FLYING HIGH

A Study on Improving Customer Satisfaction and Airline Performance

Table of Contents

1. Introduction **2**
2. Situation analysis (EDA) **2**
3. EDA data that provides a better understanding of the issue **3**
4. Reasoning and Techniques used for analysis **4**
4.1 Logistic Regression 4
4.2 Clustering 4
5. Assessment of the model(s) **4**
5.1 Logistic regression 4
5.2 Clustering 5
6. Conclusions and Recommendations **7**
7. Appendix1 that includes charts and figures on relevant analysis **9**

1. Introduction

This dataset contains consumer survey data and airline data that reflects customer relationships with the airlines. The goal is to help the airlines with improving customer satisfaction and improve their own performance.

2. Situation analysis (EDA)

The density chart shows that middle-aged customers are more satisfied than younger and older customers (fig1), who may have higher expectations or difficulties with technology.Insight: Personalizing the experience according to different age groups' preferences is crucial. Providing engaging services through social media for younger customers and simple, easy-to-use services for older customers can enhance satisfaction levels.

The Blue category has a higher percentage of unsatisfied customers compared to premium categories(fig 2) probably due to better service quality for premium customers. Insight: To improve Blue category satisfaction levels, businesses can offer incentives like discounts or special promotions to make them feel valued and increase loyalty.

Business travelers have the highest satisfaction levels, followed by mileage ticket holders and personal travelers(fig 3). It is because Business travelers receive more amenities like priority boarding ,lounge access and comfortable seating options.Insight:Personal travelers have lower expectations and fewer amenities, but offering more amenities like free meals and priority boarding can increase their satisfaction levels.

There seems to be more satisfied customers during the middle of the month (fig 4). The data suggests that timing can impact customer satisfaction levels. This could be because customers have received their paychecks and are more likely to be in a positive state of mind.Insight:

Businesses should take this into account when planning their operations and marketing campaigns to maximize customer satisfaction levels.

The difference in mean number of flights between satisfied and dissatisfied customers is significant, with dissatisfied customers taking more flights on average (24 vs 17) (fig 5). Insight:Addressing the issues faced by frequent travelers such as flight delays, cancellations, and lost baggage can help in improving their satisfaction levels. Offering timely updates, compensation, and support can go a long way in enhancing their experience and loyalty.

By looking at correlation of numerical variables,Flight time and flight distance are highly correlated (0.94), as are arrival delay and departure delay (.96) (fig6). Insight:So,we have dropped flight distance and departure delay as flight time(fig 7) and arrival delay(fig 8) show more separation between satisfied and dissatisfied respectively.

For clustering, we removed some variables such as no_of_flights, scheduled_departure_hour, because they may not add much value to the satisfaction level and to dropped a few categorical columns such as origin_state, destination_state, month_of_flight_date, flight_date to reduce computation complexity. Also, looking at correlation plot we can infer that gender binary, airline status and travel type have a comparatively more significant correlation with satisfaction level than other variables. However, as most of our variables are categorical that have been numerically encoded, it is difficult to find any significant correlation between them (fig 9).

3. EDA data that provides a better understanding of the issue

Based on EDA, we found there were variables for which the visuals for satisfied and dissatisfied customers did not show noticeable differences (i.e satisfaction and not satisfied are around 50% each).These insignificant variables were dropped from the dataset to focus on the variables that impact customer satisfaction. The final dataset includes satisfaction, airline_status, age, type of travel, day of flight date, no of flights, flight time, and arrival delay. Also, we created dummies for categorical variables and converted the binary dataset to numerical.

4. Reasoning and Techniques used for analysis

4.1 Logistic Regression

The dependent variable 'satisfaction' is binary. Therefore, we decided to use Logistic regression as it can model the probability of the binary outcome as a function of the independent variables. We used general logistic regression and since there are a large number of predictor variables, we also used backward elimination as it helps to identify the most important predictors in the model and remove the insignificant ones.

4.2 Clustering

We split the original dataset into two datasets: Satisfied Customers and Not Satisfied Customers. We applied Hierarchical Clustering and K-means Clustering on each of the two datasets in order to understand the characteristics of the customer's segment. Hierarchical clustering was used because it can make groups of the data to help in understanding the output of the algorithm. K-Means was used because it helps to find differences between the clusters.

5. Assessment of the model(s)

5.1 Logistic regression

Satisfaction(binary) is considered the dependent variable and airline_status, age, type_of_travel, day_of_flight date, no_of_flights, flight_time, and arrival_delay as the independent variable. The general Logistic regression model (LRM) identified 11 statistically significant variables(fig 10).Backward elimination using logistic regression (BWM) identified 13 statistically significant variables(fig11) (fig 14). LRM had lower AIC and BIC scores for test set and validation set in comparison to BWM(fig 12). Also, based on classification report , LRM and BWM perform similarly on the validation and test sets but the precision for the LRM model is slightly higher (90.7%) vs backward (90.3%)(fig 13). Overall based on all scores, LRM is chosen as the preferred model.

Final model(fig 14 & fig 15) is as follows which in a business context can be used to predict whether the customer will be satisfied or dissatisfied: logit (p)=

1.61+2.24*airline_status_gold+1.50*airline_status_platinum+5.36*airline_status_silver–0.64*type_of_travel_mileage_ticket–0.04*type_of_travel_personal_travel–0.60*day_of_flight_date6+1.52*day_of_flight_date15–0.93*age–0.82*no_of_flight–0.93*flight_time–0.78*arrival_delay

Keeping other predictors constant, customers with airline_status_silver have 5.36 times higher odds of being satisfied compared to customers with airline_status_blue (the reference level) followed by gold_status(2.24 times) and platinum_status(1.5 times). Customers with type_of_travel_mileage_tickets (-0.64 times) and type_of_travel_personal travel (-0.04 times) are less likely to be satisfied than type_of_travel_business (the reference level). Flight_date 15

(1.52 times) has better odds than any other date in making customers satisfied.

For the continuous variables like age, no of flights, flight time and arrival delay we can that holding all other predictors constant one unit increase in these variables decreases the odds of a customer being satisfied by a factor of 0.93,0.82,0.93 and 0.78 respectively .

5.2 Clustering

- **Hierarchical Clustering For Satisfied Customers**

Silhouette Score suggests that the Optimal Number of clusters are between 2 and 3(fig 16). On comparing Dendrograms of 2 and 3 clusters (fig 16) both looked similar with 3 showing slightly better results. Hierarchical clustering using 3 clusters suggests clusters are not evenly distributed and shows around 41%, 34% and 25% (fig 17). We found that customers in all Groups have similar average age, average shopping amount, etc(fig 18). The results from the cluster plot seem to be overlapping and hence difficult to interpret,so we performed K-means (fig 19).

- **K-Means for Satisfied Customers**

Elbow chart suggests the optimal number of clusters to be between 2 and 5. Silhouette chart suggests the optimal number of clusters to be between 2 and 3 (fig 20). The ANOVA result for 2 clusters did not show an F-value and p-value because the value of the residuals are zero which means the model perfectly explains the variation in the data. The data for 2 clusters may not be suitable for ANOVA or there is an issue with the model having 2 clusters. Considering ANOVA on 3 clusters, Departure Delay and Class both are statistically significant. and Airline Status is not statistically significant (fig 21). Also, the cluster plot for 3 clusters using K means clearly visualizing the difference between clusters. So we conclude a 3 cluster solution(fig 22)

- **Hierarchical Clustering For Not-Satisfied Customers**

Silhouette Score suggests that the Optimal Number of clusters are 2 and 3 (fig 23). On comparing dendrograms of 2 and 3 clusters both looked similar with 2 showing slightly better results (fig 24). Hierarchical clustering using 2 clusters suggests clusters are not evenly distributed and shows around 52% and 48%. For the further step, we aggregate to calculate mean values of each variable in the data grouped by the cluster assignments in hcluster (fig 25). Based on these values, we found that customers in both Groups experienced more arrival delays than departure delays.. The results from the cluster plot seem to be overlapping and hence difficult to interpret, so we performed K_Means.(fig 26).

- **K-Means for Not-Satisfied Customers**

Elbow chart suggests the optimal number of clusters to be between 2 and 5. Silhouette chart suggests the optimal number of clusters to be between 2 and 3 (Fig 27). The ANOVA result for 2 clusters did not show an F-value and p-value because the value of the residuals are zero which means the model perfectly explains the variation in the data. The data for 2 clusters may not be suitable for ANOVA or there is an issue with the model of having 2 clusters. Considering ANOVA on 3 clusters. Departure Delay and Class are both statistically significant and Airline Status is not statistically significant ((fig 28). Also, the cluster plot for 3 clusters using K means clearly visualizing the difference between clusters. So we conclude a 3 cluster solution .(fig 29)

6. Conclusions and Recommendations

Based on the coefficients in the Logistic regression model, the factors with the highest impact on customer satisfaction are:

1. Airline status (silver status has highest positive impact, followed by gold and than platinum)
2. Day of the flight (flights on 15th day have higher impact on satisfaction than other dates)
3. Arrival delay (a higher arrival delay has a negative impact on satisfaction)

The company should focus on improving their performance in these areas to increase customer satisfaction. They can offer incentives and rewards to encourage customers to achieve silver, gold, or platinum status by creating loyalty programs that incentivize customers to fly more frequently. They can also work on improving their flight schedules to minimize delays and prioritize flights on the 15th of the month. Additionally, they can work on improving their on-time performance by optimizing flight routes and providing regular updates to customers during delays, to reduce the impact of delays on customer satisfaction.

Conclusion and recommendation for clustering:

Based on Clustering Analysis, it is recommended to divide the customer base into three clusters/segments in order to understand them better and give more customized service to them.

An important observation from the analysis is that females are more satisfied than males when compared with satisfied customer base and not satisfied customer base. Also, considering age as a variable, customers that are 24 years or younger are more satisfied while not satisfied customer base comprises of the age group above 45 years.

With respect to delay time during departure, it was observed to be the same for both groups(satisfied and not satisfied) as well as through all clusters. Based on the observation that most customers shop/eat at the airport, in order to compensate for a delay in departure, discount coupons or flight points can be given.

Also, understanding the pain points of customers will help cater to them and increase their level of satisfaction. Another important recommendation is to ask for a quick customer feedback to show accountability and willingness to improve.

Appendix1 that includes charts and figures on relevant analysis

Fig1

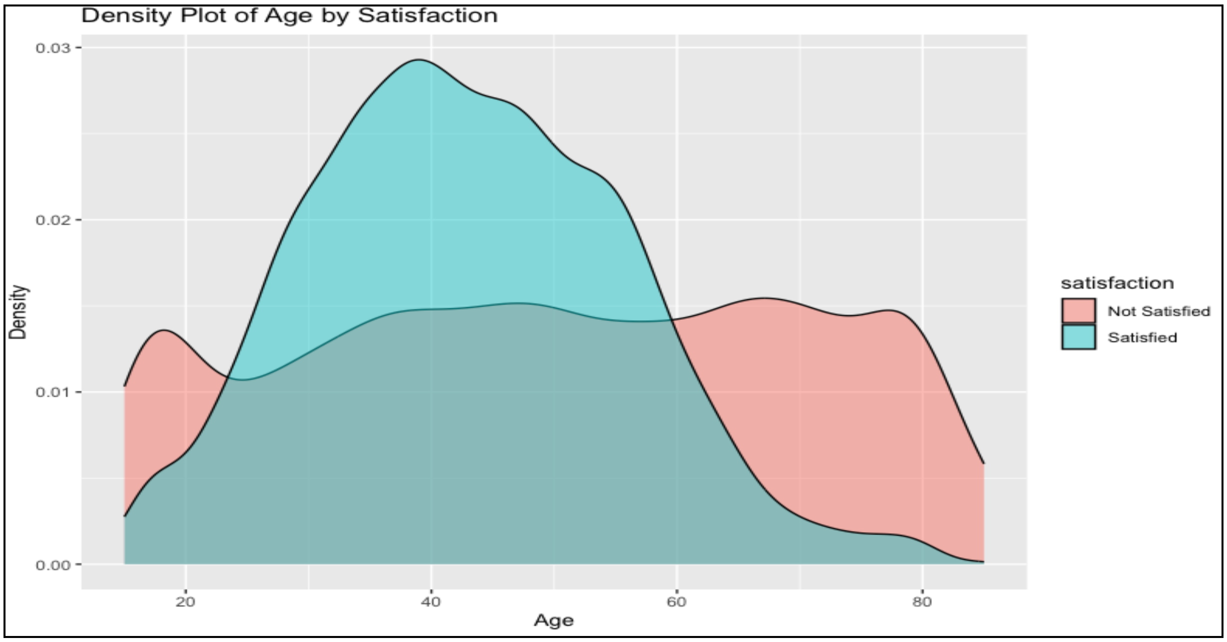


Fig 2

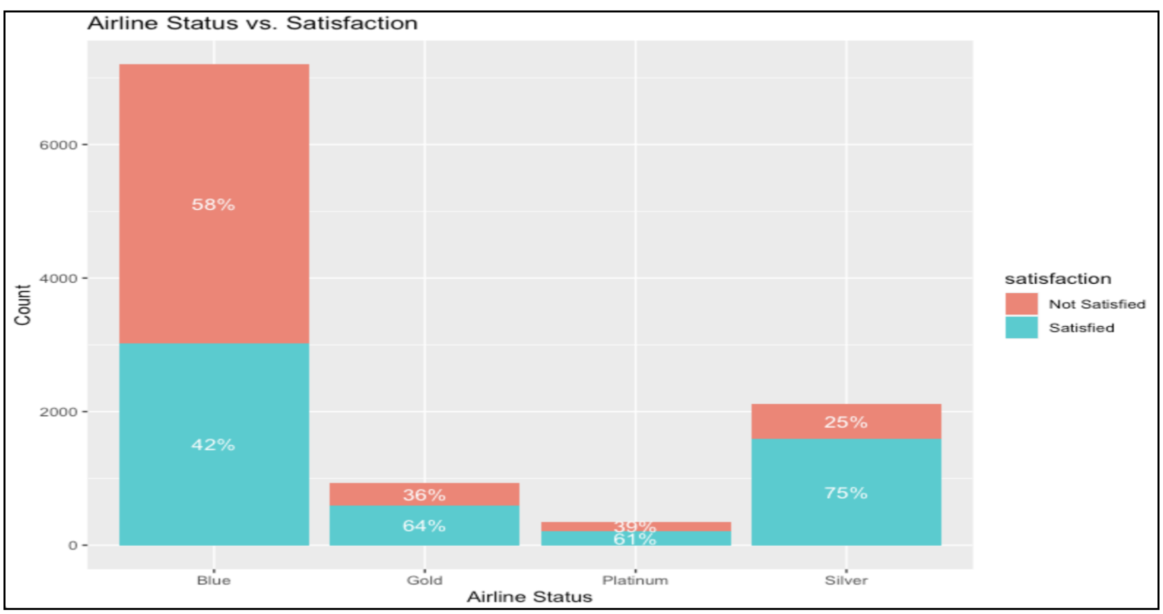


Fig 3

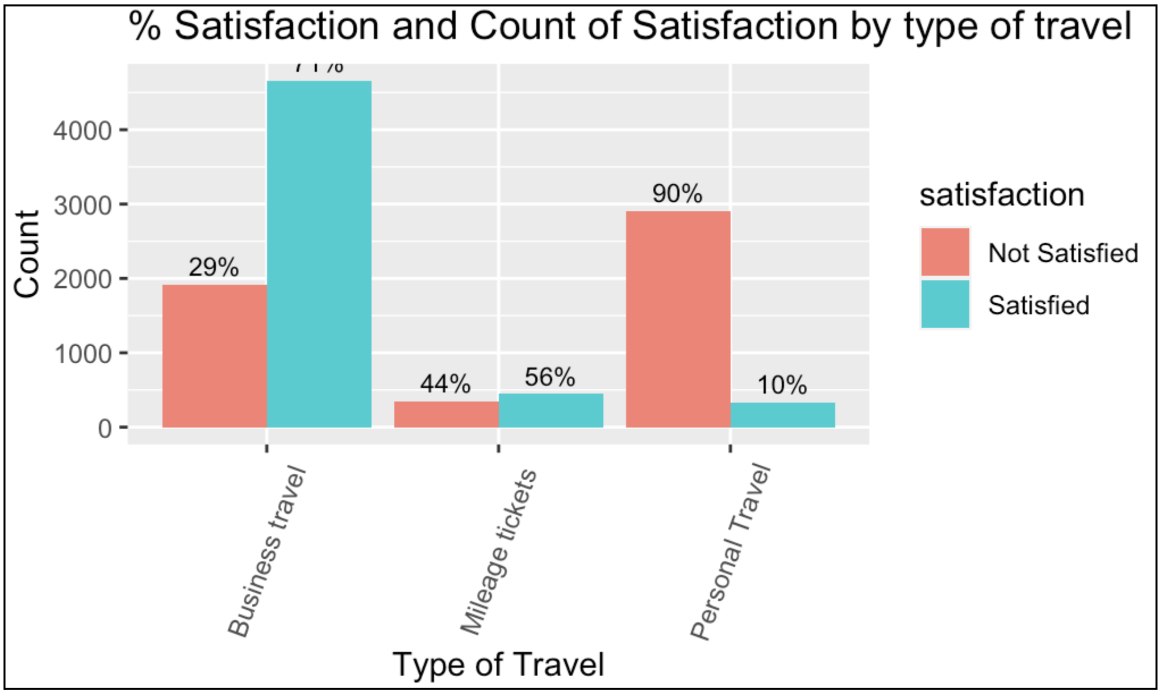

% Satisfaction and Count of Satisfaction by type of travel
Count
4000
3000
2000
1000
0
29%
90%
44%
56%
10%
satisfaction
Not Satisfied
Satisfied
Business travel
Mileage tickets
Personal Travel
Type of Travel

Fig 4

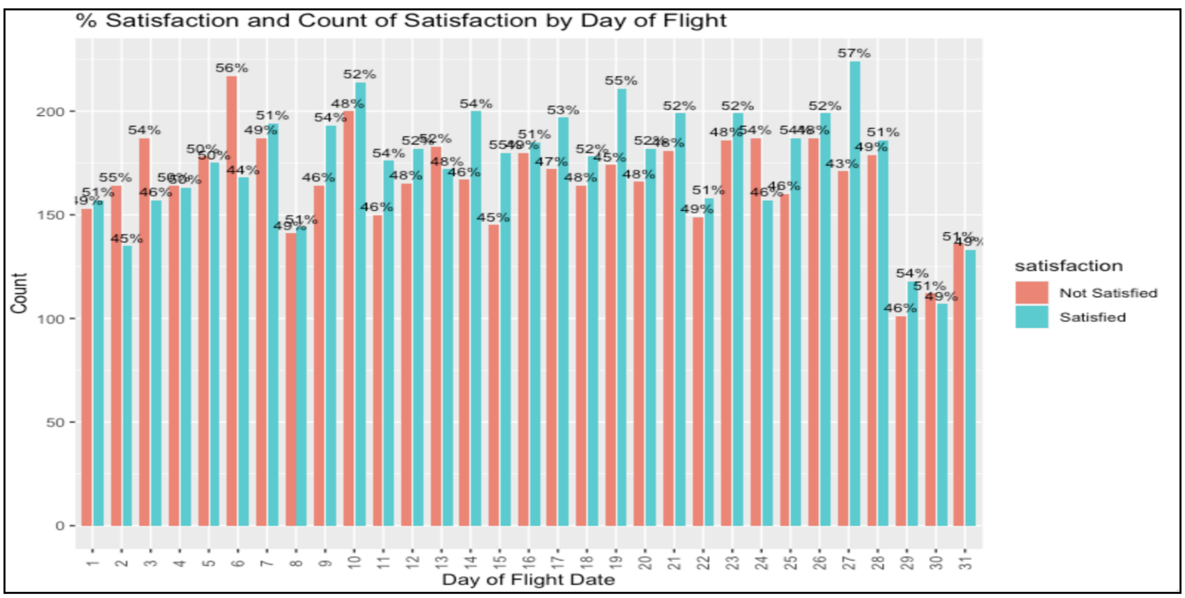

% Satisfaction and Count of Satisfaction by Day of Flight
Count
200
150
100
50
0
satisfaction
Not Satisfied
Satisfied
Day of Flight Date

Fig 5

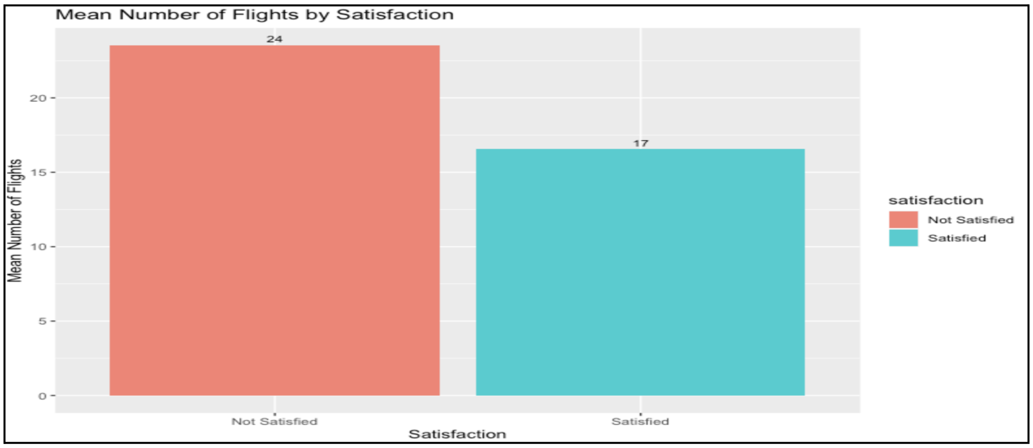

Mean Number of Flights by Satisfaction
24
17
Mean Number of Flights
20
15
10
5
0
Not Satisfied
Satisfied
Satisfaction
satisfaction
Not Satisfied
Satisfied

Fig6

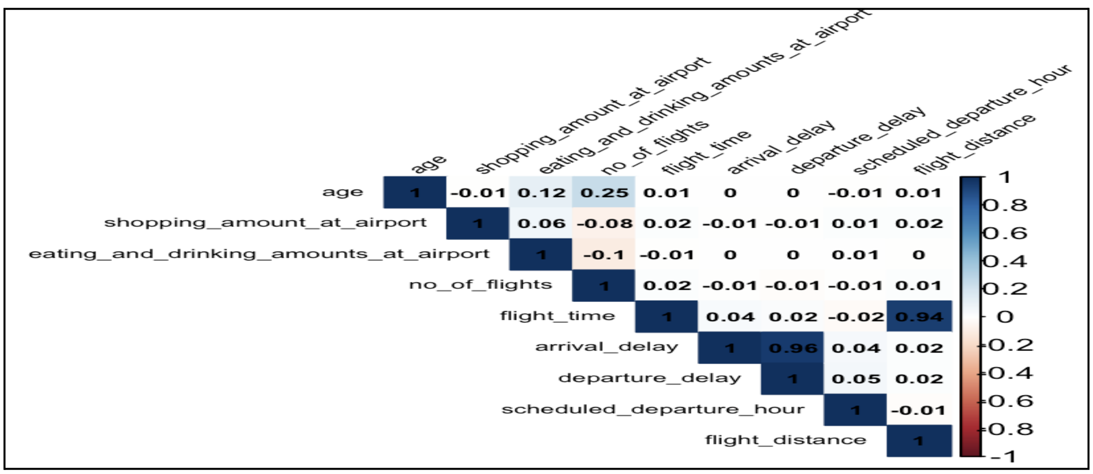

age
shopping_amount_at_airport
eating_and_drinking_amounts_at_airport
no_of_flights
flight_time
arrival_delay
departure_delay
scheduled_departure_hour
flight_distance

Fig 7

Fig 8

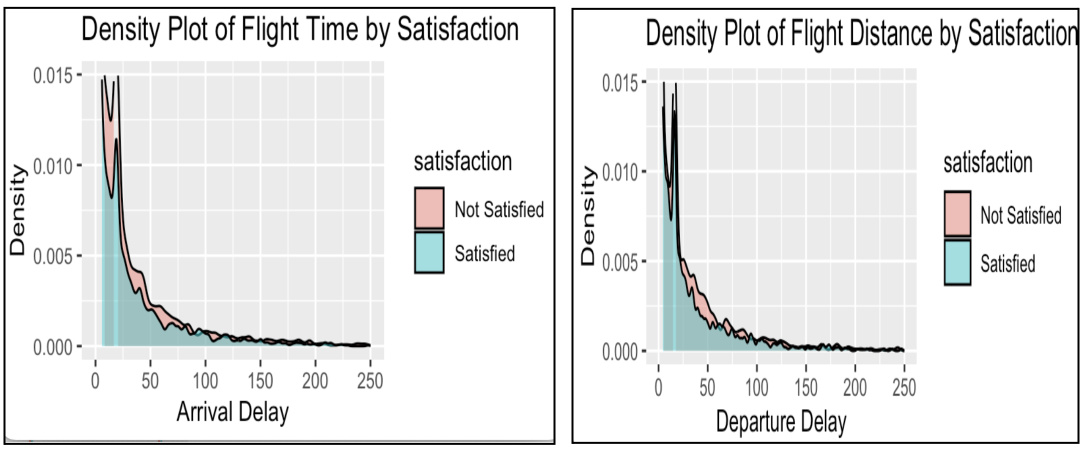


Fig 9

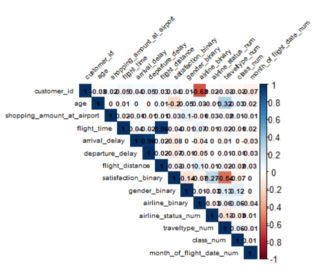


fig 10

```
                                 Estimate Std. Error z value Pr(>|z|)    
(Intercept)                     1.2318786  0.2080492   5.921 3.20e-09 ***
age                            -0.0044051  0.0019929  -2.210   0.0271 *  
no_of_flights                  -0.0139435  0.0022623  -6.163 7.12e-10 ***
flight_time                    -0.0021055  0.0008563  -2.459   0.0139 *  
arrival_delay                  -0.0055204  0.0007667  -7.200 6.00e-13 ***
airline_status_gold             0.8088252  0.1088945   7.428 1.11e-13 ***
airline_status_platinum         0.4131459  0.1590760   2.597   0.0094 ** 
airline_status_silver           1.6735749  0.0883425  18.944  < 2e-16 ***
type_of_travel_mileage_tickets -0.4514151  0.0989521  -4.562 5.07e-06 ***
type_of_travel_personal_travel -3.0616192  0.0872016 -35.110  < 2e-16 ***
day_of_flight_date2            -0.2067231  0.2435856  -0.849   0.3961    
day_of_flight_date3            -0.1274269  0.2438057  -0.523   0.6012    
day_of_flight_date4            -0.1381070  0.2449363  -0.564   0.5729    
day_of_flight_date5             0.2412217  0.2445385   0.986   0.3239    
day_of_flight_date6            -0.5136888  0.2328284  -2.206   0.0274 *  
day_of_flight_date7            -0.0966901  0.2365962  -0.409   0.6828    
day_of_flight_date8            -0.2786584  0.2479935  -1.124   0.2612    
day_of_flight_date9             0.0914987  0.2359976   0.388   0.6982    
day_of_flight_date10           -0.0856199  0.2255790  -0.380   0.7043    
day_of_flight_date11            0.1491577  0.2484427   0.600   0.5483    
day_of_flight_date12           -0.1751254  0.2360375  -0.742   0.4581    
day_of_flight_date13           -0.1752008  0.2415463  -0.725   0.4682    
day_of_flight_date14            0.0844576  0.2369068   0.357   0.7215    
day_of_flight_date15            0.4108168  0.2464838   1.667   0.0956 .  
day_of_flight_date16            0.0368302  0.2364772   0.156   0.8762    
day_of_flight_date17            0.1291792  0.2415248   0.535   0.5928    
day_of_flight_date18            0.0146469  0.2429060   0.060   0.9519    
day_of_flight_date19           -0.0289783  0.2277026  -0.127   0.8987    
day_of_flight_date20           -0.0068638  0.2416663  -0.028   0.9773    
day_of_flight_date21            0.1912465  0.2388880   0.801   0.4234    
day_of_flight_date22            0.0990800  0.2491274   0.398   0.6908    
day_of_flight_date23           -0.0584898  0.2337322  -0.250   0.8024    
day_of_flight_date24           -0.2092253  0.2384190  -0.878   0.3802    
day_of_flight_date25            0.1844876  0.2388244   0.772   0.4398    
day_of_flight_date26           -0.1590009  0.2350581  -0.676   0.4988    
day_of_flight_date27            0.3427299  0.2396588   1.430   0.1527    
day_of_flight_date28            0.0349972  0.2351503   0.149   0.8817    
day_of_flight_date29            0.0239777  0.2767890   0.087   0.9310    
day_of_flight_date30           -0.0312365  0.2765929  -0.113   0.9101    
day_of_flight_date31           -0.0858976  0.2539811  -0.338   0.7352    
age_scaled                             NA         NA      NA       NA    
no_of_flights_scaled                   NA         NA      NA       NA    
flight_time_scaled                     NA         NA      NA       NA    
arrival_delay_scaled                   NA         NA      NA       NA    
```

fig 11

```
Step:  AIC=6880.85
satisfaction_binary ~ airline_status_gold + airline_status_platinum + 
    airline_status_silver + type_of_travel_mileage_tickets + 
    type_of_travel_personal_travel + day_of_flight_date5 + day_of_flight_date6 + 
    day_of_flight_date15 + day_of_flight_date27 + age_scaled + 
    no_of_flights_scaled + flight_time_scaled + arrival_delay_scaled

                                 Df Deviance    AIC
<none>                                6852.8 6880.8
- day_of_flight_date5             1   6855.3 6881.3
- age_scaled                      1   6857.9 6883.9
- day_of_flight_date27            1   6858.0 6884.0
- flight_time_scaled              1   6858.8 6884.8
- day_of_flight_date15            1   6859.2 6885.2
- airline_status_platinum         1   6859.6 6885.6
- day_of_flight_date6             1   6862.2 6888.2
- type_of_travel_mileage_tickets  1   6872.9 6898.9
- no_of_flights_scaled            1   6891.3 6917.3
- airline_status_gold             1   6911.4 6937.4
- arrival_delay_scaled            1   6915.2 6941.2
- airline_status_silver           1   7284.6 7310.6
- type_of_travel_personal_travel  1   8666.3 8692.3
> 
```

fig12

MODEL	AIC	BIC
Logistic Regression - SelectedVar (Train)	6884.172	6966.632
Logistic Regression-Backward (Train)	6880.846	6977.049
Logistic Regression - SelectedVar (Validation)	1641.694	1706.997
Logistic Regression-Backward (Validation)	1644.356	1720.542
Logistic Regression - SelectedVar (Test)	1793.528	1859.252
Logistic Regression-Backward (Test)	1795.358	1872.036

fig 13

	Accuracy	Precision	Recall	F1
Train – Selected Variables	0.7681	0.8985	0.7199	0.7994
Train – Bwd	0.7690	0.8982	0.7211	0.8000
Validation – Selected Variables	0.7608	0.8677	0.7102	0.7811
Validation – Bwd	0.7608	0.8665	0.7107	0.7809
Test – Selected Variables	0.7538	0.9071	0.7065	0.7943
Test – Bwd	0.7538	0.9028	0.7079	0.7935

```
# Selected Variables                # Backward Selection
# > print(combined_conf_mat)        # > print(combined_conf_mat_bwd)

#      Data_Set      0      1       #      Data_Set          0      1
# 0 "Training"    "2181" "1281"     # 0 "Training-bwd"    "2189" "1273"
# 1 "Training"    "372"  "3293"     # 1 "Training-bwd"    "373"  "3292"

# 0 "Validation" "570"  "297"       # 0 "Validation-bwd" "571"  "296"
# 1 "Validation" "111"  "728"       # 1 "Validation-bwd" "112"  "727"

# 0 "Test"        "492"  "349"      # 0 "Test-bwd"        "496"  "345"
# 1 "Test"        "86"   "840"      # 1 "Test-bwd"        "90"   "836"
```

fig 14

```
> exp(coef(Train_mylogit_scaled))
                     (Intercept)             airline_status_gold
                      1.61191108                      2.23751057
         airline_status_platinum           airline_status_silver
                      1.49678920                      5.36120614
type_of_travel_mileage_tickets type_of_travel_personal_travel
                      0.64343573                      0.04736596
              day_of_flight_date6             day_of_flight_date15
                      0.60331332                      1.51743399
                      age_scaled              no_of_flights_scaled
                      0.92765534                      0.81790558
              flight_time_scaled             arrival_delay_scaled
                      0.93234719                      0.77641696
>
```

fig 15

```
logit(p) = 1.61 + 2.24*airline_status_gold + 1.50*airline_status_platinum + 5.36*airline_status_silver+
           0.64*type_of_travel_mileage_tickets + 0.04*type_of_travel_personal_travel+
           0.60*day_of_flight_date6 + 1.52*day_of_flight_date15 + 0.93*age_scaled+
           0.82*no_of_flights_scaled + 0.93*flight_time_scaled + 0.78*arrival_delay_scaled
```

fig 16

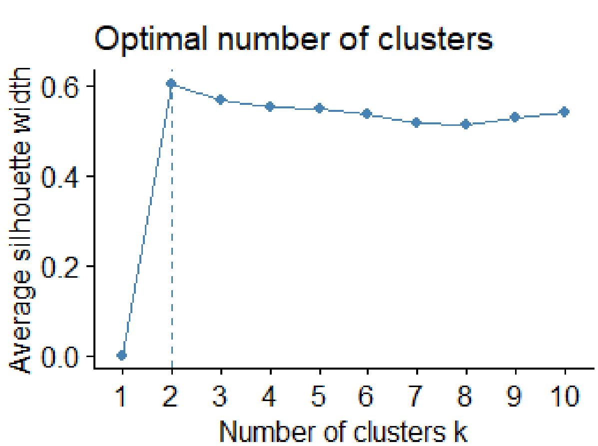


fig 16

fig 17

```
hcluster3
   1    2    3
1331 1794 2162
>
```

fig 18

```
  Group.1 X1.customer_id   X1.age X1.shopping_amount_at_airport X1.flight_time X1.arrival_delay X1.departure_delay
1       1       17696.13 43.02029                      26.67243       61.65440         11.21901           9.978588
2       2       34308.78 43.27369                      28.89688       82.10479         17.83779          14.029543
3       3       54964.70 42.60176                      29.17854       70.39223         16.66952          15.345745
  X1.flight_distance X1.satisfaction_binary X1.gender_binary X1.airline_binary X1.airline_status_num
1           408.1104                      1        0.4680691         1.0000000              2.065364
2           518.6132                      1        0.4765886         1.0000000              2.035117
3           444.9759                      1        0.5032377         0.5573543              2.110083
  X1.traveltype_num X1.class_num X1.month_of_flight_date_num
1          1.198347     1.980466                    6.049587
2          1.219621     1.987179                    5.872352
3          1.169750     1.967160                    5.790009
```

fig 19

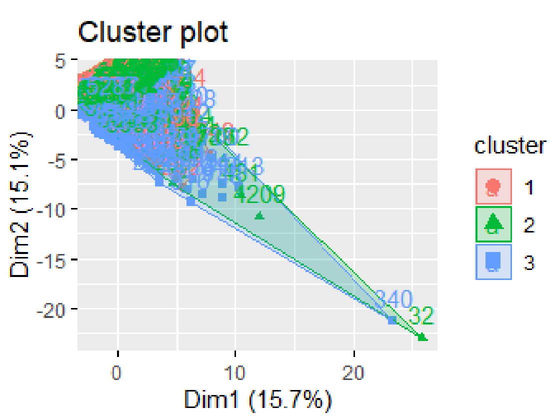


fig 20

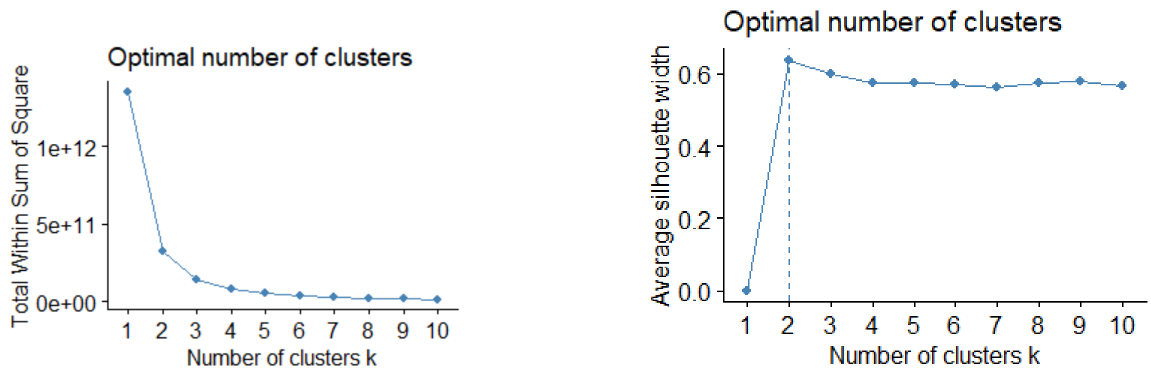


fig 21

```
> model_k3_dep<-aov(X1.departure_delay~cluster,data=kcluster3_center_data)
> summary(model_k3_dep)
            Df    Sum Sq   Mean Sq F value Pr(>F)
cluster      1 163346520 163346520   0.322  0.671
Residuals    1 506549499 506549499
> model_k3_status<-aov(X1.airline_status_num~cluster,data=kcluster3_center_data)
> summary(model_k3_status)
            Df   Sum Sq  Mean Sq F value Pr(>F)
cluster      1 0.001621 0.001621   230.3 0.0419 *
Residuals    1 0.000007 0.000007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> model_k3_class<-aov(X1.class.num~cluster,data=kcluster3_center_data)
> summary(model_k3_class)
            Df    Sum Sq   Mean Sq F value Pr(>F)
cluster      1 0.0002592 0.0002592   0.705  0.555
Residuals    1 0.0003678 0.0003678
```

fig 22

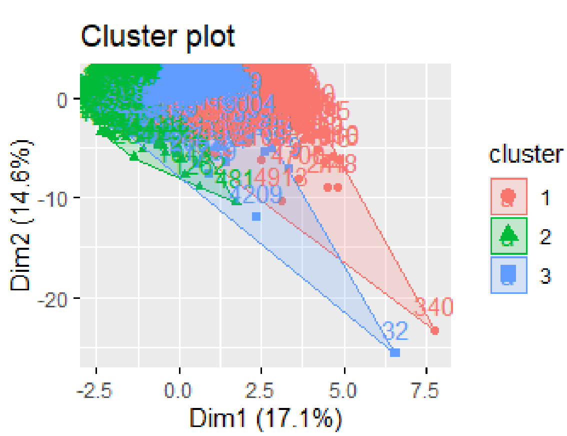


fig 23

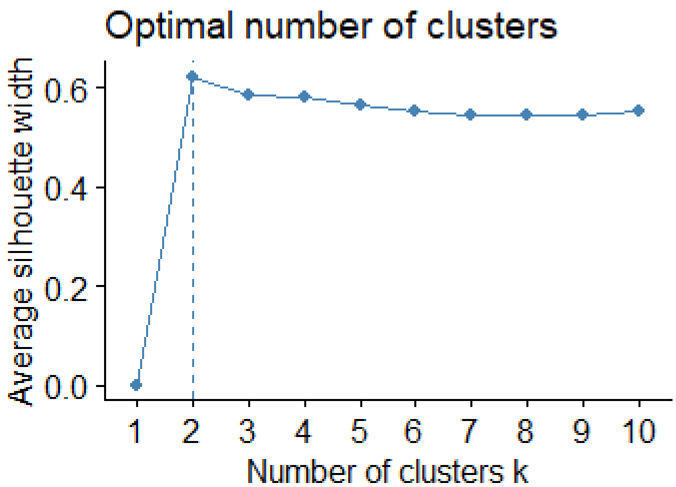


fig 24

fig 25

```
  Group.1 X0.customer_id   X0.age X0.shopping_amount_at_airport X0.flight_time X0.arrival_delay X0.departure_delay
1       1        24208.30 50.49404                      23.95596       74.96809          21.2683           18.21383
2       2        49463.06 49.23576                      26.37917       74.88527          25.1554           21.38880
  X0.flight_distance X0.satisfaction_binary X0.gender_binary X0.airline_binary X0.airline_status_num
1           482.1272                      0        0.6319149         1.0000000              1.445957
2           472.8853                      0        0.6267191         0.7493124              1.411002
  X0.traveltype_num X0.class_num X0.month_of_flight_date_num
1          2.185106     2.047234                    6.264681
2          2.171709     2.031041                    5.518664
```

fig 26

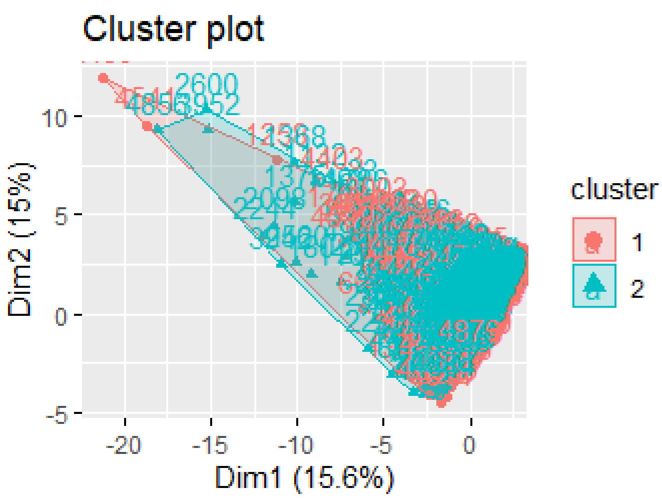


fig 27

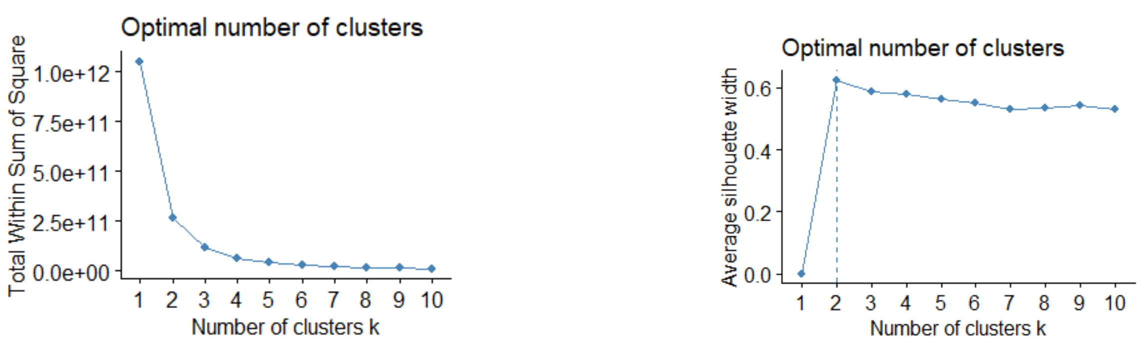
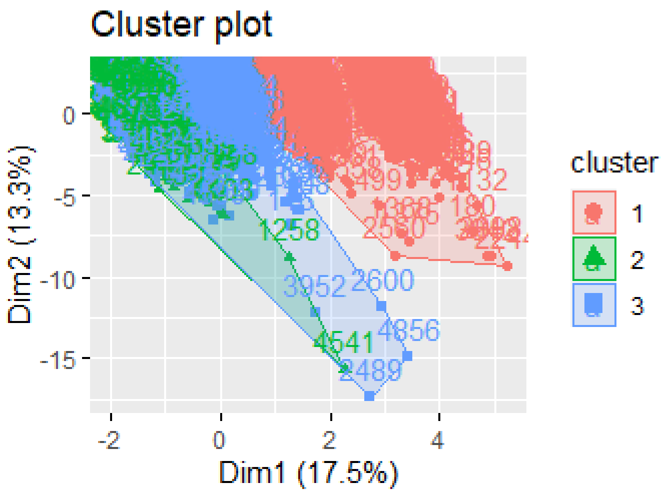


fig 28

```
> summary(model_k3_delay)
            Df    Sum Sq   Mean Sq F value Pr(>F)
cluster      1 143911012 143911012   0.332  0.667
Residuals    1 432887198 432887198
> summary(model_k3_status)
            Df   Sum Sq  Mean Sq F value Pr(>F)
cluster      1 0.001621 0.001621   230.3 0.0419 *
Residuals    1 0.000007 0.000007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> summary(model_k3_class)
            Df    Sum Sq   Mean Sq F value Pr(>F)
cluster      1 0.0002592 0.0002592   0.705  0.555
Residuals    1 0.0003678 0.0003678
```

fig 29


Cluster plot
Dim2 (13.3%)
Dim1 (17.5%)
cluster
1
2
3
1258
3952
2600
4541
4856
2489
2560
1368
180
499
132
0
-5
-10
-15
-2
0
2
4