Predicting Airline Passenger Satisfaction with Classification Algorithms

GROUP WORK

GROUP MEMBER	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

Our Analysis Project

About Flight-based customer evaluation is crucial in the aviation industry. Customer satisfaction and a comfortable experience are paramount for airlines. Understanding and meeting customer needs effectively is essential for success. Satisfied customers are more likely to return and recommend the company. Conversely, poor customer experience can lead to loss of customers and a negative impact on the business.

THIS IS OUR GOALS

- Analysis of satisfaction and dissatisfaction factors
- Improving service quality
- Boosting loyalty and strategic direction

THE DATASET

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

THE DATASET

using the (dim) function, displays the number of columns and rows in the dataset

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

using the (is null) function,It appears from the code that there are 310 empty/null values in the column"Arrival_Delay_in_Minutes"

THE DATASET

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

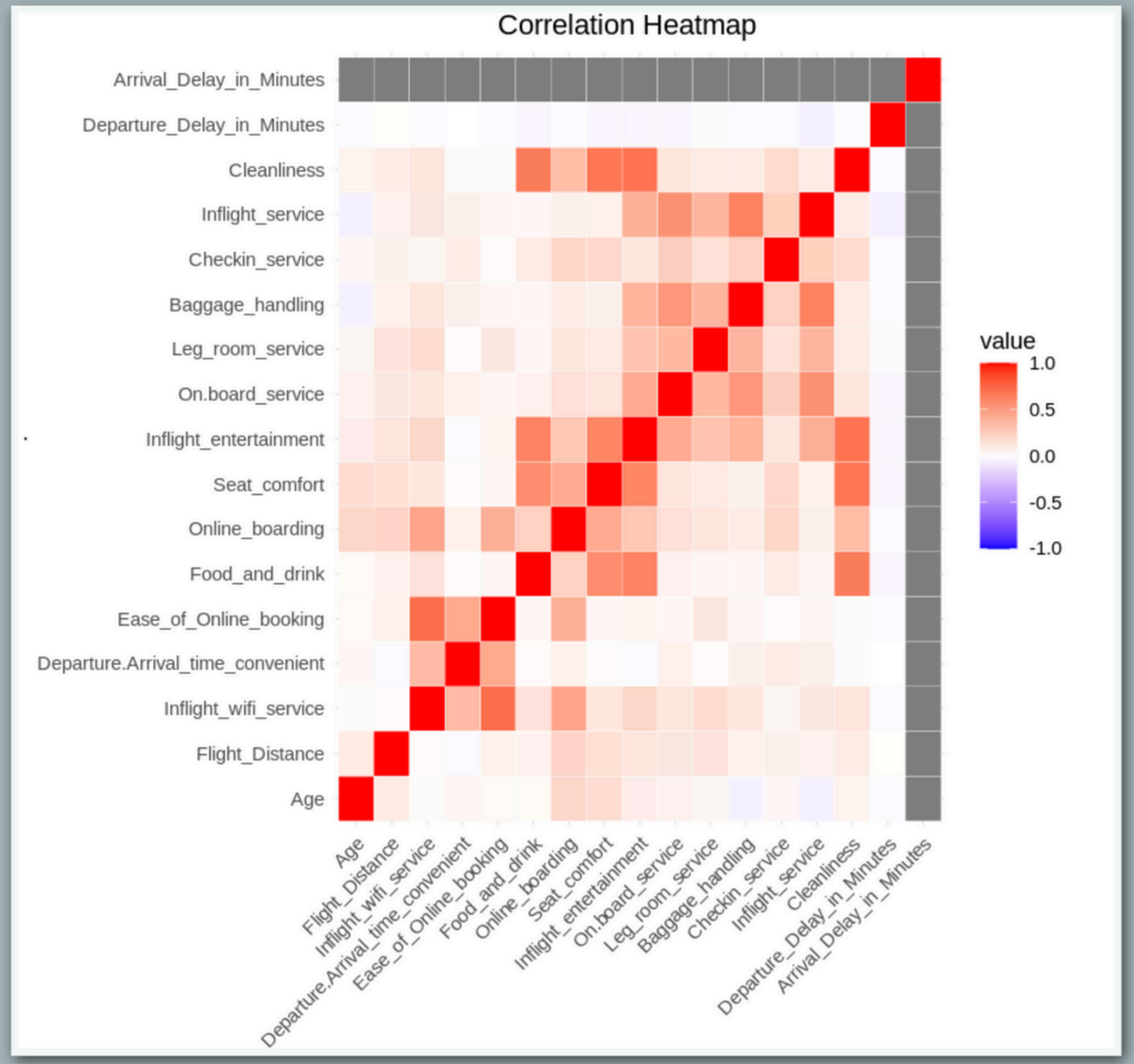
Here convert some into numerical values. For ease of modeling or analysis.

THE DATASET

```
       X          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_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
```

We used the "summary" function to explain the values such as the mean, median, minimum, and maximum values.

CORRELATION HEATMAP




```
Arrival_Delay_in_Minutes                    NA
                          Arrival_Delay_in_Minutes
```

"Arrival delay in Minutes" is gray

This indicates a weak correlation between the data because there are a significant number of missing values.

LINEAR AND NONLINEAR REGRESSION

```
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

LINEAR AND NONLINEAR REGRESSION

```
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
```

- **RMSE : 1.04512317267583**
- **R-squared : 0.47860749253182**

the RMSE and R-squared values, we found the R-squared value to be average, suggesting a potential overfitting problem

MODEL 2

the First model is better than first model

LINEAR AND NONLINEAR REGRESSION

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 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.

LINEAR AND NONLINEAR REGRESSION

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

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.

CLASSIFICATION

MODEL 1

- Food and drink,Cleanliness,
- Inflight wifi service,
- Ease of Onlinebooking

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

MODEL 2

- Ease of Online booking,Food and drink,Online boarding,Seat comfort,
- Inflight entertainment,On.board service,Leg room service,Baggage handling,
- Checkin service,Inflight service,
- Cleanliness

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

the second model is better than first model

CLUSTERING

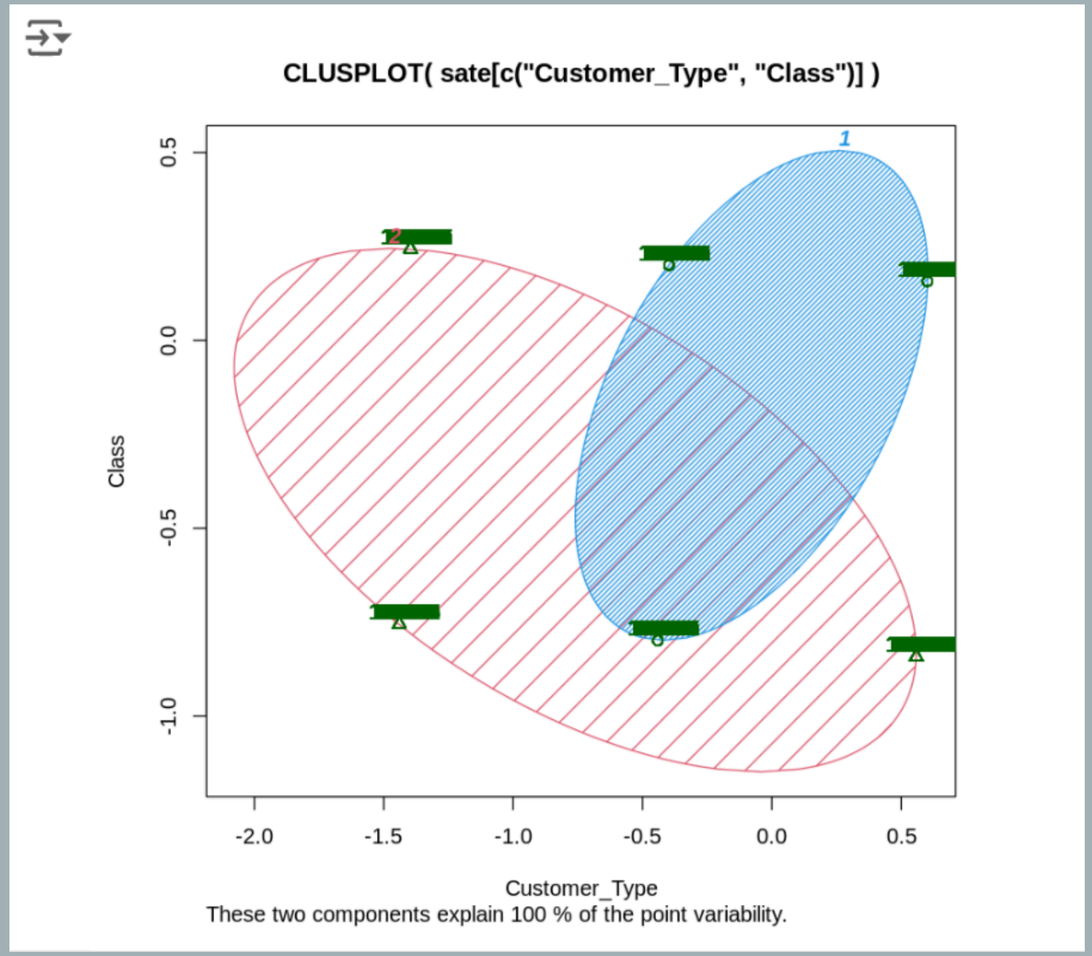
Cluster 1

- We use 2 cluster

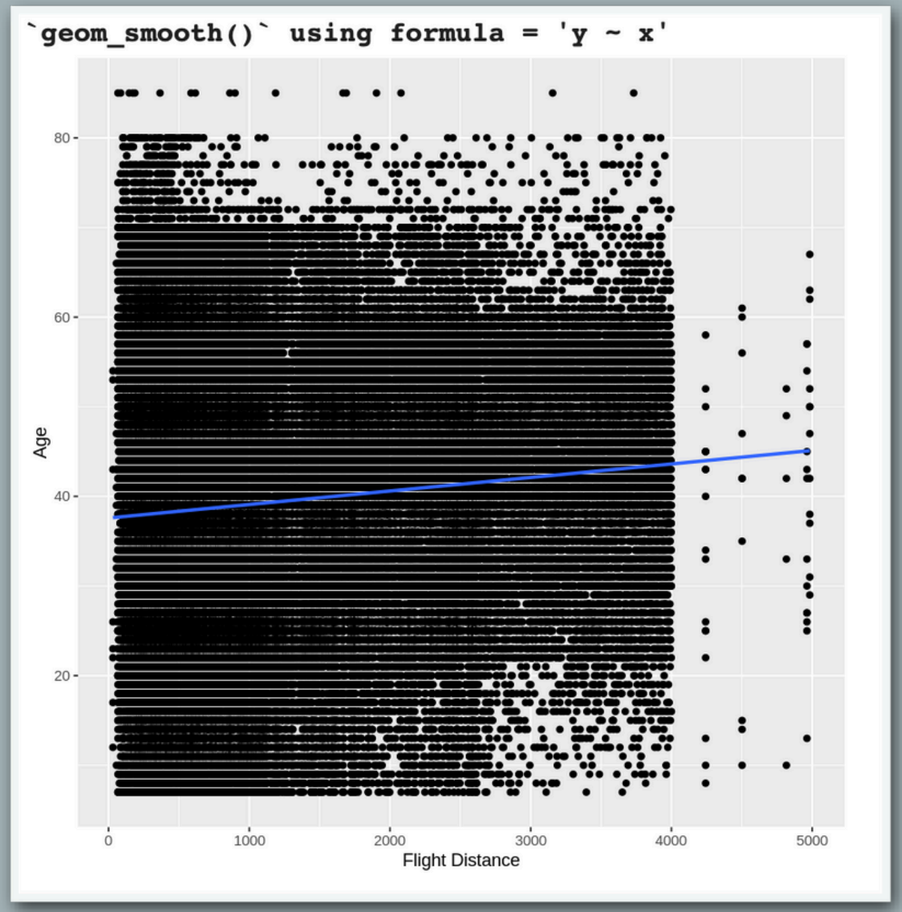
- The K-means algorithm was applied.

- On the y axis we have the Customer column and on the x axis there is the class type

- We note here that the two types (loyal, Disloyal) came together at two points, namely (Eco and Business)



ANOMALY DETECTION



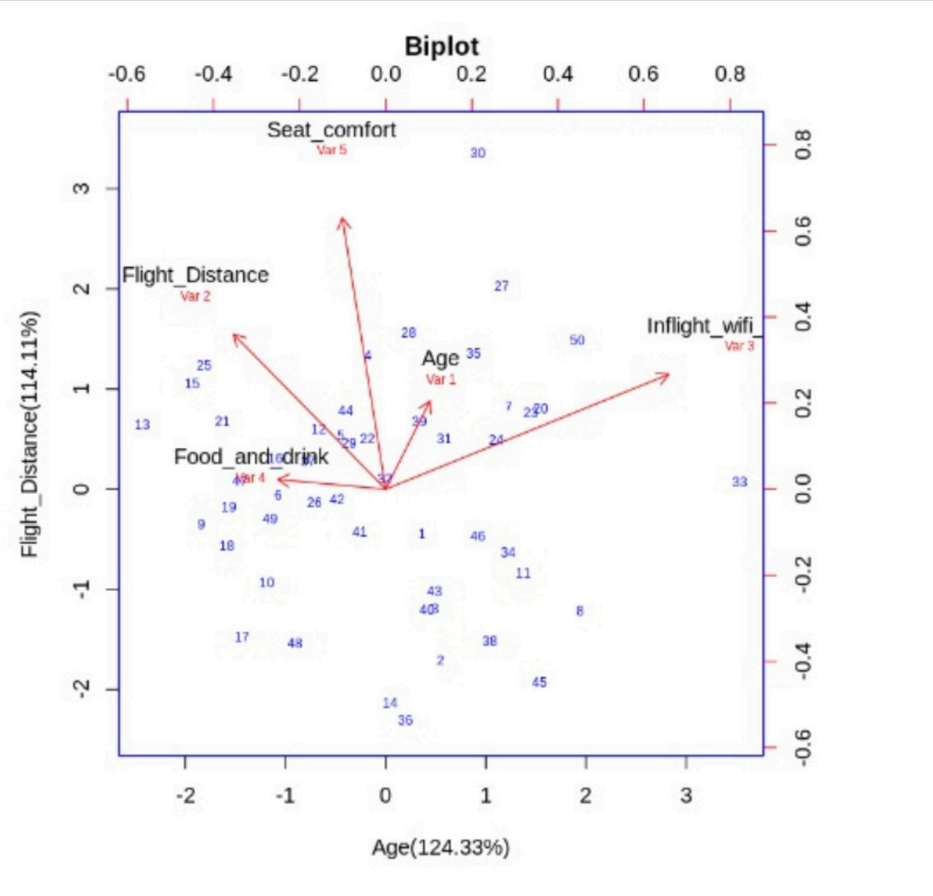
Based on the analysis our conclusions are as follows

1. older customers travel shorter distances.with exceptions.
2. The typical flight distance is around 4,000 miles, with exceptions.
3. The majority of customers fall within the 20-60 age range.

Recommendations for the airline

- Targeted offerings.
- Special services for older customers
- Targeted marketing campaigns
- Customer segmentation

PCA



From the given biplot, we can infer the following:

Positive relationships

- There are positive relationships between seat comfort, flight distance, and food and beverages.

-There is a positive relationship between in-flight Wi-Fi service and age.

Negative relationships

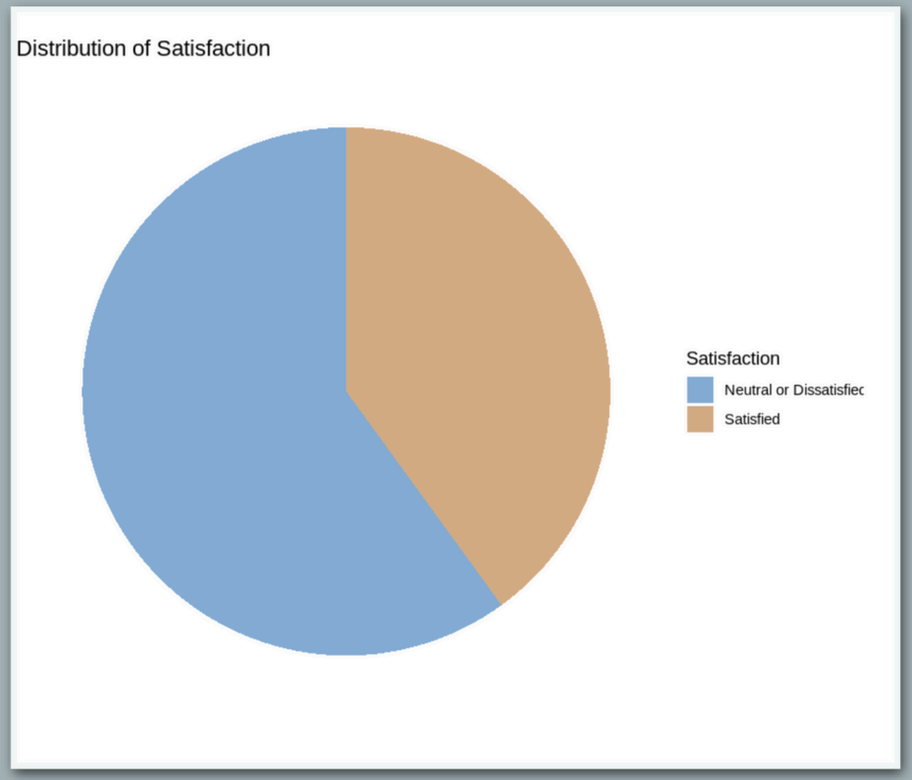
- Negative relationships exist between in-flight Wi-Fi service and flight distance, food and beverages.

- There is also a negative relationship between age and food and Drink

Weak influence

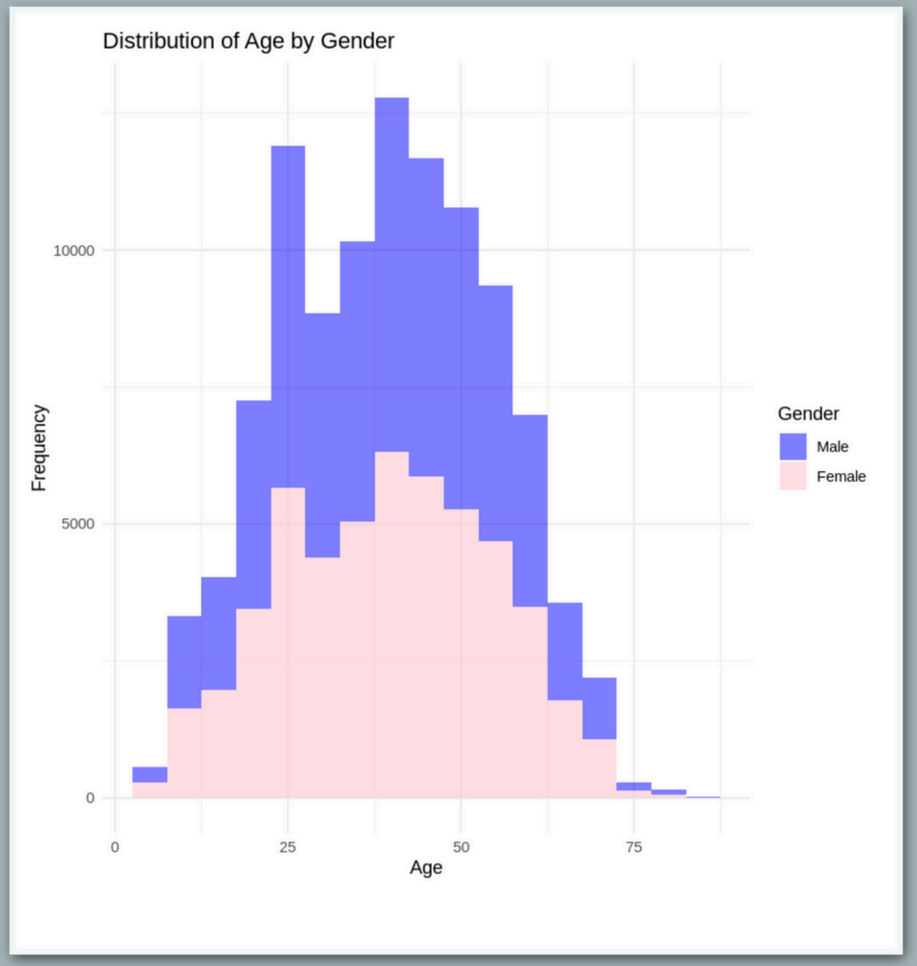
- Food and beverages have a weak influence on the components, as indicated by their proximity to 0 on the biplot.

DATA SUMMARIZATION AND VISUALIZATION



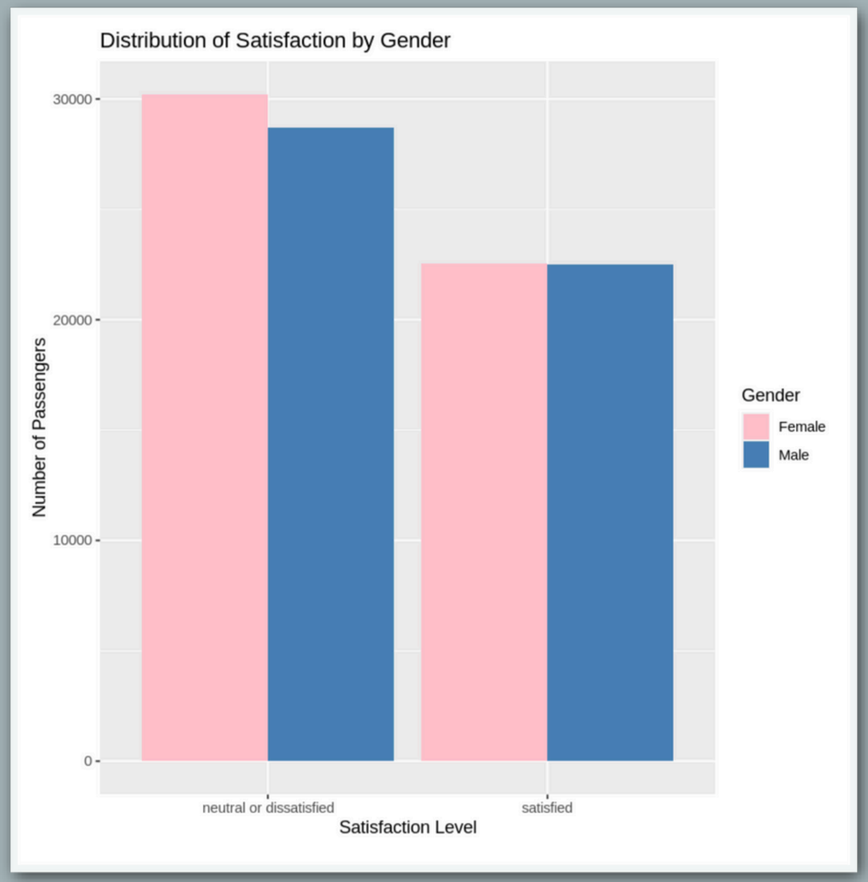
The diagram illustrates a difference in the number of **satisfied** and **dissatisfied customers**

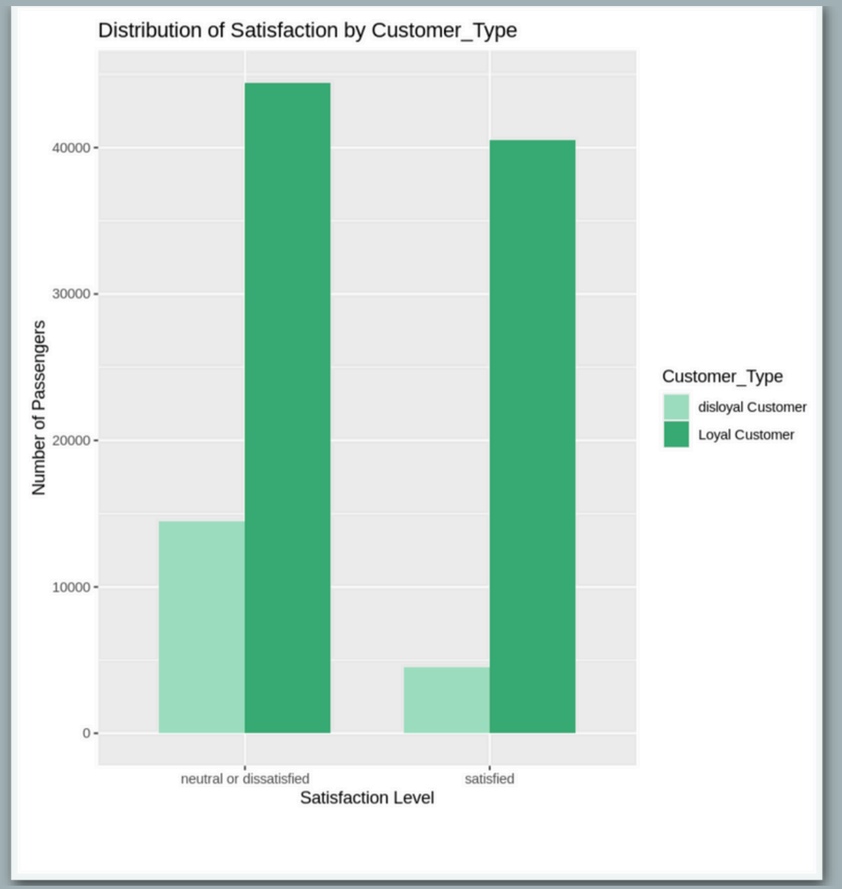
DATA SUMMARIZATION AND VISUALIZATION



This is charts show Distribution of **Age** by **Gender**

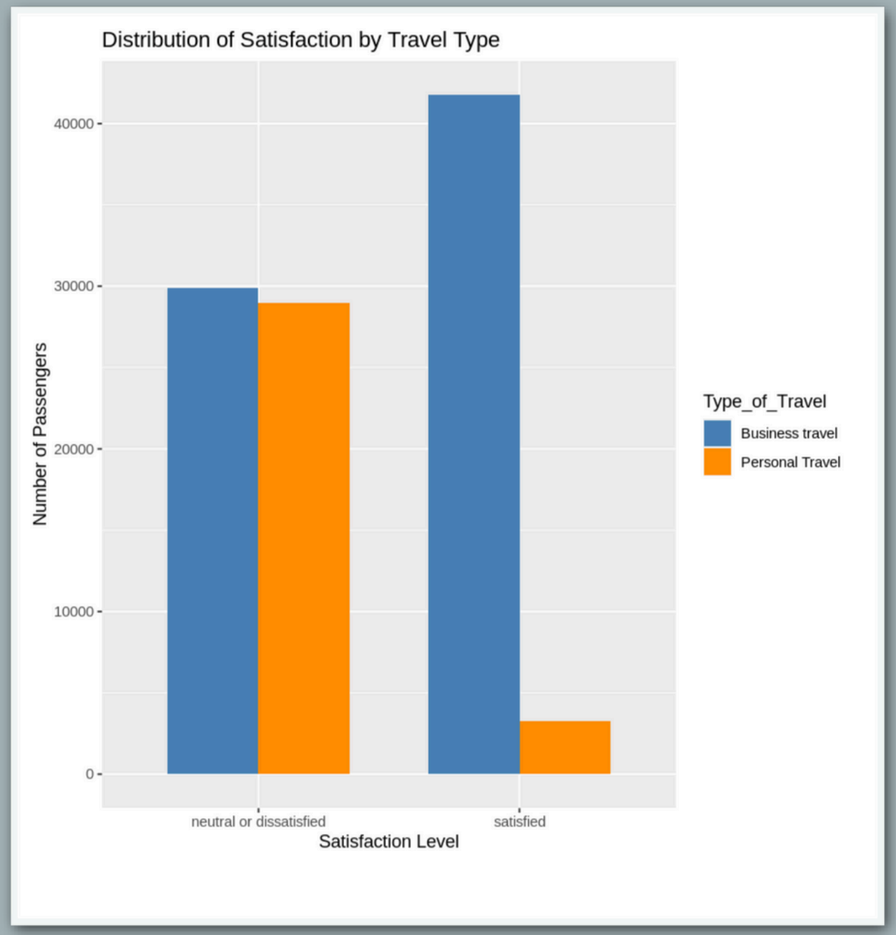
DATA SUMMARIZATION AND VISUALIZATION





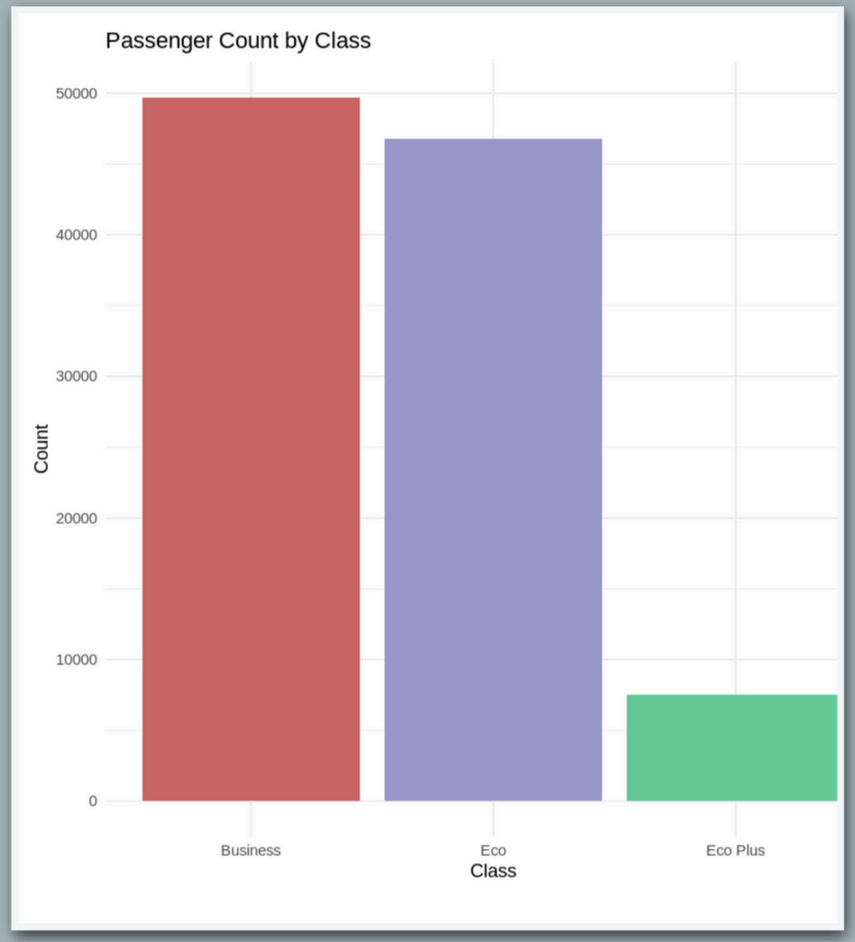
This is charts show Distribution of Satisfaction by **Gender** and **Customer Type**

DATA SUMMARIZATION AND VISUALIZATION



This is charts show Distribution of **Satisfaction** by **Travel Type**

DATA SUMMARIZATION AND VISUALIZATION



The chart show **Passenger Count** by **Class**

CHALLENGES AND BENEFITS

CHALLENGES

- CHOOSE CORRECT DATA TO WHICH ALL REQUIREMENTS CAN BE APPLIED
- UNDERSTAND GRAPHS CLEARLY AND CORRECTLY
- RACING AGAINST TIME AND ENDING WITH GOOD AND UNDERSTANDABLE RESULTS

BENEFITS

- ORGANIZING TIME AMONG GROUP MEMBERS
- UNDERSTANDING AND RESPECTING DIFFERENT POINTS OF VIEW
- QUICK UNDERSTANDING AND ANALYSIS OF DATA


Thank you