جامعة أم القرى
UMM AL-QURA UNIVERSITY

Data Analysis Project

Predicting Airline Passenger Satisfaction with Classification Algorithms

SUBMITTED BY:

NAME	ID
Joud Ahmad Al-Huthaly	444002970
Layan Adel Babkur	444002368
Reham Faisal Al-Subhi	444003014
Manar Ali Al-Subhi	444003523
Jana Abulraouf Al-Lihyani	444001382

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS
UMM AL-QURA UNIVERSITY

Table of content

- **Google collab code link:**

https://drive.google.com/drive/folders/1pqllYXwRek7dg8JGPcg_tPej6GFSRZ_b?usp=sharing

1. Introduction

Flight-based customer evaluation is a crucial element in the air transport sector, as customer satisfaction and a comfortable customer experience are paramount. Today's airlines are experiencing intense competition to attract and retain customers, so understanding customers' wants and needs and meeting them effectively is essential to the company's success and business continuity.

The importance of knowing customer satisfaction

Customer satisfaction is crucial for any airline. Customers are looking for a comfortable and smooth experience during their flights, and are greatly influenced by factors such as service, comfort, cleanliness and entertainment. If customers are satisfied with those factors, they are more likely to come back and use the company's services again, and also to recommend the company to others. On the other hand, if the customer experience is poor or unsatisfactory, it may lead to loss of customers, poor reputation and a negative impact on the business.

Based on the provided data, this is our goals

1. Analysis of satisfaction and dissatisfaction factors: Analyze the different ratings for the service factors and identify the elements contributing to overall customer satisfaction and dissatisfaction.

2. Improving service quality: Direct efforts towards enhancing various service factors such as in-flight Wi-Fi quality, ease of online booking, seat comfort, in-flight entertainment quality, and others, with the aim of increasing overall customer satisfaction.

3. Boosting loyalty and strategic direction: Utilize customer classification into "loyal" and "disloyal" categories to determine loyalty enhancement strategies and tactics to attract potential customers.

2. Exploratory Data Analysis

Attribute about dataset

- **Gender**: Contains two values, "female" and "male", representing the gender of the passenger.
- **Customer** Type: Contains two values, "loyal" and "disloyal", indicating whether the customer is loyal or disloyal to the company.
- **Age**: Contains data about the age of the passenger
- **Type of Travel**: Contains two values, "personal travel" and "business travel", representing the type of travel, either personal or business.
- **Class**: Contains three values, "eco", "eco plus", and "business", representing the travel class of the passenger.
- **Flight Distance**: Represents a rating from 0 to 5 indicating the distance of the flight.
- Inflight WIFI service: Represents a rating from 0 to 5 indicating the quality of the in-flight Wi-Fi service.
- **Departure Arrival time convenient**: Represents a rating from 0 to 5 indicating the convenience of departure and arrival times.
- **Ease of Online booking**: Represents a rating from 0 to 5 indicating the ease of the online booking process.
- **Gate location**: Represents a rating from 0 to 5 indicating the location of the gate at the airport.
- Food and drink: Represents a rating from 0 to 5 indicating the quality of food and drink on the flight.
- **Online boarding**: Represents a rating from 0 to 5 indicating the ease of online boarding procedures.
- **Seat comfort**: Represents a rating from 0 to 5 indicating the comfort of the seats on the flight.
- Inflight entertainment: Represents a rating from 0 to 5 indicating the quality of in-flight entertainment.
- **On. Board service**: Represents a rating from 0 to 5 indicating the quality of the onboard service by the crew.
- **Leg room service**: Represents a rating from 0 to 5 indicating the quality of the legroom service on the flight.
- **Baggage handling**: Represents a rating from 0 to 5 indicating the quality of the baggage handling service.
- **Check in service**: Represents a rating from 0 to 5 indicating the quality of the check-in service at the airport.
- **Inflight service**: Represents a rating from 0 to 5 indicating the quality of the service during the flight.
- **Cleanliness**: Represents a rating from 0 to 5 indicating the cleanliness of the aircraft and facilities.
- **Departure Delay in Minutes**: Contains the number of minutes by which the departure was delayed.
- **Arrival Delay in Minutes**: Contains the number of minutes by which the arrival was delayed.
- **satisfaction**: Contains three values, "Satisfied", "Neutral or Dissatisfied", reflecting the overall satisfaction level of the passenger's experience.

A data.frame

	SR	id	Gender	Customer_Type	Age	Type_of_Travel	Class	Flight_Distance	Inflight_wifi_service	Departure.Arrival_time_convenient	…	Inflight_entertainment
	<int>	<int>	<chr>	<chr>	<int>	<chr>	<chr>	<int>	<int>	<int>	…	<int>
1	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	…	5
2	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	…	1
3	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	…	5
4	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	…	2
5	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	…	3
6	5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180	3	4	…	1
7	6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2	4	…	2
8	7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4	3	…	5
9	8	79485	Female	Loyal Customer	41	Business travel	Business	853	1	2	…	1
10	9	65725	Male	disloyal Customer	20	Business travel	Eco	1061	3	3	…	2

10 × 25

On.board_service	Leg_room_service	Baggage_handling	Checkin_service	Inflight_service	Cleanliness	Departure_Delay_in_Minutes	Arrival_Delay_in_Minutes	satisfaction
<int>	<int>	<int>	<int>	<int>	<int>	<int>	<int>	<chr>
4	3	4	4	5	5	25	18	neutral or dissatisfied
1	5	3	1	4	1	1	6	neutral or dissatisfied
4	3	4	4	4	5	0	0	satisfied
2	5	3	1	4	2	11	9	neutral or dissatisfied
3	4	4	3	3	3	0	0	satisfied
3	4	4	4	4	1	0	0	neutral or dissatisfied
3	3	4	3	5	2	9	23	neutral or dissatisfied
5	5	5	4	5	4	4	0	satisfied
1	2	1	4	1	2	0	0	neutral or dissatisfied
2	3	4	4	3	2	0	0	neutral or dissatisfied

Dataset

```
[ ]  dim(sat)
```

```
103904 · 25
```

This code displays the number of columns and rows in the dataset.

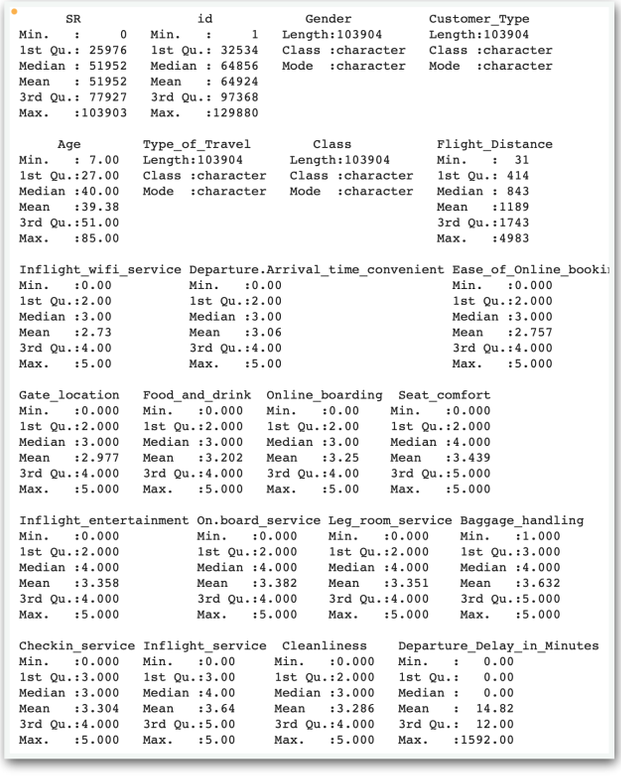
- Number of columns: 25
- Number of rows: 103,904

```
[ ]  #Show if there are duplicates
     sum(duplicated(sat))
     #Show if there are na values
     sum(is.na(sat)) #in Arrival_Delay_in_Minutes there is 310
```

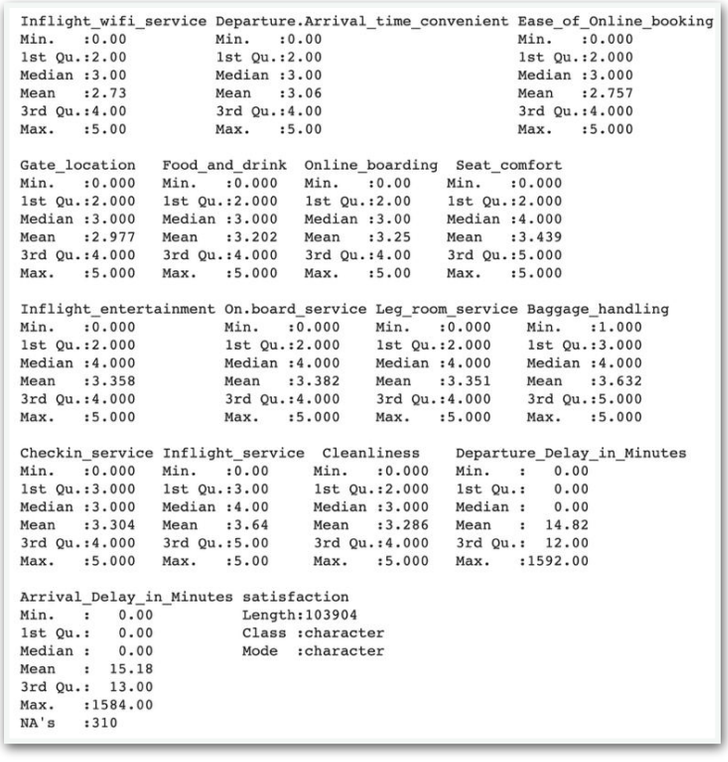
```
0
310
```

data quality was improved by checking and cleaning duplicate values and null values using the (is null) function,It appears from the code that there are 310 empty/null values in the column"Arrival_Delay_in_Minutes".

We used the "**summary**" function to explain the values such as the mean, median, minimum, and maximum values. This function provides an overview of the data distribution and basic statistics for the available variables.



```
       SR              id            Gender          Customer_Type
 Min.   :     0   Min.   :     1   Length:103904      Length:103904
 1st Qu.: 25976   1st Qu.: 32534   Class :character   Class :character
 Median : 51952   Median : 64856   Mode  :character   Mode  :character
 Mean   : 51952   Mean   : 64924
 3rd Qu.: 77927   3rd Qu.: 97368
 Max.   :103903   Max.   :129880

      Age        Type_of_Travel        Class           Flight_Distance
 Min.   : 7.00   Length:103904      Length:103904      Min.   :  31
 1st Qu.:27.00   Class :character   Class :character   1st Qu.: 414
 Median :40.00   Mode  :character   Mode  :character   Median : 843
 Mean   :39.38                                         Mean   :1189
 3rd Qu.:51.00                                         3rd Qu.:1743
 Max.   :85.00                                         Max.   :4983

 Inflight_wifi_service Departure.Arrival_time_convenient Ease_of_Online_booki
 Min.   :0.00          Min.   :0.00                       Min.   :0.000
 1st Qu.:2.00          1st Qu.:2.00                       1st Qu.:2.000
 Median :3.00          Median :3.00                       Median :3.000
 Mean   :2.73          Mean   :3.06                       Mean   :2.757
 3rd Qu.:4.00          3rd Qu.:4.00                       3rd Qu.:4.000
 Max.   :5.00          Max.   :5.00                       Max.   :5.000

 Gate_location   Food_and_drink  Online_boarding  Seat_comfort
 Min.   :0.000   Min.   :0.000   Min.   :0.00    Min.   :0.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.000
 Median :3.000   Median :3.000   Median :3.00    Median :4.000
 Mean   :2.977   Mean   :3.202   Mean   :3.25    Mean   :3.439
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.000

 Inflight_entertainment On.board_service Leg_room_service Baggage_handling
 Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000
 1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000
 Median :4.000          Median :4.000    Median :4.000    Median :4.000
 Mean   :3.358          Mean   :3.382    Mean   :3.351    Mean   :3.632
 3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000
 Max.   :5.000          Max.   :5.000    Max.   :5.00     Max.   :5.000

 Checkin_service Inflight_service  Cleanliness    Departure_Delay_in_Minutes
 Min.   :0.000   Min.   :0.00      Min.   :0.000   Min.   :   0.00
 1st Qu.:3.000   1st Qu.:3.00      1st Qu.:2.000   1st Qu.:   0.00
 Median :3.000   Median :4.00      Median :3.000   Median :   0.00
 Mean   :3.304   Mean   :3.64      Mean   :3.286   Mean   :  14.82
 3rd Qu.:4.000   3rd Qu.:5.00      3rd Qu.:4.000   3rd Qu.:  12.00
 Max.   :5.000   Max.   :5.00      Max.   :5.000   Max.   :1592.00
```



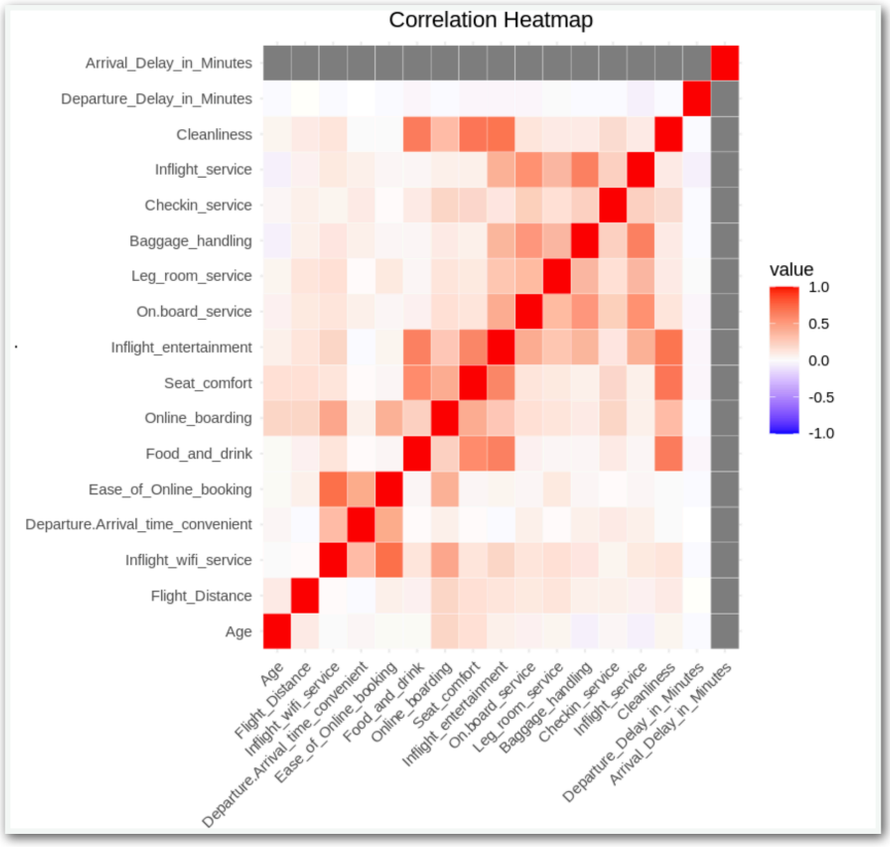
```
 Inflight_wifi_service Departure.Arrival_time_convenient Ease_of_Online_booking
 Min.   :0.00          Min.   :0.00                       Min.   :0.000
 1st Qu.:2.00          1st Qu.:2.00                       1st Qu.:2.000
 Median :3.00          Median :3.00                       Median :3.000
 Mean   :2.73          Mean   :3.06                       Mean   :2.757
 3rd Qu.:4.00          3rd Qu.:4.00                       3rd Qu.:4.000
 Max.   :5.00          Max.   :5.00                       Max.   :5.000

 Gate_location   Food_and_drink  Online_boarding  Seat_comfort
 Min.   :0.000   Min.   :0.000   Min.   :0.00    Min.   :0.000
 1st Qu.:2.000   1st Qu.:2.000   1st Qu.:2.00    1st Qu.:2.000
 Median :3.000   Median :3.000   Median :3.00    Median :4.000
 Mean   :2.977   Mean   :3.202   Mean   :3.25    Mean   :3.439
 3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:4.00    3rd Qu.:5.000
 Max.   :5.000   Max.   :5.000   Max.   :5.00    Max.   :5.000

 Inflight_entertainment On.board_service Leg_room_service Baggage_handling
 Min.   :0.000          Min.   :0.000    Min.   :0.000    Min.   :1.000
 1st Qu.:2.000          1st Qu.:2.000    1st Qu.:2.000    1st Qu.:3.000
 Median :4.000          Median :4.000    Median :4.000    Median :4.000
 Mean   :3.358          Mean   :3.382    Mean   :3.351    Mean   :3.632
 3rd Qu.:4.000          3rd Qu.:4.000    3rd Qu.:4.000    3rd Qu.:5.000
 Max.   :5.000          Max.   :5.000    Max.   :5.000    Max.   :5.000

 Checkin_service Inflight_service  Cleanliness    Departure_Delay_in_Minutes
 Min.   :0.000   Min.   :0.00      Min.   :0.000   Min.   :   0.00
 1st Qu.:3.000   1st Qu.:3.00      1st Qu.:2.000   1st Qu.:   0.00
 Median :3.000   Median :4.00      Median :3.000   Median :   0.00
 Mean   :3.304   Mean   :3.64      Mean   :3.286   Mean   :  14.82
 3rd Qu.:4.000   3rd Qu.:5.00      3rd Qu.:4.000   3rd Qu.:  12.00
 Max.   :5.000   Max.   :5.00      Max.   :5.000   Max.   :1592.00

 Arrival_Delay_in_Minutes satisfaction
 Min.   :   0.00          Length:103904
 1st Qu.:   0.00          Class :character
 Median :   0.00          Mode  :character
 Mean   :  15.18
 3rd Qu.:  13.00
 Max.   :1584.00
 NA's   :310
```

Here convert some into numerical values. For ease of modeling or analysis.

Class <chr>	Customer_Type <chr>
Eco Plus	Loyal Customer
Business	disloyal Customer
Business	Loyal Customer
Business	Loyal Customer
Business	Loyal Customer
Eco	Loyal Customer

Customer_Type1 <dbl>	Class1 <dbl>
0	1
1	2
0	2
0	2
0	2
0	3

satisfaction <chr>	satisfaction.1 <dbl>
neutral or dissatisfied	0
neutral or dissatisfied	0
satisfied	1
neutral or dissatisfied	0
satisfied	1
neutral or dissatisfied	0



Correlation Heatmap

```
Departure_Delay_in_Minutes                1.000000000
Arrival_Delay_in_Minutes                           NA
                                Arrival_Delay_in_Minutes
```

In the heatmap, we noticed that the color representing "Arrival delay in Minutes" is gray. This indicates a weak correlation between the data because there are a significant number of missing values.

3. Linear and Nonlinear Regression

The main goal of regression analysis is to predict the value of the dependent variable based on the values of the independent variables, we built two models.

First The null variables in "arrival time in minutes" have been removed, and a new name has been assigned to the data, which is "Sate" instead of "Sat".

```
Call:
lm(formula = Arrival_Delay_in_Minutes ~ Departure_Delay_in_Minutes,
    data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-47.627  -2.114  -0.695  -0.455 236.331

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                0.6952628  0.0374460   18.57   <2e-16 ***
Departure_Delay_in_Minutes 0.9799573  0.0009214 1063.53   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.05 on 82873 degrees of freedom
Multiple R-squared:  0.9317,    Adjusted R-squared:  0.9317
F-statistic: 1.131e+06 on 1 and 82873 DF,  p-value: < 2.2e-16
10.2447414581709
0.93215337092099
```

MODEL.1

- RMSE : 10.2447414581709
- R-squared : 0.93215337092099

We calculated RMSE and R-squared. The r-squared value was close to 1, which means that the model is excellent.

```
Call:
lm(formula = Inflight_entertainment ~ Cleanliness, data = train_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.8590 -0.4519  0.1410  0.4375  3.2517

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.044742   0.009004   116.0   <2e-16 ***
Cleanliness 0.703558   0.002546   276.3   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.962 on 82873 degrees of freedom
Multiple R-squared:  0.4795,    Adjusted R-squared:  0.4795
F-statistic: 7.636e+04 on 1 and 82873 DF,  p-value: < 2.2e-16
1.04512317267583
0.47860749253182
```

MODEL.2

- RMSE : 1.04512317267583
- R-squared : 0.47860749253182

We calculated RMSE and R-squared. The r-squared value was average, which means there could be an overfitting problem.

The first model was developed by using **'K-fold'**. The results were obtained. fold 5 is the best because of the lower RMSE, Rsquared is closer to one, and the MAE is lower.

RMSE	Rsquared	MAE	Resample
<dbl>	<dbl>	<dbl>	<chr>
10.279420	0.9299078	5.290517	Fold1
9.886820	0.9342961	5.223918	Fold2
10.317832	0.9277466	5.286919	Fold3
10.042280	0.9307162	5.302763	Fold4
9.866808	0.9378106	5.236008	Fold5

Model. 1

The second model was developed by using **'K-fold'**. The results were obtained. fold 4 is the best because of the lower RMSE, Rsquared is closer to one, and the MAE is lower.

RMSE	Rsquared	MAE	Resample
<dbl>	<dbl>	<dbl>	<chr>
0.9628009	0.4798040	0.6895171	Fold1
0.9642706	0.4763215	0.6908447	Fold2
0.9664958	0.4737523	0.6893402	Fold3
0.9520175	0.4904740	0.6853942	Fold4
0.9677151	0.4724073	0.6914537	Fold5

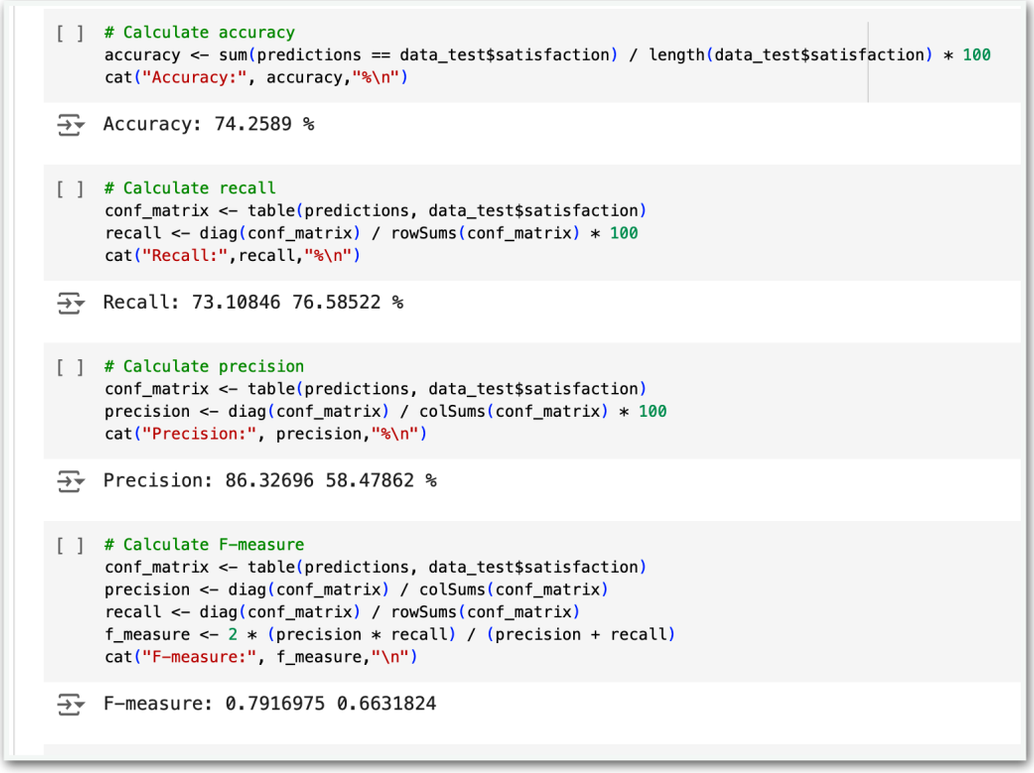
Model. 2

4. Classification.

The goal of this classification is to build a classification model used to predict passenger satisfaction on airlines. This is done by dividing the training data into a training set and a test set, then training a classification model using an iterative partition analysis algorithm called "Recursive Partitioning". After training, the model is used to predict the level of passenger satisfaction in the test set, and the model's performance is then evaluated by comparing the predictions with the actual values.

We designed two models to predict passenger satisfaction with airlines.

The first model was based on several services provided: food and drink, cleanliness, Wi-Fi service on board, and ease of booking online.



```
[ ]  # Calculate accuracy
     accuracy <- sum(predictions == data_test$satisfaction) / length(data_test$satisfaction) * 100
     cat("Accuracy:", accuracy,"%\n")

     Accuracy: 74.2589 %

[ ]  # Calculate recall
     conf_matrix <- table(predictions, data_test$satisfaction)
     recall <- diag(conf_matrix) / rowSums(conf_matrix) * 100
     cat("Recall:",recall,"%\n")

     Recall: 73.10846 76.58522 %

[ ]  # Calculate precision
     conf_matrix <- table(predictions, data_test$satisfaction)
     precision <- diag(conf_matrix) / colSums(conf_matrix) * 100
     cat("Precision:", precision,"%\n")

     Precision: 86.32696 58.47862 %

[ ]  # Calculate F-measure
     conf_matrix <- table(predictions, data_test$satisfaction)
     precision <- diag(conf_matrix) / colSums(conf_matrix)
     recall <- diag(conf_matrix) / rowSums(conf_matrix)
     f_measure <- 2 * (precision * recall) / (precision + recall)
     cat("F-measure:", f_measure,"\n")

     F-measure: 0.7916975 0.6631824
```

Model.1

- Accuracy: 74.2589 %
- Recall: 73.10846 76.58522 %
- Precision: 86.32696 58.47862 %
- F-measure: 0.7916975 0.6631824

The second model was based on several other additional services: ease of booking online, food and drinks, and boarding. Online, seat comfort, in-flight entertainment, in-flight service, leg room service, baggage handling, check-in service, in-flight service, cleanliness

```
[ ]  # Calculate accuracy
     accuracy <- sum(predictions == data_test$satisfaction) / length(data_test$satisfaction) * 100
     cat("Accuracy:", accuracy,"%\n")

     Accuracy: 81.88643 %

[ ]  # Calculate Recall
     conf_matrix <- table(predictions, data_test$satisfaction)
     recall <- diag(conf_matrix) / rowSums(conf_matrix) * 100
     cat("Recall:",recall,"%\n")

     Recall: 81.47246 82.54067 %

[ ]  # Calculate precision
     conf_matrix <- table(predictions, data_test$satisfaction)
     precision <- diag(conf_matrix) / colSums(conf_matrix) * 100
     cat("Precision:", precision,"%\n")

     Precision: 88.05945 73.81455 %

[ ]  # Calculate F-measure
     conf_matrix <- table(predictions, data_test$satisfaction)
     precision <- diag(conf_matrix) / colSums(conf_matrix)
     recall <- diag(conf_matrix) / rowSums(conf_matrix)
     f_measure <- 2 * (precision * recall) / (precision + recall)
     cat("F-measure:", f_measure,"\n")

     F-measure: 0.8463799 0.7793411
```

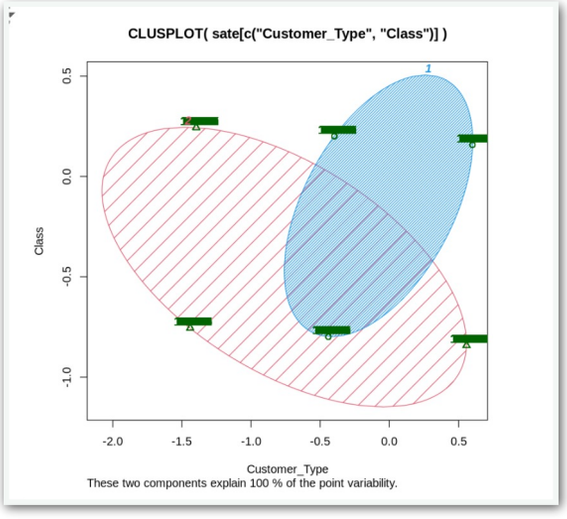
Model. 2

- Accuracy: 81.88643 %
- Recall: 81.47246 82.54067 %
- Precision: 88.05945 73.81455 %
- F-measure: 0.8463799 0.7793411

After comparing the two models by calculating the value of accuracy ,recall, precision ,f-measure it was noted that the second model is better than the first model

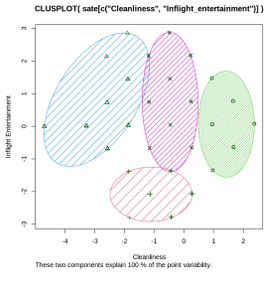
5. Clustering

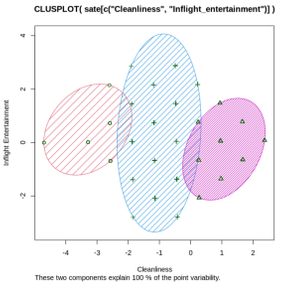
Cluster is the of dividing unlabeled data or data points into different groups so that similar data points fall into the same group compared to different ones.
Here we used K-means algorithm
In this method, data points are assigned to initial groups such that the sum of the squared distances between the data points and the centroid in the model
We have chosen two columns, Class Column and Customer Column, to show that a repeat customer is booking which class and a new customer is interested in which class.
This aims to improve the quality of other categories that are not requested by both parties
At first we faced a problem, which is that the data is textual and must be numeric, and we solved the problem by converting the textual data into numeric data (0 and 1).
The class column has only two classes: loyal Customer and disloyal Customer
The custom column has 3 categories: Business Class, Eco plus and Eco, which we have converted to 0, 1 and 2.

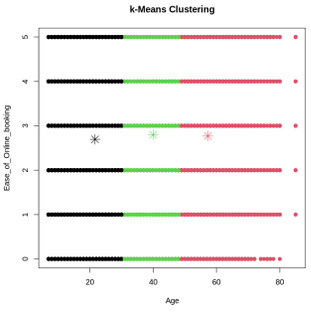


In general, similar points are placed together as shown in the figure. The graphic indicates that all types are close to each other, which leads us to the conclusion that all categories are desirable by both parties (loyal, traitor)

There is more than one model that has been tried, but it is difficult for us to explain it





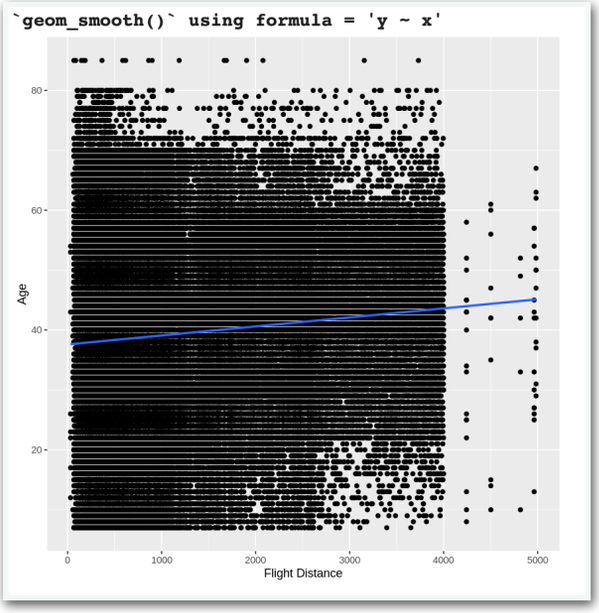


6. Anomaly Detection

Anomaly detection involves identifying patterns or points in data that deviate significantly from the norm. It helps detect errors, unusual events, or unexpected patterns in a dataset. In our analysis, we utilized **outlier detection** which aims to identify values or points in a dataset that deviate significantly from the general pattern or other patterns within the dataset. By using outlier detection methods, we aimed to identify customers who exhibit unusual flight distances relative to their age.

we focused on the relationship between flight distance and customer age to identify anomalies. We used statistical methods like standard deviation and mean to calculate the average age and identify customers outside a reasonable range of deviation.

This helps us improve our understanding of customers' needs and better meet them in the future.



Based on the analysis our conclusions are as follows

1. Age affects travel preferences:older customers travel shorter distances.with exceptions.

2. Normal flight distance range: The typical flight distance is around 4,000 miles, with exceptions.

3. Majority age group: The majority of customers fall within the 20-60 age range.

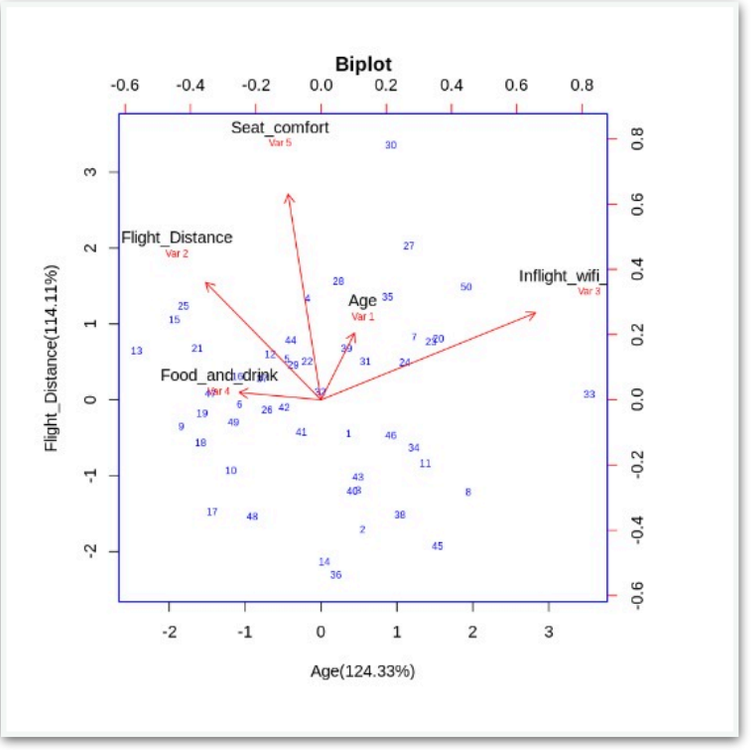
Recommendations for the airline

- Targeted offerings: Create tailored offers for younger customers preferring shorter flights.
- Special services for older customers: Provide assistance and accommodations for older customers who travel shorter distances.
- Targeted marketing campaigns: Design campaigns appealing to the 25-80 age group, emphasizing the benefits of traveling within the normal flight distance range.
- Customer segmentation: Segment customers based on age and travel patterns for personalized promotions.

These insights will help the airline better understand customer needs and improve services to enhance satisfaction and loyalty.

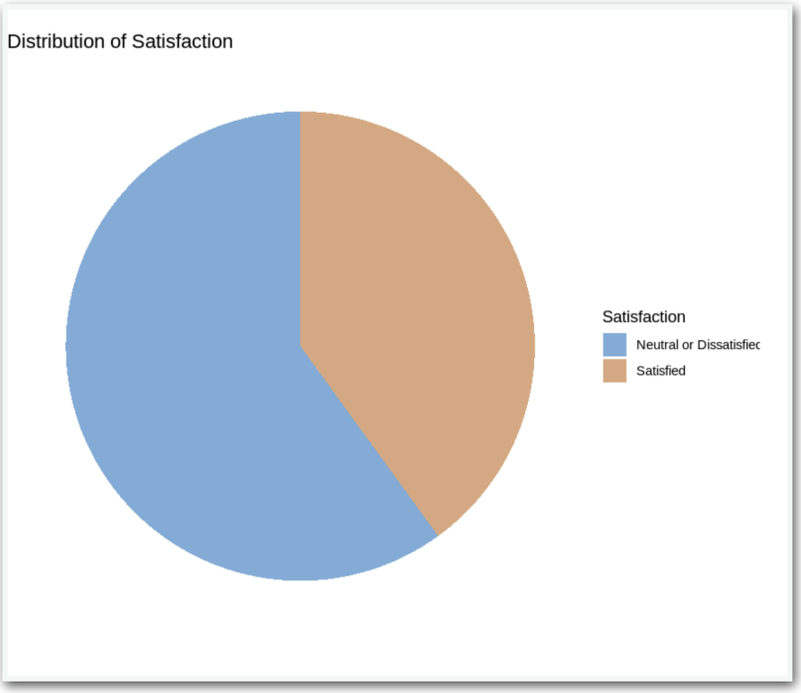
7. Principal Component Analysis (PCA)

Principal component analysis, or PCA, is a dimensionality reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.
(PCA) plot shows similarities between groups of samples in a data set. Each point on a PCA plot represents a correlation between an initial variable and the first and second principal components.
Points that are close to each other in the biplot represent observations with similar values.
a **classic loading biplot** enables you to visualize high-dimensional data by using a two-dimensional graph.



As we can see in this **biplot** there are positive related relationship between seat comfort, flight distance. food and drink, also between Inflight Wi-Fi service and Age and there is another relationship between Age, flight distance, seat comfort because the angle between them is smaller than 90° so it's a positive related
The smaller the angle the stronger that relationship is.
and negatively related relationship between Inflight Wi-Fi service and Flight distance, food and drink, seat comfort also between Age and food and drink because the angle between them is larger than 90°.
Coefficients close to -1 or 1 indicate that the variable strongly influences the component. Coefficients close to 0 indicate that the variable has a weak influence on the component. Food and drink have a weak influence on the component because its shows at 0 at the biplot, There is no Outliers

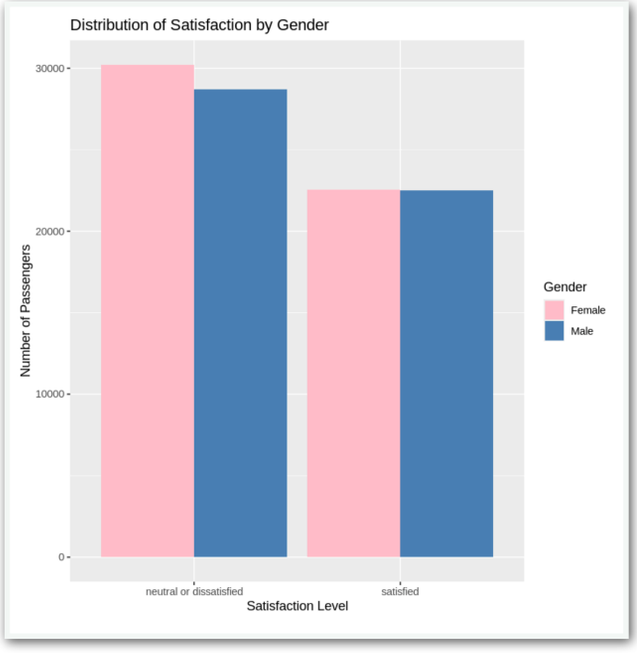
8. Data Summarization and Visualization



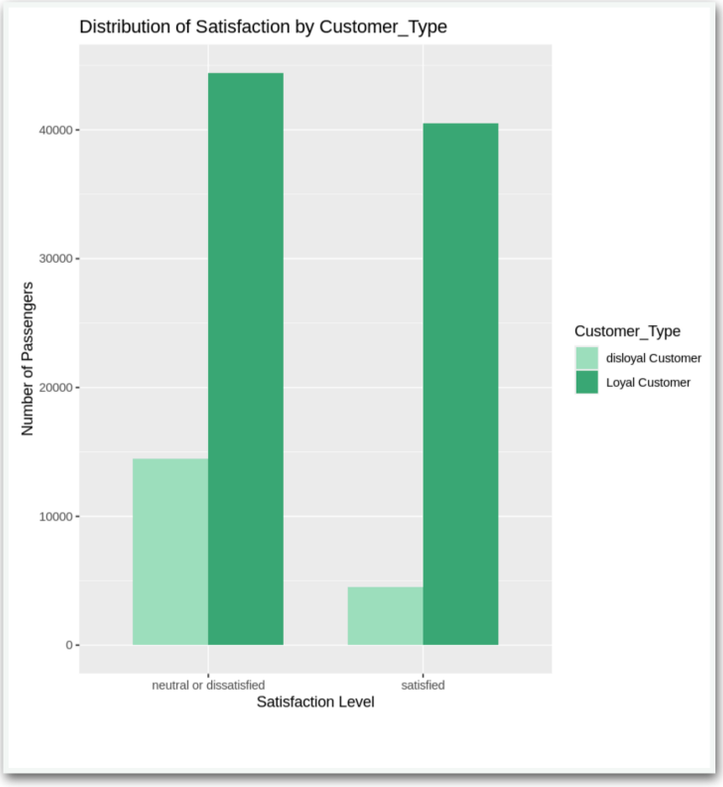
The diagram illustrates a difference in the number of satisfied and dissatisfied customers. It is evident that the number of dissatisfied customers is significantly high. The upcoming diagrams will provide a detailed analysis of the possible reasons behind this situation.

Gender <chr>	Type_of_Travel <chr>	con <int>
Female	Business travel	36528
Female	Personal Travel	16199
Male	Business travel	35127
Male	Personal Travel	16050

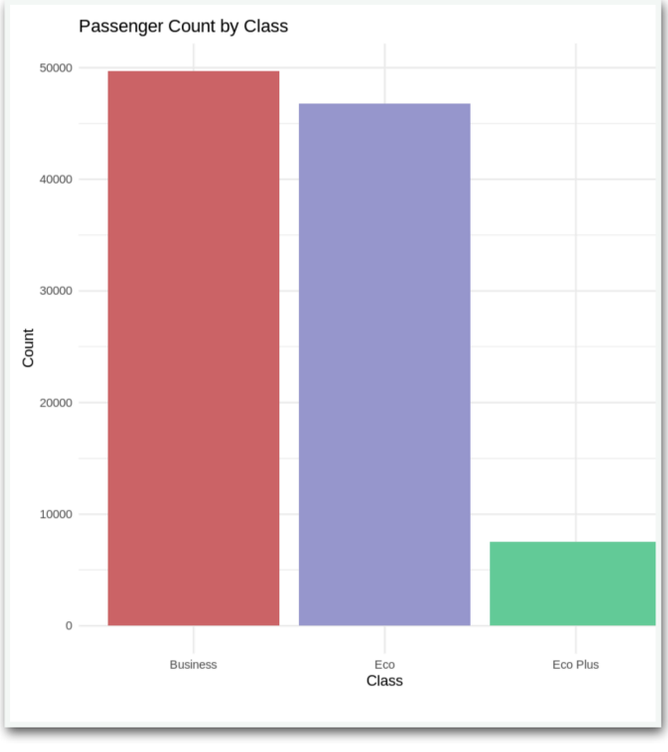
In the chart at the beginning, women and senior clients actually show satisfaction, but women are less intentionally unacknowledged or neutral than males. This indicates the importance of meeting women's needs and the necessity of their services on board.



```
disloyal Customer    Loyal Customer
            18981             84923
```

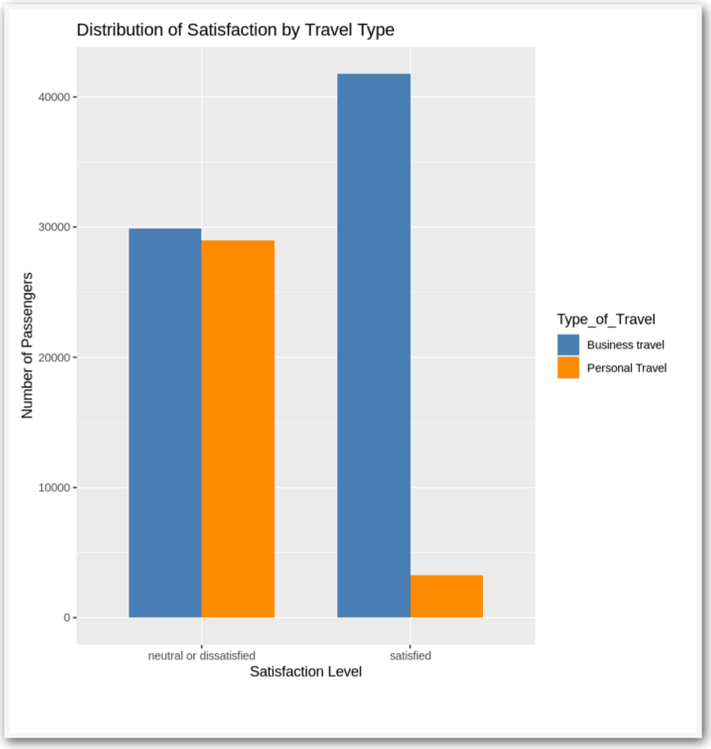


In the graph, we notice that loyal customers show a higher level of satisfaction compared to disloyal customers. Loyal customers also appear to be more dissatisfied or neutral than disloyal customers.
It is worth noting that the majority of disloyal customers show a neutral dissatisfied
We conclude from this that it is better to provide offers and additional benefits to loyalists. These offers can include loyalty programs, special offers, service features, and exclusive merchandising. By offering such offers, you reward loyal customers with satisfaction and the possibility of them becoming repeat customers.



We note here that the majority of customers choose Economy or Business class. It is worth noting that the number of customers who chose the advanced economy class (Eco Plus) is less than 10,000.
To solve the problem, the following actions can be taken:

1. Improving the added value in advanced economy class.
2. Directly targeting potential customers and promoting the additional benefits of the degree.
3. Providing special offers and discounts on advanced economy class tickets.
4. Improving customer experience in advanced economy class.
5. Regularly analyze choices and respond to improve services and marketing.



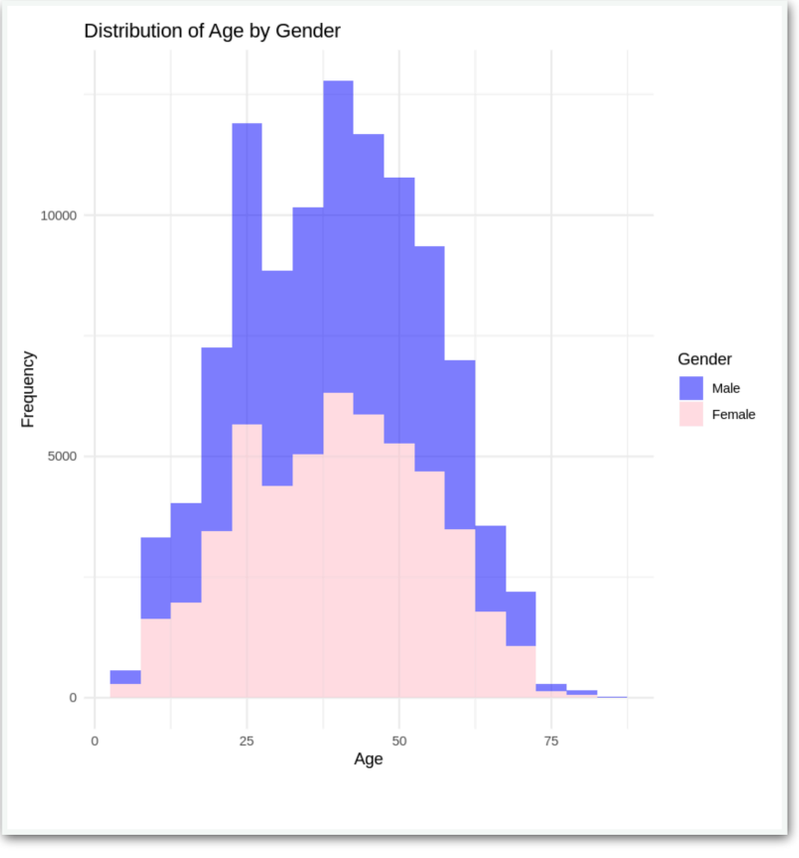
The analysis indicates that the majority of customers on personal and business flights are dissatisfied, but customers on business flights are shown to be more satisfied than customers on personal flights.

1. Improving the quality of service in personal flights

2. Improving customer experience on business trips
Additional services such as Wi-Fi on board and facilities for customers to do their work during the trip can be provided.

3. Listening to customer feedback: There must be a mechanism to effectively collect and respond to customer feedback and inquiries, so as to improve services and meet customer expectations.

This plot illustrates the age distribution by gender, showing that most men are in their thirties and around the age of 25, while women are primarily in their thirties and forties. It provides insights into the prevalent age groups for each gender.



References

[1] Klein, T. (2020) Airline passenger satisfaction, Kaggle. Available at: https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

[2] Alvin, T.P. (2024) Predicting satisfaction of airline passengers with classification, Medium. Available at: https://towardsdatascience.com/predicting-satisfaction-of-airline-passengers-with-classification-76f1516e1d16

[3] Zach. (2020, November 4). K-Fold Cross Validation in R (Step-by-Step). Statology. https://www.statology.org/k-fold-cross-validation-in-r/

[5]“Poe - Fast, Helpful AI Chat.” Poe.com, poe.com.

[6] *Customer Satisfaction analysis, Modelling, | R*. (n.d.). Kaggle.com. Retrieved May 17, 2024, from https://www.kaggle.com/code/khsamaha/customer-satisfaction-analysis-modelling-r/input

[7]Team, B. (2018, September 18). How to read PCA biplots and scree plots. Medium. https://bioturing.medium.com/how-to-read-pca-biplots-and-scree-plots-186246aae063

[8] https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/multivariate/how-to/principal-components/interpret-the-results/all-statistics-and-graphs/

[9]datacamp, https://www.datacamp.com/tutorial/detect-anomalies-anomalize-r. (n.d.).