MSE Project

Flight Price Prediction

SHUBHAM BATHWAL – A21031
DEBOPRIYO MALLICK – A21010

FEBRUARY 25

PRAXIS BUSINESS SCHOOL
BATCH – DS AUG 2021

CONTENTS

INTRODUCTION

The objective of the study is to analyse the flight booking dataset obtained from “Ease My Trip” website and to conduct various statistical hypothesis tests in order to get meaningful information from it. The 'Linear Regression' statistical algorithm would be used to train the dataset and predict a continuous target variable. 'Easemytrip' is an internet platform for booking flight tickets, and hence a platform that potential passengers use to buy tickets. A thorough study of the data will aid in the discovery of valuable insights that will be of enormous value to passengers.

The model created with Linear Regression will also be used to predict the price of a plane ticket.

RESEARCH QUESTIONS:

The aim of our study is to answer the below research questions:

a) Does price vary with Airlines?
b) How is the price affected when tickets are bought in just 1 or 2 days before departure?
c) Does ticket price change based on the departure time and arrival time?
d) How the price changes with change in Source and Destination?
e) How does the ticket price vary between Economy and Business class?

DATA COLLECTION AND METHODOLOGY

Octoparse scraping tool was used to extract data from the website. Data was collected in two parts: one for economy class tickets and another for business class tickets. A total of 300261 distinct flight booking options was extracted from the site. Data was collected for 50 days, from February 11th to March 31st, 2022.

Data source was secondary data and was collected from Ease my trip website.

For Hypothesis testing, random sampling was done to select 10% of the population data and due care was taken to ensure that sample parameters represent population parameters**.**

DATASET

Dataset contains information about flight booking options from the website Easemytrip for flight travel between India's top 6 metro cities. There are 300261 datapoints and 11 features in the cleaned dataset.

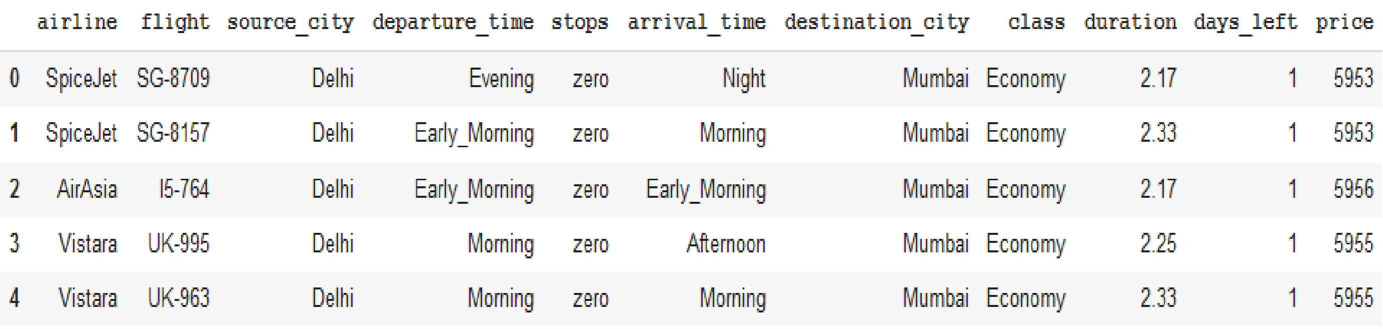
FEATURES

The various features of the cleaned dataset are explained below:

1) **Airline**: The name of the airline company is stored in the airline column. It is a categorical feature having 6 different airlines.
2) **Flight**: Flight stores information regarding the plane's flight code. It is a categorical feature.
3) **Source City**: City from which the flight takes off. It is a categorical feature having 6 unique cities.
4) **Departure Time**: This is a derived categorical feature obtained created by grouping time periods into bins. It stores information about the departure time and have 6 unique time labels.
5) **Stops**: A categorical feature with 3 distinct values that stores the number of stops between the source and destination cities.
6) **Arrival Time**: This is a derived categorical feature created by grouping time intervals into bins. It has six distinct time labels and keeps information about the arrival time.
7) **Destination City**: City where the flight will land. It is a categorical feature having 6 unique cities.
8) **Class**: A categorical feature that contains information on seat class; it has two distinct values: Business and Economy.
9) **Duration**: A continuous feature that displays the overall amount of time it takes to travel between cities in hours.
10)**Days Left**: This is a derived characteristic that is calculated by subtracting the trip date by the booking date.
11) **Price**: Target variable stores information of the ticket price.

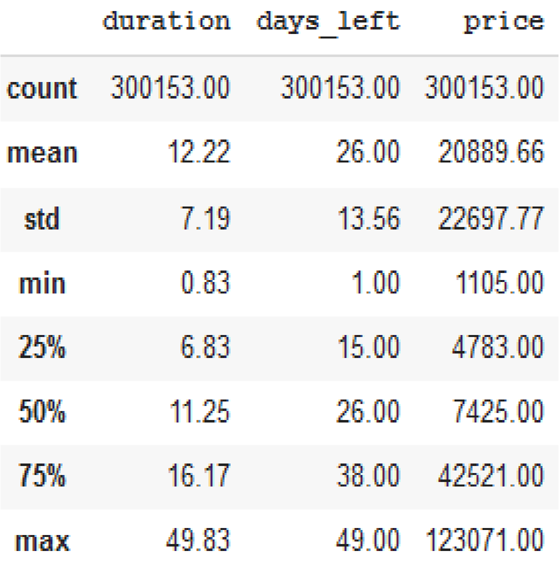
EXPLORATORY DATA ANALYSIS

Instance of the Dataset



	airline	flight	source_city	departure_time	stops	arrival_time	destination_city	class	duration	days_left	price
0	SpiceJet	SG-8709	Delhi	Evening	zero	Night	Mumbai	Economy	2.17	1	5953
1	SpiceJet	SG-8157	Delhi	Early_Morning	zero	Morning	Mumbai	Economy	2.33	1	5953
2	AirAsia	I5-764	Delhi	Early_Morning	zero	Early_Morning	Mumbai	Economy	2.17	1	5956
3	Vistara	UK-995	Delhi	Morning	zero	Afternoon	Mumbai	Economy	2.25	1	5955
4	Vistara	UK-963	Delhi	Morning	zero	Morning	Mumbai	Economy	2.33	1	5955

Basic Description:

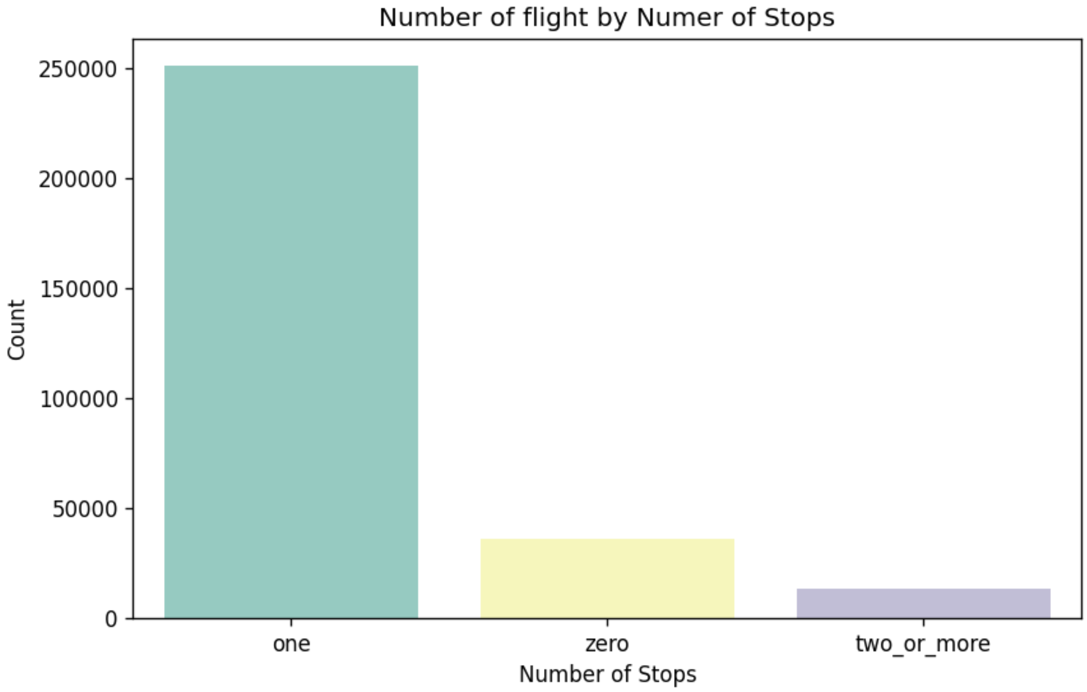


	duration	days_left	price
count	300153.00	300153.00	300153.00
mean	12.22	26.00	20889.66
std	7.19	13.56	22697.77
min	0.83	1.00	1105.00
25%	6.83	15.00	4783.00
50%	11.25	26.00	7425.00
75%	16.17	38.00	42521.00
max	49.83	49.00	123071.00

1. 50% of the flight takes less than 11hrs 15 min, mean duration is 12hrs 12min which is very close to median, so the duration is slightly right skewed.
2. 50% of the flight ticket costs less than Rs7500, mean price is Rs20889 which is very large compared to median price, so the price is highly right skewed.

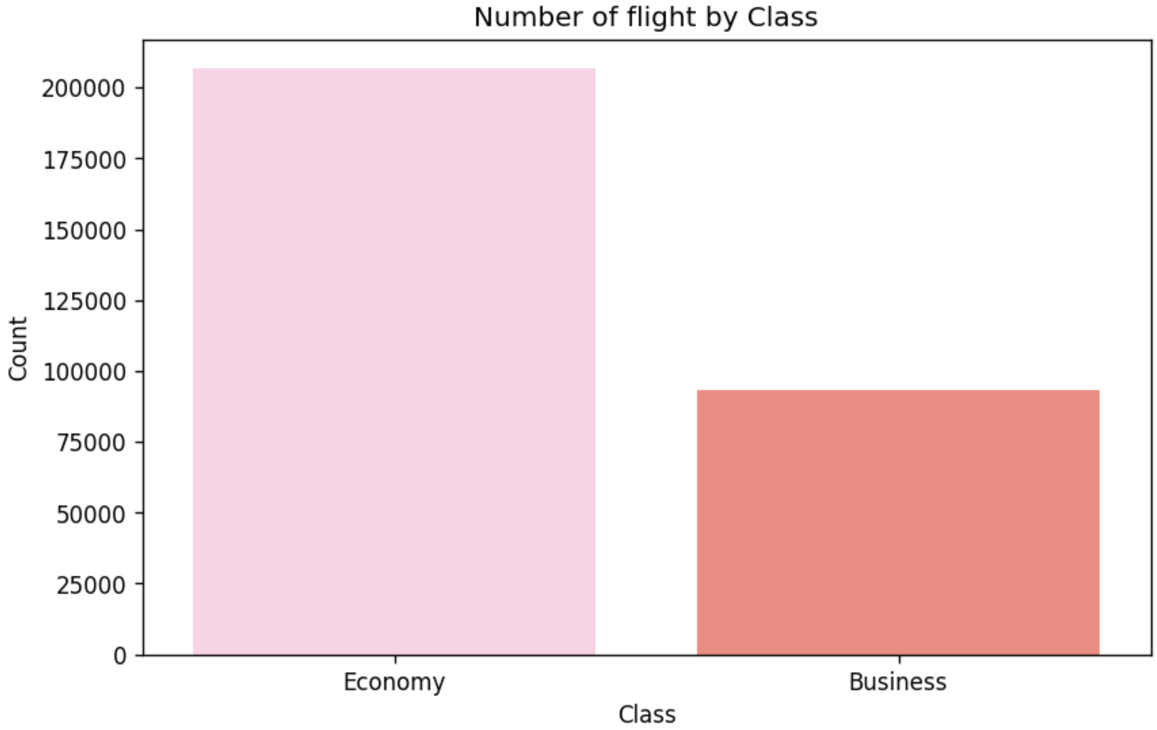
Graphical Analysis:

1. Number of flights by Stops



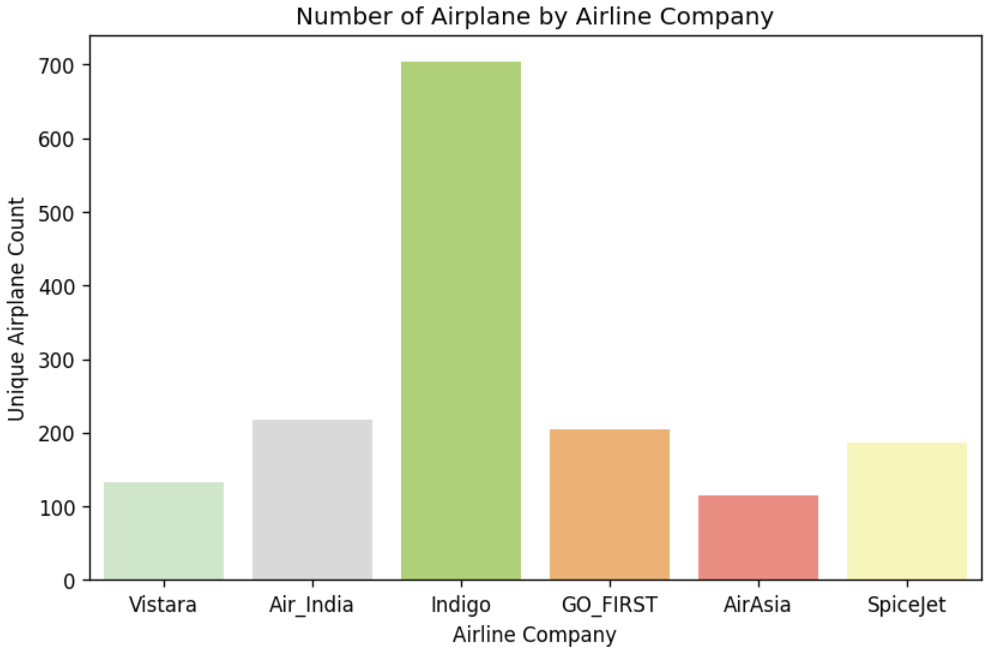
Between the origin and destination city, most flights make at least one stop.

2. Number of Flights by Class



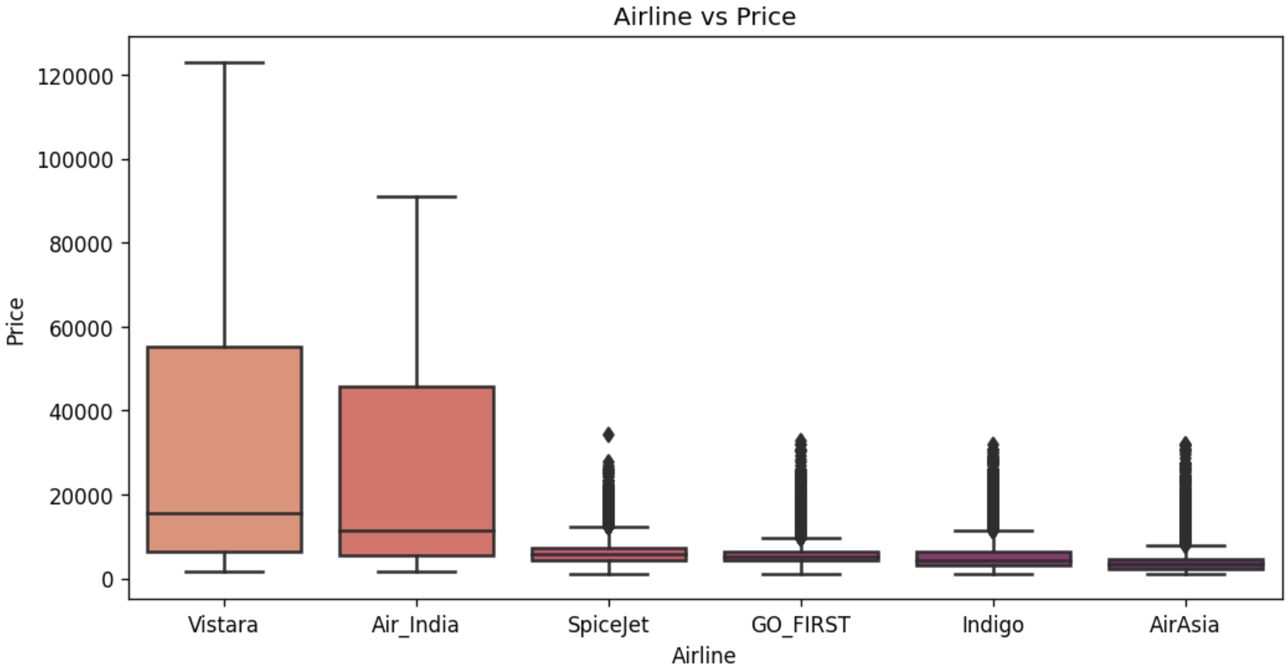
In India, majority of travel is done in Economy Class.

3. Number of airplanes by each airline companies:



Indigo has the largest number of planes that operates between cities.

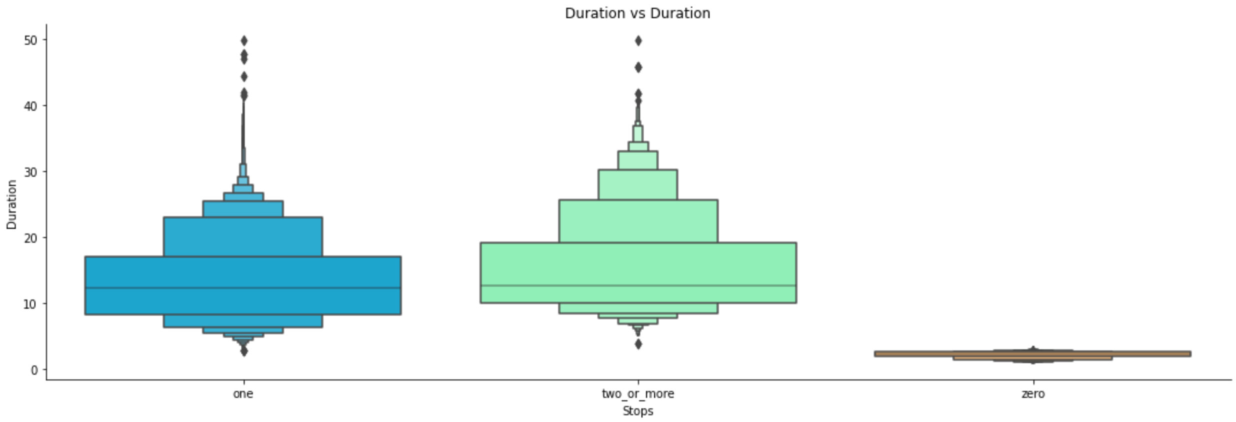
4. Airline vs Price



From the graph, we observe that Vistara and Air India have the highest median prices. Apart from them, practically every airline has comparable median price. Vistara and Air India offers wide range of options in terms of price.

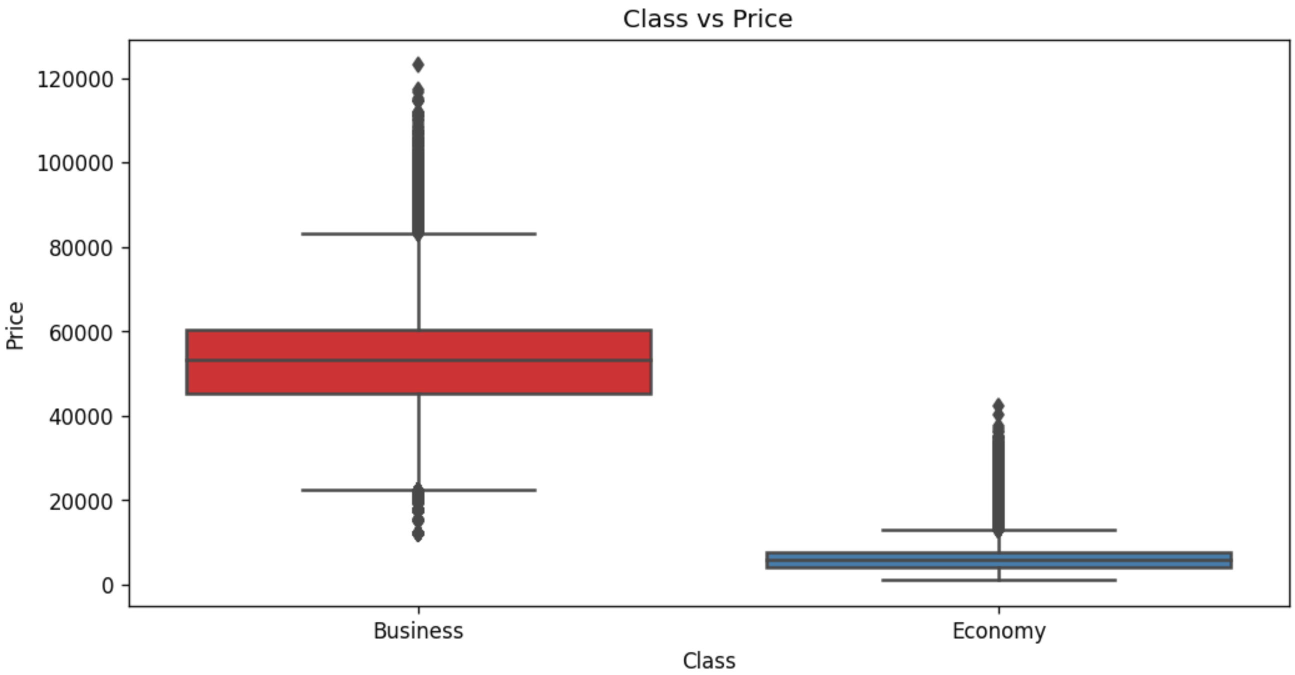
Higher Price of Vistara and Air India is attributed to the fact they are the airline company that offers Business class services in India.

5. Stops vs Duration



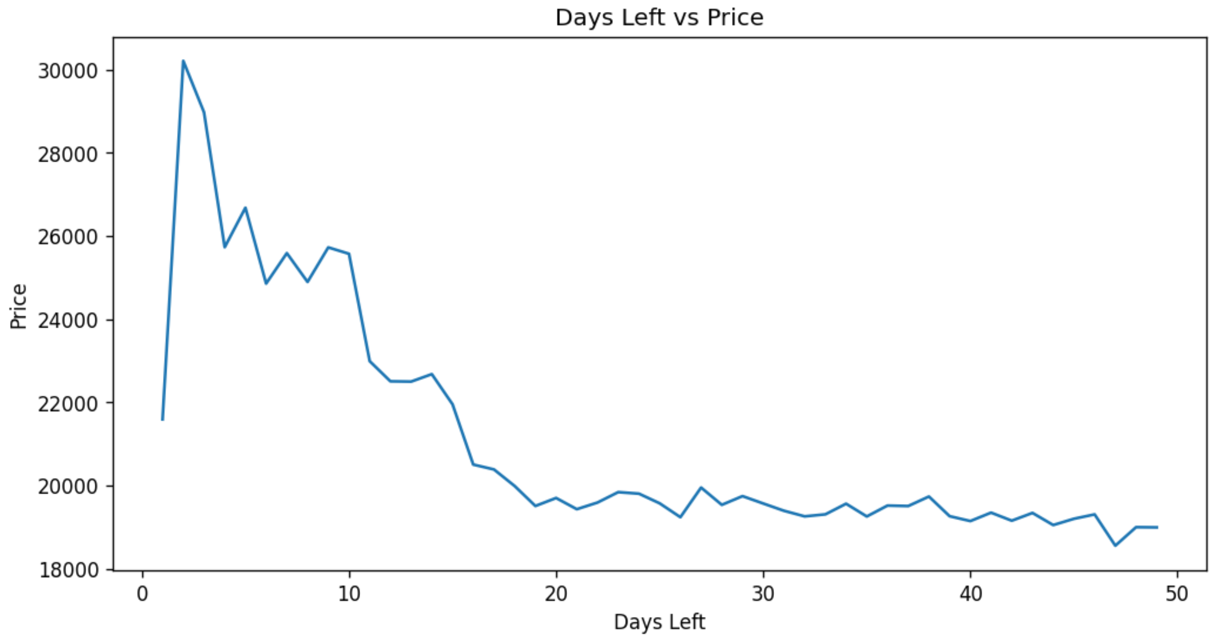
From the graph it is evident that flight without stops takes the least time.

6. Class vs Price



From the graph it can be seen that Economy class flights are cheaper than business class flights.

7. Days Left vs Price



The cost of a flight varies greatly depending on how early you book it.
However, after a certain amount of time, it reaches saturation.

HYPOTHESIS TESTING

For the purpose of hypothesis testing, we are taking a sample of 3000 datapoints (approximately 1% of the population) and performing test on it.

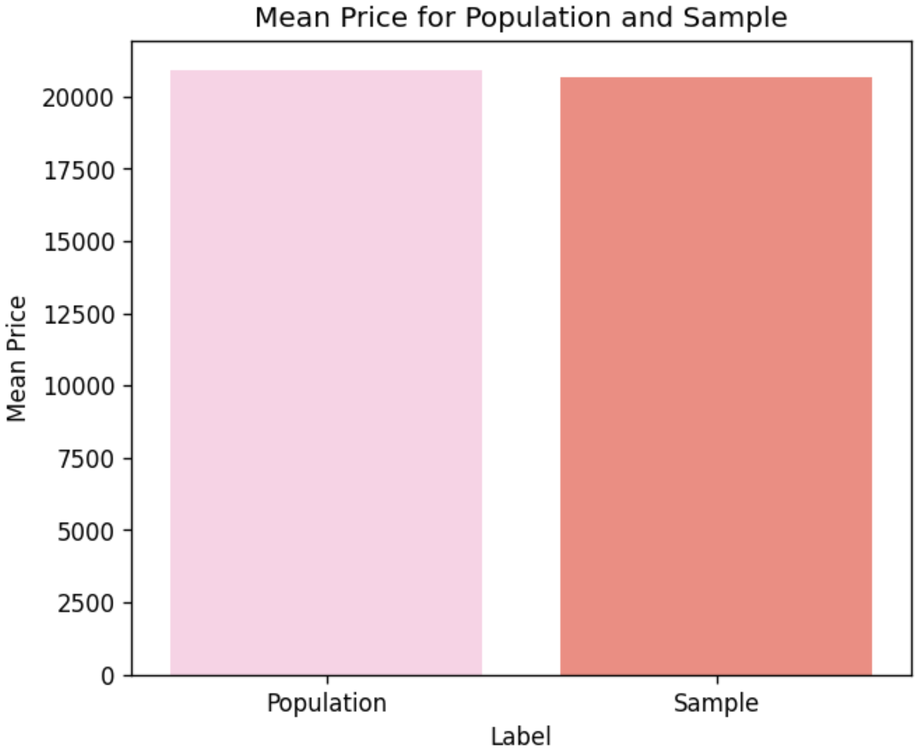
T test

One Sample T test

In one sample t test we compare the mean of one sample against the set mean (population mean).

1) **To check whether the sample dataset is a significant representative of the population**

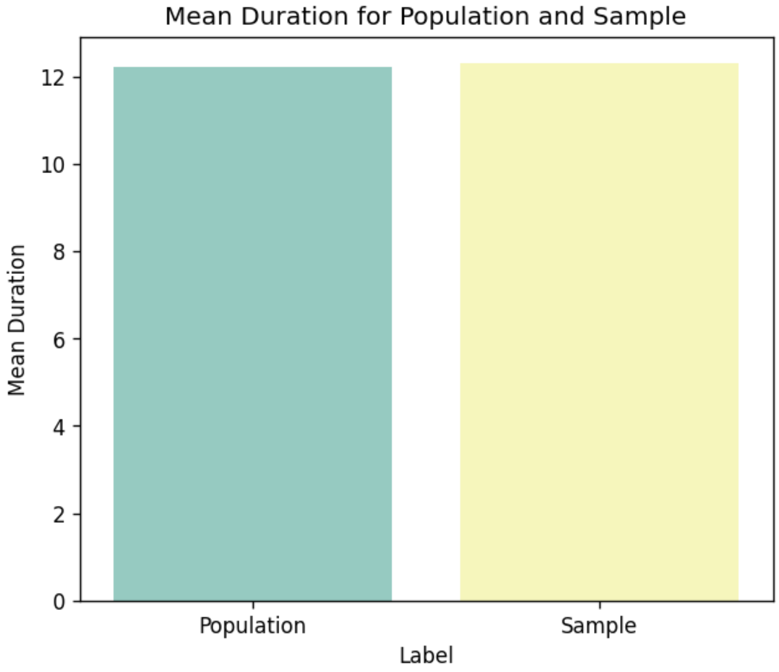
   i) **Comparison between mean sample price and mean population price**.



```
Population Mean Price:  20889.66
Sample Mean Price:  20673.0

Doing t-test to compare means of sample and population
Null Hypothesis: Mean Price are same.
Alternate Hypothesis: Mean Price are different

T-Statistics: -0.524, p-value:0.6

At 95% Confidence:
Failed to reject the Null Hypothesis
```

Mean Price of Sample and Population are not significantly different.

ii) Comparison between mean sample duration and mean population duration.



```
Population Mean Duration:  12.22
Sample Mean Price:  12.3

Doing t-test to compare means of sample and population
Null Hypothesis: Mean Duration are same.
Alternate Hypothesis: Mean Duration are different

T-Statistics:0.571, p-value:0.5682

At 95% Confidence:
Failed to reject the Null Hypothesis
```

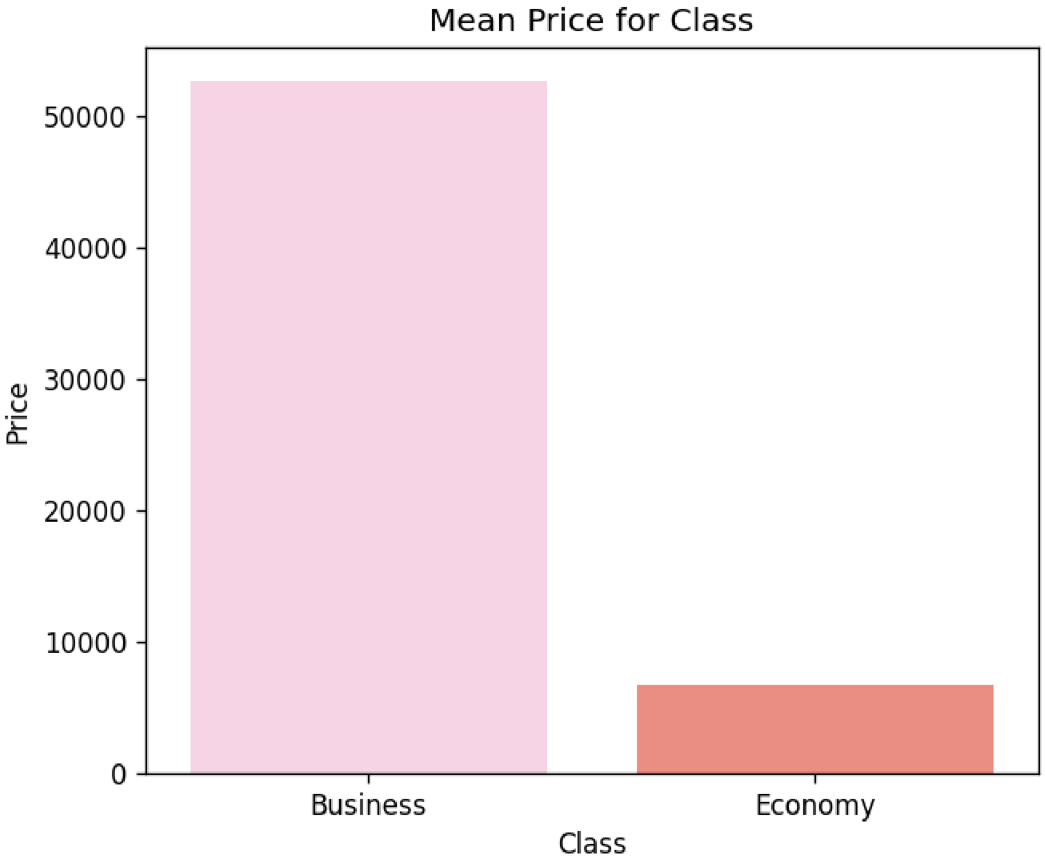
Mean Duration of Sample and Population are not significantly different.

From One Sample T test we can conclude that the sample is a good representation of the population, and we can use the sample to make hypothesis about the population.

Two Sample T test

Two sample t-test is a type of inferential statistic used to determine if there is a significant difference between the means of two samples.

1. **A) To check whether there is any significant difference between price and ticket's class (Economy and Business)**

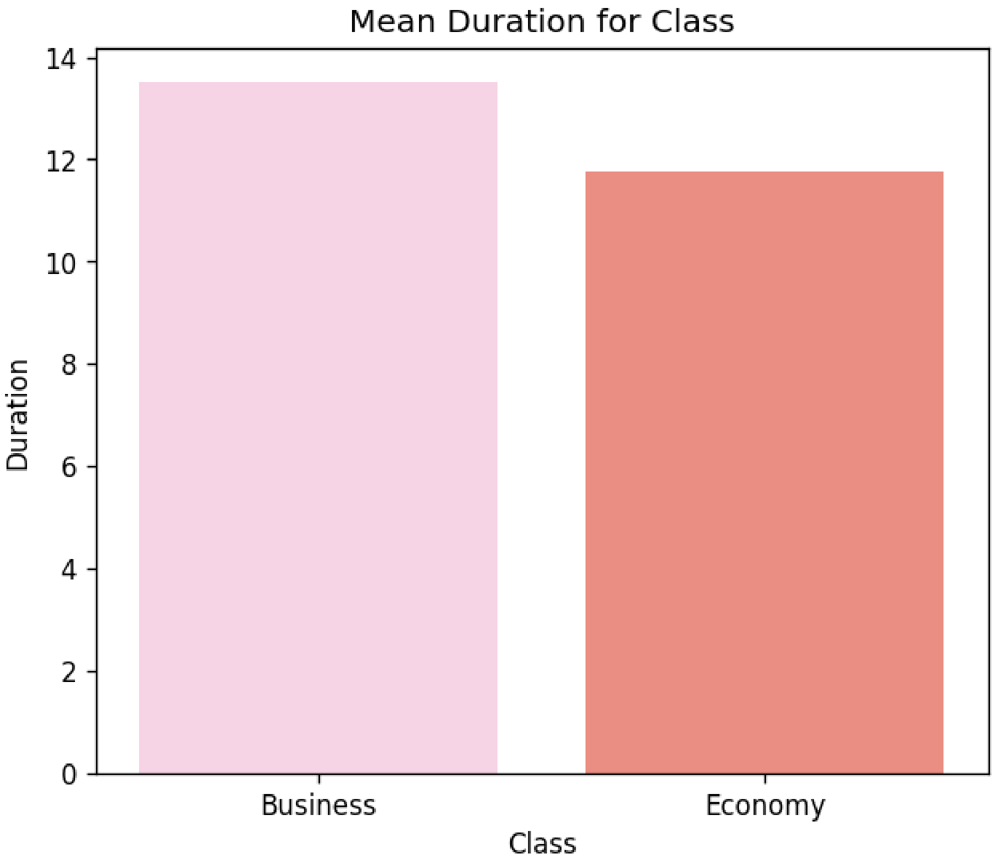


The bar plot shows that the mean ticket price of business class is higher than the mean ticket price of economy class tickets. T Test is conducted to validate the difference.

```
Doing t-test to compare means of the two class
Null Hypothesis: Mean Price for both classes are same.
Alternate Hypothesis: Mean Price for both classes are different

T-Statistics: -100.863, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the price of business class tickets is significantly greater than price of economy class tickets. This information can be used by passengers to buy cheaper tickets by compromising comfort of travel.

1. **B) To check whether there is any significant difference between duration and ticket's class (Economy and Business)**

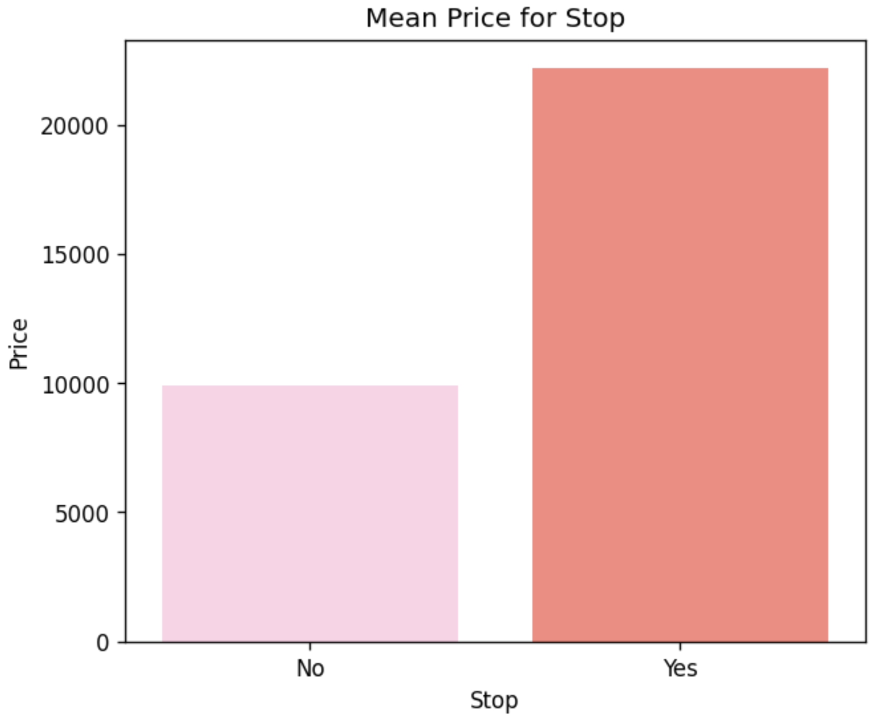


The bar plot shows that the mean duration taken by flights of business class is higher than the mean duration taken by flights of economy class. T Test is conducted to validate the difference.

```
Doing t-test to compare means of the two class
Null Hypothesis: Mean Duration for both classes are same.
Alternate Hypothesis: Mean Duration for both classes are different

T-Statistics: -6.249, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the duration of business class flights is significantly greater than the duration of economy class flights. This information can be used by passengers to buy tickets of shorter duration.

2. **A) To check whether there is any significant difference between price and stops (No stop and One or more stops)**

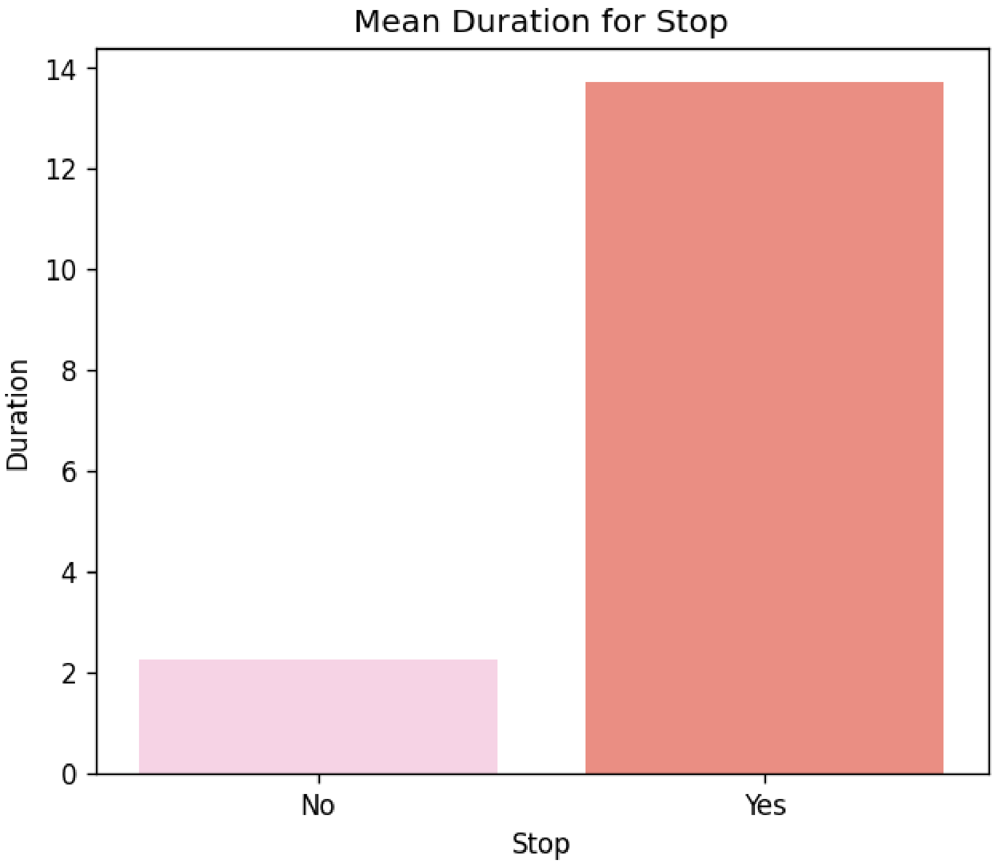


The bar plot shows that the mean ticket price of flights with stops is higher than the mean ticket price of flights without any stop. T Test is conducted to validate the difference.

```
Doing t-test to compare means of the two stops
Null Hypothesis: Mean Price for both stops are same.
Alternate Hypothesis: Mean Price for both stops are different

T-Statistics: -16.637, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the price of flights with stops is significantly greater than price of price of flights without any stops. This information can be used by passengers to buy cheaper tickets by booking flights without any stop.

2. **B) To check whether there is any significant difference between duration and stops (No stop and One or more stops)**



The bar plot shows that the mean duration with stops is higher than the mean duration without any stop. T Test is conducted to validate the difference.

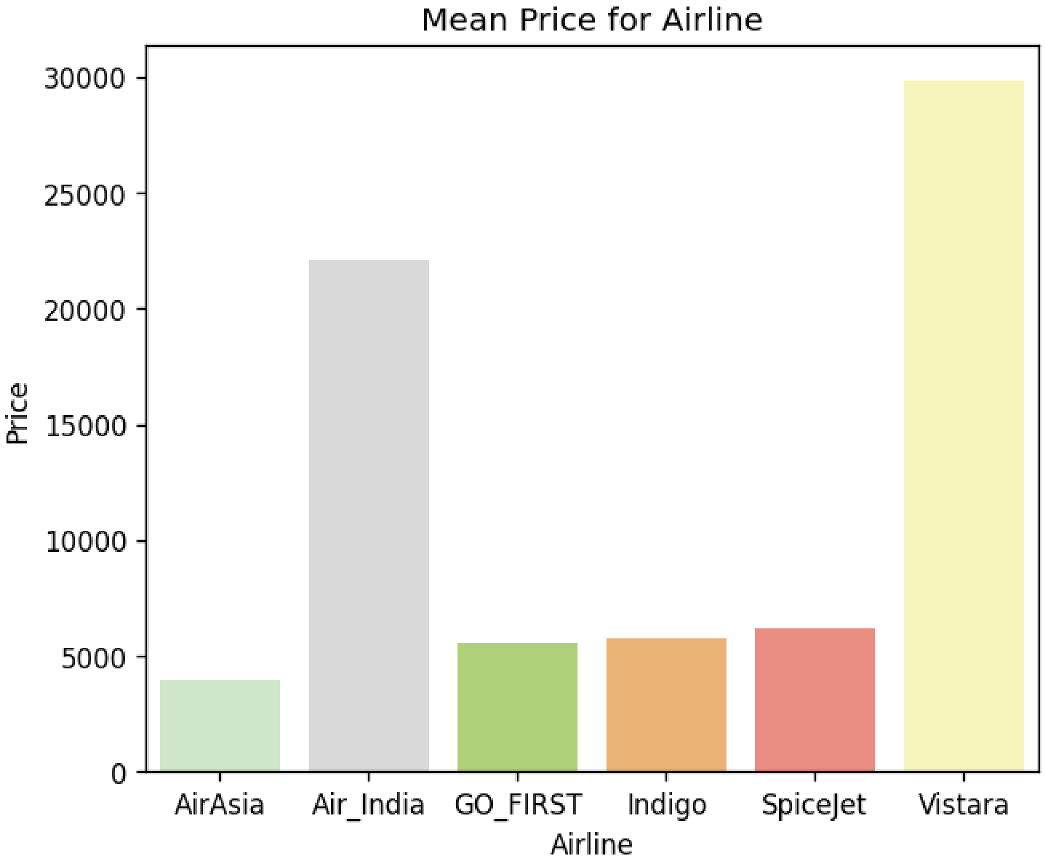
```
Doing t-test to compare means of the two stops
Null Hypothesis: Mean Duration for both stops are same.
Alternate Hypothesis: Mean Duration for both stops are different

T-Statistics: -86.988, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the duration with stops is significantly greater than duration without any stops. This information can be used by passengers to buy tickets of small duration.

ANOVA

Analysis of variance (ANOVA) is a collection of statistical models and their associated estimation procedures used to analyse the differences among means for more than 2 samples.

1. **To check whether there is any significant difference between ticket price for different airlines.**

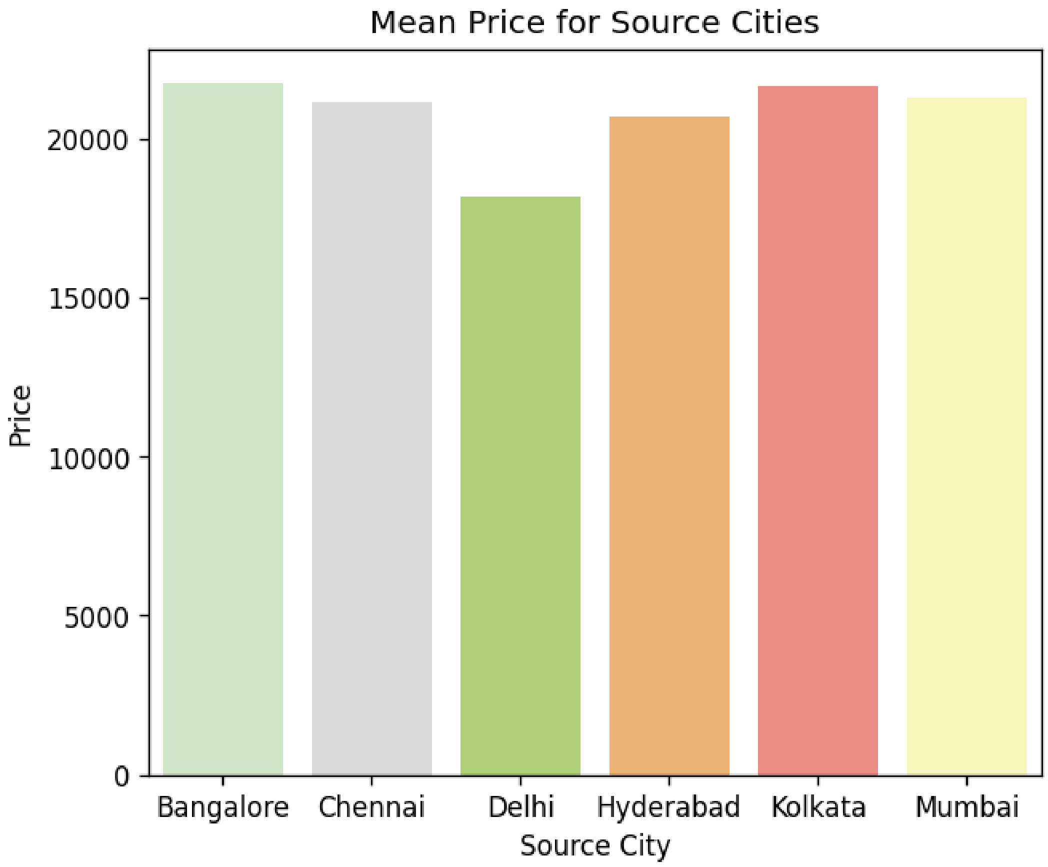


From the bar plot it can be seen that Vistara airline and Air India airline that also provides business class services has relatively higher mean ticket price. ANOVA is conducted to check whether the difference is significant.

```
ANOVA test to compare means
Null Hypothesis: Mean Price for all airlines are same.
Alternate Hypothesis: Mean Price for all airlines are different

F-Statistics:156.179, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the ticket price for different airlines is significantly different.

2. **To check whether there is any significant difference between ticket price for different source city**.



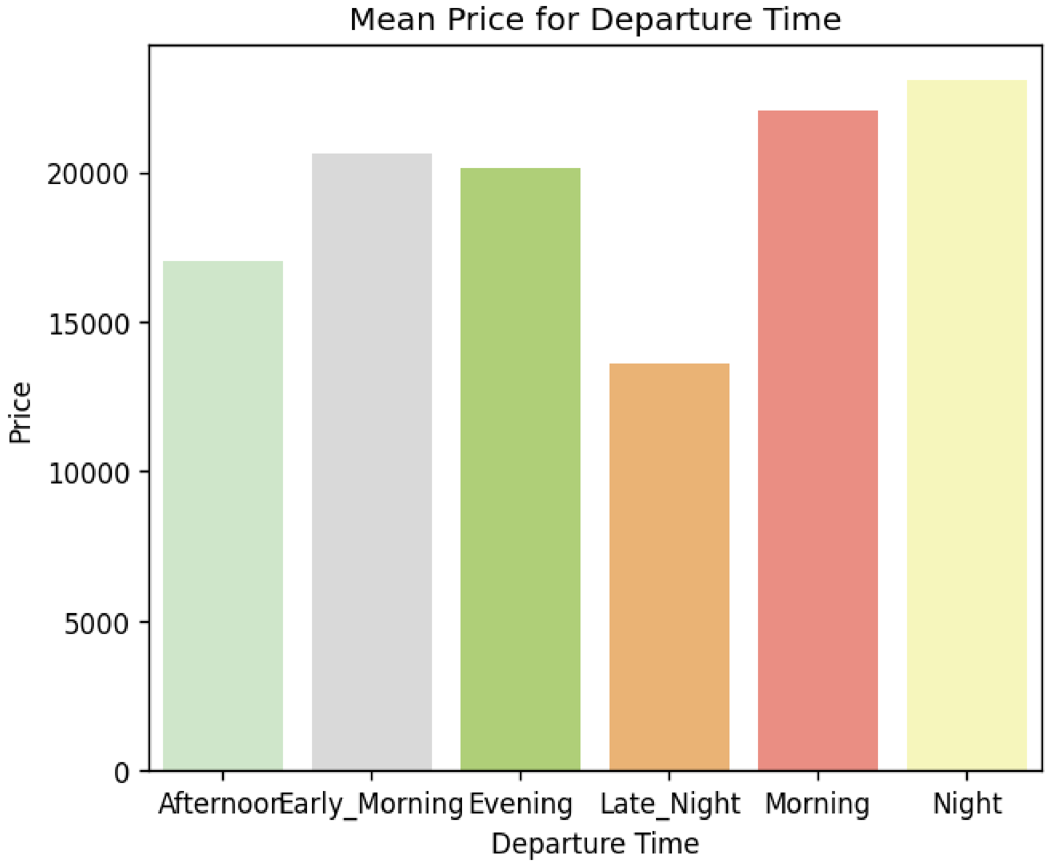
From the bar plot it can be seen that the mean ticket price for flights originating from different cities are almost the same. ANOVA is conducted to check whether the difference is significant.

```
Anova test to compare means
Null Hypothesis: Mean Price for all source cities are same.
Alternate Hypothesis: Mean Price for all source cities are different

F-Statistics:2.045, p-value:0.0695

At 95% Confidence:
Failed to reject the Null Hypothesis
```

This shows that the at confidence of 95% the ticket price is not determined the flight's source city.

(Similar test was done to see if the price depends on the flight destination city and at 95% confidence no significant difference in price was observed.)

3. To check whether there is any significant difference between ticket price for different departure time.



From the bar plot it can be seen that the mean ticket price for flights departing at different times are different. ANOVA is conducted to check whether the difference is significant.

```
ANOVA test to compare means
Null Hypothesis: Mean Price for all Departure Time are same.
Alternate Hypothesis: Mean Price for all Departure Time are different

F-Statistics:4.362, p-value:0.0006

At 95% Confidence:
Reject the Null Hypothesis
```

This shows that the at confidence of 95% the ticket price is determined by the flight's departure time.

(Similar test was done to see if the price depends on the flight arrival time and at 95% confidence no significant difference in price was observed.)

Chi Square Test

A chi-square test for independence compares two variables in a contingency table to see if they are related.

1. To check whether for the class and the airline are related.

```
Contingency Table:
```

```
class      Business  Economy  Total
airline
AirAsia           0      163    163
Air_India       299      495    794
GO_FIRST          0      243    243
Indigo            0      392    392
SpiceJet          0       76     76
Vistara         615      717   1332
Total           914     2086   3000
```

```
Chi-Square test to check relationship
Null Hypothesis:  The two categorical variables airline and class have no
relationship.
Alternate Hypothesis: There is a relationship between two categorical variables
airline and class.

Chi-Statistics:557.404, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

Hence there is a relationship between airline and class variable.

2. To check whether for the class and stops are related.

```
Contingency Table:
```

```
stops       No   Yes  Total
airline
AirAsia     36   127    163
Air_India   52   742    794
GO_FIRST    47   196    243
Indigo     101   291    392
SpiceJet    22    54     76
Vistara    112  1220   1332
Total      370  2630   3000
```

```
Chi-Square test to check relationship
Null Hypothesis:  The two categorical variables airline and stops have no
relationship.
Alternate Hypothesis: There is a relationship between two categorical variables
airline and stops.
```

```
Chi-Statistics:153.738, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

Hence there is a relationship between airline and stops variable.

3. To check whether for the source city and stops are related.

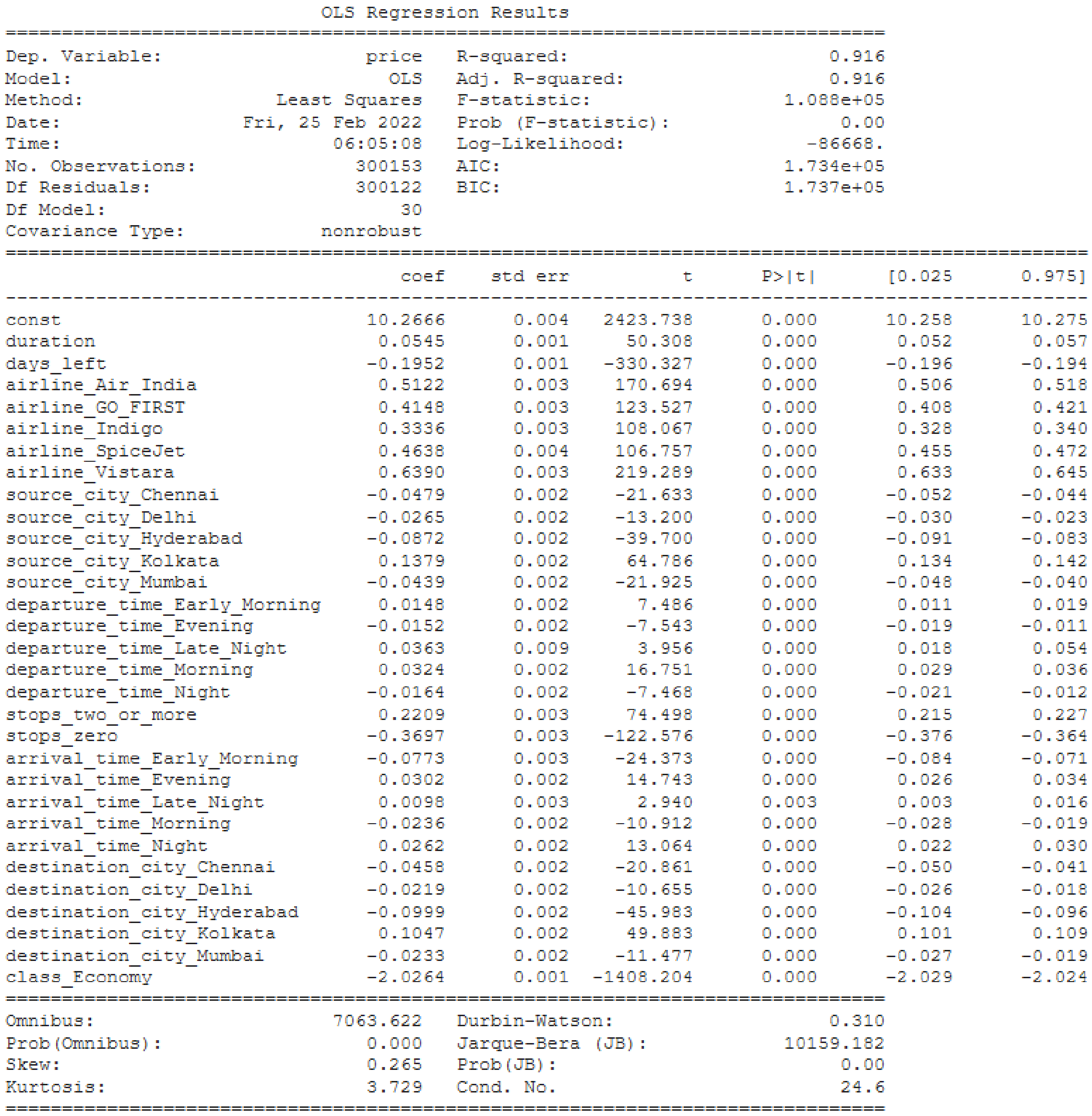
```
Contingency Table:
```

```
stops          No   Yes  Total
source_city
Bangalore      47   475    522
Chennai        36   333    369
Delhi         120   502    622
Hyderabad      34   395    429
Kolkata        49   401    450
Mumbai         84   524    608
Total         370  2630   3000
```

```
Chi-Square test to check relationship
Null Hypothesis:  The two categorical variables source_city and stops have no
relationship.
Alternate Hypothesis: There is a relationship between two categorical variables
source_city and stops.

Chi-Statistics:45.294, p-value:0.0

At 95% Confidence:
Reject the Null Hypothesis
```

Hence there is a relationship between source city and stops variable.

MULTIPLE LINEAR REGRESSION



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.916
Model:                            OLS   Adj. R-squared:                  0.916
Method:                 Least Squares   F-statistic:                 1.088e+05
Date:                Fri, 25 Feb 2022   Prob (F-statistic):               0.00
Time:                        06:05:08   Log-Likelihood:                -86668.
No. Observations:              300153   AIC:                         1.734e+05
Df Residuals:                  300122   BIC:                         1.737e+05
Df Model:                          30
Covariance Type:            nonrobust
================================================================================================
                                   coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const                           10.2666      0.004   2423.738      0.000      10.258      10.275
duration                         0.0545      0.001     50.308      0.000       0.052       0.057
days_left                       -0.1952      0.001   -330.327      0.000      -0.196      -0.194
airline_Air_India                0.5122      0.003    170.694      0.000       0.506       0.518
airline_GO_FIRST                 0.4148      0.003    123.527      0.000       0.408       0.421
airline_Indigo                   0.3336      0.003    108.067      0.000       0.328       0.340
airline_SpiceJet                 0.4638      0.004    106.757      0.000       0.455       0.472
airline_Vistara                  0.6390      0.003    219.289      0.000       0.633       0.645
source_city_Chennai             -0.0479      0.002    -21.633      0.000      -0.052      -0.044
source_city_Delhi               -0.0265      0.002    -13.200      0.000      -0.030      -0.023
source_city_Hyderabad           -0.0872      0.002    -39.700      0.000      -0.091      -0.083
source_city_Kolkata              0.1379      0.002     64.786      0.000       0.134       0.142
source_city_Mumbai              -0.0439      0.002    -21.925      0.000      -0.048      -0.040
departure_time_Early_Morning     0.0148      0.002      7.486      0.000       0.011       0.019
departure_time_Evening          -0.0152      0.002     -7.543      0.000      -0.019      -0.011
departure_time_Late_Night        0.0363      0.009      3.956      0.000       0.018       0.054
departure_time_Morning           0.0324      0.002     16.751      0.000       0.029       0.036
departure_time_Night            -0.0164      0.002     -7.468      0.000      -0.021      -0.012
stops_two_or_more                0.2209      0.003     74.498      0.000       0.215       0.227
stops_zero                      -0.3697      0.003   -122.576      0.000      -0.376      -0.364
arrival_time_Early_Morning      -0.0773      0.003    -24.373      0.000      -0.084      -0.071
arrival_time_Evening             0.0302      0.002     14.743      0.000       0.026       0.034
arrival_time_Late_Night          0.0098      0.003      2.940      0.003       0.003       0.016
arrival_time_Morning            -0.0236      0.002    -10.912      0.000      -0.028      -0.019
arrival_time_Night               0.0262      0.002     13.064      0.000       0.022       0.030
destination_city_Chennai        -0.0458      0.002    -20.861      0.000      -0.050      -0.041
destination_city_Delhi          -0.0219      0.002    -10.655      0.000      -0.026      -0.018
destination_city_Hyderabad      -0.0999      0.002    -45.983      0.000      -0.104      -0.096
destination_city_Kolkata         0.1047      0.002     49.883      0.000       0.101       0.109
destination_city_Mumbai         -0.0233      0.002    -11.477      0.000      -0.027      -0.019
class_Economy                   -2.0264      0.001  -1408.204      0.000      -2.029      -2.024
==============================================================================
Omnibus:                     7063.622   Durbin-Watson:                   0.310
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            10159.182
Skew:                           0.265   Prob(JB):                         0.00
Kurtosis:                       3.729   Cond. No.                         24.6
==============================================================================
```

```
R-squared: 0.9158
Mean Absolute error:  0.2543
Root Mean Squared Error: 0.3229
```

Inference Drawn,

1. P-Value of F-statistics is significant saying that the at least one beta coefficient of the independent feature is nonzero and that we may go ahead to analyse the regression output.
2. R-Squared value is 0.916, saying that about 91.6% variance in target variable is explained by all the independent variables.
3. Adjusted R-Squared value is same as R-Squared, telling us that all the independent variables is adding something to explain the target variable. But this inference may not be significant as the dataset has about 3 lakh datapoints.
4. All the variable coefficient is significant as the p-Value is less than 0.05.

Broad Inference:

Dummy Variable Inference:

1. Airlines:
    a. When all the airline dummy variable is zero then it will give the price for the airline AirAsia.
    b. We see that the Vistara airline has the maximum coefficient (dummy intercept) of 0.639 telling us that log price goes up by a value of 0.639 as we change the airline from AirAsia to Vistara.
    c. Since all the coefficient for airline is positive, it tells us that AirAsia has the minimum ticket price among all the airlines.

2. Source City:
    a. When all the source city dummy variable is zero then it will give the price for the flights from Bangalore.
    b. We see that the source city Kolkata has the maximum coefficient (dummy intercept) of 0.138 telling us that log price goes up by a value of 0.138 if the flight originates from Kolkata in place of Bangalore.
    c. Source City Hyderabad has the most negative coefficient, telling us that flight originating from Hyderabad cost less.

3. Destination City:
    a. When all the destination city dummy variable is zero then it will give the price for the flights landing in Bangalore.
    b. We see that the Destination city Kolkata has the maximum coefficient (dummy intercept) of 0.105 telling us that log price goes up by a value of 0.105 if the flight lands in Kolkata in place of Bangalore.
    c. Source City Hyderabad has the most negative coefficient, telling us that flight landing in Hyderabad cost less.

4. Stops:
    a. When all the stops dummy variable is zero then it will give the price for the flights with one stop.
    b. We see that the Flights with stops two or more has the maximum coefficient (dummy intercept) of 0.221 telling us that log price goes up by a value of 0.221 if the flight number of stops increases from one to two or more.
    c. Flights with no stop has the most negative coefficient, telling us that flight having no stop cost less.

5. Departure Time:
    a. When all the Departure Time dummy variable is zero then it will give the price for the flights departing at Afternoon.
    b. We see that the Flights departing at late night has the maximum coefficient (dummy intercept) of 0.036 telling us that log price goes up by a value of 0. 036 if the flight departs at late night instead of afternoon.
    c. Flights departing at night has the most negative coefficient, telling us that night travel is cheaper.

6. Arrival Time:
    a. When all the Arrival Time dummy variable is zero then it will give the price for the flights arriving at location in Afternoon.
    b. We see that the Flights arriving in evening has the maximum coefficient (dummy intercept) of 0.030 telling us that log price goes up by a value of 0. 030 if the flight arrives at location in evening instead of afternoon.
    c. Flights arriving at location in Early Morning has the most negative coefficient.

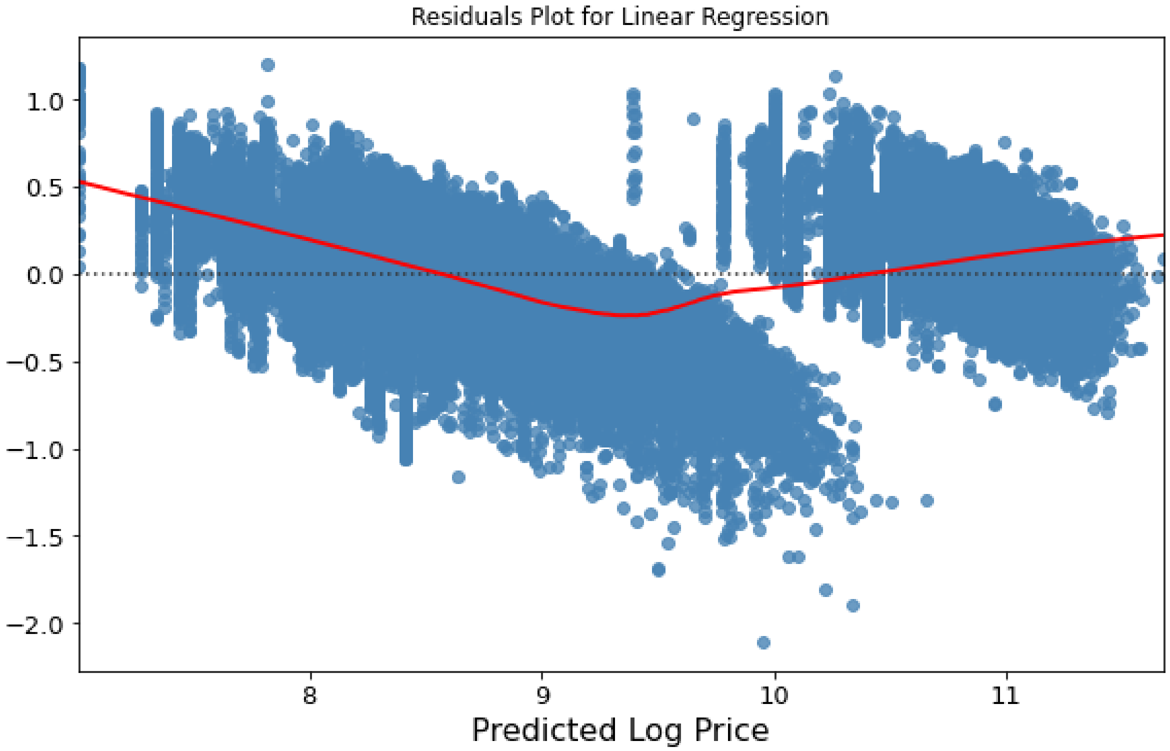
7. Class:
    a. When the Class dummy variable is zero then it will give flight price for Business Class.
    b. We see that the Flights of Economy Class has a negative coefficient (dummy intercept) of -2.026 telling us that log price goes down significantly if the flight class changes from Business to Economy.

Continuous Variable Inference:

1. Days Left has a negative slope, telling us that with increase in number of days between booking and travel, price goes down.
2. Duration has a small positive slope, telling us that with increase in flight's duration price goes up.

ASSUMPTIONS CHECK:

1. **Linearity:**

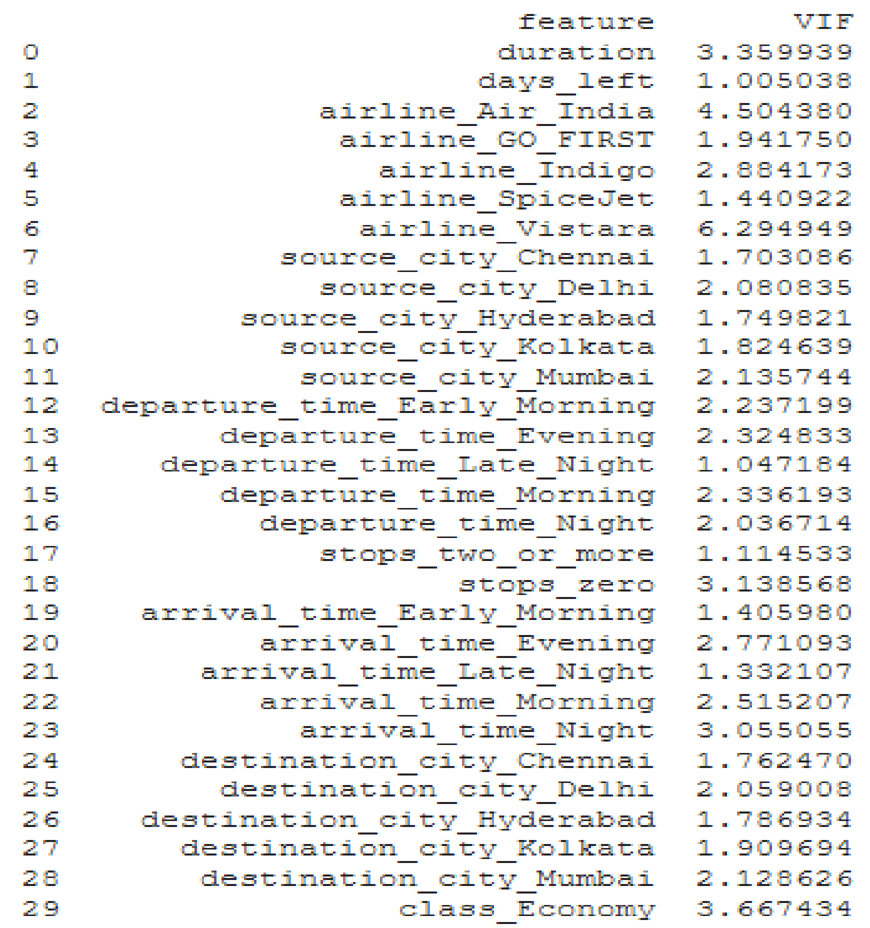


Points are asymmetrically distributed around the horizontal line in the Residual vs Predicted. The inspection of the plot shows that the linearity assumption is not satisfied.
Also, residuals have varying variance, may be heteroskedastic.

2. **Mean of the residuals is zero:**

Mean of residuals: -0.0009325520886841913
Very Small, can be assumed to be zero.

3. **No Multicollinearity:**



```
                         feature       VIF
0                       duration  3.359939
1                      days_left  1.005038
2              airline_Air_India  4.504380
3               airline_GO_FIRST  1.941750
4                 airline_Indigo  2.884173
5               airline_SpiceJet  1.440922
6                airline_Vistara  6.294949
7            source_city_Chennai  1.703086
8              source_city_Delhi  2.080835
9          source_city_Hyderabad  1.749821
10           source_city_Kolkata  1.824639
11            source_city_Mumbai  2.135744
12  departure_time_Early_Morning  2.237199
13        departure_time_Evening  2.324833
14     departure_time_Late_Night  1.047184
15        departure_time_Morning  2.336193
16          departure_time_Night  2.036714
17             stops_two_or_more  1.114533
18                    stops_zero  3.138568
19    arrival_time_Early_Morning  1.405980
20          arrival_time_Evening  2.771093
21       arrival_time_Late_Night  1.332107
22          arrival_time_Morning  2.515207
23            arrival_time_Night  3.055055
24      destination_city_Chennai  1.762470
25        destination_city_Delhi  2.059008
26    destination_city_Hyderabad  1.786934
27      destination_city_Kolkata  1.909694
28       destination_city_Mumbai  2.128626
29                 class_Economy  3.667434
```

All Below 10.
No
Multicollinearity

4. **Homoscedasticity:**

Performing White’s test:

White’s test uses the following null and alternative hypotheses:

1. Null (H0): Homoscedasticity is present (residuals are equally scattered)
2. Alternative (HA): Heteroscedasticity is present (residuals are not equally scattered)

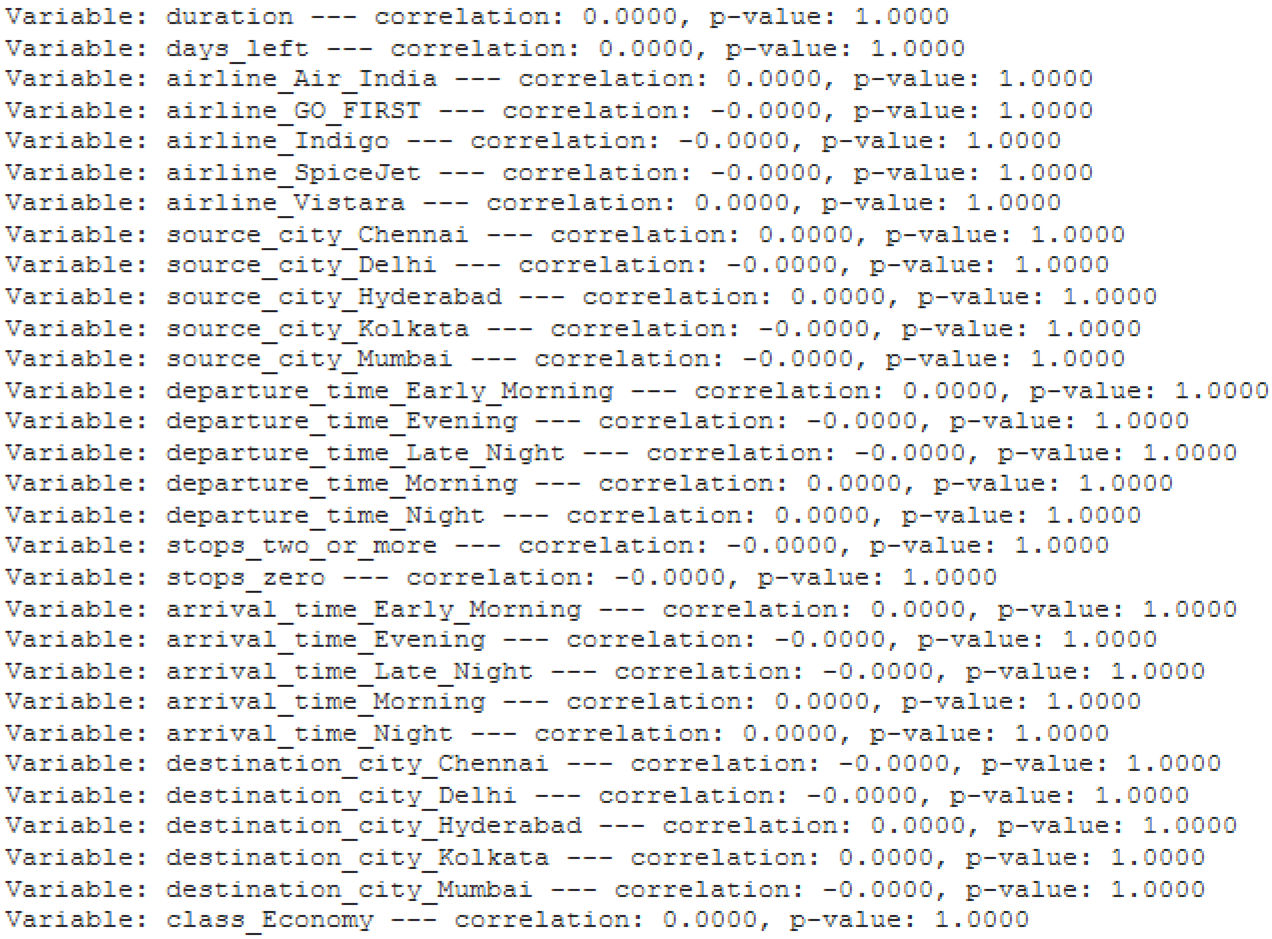
'Test Statistic': 48486.187869696,   'Test Statistic p-value': 0.0
'F-Statistic': 143.29953751437176,   'F-Test p-value': 0.0

Rejecting the Null Hypothesis, Heteroscedasticity is present.

5. **No Correlation between Features and Residuals:**

Pearson r correlation test:

1. Null(H0): No Correlation present.
2. Alternate(H1): Correlation present.

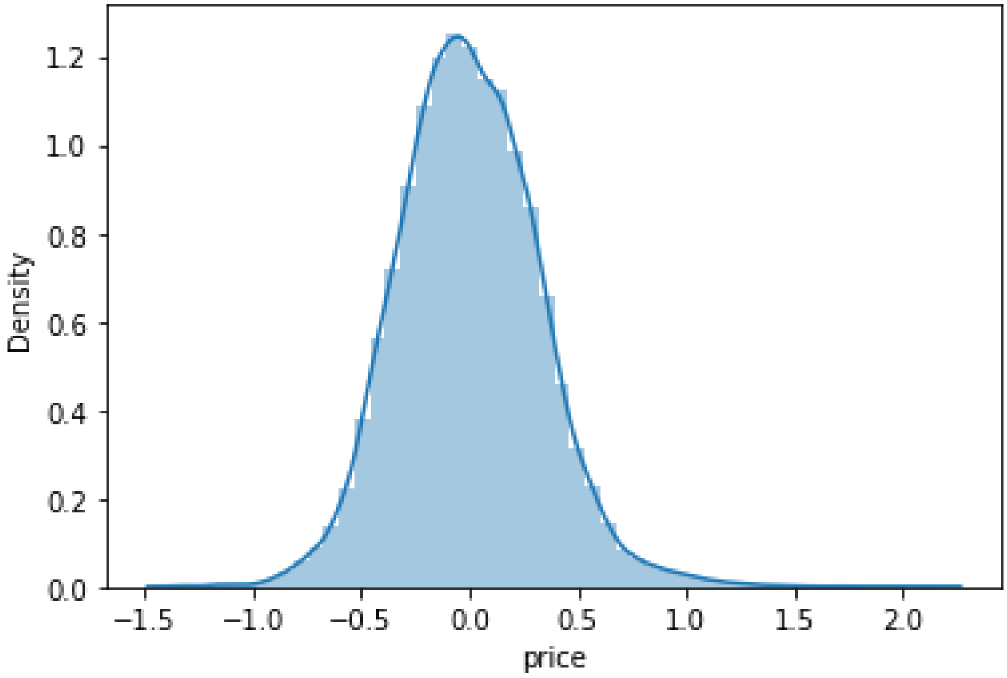


```
Variable: duration --- correlation: 0.0000, p-value: 1.0000
Variable: days_left --- correlation: 0.0000, p-value: 1.0000
Variable: airline_Air_India --- correlation: 0.0000, p-value: 1.0000
Variable: airline_GO_FIRST --- correlation: -0.0000, p-value: 1.0000
Variable: airline_Indigo --- correlation: -0.0000, p-value: 1.0000
Variable: airline_SpiceJet --- correlation: -0.0000, p-value: 1.0000
Variable: airline_Vistara --- correlation: 0.0000, p-value: 1.0000
Variable: source_city_Chennai --- correlation: 0.0000, p-value: 1.0000
Variable: source_city_Delhi --- correlation: -0.0000, p-value: 1.0000
Variable: source_city_Hyderabad --- correlation: 0.0000, p-value: 1.0000
Variable: source_city_Kolkata --- correlation: -0.0000, p-value: 1.0000
Variable: source_city_Mumbai --- correlation: -0.0000, p-value: 1.0000
Variable: departure_time_Early_Morning --- correlation: 0.0000, p-value: 1.0000
Variable: departure_time_Evening --- correlation: -0.0000, p-value: 1.0000
Variable: departure_time_Late_Night --- correlation: -0.0000, p-value: 1.0000
Variable: departure_time_Morning --- correlation: 0.0000, p-value: 1.0000
Variable: departure_time_Night --- correlation: 0.0000, p-value: 1.0000
Variable: stops_two_or_more --- correlation: -0.0000, p-value: 1.0000
Variable: stops_zero --- correlation: -0.0000, p-value: 1.0000
Variable: arrival_time_Early_Morning --- correlation: 0.0000, p-value: 1.0000
Variable: arrival_time_Evening --- correlation: -0.0000, p-value: 1.0000
Variable: arrival_time_Late_Night --- correlation: -0.0000, p-value: 1.0000
Variable: arrival_time_Morning --- correlation: 0.0000, p-value: 1.0000
Variable: arrival_time_Night --- correlation: 0.0000, p-value: 1.0000
Variable: destination_city_Chennai --- correlation: -0.0000, p-value: 1.0000
Variable: destination_city_Delhi --- correlation: -0.0000, p-value: 1.0000
Variable: destination_city_Hyderabad --- correlation: 0.0000, p-value: 1.0000
Variable: destination_city_Kolkata --- correlation: 0.0000, p-value: 1.0000
Variable: destination_city_Mumbai --- correlation: -0.0000, p-value: 1.0000
Variable: class_Economy --- correlation: 0.0000, p-value: 1.0000
```

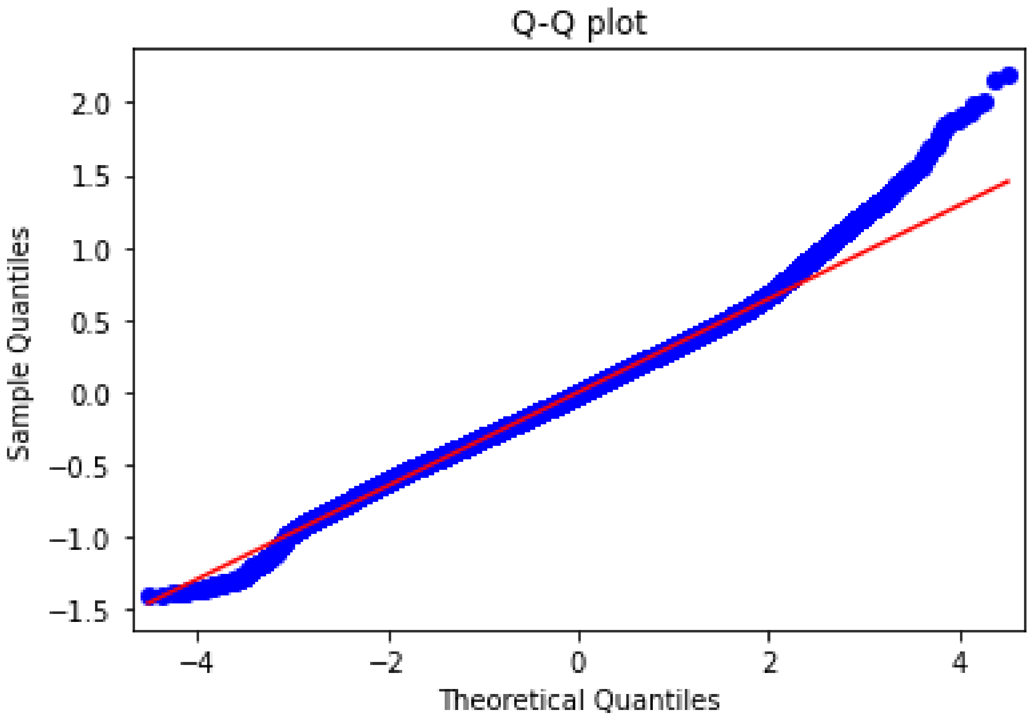
For all the variables we reject the accept the null hypothesis.
Hence, no correlation found between variable and residual.

6. **Normality of the residuals:**

Residual Distribution Plot-



QQ Plot-



Jarque-Bera test ---- statistic: 10159.1824, p-value: 0.0
Shapiro-Wilk test ---- statistic: 0.9942, p-value: 0.0000
Kolmogorov-Smirnov test ---- statistic: 0.2580, p-value: 0.0000

1. Null (H0): Normally Distributed
2. Alternate(H1): Not normal distribution

From the results above we can infer that the residuals do not follow Gaussian distribution. From the shape of the QQ plot, as well as rejecting the null hypothesis in all statistical tests

7. Residuals are not autocorrelated:

The Durbin Watson test reports a test statistic, with a value from 0 to 4, where:

- 2 is no autocorrelation.
- 0 to <2 is positive autocorrelation
- >2 to 4 is negative autocorrelation

Durbin-Watson: 2.009205391097966
Little to no autocorrelation

Assumption satisfied; residuals are not autocorrelated.

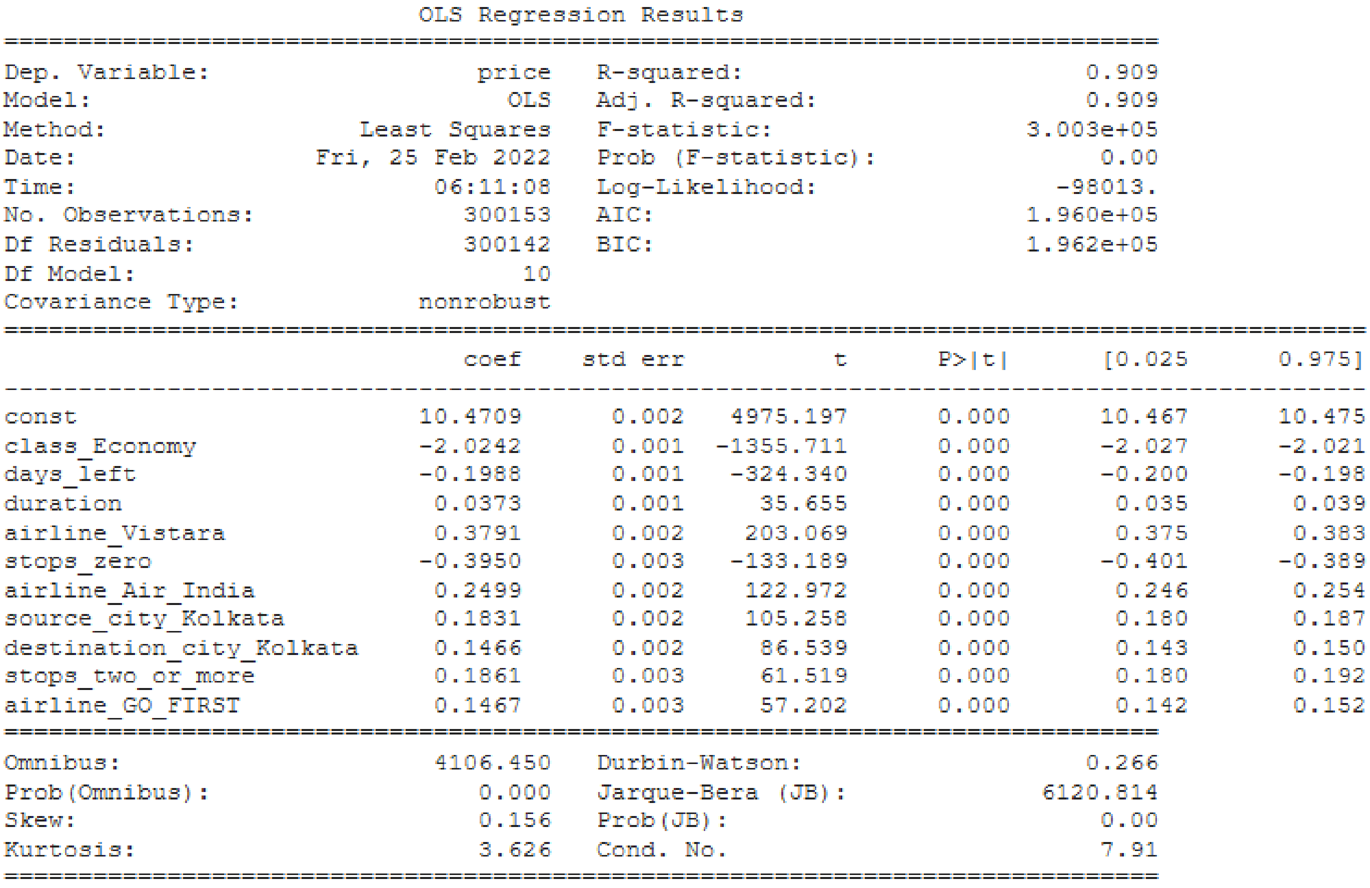
Many of the OLS assumptions were unsatisfied, Hence the OLS regression results are invalid.

STEPWISE REGRESSION:

Stepwise Regression was performed to see if all the variables are adding equally to explain the target.

Inference: We notice that 10 independent variables explain almost 90.9% of the variability in target variable. And only 0.7% of extra variability is explained by adding other 20 variables.

Regression output with 10 independent variables:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.909
Model:                            OLS   Adj. R-squared:                  0.909
Method:                 Least Squares   F-statistic:                 3.003e+05
Date:                Fri, 25 Feb 2022   Prob (F-statistic):               0.00
Time:                        06:11:08   Log-Likelihood:                -98013.
No. Observations:              300153   AIC:                         1.960e+05
Df Residuals:                  300142   BIC:                         1.962e+05
Df Model:                          10
Covariance Type:            nonrobust
============================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                      10.4709      0.002   4975.197      0.000      10.467      10.475
class_Economy              -2.0242      0.001  -1355.711      0.000      -2.027      -2.021
days_left                  -0.1988      0.001   -324.340      0.000      -0.200      -0.198
duration                    0.0373      0.001     35.655      0.000       0.035       0.039
airline_Vistara             0.3791      0.002    203.069      0.000       0.375       0.383
stops_zero                 -0.3950      0.003   -133.189      0.000      -0.401      -0.389
airline_Air_India           0.2499      0.002    122.972      0.000       0.246       0.254
source_city_Kolkata         0.1831      0.002    105.258      0.000       0.180       0.187
destination_city_Kolkata    0.1466      0.002     86.539      0.000       0.143       0.150
stops_two_or_more           0.1861      0.003     61.519      0.000       0.180       0.192
airline_GO_FIRST            0.1467      0.003     57.202      0.000       0.142       0.152
==============================================================================
Omnibus:                     4106.450   Durbin-Watson:                   0.266
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6120.814
Skew:                           0.156   Prob(JB):                         0.00
Kurtosis:                       3.626   Cond. No.                         7.91
==============================================================================
```

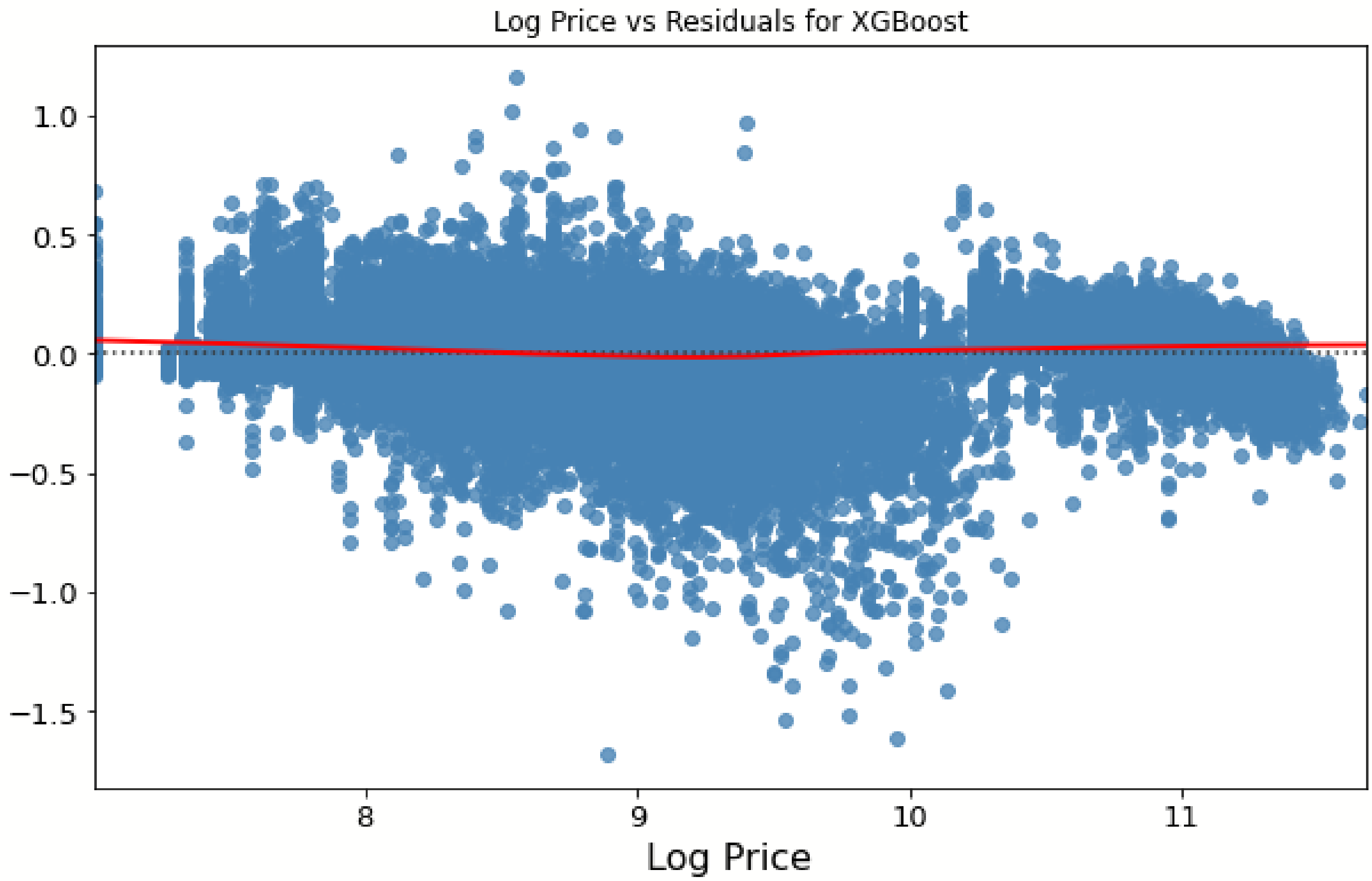
Mean Absolute Error: 0.26323902322709436
Root Mean Squared Error: 0.33541207995385613

FINAL MODEL:

After comparing different regression algorithms, XGBoost Regressor was found to have the best result. The model was tuned with the help of Random Search CV.

Algorithm with tuned hyperparameters:

```
XGBRegressor (colsample_bylevel= 0.6,
              colsample_bytree= 0.7,
              learning_rate= 0.02,
              max_depth = 12,
              n_estimators= 1000,
              subsample= 0.6).
```



R-squared: 0.9844
Mean Absolute error: 0.0875
Root Mean Squared Error: 0.1392

XGBoost is giving results far better than Linear Regression.

CONCLUSION

Our research yielded some interesting observations regarding the factors impacting the price.

The results are statistically validated with hypothesis testing with 5% significance.

We will conclude by summarizing the answers to our research questions:

1. Ticket price depends on various parameters.
2. Tickets should be bought well in advance to get better deal. Tickets purchased 1-2 days before departure date costs much more than the tickets bought more than 30 days.
3. Ticket prices vary depending on departure time. Flights departing late at night are less expensive.
4. Tickets Price costs more if flight originates from or lands in Kolkata. If Hyderabad is the source or destination city, then the ticket price is the lowest.
5. Business class tickets are much more expensive than Economy class tickets.