FLIP ROBO

FLIGHT PRICE PREDICTION PROJECT

Submitted by:
ANISH ANTONY

ABSTRACT:

Traveling in an airplane is one of the most exciting and expensive modes of travel. Since it is expensive, we must plan well accordingly. In this project we will see how the ticket vary based on advance booking of the flight and also how the price varies as the dates come closer. In this project we will be predicting the flight price from machine learning models. To predict this model, we need to scrape the flight booking details for two months from online websites using selenium web driver, and the data have preprocessed, trained and tested using the regression algorithms. Then it is hypertuned and best algorithm with best parameters is obtained and finally the ticket price of the flight is predicted.

Keywords: Flight ticket, Selenium, Data cleaning, Ticket Price, Regression

CHAPTER I

INTRODUCTION

1.1 Business Problem Framing:

Problem Description:

Anyone who has booked a flight ticket knows how unexpectedly the prices vary. The cheapest available ticket on a given flight gets more and less expensive over time. This usually happens as an attempt to maximize revenue based on - 1. Time of purchase patterns (making sure last-minute purchases are expensive) 2. Keeping the flight as full as they want it (raising prices on a flight which is filling up in order to reduce sales and hold back inventory for those expensive last-minute expensive purchases) So, we have to work on a project where you collect data of flight fares with other features and work to make a model to predict fares of flights.

Business Objectives:

As a Data scientist it is required to apply some data science techniques for the flight ticket price with the available independent variables. That should help the customer to understand how exactly the prices vary with the independent variables. They can accordingly plan book or postpone travel, the purchase strategy etc.

1. Data Collection

We have scraped at least 9929 rows of data. In this section you have to scrape the data of flights from different websites (yatra.com, and makemytrip.com). The number of columns for data is close to 35 based on the data uniformity and availability, the columns are fixed. Generally, these columns are airline name, date of journey, source, destination, route, departure time, arrival time, duration, total stops and the target variable price. it completely depends on the website from which you are fetching the data.

2. Data Analysis:

After cleaning the data, you have to do some analysis on the data. Do airfares change frequently? Do they move in small increments or in large jumps? Do they tend to go up or down over time? What is the best time to buy so that the consumer can save the most by taking the least risk? Does price increase as we get near to departure date? Is Indigo cheaper than Jet Airways? Are morning flights expensive?

3. Model Building:

After collecting the data, you need to build a machine learning model. Before model building do all data pre-processing steps. Try different models with different hyper parameters and select the best model.

Follow the complete life cycle of data science. Include all the steps like

1. Data Cleaning

2. Exploratory Data Analysis

3. Data Pre-processing

4. Model Building

5. Model Evaluation

6. Selecting the best model

1.3 Review of Literature

From the article "Flight Fare Prediction System Using Machine Learning" published in Ijraset Journal For Research in Applied Science and Engineering Technology

Neel Bhosale majorly targeted to uncover underlying trends of flight prices in India using historical data and also to suggest the best time to buy a flight ticket. He made a comparison study among various models in predicting the optimal time to buy the flight ticket and the amount that can be saved if done so. He found that the trends of the prices are highly sensitive to the route, month of departure, day of departure, time of departure, whether the day of departure is a holiday and airline carrier.

Highly competitive routes like most business routes (tier 1 to tier 1 cities like Mumbai-Delhi) had a non-decreasing trend where prices increased as days to departure decreased, however other routes (tier 1 to tier 2 cities like Delhi - Guwahati) had a specific time frame where the prices are minimum.

Moreover, the data also uncovered two basic categories of airline carriers operating in India – the economical group and the luxurious group, and in most cases, the minimum priced flight was a member of the economical group. The data also validated the fact that, there are certain time-periods of the day where the prices are expected to be maximum.

From the article “A Framework for Airfare Price Prediction: A Machine Learning Approach”

Tianyi Wang predicted that the price of an airline ticket is affected by a number of factors, such as flight distance, purchasing time, fuel price, etc. Each carrier has its own proprietary rules and algorithms to set the price accordingly.

Recent advance in Artificial Intelligence (AI) and Machine Learning (ML) makes it possible to infer such rules and model the price variation. This paper proposes a novel application based on two public data sources in the domain of air transportation: the Airline Origin and Destination Survey (DB1B) and the Air Carrier Statistics database (T-100). His proposed framework combines the two databases, together with macroeconomic data, and uses machine learning algorithms to model the quarterly average ticket price based on different origin and destination pairs, as known as the market segment.

1.4 Motivation for the Problem Undertaken

Describe your objective behind to make this project, this domain and what is the motivation behind.

Data is collected from the websites which sell the flight tickets so only limited information can be accessed. Machine Learning algorithms are applied on the dataset to predict the dynamic fare of flights. This gives the predicted values of flight fare to get a flight ticket at minimum cost.

CHAPTER II

Analytical Problem Framing

2.1 Mathematical/ Analytical Modeling of the Problem:

Machine Learning:

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists chooses to use depends on what type of data they want to predict.

Supervised learning:

In this type of machine learning, data scientists supply algorithms with labeled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Supervised learning can be separated into two types of problems when data mining—classification and regression

- Common classification algorithms are linear classifiers, support vector machines (SVM), decision trees, k-nearest neighbor, and random forest, etc.
- Linear regression, logistical regression, and polynomial regression are popular regression algorithms.

Unsupervised learning:

This type of machine learning involves algorithms that train on unlabelled data. The algorithm scans through data sets looking for any meaningful connection. The data that algorithms train on as well as the predictions or recommendations they output are predetermined.

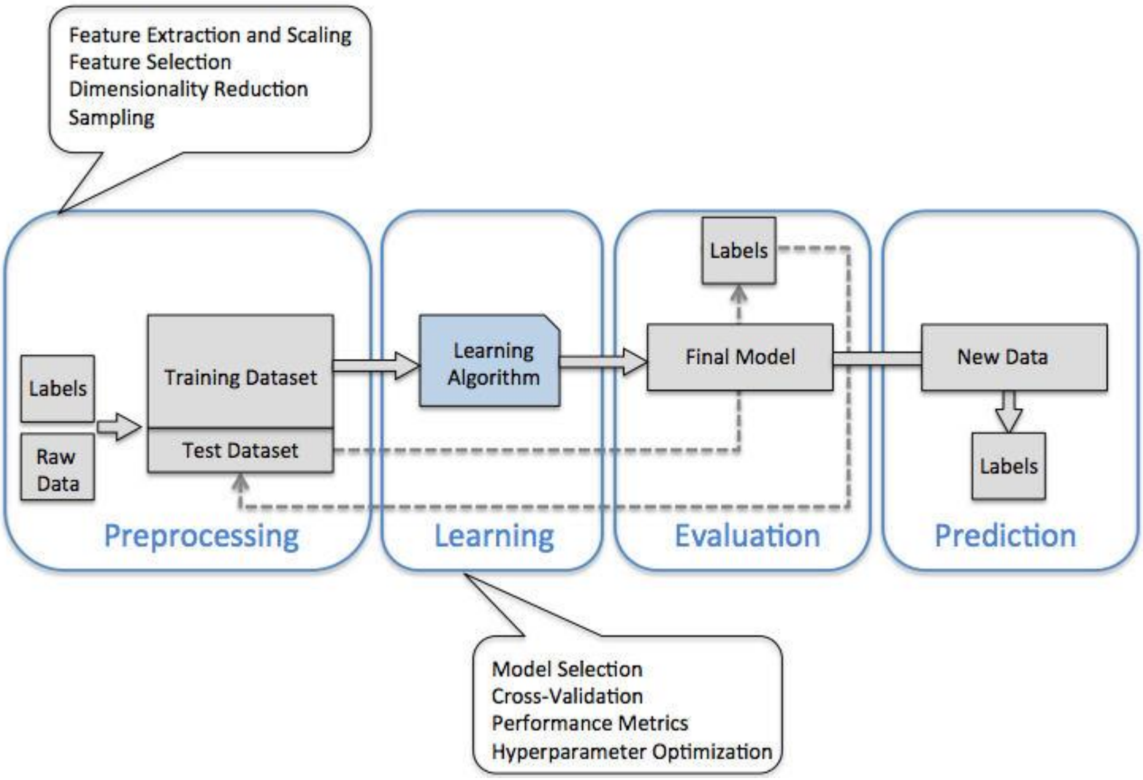
Popular un-supervised algorithms are K-means clustering, affinity propagation etc.

Semi-supervised learning:

This approach to machine learning involves a mix of the two preceding types. Data scientists may feed an algorithm mostly labeled training data, but the model is free to explore the data on its own and develop its own understanding of the data set.

Reinforcement learning:

Data scientists typically use reinforcement learning to teach a machine to complete a multi-step process for which there are clearly defined rules. Data scientists program an algorithm to complete a task and give it positive or negative cues as it works out how to complete a task. But for the most part, the algorithm decides on its own what steps to take along the way.



Linear regression:

Linear regression attempts to model the relationship between two variables by fitting a linear equation to observed data. One variable is considered to be an explanatory variable, and the other is considered to be a dependent variable.

A linear regression line has an equation of the form Y = a + bX, where X is the explanatory variable and Y is the dependent variable. The slope of the line is b, and a is the intercept (the value of y when x = 0).

Random forest Algorithm:

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithms of the same type i.e., multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks.

Statistical Analysis:

Statistical analysis, or statistics, involves collecting, organizing and analysing data based on established principles to identify patterns and trends.

Predictive analysis:

Predictive analysis uses powerful statistical algorithms and machine learning tools to predict future events and behaviour based on new and historical data trends. It is important to note that predictive analysis can only make hypothetical forecasts and the quality of the predictions depends on the accuracy of the underlying data sets.

The terms used for statistical analysis are:

Mean	$\overline{x} = \frac{\sum x}{n}$	x = Observations given n = Total number of observations
Median	If n is odd, then $M = \frac{n+1}{2th}$ term If n is even, then $M = \frac{\left(\frac{n}{2}\right)th\ term+\ \left(\frac{n}{2}+1\right)th\ term}{2}$	n = Total number of observations
Mode	The value which occurs most frequently	
Variance	$= \sigma^2 = \sum \frac{(x-x)^2}{n}$	x = Observations given = Mean n = Total number of observations

Standard Deviation	$S = \sigma = \sqrt{\sum \frac{(x-x)^2}{n}}$	x = Observations given $\bar{x}$ = Mean n = Total number of observations

Z score = $\frac{x-\bar{x}}{\sigma}$

Where,

x = Standardized random variable

$\bar{x}$ = Mean

σ = Standard deviation.

Quartile Formula:

When the set of observation is arranged in an ascending order, then the 25th percentile is given as:

$$Q_1 = (\frac{n+1}{4})^{th} \text{ term}$$

The second quartile or the 50th percentile or the Median is given as:

$$Q_2 = (\frac{n+1}{2})^{th} \text{ term}$$

The third Quartile of the 75th Percentile (Q3) is given as:

$$Q_2 = (\frac{3(n+1)}{4})^{th} \text{ term}$$

IQR = Upper Quartile – Lower Quartile

Regression:

Regression is a statistical technique used to find a relationship between a dependent variable and an independent variable. It helps track how changes in one variable affect changes in another or the effect of one on the other. Regression can show whether the relationship between two variables is weak, strong or varies over a time interval. The regression formula is:

$$Y = a + b(x)$$

Y represents the independent variable, or the data used to predict the dependent variable

x represents the dependent variable which is the variable you want to measure

a represents the y-intercept or the value of y when x equals zero

b represents the slope of the regression graph

Hypothesis testing:

Hypothesis testing is used to test if a conclusion is valid for a specific data set by comparing the data against a certain assumption. The result of the test can nullify the hypothesis, where it is called the null hypothesis or hypothesis 0. Anything that violates the null hypothesis is called the first hypothesis or hypothesis 1.

2.2 Data Sources and their formats:

1. Data Collection Phase

In this project we have scrapped close to 9329 flight ticket booking details of 35 features. Initially, we surfed the internet for flight ticket prices. We searched through the online ticketing websites. The data has been scraped from websites such as Yatra and MakeMyTrip. The data has been scrapped from different locations such as Chennai, Bangalore and Mumbai.

Scrape the necessary details and transform it into a dataframe using python in the jupyter notebook. Combine all individual dataframes and merge into a single dataframe

2. Data Cleaning:

We need to need to clean the dataset since the data is collected from different websites.

For use of access, we extract all the 19 features into 3 columns, because since webscraping for 9329 data takes much more time. These three columns are then split into 19 features that’s the way the data is available

Name column contains the Model, Brand Year and name of the vehicle and it is split into the respective features.

The features used in the dataset are:

- Flight Company
- Flight No
- Flight Class
- Flight Model
- Travel
- Departure Time
- Departure Day
- Departure Date
- Departure Month
- Departure Duration
- Arrival Time
- Arrival Day
- Meal
- Flight Ticket Price
- Luggage
- Connecting Planes
- No of days

The categorical features were standardized and made uniform for all data

Dataset Description:

Columns	**Datatype**	**Unique values**	**Mode Mean**	**Values**
1	Flight Company	object	9	Vistara Premium Economy
2	Flight No	object	446	UK832
3	Flight Class	object	3	Economy
4	Flight Model	object	19	Vistara, Airbus A32-100
5	Travel	object	6	Bengaluru to Mumbai

6	Departure Time	object	22	7
7	Departure Day	object	7	Sat
8	Departure Date	object	28	29
9	Departure Month	object	2	Aug
10	Departure Duration	int32	68	126.74935
11	Arrival Time	int32	23	15.164451
12	Arrival Day	object	7	Sat
13	Meal	object	2	Free Meal
14	Flight Ticket Price	int64	652	19838.348439
15	Luggage	object	5	15
16	Connecting Planes	object	3	Has one connecting plane
17	No of days	int32	5	19.52242

Some of the features were redundant, when comparing different websites, the features were fixed based on the requirement and importance and others were removed.

2.3 Data Preprocessing Done:

Exploratory Data Analysis (EDA):

Some of the features have ‘Not available’ as feature values. It has to replace with mode or mean based on the dataset datatype.

2.4 State the set of assumptions (if any) related to the problem under consideration:

Assumptions for data collections:

Since the data extracted needed to cover most major features, different locations and websites. The data is standardized with respect to yatra website since it most features and different location and other websites data has been benchmarked with these websites. Due to this some of the data has been removed. This data is taken from Mumbai, Bengaluru and Chennai and prices will be dependent on the location.

2.5 Hardware and Software Requirements and Tools Used:

Hardware – PC Windows 10, 4 GB Ram

Software – Google chrome, MS Excel, Python, Selenium webdriver

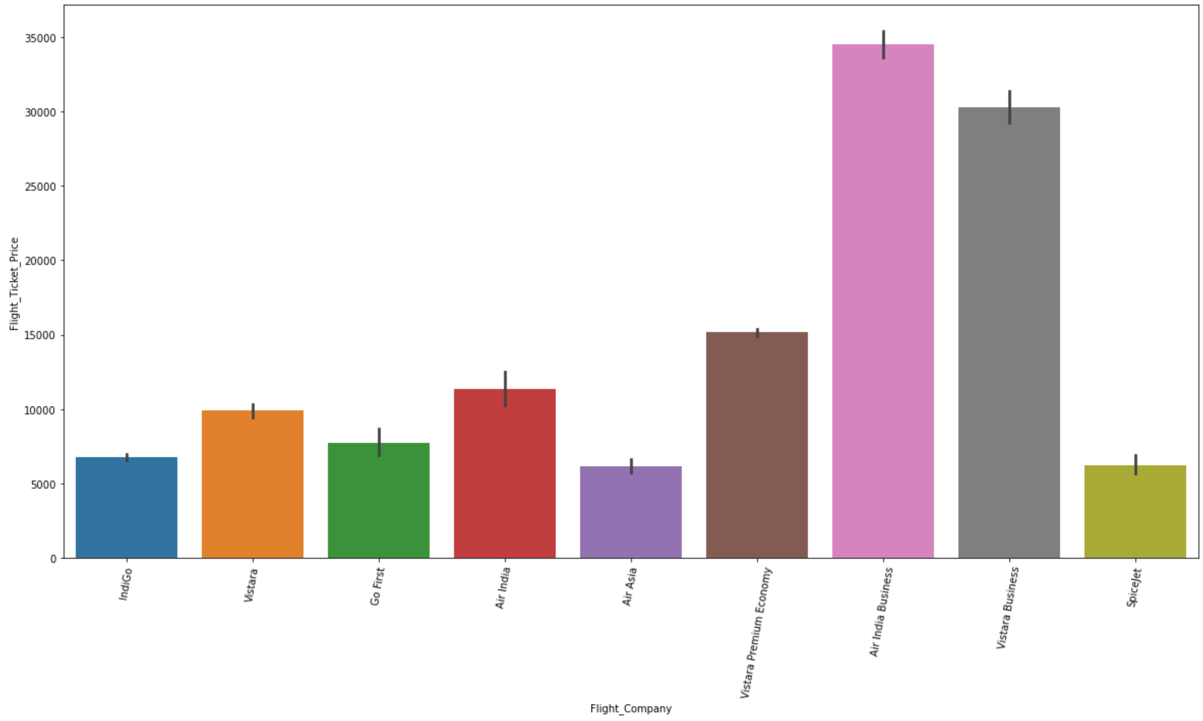
Libraries – Pandas, NumPy, Matplotlib, Seaborn, sklearn, SciPy. Stats

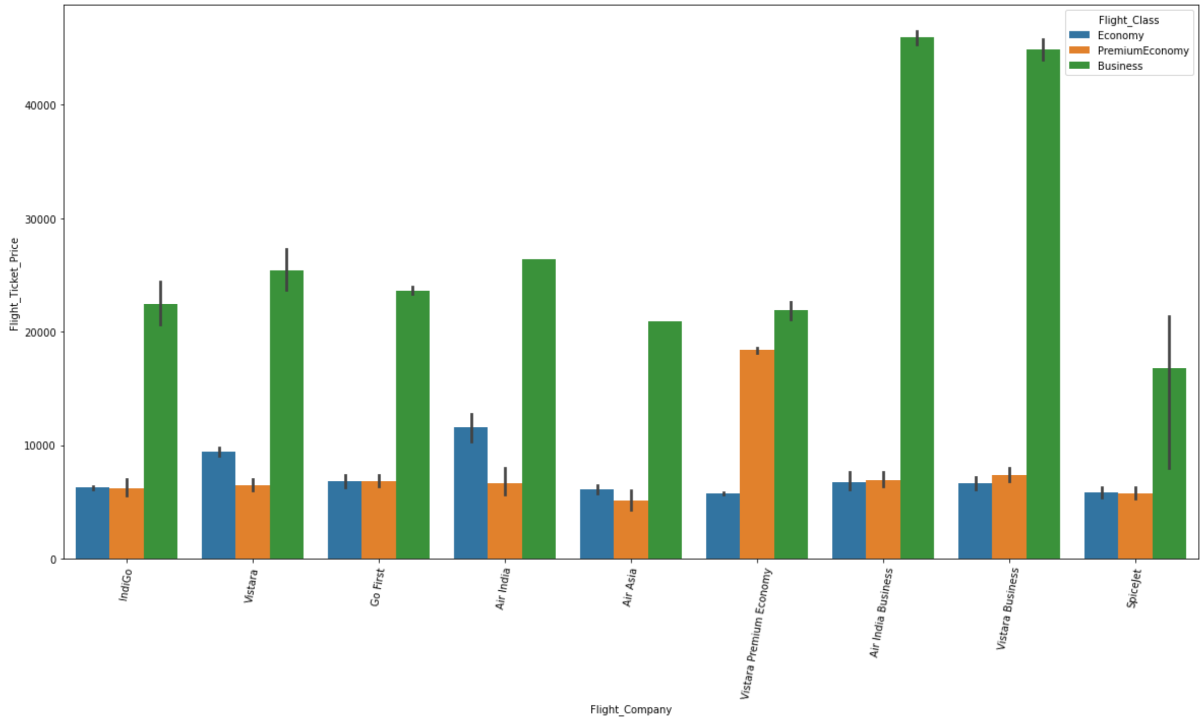
- Browsing – Google Chrome
- Webscraping – Python, Selenium webdriver
- Data cleaning – Python, Pandas, NumPy & SciPy. Stats
- Data visualization – Matplotlib & Seaborn
- Machine learning – Sklearn

2.6 Data Inputs- Logic- Output Relationships:

1.Flight Company:

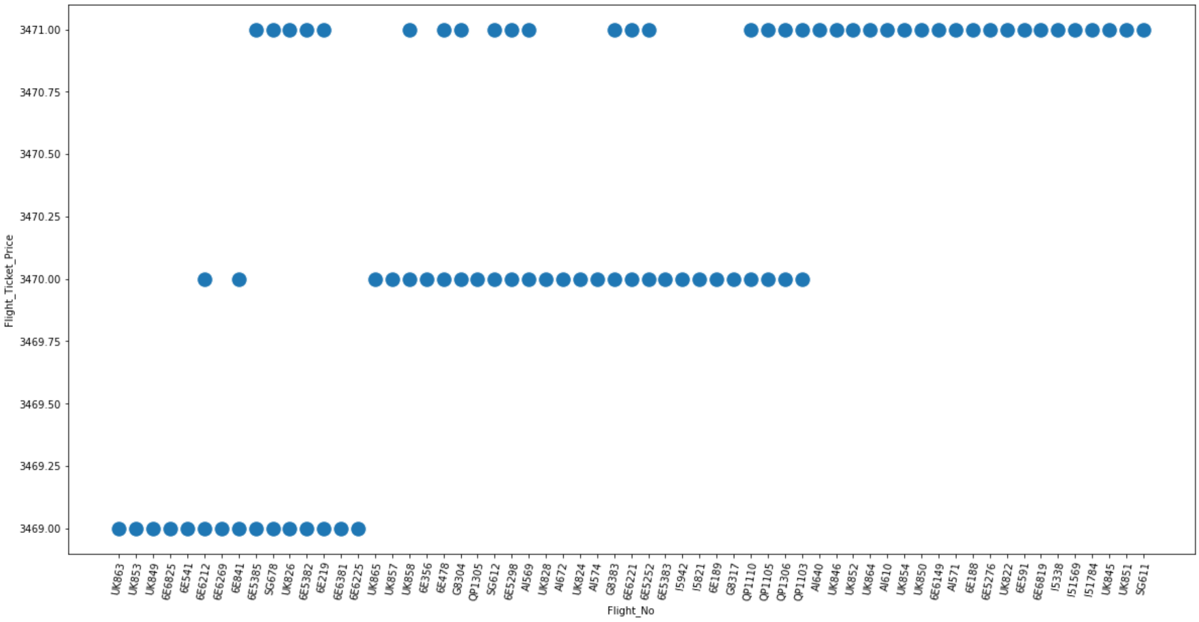
From the below plots, we find that the Indigo, Air Asia and Spice jet flights are cheaper. While Business class flights are generally expensive.





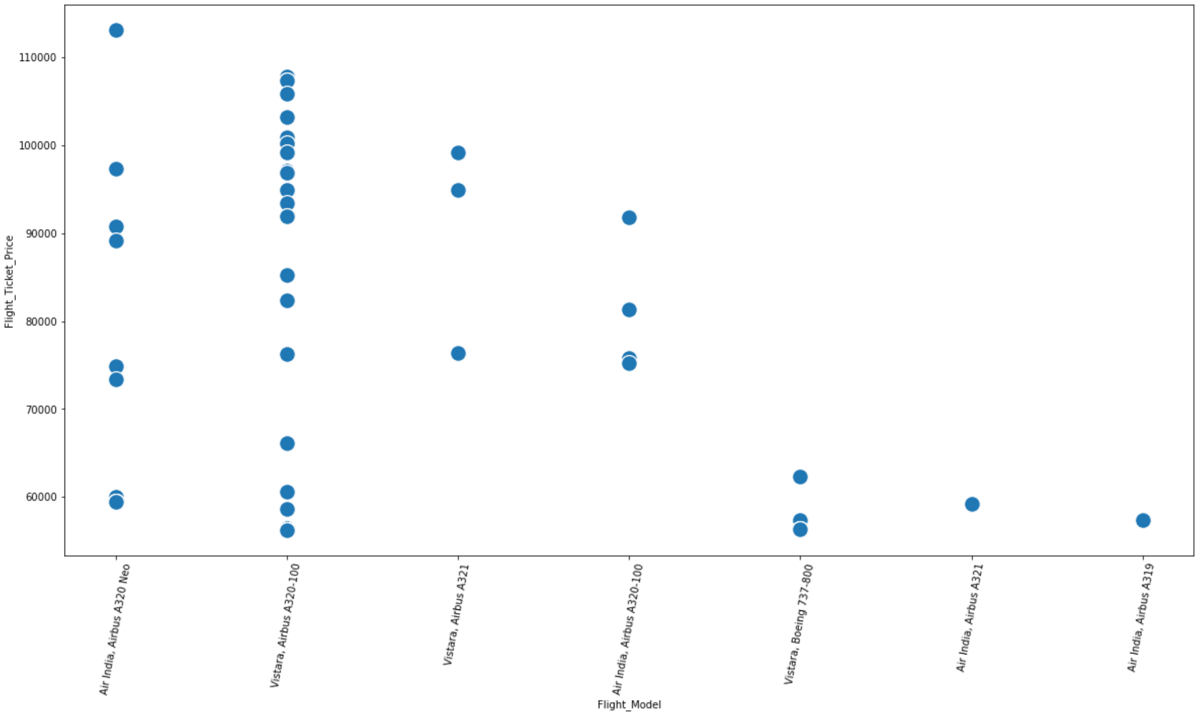
2. Flight No:

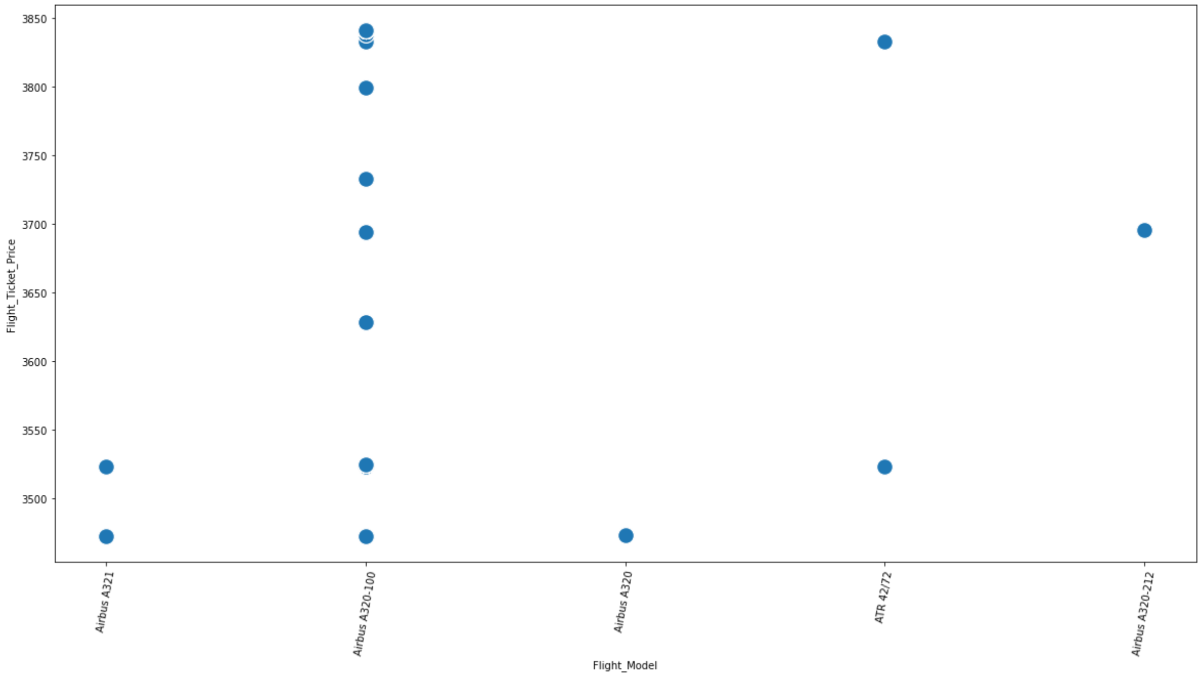
Comparing the Flight, no vs Flight price plot, AI683 is more costly and UK863 is the cheapest.



3.Flight Model:

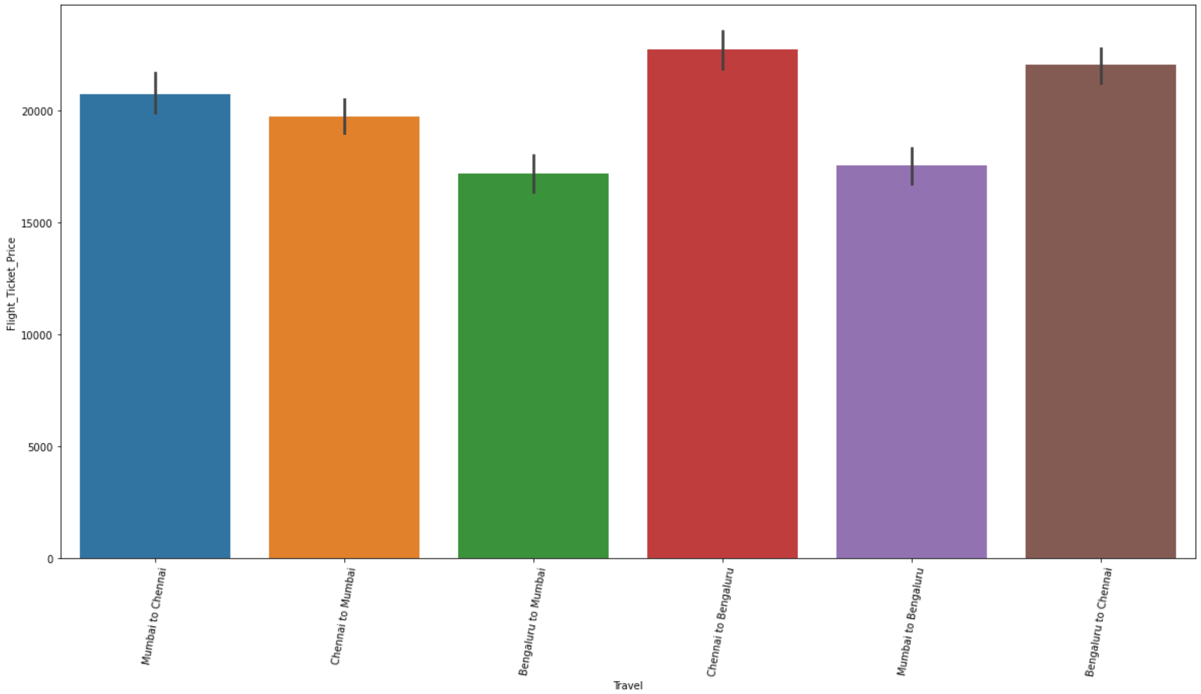
Comparing the plots, Air India Air bus A320 Neo is costly and Airbus 321 is cheaper





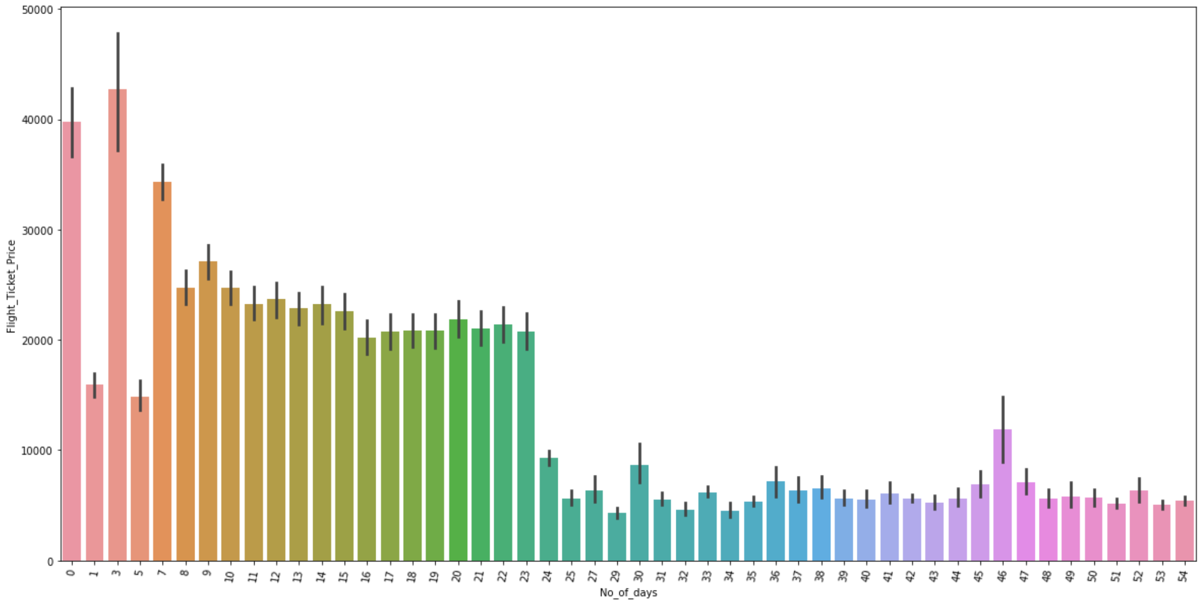
4. Travel:

Comparing the plots, Chennai to Bengaluru is costlier and Bengaluru to Mumbai is cheaper.



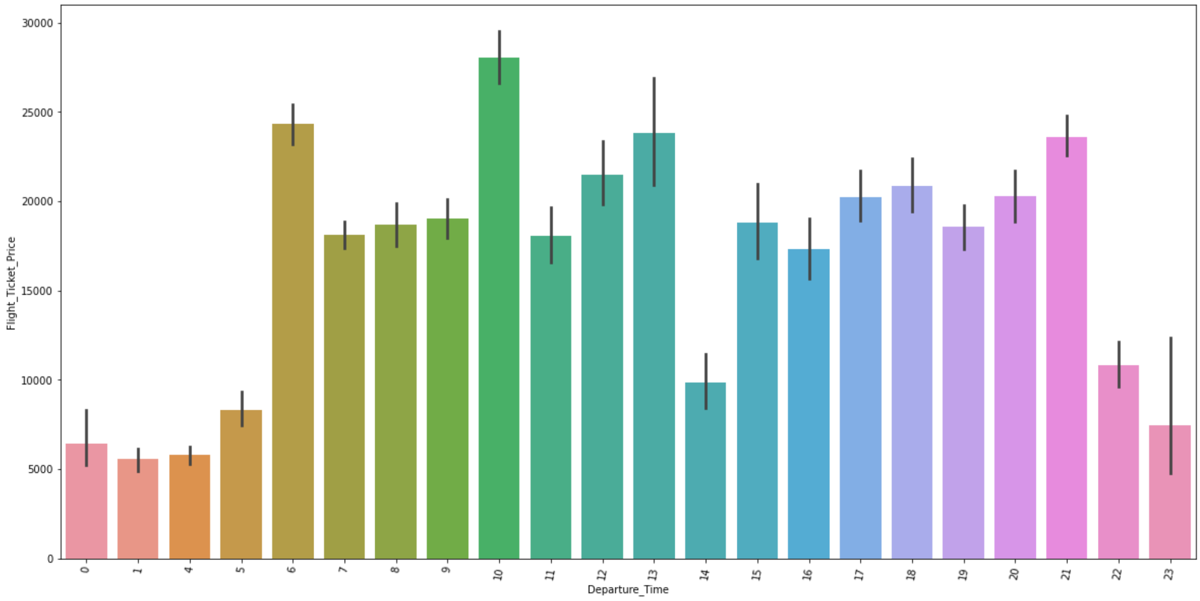
5.Counting No of Days:

From the plots we find that as the day progresses , the cost decreases



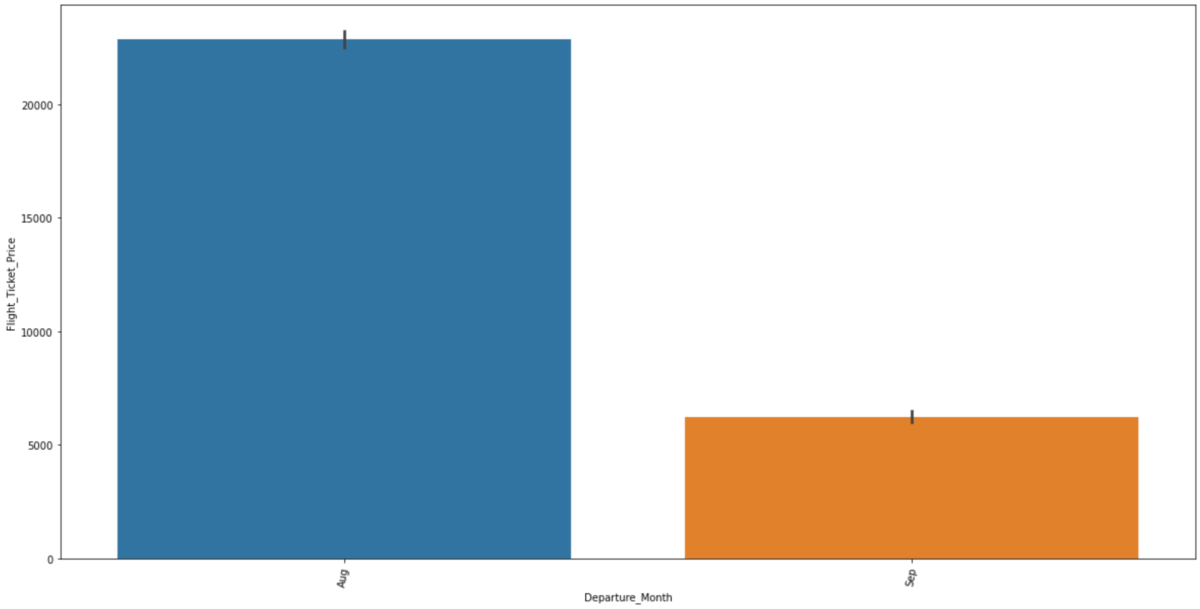
6. Departure Time:

From the plot, we find that the night flights are cheaper than the day.



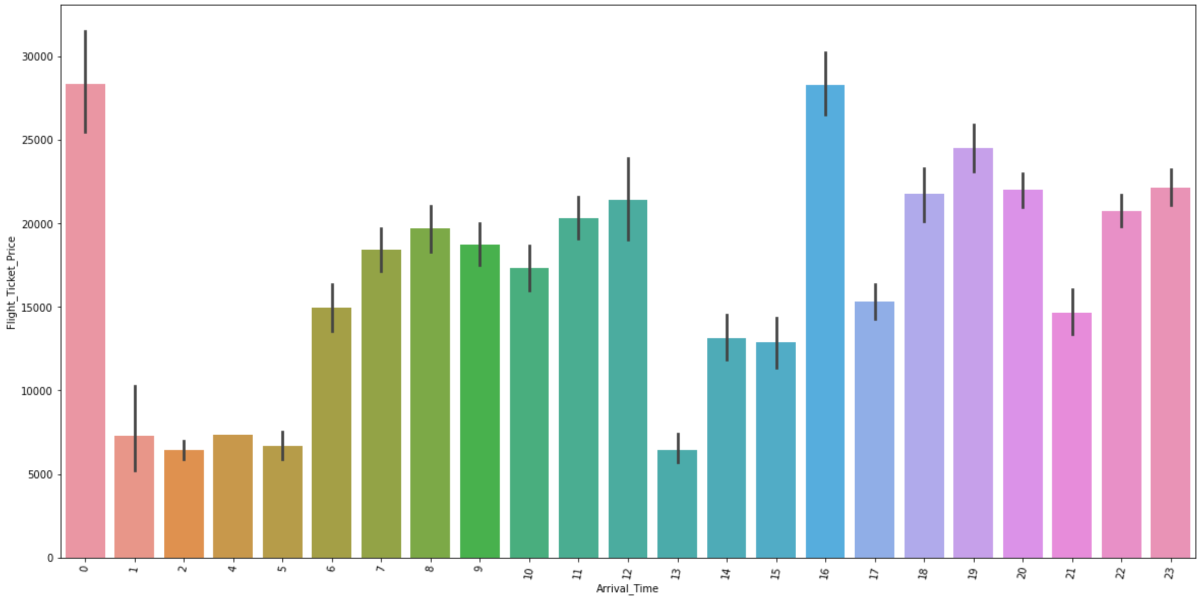
7.Departure Month:

From the plots the current month is always costlier than the next month



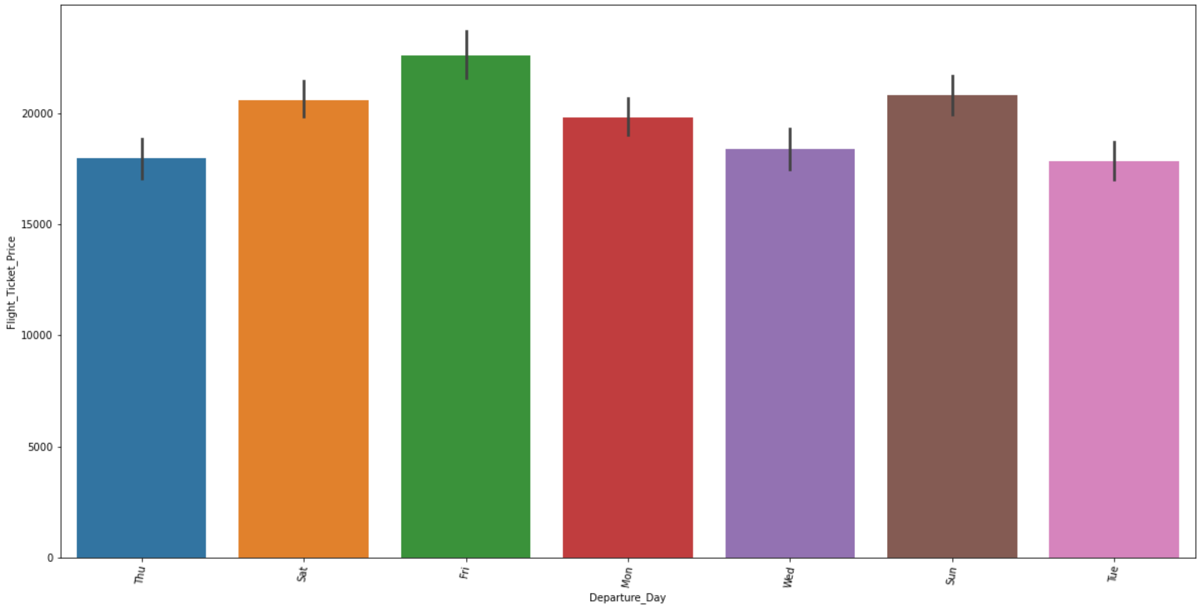
8.Arrival Time:

From the plot the flight arriving between midnight and early morning is cheap



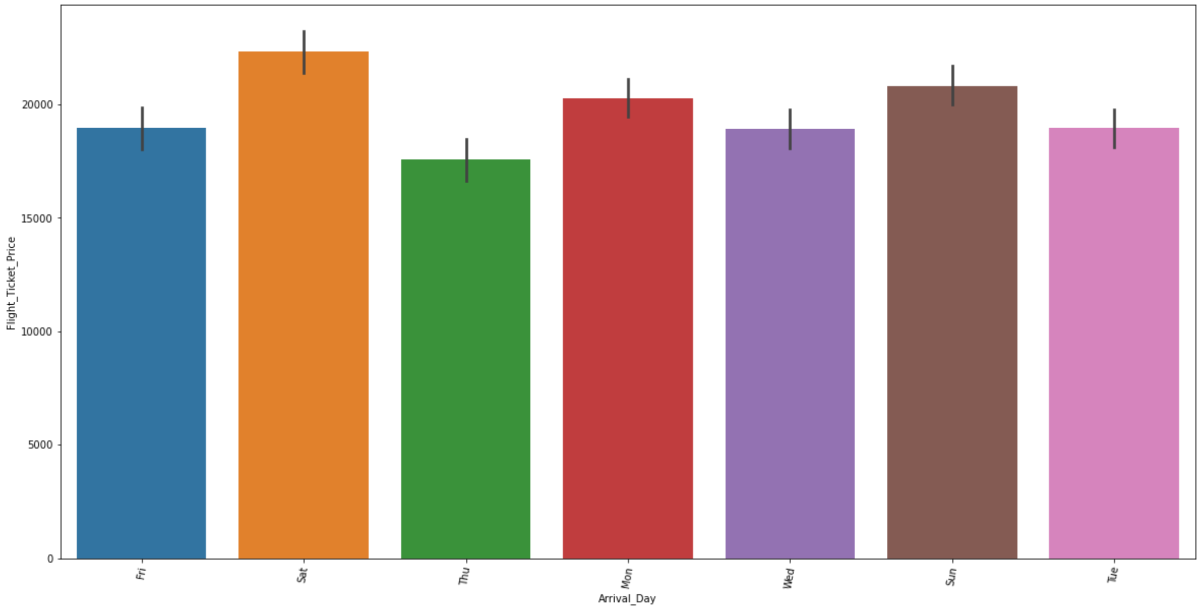
9.Departure Day:

From the plot, we find that weekend flights are costlier when compared mid of the week



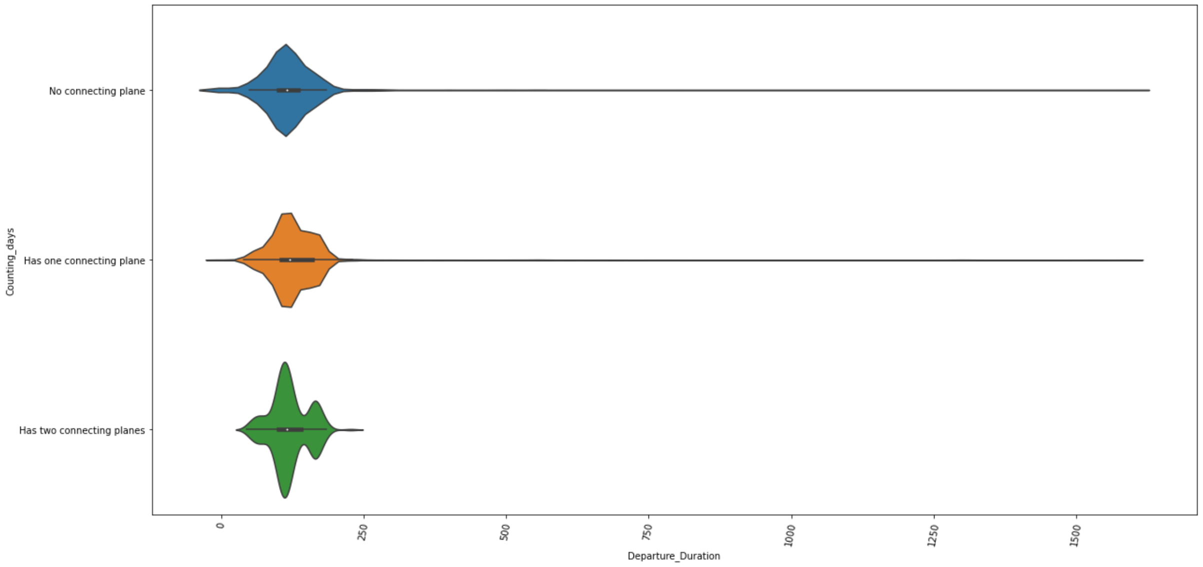
10.Arrival Day:

Similar trend compared to the previous plot,
Departure: Friday, Saturday & Sunday
Arrival: Saturday, Sunday and Monday



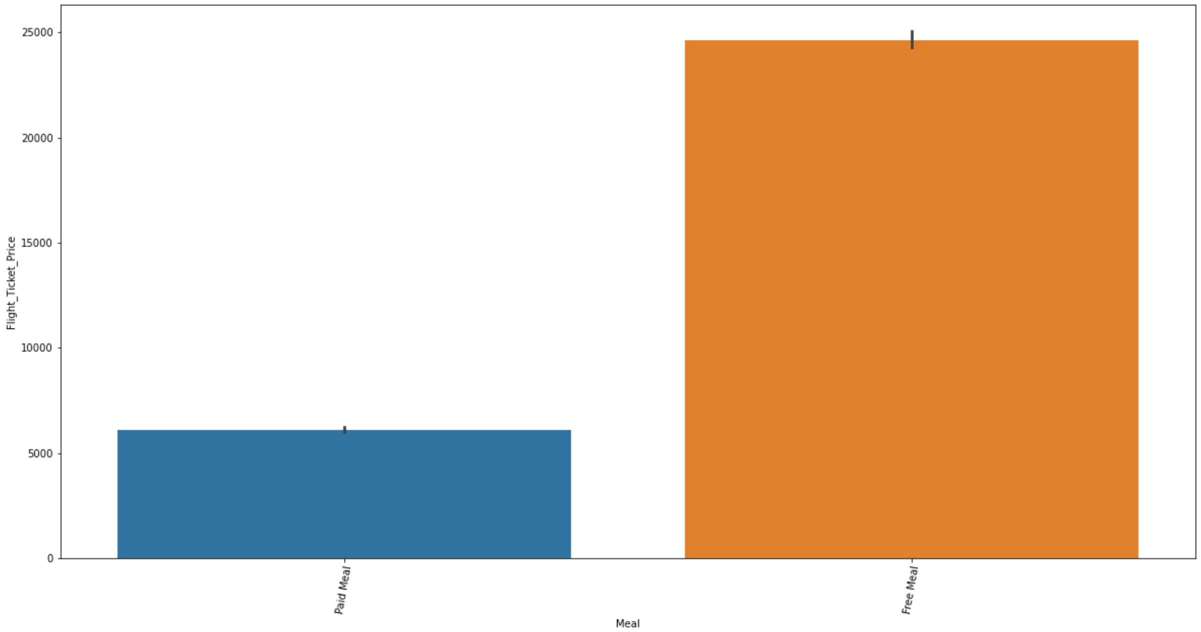
11.Duration:

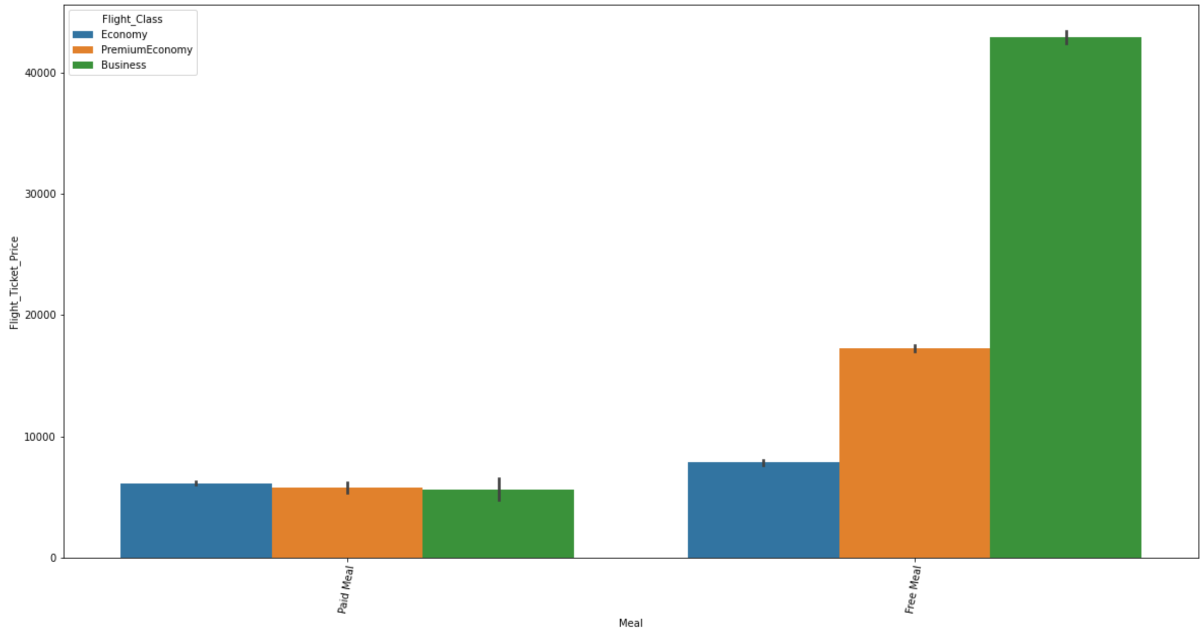
Duration is dependent upon connecting planes. If it has connecting planes then the total duration is dependent upon the connecting flight gap.

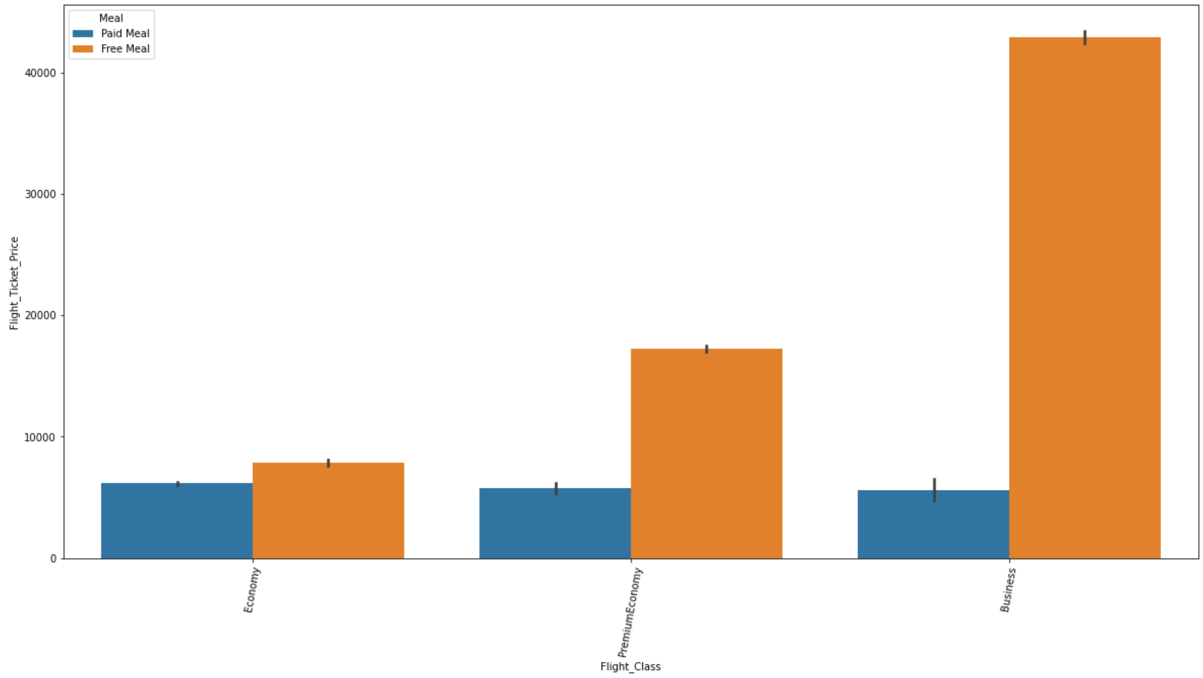


12.Meal:

From the plots, meal is dependent on the luggage, and class of the flight. For free meal the ticket price is more.

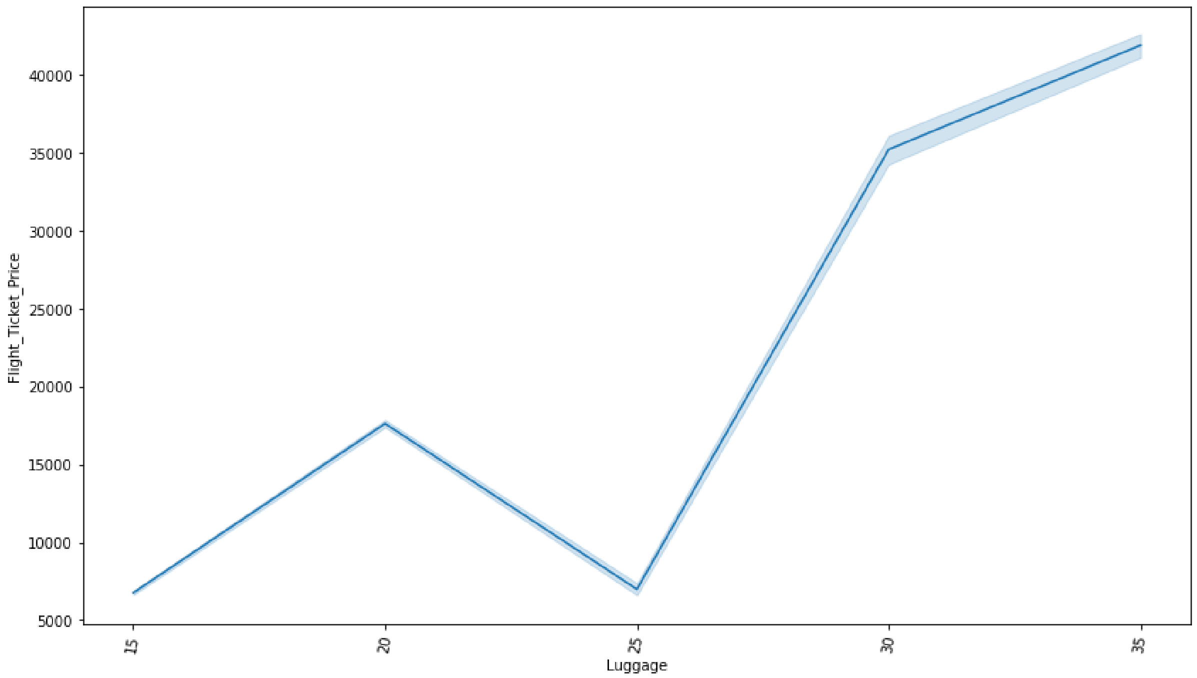


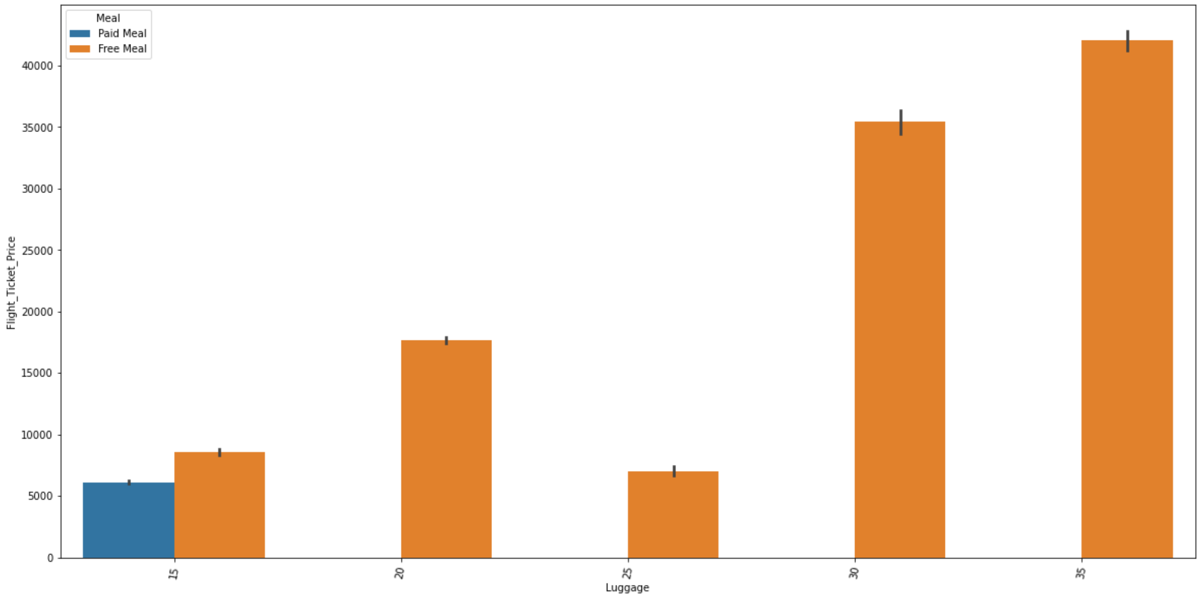


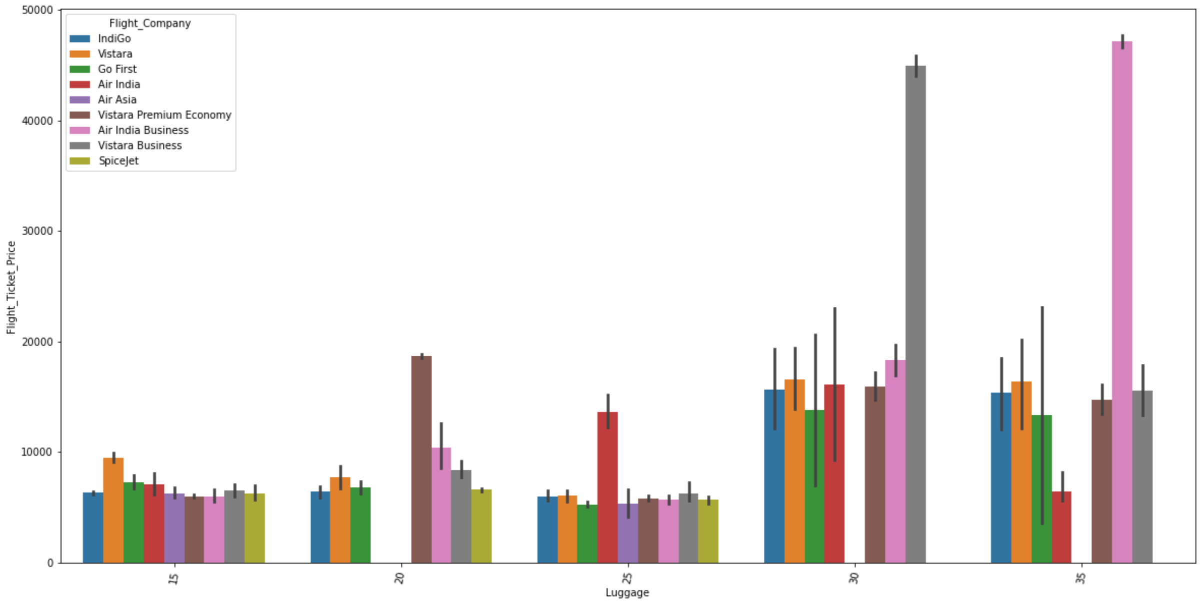


13.Luggage:

From the plots,15 and 25 kgs are cheap and once the load increases the ticket price also increases.

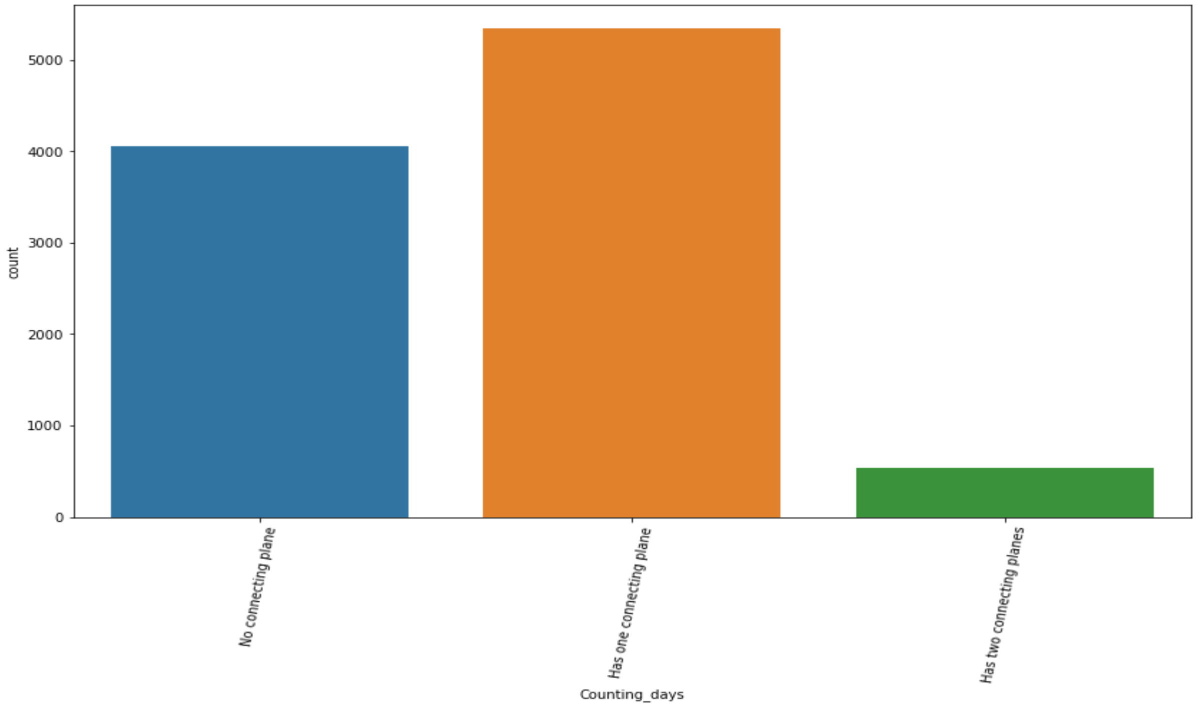

Flight_Ticket_Price
40000
35000
30000
25000
20000
15000
10000
5000
15
20
25
30
35
Luggage

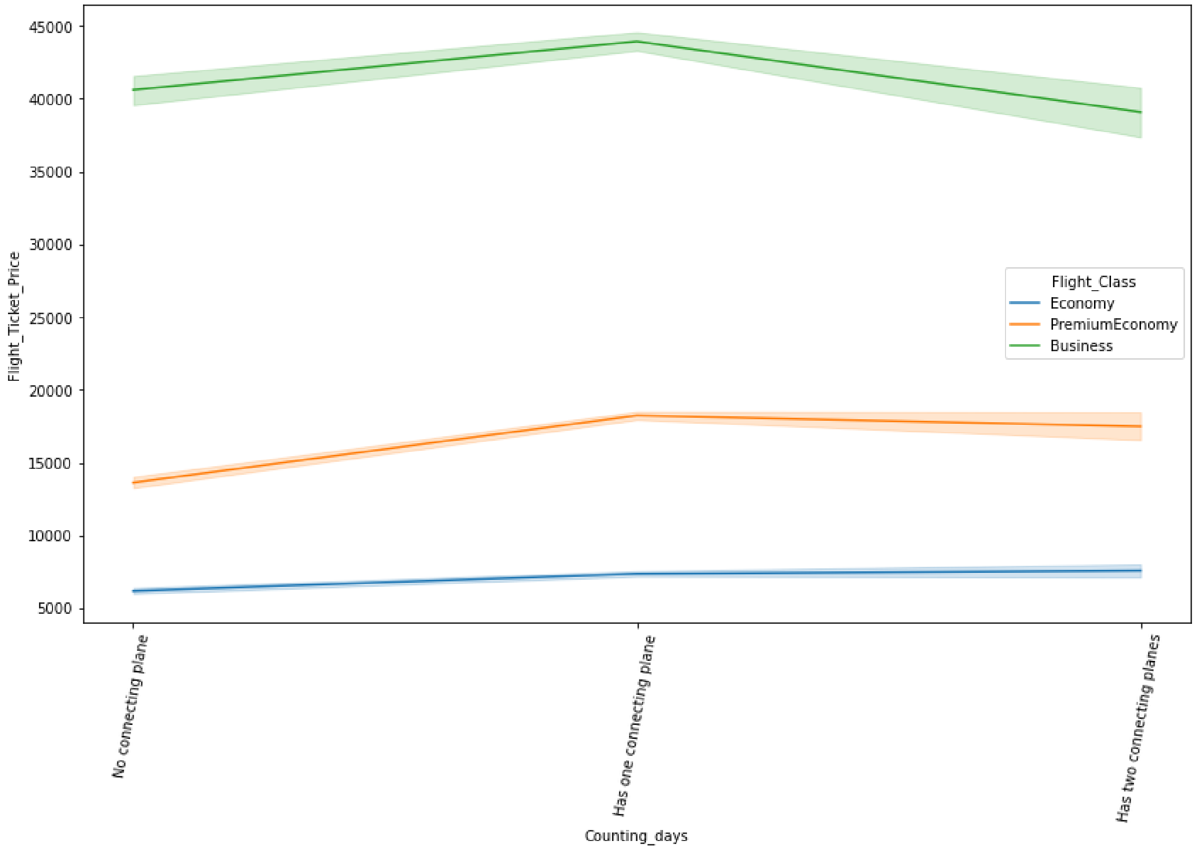

Meal
Paid Meal
Free Meal
Flight_Ticket_Price
40000
35000
30000
25000
20000
15000
10000
5000
0
15
20
25
30
35
Luggage



14.Connecting Plane:

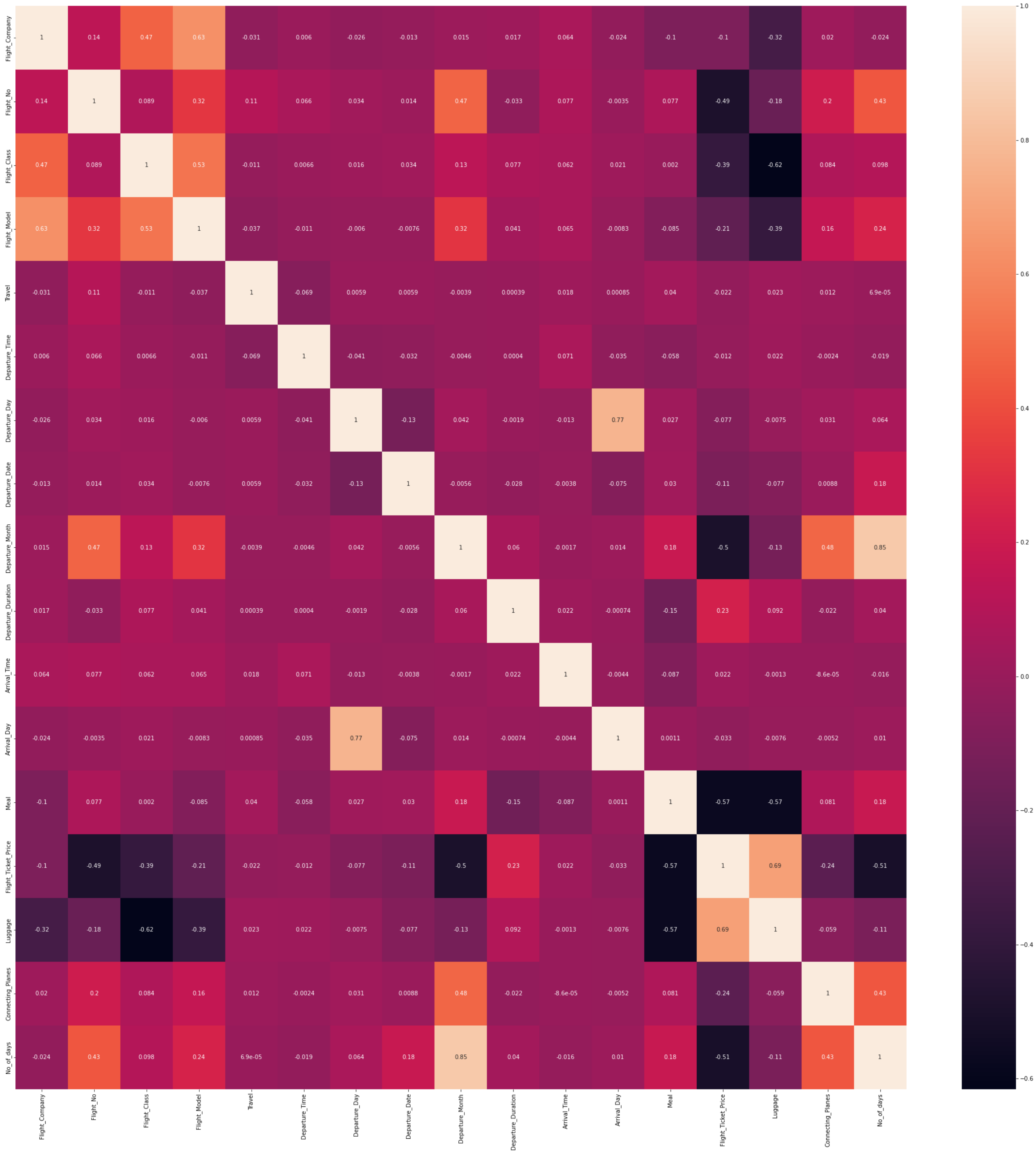
From the plots, we have more flight with a single connecting plane, it affects the duration also.





Once preprocessing is completed, the categorical data is converted to continuous data using label encoding.

Now the dataframe is checked for correlation and heatmap is shown below:



Now the dataset is checked for presence of outliers. Here we use two methods

1. IQR Method

From this method we have 1061 rows of outliers

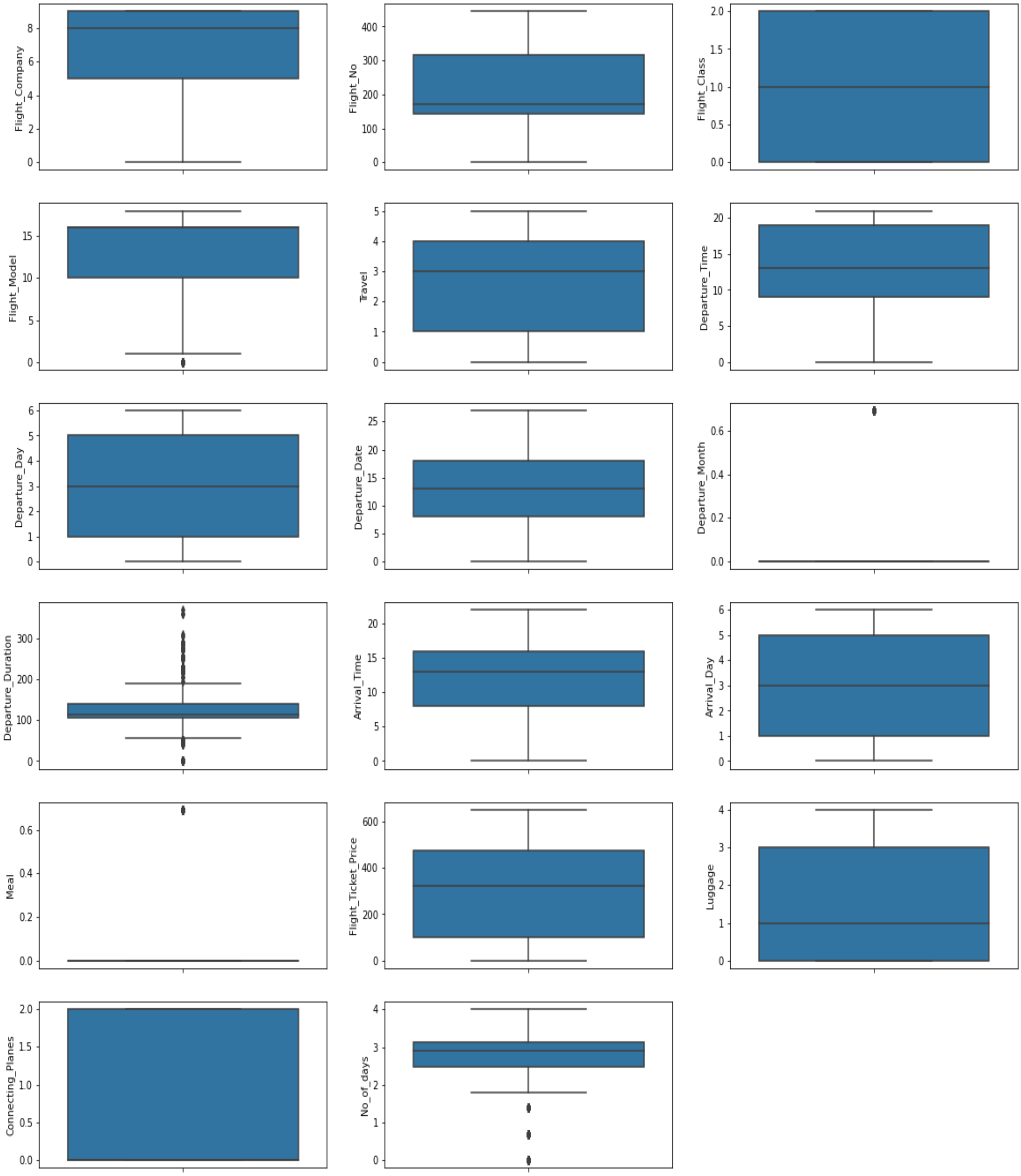
2. Z-Score Method

Using Z-Score method we detected 95 outliers

From the Z-Score method’s dataframe is used further.

The new dataframe has the shape of 9835 rows and 17 columns

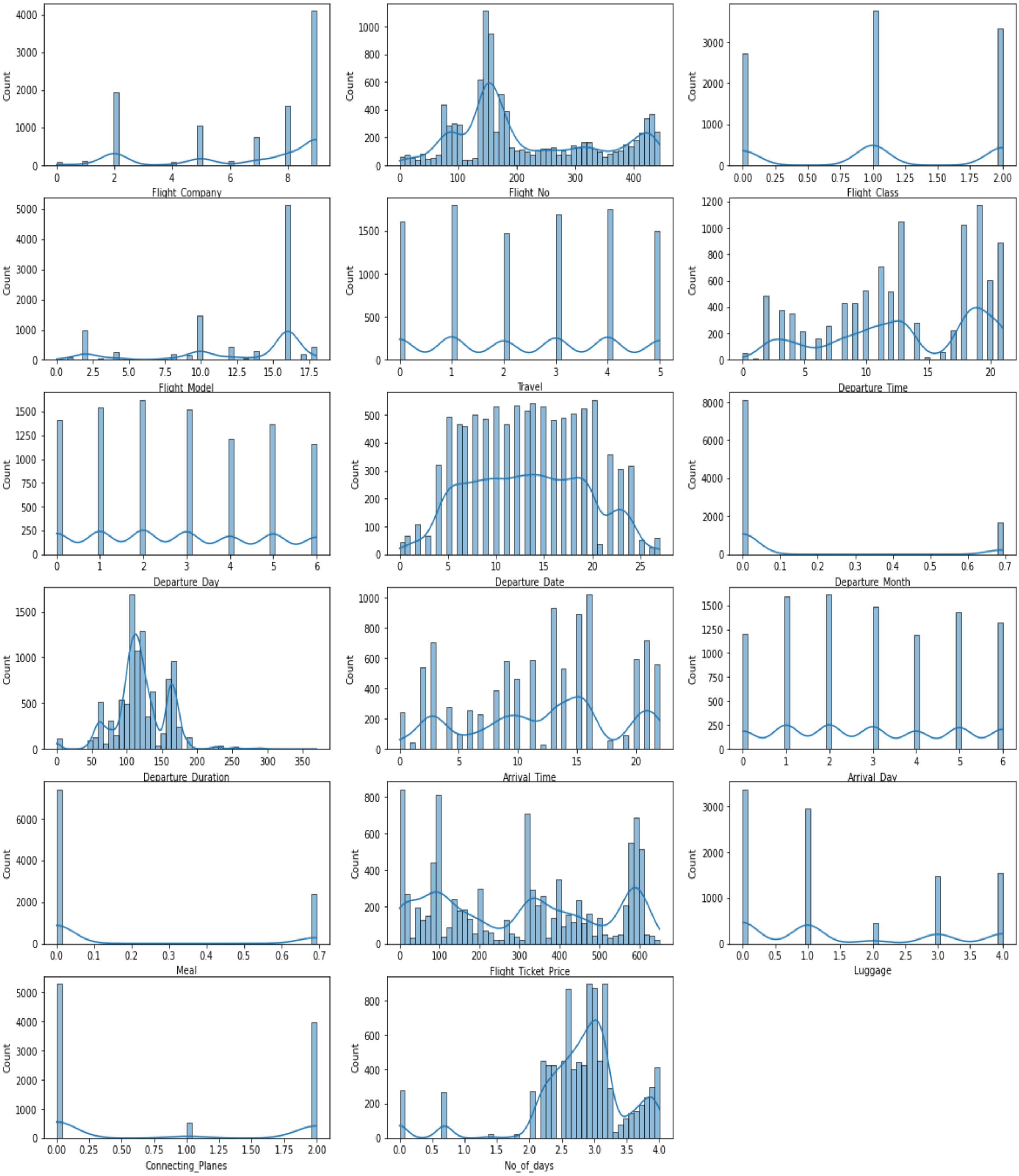
After removing outliers, we plotted the box-plot



Skewness:

Histogram:

A histogram is plot to check whether the features are normally distributed or not.



CHAPTER III

Model/s Development and Evaluation

3.1 Identification of possible problem-solving approaches (methods)

Basic Parameters:

1. Standardization:

Standardization is another scaling technique where the values are centered around the mean with a unit standard deviation. This means that the mean of the attribute becomes zero and the resultant distribution has a unit standard deviation.

$$X' = \frac{X - \mu}{\sigma}$$

2. Train Test data:

Train/Test is a method to measure the accuracy of your model. It is called Train/Test because we split the data set into two sets: a training set and a testing set. 80% for training, and 20% for testing. We train the model using the training set. We test the model using the testing set.

3. Linear regression

Linear Regression is an ML algorithm used for supervised learning. Linear regression performs the task to predict a dependent variable(target) based on the given independent variable(s). So, this regression technique finds out a linear relationship between a dependent variable and the other given independent variables. Hence, the name of this algorithm is Linear Regression.

5. Random Forest Regressor

Random Forests are an ensemble(combination) of decision trees. It is a Supervised Learning algorithm used for classification and regression. The input data is passed through multiple decision trees. It executes by constructing a different number of decision trees at training time and outputting the class that is the mode of the classes (for classification) or mean prediction (for regression) of the individual trees.

3.2 Testing of Identified Approaches (Algorithms):

Listing down all the algorithms used for the training and testing.

- Linear Regression
- Gradient Boosting Regressor
- AdaBoost Regressor
- Decision Tree Regressor
- KNeighbors Regressor
- Extra Trees Regressor
- Random Forest Regressor

3.3 Run and evaluate selected models:

From the above, the model is scaled using standard scaler, looped with the above methods and best model is obtained.

From this the best model is Extra Trees Regressor from the Random state 90. Now this model is hypertuned and best parameters is obtained

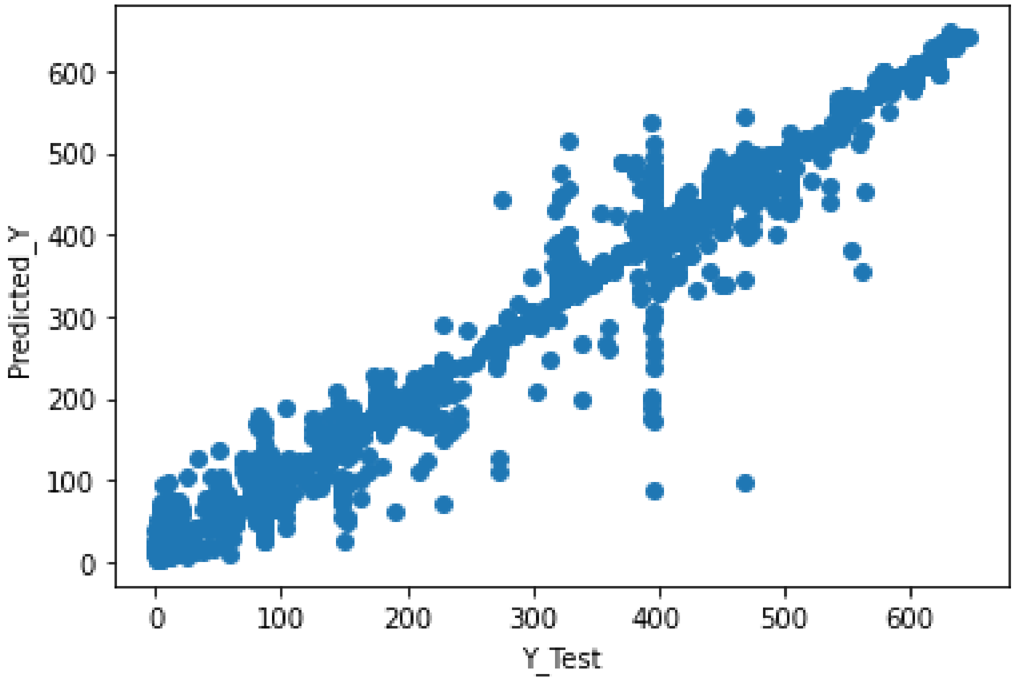
3.4 Key Metrics for success in solving problem under consideration:

Accuracy Parameter:

- R2 Score: 98.48414998011373
- Mean Absolute Error: 10.476769408502774
- Mean squared Error: 654.353253460798
- Root Mean Absolute Error: 3.236783806265530

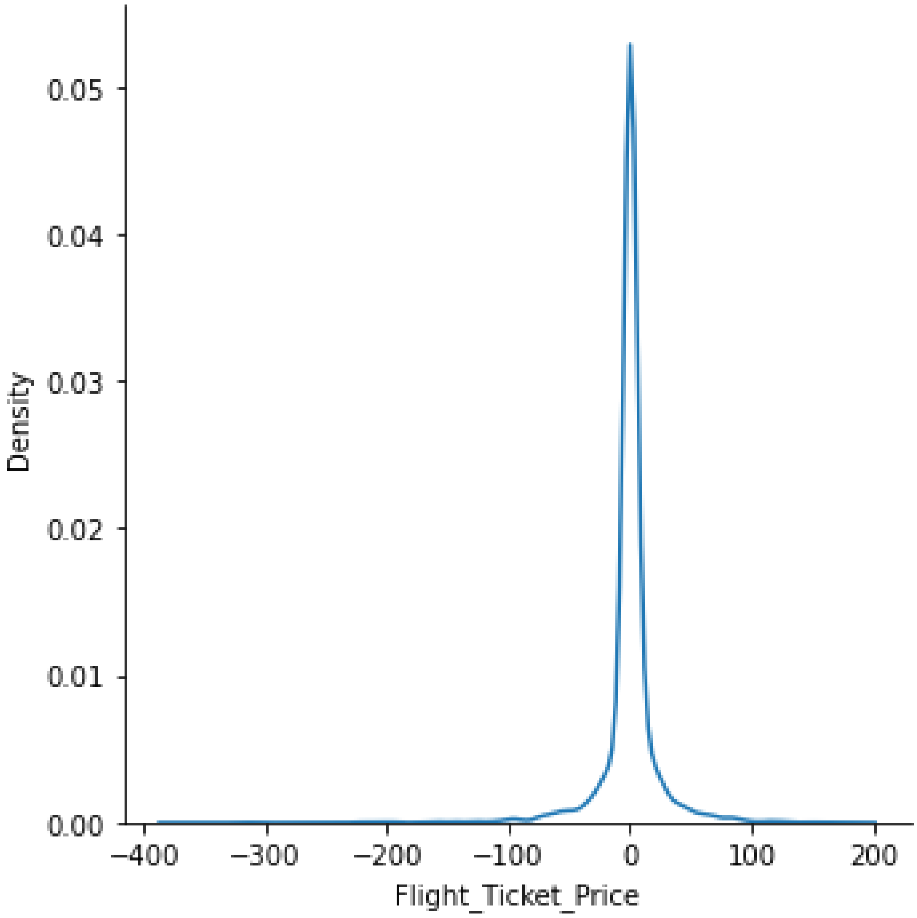
The best model is obtained by Hypertuning the existing models.

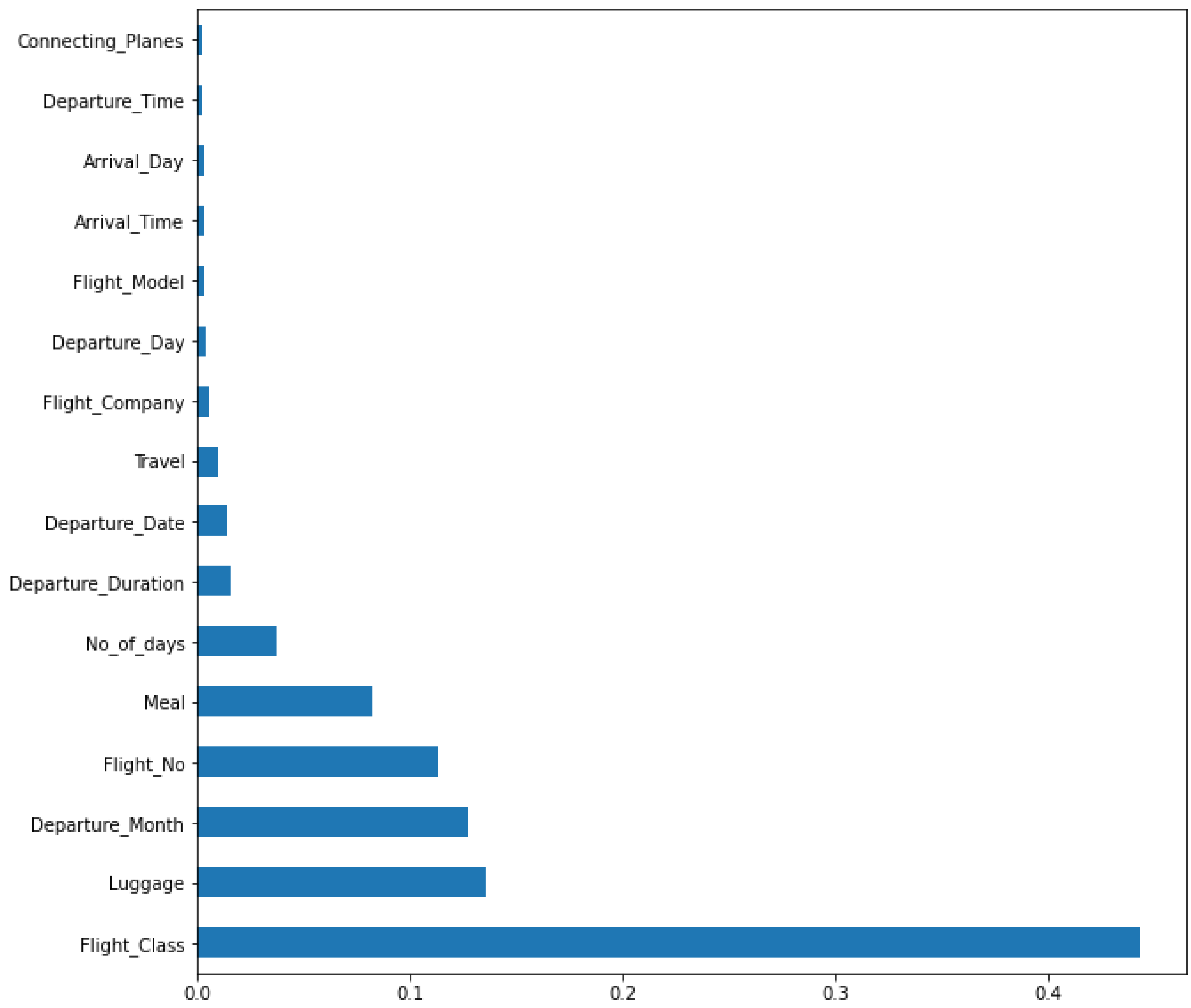
3.5 Visualizations:



We get almost a linear plot

Distribution of Predicted Values:





Regarding the feature importance's, registration year dominates the selling price more.

3.6 Interpretation of the Results

- From the results, we find that this problem can be solved by a regression method and selling price can be predicted.
- The skewness is not reduced as most of the data is categorical
- Flight class dominates the flight price more.

CHAPTER IV
CONCLUSION

4.1 Key Findings and Conclusions of the Study:

- This dataset has been taken from 2 websites of this, Yatra and MakeMyTrip constitutes the majority of data
- Since the target feature is continuous data, this problem can be solved by regression algorithms
- Extra Forest Regression gives an R2 score 0.98
- Flight class dominates the flight price more.

4.2 Learning Outcomes of the Study in respect of Data Science:

- It gives a deep learning of Selenium webscraping
- It emphasizes the importance of data cleaning
- Data uniformity and the importance of features required
- How data collection affects the results
- Role of data cleaning, feature selection, etc.

4.3 Limitations of this work and Scope for Future Work:

- Initially from the 2 websites many columns have been web scrapped, but due to the non-uniformity of data, and the features, some of the data have been removed
- Skewness have not been reduced

In feature, data will be gathered from more websites and more algorithms will be used along with different methods of scaling.

In future, more features will be added in predicting the flight ticket price.

*Note:

In order to find the pricing pattern, we need more days of data, so we took 2 months data.

There was a non-uniformity between the data in case of connecting plane and duration was mostly dependent on it so, in this project there will be outliers

Data was not be scrapped from official websites, During the scrapping process, two types of codes will be present because webscraping was done in two machines which has two different chrome driver versions

Future Work:

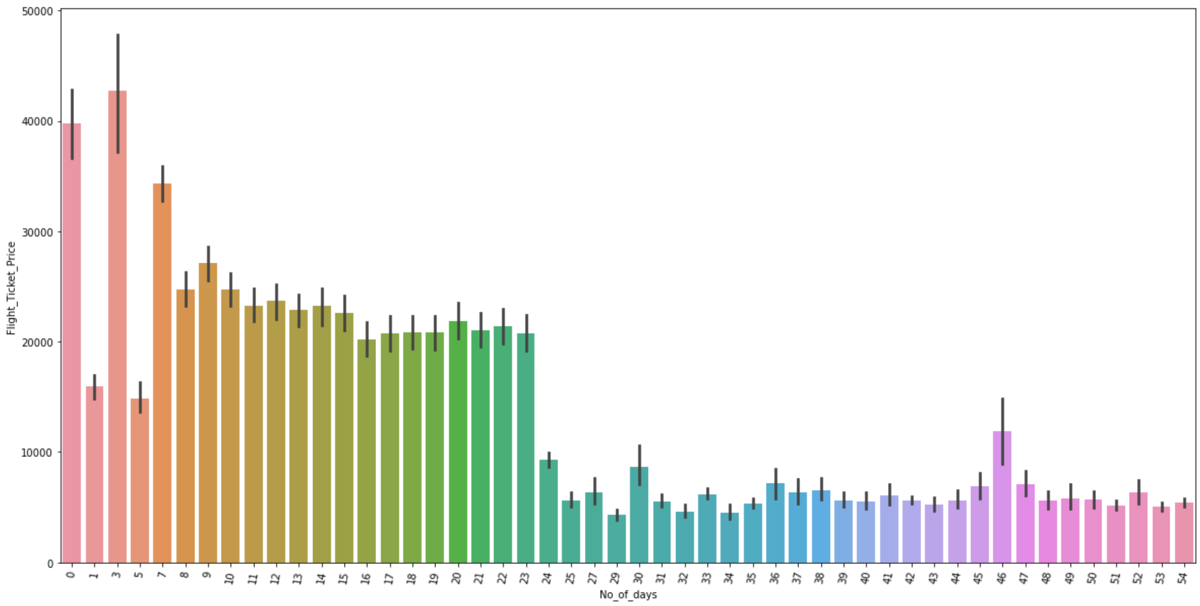
In future, data will be scrapped from more websites and more analysis will be done in the connecting plane area and the duration will be much more accurate

Data Analysis:

After cleaning the data, you have to do some analysis on the data.

1. Do airfares change frequently?

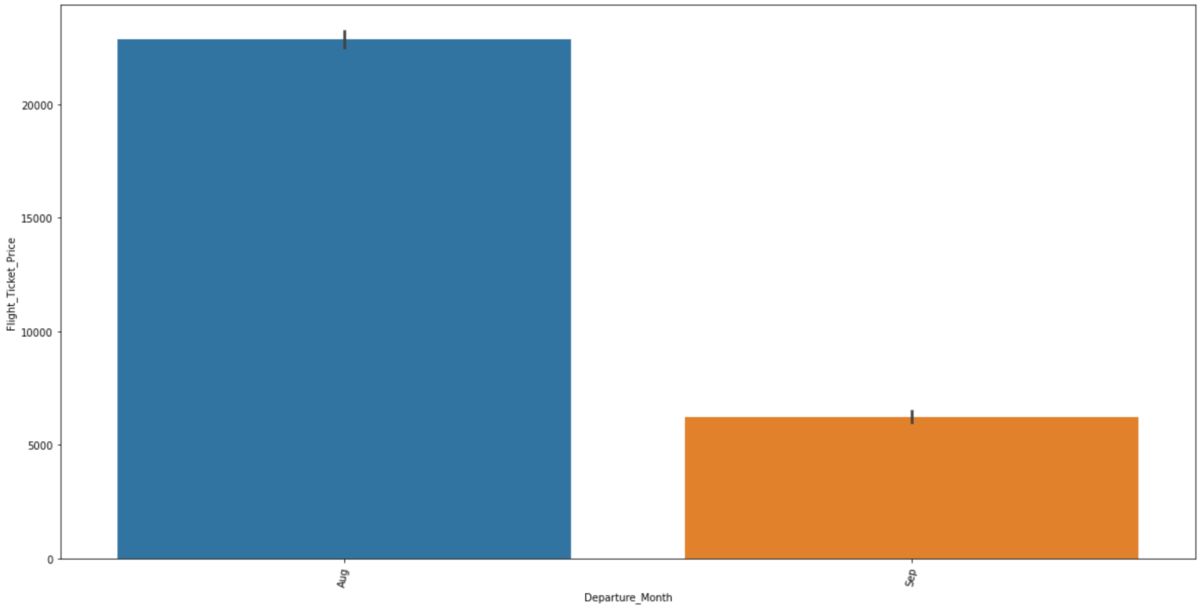
- ❖ No, it varies during weekends, eve of festivals
- ❖ Departure Flights: Friday, Saturday & Sunday
- ❖ Arrival Flights: Saturday, Sunday and Month
- ❖ Also based on flight class



2. Do they move in small increments or in large jumps?

- ❖ Generally, in small increments after 20 days it goes down and stays stable after that.

3. Do they tend to go up or down over time?



- As days increases the price decreases after around 24 days.

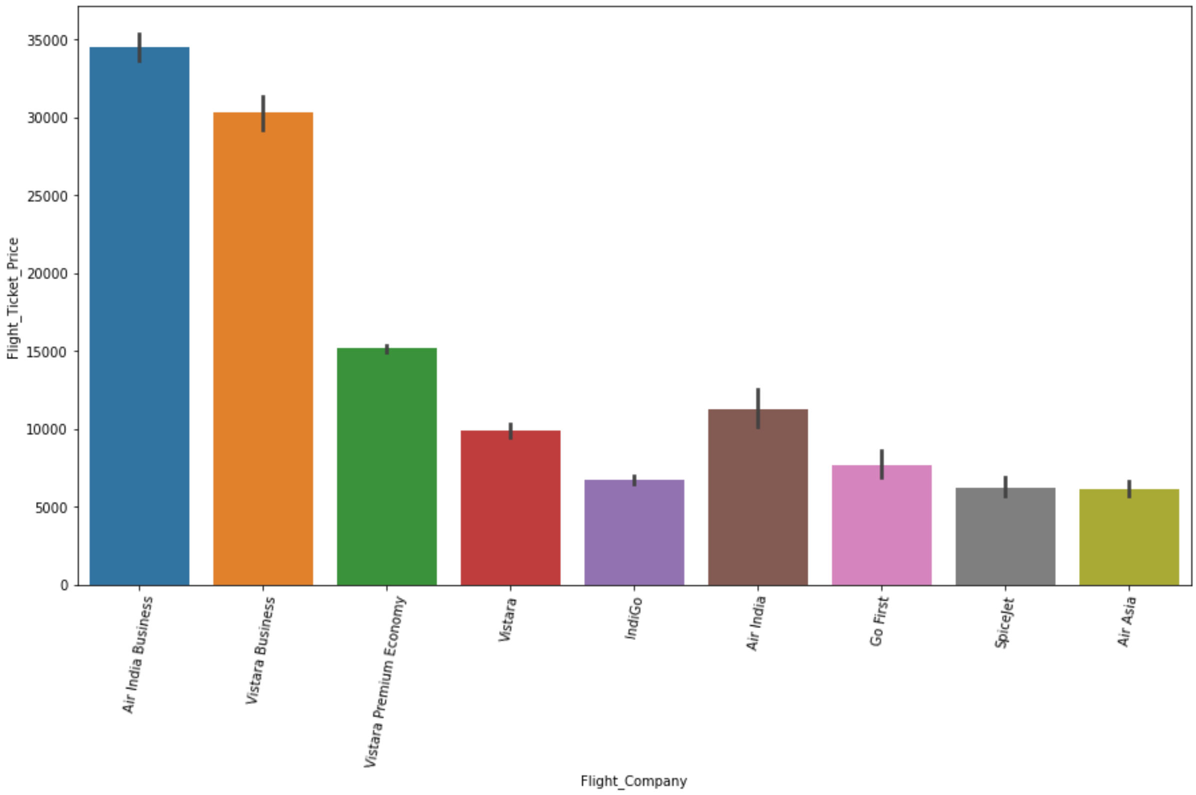
4. What is the best time to buy so that the consumer can save the most by taking the least risk?

- Need to plan before 25 days to get flight prices economical,
- Last minute flight bookings are costly
- Flight class is dependent on user’s convenience

5. Does price increase as we get near to departure date?

- Yes, last minute booking is costly

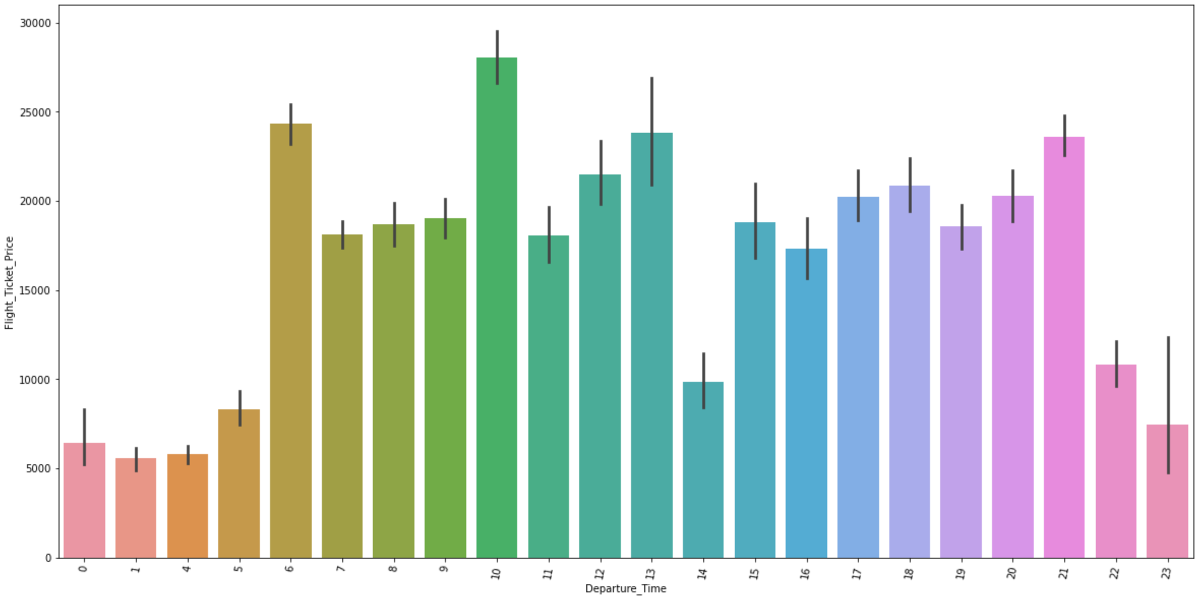
6. Is Indigo cheaper than Jet Airways?



- ❖ We could not find Jet airways but Indigo, SpiceJet, Air Asia is cheaper

7. Are morning flights expensive?

Flights after midnight to 6 AM is cheaper, after 6 the price increases.


Flight_Ticket_Price
Departure_Time