. Another limitation could be whether these recommendations can be generalized to the whole population of airplane riders. More information is needed on how the sample was collected and when it was taken. Interviewer falsification is a big issue and can skew the results. Passengers can rate their rides higher if there is an incentive, if they were sampled right after their ride, or if there was a long delay before they were asked all these questions allowing them time to forget details in their flight experience. One more variable to also consider is a frequent flyers distinction per rider, especially for stakeholders such as the flights themselves whom this analysis will benefit. As riders fly more often, their satisfaction and specific elements of what they are satisfied with could be changed. With all these variables, the analysis could be improved further on future iterations.

Conclusion

In conclusion, through multiple data analysis techniques and working with all the variables we have in our dataset, our group has determined the specific variables that are the most important for airline passengers. For the rating variables, they are Inflight entertainment, Ease of Online booking, On-board service, Online support, Leg room service, Seat comfort, and Online boarding. In terms of the demographics that tend to be more satisfied are Gender (female), Customer Type (Loyal), and Class (Business). The airline organization should make improvements on departure delay and gate location and continue to focus on technology based services and physical seating. With in-depth analysis of many different demographic groups, we were able to find correlation aspects between each demographic. Using the machine learning model, we were also able to accurately predict a satisfied passenger with 92.8% accuracy.

The next steps are for this airline organization to specifically work on these aspects of the rating variables we mentioned. By increasing their short-term investments into the aspects of the flight experience that are important to passengers, they will be able to increase passenger satisfaction and reap financial rewards long term. If possible, this airline should also invest time and effort to collect more data that could help understand the customers more, such as ticket price or destination of travel. Other important aspects that the company can focus on is making sure that the economy class passengers are better accommodated for in the aspects that they find important and ultimately creating a path to help disloyal customers convert to

loyal customers. These are some recommendations that the company can consider that can impact the current and future success of their organization.

Acknowledgement

Name	**Percent Contribution**
George Sayegh	**100%**
George Gu	**100%**
Sai Ponnapalli	**100%**
Thomas Nguyen	**100%**
Jooho Kim	**100%**
Obaid Abbasi	**100%**
Adriana van Zanten	**100%**

Bibliography

1. “Airlines Customer Satisfaction | Kaggle.” *Kaggle: Your Machine Learning and Data Science Community*, https://www.kaggle.com/datasets/sjleshrac/airlines-customer-satisfaction. Accessed 5 Dec. 2022.

2. “Air traffic by the numbers.” *Federal Aviation Administration*, https://www.faa.gov/air_traffic/by_the_numbers. Accessed 5 Dec. 2022.