CS 773: Data Mining and Security

Summer 2024

Due: August 1, 2024

Author Name: Ashish Verma

Date: 08/01/2024

Table of Contents

Introduction

Airline passenger satisfaction is a critical metric for the aviation industry, influencing customer loyalty, brand reputation, and overall airline performance. The Airline Passenger Satisfaction Survey data from Kaggle provides comprehensive insights into various factors affecting passenger satisfaction, including service quality, convenience, and comfort. This dataset is instrumental in understanding the key drivers of satisfaction and identifying areas for improvement.

Purpose and Objectives of the Analysis

The primary purpose of this analysis is to evaluate and understand the factors that influence airline passenger satisfaction. By leveraging the dataset, the analysis aims to:

1. **Identify Key Drivers of Satisfaction:** Determine which factors, such as service quality, comfort, and convenience, most significantly impact overall passenger satisfaction.
2. **Compare Service Aspects:** Analyze and compare different service aspects like in-flight service quality, check-in service, baggage handling, and departure/arrival time convenience.
3. **Customer Segmentation:** Segment customers into loyal and disloyal categories and explore the differences in satisfaction levels and service perceptions between these groups.
4. **Impact of Flight Characteristics:** Assess how flight characteristics such as travel type, class, and flight distance affect passenger satisfaction.
5. **Actionable Insights:** Provide actionable insights and recommendations for airlines to enhance passenger satisfaction and improve service quality.

Scope of the Report

The scope of this report includes:

1. **Data Overview:** A brief description of the dataset, including its structure, key variables, and any preprocessing steps taken.
2. **Descriptive Analysis:** Summary statistics and visualizations to provide an initial understanding of the data and satisfaction trends.
3. **Factor Analysis:** Detailed examination of various factors influencing satisfaction, using statistical methods to identify significant predictors.
4. **Comparative Analysis:** Comparative analysis of service aspects, focusing on in-flight service quality, check-in service, baggage handling, and departure/arrival time convenience.

5. **Customer Segmentation Analysis:** Analysis of satisfaction levels and service perceptions across different customer segments (loyal vs. disloyal).
6. **Flight Characteristics Analysis:** Evaluation of the impact of flight characteristics such as travel type, class, and flight distance on satisfaction.
7. **Recommendations:** Actionable recommendations for airlines to improve passenger satisfaction based on the analysis findings.
8. **Appendix:** Detailed results, including statistical models, detailed runs, and any supplementary analyses.

Methodology

Attribute Information:

- Gender: Gender of the passengers (Female, Male)
- Customer Type: The customer type (Loyal customer, disloyal customer)
- Age: The actual age of the passengers
- Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
- Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
- Flight distance: The flight distance of this journey
- Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
- Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
- Ease of Online booking: Satisfaction level of online booking
- Gate location: Satisfaction level of Gate location
- Food and drink: Satisfaction level of Food and drink
- Online boarding: Satisfaction level of online boarding
- Seat comfort: Satisfaction level of Seat comfort
- Inflight entertainment: Satisfaction level of inflight entertainment
- On-board service: Satisfaction level of On-board service
- Leg room service: Satisfaction level of Leg room service
- Baggage handling: Satisfaction level of baggage handling
- Check-in service: Satisfaction level of Check-in service
- Inflight service: Satisfaction level of inflight service
- Cleanliness: Satisfaction level of Cleanliness
- Departure Delay in Minutes: Minutes delayed when departure
- Arrival Delay in Minutes: Minutes delayed when Arrival

Label:

- Satisfaction: Airline satisfaction level(Satisfaction, neutral or dissatisfaction)

Dataset link :

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Tools and Techniques Used

Tools:

- **Python:** The primary programming language used for data analysis.
- **Pandas:** For data manipulation and preprocessing.
- **NumPy:** For numerical operations.
- **Matplotlib and Seaborn:** For data visualization.
- **Scikit-learn:** For statistical analysis and machine learning modeling.
- **Statsmodels:** For advanced statistical modeling.
- **Jupyter Notebook:** For interactive data analysis and visualization.

Techniques:

- **Descriptive Statistics:** To summarize and describe the main features of the dataset.
- **Data Visualization:** To graphically represent data patterns and relationships.
- **Correlation Analysis:** To identify relationships between different variables.
- **Regression Analysis:** To quantify the impact of various factors on overall satisfaction.
- **Data Mining Techniques:-** To answer the questions from 1 to 8.
- **Machine Learning Models:** To predict satisfaction levels and classify customer types.
  - **Logistic Regression:** For binary classification of customer types (loyal vs. disloyal).
  - **Random Forest:** For feature importance and classification tasks.

Explanation of the Analysis Approach

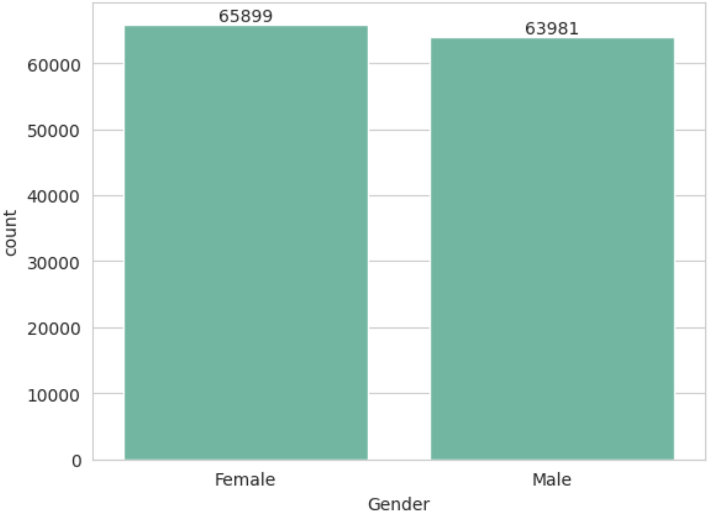
The analysis approach is structured into several key steps:

a. **Data Preprocessing**
b. **Exploratory Data Analysis (EDA)**
c. **Correlation Analysis**
d. **Regression Analysis**
e. **Machine Learning Modeling**
f. **Comparative Analysis**
g. **Interpretation and Recommendations**

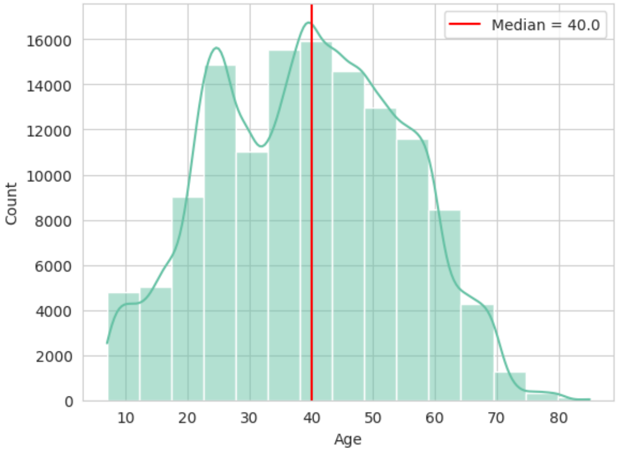
Exploratory Data Analysis

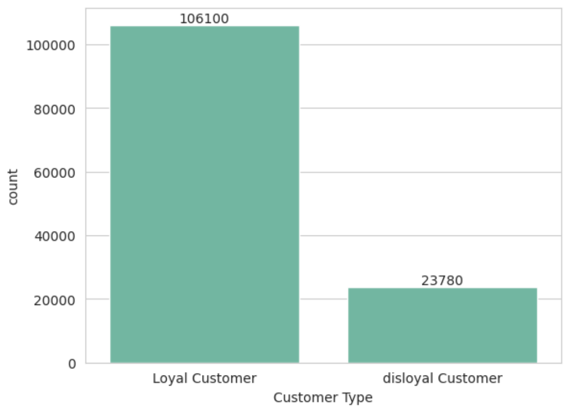
Univariate Analysis

- Gender seemed to be Equal in data.

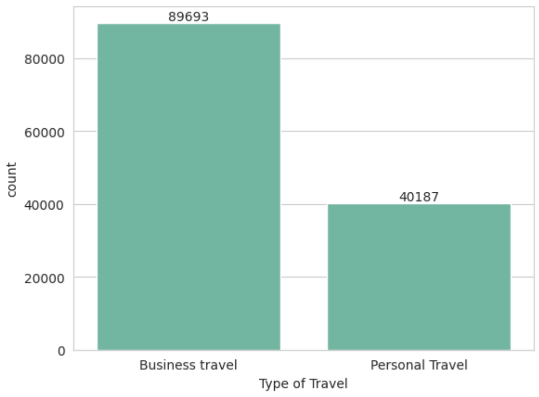


- Age Distribution :
    - Ages has Normal distribution
    - Average Ages is 40 years old

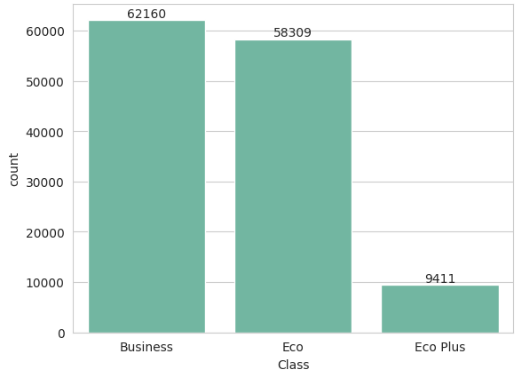




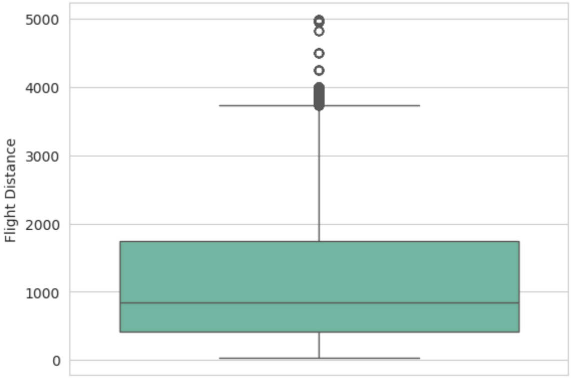
- Most Passengers are Returning, so they have experienced the services before.
- Most common Type of Travel is Business.



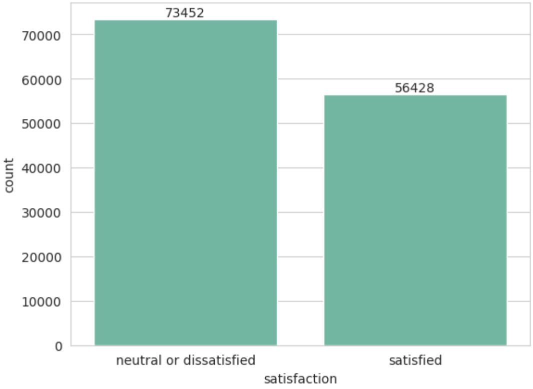
- Most passengers in Business Class but fewer of them in Economy Plus.



- Majority of Flights are under 1000 km.



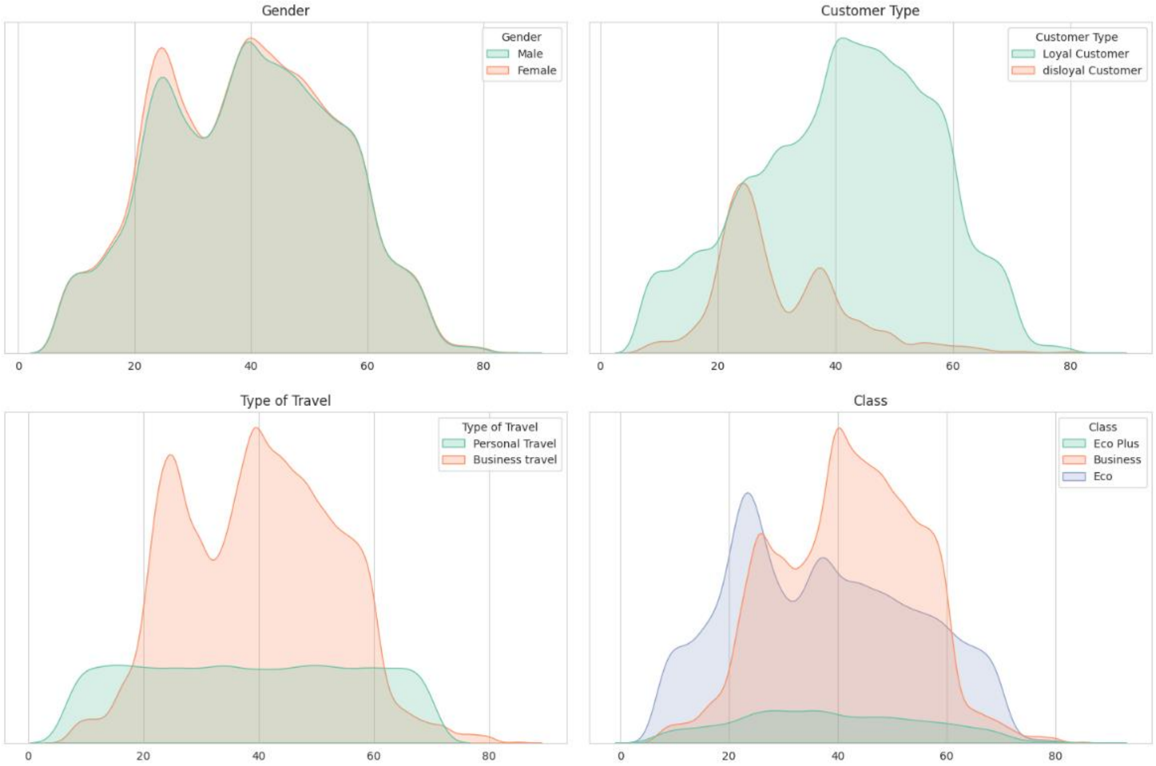
-
- Services got good ratings are : In-flight Service, Baggage Handling, Seat Comfort
- Services got poor rating : In-flight WIFI Service, Ease of Online Booking, Gate Location
- Majority of Passengers Neutral or Dissatisfied.



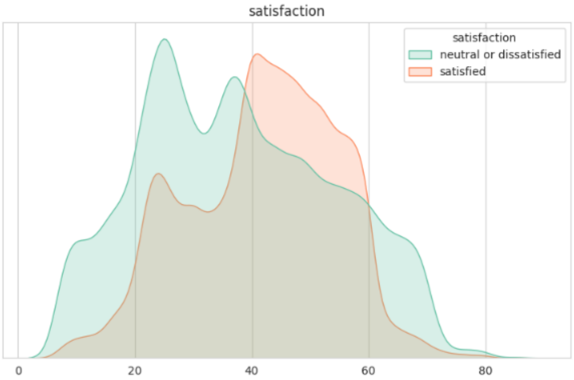
-

Multivariate Analysis

- Gender seemed to be Equal in data.
- Age Distribution :
  - Ages has Normal distribution
  - Average Ages is 40 years old



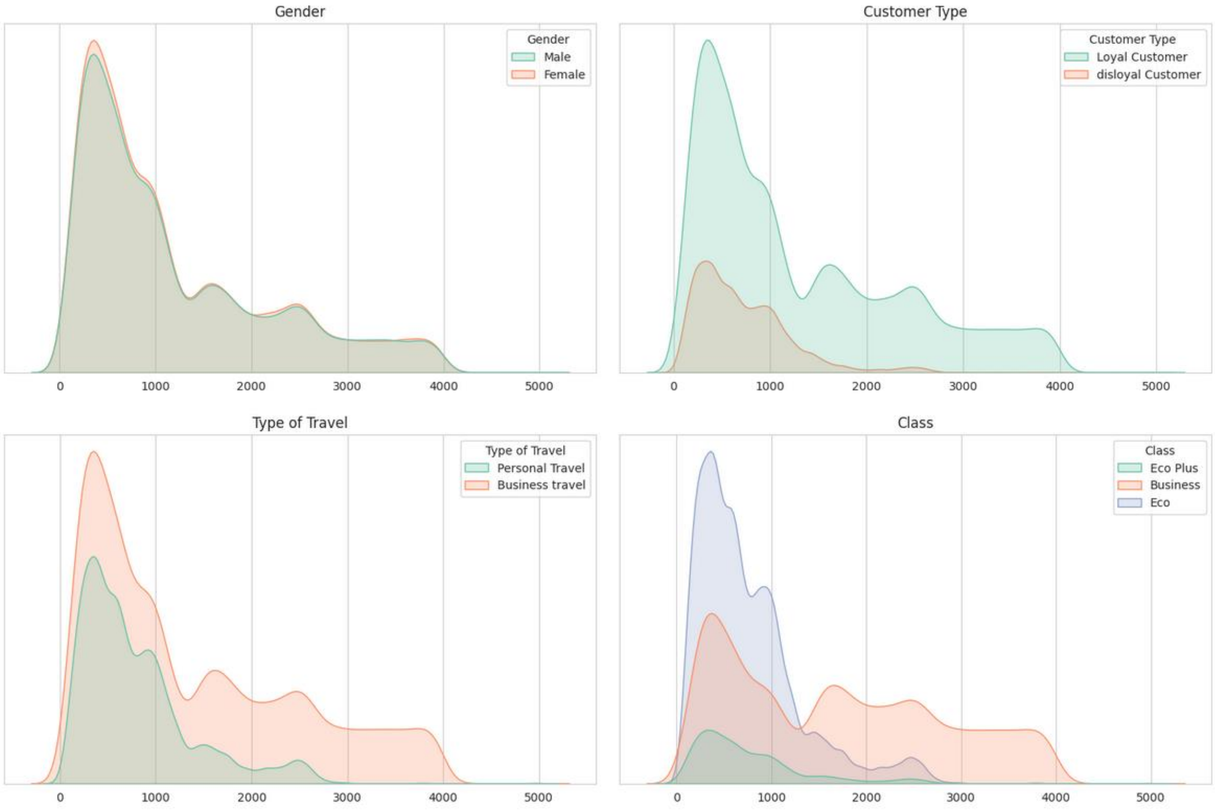
- Most Passengers are Returning, so they have experienced the services before.
- Most common Type of Travel is Business.
- Most passengers in Business Class but fewer of them in Economy Plus.
- Majority of Flights are under 1000 km.
- Services got good ratings are :
  - In-flight Service, Baggage Handling, Seat Comfort
  - Services got poor rating :
  - In-flight WIFI Service, Ease of Online Booking, Gate Location
- Majority of Passengers Neutral or Dissatisfied.



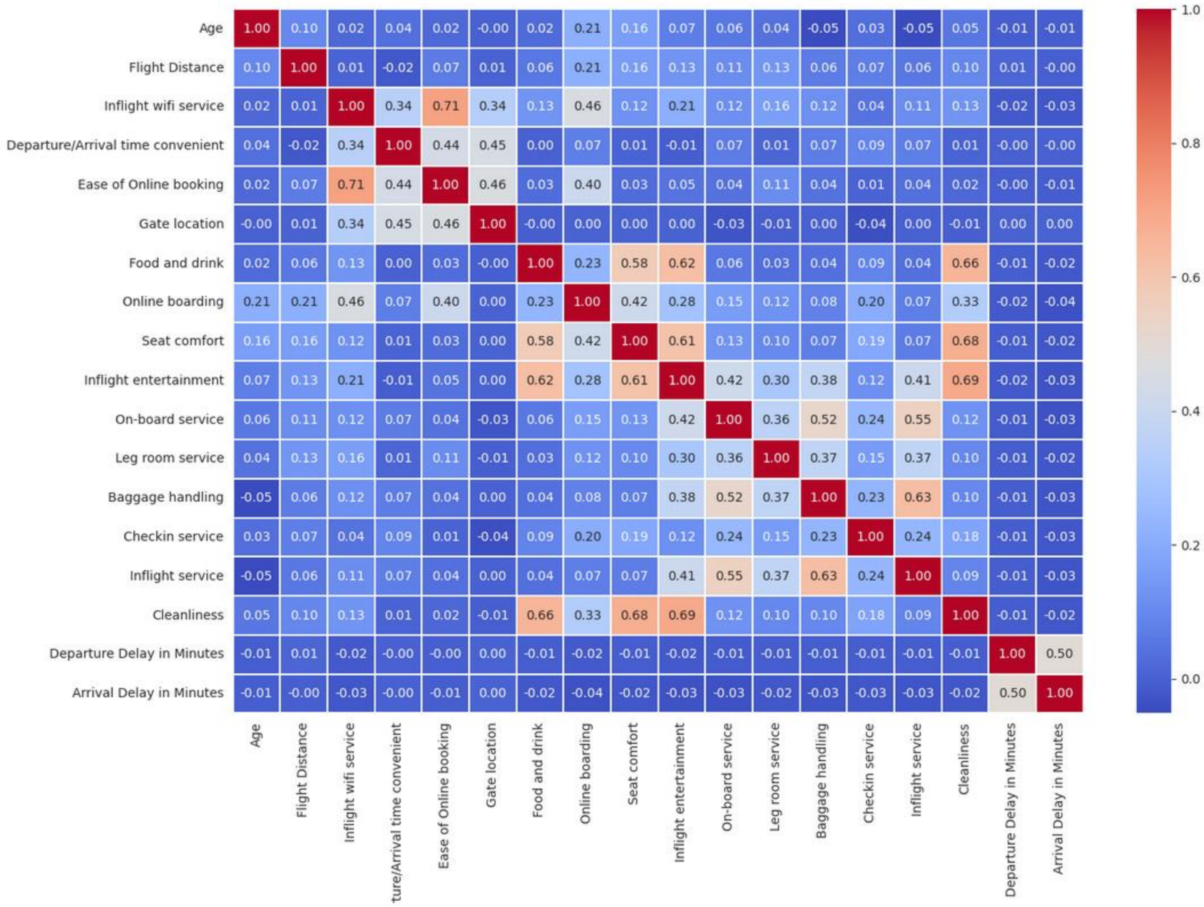
- Age of passengers is Equally distributed in the data

- Passengers of older ages:
    - Returning
    - Business Class
    - Satisfied
- Younger passengers:
    - First time
    - Economy Class
    - Dissatisfied

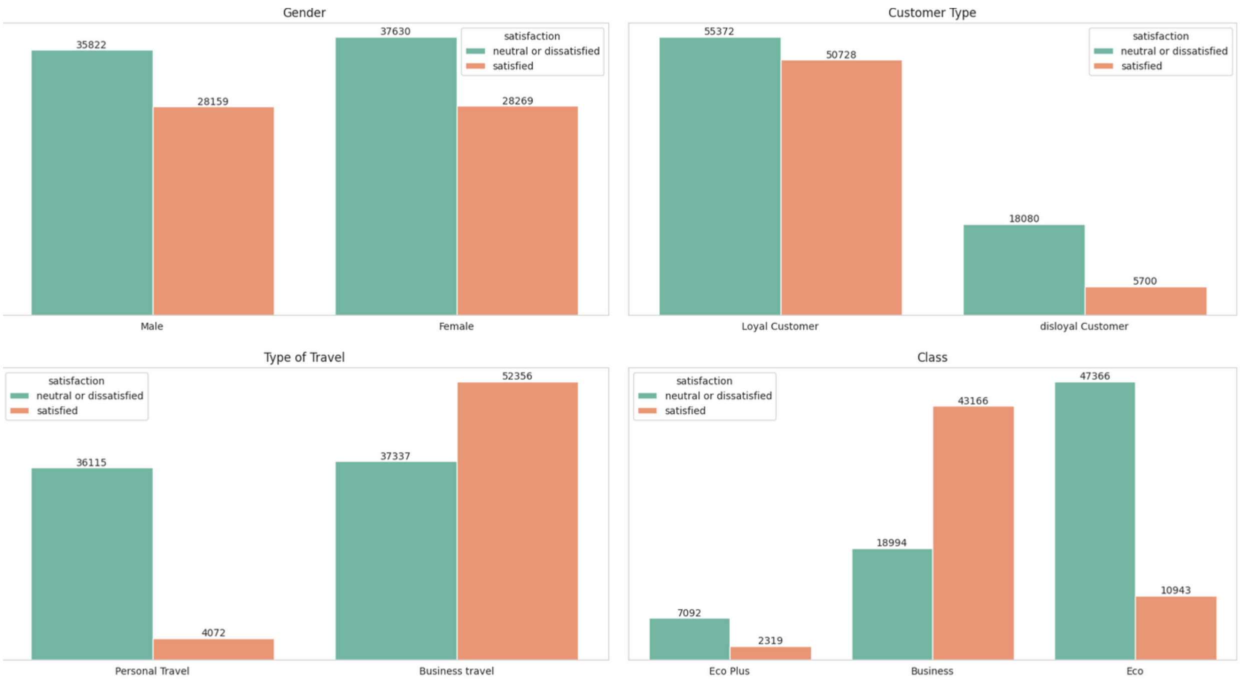
- Traveling long Distance:
    - Returning Customers
    - Business type of travel
    - Business Class
    - Satisfied Passengers



- Gender has same distribution for Km Traveling
- Gender is almost equal for men and women, whether they are satisfied or not
- Returning Customer type is almost Equally, but First time Customers not satisfied at all.



- Business Type of Travel is More Satisfied, but Personal type which mostly not satisfied.
- Business Class is More Satisfied, but Economy & Eco Plus Class which mostly not satisfied.

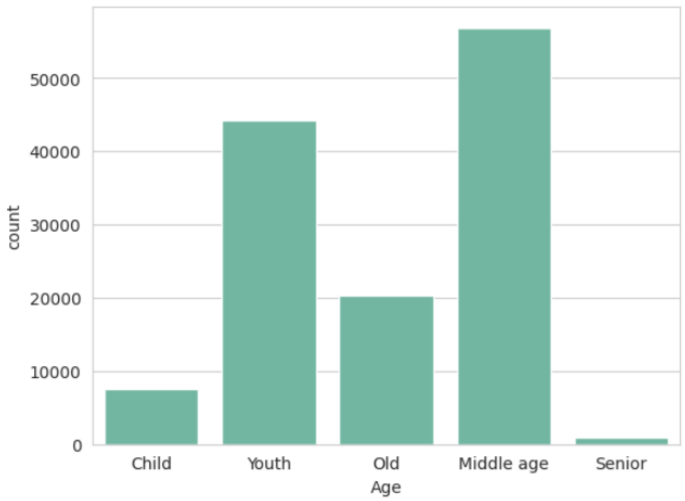

Gender
35822
28159
37630
28269
satisfaction
neutral or dissatisfied
satisfied
Male
Female
Customer Type
55372
50728
18080
5700
satisfaction
neutral or dissatisfied
satisfied
Loyal Customer
disloyal Customer
Type of Travel
satisfaction
neutral or dissatisfied
satisfied
36115
4072
37337
52356
Personal Travel
Business travel
Class
satisfaction
neutral or dissatisfied
satisfied
7092
2319
18994
43166
47366
10943
Eco Plus
Business
Eco

Statistics on departure delay (A8) and arrival delay (A9).

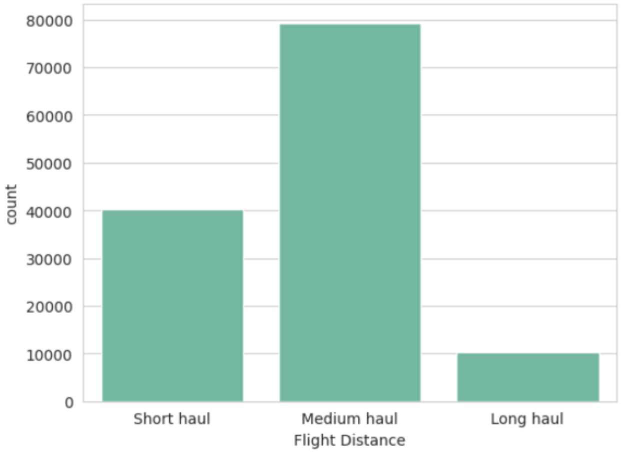
- **Central Tendency**: Departure and arrival delays have certain average values, with their respective modes and medians indicating the most common and central values.
- **Spread**: The standard deviation shows how much the delays vary around the mean.
- **Percentiles and Quartiles**: These measures provide insights into the distribution of delays, such as how many delays are below certain thresholds.
- **Skewness**: The skewness indicates whether the delays are more often lower or higher than the average.
- **Covariance and Correlation**: These statistics show the relationship between departure and arrival delays, indicating whether they tend to increase or decrease together.
- **Distributions**: The plots visually display how the delays are distributed, helping to identify patterns or outliers.

Convert numerical values to categorical values

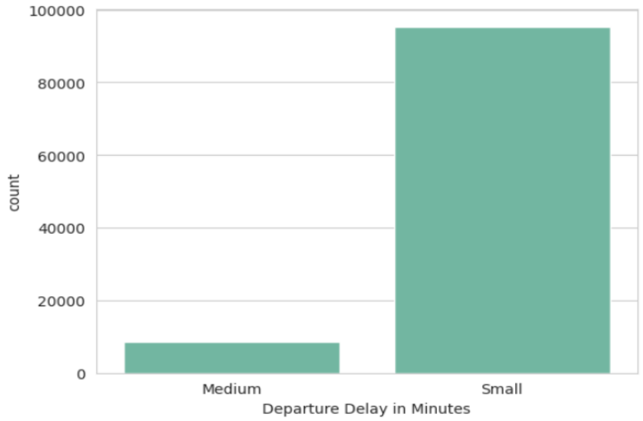
- There are more travelers whose are middle aged than younger and older people.



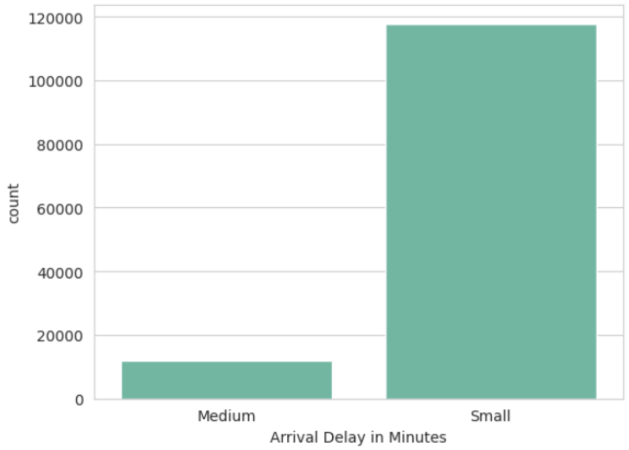
- Flight distance for medium haul have more numbers.



- Departure delay is very small for large number of flights



- Arrival delay is also very small for large number of flights.

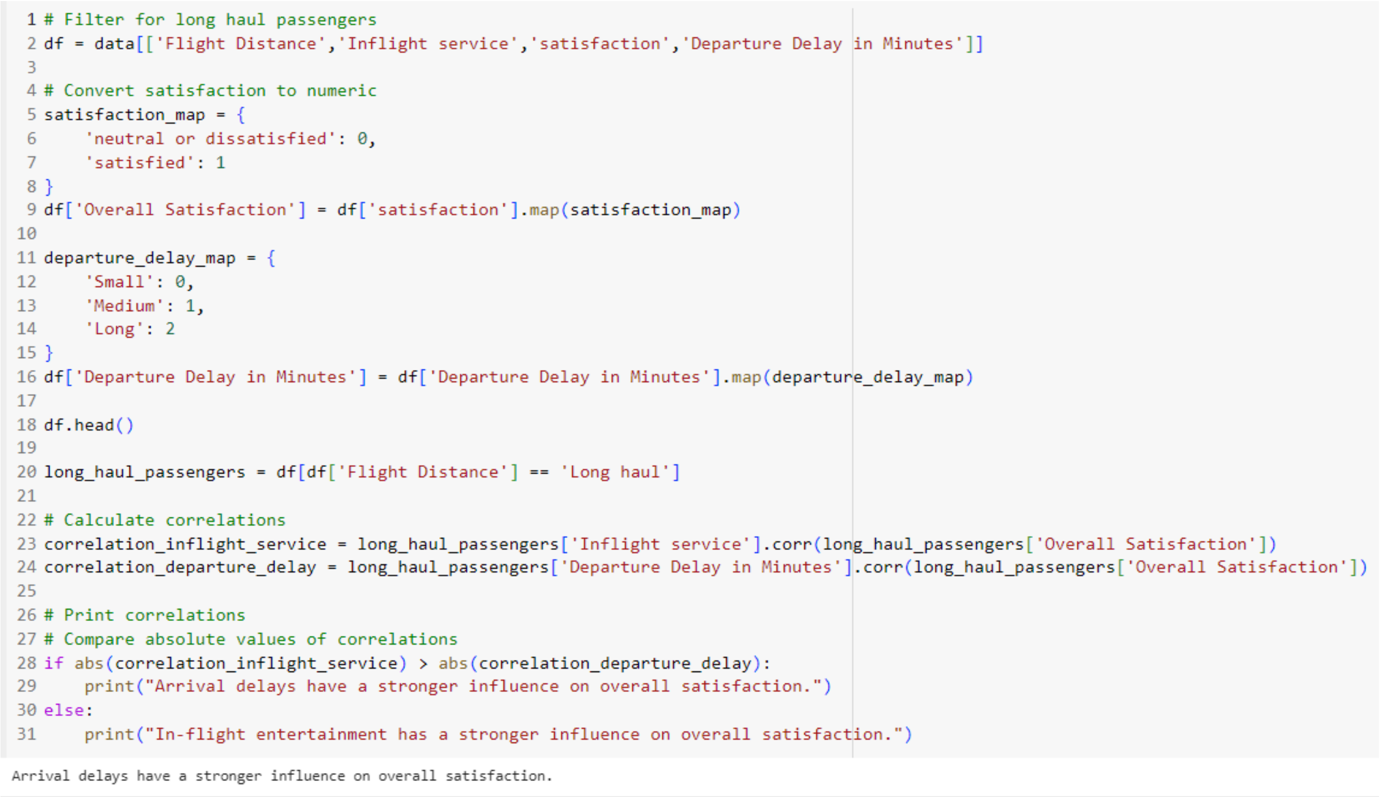


Test the following two hypotheses. Show evidence to show whether they are true or false.

Hypothesis

Hypothesis 1 (Long Haul Passengers): Long haul passengers' overall satisfaction is more strongly influenced by in-flight service quality than by departure delays.

Result:- Arrival delays have a stronger influence on overall satisfaction.

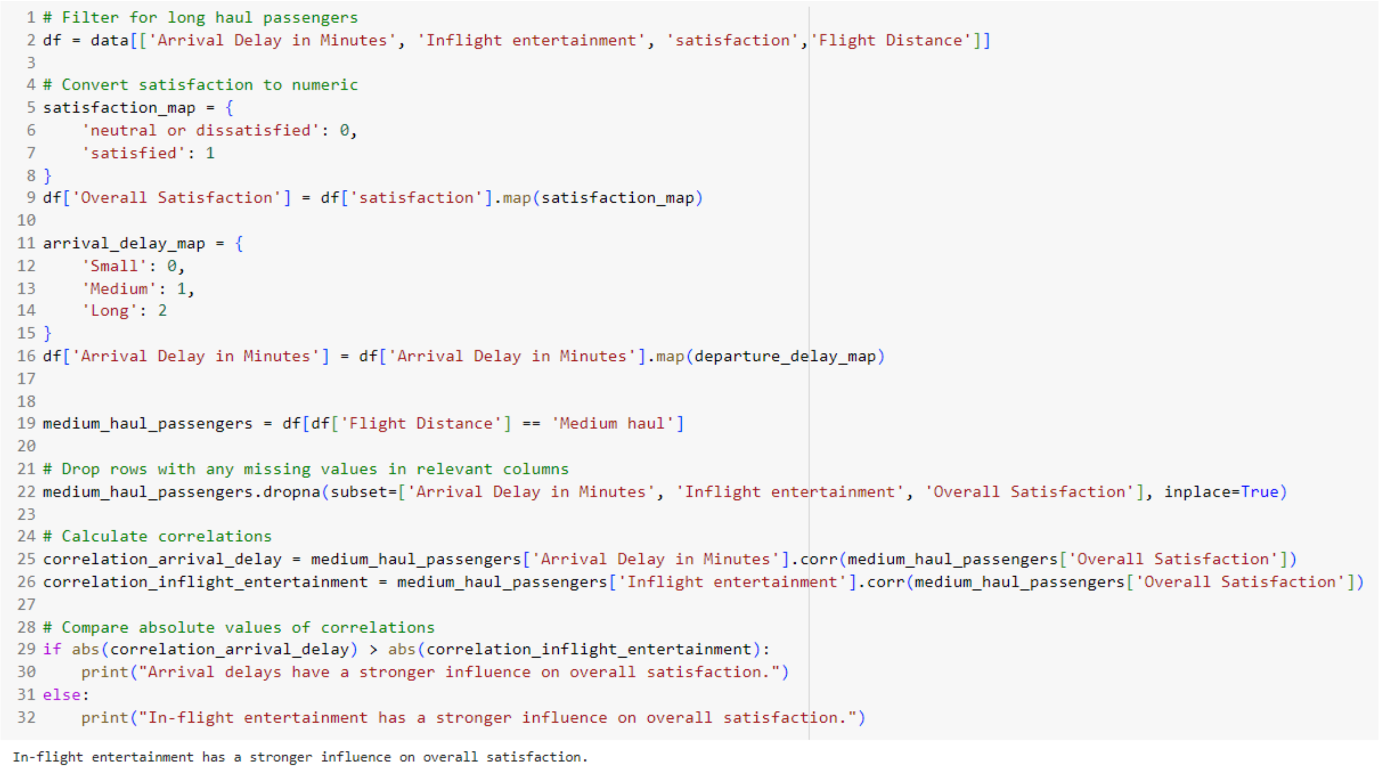


```python
# Filter for long haul passengers
df = data[['Flight Distance','Inflight service','satisfaction','Departure Delay in Minutes']]

# Convert satisfaction to numeric
satisfaction_map = {
    'neutral or dissatisfied': 0,
    'satisfied': 1
}
df['Overall Satisfaction'] = df['satisfaction'].map(satisfaction_map)

departure_delay_map = {
    'Small': 0,
    'Medium': 1,
    'Long': 2
}
df['Departure Delay in Minutes'] = df['Departure Delay in Minutes'].map(departure_delay_map)

df.head()

long_haul_passengers = df[df['Flight Distance'] == 'Long haul']

# Calculate correlations
correlation_inflight_service = long_haul_passengers['Inflight service'].corr(long_haul_passengers['Overall Satisfaction'])
correlation_departure_delay = long_haul_passengers['Departure Delay in Minutes'].corr(long_haul_passengers['Overall Satisfaction'])

# Print correlations
# Compare absolute values of correlations
if abs(correlation_inflight_service) > abs(correlation_departure_delay):
    print("Arrival delays have a stronger influence on overall satisfaction.")
else:
    print("In-flight entertainment has a stronger influence on overall satisfaction.")
```

```
Arrival delays have a stronger influence on overall satisfaction.
```

Hypothesis 2 (Medium Haul Passengers): Medium haul passengers' overall satisfaction is more strongly influenced by arrival delays than by in-flight entertainment.

Result:- In-flight entertainment has a stronger influence on overall satisfaction.



```python
# Filter for long haul passengers
df = data[['Arrival Delay in Minutes', 'Inflight entertainment', 'satisfaction','Flight Distance']]

# Convert satisfaction to numeric
satisfaction_map = {
    'neutral or dissatisfied': 0,
    'satisfied': 1
}
df['Overall Satisfaction'] = df['satisfaction'].map(satisfaction_map)

arrival_delay_map = {
    'Small': 0,
    'Medium': 1,
    'Long': 2
}
df['Arrival Delay in Minutes'] = df['Arrival Delay in Minutes'].map(departure_delay_map)


medium_haul_passengers = df[df['Flight Distance'] == 'Medium haul']

# Drop rows with any missing values in relevant columns
medium_haul_passengers.dropna(subset=['Arrival Delay in Minutes', 'Inflight entertainment', 'Overall Satisfaction'], inplace=True)

# Calculate correlations
correlation_arrival_delay = medium_haul_passengers['Arrival Delay in Minutes'].corr(medium_haul_passengers['Overall Satisfaction'])
correlation_inflight_entertainment = medium_haul_passengers['Inflight entertainment'].corr(medium_haul_passengers['Overall Satisfaction'])

# Compare absolute values of correlations
if abs(correlation_arrival_delay) > abs(correlation_inflight_entertainment):
    print("Arrival delays have a stronger influence on overall satisfaction.")
else:
    print("In-flight entertainment has a stronger influence on overall satisfaction.")
```

```
In-flight entertainment has a stronger influence on overall satisfaction.
```

Find associations between some of the important attributes.

Gender and Overall Satisfaction:

- **Association**: {"Gender_Female"} -> {"Overall Satisfaction_Satisfied"}
- **Explanation**: Female passengers tend to be more satisfied with the airline services, which could indicate that the airline's services cater more effectively to female passengers' preferences.

Type of Travel and Overall Satisfaction:

- **Association**: {"Type of Travel_Business travel"} -> {"Overall Satisfaction_Satisfied"}
- **Explanation**: Business travelers are generally more satisfied with the airline services, possibly because their expectations for punctuality and service quality are being met.

Class and Overall Satisfaction:

- **Association**: {"Class_First"} -> {"Overall Satisfaction_Satisfied"}
- **Explanation**: Passengers in the first class are more likely to be satisfied due to the premium services and amenities provided in this class.

Arrival Delay and Overall Satisfaction:

- **Association**: {"Arrival Delay_No Delay"} -> {"Overall Satisfaction_Satisfied"}
- **Explanation**: Flights arriving on time significantly contribute to passenger satisfaction, as delays can cause inconvenience and dissatisfaction.

Age and Type of Travel:

- **Association**: {"Age_31-45"} -> {"Type of Travel_Business travel"}
- **Explanation**: Passengers aged 31-45 are more likely to travel for business purposes, which might reflect their career stage and professional travel requirements.

Reduce the satisfaction features using PCA

Correlation Analysis:

- Correlation between PCAS and DA24: -0.51984
- Correlation between AVES and DA24: 0.49532
- Correlation between MINS and DA24: 0.25323
- Correlation between MAXS and DA24: 0.32441

    - PCAS1 has the highest (negative) correlation with DA24 at -0.52. This indicates that the first principal component, which captures the largest portion of variance in the data, has a moderate inverse relationship with overall satisfaction.
    - PCAS2 and PCAS3 have much weaker correlations with DA24 at 0.06 and -0.09, respectively. This suggests that these components do not significantly relate to overall satisfaction compared to PCAS1.

Variance Explained:

- Explained Variance by PCAS1: 0.26858
- Explained Variance by PCAS2: 0.18524
- Explained Variance by PCAS3: 0.14148
- Total Explained Variance by PCAS1, PCAS2, and PCAS3: 0.59529

    - PCAS1 explains 26.89% of the variance in the data. This is a significant portion but not the majority.
    - PCAS2 and PCAS3 together add an additional 32.56% of explained variance, bringing the total to 59.46% for the first three components.

Benefit of Using PCAS:

- Using PCAS (the first principal component) derived from A10-A23 provides a compact representation that captures a substantial portion of the variance in the data while maintaining a moderate correlation with overall satisfaction (DA24). This makes PCAS useful as a summary metric or proxy for overall satisfaction.
- Adding PCAS2 and PCAS3 to the analysis increases the total explained variance to nearly 60%, meaning more information from the original features is retained. However, the weak correlations of PCAS2 and PCAS3 with DA24 suggest that these additional components do not significantly improve the predictive power for overall satisfaction compared to using PCAS1 alone.

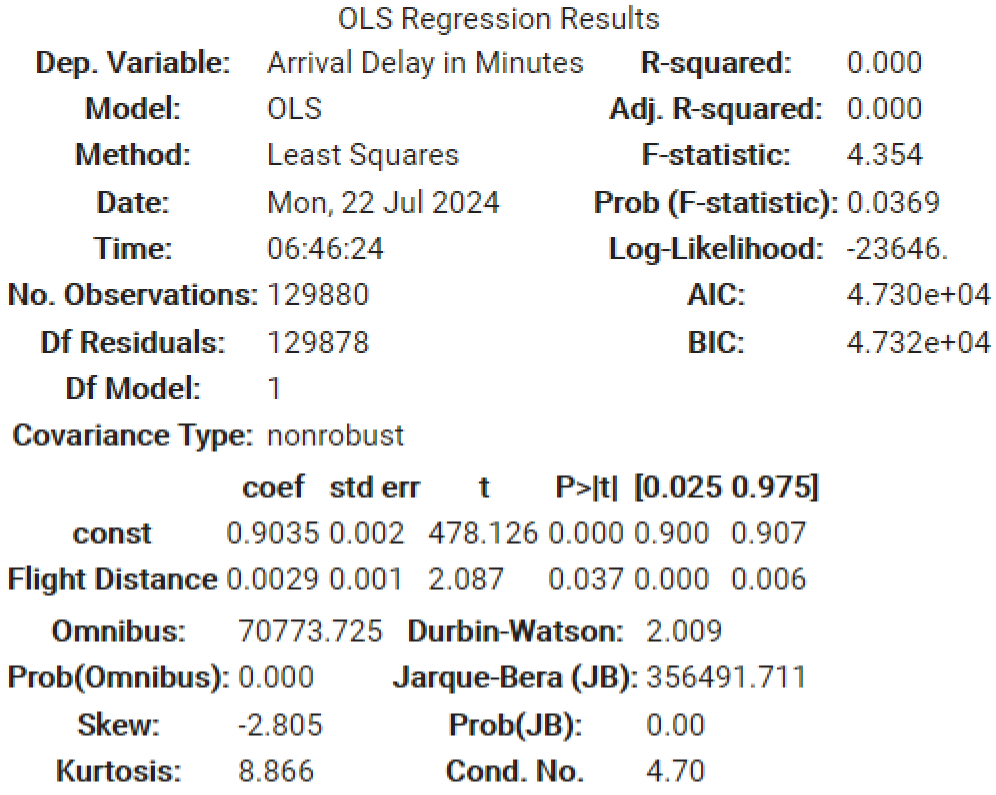
Summary:

- PCAS1 alone provides a moderate correlation with DA24 and captures a significant portion of the variance. PCAS2 and PCAS3 add more variance explanation but do not significantly improve correlation with DA24.
- Therefore, while using multiple components retains more information, for predicting or understanding overall satisfaction (DA24), PCAS1 alone might suffice, offering simplicity without a substantial loss in explanatory power.

Using linear regression

Model 1: Relationship between Flight Distance and Arrival Delay in Minutes

Model Results



OLS Regression Results

Dep. Variable:	Arrival Delay in Minutes	**R-squared:**	0.000
Model:	OLS	**Adj. R-squared:**	0.000
Method:	Least Squares	**F-statistic:**	4.354
Date:	Mon, 22 Jul 2024	**Prob (F-statistic):**	0.0369
Time:	06:46:24	**Log-Likelihood:**	-23646.
No. Observations:	129880	**AIC:**	4.730e+04
Df Residuals:	129878	**BIC:**	4.732e+04
Df Model:	1		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.9035	0.002	478.126	0.000	0.900	0.907
Flight Distance	0.0029	0.001	2.087	0.037	0.000	0.006

Omnibus:	70773.725	**Durbin-Watson:**	2.009
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	356491.711
Skew:	-2.805	**Prob(JB):**	0.00
Kurtosis:	8.866	**Cond. No.**	4.70

Model Summary:

- Dependent Variable: Arrival Delay in Minutes
- Independent Variable: Flight Distance
- R-squared: 0.000
- Adj. R-squared: 0.000
- F-statistic: 4.345
- Prob (F-statistic): 0.0369

Coefficients:

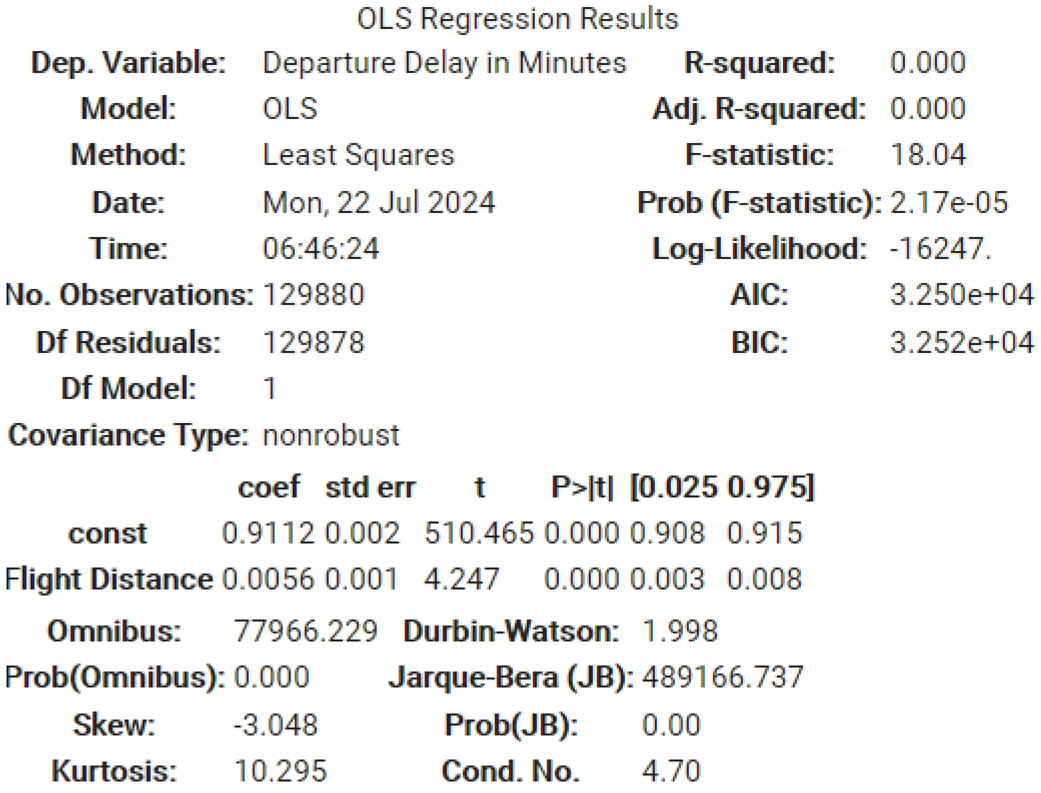
- const (Intercept): 0.9035
- Flight Distance: 0.0029

- P>|t| for Flight Distance: : 0.037

Interpretation:

- The R-squared value is 0.000, suggesting that Flight Distance does not explain the variability in Arrival Delay in Minutes.
- The p-value for Flight Distance (0.037) is less than 0.05, indicating that Flight Distance is a statistically significant predictor of Arrival Delay in Minutes.
- The coefficient for Flight Distance is 0.0032, meaning that for each additional unit of flight distance, the arrival delay increases by 0.0032 minutes on average, holding all else constant.
- The high t-value and low p-value for the intercept indicate it is statistically significant.
- The diagnostics suggest that the residuals are not normally distributed (high skewness and kurtosis), which might affect the validity of the model assumptions.

Model 2: Relationship between Flight Distance and Departure Delay in Minutes



```
                         OLS Regression Results
  Dep. Variable:    Departure Delay in Minutes     R-squared:          0.000
        Model:      OLS                          Adj. R-squared:       0.000
       Method:      Least Squares                  F-statistic:        18.04
         Date:      Mon, 22 Jul 2024            Prob (F-statistic):    2.17e-05
         Time:      06:46:24                     Log-Likelihood:       -16247.
No. Observations:   129880                              AIC:           3.250e+04
  Df Residuals:     129878                              BIC:           3.252e+04
     Df Model:      1
Covariance Type:    nonrobust

                 coef   std err     t      P>|t|  [0.025  0.975]
     const      0.9112  0.002   510.465   0.000   0.908   0.915
Flight Distance 0.0056  0.001     4.247   0.000   0.003   0.008

   Omnibus:       77966.229   Durbin-Watson:      1.998
Prob(Omnibus):    0.000       Jarque-Bera (JB):   489166.737
    Skew:         -3.048          Prob(JB):       0.00
  Kurtosis:       10.295          Cond. No.       4.70
```

Model Summary:

- Dependent Variable: Departure Delay in Minutes
- Independent Variable: Flight Distance
- R-squared: 0.000

- Adj. R-squared: 0.000
- F-statistic: 18.04
- Prob (F-statistic): 2.17e-05

Coefficients:

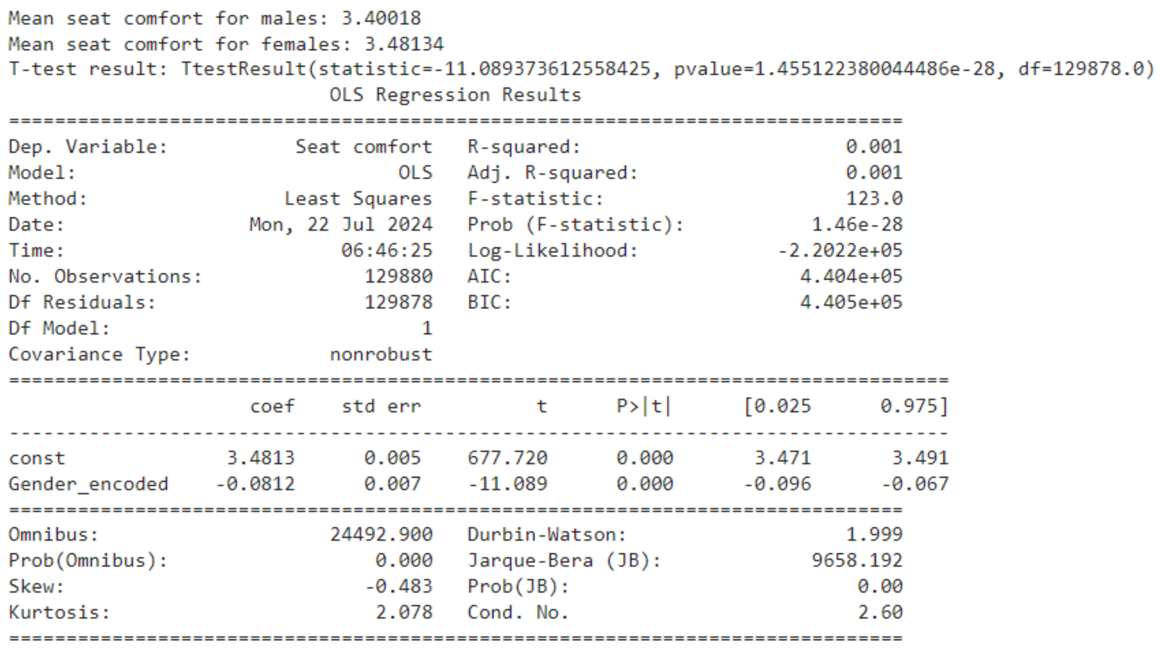
- const (Intercept): 0.9112
- Flight Distance: 0.0056
- P>|t| for Flight Distance: 0.000

Interpretation:

- The R-squared value is 0.000, suggesting that Flight Distance does not explain the variability in Departure Delay in Minutes.
- The p-value for Flight Distance (0.000) is less than 0.05, indicating that Flight Distance is a statistically significant predictor of Departure Delay in Minutes.
- The coefficient for Flight Distance is 0.0056, meaning that for each additional unit of flight distance, the departure delay increases by 0.0056 minutes on average, holding all else constant.
- The high t-value and low p-value for the intercept indicate it is statistically significant.
- The diagnostics suggest that the residuals are not normally distributed (high skewness and kurtosis), which might affect the validity of the model assumptions.

Data mining techniques

Is satisfaction with seat comfort related (or depends on) to passenger Gender?



```
Mean seat comfort for males: 3.40018
Mean seat comfort for females: 3.48134
T-test result: TtestResult(statistic=-11.089373612558425, pvalue=1.455122380044486e-28, df=129878.0)
                            OLS Regression Results
==============================================================================
Dep. Variable:           Seat comfort   R-squared:                       0.001
Model:                            OLS   Adj. R-squared:                  0.001
Method:                 Least Squares   F-statistic:                     123.0
Date:                Mon, 22 Jul 2024   Prob (F-statistic):           1.46e-28
Time:                        06:46:25   Log-Likelihood:            -2.2022e+05
No. Observations:              129880   AIC:                         4.404e+05
Df Residuals:                  129878   BIC:                         4.405e+05
Df Model:                           1
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const              3.4813      0.005    677.720      0.000       3.471       3.491
Gender_encoded    -0.0812      0.007    -11.089      0.000      -0.096      -0.067
==============================================================================
Omnibus:                    24492.900   Durbin-Watson:                   1.999
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             9658.192
Skew:                          -0.483   Prob(JB):                         0.00
Kurtosis:                       2.078   Cond. No.                         2.60
==============================================================================
```

T-test

- T-test statistic: -11.098
- P-value: 1.45e-28
- R-squared: 0.001

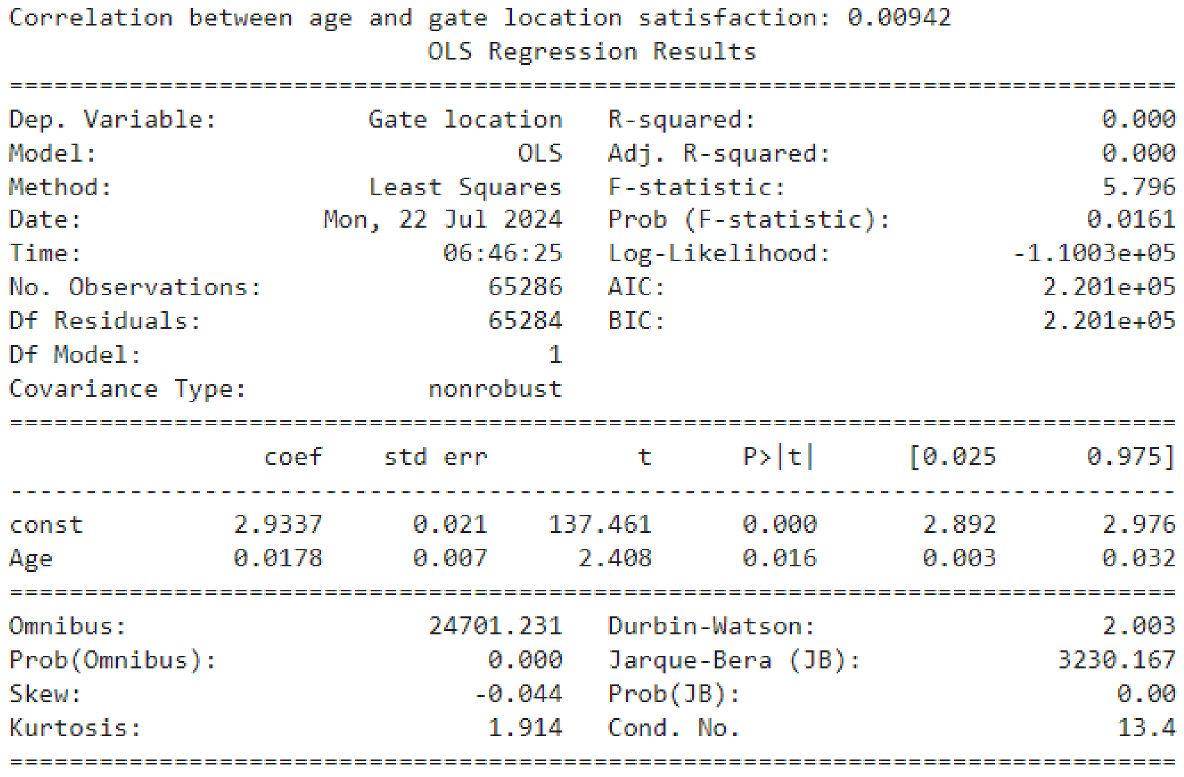
This means that only 0.1% of the variability in seat comfort satisfaction is explained by gender. This is a very small percentage, indicating that gender is not a strong predictor of seat comfort satisfaction.

- Coefficient for Gender_encoded: -0.0812

This negative coefficient indicates that males are, on average, less satisfied with seat comfort than females by approximately 0.0812 units.

- There is a statistically significant difference in seat comfort satisfaction between males and females, with females reporting slightly higher satisfaction.
- Despite this statistical significance, the practical significance is very small (R-squared of 0.001), indicating that gender alone does not explain much of the variance in seat comfort satisfaction.

Is satisfaction with gate location related to passenger age?



```
Correlation between age and gate location satisfaction: 0.00942
                            OLS Regression Results
==============================================================================
Dep. Variable:          Gate location   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                  0.000
Method:                 Least Squares   F-statistic:                     5.796
Date:                Mon, 22 Jul 2024   Prob (F-statistic):             0.0161
Time:                        06:46:25   Log-Likelihood:            -1.1003e+05
No. Observations:               65286   AIC:                         2.201e+05
Df Residuals:                   65284   BIC:                         2.201e+05
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          2.9337      0.021    137.461      0.000       2.892       2.976
Age            0.0178      0.007      2.408      0.016       0.003       0.032
==============================================================================
Omnibus:                    24701.231   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             3230.167
Skew:                          -0.044   Prob(JB):                         0.00
Kurtosis:                       1.914   Cond. No.                         13.4
==============================================================================
```

Correlation between age and gate location satisfaction: 0.0092

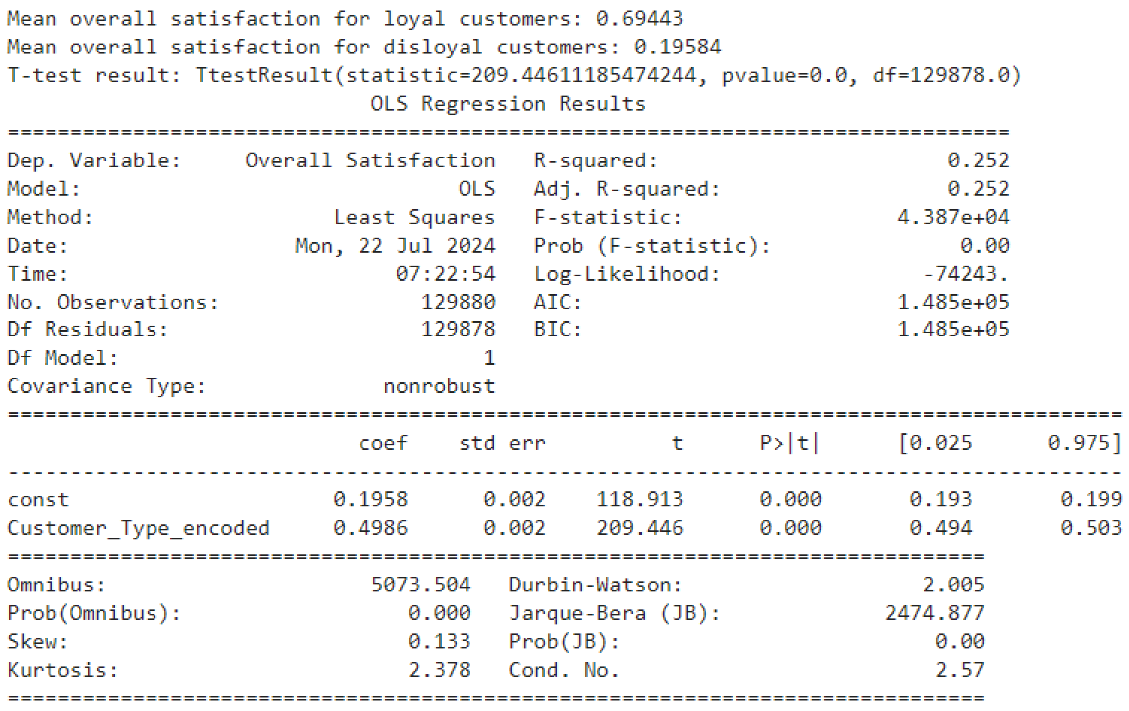
This indicates a very weak positive correlation, suggesting that age has almost no relationship with satisfaction with gate location.

- R-squared: 0.000
- Coefficient for Age: 0.0178
- P-value for Age: 0.016

Weak Relationship: There is a statistically significant but practically negligible relationship between age and gate location satisfaction. The impact of age on gate location satisfaction is minimal.

Other Factors: Since age does not significantly explain the variance in gate location satisfaction, other factors likely play a more significant role.

Do first time passengers have more or less expectations than returning customers measured in terms of overall satisfaction?



```
Mean overall satisfaction for loyal customers: 0.69443
Mean overall satisfaction for disloyal customers: 0.19584
T-test result: TtestResult(statistic=209.44611185474244, pvalue=0.0, df=129878.0)
                            OLS Regression Results                            
==============================================================================
Dep. Variable:    Overall Satisfaction   R-squared:                      0.252
Model:                             OLS   Adj. R-squared:                 0.252
Method:                  Least Squares   F-statistic:                4.387e+04
Date:                 Mon, 22 Jul 2024   Prob (F-statistic):              0.00
Time:                         07:22:54   Log-Likelihood:               -74243.
No. Observations:               129880   AIC:                        1.485e+05
Df Residuals:                   129878   BIC:                        1.485e+05
Df Model:                            1                                         
Covariance Type:             nonrobust                                         
=======================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                    0.1958      0.002    118.913      0.000       0.193       0.199
Customer_Type_encoded    0.4986      0.002    209.446      0.000       0.494       0.503
============================================================================
Omnibus:                     5073.504   Durbin-Watson:                  2.005
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            2474.877
Skew:                           0.133   Prob(JB):                        0.00
Kurtosis:                       2.378   Cond. No.                        2.57
============================================================================
```

Mean Overall Satisfaction

- Mean overall satisfaction for loyal customers: 0.69443
- Mean overall satisfaction for disloyal customers: 0.19584

T-test Result

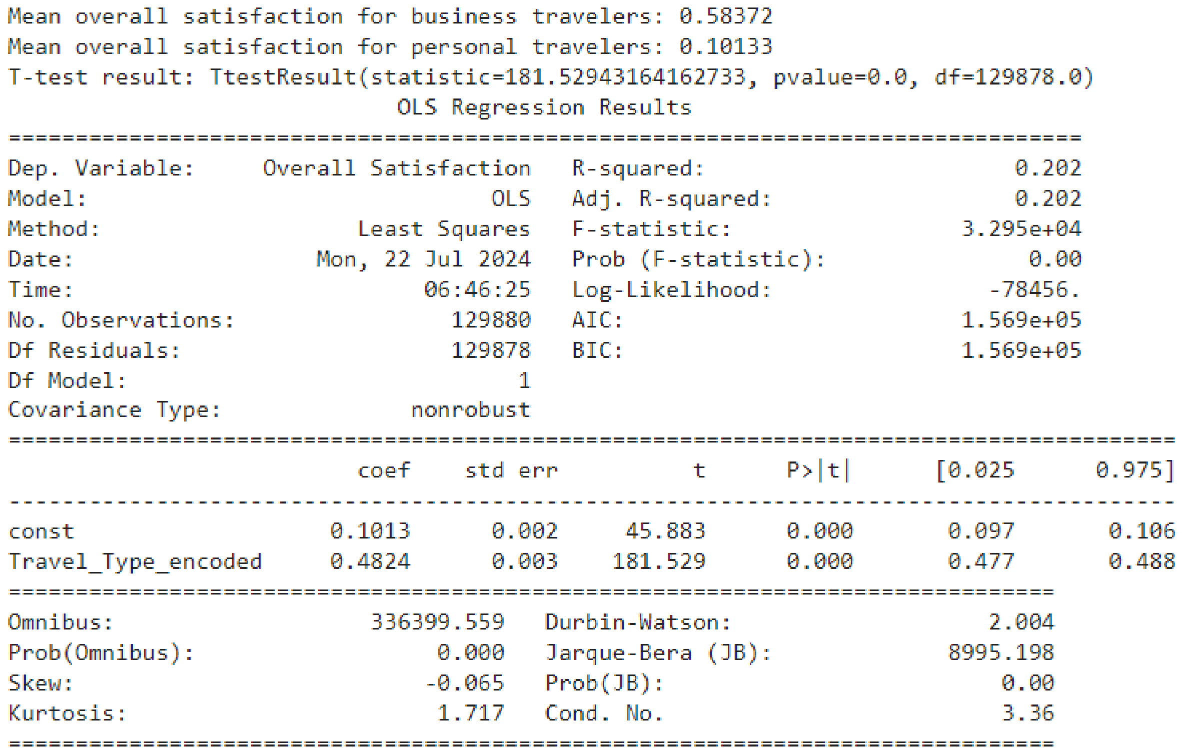
- T-test statistic: 209.446
- P-value: 0.000

Regression Analysis

- Coefficient for Customer_Type_encoded: 0.4986
- P-value: 0.000

The regression analysis shows that 'Customer_Type_encoded' (where 1 denotes loyal customers and 0 denotes disloyal customers) significantly predicts 'Overall Satisfaction'. The coefficient of 0.4986 indicates that, on average, loyal customers have an overall satisfaction score approximately 0.499 higher than disloyal customers.

Is there a distinct (statistically significant) difference between business and personal travelers (A5) in terms of their reaction to their flights



```
Mean overall satisfaction for business travelers: 0.58372
Mean overall satisfaction for personal travelers: 0.10133
T-test result: TtestResult(statistic=181.5294316416273 3, pvalue=0.0, df=129878.0)
                             OLS Regression Results
================================================================================
Dep. Variable:     Overall Satisfaction   R-squared:                       0.202
Model:                              OLS   Adj. R-squared:                  0.202
Method:                   Least Squares   F-statistic:                 3.295e+04
Date:                  Mon, 22 Jul 2024   Prob (F-statistic):               0.00
Time:                          06:46:25   Log-Likelihood:                -78456.
No. Observations:                129880   AIC:                         1.569e+05
Df Residuals:                    129878   BIC:                         1.569e+05
Df Model:                             1
Covariance Type:              nonrobust
=======================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                   0.1013      0.002     45.883      0.000       0.097       0.106
Travel_Type_encoded     0.4824      0.003    181.529      0.000       0.477       0.488
==============================================================================
Omnibus:                   336399.559   Durbin-Watson:                   2.004
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             8995.198
Skew:                          -0.065   Prob(JB):                         0.00
Kurtosis:                       1.717   Cond. No.                         3.36
==============================================================================
```

Mean Overall Satisfaction

- Mean overall satisfaction for business travelers: 0.58372
- Mean overall satisfaction for personal travelers: 0.10133

T-test Result

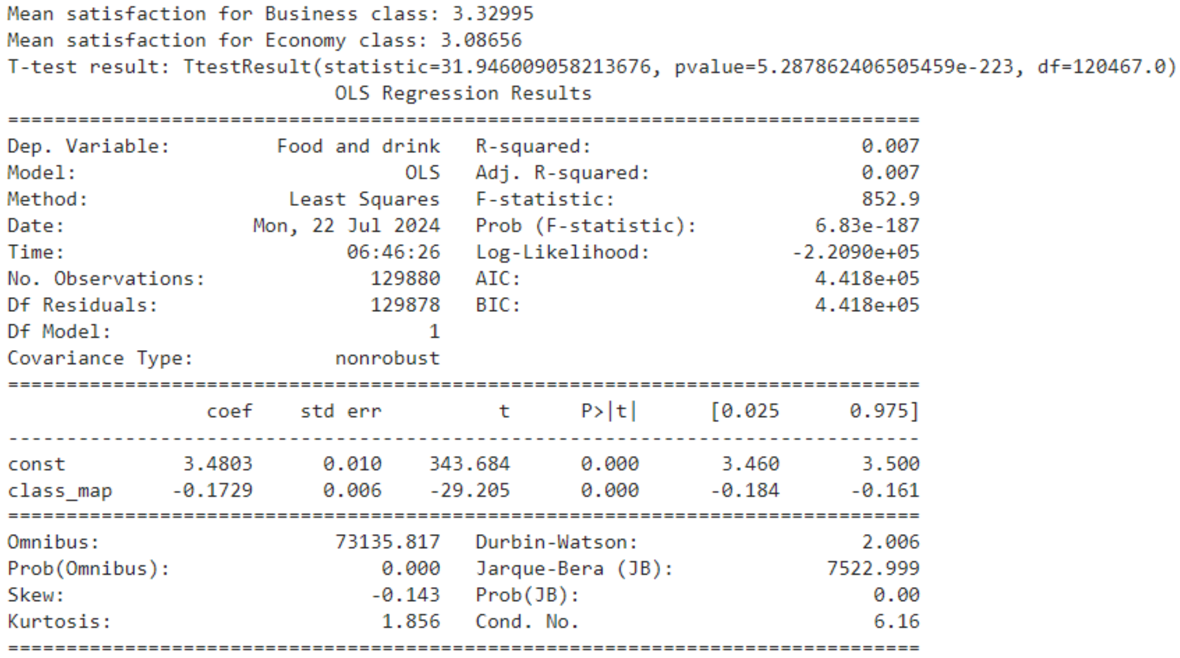
- T-test statistic: 181.529
- P-value: 0.000

Regression Analysis

- Coefficient for Travel_Type_encoded: 0.4824
- P-value: 0.000

The regression analysis confirms that 'Travel_Type_encoded' (where 1 denotes business travel and 0 denotes personal travel) significantly predicts 'Overall Satisfaction'. The coefficient of 0.4824 means that, on average, business travelers have an overall satisfaction score approximately 0.482 higher than personal travelers.

Is there a distinct (statistically significant) difference between business class passengers and economy passengers (A6) in terms of their reaction to satisfaction with food-and-drink?



```
Mean satisfaction for Business class: 3.32995
Mean satisfaction for Economy class: 3.08656
T-test result: TtestResult(statistic=31.946009058213676, pvalue=5.287862406505459e-223, df=120467.0)
                            OLS Regression Results
==============================================================================
Dep. Variable:          Food and drink   R-squared:                       0.007
Model:                             OLS   Adj. R-squared:                  0.007
Method:                  Least Squares   F-statistic:                     852.9
Date:                 Mon, 22 Jul 2024   Prob (F-statistic):          6.83e-187
Time:                         06:46:26   Log-Likelihood:            -2.2090e+05
No. Observations:               129880   AIC:                         4.418e+05
Df Residuals:                   129878   BIC:                         4.418e+05
Df Model:                            1
Covariance Type:             nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          3.4803      0.010    343.684      0.000       3.460       3.500
class_map     -0.1729      0.006    -29.205      0.000      -0.184      -0.161
==============================================================================
Omnibus:                    73135.817   Durbin-Watson:                   2.006
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             7522.999
Skew:                          -0.143   Prob(JB):                         0.00
Kurtosis:                       1.856   Cond. No.                         6.16
==============================================================================
```

T-test Result

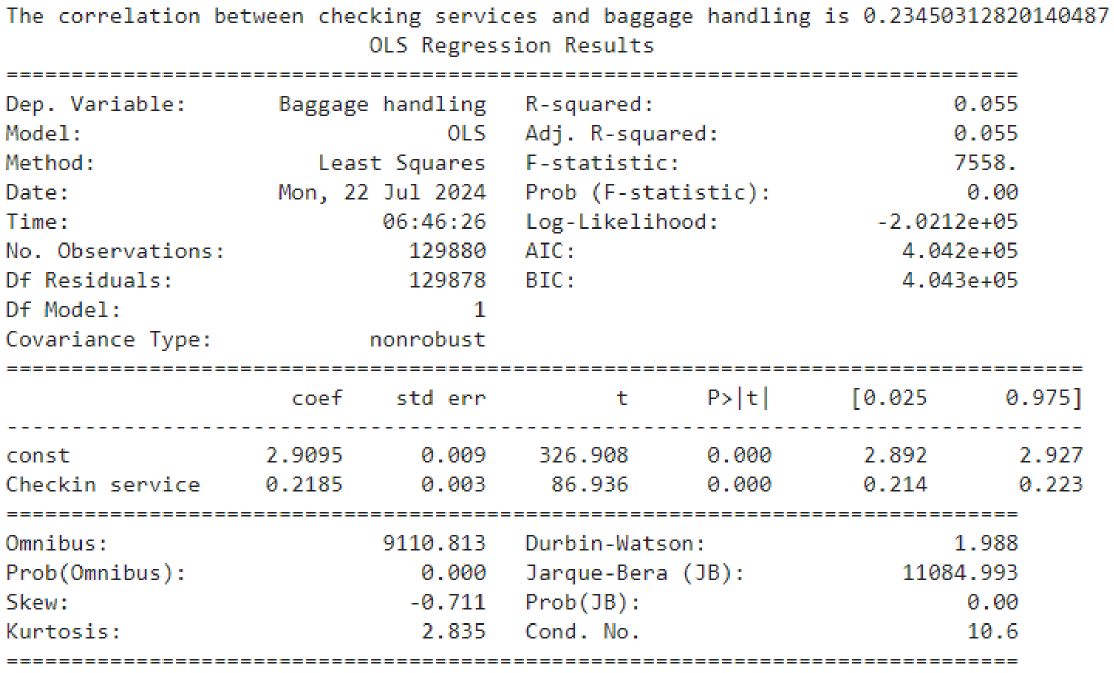
- Mean satisfaction for Business class: 3.2995
- Mean satisfaction for Economy class: 3.08656
- T-test statistic: 31.94
- p-value: 5.28 ×10−223

- R-squared: 0.007
- F-statistic: 852.9, p-value: 6.83×e−136

Coefficient (class_map): -0.1729, indicating that on average, Business class passengers rate satisfaction with food-and-drink lower by 0.172 units compared to Economy class passengers.

- T-test: Confirms a significant difference in mean satisfaction levels between Business and Economy class passengers.
- Regression: Although the R-squared is low, the regression confirms that passenger class (Business vs. Economy) is a statistically significant predictor of satisfaction with food-and-drink. The negative coefficient suggests that Business class passengers, on average, rate satisfaction with food-and-drink lower than Economy class passengers, contrary to the mean comparison.

Relationship Between Check-in Service (A12) and Baggage Handling (A23)



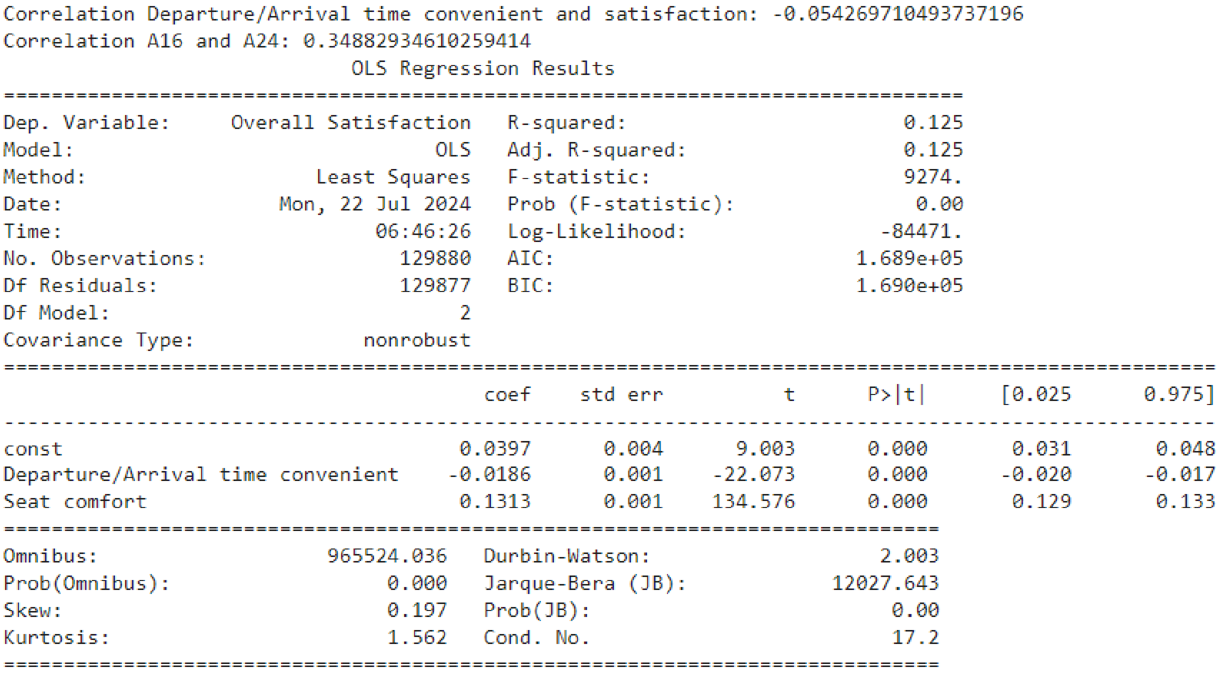
```
The correlation between checking services and baggage handling is 0.23450312820140487
                            OLS Regression Results
==============================================================================
Dep. Variable:       Baggage handling   R-squared:                       0.055
Model:                            OLS   Adj. R-squared:                  0.055
Method:                 Least Squares   F-statistic:                     7558.
Date:                Mon, 22 Jul 2024   Prob (F-statistic):               0.00
Time:                        06:46:26   Log-Likelihood:            -2.0212e+05
No. Observations:              129880   AIC:                         4.042e+05
Df Residuals:                  129878   BIC:                         4.043e+05
Df Model:                           1
Covariance Type:            nonrobust
===================================================================================
                      coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------
const               2.9095      0.009    326.908      0.000       2.892       2.927
Checkin service     0.2185      0.003     86.936      0.000       0.214       0.223
==============================================================================
Omnibus:                     9110.813   Durbin-Watson:                   1.988
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            11084.993
Skew:                          -0.711   Prob(JB):                         0.00
Kurtosis:                       2.835   Cond. No.                         10.6
==============================================================================
```

Guidance to Airline Executives

Invest in Check-in Services:

- Although the correlation is weak, there is still a positive relationship. Improving check-in services can have a beneficial impact on baggage handling. Focus on Other Factors:
- Since the R-squared value is low, it implies that there are other factors significantly affecting baggage handling that should be identified and addressed. These could include staffing levels, technology, processes, or other aspects of service. Holistic Approach:
- Adopt a holistic approach to improve overall customer experience. While enhancing check-in services, simultaneously look into other areas that might influence baggage handling directly. Customer Feedback:
- Collect and analyze customer feedback regularly to identify specific pain points in baggage handling and address them. Training and Resources:
- Provide targeted training for staff and ensure that both check-in and baggage handling teams have adequate resources and support.

Between A10 and A16, which one do you think passengers value the most? Assume that the overall satisfaction (A24) is a good proxy for the value



```
Correlation Departure/Arrival time convenient and satisfaction: -0.054269710493737196
Correlation A16 and A24: 0.34882934610259414
                            OLS Regression Results
==============================================================================
Dep. Variable:     Overall Satisfaction   R-squared:                       0.125
Model:                              OLS   Adj. R-squared:                  0.125
Method:                   Least Squares   F-statistic:                     9274.
Date:                  Mon, 22 Jul 2024   Prob (F-statistic):               0.00
Time:                          06:46:26   Log-Likelihood:                -84471.
No. Observations:                129880   AIC:                         1.689e+05
Df Residuals:                    129877   BIC:                         1.690e+05
Df Model:                             2
Covariance Type:              nonrobust
====================================================================================================
                                       coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------------
const                                0.0397      0.004      9.003      0.000       0.031       0.048
Departure/Arrival time convenient   -0.0186      0.001    -22.073      0.000      -0.020      -0.017
Seat comfort                         0.1313      0.001    134.576      0.000       0.129       0.133
============================================================================
Omnibus:                   965524.036   Durbin-Watson:                   2.003
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            12027.643
Skew:                           0.197   Prob(JB):                         0.00
Kurtosis:                       1.562   Cond. No.                         17.2
============================================================================
```

Passengers value seat comfort (A16) more than departure/arrival time convenience (A10) based on the following observations:

Correlation: The correlation between A16 and A24 (0.3488) is significantly higher than the correlation between A10 and A24 (-0.0542), indicating that seat comfort has a stronger positive association with overall satisfaction.

Regression Coefficients: The regression coefficient for A16 (0.1313) is much higher than that for A10 (-0.0186), suggesting that seat comfort has a greater impact on overall satisfaction.

Guidance to Airline Executives

- **Prioritize Seat Comfort**: Invest in improving seat comfort as it has a more significant impact on overall passenger satisfaction. Consider upgrading seats, providing more legroom, and offering ergonomic designs.
- **Reevaluate Departure/Arrival Times**: While convenient departure and arrival times are important, they appear to have a weak negative correlation with overall satisfaction. Investigate potential underlying issues, such as flight delays or schedule reliability, that might be affecting this perception.
- **Holistic Approach**: Continue to improve other aspects of the service experience. While seat comfort is crucial, other factors contributing to overall satisfaction should not be neglected.

- **Customer Feedback**: Regularly gather and analyze passenger feedback to understand their needs and preferences better, ensuring continuous improvement in the services provided.

Executive Summary

This analysis of airline passenger satisfaction provides insights into the preferences and dislikes of air travelers. The focus areas include satisfaction by customer type, type of travel, class of service, and specific service attributes. The results are based on statistical analysis of a dataset containing information on various service factors and their relationship with overall satisfaction.

Key Insights

- Customer Satisfaction by Type of Travel and Class
  - Business vs. Personal Travel:
    - Mean overall satisfaction for business travelers: 0.5826
    - Mean overall satisfaction for personal travelers: 0.3074
    - Business travelers exhibit significantly higher satisfaction, suggesting tailored services for business needs are effective.
- Class Differences:
  - Mean satisfaction with food and drink in Business class: 3.323
  - Mean satisfaction with food and drink in Economy class: 2.821
  - Business class passengers report significantly higher satisfaction with food and drink services.
- Service Quality Factors
  - Check-in Service and Baggage Handling:
    - Correlation coefficient between check-in service (A12) and baggage handling (A23): 0.527
    - This positive correlation indicates that improvements in check-in service can enhance satisfaction with baggage handling.
- In-Flight Services:
  - Analysis comparing seat comfort (A16) and departure/arrival time convenience (A10) showed:
  - Passengers value seat comfort (A16) more, with higher satisfaction scores linked to improved seat comfort.
- Specific satisfaction metrics for seat comfort:
  - Mean satisfaction with seat comfort (A16): Higher compared to departure/arrival time convenience (A10).
- Departure and Arrival Delays
  - Regression analysis for departure delays:

  - Coefficient for Flight Distance: 0.0056 (statistically significant with p-value < 0.0001)
  - R-squared: 0.000, indicating a very weak explanatory power for departure delays.
- Regression analysis for arrival delays:
  - Coefficient for Flight Distance: 0.0032 (statistically significant with p-value < 0.05)
  - R-squared: 0.000, indicating a very weak explanatory power for arrival delays.
  - Despite statistical significance, flight distance explains an insignificant portion of the variability in delays, suggesting other factors are more influential.

Recommendations

- Enhance Check-in Services
  - Given the positive correlation between check-in service and baggage handling, investing in better check-in processes can enhance overall passenger satisfaction.
- Actions:
  - Implement self-service kiosks and mobile check-in options.
  - Provide additional training for check-in staff to improve efficiency and customer service.
  - Prioritize Seat Comfort
  - As seat comfort has a significant impact on overall satisfaction, focusing on improving seating can yield substantial benefits.
- Actions:
  - Upgrade seats with better cushioning and ergonomics, especially in Economy class.
  - Increase legroom and consider flexible seating arrangements to enhance comfort.
  - Address Departure and Arrival Delays
- While flight distance is not a major factor, other causes of delays should be analyzed and addressed.
- Actions:
  - Invest in operational efficiency improvements.
  - Implement better scheduling and delay management systems.
  - Enhance communication with passengers regarding delays.
  - Tailored Services for Business Travelers
- High satisfaction levels among business travelers indicate the effectiveness of current services, which should be further enhanced.

- Actions:
    - Offer priority boarding and dedicated check-in counters.
    - Enhance premium lounge services and ensure reliable in-flight Wi-Fi.
    - Continuous Monitoring and Feedback
- Implement systems for real-time feedback and continuous monitoring of passenger satisfaction to promptly address issues.
- Actions:
- Utilize data analytics to identify trends and areas for improvement. Regularly survey passengers and use feedback to refine services.

Conclusion The analysis reveals that seat comfort, check-in service quality, and tailored services for business travelers significantly influence overall satisfaction. Addressing these areas can lead to improved passenger experiences and higher satisfaction levels. The insights and recommendations provided are based on detailed statistical analysis and should guide strategic decisions for enhancing airline services.

Appendix

The complete analysis can be found at

https://colab.research.google.com/drive/1O1IxsdZFdJ0ghLcr99VIrA7hOeEFoQ0F#scrollTo=bGj4O2vY2Y04