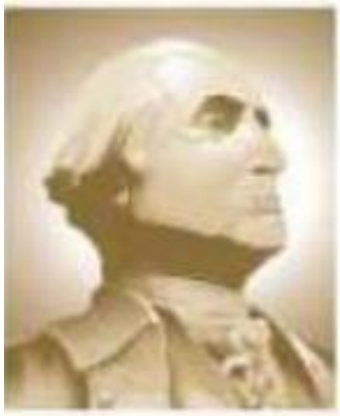


THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

Visualization of Complex Data

DATS 6401

Final Term Project (FTP)

Columbian college of arts and sciences

Instructor: Reza Jafari

Date: 4/26/2024

Author: Sanika Narayanpethkar

Gwid: G40709114

ABSTRACT

For my project, I investigated “Airline Passenger Satisfaction” using a dataset of 103,904 reviews sourced from Kaggle. The study delved into passenger’s demographic details, including age, gender, travel purpose, and class, along with factors like arrival/departure delay, flight distance, and service ratings (ranging from 0 to 5). I aimed to assess passenger satisfaction comprehensively, evaluating aspects like in-flight Wi-Fi, check-in service, cleanliness, seat comfort, food, and more. My analysis utilized diverse visualizations and statistical methods. The goal was to visualize the proposed enhancements, striving to enhance overall passenger experiences, ensuring hassle-free and memorable journeys, and making every investment worthwhile.

INTRODUCTION

My motivation for this project was driven by my experience as an international student grappling with the uncertainties of air travel. Intrigued by data showing that 52% of US air travelers are dissatisfied, I saw an opportunity to delve into this issue given its potential profitability. Through visual analysis, I sought to understand the factors influencing customer satisfaction in the airline industry, which is not only a major global transportation sector but also intensely competitive.

Through the visualizations, I observed trends and patterns that highlighted the importance of service quality in the airline industry. These insights are critical as airlines strive to differentiate themselves from competitors and enhance passenger experiences. My analysis, presented through various graphs and charts, showed how different service elements correlate with overall customer satisfaction.

In the data visualizations, I also explored the competitive landscape of the airline industry, which is growing rapidly and is characterized by aggressive marketing strategies such as price wars and loyalty programs. This shift towards consumer-oriented strategies reflects the dynamic nature of the market, which our visualizations helped to clarify.

DESCRIPTION OF DATASET

For the "Airline Passenger Satisfaction" project, I utilized a detailed dataset sourced from Kaggle, which includes responses from 103,904 airline customers. This dataset is particularly valuable as it covers a wide range of attributes relevant to airline service experiences. Attributes include demographic information such as gender, alongside operational aspects like travel type (business or leisure), class of service (Economy, Business, or First Class), flight distance, and specific service ratings including check-in process and overall hospitality.

Description of the Dataset:

Dependent Variable: The primary dependent variable in this analysis is satisfaction, which is categorized as either satisfied or dissatisfied. This binary variable serves as a measure of overall customer satisfaction based on various service attributes.

Independent Variables: The independent variables include:

Gender: Categorical variable indicating the passenger's gender.

Customer Type: Indicates whether the customer is a returning or first-time flyer.

Age: Continuous variable representing the passenger's age.

Type of Travel: Categorical variable distinguishing between personal and business travel.

Class: The service class chosen by the passenger.

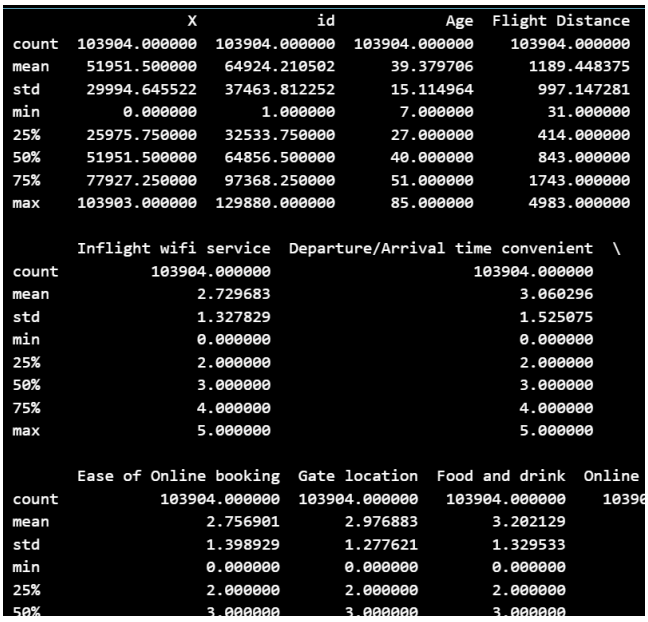
Flight Distance: Continuous variable showing the length of the flight in miles.

Service quality metrics such as Inflight Wi-Fi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Check-in service, Inflight service, and Cleanliness. Each of these is rated on an ordinal scale.

The importance of this dataset in the industry is profound, as it provides actionable insights that can directly influence policy making, marketing strategies, and operational adjustments, all aimed at enhancing passenger satisfaction in a highly competitive market. By focusing on the variables that significantly impact satisfaction, airlines can strategically allocate resources to areas that will maximize customer approval and retention.

PRE-PROCESSING DATASET

In the "Airline Passenger Satisfaction" project, my dataset preprocessing was aimed at enhancing the quality of our data for more accurate analysis and insights. Initially, I conducted a comprehensive assessment of the dataset to understand its structure and content, utilizing df.describe() to obtain a summary of the statistics and df.isnull().sum() to identify missing values.



	X	id	Age	Flight Distance
count	103904.000000	103904.000000	103904.000000	103904.000000
mean	51951.500000	64924.210502	39.379706	1189.448375
std	29994.645522	37463.812252	15.114964	997.147281
min	0.000000	1.000000	7.000000	31.000000
25%	25975.750000	32533.750000	27.000000	414.000000
50%	51951.500000	64856.500000	40.000000	843.000000
75%	77927.250000	97368.250000	51.000000	1743.000000
max	103903.000000	129880.000000	85.000000	4983.000000

	Inflight wifi service	Departure/Arrival time convenient	\
count	103904.000000	103904.000000	
mean	2.729683	3.060296	
std	1.327829	1.525075	
min	0.000000	0.000000	
25%	2.000000	2.000000	
50%	3.000000	3.000000	
75%	4.000000	4.000000	
max	5.000000	5.000000	

	Ease of Online booking	Gate location	Food and drink	Online
count	103904.000000	103904.000000	103904.000000	10390
mean	2.756901	2.976883	3.202129	
std	1.398929	1.277621	1.329533	
min	0.000000	0.000000	0.000000	
25%	2.000000	2.000000	2.000000	
50%	3.000000	3.000000	3.000000	

Fig 1

```
X                                     0
id                                    0
Gender                                0
Customer Type                         0
Age                                   0
Type of Travel                        0
Class                                 0
Flight Distance                       0
Inflight wifi service                 0
Departure/Arrival time convenient     0
Ease of Online booking                0
Gate location                         0
Food and drink                        0
Online boarding                       0
Seat comfort                          0
Inflight entertainment                0
On-board service                      0
Leg room service                      0
Baggage handling                      0
Checkin service                       0
Inflight service                      0
Cleanliness                           0
Departure Delay in Minutes            0
Arrival Delay in Minutes            310
satisfaction                          0
...
Departure Delay in Minutes          0
```

Fig 2

DATA CLEANING PROCESS:

Handling Missing Values:

I removed all rows with missing values using df.dropna(inplace=True), ensuring that my dataset only contained complete records. This was crucial as missing data could skew our analysis and lead to inaccurate conclusions about passenger satisfaction.

```
X                                  0
id                                 0
Gender                             0
Customer Type                      0
Age                                0
Type of Travel                     0
Class                              0
Flight Distance                    0
Inflight wifi service              0
Departure/Arrival time convenient  0
Ease of Online booking             0
Gate location                      0
Food and drink                     0
Online boarding                    0
Seat comfort                       0
Inflight entertainment             0
On-board service                   0
Leg room service                   0
Baggage handling                   0
Checkin service                    0
Inflight service                   0
Cleanliness                        0
Departure Delay in Minutes         0
Arrival Delay in Minutes           0
satisfaction                       0
dtype: int64
```

Fig 3

Outlier Detection and Removal:

Outliers can often distort the results of a dataset, especially in statistical analyses. We used the Interquartile Range (IQR) method to detect and remove outliers from key numerical columns such as 'Flight Distance', 'Departure Delay in Minutes', and 'Arrival Delay in Minutes'. For each of these columns, we calculated the first and third quartiles (Q1, Q3) and the IQR. Data points lying more than 1.5 times the IQR from the Q1 and Q3 were considered outliers and were excluded from further analysis.

1) Flight distance:

```
Q1 and Q3 of the Flight Distance is 414.00  & 1743.00 .
IQR for the Flight Distance is 1329.00 .
Any Flight Distance < -1579.50  and Flight Distance > 3736.50  is an outlier.
```

Fig 4

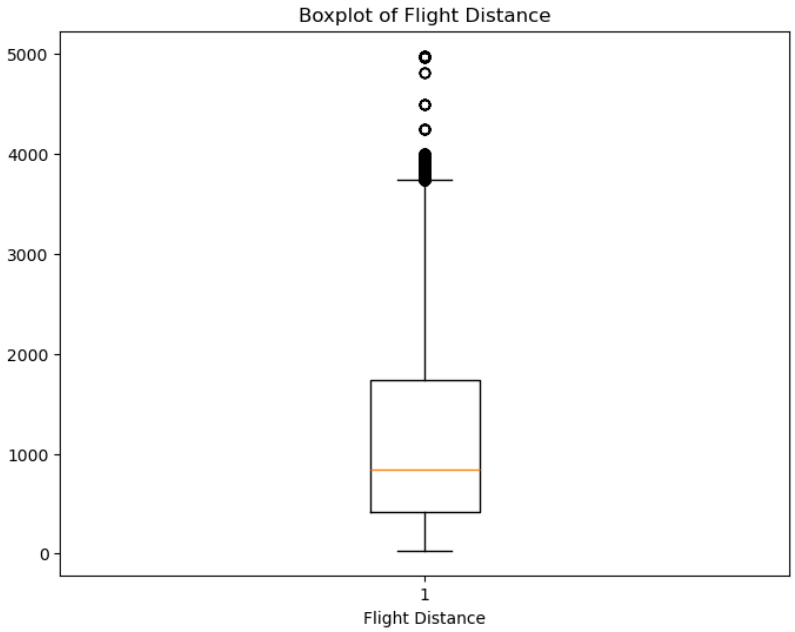


Fig 5

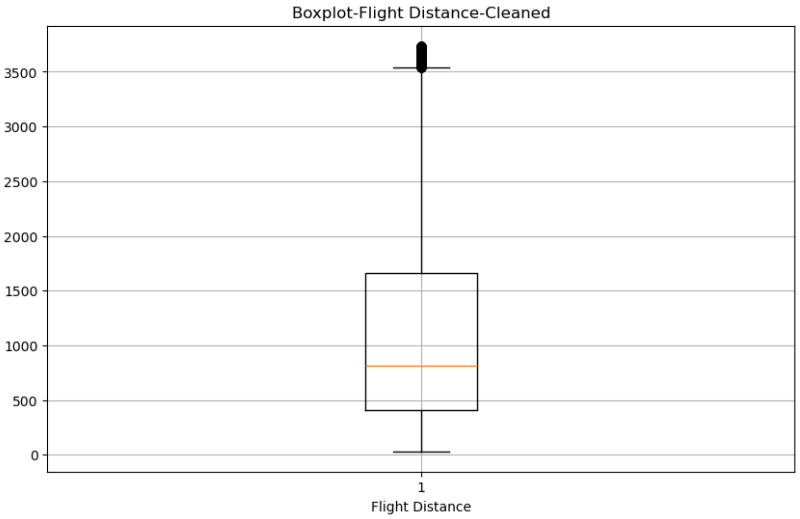


Fig 6

The initial analysis showed a wide range of flight distances, from 31 to 4983 kilometers. The presence of outliers was significant, especially in flights longer than is typical for much of the data.

Post-removal, the data showed a more consistent range that aligns closely with typical commercial flight distances, improving the accuracy of any statistical or machine learning analysis aimed at understanding factors influencing passenger satisfaction related to flight duration.

2) Departure delay in minutes:

```
Q1 and Q3 of the Departure Delay is 0.00  & 12.00 .
IQR for the Departure Delay is 12.00 .
Any Departure_Delay < -18.00  and Departure_Delay > 30.00  is an outlier.
```

Fig 7

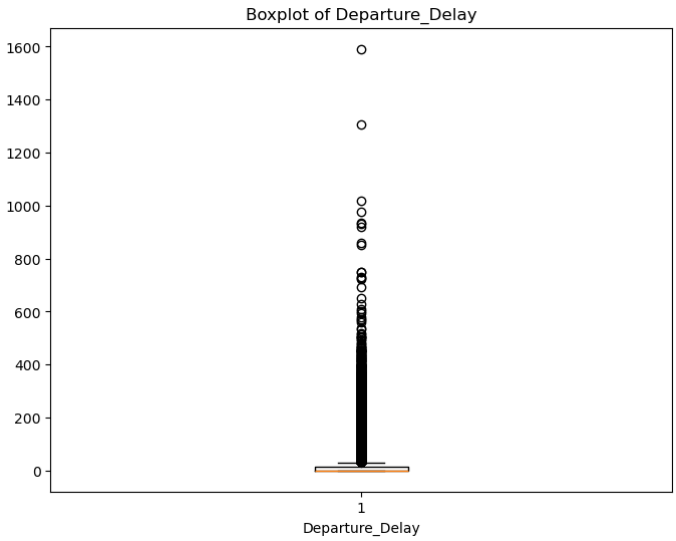


Fig 8

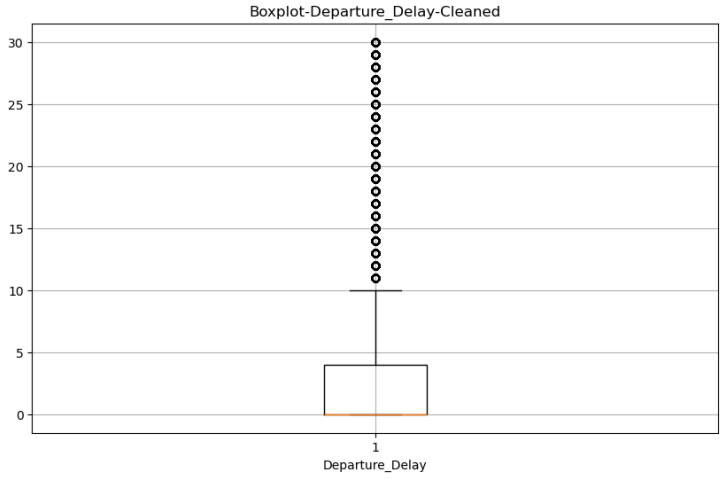


Fig 9

Delays at departure had a broad range, with some extreme outliers that could potentially represent unusual circumstances such as technical issues or severe weather conditions. Removing these outliers helped in focusing the analysis on more common scenarios, which are more indicative of the typical airline operational efficiencies and passenger experiences.

3) Arrival delay in minutes:

```
Q1 and Q3 of the Arrival_Delay is 0.00  & 12.00 .
IQR for the Arrival_Delay is 12.00 .
Any Arrival_Delay < -19.50  and Arrival_Delay > 32.50  is an outlier.
```

Fig 10

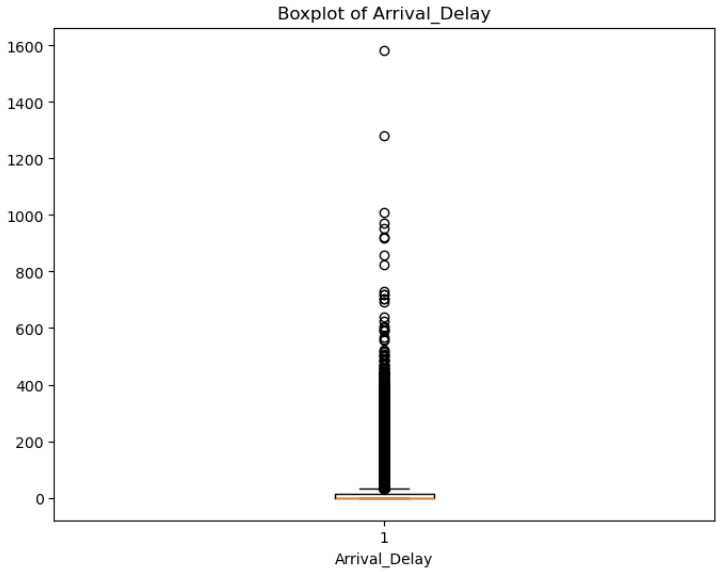


Fig 11

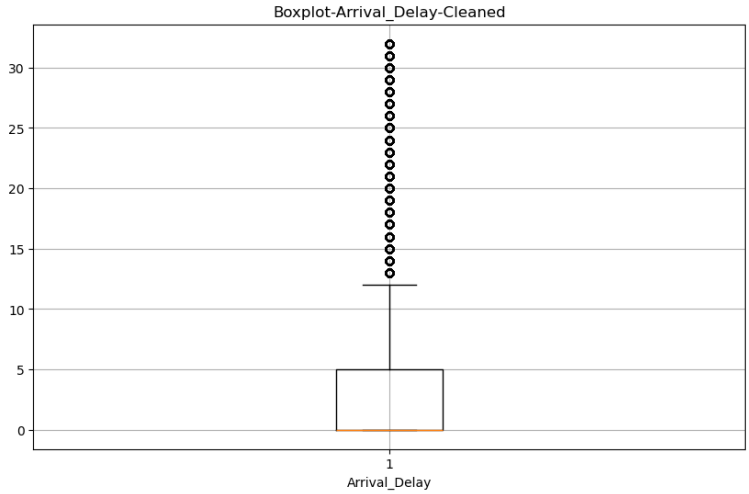


Fig 12

Like departure delays, arrival delays also had outliers that were markedly higher than the norm. These outliers can distort average delay times and misrepresent the typical passenger experience. Cleaning these outliers provided a dataset that better reflects the usual operational performance of airlines and allows for a more accurate assessment of factors that regularly impact passenger satisfaction regarding timeliness.

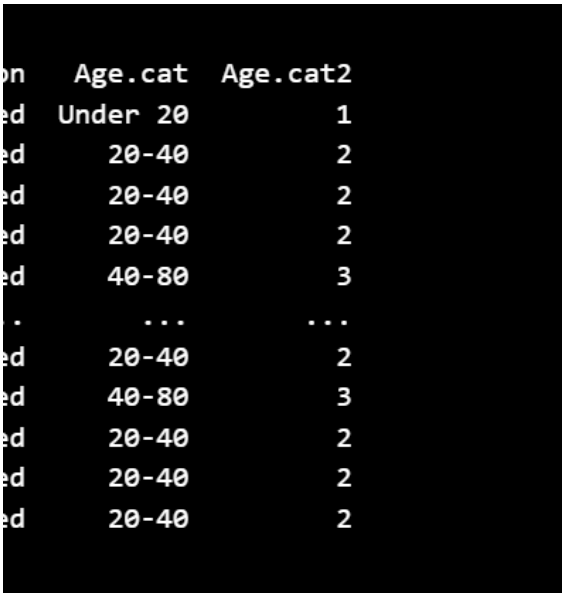
DATA TRANSFORMATION FOR ANALYSIS:

Infinitesimal Values:

We replaced infinite values with NaN (np.inf, -np.inf) to standardize the dataset format.

Categorization of Age:

We added a new categorical feature, 'Age.cat', to segment passengers into age groups ('Under 20', '20-40', '40-80', 'above 80'), facilitating age-related analysis.



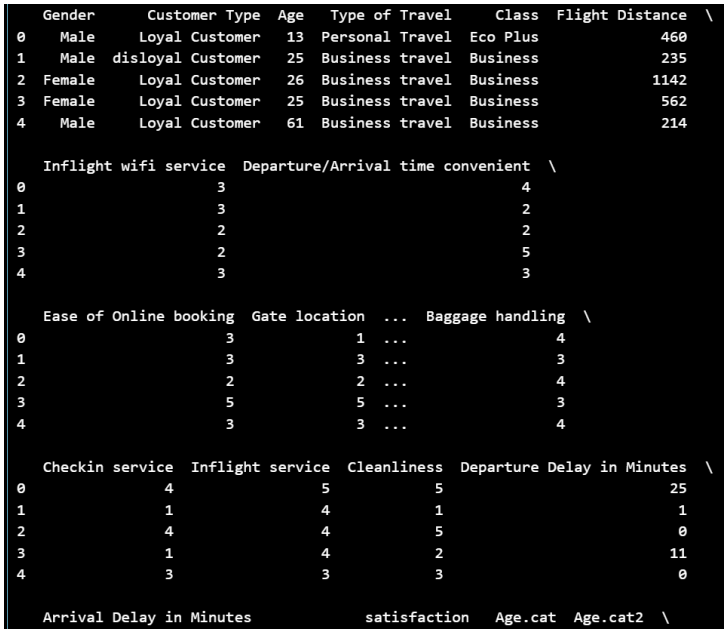
	Age.cat	Age.cat2
	Under 20	1
	20-40	2
	20-40	2
	20-40	2
	40-80	3
	...	...
	20-40	2
	40-80	3
	20-40	2
	20-40	2
	20-40	2

Fig 13

Numeric Transformation:

The 'satisfaction' variable was transformed into a numeric format where 'satisfied' was encoded as 1 and others as 0. This enabled us to use statistical methods that require numeric input.

First Few Columns of The Dataset:



```
   Gender      Customer Type  Age   Type of Travel     Class  Flight Distance  \
0    Male     Loyal Customer   13  Personal Travel  Eco Plus              460
1    Male  disloyal Customer   25  Business travel  Business              235
2  Female     Loyal Customer   26  Business travel  Business             1142
3  Female     Loyal Customer   25  Business travel  Business              562
4    Male     Loyal Customer   61  Business travel  Business              214

   Inflight wifi service  Departure/Arrival time convenient  \
0                      3                                  4
1                      3                                  2
2                      2                                  2
3                      2                                  5
4                      3                                  3

   Ease of Online booking  Gate location  ...  Baggage handling  \
0                       3              1  ...                 4
1                       3              3  ...                 3
2                       2              2  ...                 4
3                       5              5  ...                 3
4                       3              3  ...                 4

   Checkin service  Inflight service  Cleanliness  Departure Delay in Minutes  \
0                4                 5            5                          25
1                1                 4            1                           1
2                4                 4            5                           0
3                1                 4            2                          11
4                3                 3            3                           0

   Arrival Delay in Minutes            satisfaction   Age.cat  Age.cat2  \
```

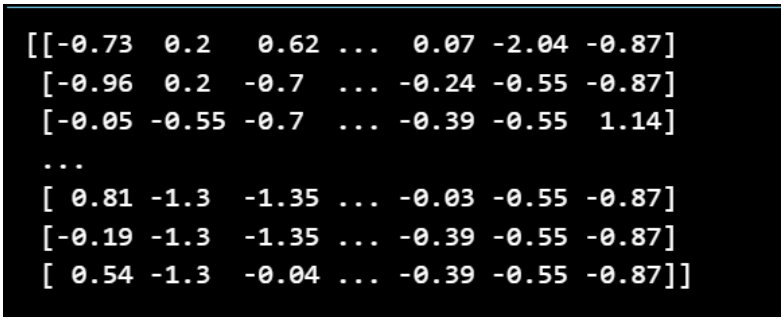
Fig 14

PRINCIPAL COMPONENT ANALYSIS:

I have conducted a Principal Component Analysis (PCA) on the pre-processed dataset to reduce the dimensionality of our features while retaining most of the variance within the data. The goal of PCA is to identify and quantify the most meaningful basis to re-express our dataset, providing us with the insight needed to understand the underlying structure of the data and to create a more efficient, informative representation.

STANDARDIZATION:

We began by standardizing the features to have a mean of 0 and a standard deviation of 1. This is an essential step as PCA is influenced by the scale of the data.



```
[[-0.73  0.2   0.62 ...  0.07 -2.04 -0.87]
 [-0.96  0.2  -0.7  ... -0.24 -0.55 -0.87]
 [-0.05 -0.55 -0.7  ... -0.39 -0.55  1.14]
 ...
 [ 0.81 -1.3  -1.35 ... -0.03 -0.55 -0.87]
 [-0.19 -1.3  -1.35 ... -0.39 -0.55 -0.87]
 [ 0.54 -1.3  -0.04 ... -0.39 -0.55 -0.87]]
```

Fig 15

OLD VS NEW SINGULAR VALUES:

The new singular values are likely lower than the original ones for the latter components because PCA has removed the dimensions that contributed less to the data variance, leaving us with components that have a greater impact.

```
singular values are
[662.36 494.79 474.98 450.89 381.66 315.87 310.31 300.77 267.71 243.44
 226.45 218.92 208.96 194.93 182.66 174.09 159.4  138.51  59.76]
```

Fig 16

```
the new singular values are
[662.36 494.79 474.98 450.89 381.66 315.87 310.31 300.77 267.71 243.44
 226.45 218.92 208.96 194.93 182.66]
```

Fig 17

OLD VS NEW CONDITIONAL NUMBER:

The condition number has significantly improved from 11.08 to 3.63, indicating a reduction in the multicollinearity within the dataset. This reflects a dataset that has been effectively regularized through PCA, potentially enhancing the performance of subsequent predictive modeling.

```
conditional number is
11.08
```

Fig 18

```
the new conditional number is
3.63
```

Fig 19

In conclusion, PCA has allowed us to streamline the feature set by focusing on the components that carry the most information. This reduction not only clarifies the patterns within the data but also prepares us for more effective data modeling and visualization.

GRAPH:

The graph you've provided represents the cumulative explained variance ratio as a function of the number of principal components retained in the PCA analysis. This plot is a critical tool to

determine the number of components necessary to capture a substantial portion of the variance within the data.

In principal component analysis, the initial steep slope of the explained variance curve signifies the dominance of a few key factors shaping the dataset. As the curve reaches an elbow point, typically around the fifth component, diminishing returns in variance explanation become apparent. Beyond the tenth component, the curve plateaus, indicating minimal gains in capturing additional information. To select an appropriate number of components for a specific variance threshold, such as 95%, one would pinpoint the point where the cumulative explained variance ratio surpasses this threshold, typically within the first few components.

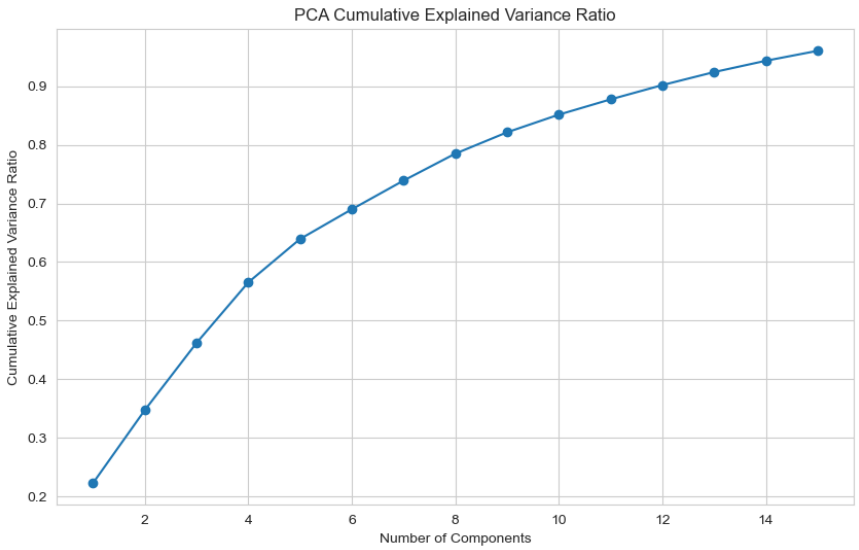


Fig 20

NORMALITY TEST:

The Shapiro-Wilk test has been used to assess the normality of various variables within our dataset. This test evaluates the null hypothesis that a data sample comes from a normally distributed population.

The Shapiro-Wilk test was conducted for all variables, yielding statistics ranging from 0.48 to 1.0. Despite some statistics approaching 1, all corresponding p-values were 0.0, indicating a significant deviation from normality at the 0.05 alpha level. Consequently, we reject the assumption of normality for each variable, concluding that none adheres to a Gaussian distribution based on this analysis.

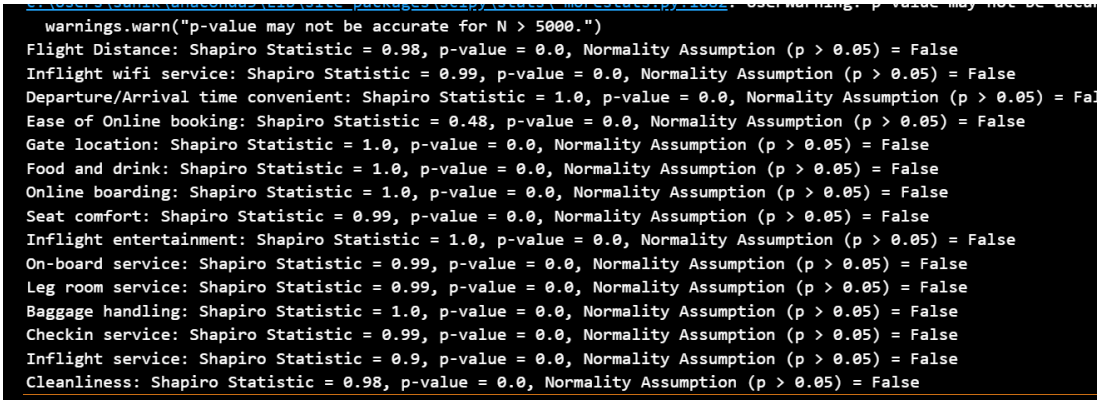


Fig 21

HEATMAP AND PEARSONS CORRELATION COEFFICIENT MATRIX

High correlation coefficients near 1 or -1 indicate strong linear relationships, like the connection between "Online boarding" and "satisfaction_numeric" suggesting higher online boarding ratings align with higher overall satisfaction. Conversely, correlations near 0 signify weak or no linear relationships, seen in variables like "Gate location" that may not directly impact customer satisfaction. Negative correlations indicate inverse relationships, and identifying multicollinearity between independent variables, such as "Inflight Wi-Fi service" and "Ease of Online booking," is crucial for optimizing machine learning models sensitive to such issues.

We have a table below to show the relationship between the dependent variable and other independent variables.

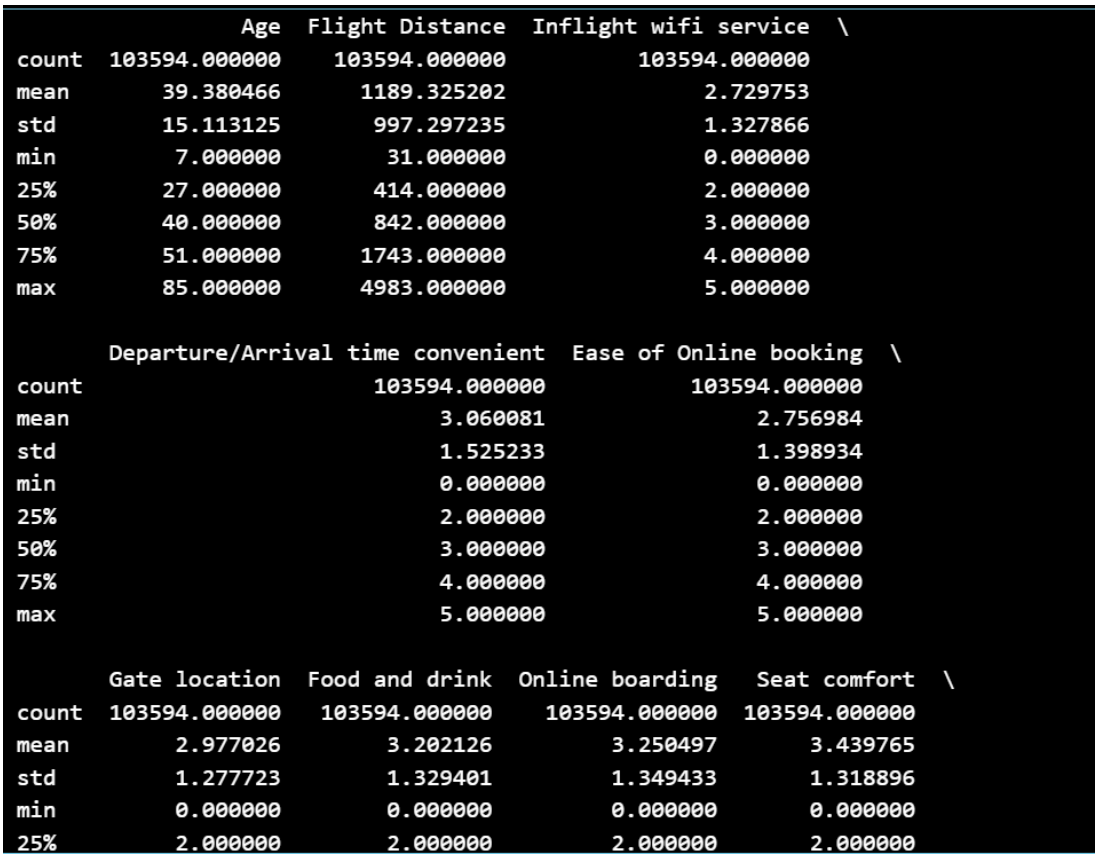
Variable pair	Correlation coefficient	Interpretation
Online boarding & Satisfaction	0.5	NOT CORRELATED
Inflight Wi-Fi service & Satisfaction	0.1	FAIRLY CORRELATED
Seat comfort & Satisfaction	0.4	NOT CORRELATED ENOUGH
Inflight entertainment & Satisfaction	0.3	MODERATELY CORRELATED
Cleanliness & Satisfaction	0.3	MODERATELY CORRELATED
Baggage handling & Satisfaction	0.1	FAIRLY CORRELATED
Food and drink & Satisfaction	0.1	FAIRLY CORRELATED
Inflight service & Satisfaction	0.1	FAIRLY CORRELATED

Flight Distance & Satisfaction	0.1	MODERATELY CORRELATED
Departure delay & Satisfaction	0	HIGHLY CORRELATED
Arrival Delay & Satisfaction	0	HIGHLY CORRELATED
Age & Satisfaction	0.2	FAIRLY CORRELATED
Gate Location & Satisfaction	0	HIGHLY CORRELATED
Arrival/Departure Time Convenient	0	HIGHLY CORRELATED

Table 1

STATISTICS

The dataset provides insights into various aspects of passenger experiences and demographics. The mean age of passengers is around 39 years, with a standard deviation of approximately 15 years, indicating a somewhat varied age distribution. Flight distances range widely from 31 to 4983 miles, with a noticeable peak at shorter distances, suggesting a higher frequency of shorter flights. Additionally, satisfaction ratings for different services like online booking, seat comfort, and cleanliness hover around the mid-range of 3 to 4, indicating a generally satisfactory but not exceptional experience. However, notable departures and arrival delays are present, with the mean delay being around 15 minutes, but some instances reaching over 1500 minutes, potentially impacting overall customer satisfaction.



	Age	Flight Distance	Inflight wifi service	\
count	103594.000000	103594.000000	103594.000000	
mean	39.380466	1189.325202	2.729753	
std	15.113125	997.297235	1.327866	
min	7.000000	31.000000	0.000000	
25%	27.000000	414.000000	2.000000	
50%	40.000000	842.000000	3.000000	
75%	51.000000	1743.000000	4.000000	
max	85.000000	4983.000000	5.000000	

	Departure/Arrival time convenient	Ease of Online booking	\
count	103594.000000	103594.000000	
mean	3.060081	2.756984	
std	1.525233	1.398934	
min	0.000000	0.000000	
25%	2.000000	2.000000	
50%	3.000000	3.000000	
75%	4.000000	4.000000	
max	5.000000	5.000000	

	Gate location	Food and drink	Online boarding	Seat comfort	\
count	103594.000000	103594.000000	103594.000000	103594.000000	
mean	2.977026	3.202126	3.250497	3.439765	
std	1.277723	1.329401	1.349433	1.318896	
min	0.000000	0.000000	0.000000	0.000000	
25%	2.000000	2.000000	2.000000	2.000000	

Fig 22

The KDE analysis of the age distribution reveals a primary peak in the mid-30s to 40s range, indicating a significant portion of passengers falling within this demographic. Additionally, smaller secondary peaks in the 20s and 50-60s age groups suggest varying proportions of passengers across different age brackets. The KDE's spread showcases a broad range of ages, tapering off in density beyond 60, indicating a decline in the number of older passengers.

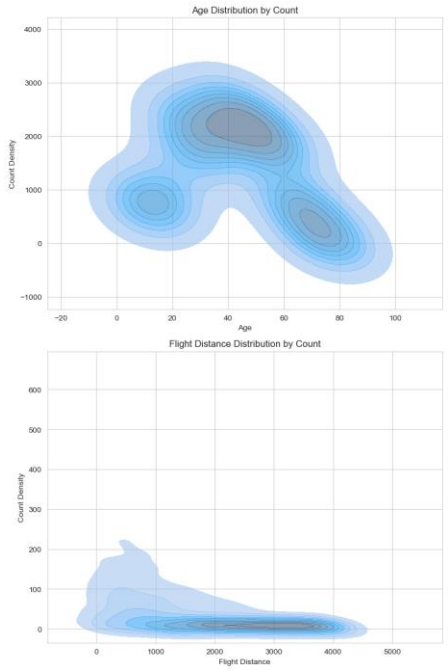


Fig 23

The Flight Distance Distribution KDE reveals a right-skewed pattern, emphasizing a prevalence of shorter flights over longer ones. The presence of a long tail indicates occasional very long flights taken by passengers. Moreover, the multiple modes in the distribution hint at distinct clusters of flight distances, possibly representing short-haul, medium-haul, and long-haul flights.

DATA VISUALISATION

PASSENGER AGE DISTRIBUTION

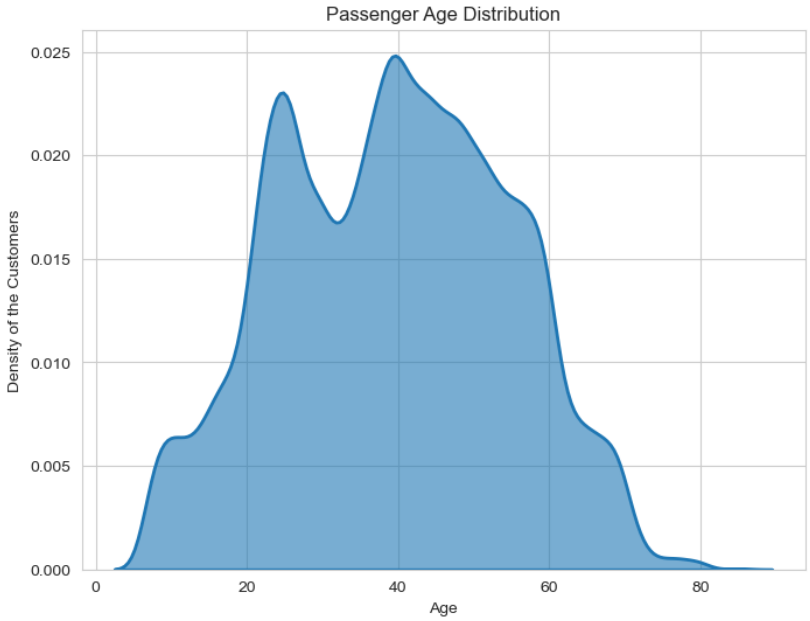


Fig 24

The dataset spans a wide age range, capturing responses from individuals as young as 7 and as old as 85. Most respondents fall within the age bracket of 20 to 60 years, signifying that the dataset primarily represents the adult population. This diverse age distribution underscores the dataset's comprehensive nature, providing insights into passenger satisfaction across various ages.

PASSENGERS VS GENDER

Fig 25

The gender distribution bar plot shows a significant gender disparity, with a higher proportion of female respondents than males, potentially impacting overall satisfaction trends among surveyed airline passengers.

SUBPLOT OF CATEGORICAL COLUMNS VS PASSENGERS

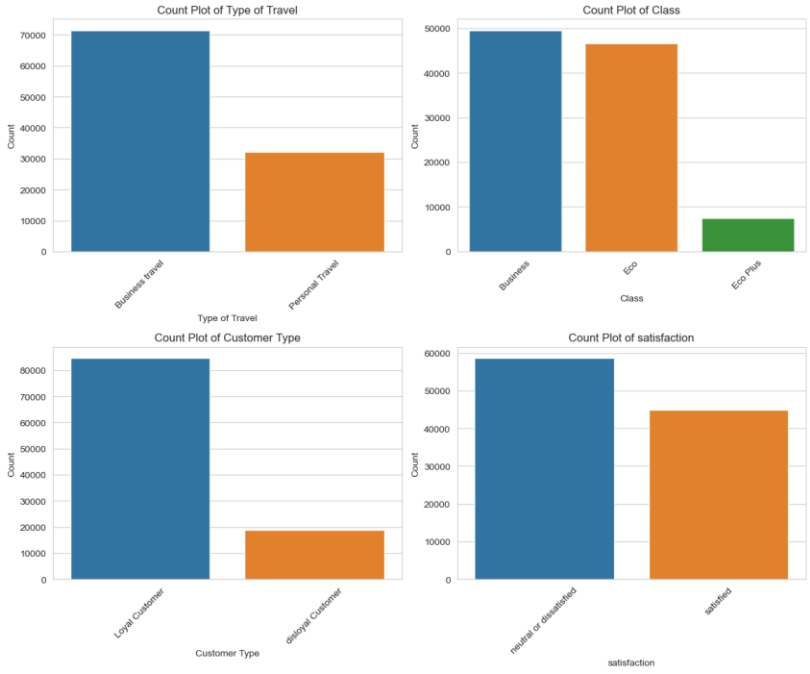


Fig 26

1) Type of travel: business travel is popular as compared to personal travel.
2) Class: very less people travel from economy class, as compared to eco plus and business, with business being the top.
3) Customer type: loyal customers are more than disloyal customers.
4) Satisfaction: the number of dissatisfied customers exceeds the number of satisfied customers.

PASSENGERS VS CATEGORICAL COLUMNS WITH CUSTOMER SATISFACTION

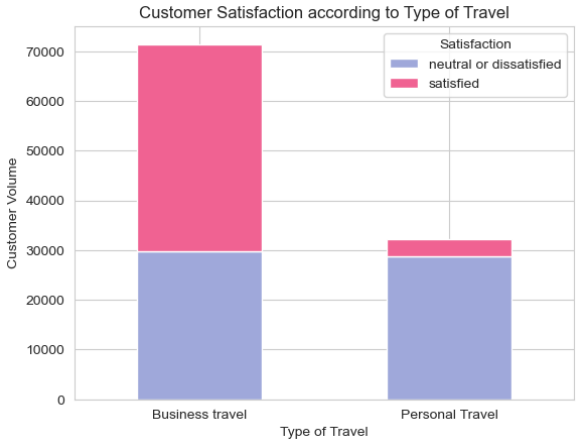


Fig 27

Fig 28

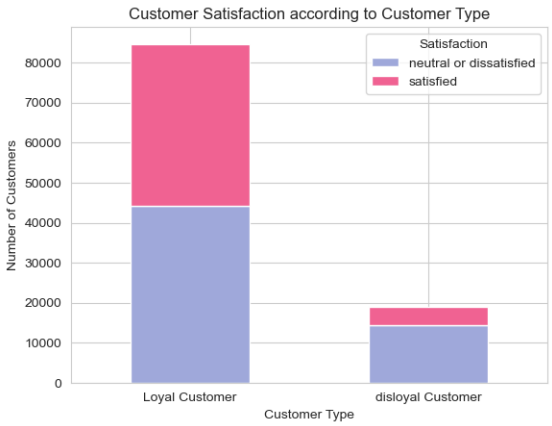


Fig 29

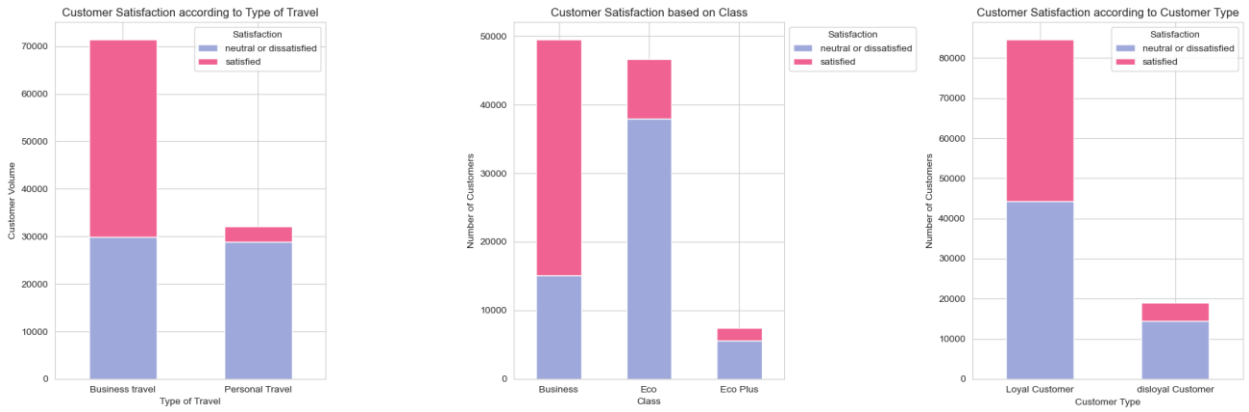


Fig 30

First Graph: The number of satisfied customers is more in business travel as compared to personal travel.

Second Graph: we can see that the number of dissatisfied customers is way more in eco and eco plus as compared to satisfied customers. Similarly, in business class there are more satisfied customers than dissatisfied customers.

Third Graph: satisfied and dissatisfied are somewhat equal under loyal customers, however under disloyal customers dissatisfied are more.

PIE CHART BETWEEN CLASS AND GENDER

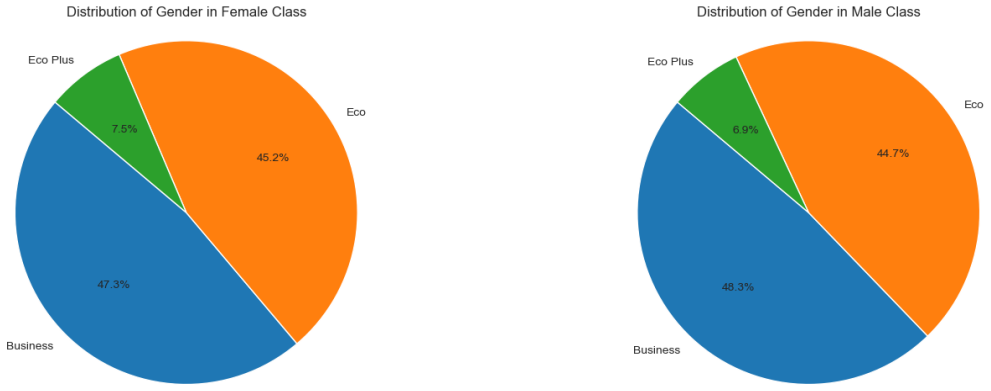


Fig 31

The first pie chart shows the distribution of females in class. There are about 7.5% females in eco plus (it is having the least number of passengers). 45.2% female passengers in economy and 47.3% females in business class (comprises the highest number female passengers).

The second pie chart shows the distribution of males in class. There are about 6.9% males in eco plus (it is having the least number of passengers). 44.7% male passengers in economy and 48.3% males in business class (comprises the highest number male passengers).

PIE CHART BETWEEN TYPE OF TRAVEL AND CLASS

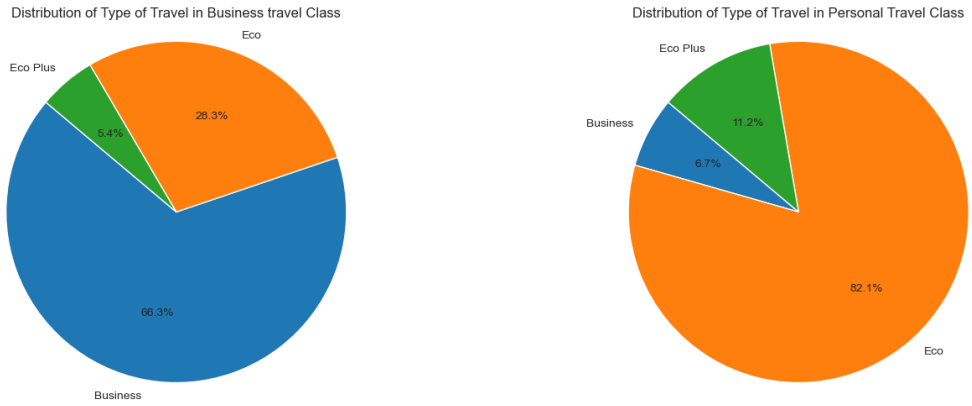


Fig 32

The first pie chart shows the distribution of business as purpose of travel in class. There are about 5.4% passengers in eco plus (it is having the least number of passengers). 28.3% passengers in economy and 66.3% passengers in business class (comprises the highest number passengers). Indicating that for purposes of business, people travel more in business class.

The second pie chart shows the distribution of personal as purpose of travel in class. There are about 11.2% passengers in eco plus. 6.7% passengers in business (it has the least number of passengers) and 82.1% passengers in economy class (comprises the highest number passengers). Indicating that for purposes of personal travel, people prefer economy class.

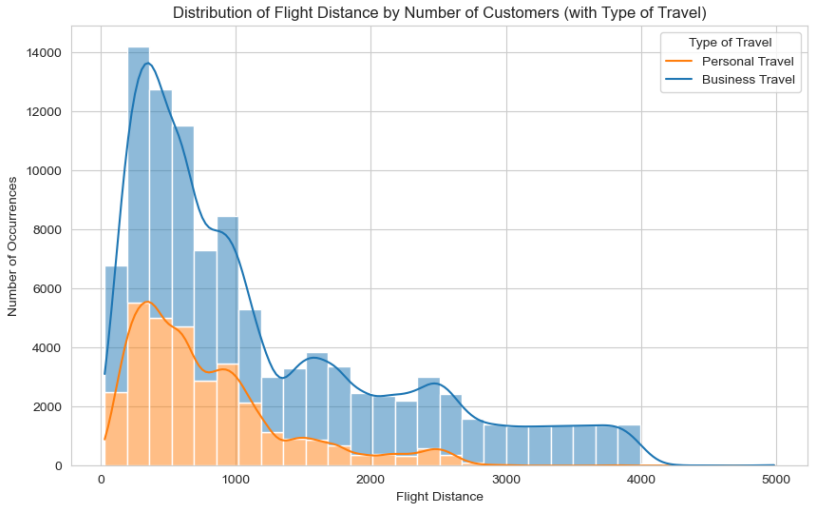


Fig 33

The analysis of the flight distance graph indicates a notable trend: as the flight distance increases, there is a significant decrease in the number of customers traveling. This pattern suggests a preference among passengers for shorter-distance flights, potentially reflecting differing travel purposes. Specifically, individuals traveling for personal reasons exhibit a tendency to opt for shorter distances compared to those traveling for business purposes. These observations underscore the influence of travel purposes on the choice of flight distance among passengers.

BOX AND VIOLIN PLOT OF SEAT COMFORT BY CLASS

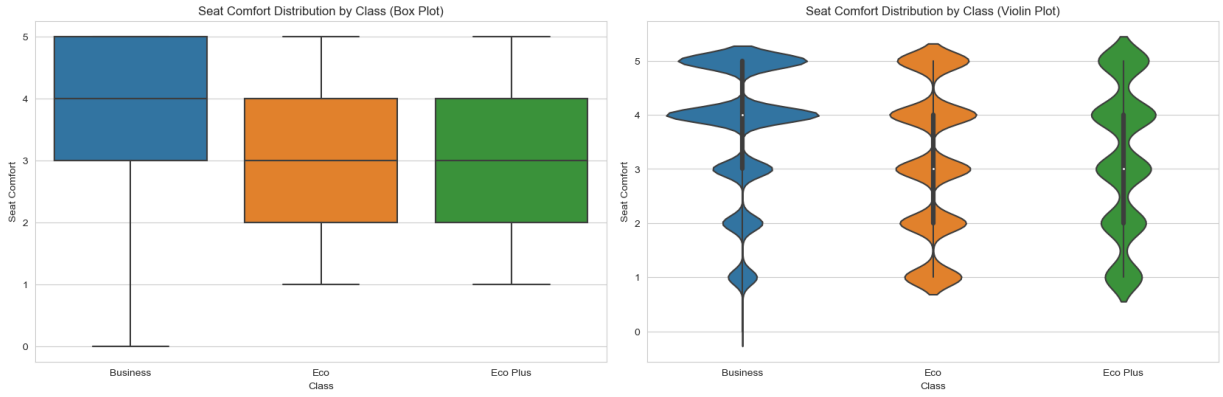


Fig 34

These graphs tell us that passengers traveling by business class have rated 4 and 5 for seat comfort which is well understood. However, passengers traveling by eco and eco plus are somewhat neutral with feedback of seat comfort.

VIOLIN PLOT OF DIFFERENT INFLIGHT SERVICES BY SATISFACTION

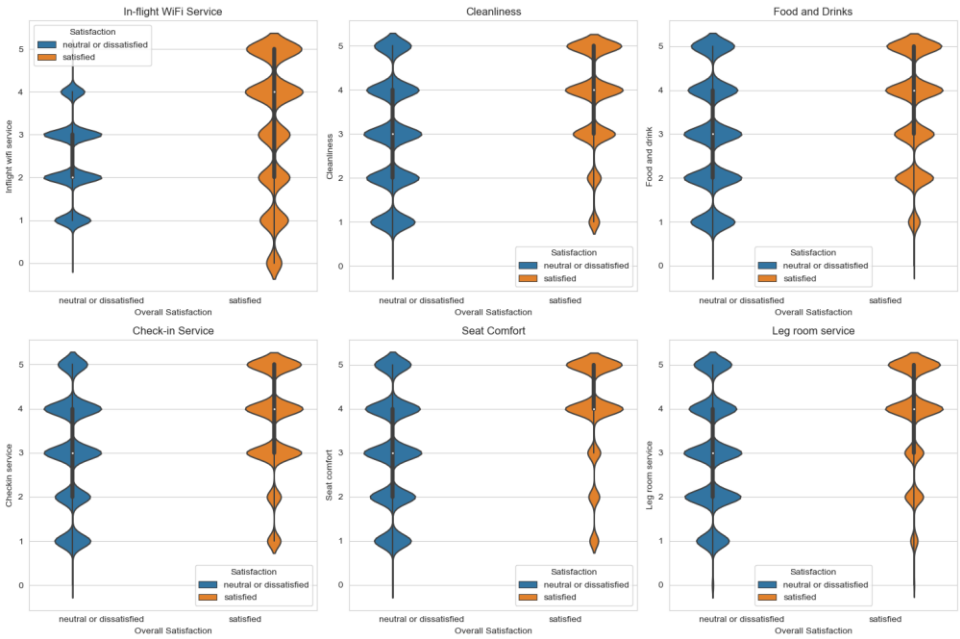


Fig 35

In-flight Wi-Fi Service: Higher ratings for in-flight Wi-Fi are associated with satisfaction, indicating its importance to the overall passenger experience.

Cleanliness: Cleanliness shows a clear distinction between satisfied and dissatisfied passengers, emphasizing its impact on satisfaction.

Food and Drinks: Food and drink quality appears to have a broad distribution among satisfied passengers, suggesting varied expectations or experiences.

Check-in Service: A smoother distribution in check-in service ratings among satisfied passengers might indicate consistent expectations being met.

Seat Comfort: Seat comfort has a strong influence on passenger satisfaction, with satisfied passengers often rating it higher.

Leg Room Service: Leg room service has a notable impact on satisfaction, with dissatisfied passengers often giving lower ratings.

COUNT PLOT OF CHECK IN SERVICE SATISFACTION BY TYPE OF TRAVEL

Fig 36

There are a greater number of dissatisfied customers under loyal type of customer as compared to satisfied.

Similarly, there are a greater number of dissatisfied customers under disloyal types of customers as well indicating that airlines need to fix the issue related to check in services for better passenger experience.

COUNT PLOT OF TYPE OF TRAVEL VS SATISFACTION

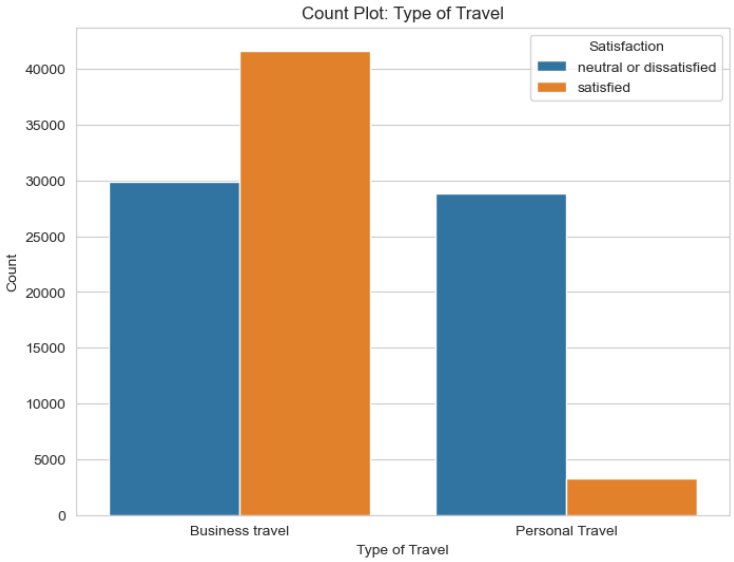


Fig 37

From this plot we can see that there are a greater number of satisfied people when they travel for business purposes as compared to dissatisfied.

However, the number of dissatisfied passengers traveling for personal purposes is way higher than satisfied passengers.

JOINT PLOT: AGE VS FLIGHT DISTANCE WITH GENDER

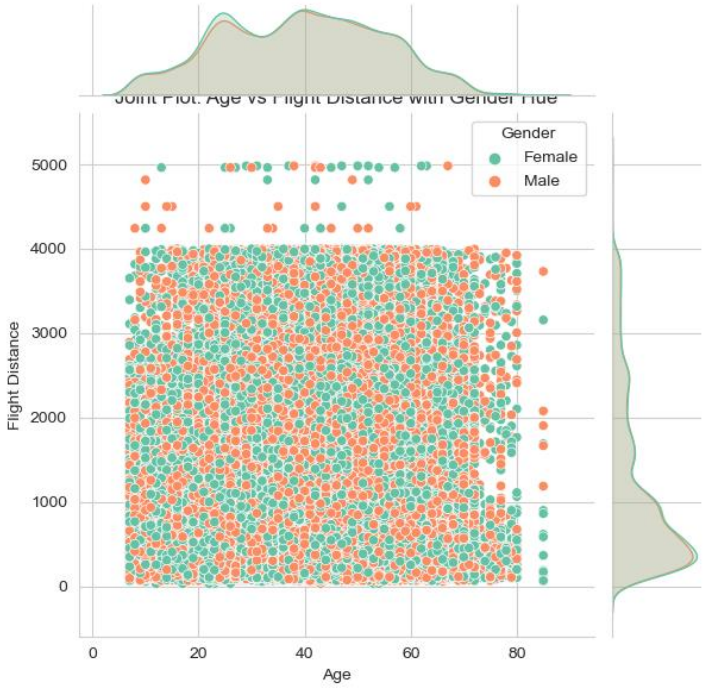


Fig 38

Both genders are equally distributed across various ages and flight distances, indicating that gender is not a distinguishing factor for these variables. The density along the middle shows that most passengers, regardless of gender, are in the middle-age range and take flights of shorter distances. As flight distance increases, there is a corresponding decrease in the proportion of older passengers traveling. This trend suggests that the younger population may demonstrate a greater capacity or inclination to cover longer distances comfortably.

GENDER VS SATISFACTION

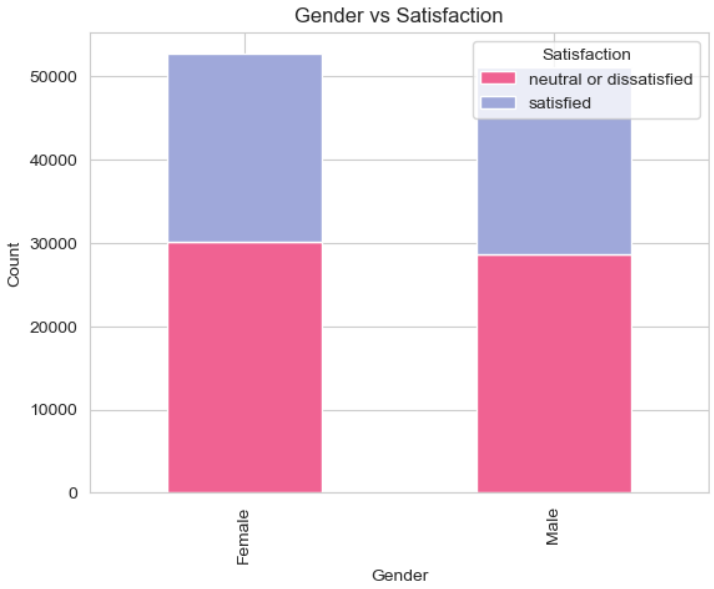


Fig 39

In both male and female, the number of dissatisfied passengers is slightly more than number of satisfied passengers.

LINE PLOT OF AGE VS FLIGHT DISTANCE

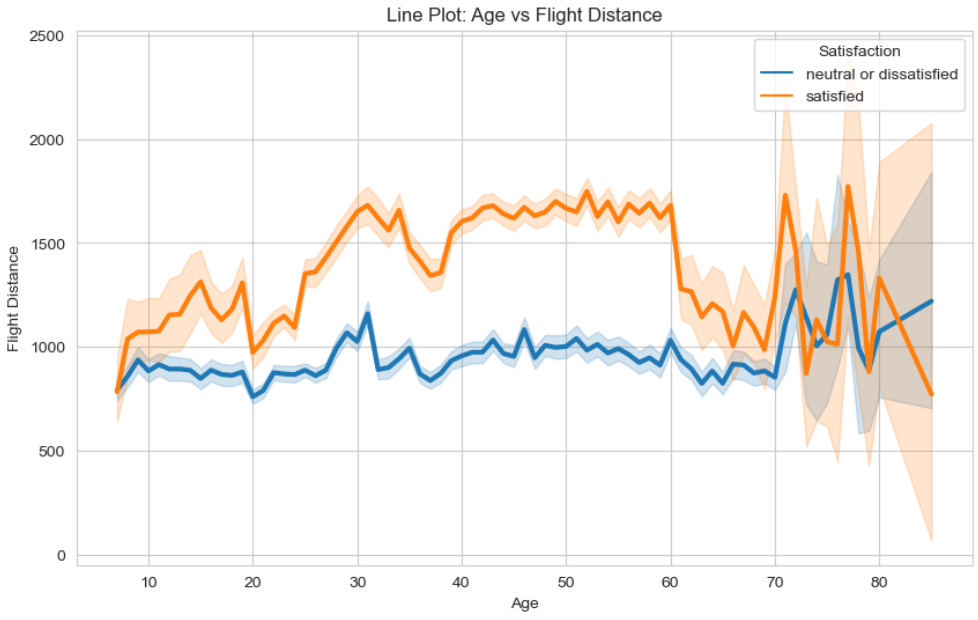


Fig 40

The line plot indicates variability in flight distance across different ages, with no consistent trend linking age to flight distance for both satisfied and neutral or dissatisfied passengers. Both groups exhibit fluctuations in flight distance with age, yet there's a convergence of flight distances across satisfaction levels in the 60-70 age range. Additionally, the shaded areas, which likely represent confidence intervals or data variability, suggest more variability in flight distances chosen by satisfied customers. Overall, passenger satisfaction does not seem to be strongly associated with age or the flight distances they travel.

QQPLOT OF FLIGHT DISTANCE AND DEPARTURE DELAY IN MINUTES

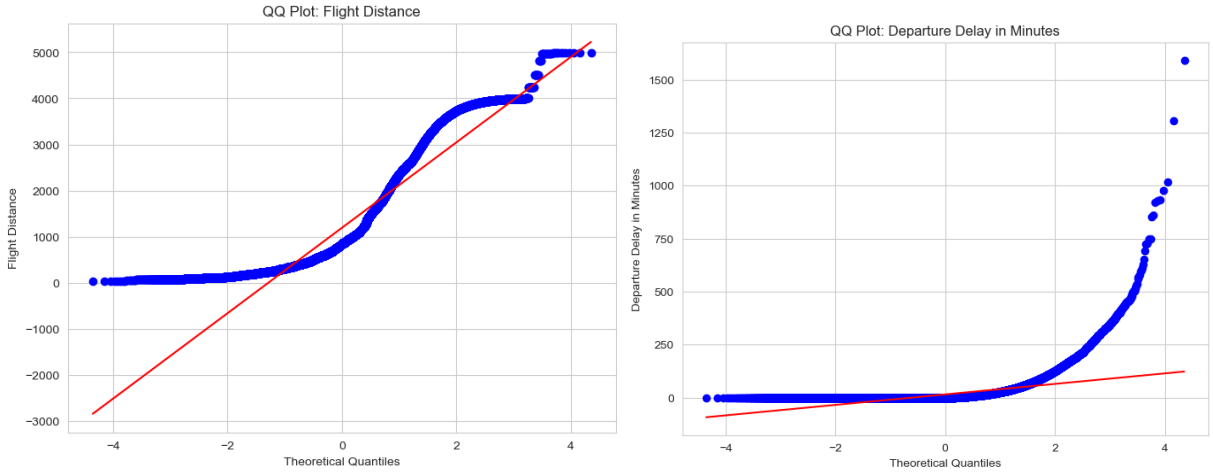


Fig 41

For the 'Flight Distance', the data shows a rightward (positive) curvature away from the normal distribution line, indicating a right-skewed distribution with a heavy tail—meaning there are a notable number of flights with distances much longer than the average.

In the case of 'Departure Delay in Minutes', the extreme upward deviation at the higher quantiles also suggests a right-skewed distribution, where most delays are short, but there are a significant number of instances with much longer than typical delays.

These patterns suggest that neither variable is normally distributed, and their skewness could be indicative of operational factors in the airline industry, such as the prevalence of shorter flights and the occasional significant delay.

MULTIVARIATE BOXPLOT AND VIOLIN PLOT OF SERVICES WITH SATISFACTION

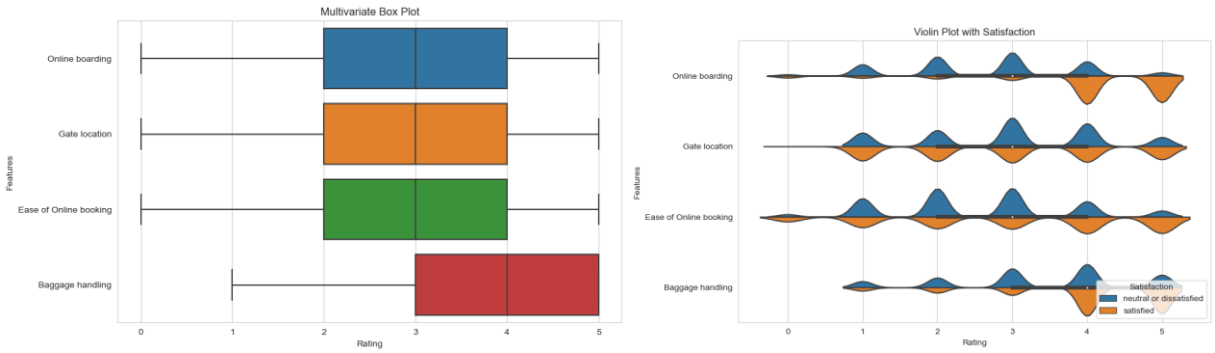


Fig 42

The violin and box plot visualizes the distribution of customer satisfaction ratings for different service features. The plots indicate that higher ratings for online boarding, ease of online booking, and baggage handling are more commonly associated with satisfaction, as shown by the wider sections of the violin for satisfied customers. Gate location, however, shows less distinction between satisfied and neutral or dissatisfied customers, suggesting it might be less influential on overall satisfaction. Each feature's distribution shows peaks at different rating levels, with the satisfaction group tending to have peaks at higher ratings, indicating that quality in these service areas is a potential contributor to overall passenger satisfaction.

AREA PLOT OF SATISFACTION RATING OFINFLIGHT WIFI SERVICE AND CLEANLINESS

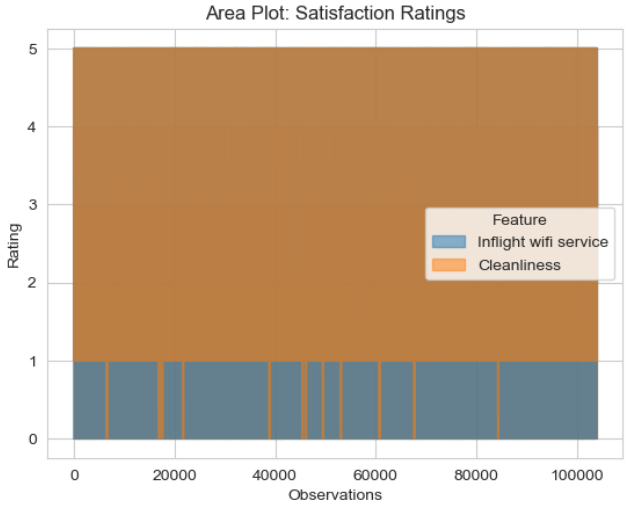


Fig 43

The area plot presents the aggregate satisfaction ratings for in-flight Wi-Fi service and cleanliness. It appears that cleanliness has consistently high ratings across many observations, dominating over in-flight Wi-Fi service, which has comparatively lower ratings. This visualization suggests that passengers generally rate cleanliness more favorably than in-flight Wi-Fi service, indicating a potential area of improvement for airlines in the Wi-Fi service offered.

3D Scatter Plot of Age vs Flight Distance vs Seat Comfort

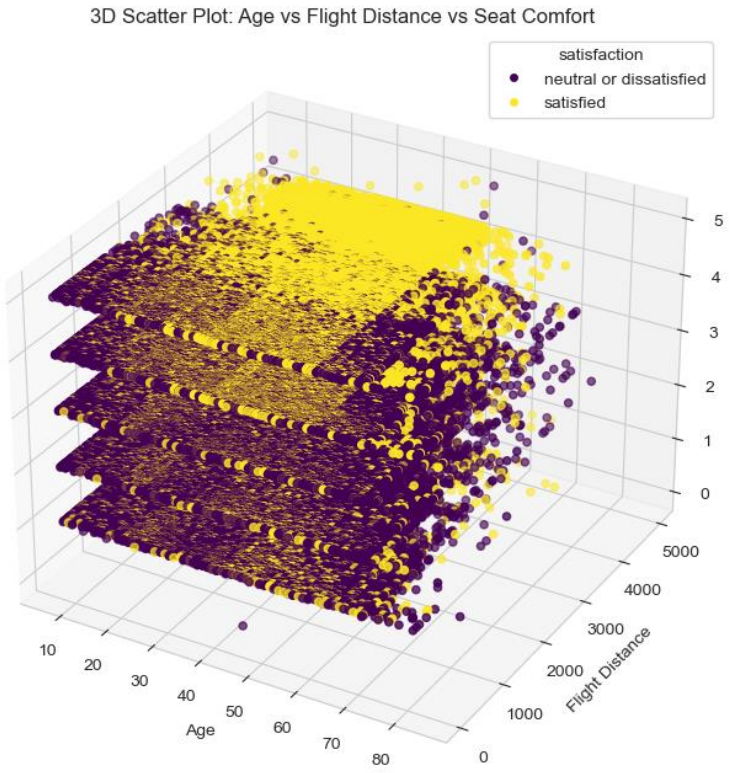


Fig 44

The middle region shows a somewhat normal distribution of satisfied and dissatisfied customers under, age, flight distance and seat comfort. The dissatisfaction tends to occur more frequently as passengers age and as flight distance increases.

HEXBIN PLOT OF AGE VS FLIGHT DISTANCE

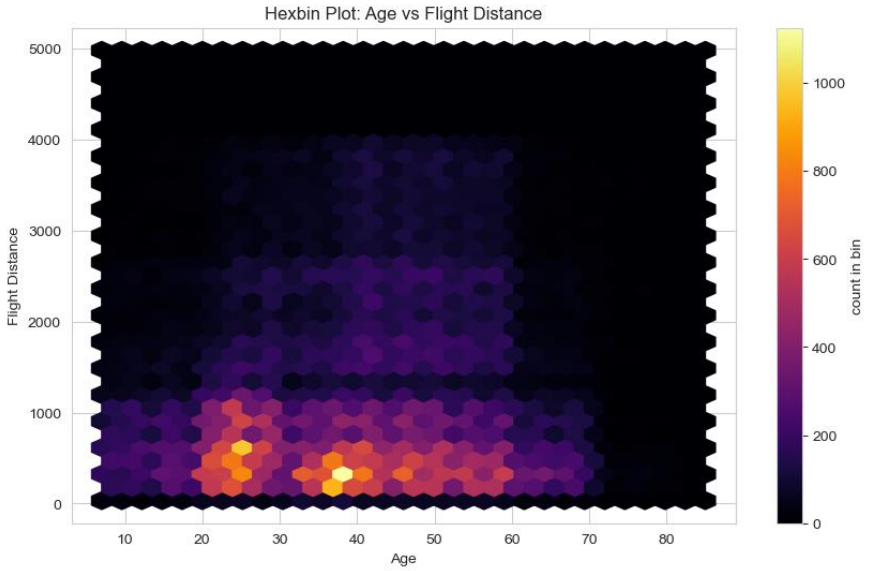


Fig 45

The hexbin plot illustrates the density of observations within the dataset for age versus flight distance. Darker hexagons, especially in the lower age ranges and shorter flight distances, show a higher concentration of passengers, indicating that most passengers are younger and travel shorter distances. The lighter hexagons in the higher age and flight distance regions suggest fewer observations, indicating that older passengers and longer flights are less common within the dataset.

STRIP PLOT: TYPE OF TRAVEL WITH DIFFERENT ATTRIBUTES

Fig 46

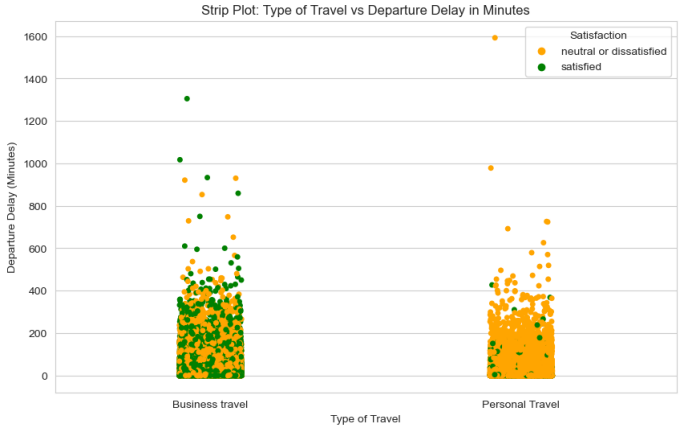


Fig 47

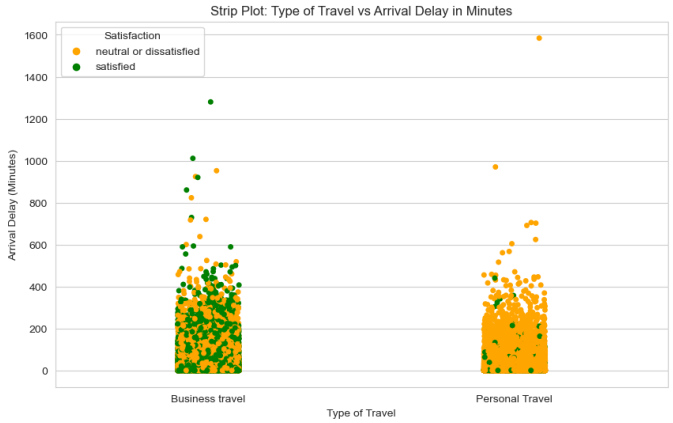


Fig 48

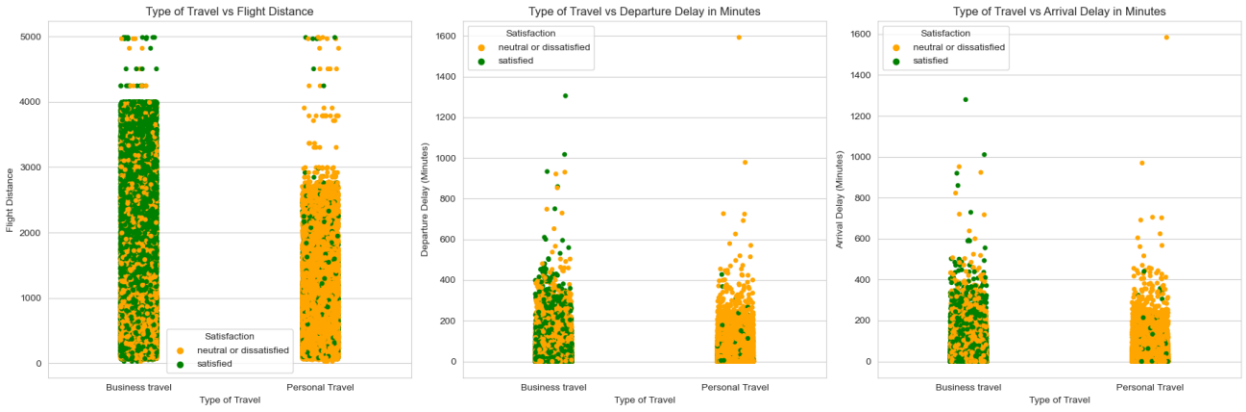


Fig 49

The scatter plots contrast the flight distance and delays (both departure and arrival) against the type of travel, segmented by overall passenger satisfaction. Business travelers show a wide range of flight distances but with more satisfaction noted at all levels compared to personal travel. For both departure and arrival delays, longer delays show a concentration of neutral or dissatisfied

passengers, indicating that delays, regardless of travel type, are likely to negatively impact passenger satisfaction.

RADAR CHARTS OF INFLIGHT SERVICE, CLEANLINESS, DEPARTURE/ARRIVAL TIME CONVENIENT

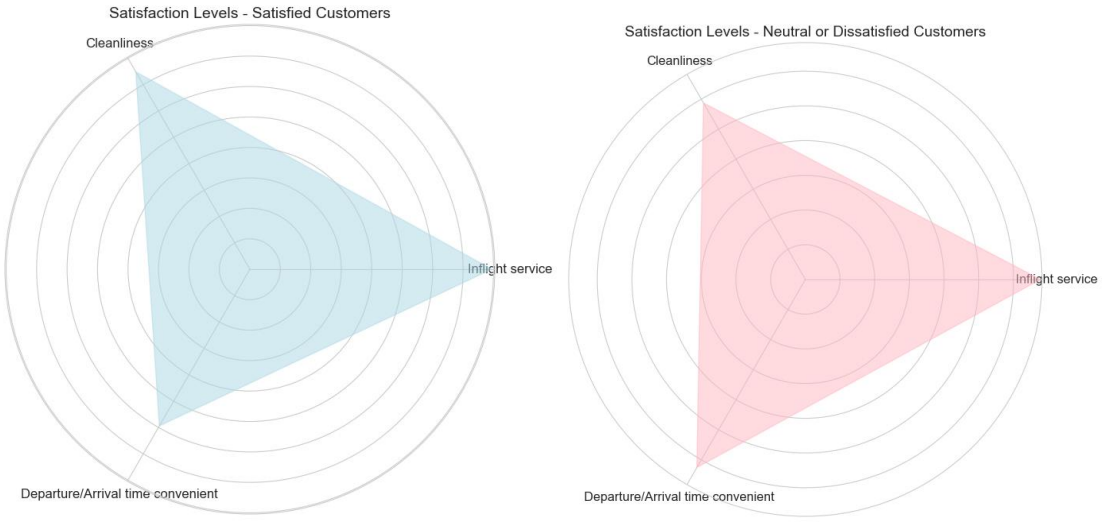


Fig 50

The radar charts compare levels of customer satisfaction in terms of inflight service, cleanliness, and convenience of departure/arrival times. For neutral or dissatisfied customers, in-flight service is the most lacking feature, followed by departure/arrival time convenience and cleanliness. Conversely, for satisfied customers, in-flight service also scores the highest, suggesting that this feature is critical to passenger satisfaction. Both charts highlight in-flight service as a significant factor, indicating it has a strong influence on the overall satisfaction levels of customers.

PAIR PLOT

Fig 51

The pair plot visualizes the relationships between age, flight distance, and service ratings for inflight Wi-Fi, food and drink, and seat comfort. Age and flight distance show a dense clustering without a distinct pattern, indicating that passengers of all ages are traveling across a range of distances. Service ratings appear discreetly banded due to their categorical nature and indicating that higher ratings in one service area correspond to higher ratings in another, suggesting that passenger perceptions of these services are independent of each other.

REGRESSION PLOT OF AGE VS ARRIVAL DELAY IN MINUTES

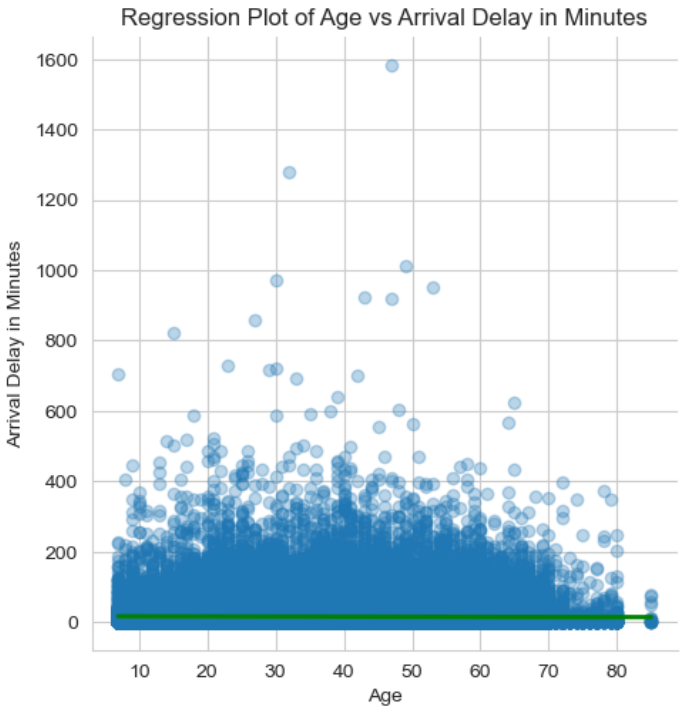


Fig 52

The regression plot indicates that there is no significant correlation between passengers' age and the arrival delay of their flights, as suggested by the flat regression line. The scatter of points shows a wide variation in arrival delays across all ages, with a few extreme delays likely causing some minor upward skew.

RUG PLOT OF ARRIVAL DELAY IN MINUTES

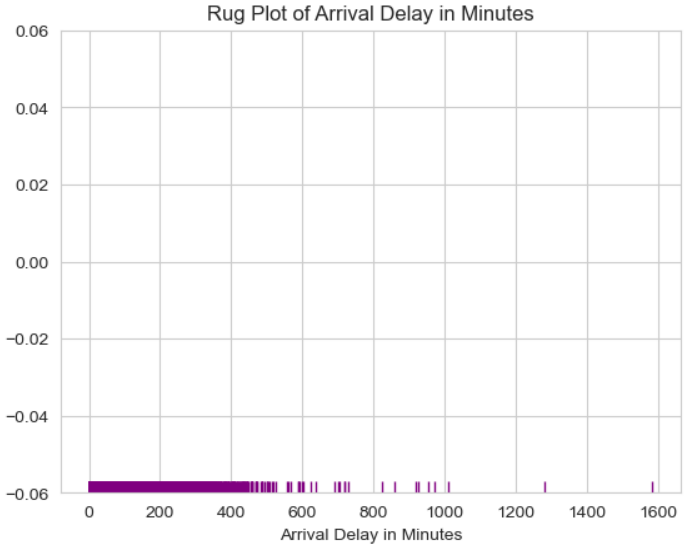


Fig 53

The rug plot visualizes individual data points for arrival delays in minutes along a single axis, showing the frequency of delays at different durations. Most delays are clustered at the lower end of the scale, indicating that shorter delays are much more common than longer ones. The few marks spread out towards the higher end of the axis suggest that while long delays are relatively rare, they do occur, with a noticeable gap between common and extreme delay times.

DASHBOARD

LINK: https://dashapp-vfdfwm4wta-ue.a.run.app/

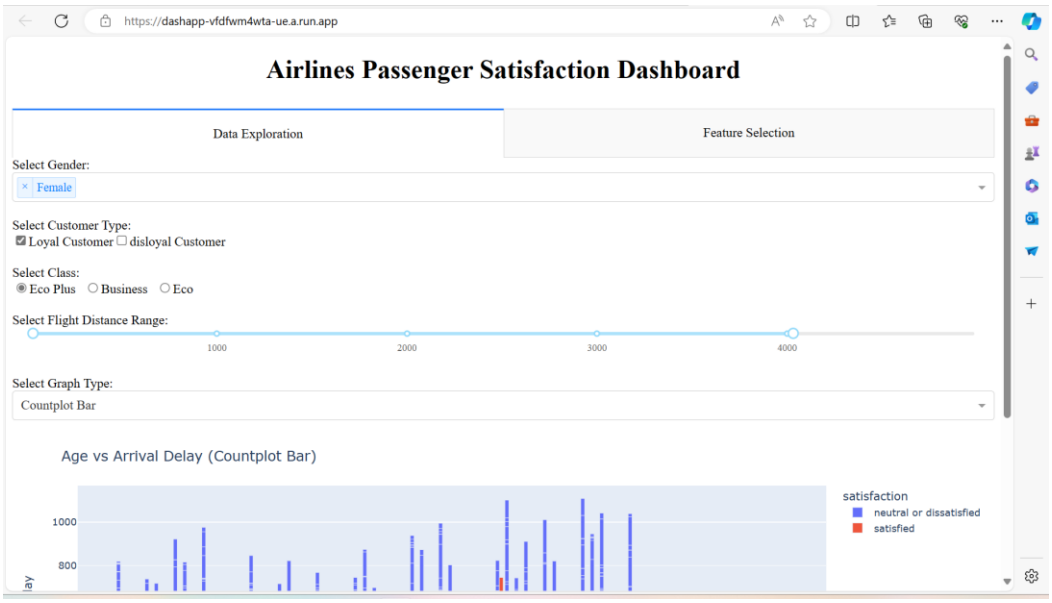


Fig 54

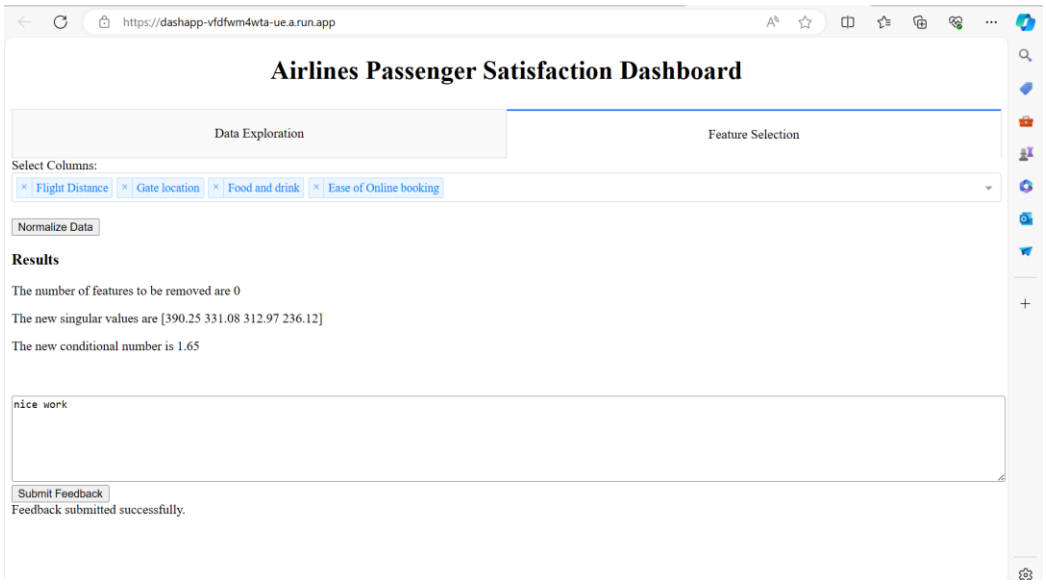


Fig 55

The provided Dash application, which is a web-based dashboard for visualizing airline passenger satisfaction data. The application includes features for data exploration, normalization, and feedback submission. It's designed to be deployed on Google Cloud Platform (GCP), leveraging GCP's scalable and secure infrastructure.

CONCLUSION

The various graphs created in this project provide a comprehensive visual exploration of the factors influencing airline passenger satisfaction. I learned that several factors, such as flight distance, age, service quality (in-flight Wi-Fi, food and drink, seat comfort), and travel type, have varying impacts on satisfaction. The visualizations also underscored the complexity of these relationships, where no single factor guarantees satisfaction, and how each element contributes differently to the overall experience.

The created Python dashboard serves as a dynamic and interactive tool that enables users to explore the dataset through customizable visualizations. With functionalities such as filtering by demographics, selecting specific flight characteristics, and comparing satisfaction levels, it allows users to identify trends and patterns that may not be immediately apparent from raw data, aiding in data-driven decision-making processes.

The functionality of the created app appears to be robust, providing users with a versatile platform to analyze the airline dataset interactively. The app includes features for normalizing data, applying PCA for feature reduction, and downloading user feedback. Its design facilitates user engagement and accessibility, making it a functional and valuable tool for stakeholders interested in understanding customer satisfaction drivers in the airline industry.

APPENDIX

```
Run Cell | Run Below | Debug Cell | Go to [1]
#%%
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
import dash as dash
from dash import html, dcc
from dash.dependencies import Input, Output, State
from pandas_datareader import data
from datetime import date
from numpy.linalg import cond, svd
from scipy import stats

Run Cell | Run Above | Debug Cell | Go to [2]
# %%
file_path = "Airlines_train.csv"
df = pd.read_csv(file_path)

Run Cell | Run Above | Debug Cell | Go to [3]
#%%
print(df.describe())

Run Cell | Run Above | Debug Cell | Go to [4]
#%%
print(df.isnull().sum())

df.dropna(inplace=True)

print(df.isnull().sum())

```

```
Run Cell | Run Above | Debug Cell | Go to [5]
#%%
df.replace([np.inf, -np.inf], np.nan, inplace=True)
Run Cell | Run Above | Debug Cell | Go to [6]
#%%
df = df.drop(columns=['X', 'id'])

Run Cell | Run Above | Debug Cell | Go to [7]
#%%
df['Age.cat'] = None

df.loc[df['Age'] <= 20, 'Age.cat'] = 'Under 20'
df.loc[(df['Age'] >= 21) & (df['Age'] <= 40), 'Age.cat'] = '20-40'
df.loc[(df['Age'] >= 41) & (df['Age'] <= 80), 'Age.cat'] = '40-80'
df.loc[df['Age'] >= 80, 'Age.cat'] = 'above 80'

df['Age.cat'] = pd.Categorical(df['Age.cat'], categories=['Under 20', '20-40', '40-80', 'above 80'])

df['Age.cat2'] = None

df.loc[df['Age'] <= 20, 'Age.cat2'] = 1
df.loc[(df['Age'] >= 21) & (df['Age'] <= 40), 'Age.cat2'] = 2
df.loc[(df['Age'] >= 41) & (df['Age'] <= 80), 'Age.cat2'] = 3
df.loc[df['Age'] >= 80, 'Age.cat2'] = 4

print(df)

Run Cell | Run Above | Debug Cell | Go to [8]
#%%

df['satisfaction_numeric'] = df['satisfaction'].apply(lambda x: 1 if x == 'satisfied' else 0)

```

```
57
58
    Run Cell | Run Above | Debug Cell | Go to [9]
59  #%%
60  df['satisfaction'] = df['satisfaction'].astype('category')
61  df['Class'] = df['Class'].astype('category')
62  df['Type of Travel'] = df['Type of Travel'].astype('category')
63  df['Gender'] = df['Gender'].astype('category')
64  df['Customer Type'] = df['Customer Type'].astype('category')
65  df['Age.cat']=df['Age.cat'].astype('category')
66  df['Age.cat2']=df['Age.cat2'].astype('int64')
    Run Cell | Run Above | Debug Cell | Go to [10]
67  #%%
68  print(df.dtypes)
69
    Run Cell | Run Above | Debug Cell | Go to [11]
70  # %%
71  q1_Flight_Distance = df['Flight Distance'].quantile(0.25)
72  q3_Flight_Distance = df['Flight Distance'].quantile(0.75)
73  iqr_Flight_Distance= q3_Flight_Distance - q1_Flight_Distance
74  lower_bound = q1_Flight_Distance - 1.5 * iqr_Flight_Distance
75  upper_bound = q3_Flight_Distance + 1.5 * iqr_Flight_Distance
76
77  print(f"Q1 and Q3 of the Flight Distance is {q1_Flight_Distance:.2f}  & {q3_Flight_Distance:.2f} .")
78  print(f"IQR for the Flight Distance is {iqr_Flight_Distance:.2f} .")
79  print(f"Any Flight Distance < {lower_bound:.2f}  and Flight Distance > {upper_bound:.2f}  is an outlier.")
80
81  plt.figure(figsize=(8, 6))
82  plt.boxplot(df['Flight Distance'])
83  plt.xlabel('Flight Distance')
84  plt.title('Boxplot of Flight Distance')
85
86  plt.show()
```

```
87
88  cleaned_df = df[(df['Flight Distance'] >= lower_bound) & (df['Flight Distance'] <= upper_bound)]
89
90  plt.figure(figsize=(10, 6))
91  plt.boxplot(cleaned_df['Flight Distance'])
92  plt.title('Boxplot-Flight Distance-Cleaned')
93  plt.xlabel('Flight Distance')
94  plt.grid(True)
95  plt.show()
96
    Run Cell | Run Above | Debug Cell | Go to [12]
97  #%%
98  q1_Departure_Delay = df['Departure Delay in Minutes'].quantile(0.25)
99  q3_Departure_Delay= df['Departure Delay in Minutes'].quantile(0.75)
100 iqr_Departure_Delay = q3_Departure_Delay - q1_Departure_Delay
101 lower_bound = q1_Departure_Delay - 1.5 * iqr_Departure_Delay
102 upper_bound = q3_Departure_Delay + 1.5 * iqr_Departure_Delay
103
104 print(f"Q1 and Q3 of the Departure Delay is {q1_Departure_Delay:.2f}  & {q3_Departure_Delay:.2f} .")
105 print(f"IQR for the Departure Delay is {iqr_Departure_Delay:.2f} .")
106 print(f"Any Departure_Delay < {lower_bound:.2f}  and Departure_Delay > {upper_bound:.2f}  is an outlier.")
107
108 plt.figure(figsize=(8, 6))
109 plt.boxplot(df['Departure Delay in Minutes'])
110 plt.xlabel('Departure_Delay')
111 plt.title('Boxplot of Departure_Delay')
112
113 plt.show()
114
115 cleaned_df = df[(df['Departure Delay in Minutes'] >= lower_bound) & (df['Departure Delay in Minutes'] <= upp
116
117 plt.figure(figsize=(10, 6))
118 plt.boxplot(cleaned_df['Departure Delay in Minutes'])
```

```
118   plt.boxplot(cleaned_df['Departure Delay in Minutes'])
119   plt.title('Boxplot-Departure_Delay-Cleaned')
120   plt.xlabel('Departure_Delay')
121   plt.grid(True)
122   plt.show()
123
      Run Cell | Run Above | Debug Cell | Go to [13]
124   #%%
125   q1_Arrival_Delay = df['Arrival Delay in Minutes'].quantile(0.25)
126   q3_Arrival_Delay= df['Arrival Delay in Minutes'].quantile(0.75)
127   iqr_Arrival_Delay = q3_Arrival_Delay - q1_Arrival_Delay
128   lower_bound = q1_Arrival_Delay - 1.5 * iqr_Arrival_Delay
129   upper_bound = q3_Arrival_Delay + 1.5 * iqr_Arrival_Delay
130
131   print(f"Q1 and Q3 of the Arrival_Delay is {q1_Departure_Delay:.2f}  & {q3_Departure_Delay:.2f} .")
132   print(f"IQR for the Arrival_Delay is {iqr_Departure_Delay:.2f} .")
133   print(f"Any Arrival_Delay < {lower_bound:.2f}  and Arrival_Delay > {upper_bound:.2f}  is an outlier.")
134
135   plt.figure(figsize=(8, 6))
136   plt.boxplot(df['Arrival Delay in Minutes'])
137   plt.xlabel('Arrival_Delay')
138   plt.title('Boxplot of Arrival_Delay')
139
140   plt.show()
141
142   cleaned_df = df[(df['Arrival Delay in Minutes'] >= lower_bound) & (df['Arrival Delay in Minutes'] <= upper_bou
143
144   plt.figure(figsize=(10, 6))
145   plt.boxplot(cleaned_df['Arrival Delay in Minutes'])
146   plt.title('Boxplot-Arrival_Delay-Cleaned')
147   plt.xlabel('Arrival_Delay')
148   plt.grid(True)
149   plt.show()
      Run Cell | Run Above | Debug Cell | Go to [14]
```

```
147   plt.xlabel('Arrival_Delay')
148   plt.grid(True)
149   plt.show()
      Run Cell | Run Above | Debug Cell
150   #%%
151   print(df.head(5))
152   print(df.describe())
153   sns.set_style('whitegrid')
154
      Run Cell | Run Above | Debug Cell | Go to [15]
155   #%%
156   plt.figure(figsize=(8, 6))
157   sns.kdeplot(data=df, x='Age', alpha=0.6, linewidth=2, fill=True)
158   plt.title('Passenger Age Distribution')
159   plt.xlabel('Age')
160   plt.ylabel('Density of the Customers')
161   plt.show()
      Run Cell | Run Above | Debug Cell | Go to [16]
162   # %%
163   gender_counts = df['Gender'].value_counts().reset_index()
164   gender_counts.columns = ['Gender', 'Count']
165
166   colors = {'Male': '#F06292', 'Female': '#9FA8DA', 'Other': 'gray'}
167
168   plt.bar(gender_counts['Gender'], gender_counts['Count'], color=[colors[gender] for gender in gender_co
169   plt.title('Gender Distribution')
170   plt.xlabel('Gender')
171   plt.ylabel('Number of Customers')
172   plt.xticks(rotation=45)
173   plt.grid(axis='y')
174   plt.show()
175
      Run Cell | Run Above | Debug Cell | Go to [17]
176   # %%
```

project.py > ...

```
Run Cell | Run Above | Debug Cell | Go to [17]
# %%
columns = ['Type of Travel', 'Class', 'Customer Type', 'satisfaction']

fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(12, 10))

for col, ax in zip(columns, axes.flatten()):
    sns.countplot(x=col, data=df, ax=ax)
    ax.set_title(f'Count Plot of {col}')
    ax.set_xlabel(col)
    ax.set_ylabel('Count')
    ax.tick_params(axis='x', rotation=45)

plt.tight_layout()
plt.show()

Run Cell | Run Above | Debug Cell | Go to [18]
# %%
color_palette = {'satisfied': '#F06292', 'neutral or dissatisfied': '#9FA8DA', 'Other': 'gray'}

plt.figure(figsize=(8, 6))
df.groupby(['Type of Travel', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_pa
plt.title('Customer Satisfaction according to Type of Travel')
plt.xlabel('Type of Travel')
plt.ylabel('Customer Volume')
plt.legend(title='Satisfaction', bbox_to_anchor=(1, 1))
plt.xticks(rotation=0)
plt.show()
Run Cell | Run Above | Debug Cell | Go to [19]
# %%
plt.figure(figsize=(8, 6))
df.groupby(['Class', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_palette.get
plt.title('Customer Satisfaction based on Class')
plt.xlabel('Class')
```

```
206   plt.xlabel('Class')
207   plt.ylabel('Number of Customers')
208   plt.legend(title='Satisfaction', bbox_to_anchor=(1, 1))
209   plt.xticks(rotation=0)
210   plt.show()
      Run Cell | Run Above | Debug Cell | Go to [20]
211   # %%
212   plt.figure(figsize=(8, 6))
213   df.groupby(['Customer Type', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_pale
214   plt.title('Customer Satisfaction according to Customer Type')
215   plt.xlabel('Customer Type')
216   plt.ylabel('Number of Customers')
217   plt.legend(title='Satisfaction', bbox_to_anchor=(1, 1))
218   plt.xticks(rotation=0)
219   plt.show()
220
      Run Cell | Run Above | Debug Cell | Go to [21]
221   #%%
222
223   color_palette = {'satisfied': '#F06292', 'neutral or dissatisfied': '#9FA8DA', 'Other': 'gray'}
224
225   fig, axs = plt.subplots(1, 3, figsize=(18, 6))
226
227   df.groupby(['Type of Travel', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_pa
228   axs[0].set_title('Customer Satisfaction according to Type of Travel')
229   axs[0].set_xlabel('Type of Travel')
230   axs[0].set_ylabel('Customer Volume')
231   axs[0].legend(title='Satisfaction', bbox_to_anchor=(1, 1))
232   axs[0].tick_params(axis='x', rotation=0)
233
234   df.groupby(['Class', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_palette.get
235   axs[1].set_title('Customer Satisfaction based on Class')
236   axs[1].set_xlabel('Class')
237   axs[1].set_ylabel('Number of Customers')
```

project.py > ...

```
axs[1].legend(title='Satisfaction', bbox_to_anchor=(1, 1))
axs[1].tick_params(axis='x', rotation=0)

df.groupby(['Customer Type', 'satisfaction']).size().unstack().plot(kind='bar', stacked=True, color=[color_pal
axs[2].set_title('Customer Satisfaction according to Customer Type')
axs[2].set_xlabel('Customer Type')
axs[2].set_ylabel('Number of Customers')
axs[2].legend(title='Satisfaction', bbox_to_anchor=(1, 1))
axs[2].tick_params(axis='x', rotation=0)

plt.tight_layout()
plt.show()


Run Cell | Run Above | Debug Cell | Go to [22]
# %%
class_gender_counts = df.groupby(['Class', 'Gender']).size().unstack()

fig, axes = plt.subplots(1, 2, figsize=(15, 5))  # Create subplots for each class

for i, (class_name, class_data) in enumerate(class_gender_counts.items()):
    ax = axes[i]
    ax.pie(class_data, labels=class_data.index, autopct='%1.1f%%', startangle=140)
    ax.set_title(f'Distribution of Gender in {class_name} Class')
    ax.axis('equal')
plt.tight_layout()

plt.show()

Run Cell | Run Above | Debug Cell | Go to [23]
# %%
class_travel_counts = df.groupby(['Class', 'Type of Travel']).size().unstack()

fig, axes = plt.subplots(1, 2, figsize=(15, 5))
```

project.py > ...

```
270
271 for i, (class_name, class_data) in enumerate(class_travel_counts.items()):
272     ax = axes[i] if len(class_travel_counts) > 1 else axes  # Use single axis if only one 'Class' category
273     ax.pie(class_data, labels=class_data.index, autopct='%1.1f%%', startangle=140)
274     ax.set_title(f'Distribution of Type of Travel in {class_name} Class')
275     ax.axis('equal')
276 plt.tight_layout()
277
278 plt.show()
    Run Cell | Run Above | Debug Cell | Go to [24]
279 # %%
280
281 plt.figure(figsize=(10, 6))
282 sns.histplot(data=df, x='Flight Distance', hue='Type of Travel', kde=True, bins=30, stat="count", multiple='sta
283 plt.title('Distribution of Flight Distance by Number of Customers (with Type of Travel)')
284 plt.xlabel('Flight Distance')
285 plt.ylabel('Number of Occurrences')
286 plt.legend(title='Type of Travel', labels=['Personal Travel', 'Business Travel'])
287
288 plt.show()
289
290
    Run Cell | Run Above | Debug Cell | Go to [25]
291 # %%
292 plt.figure(figsize=(10, 6))
293 sns.boxplot(data=df, x='Class', y='Seat comfort')
294 plt.title('Seat Comfort Distribution by Class (Box Plot)')
295 plt.xlabel('Class')
296 plt.ylabel('Seat Comfort')
297 plt.show()
298
299 plt.figure(figsize=(10, 6))
300 sns.violinplot(data=df, x='Class', y='Seat comfort')
301 plt.title('Seat Comfort Distribution by Class (Violin Plot)')
```

```
project.py > ...
    plt.xlabel('Class')
    plt.ylabel('Seat Comfort')
    plt.show()

    Run Cell | Run Above | Debug Cell | Go to [26]
Run and Debug (Ctrl+Shift+D)
    aspects = ['Inflight wifi service', 'Cleanliness', 'Food and drink', 'Checkin service', 'Seat comfort', 'Leg r
    titles = ['In-flight WiFi Service', 'Cleanliness', 'Food and Drinks', 'Check-in Service', 'Seat Comfort', 'Leg

    fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(15, 10))

    axes = axes.flatten()

    for i, aspect in enumerate(aspects):
        sns.violinplot(ax=axes[i], data=df, x='satisfaction', y=aspect, hue='satisfaction')
        axes[i].set_title(titles[i])
        axes[i].set_xlabel('Overall Satisfaction')
        axes[i].set_ylabel(aspect)
        axes[i].legend(title='Satisfaction')

    plt.tight_layout()
    plt.show()
    Run Cell | Run Above | Debug Cell | Go to [27]
    # %%

    plt.figure(figsize=(10, 6))
    sns.countplot(data=df, x='Customer Type', hue='satisfaction', palette='Set2')
    plt.title('Count of Checkin Service Satisfaction by Type of Customer')
    plt.xlabel('Type of Customer')
    plt.ylabel('Count')
    plt.legend(title='Satisfaction', bbox_to_anchor=(1, 1))

    plt.show()
```

```

    Run Cell | Run Above | Debug Cell | Go to [28]
    # %%
    plt.figure(figsize=(8, 6))
    sns.countplot(data=df, x='Type of Travel', hue='satisfaction')
    plt.title('Count Plot: Type of Travel')
    plt.xlabel('Type of Travel')
    plt.ylabel('Count')
    plt.legend(title='Satisfaction')
    plt.show()
    Run Cell | Run Above | Debug Cell | Go to [29]
    # %%

    selected_columns = ['Age', 'Gender', 'Flight Distance']
    filtered_df = df[selected_columns]

    g = sns.jointplot(data=filtered_df, x='Age', y='Flight Distance', hue='Gender', kind='scatter', palette='Set2')
    plt.title('Joint Plot: Age vs Flight Distance with Gender Hue')
    plt.show()


    Run Cell | Run Above | Debug Cell | Go to [30]
    # %%
    grouped_data = df.groupby(['Gender', 'satisfaction']).size().unstack()

    grouped_data_norm = grouped_data.div(grouped_data.sum(axis=1), axis=0)

    plt.figure(figsize=(8, 6))
    grouped_data.plot(kind='bar', stacked=True, color=['#F06292', '#9FA8DA'])
    plt.title('Gender vs Satisfaction')
    plt.xlabel('Gender')
    plt.ylabel('Count')
    plt.legend(title='Satisfaction')
```

```
plt.show()

Run Cell | Run Above | Debug Cell | Go to [31]
# %%

columns = ['Age', 'Flight Distance']

fig, axes = plt.subplots(nrows=len(columns), ncols=1, figsize=(8, 6 * len(columns)))

for idx, col in enumerate(columns):
    count_data = df[col].value_counts().reset_index()
    count_data.columns = [col, 'Count']

    sns.kdeplot(data=count_data, x=col, y='Count', fill=True, alpha=0.6, ax=axes[idx], linewidth=1)
    axes[idx].set_title(f'{col} Distribution by Count')
    axes[idx].set_xlabel(col)
    axes[idx].set_ylabel('Count Density')

plt.tight_layout()
plt.show()



Run Cell | Run Above | Debug Cell | Go to [32]
# %%
plt.figure(figsize=(10, 6))
sns.lineplot(data=df, x='Age', y='Flight Distance', hue='satisfaction', linewidth=3)
plt.title('Line Plot: Age vs Flight Distance')
plt.xlabel('Age')
plt.ylabel('Flight Distance')
plt.legend(title='Satisfaction')
plt.show()
```

```
Run Cell | Run Above | Debug Cell | Go to [33]
#%%
plt.figure(figsize=(8, 6))
stats.probplot(df['Flight Distance'], dist="norm", plot=plt)
plt.title('QQ Plot: Flight Distance')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Flight Distance')
plt.show()

Run Cell | Run Above | Debug Cell | Go to [34]
#%%
plt.figure(figsize=(8, 6))
stats.probplot(df['Departure Delay in Minutes'], dist="norm", plot=plt)
plt.title('QQ Plot: Departure Delay in Minutes')
plt.xlabel('Theoretical Quantiles')
plt.ylabel('Departure Delay in Minutes')
plt.show()
Run Cell | Run Above | Debug Cell | Go to [35]
# %%
plt.figure(figsize=(10, 6))
sns.boxplot(data=df[['Online boarding', 'Gate location', 'Ease of Online booking', 'Baggage handling']], orient
plt.title('Multivariate Box Plot')
plt.xlabel('Rating')
plt.ylabel('Features')
plt.show()

Run Cell | Run Above | Debug Cell | Go to [36]
# %%

features = ['Online boarding', 'Gate location', 'Ease of Online booking', 'Baggage handling', 'satisfaction']

melted_df = df[features].melt(id_vars=['satisfaction'], var_name='Feature', value_name='Rating')

sns.violinplot(data=melted_df, x='Rating', y='Feature', hue='satisfaction', split=True)
plt.title('Violin Plot with Satisfaction')
plt.xlabel('Rating')
plt.ylabel('Features')
plt.legend(title='Satisfaction', loc='lower right')
plt.show()


Run Cell | Run Above | Debug Cell | Go to [37]
# %%
plt.figure(figsize=(10, 6))
df[['Inflight wifi service', 'Cleanliness']].plot.area(stacked=False, alpha=0.5)
plt.title('Area Plot: Satisfaction Ratings')
plt.xlabel('Observations')
plt.ylabel('Rating')
plt.legend(title='Feature' )
plt.show()

Run Cell | Run Above | Debug Cell | Go to [38]
#%%
from mpl_toolkits.mplot3d import Axes3D

fig = plt.figure(figsize=(10, 8))
ax = fig.add_subplot(111, projection='3d')

scatter = ax.scatter(df['Age'], df['Flight Distance'], df['Seat comfort'], c=df['satisfaction'].cat.codes, cma

legend_labels = df['satisfaction'].cat.categories.tolist()
legend_handles = [plt.Line2D([0], [0], marker='o', color='w', markerfacecolor=scatter.cmap(scatter.norm(level)
ax.legend(handles=legend_handles, title='satisfaction')

ax.set_xlabel('Age')
ax.set_ylabel('Flight Distance')
```

```
459
460 plt.show()
461
    Run Cell | Run Above | Debug Cell | Go to [39]
462 # %%
463 plt.figure(figsize=(10, 6))
464 plt.hexbin(x=df['Age'], y=df['Flight Distance'], gridsize=30, cmap='inferno')
465 plt.colorbar(label='count in bin')
466 plt.title('Hexbin Plot: Age vs Flight Distance')
467 plt.xlabel('Age')
468 plt.ylabel('Flight Distance')
469 plt.show()
    Run Cell | Run Above | Debug Cell | Go to [40]
470 # %%
471 custom_palette = {'satisfied': 'green', 'neutral or dissatisfied': 'orange', 'Other': 'gray'}
472 plt.figure(figsize=(10, 6))
473 sns.stripplot(data=df, x='Type of Travel', y='Flight Distance', hue='satisfaction', palette=custom_pale
474 plt.title('Strip Plot: Type of Travel vs Flight Distance')
475 plt.xlabel('Type of Travel')
476 plt.ylabel('Flight Distance')
477 plt.legend(title='Satisfaction')
478 plt.show()
479
    Run Cell | Run Above | Debug Cell | Go to [41]
480 #%%
481 plt.figure(figsize=(10, 6))
482 sns.stripplot(data=df, x='Type of Travel', y='Departure Delay in Minutes', hue='satisfaction', palette
483 plt.title('Strip Plot: Type of Travel vs Departure Delay in Minutes')
484 plt.xlabel('Type of Travel')
485 plt.ylabel('Departure Delay (Minutes)')
486 plt.legend(title='Satisfaction')
487 plt.show()
488
```

```
488
    Run Cell | Run Above | Debug Cell | Go to [42]
489 #%%
490 plt.figure(figsize=(10, 6))
491 sns.stripplot(data=df, x='Type of Travel', y='Arrival Delay in Minutes', hue='satisfaction', palette=custom_p
492 plt.title('Strip Plot: Type of Travel vs Arrival Delay in Minutes')
493 plt.xlabel('Type of Travel')
494 plt.ylabel('Arrival Delay (Minutes)')
495 plt.legend(title='Satisfaction')
496 plt.show()
497
    Run Cell | Run Above | Debug Cell | Go to [43]
498 #%%
499
500 fig, axes = plt.subplots(1, 3, figsize=(18, 6))
501
502 sns.stripplot(data=df, x='Type of Travel', y='Flight Distance', hue='satisfaction', palette=custom_palette, j
503 axes[0].set_title('Type of Travel vs Flight Distance')
504 axes[0].set_xlabel('Type of Travel')
505 axes[0].set_ylabel('Flight Distance')
506 axes[0].legend(title='Satisfaction')
507
508 sns.stripplot(data=df, x='Type of Travel', y='Departure Delay in Minutes', hue='satisfaction', palette=custom
509 axes[1].set_title('Type of Travel vs Departure Delay in Minutes')
510 axes[1].set_xlabel('Type of Travel')
511 axes[1].set_ylabel('Departure Delay (Minutes)')
512 axes[1].legend(title='Satisfaction')
513
514 sns.stripplot(data=df, x='Type of Travel', y='Arrival Delay in Minutes', hue='satisfaction', palette=custom_p
515 axes[2].set_title('Type of Travel vs Arrival Delay in Minutes')
516 axes[2].set_xlabel('Type of Travel')
517 axes[2].set_ylabel('Arrival Delay (Minutes)')
518 axes[2].legend(title='Satisfaction')
```

```
519
520 plt.tight_layout()
521 plt.show()
522
523
    Run Cell | Run Above | Debug Cell | Go to [44]
524 #%%
525 categories = ['Inflight service', 'Cleanliness', 'Departure/Arrival time convenient']
526 values_satisfied = [df[df['satisfaction'] == 'satisfied'][cat].mean() for cat in categories]
527 values_neutral_dissatisfied = [df[df['satisfaction'] == 'neutral or dissatisfied'][cat].mean() for cat in cate
528
529 N = len(categories)
530 angles = np.linspace(0, 2 * np.pi, N, endpoint=False).tolist()
531
532 values_satisfied += values_satisfied[:1]
533 values_neutral_dissatisfied += values_neutral_dissatisfied[:1]
534 angles += angles[:1]
535
536 fig, ax = plt.subplots(figsize=(8, 8), subplot_kw=dict(polar=True))
537 ax.fill(angles, values_satisfied, color='lightblue', alpha=0.5)
538 ax.set_yticklabels([])
539 ax.set_xticks(angles[:-1])
540 ax.set_xticklabels(categories, fontsize=12)
541 ax.set_title('Satisfaction Levels - Satisfied Customers', fontsize=14)
542 ax.grid(True)
543
544 fig, ax = plt.subplots(figsize=(8, 8), subplot_kw=dict(polar=True))
545 ax.fill(angles, values_neutral_dissatisfied, color='lightpink', alpha=0.5)
546 ax.set_yticklabels([])
547 ax.set_xticks(angles[:-1])
548 ax.set_xticklabels(categories, fontsize=12)
549 ax.set_title('Satisfaction Levels - Neutral or Dissatisfied Customers', fontsize=14)
550 ax.grid(True)
```

```
551
552 plt.show()
    Run Cell | Run Above | Debug Cell | Go to [45]
553 #%%
554 selected_columns = ['Age', 'Flight Distance', 'Inflight wifi service', 'Food and drink', 'Seat comfort']
555 sns.pairplot(df[selected_columns])
556 plt.suptitle('Pair Plot of Selected Columns')
557 plt.show()
    Run Cell | Run Above | Debug Cell | Go to [46]
558 # %%
559 sns.lmplot(x='Age', y='Arrival Delay in Minutes', data=df, scatter_kws={'alpha': 0.3}, line_kws={'color': 'gree
560 plt.title('Regression Plot of Age vs Arrival Delay in Minutes')
561 plt.xlabel('Age')
562 plt.ylabel('Arrival Delay in Minutes')
563 plt.show()
564
565
    Run Cell | Run Above | Debug Cell | Go to [47]
566 # %%
567 sns.rugplot(df['Arrival Delay in Minutes'], color='purple')
568 plt.title('Rug Plot of Arrival Delay in Minutes')
569 plt.xlabel('Arrival Delay in Minutes')
570 plt.show()
571
    Run Cell | Run Above | Debug Cell | Go to [48]
572 # %%
573 features=df.drop(columns=['Gender', 'Customer Type', 'Type of Travel', 'Class', 'satisfaction', 'Age.cat', 'Age
574 scaler=StandardScaler()
575 scaled_f=scaler.fit_transform(features)
576 print(np.round(scaled_f, 2))
    Run Cell | Run Above | Debug Cell | Go to [49]
577 # %%
578 cm=pd.DataFrame(scaled_f, columns=features.columns).corr()
579 plt.figure(figsize=(10,8))
```

```
583    u, s, v=svd(scaled_f, full_matrices=False)
584    cond_num=cond(scaled_f)
585    print('singular values are')
586    print(np.round(s,decimals=2))
587    print('conditional number is ')
588    print(np.round(cond_num,decimals=2))
       Run Cell | Run Above | Debug Cell | Go to [51]
589    # %%
590    pca=PCA(n_components=0.95)
591    pca.fit(scaled_f)
592    explained_variance=pca.explained_variance_ratio_
593    print(np.round(explained_variance, 2))
594    print('\n')
595
596
597    components=pca.n_components_
598    features_to_be_removed=scaled_f.shape[1]-components
599    print('the number of features to be removed are', features_to_be_removed)
       Run Cell | Run Above | Debug Cell | Go to [52]
600    # %%
601    transformed_f=pca.transform(scaled_f)
602    u_reduced, s_reduced, v_reduced=svd(transformed_f, full_matrices=False)
603    print('the new singular values are')
604    print(np.round(s_reduced, 2))
605    print('\n')
606    cond_num_reduced=cond(transformed_f)
607    print('the new conditional number is')
608    print(np.round(cond_num_reduced, 2))
       Run Cell | Run Above | Debug Cell | Go to [53]
609    # %%
610    cm_reduced=pd.DataFrame(transformed_f).corr()
611    plt.figure(figsize=(10,8))
612    sns.heatmap(cm_reduced, annot=True, fmt='.2f')
613    plt.show()
```

```
plt.show()

Run Cell | Run Above | Debug Cell | Go to [54]
# %%
cumulative_explained_variance_ratio = np.cumsum(pca.explained_variance_ratio_)

plt.figure(figsize=(10, 6))
plt.plot(range(1, len(cumulative_explained_variance_ratio) + 1), cumulative_explained_variance_ratio, marker='
plt.xlabel('Number of Components')
plt.ylabel('Cumulative Explained Variance Ratio')
plt.title('PCA Cumulative Explained Variance Ratio')
plt.grid(True)
plt.show()
Run Cell | Run Above | Debug Cell | Go to [55]
# %%
from scipy.stats import shapiro

feature_names = features.columns.tolist()

shapiro_results = {}
for col in range(transformed_f.shape[1]):
    feature_name = feature_names[col]
    shapiro_stat, shapiro_pvalue = shapiro(transformed_f[:, col])
    shapiro_results[feature_name] = {
        'Shapiro Statistic': round(shapiro_stat, 2),
        'p-value': round(shapiro_pvalue, 2),
        'Normal (p > 0.05)': shapiro_pvalue > 0.05
    }

for feature, result in shapiro_results.items():
    print(f"{feature}: Shapiro Statistic = {result['Shapiro Statistic']}, p-value = {result['p-value']}, Norma
```

```
Run Cell | Run Above | Debug Cell | Go to [56]
# %%
scaled_f_df = pd.DataFrame(scaled_f)
sns.clustermap(scaled_f_df.corr(), cmap='coolwarm', linewidths=0.5)
plt.title('Cluster Map of Correlation Matrix')
plt.show()

Run Cell | Run Above | Debug Cell
# %%
```

REFERENCES

1. Kimes, S. E., & Wirtz, J. (2003). Has Revenue Management Become Acceptable? Findings from an International Study on the Perceived Fairness of Rate Fences. Journal of Service Research, 6(2), 125–135. doi: 10.1177/1094670503257031.
2. Lovelock, C. H., & Wirtz, J. (2011). Services Marketing: People, Technology, Strategy. 7th ed. Prentice Hall. This book provides a comprehensive look into service marketing and management with relevant applications to airline services.
3. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. This text offers a detailed explanation of various statistical methods that can be applied to data analysis in the context of customer satisfaction and predictive modeling.
4. McKinney, W. (2017). Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython. O'Reilly Media. An essential guide for data analysis in Python, including the use of libraries such as Pandas and Seaborn, which can be employed for creating a dashboard and analyzing airline passenger satisfaction data.
5. Oliver, R. L. (2010). Satisfaction: A Behavioral Perspective on the Consumer. 2nd ed. M.E. Sharpe. This book provides an in-depth discussion of consumer satisfaction, how it can be measured, and the implications of service quality, which are critical to understanding airline passenger satisfaction.