MSCI 562 – Intelligent Data Analysis and Visualisation

Individual Coursework - 1

Submitted by,
36342582

Table of Contents

Executive Summary:

The Report highlights the data understanding phase of CRISP-DM where the initial analysis is carried out to understand the structure of the data to tackle the airline problem. From the collected survey data, it could be observed that 57% of passengers gave neutral/dissatisfied response. To identify the factors, initially visual inspection was undertaken, where some of the potential correlations between the age and customertype was revealed, but on further examination these variable seems to be insignificant.

In the next Stage, Information Value & WoE concept was applied to identify the predictor variable which contributes to the target. OnlineBoarding, Inflightwifiservice and Inflight entertainment were identified as key factors. Relationship among the predictor variable were explored using the correlation matrix from which strong association between Departure delay and Arrivaldelay was identified. TypeofTravel and Class variable also had a moderate relationship among them.

In the final stage, MDS technique was done to identify the most common characteristic from which the TypeofTravel, Class,OnlineBoarding,SeatComfort features were identified.

Introduction:

The purpose of this data analysis report is to examine the level of customer satisfaction within a UK- based airline by identifying factors that drives customer satisfaction and recommending strategies to enhance service quality. The airline conducted survey during flights which captures a range of information on pre-boarding and onboarding services, in addition to personal and flight-related details. It is to be noted that most of the captured data are categorical variables in the given dataset.

The initial section of report discusses about the preliminary analysis of the given data to identify any data quality issues, evaluate current service quality, and explore relationship between the features. Then Weight of Evidence and Information Value concept is utilised to uncover the predictive power of individual variable on Customer Satisfaction. Furthermore, the next section examines the relationship among predictor variables and investigates the similarities among predictors using correlation matrix to unveil any underlying pattern in data. At the end, Multidimensional Scaling - dimensionality reduction technique is employed to identify the common characteristic among the passengers.

The report concludes by summarizing the important findings of the given dataset and suggest ways to improve their airline service based on the data analysis.

1.Data Investigation & Relationship:

The given dataset consists of 24 features categorised as personal details, flight details, pre-boarding and onboarding services with the target variable being satisfaction. Before diving into analysis, the data quality is checked based on outliers, inconsistent values, and missing values. Missing values in the Arrival Delay in Minutes, Pre-boarding, and Onboarding service (denoted by 0 & NA) are omitted (on basis of small proportion of dataset) as they have negligible effect on analysis. In the Figure 1.1, points concentrated outside the whisker (outliers) are considered for analysis as they might hold important information.

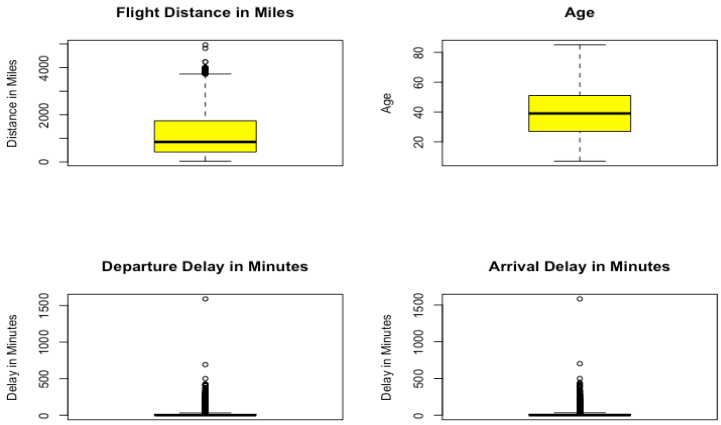


Figure 1.1 Box Plots of Numerical Variables

On initial analysis, it could be found that the overall 57% of the customer gave neutral or dissatisfied response as per the Figure 1.2 below. There could be numbers of factors influencing the response which would be discussed later.

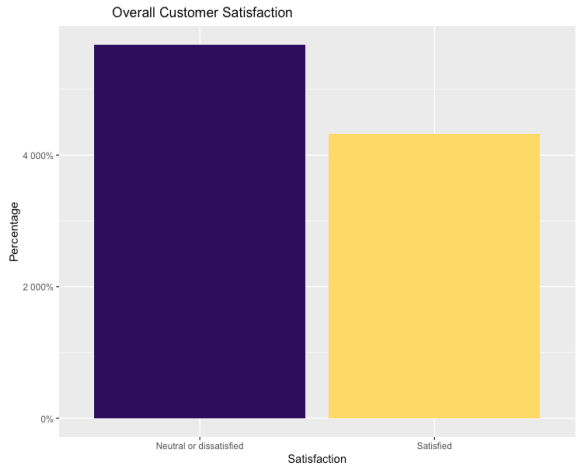


Figure 1.2 Overall Customer Satisfaction.

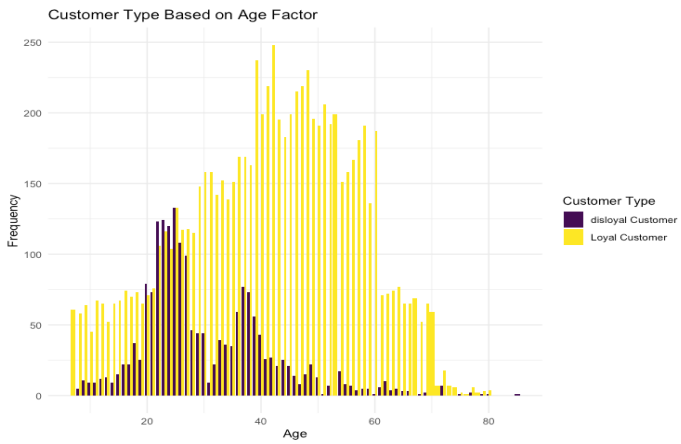


Figure 1.3 Customer Type Based on Age

The data consist of range of customers who are categorised as loyal and disloyal. It can be noted from Figure 1.3, loyal customers are mostly middle aged (35-60) while disloyal customers are mostly younger (20-30). From the perspective of customer satisfaction (Refer Figure 1.4), most of the middle-aged passengers are satisfied while younger passengers are neutral or dissatisfied with flight service. This shows that many disloyal customers had neutral/negative experience, while some loyal customers also had issues with service.

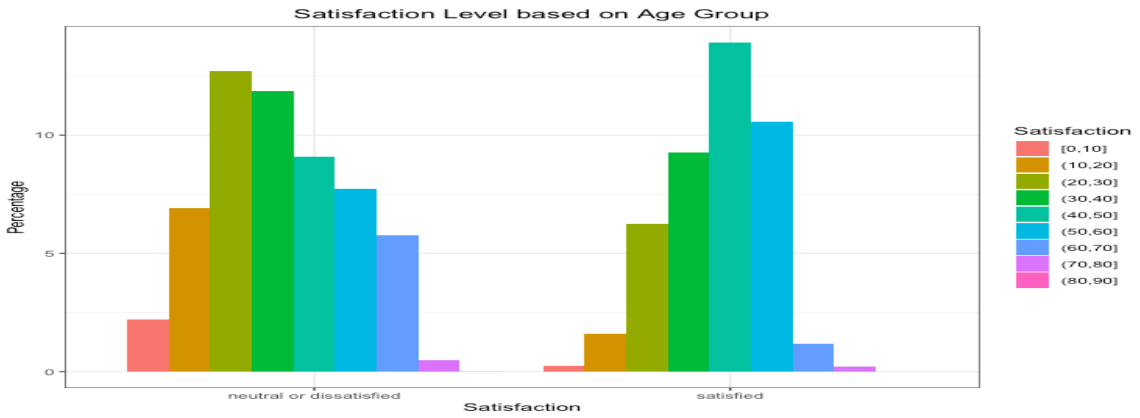


Figure 1.4 Satisfaction Level based on Age Group

One of the basic assumptions is that flight delays lead to negative response. Figure 1.5, shows that 60% of passengers give neutral/negative response when there is delay, validating the assumption. From this we can infer that, delay in Flights may contribute for negative response.

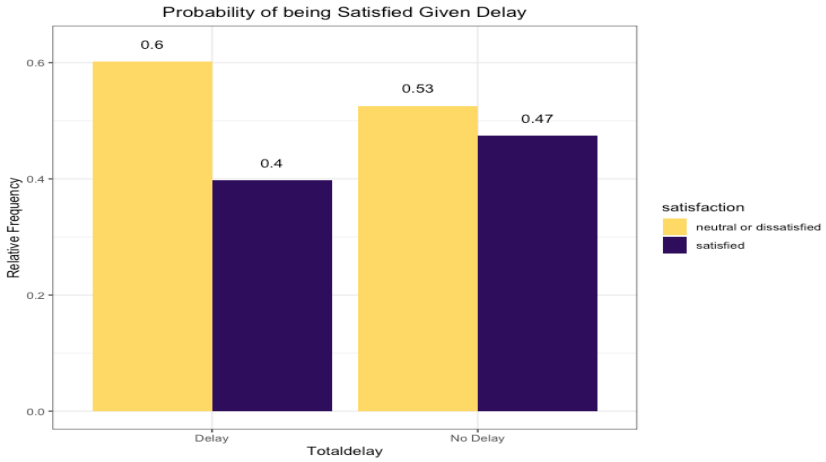
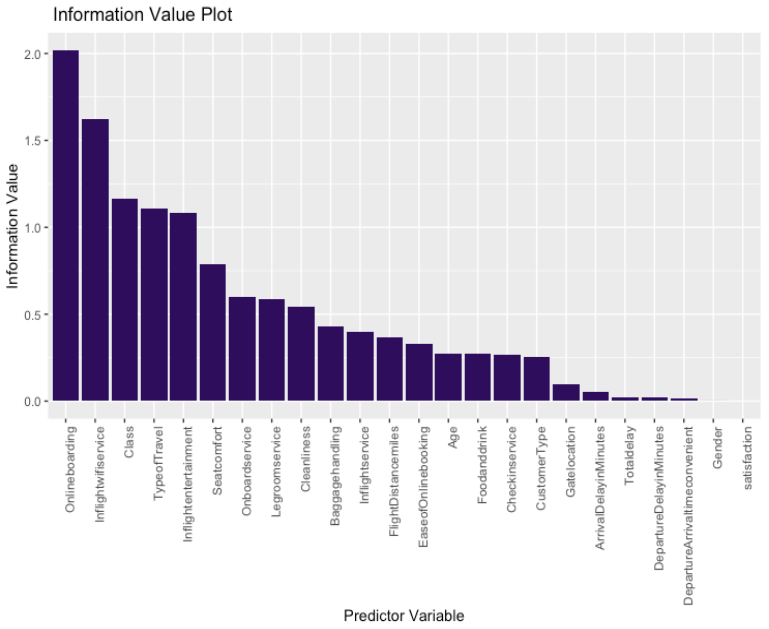


Figure 1.5 P(Satisfaction|Delay)



ummary

Variable	IV
Onlineboarding	2.0185356990
Inflightwifiservice	1.6235660422
Class	1.1638013571
TypeofTravel	1.1079665953
Inflightentertainment	1.0851876124
Seatcomfort	0.7874318924
Onboardservice	0.6015419638
Legroomservice	0.5895404147
Cleanliness	0.5446843320
Baggagehandling	0.4288034890
Inflightservice	0.3991621490
FlightDistancemiles	0.3685550798
EaseofOnlinebooking	0.3271910081
Age	0.2748109685
Foodanddrink	0.2740085276
Checkinservice	0.2683853507
CustomerType	0.2509723963
Gatelocation	0.0980386425
ArrivalDelayinMinutes	0.0521272744
Totaldelay	0.0239874035
DepartureDelayinMinutes	0.0211219973
DepartureArrivaltimeconvenient	0.0159760177
Gender	0.0004477391

Figure 1.6 IV Values & Plot

On a separate note, Weights of Evidence & Information Value method is utilised to determine the independent contribution of each feature to customer satisfaction variable.Generally, the IV value implies the variable informativeness. Thus, from above figure 1.6, we can infer that features – OnlineBoarding, InflightWifi Service, ClassType, Typeoftravel, InflightEntertainment are the important factors for customer Satisfaction while seatcomfort, onboardservice, Legroomservice have moderate impact. Meanwhile, customertype, gatelocation, food&drink are only marginally informative. Figure 1.7 analyses the WoE of ClassType & TypeofTravel

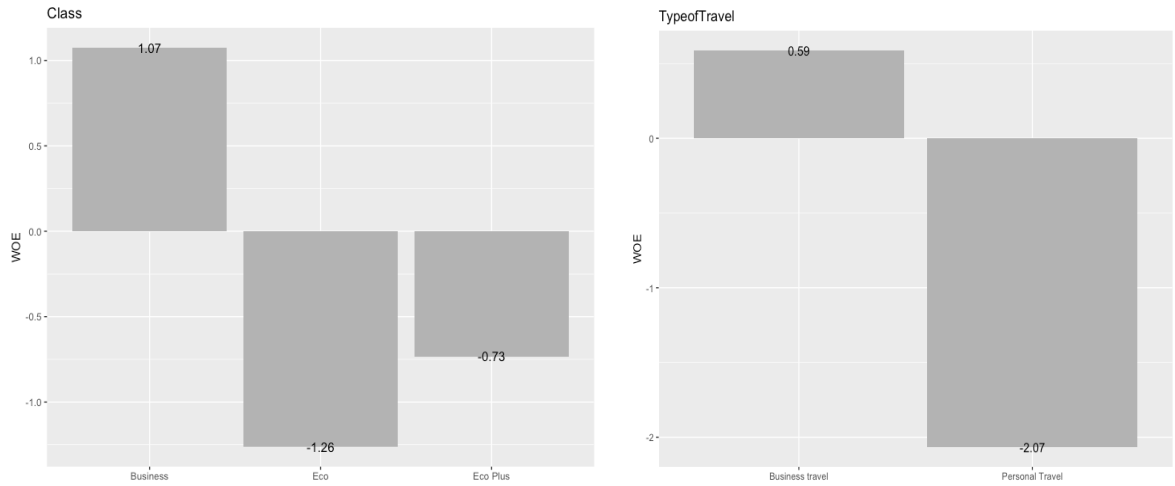


Figure 1.7 WOE Plot -Class & TravelType

It can be inferred that passengers travelling in Business class & for business purpose tend to be satisfied while people travelling for personal reasons and in Eco & Eco Plus tend to be dissatisfied or neutral. From this it can be inferred that Class & Type of Travel are significant factor for customer satisfaction. On the other hand, features like OnlineBoarding, InflightWifiService, are significant based on the Figure 1.8,1.9,1,10 below,

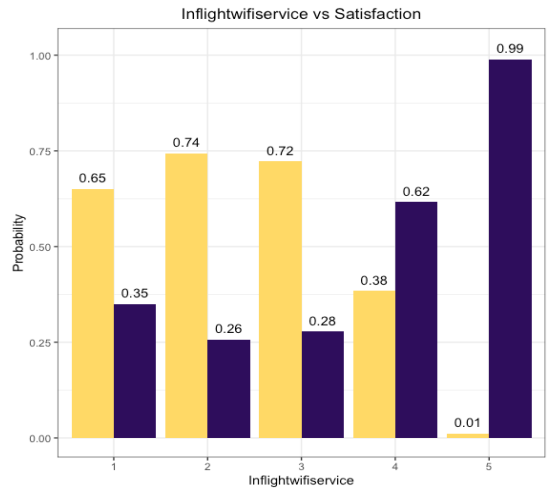


Figure 1.8 Probability Plot of InflightWifiService

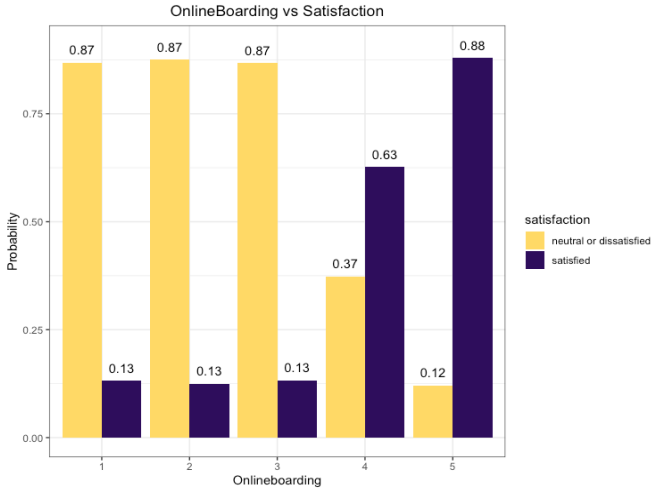


Figure 1.9 Probability Plot for Online Boarding

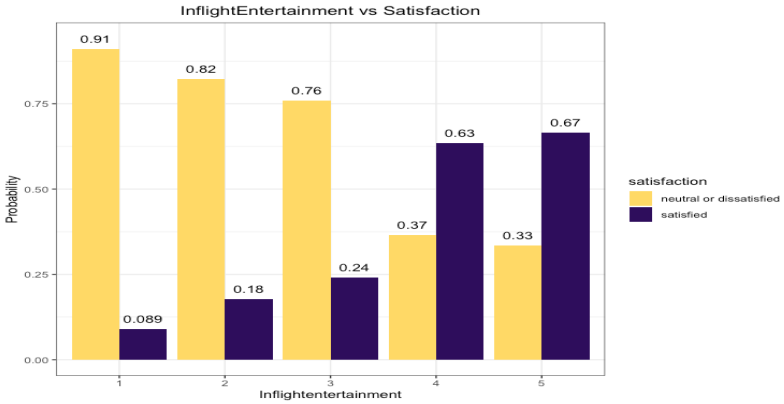


Figure 1.10 probability plot for InflightEntertainment

The satisfaction factor varies widely among customers for important variables like OnlineBoarding, Inflightwifiservice, InflightEntertainment, as illustrated in the figures 1.8,1.9,1.10. Thus, this indicates that these variables influence the customer satisfaction significantly. However, the features like Gatelocation, Checkinservice survey response does not show significant variation among the customers, as passengers have mostly equal opinion on the service which makes it difficult to analyse the target. Refer the Figure 1.11, 1.12 for validation of above statement.

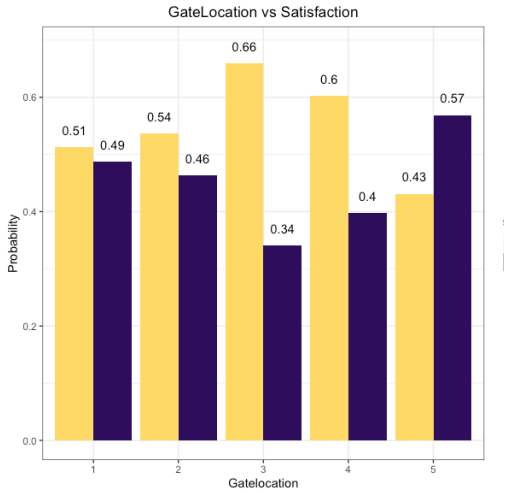


Figure 1.11 GateLocation vs Satisfaction.

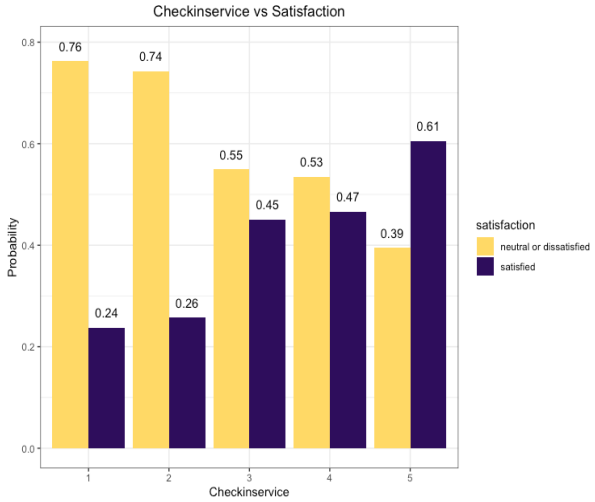


Figure 1.12 Checkinservice vs Satisfaction

From the Information Value & WoE concept we could determine few features as significant for the target variable but that doesn't mean that these results are conclusive. There are few variables when combined with other has the strong influence on target. These relations among predictor variable are analysed in the next section.

2. Correlation of Variables:

At first, to determine the relationship of variables lets analyse the spread function on the data shown in the Figure 2.1. it can be noted that variables TypeofTravel, Class, FlightDistancemiles seems to be correlated and it can be inferred from the graph that most of the personal type passengers opt this airline for short distance travel while long distance travel is adopted by business travellers. Combining the analysis seen before, it can be said that business class services are satisfactory as passengers choose this airline for long business travel. There seems to be linear relationship between DepartureDelayinMinutes and ArrivalDelayinMinutes and

there is a chance that above variables might convey similar information. The variable Gender, CustomerType doesn't have any strong relationship among the other variables in the data.

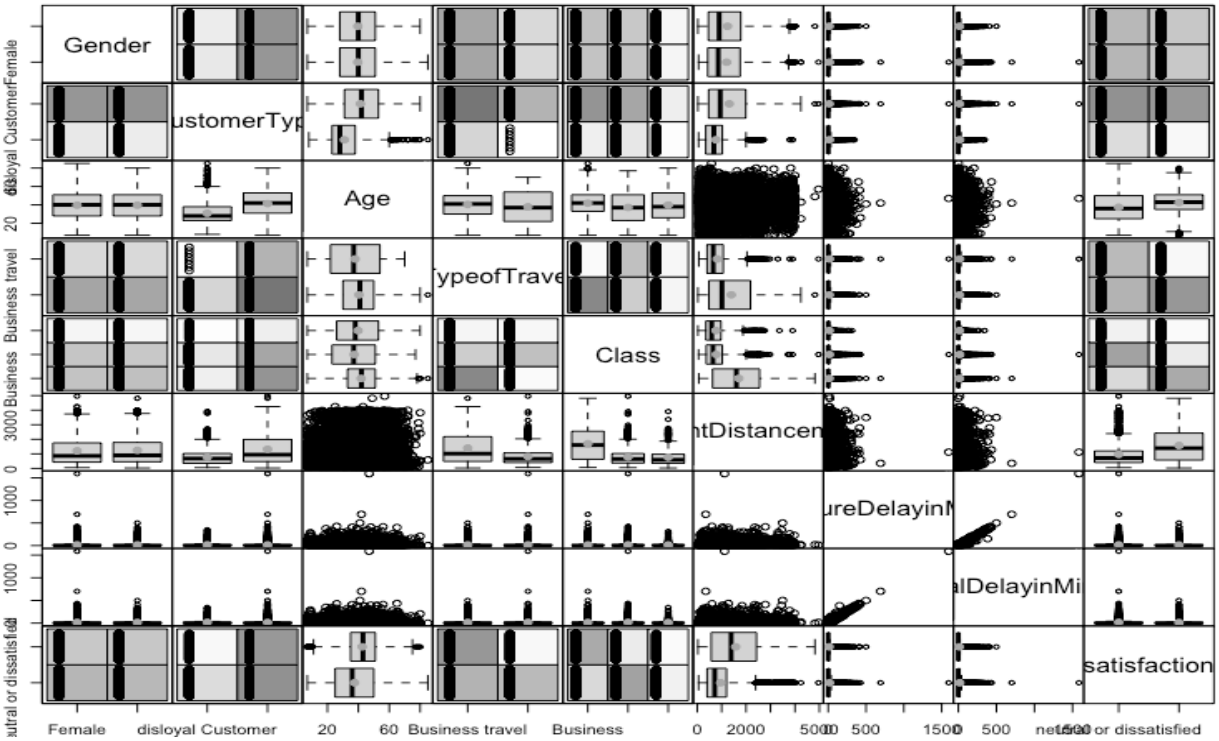


Figure 2.1 Spread Function on Dataset

The findings mentioned above is inferred visually, to validate that association function is utilised to determine the correlation matrix of the variables. The below figure 2.2 shows correlation matrix.

	Gender	CustomerType	Age	TypeofTravel	Class	FlightDistancemiles
Gender	1.000000000	0.02910214	0.001082110	0.00000000	0.01760785	0.018163304
CustomerType	0.029102144	1.00000000	0.258211238	0.28391889	0.12508723	0.202262497
Age	0.001082110	0.25821124	1.000000000	0.07676857	0.14636816	0.084117601
TypeofTravel	0.000000000	0.28391889	0.076768574	1.00000000	0.55458355	0.262921813
Class	0.017607854	0.12508723	0.146368162	0.55458355	1.00000000	0.461070439
FlightDistancemiles	0.018163304	0.20226250	0.084117601	0.26292181	0.46107044	1.000000000

Figure 2.2 Correlation Matrix - Class, FlightDistance, Travel Type

As inferred earlier, type of travel and class are moderately correlated while FlightDistancemiles is moderately correlated with class but weakly correlated with the TypeofTravel. Analysing other variables in the data using the correlation plot shown in Figure 2.3, there are three clusters observed where two clusters are moderately correlated and the other one has high association. The first cluster consist of TypeofTravel, Class, FlightDistancemiles. The second cluster has Inflightwifiservice, Departurearrivaltimeconvenient, EaseofOnlinebooking, Gatelocation but these variables are weakly associated. The final cluster is strongly associated which can be numerically verified in the Figure 2.4. If DepartureDelay and ArrivalDelay gives the similar information, then these variables are excluded in model to avoid multicollinearity.

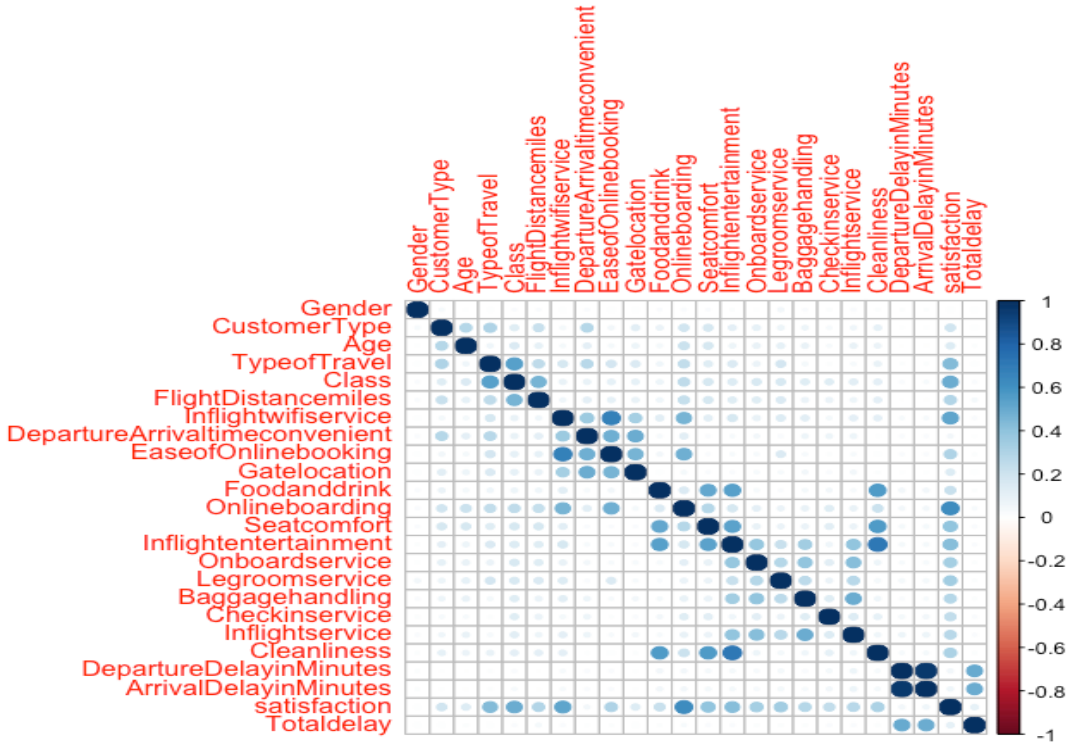


Figure 2.3 Correlation Plot

	DepartureDelayinMinutes	ArrivalDelayinMinutes
Class	−0.006968106	−0.013380453
FlightDistancemiles	0.011802031	0.014157850
Inflightwifiservice	0.030794952	0.038103764
DepartureArrivaltimeconvenient	−0.004047768	−0.007707911
EaseofOnlinebooking	0.023651324	0.025638739
Gatelocation	0.019766510	0.016050116
Foodanddrink	0.020589522	0.022033138
Onlineboarding	0.010044392	0.012120792
Seatcomfort	0.023871443	0.028889893
Inflightentertainment	0.044057225	0.048114323
Onboardservice	0.047923420	0.052379169
Legroomservice	0.047588722	0.048625255
Baggagehandling	0.030457281	0.035295950
Checkinservice	0.045625870	0.048876777
Inflightservice	0.028811439	0.031191087
Cleanliness	0.047251355	0.054261273
DepartureDelayinMinutes	0.073884468	0.080693033
ArrivalDelayinMinutes	0.037485331	0.042704896
satisfaction	1.000000000	0.966786400
	0.966786400	1.000000000
	0.055093632	0.062364609

Figure 2.4 Correlation between Departure & Arrival Delay

One of obvious relationship that can be seen (in Figure 2.3) is between FoodandDrink, InflightEntertainment, Cleanliness. There is one more variable that correlates with cleanliness – Seat Comfort. This validates the basic assumption that basic necessities (Foodanddrink, SeatComfort) are rated according to the cleanliness maintained inside the flight. It can be seen from Figure 2.3 that Age, CustomerType, Checkinservice doesn't influence any other variables this finding coincides with the Information value findings observed in the previous section.

3. Dimensionality Reduction Technique:

In the earlier part of the report, we applied method like IV & WoE to find the significant predictor variables for customer satisfaction and correlation to find relationship among variables. This section focuses primarily on finding the common characteristic among the passengers. Since in the given dataset most of significant variables are categorical, Principal Component Analysis (PCA) cannot be utilised as it requires numerical variables for deeper analysis. Thus, Multidimensional Scaling (MDS) is utilised in the given dataset to identify the common characteristic.

To work with MDS, dissimilarity matrix is calculated using Gower's coefficient. This coefficient is appropriate for this data as these contains categorical variables. Then number of dimensions for this data is identified using the stress plot displayed in Figure 3.1.

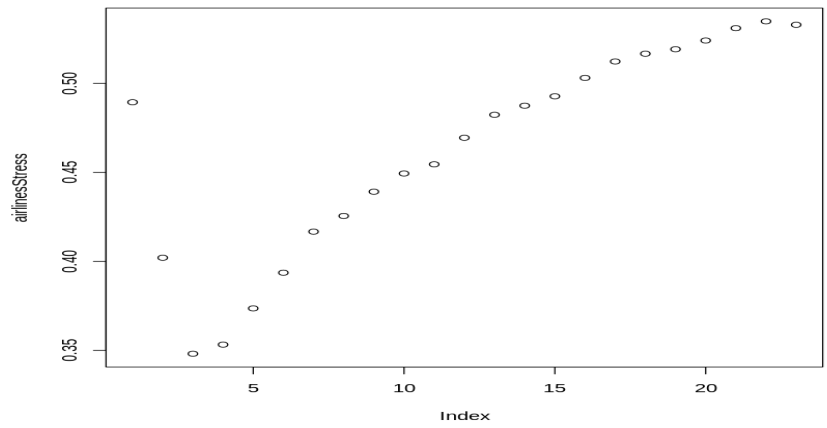


Figure 3.1 Stress Plot

From the figure above, it could be seen that the dimension 3 has the lowest stress metric and thus optimal number of dimensions is 3 for the MDS. With the obtained dimension, MDS is produced on the given dataset.

Figure 3.2 3D plot of MDS

From the Output produced in Figure 3.2, it can be observed that two clusters of satisfied and neutral/dissatisfied are formed. Thus, these distinct groupe of satisfaction shows that

characteristic of passengers is well separated by dissimilarities. Observing the plot from D1 perspective, most of satisfied falls in negative scale while purple lie n positive scale.

To identify the most significant characteristic of the passengers, association matrix is determined from the MDS. Based on the output in Figure 3.3, it is noticed that TypeofTravel,

	Age	TypeofTravel	Class	FlightDistancemiles	Inflightwifiservice
D1	-0.174217010	0.50847241	0.64929309	-0.376622365	0.489762754
D2	-0.028767402	0.65838990	0.50914015	-0.229267084	0.417343490
D3	0.007517874	0.22215671	0.17343779	-0.084139633	0.495024581

	Onlineboarding	Seatcomfort	Inflightentertainment	Onboardservice
D1	0.68036772	0.620743210	0.70330763	0.48934791
D2	0.16920287	0.235783094	0.29709584	0.15976343
D3	0.17978802	0.407703637	0.41854479	0.08152273

Figure 3.3 Association Matrix of MDS

Class, OnlineBoarding, SeatComfort, InflightEntertainment seems to hold high correlation with the Dimension data while Age and FlightDistancemiles give a weak negative correlation. From this it can be concluded that most common characteristic among the passengers depend on the Features - Class, OnlineBoarding, SeatComfort, TypeofTravel.

Conclusion:

On initial analysis, some underlying pattern related to age & CustomerType was identified visually but on the later stage of analysis it was known that age & CustomerType doesn't have a significant impact on the customer satisfaction. On the whole, it can be noticed that percentage of neutral/negative response seems to be slightly higher with respect to Quality of service.

In next phase of analysis, relationship among the predictor variables is identified where DepartureDelay and ArrivalDelay had high association. It was also noted that based on the Traveldistance, passengers chose the TravelType and Class in this airline. Most of the people opt this airline for short distance travel. From the Information value & WoE it can be seen that OnlineBoarding, Inflightwifiservice, InflightEntertainment contribute to customer satisfaction.

To determine the most common characteristics among the passengers, MDS technique was applied. The stress plot was utilised to obtain the dimensions for MDS after which the most significant features influencing the customer satisfaction was identified as TypeofTravel, Class, OnlineBoarding, SeatComfort. The other features in the dataset does impact customer satisfaction in a minimal way.

Thus, to improve the overall service quality of airline, it is recommended that airline invest in Digital to improve their online services and implement some entertainment during travel which is useful for youngsters. If the seatcomfort issue is taken care of by the airline, then there would be scope for new customers.