FLIGHT PRICE PREDICTION

Submitted by:

PRANJAL GANDOTRA

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

https://www.geeksforgeeks.org/

https://github.com/

https://www.mckinsey.com/

https://www.counterpointresearch.com/

INTRODUCTION

• Business Problem Framing

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, you 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.

• Conceptual Background of the Domain Problem

A good knowledge of how flights, airline operate will be very useful for the project.

• Review of Literature

A lot of research is available on flight prices. The following articles were useful in understanding the problem ebetter.

https://www.sciencedirect.com/science/article/pii/S131915781830884X

https://www.researchgate.net/publication/337821411_Predicting_Flight_Prices_in_India

• Motivation for the Problem Undertaken

Everybody wants to book cheaper flights; but to book cheaper flights one must understand how airlines evaluate the prices of the flights.

Analytical Problem Framing

- ## Mathematical/ Analytical Modeling of the Problem

Inbuilt function such as standardising and log will be used in tackling this problem.

R-square is a comparison of residual sum of squares *(SS_{res})* with total sum of squares*(SS_{tot})*. Total sum of squares is calculated by summation of squares of perpendicular distance between data points and the average line.

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Where SS_{res} is the residual sum of squares and SS_{tot} is the total sum of squares.

R-square is the main metric which I will use in this regression analysis.

Concordance index was also used. The concordance index or c-index is a metric to evaluate the predictions made by an algorithm. It is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs.

- ## Data Sources and their formats

  The data was scraped from makemytrip website; data was scraped for 72 different flight routes, between the 9 most popular domestic flight routes in India.

  This is the list of 9 most popular airpot codes in india. the data will be scraped for all these cities out going and incoming.

  The cities included are (in order):

- Delhi
- Bombay
- Bengaluru
- Chennai
- Kolkata
- Hydrebad
- Cochin

- Ahemdabad
- Pune

```
df.head() # First 5 elements
```

	Airline name	Departure city	Arrival city	Departure time	Arrival time	Total stops	Price
0	AirAsia	New Delhi	Mumbai	09:35	20:35	1 stop via Bengaluru	₹ 5,953
1	Go First	New Delhi	Mumbai	02:00	04:15	Non stop	₹ 5,954
2	Go First	New Delhi	Mumbai	07:00	09:10	Non stop	₹ 5,954
3	Go First	New Delhi	Mumbai	08:00	10:10	Non stop	₹ 5,954
4	Go First	New Delhi	Mumbai	10:30	12:50	Non stop	₹ 5,954

- # Data Preprocessing Done

The ‘Total stops’ column contained information about the number of stops and the route taken.

The information was extracted and put in different columns.

The price information was also truncated and converted into integer format.

Airlines with very few flights were simplified and flights contain two or more airlines were simplified to the first airline as well.

The departure and arrival time were divided into the four phases of a day.

- Early morning (00:00 – 06:00)
- Late morning (06:00 – 12:00)
- Early evening (12:00 – 18:00)
- Late evening (18:00 – 24:00)

```
df.head()
```

	Airline name	Departure city	Arrival city	Route dummy	Price	Total stops	duration	Departure	Arrival
0	AirAsia	New Delhi	Mumbai	Bengaluru	5953	1	660	Late Morning	Late Evening
1	Go First	New Delhi	Mumbai		5954	0	135	Early Morning	Early Morning
2	Go First	New Delhi	Mumbai		5954	0	130	Late Morning	Late Morning
3	Go First	New Delhi	Mumbai		5954	0	130	Late Morning	Late Morning
4	Go First	New Delhi	Mumbai		5954	0	140	Late Morning	Early Evening

- ## Hardware and Software Requirements and Tools Used

Pandas, Seaborn and sickit libraries were used throughout the project.

Model/s Development and Evaluation

- ## Identification of possible problem-solving approaches (methods)

  Regression and co relation.

In statistical modelling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables

- ## Testing of Identified Approaches (Algorithms)

- Decision tree regression
- Random forest regression
- Support vector regression
- Lasso regression
- Linear regression

Natural log, and min-max scaling

And finally hyper parameter tuning

- ## Run and Evaluate selected models

  All R-square values are from cross-validation of 5 samples

Algorithm (regression)	*R-square value*
Decision tree	*0.39556319719295796*
Random forest	*0.5725469207884493*
SVR	*0.36473493779213995*
Lasso	*0.09062796471381451*
Linear	*0.43660242440566654*

And as per this data random forest was chosen as the best model; further hyper parameter tuning was performed.

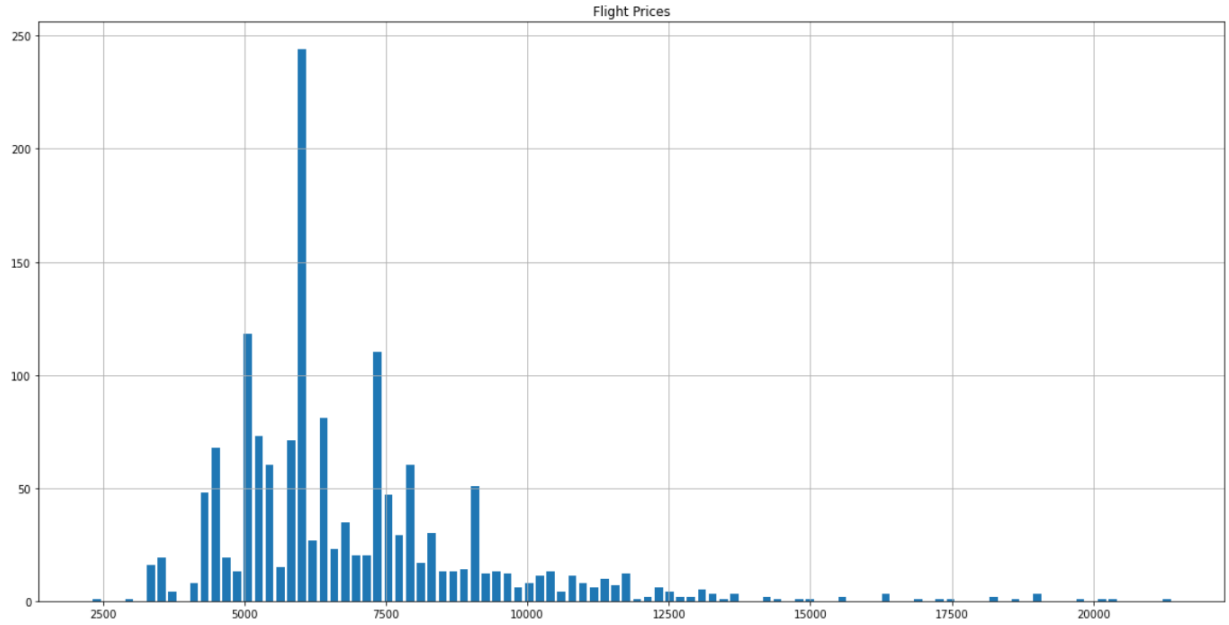
Final model: r-squared = 0.6177385219716642

Which is an improvement from all previous attempts.

- # Key Metrics for success in solving problem under consideration

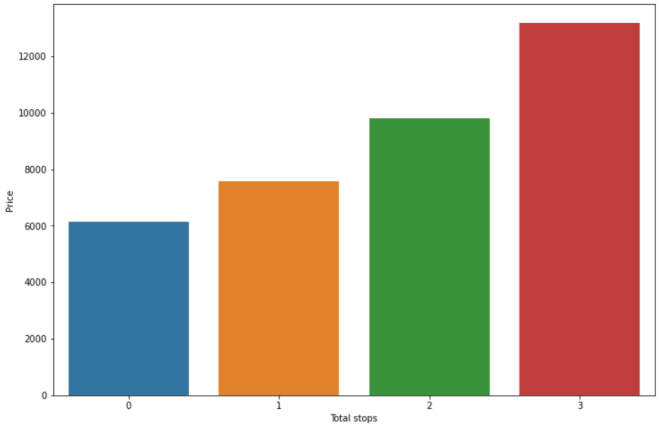
R-square was used to determine the success if an algorithm performed well or not.

- # Visualizations

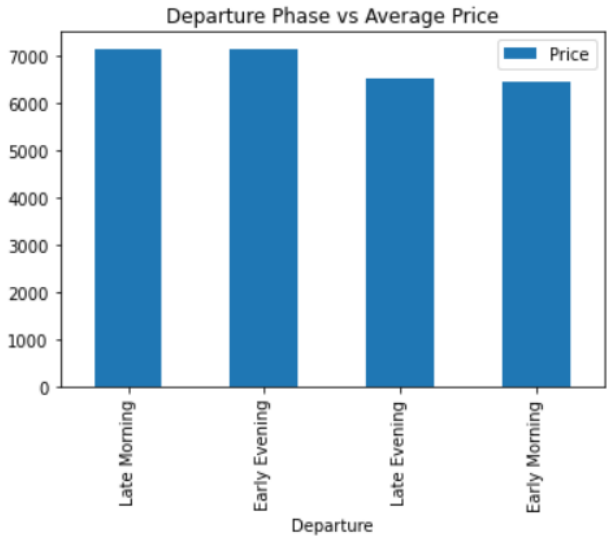


This is a histogram plot of price distribution of the flight prices.

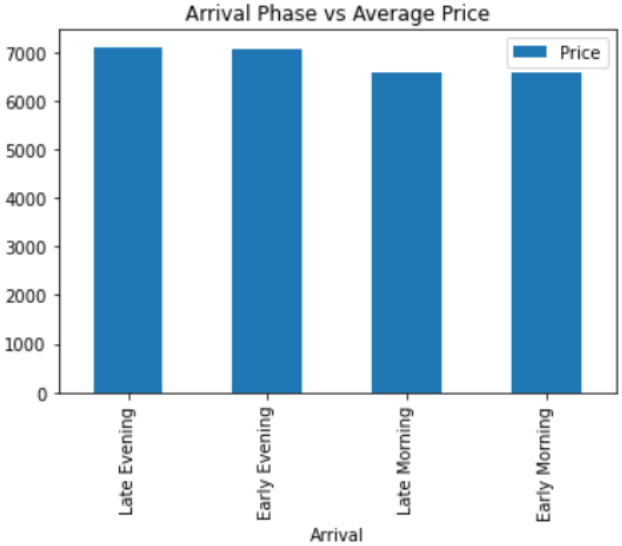
We can see that most of the flight prices range between Rs. 3000 and Rs. 10000. There is a right skew to the flight prices.



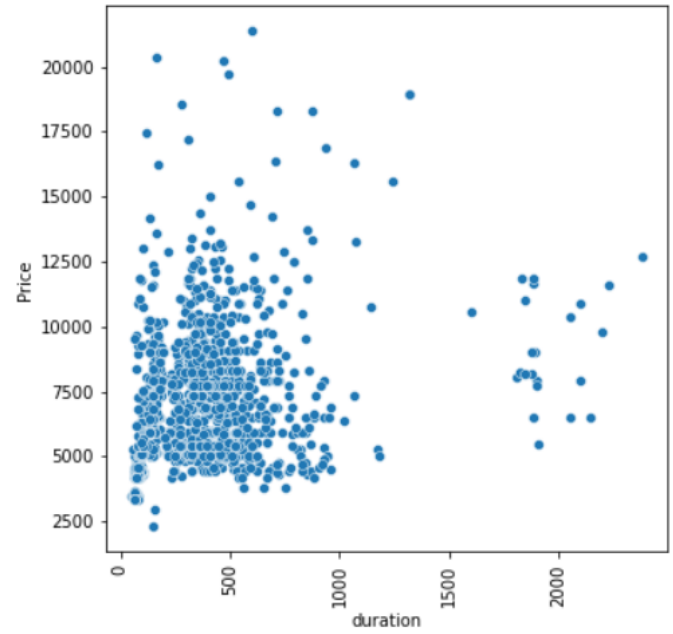
Plot of average price compared to how many stops the flight had.



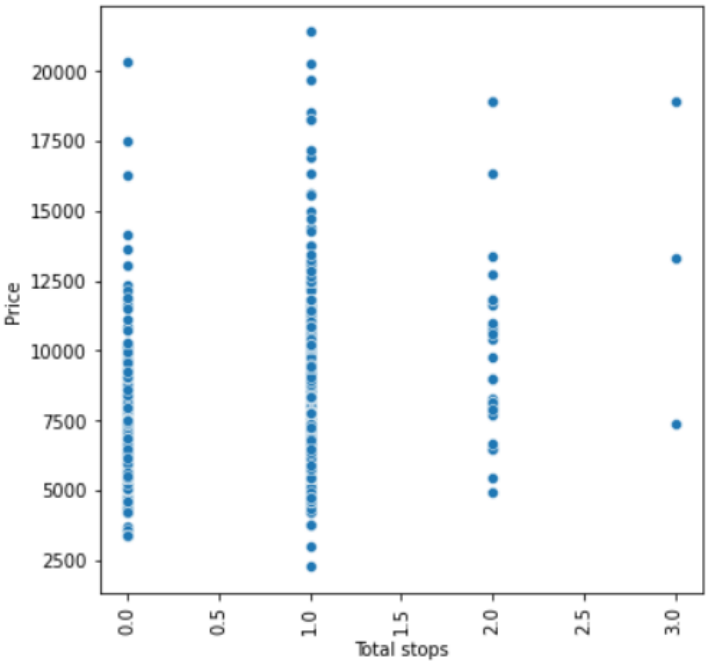
Departure phase vs average price.



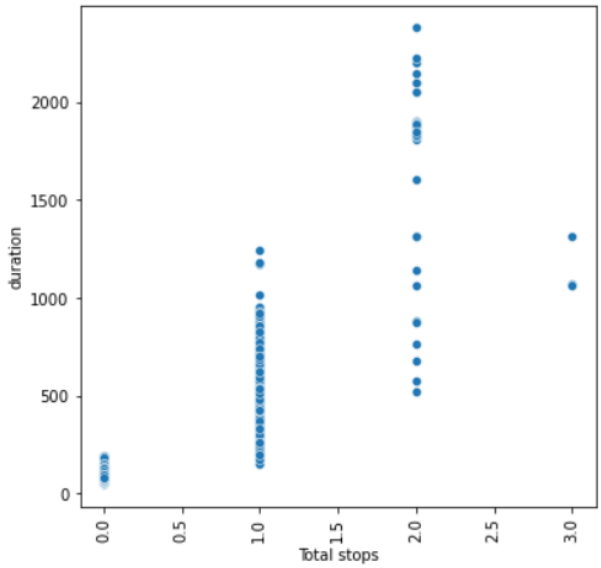
Arrival phase vs average price.



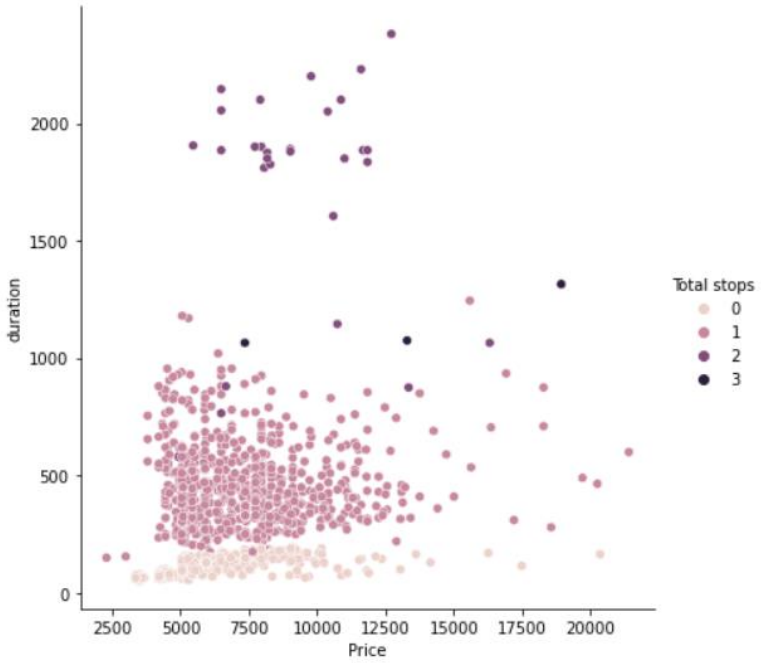
Scatter plot of flight price vs duration of flight (in minutes)



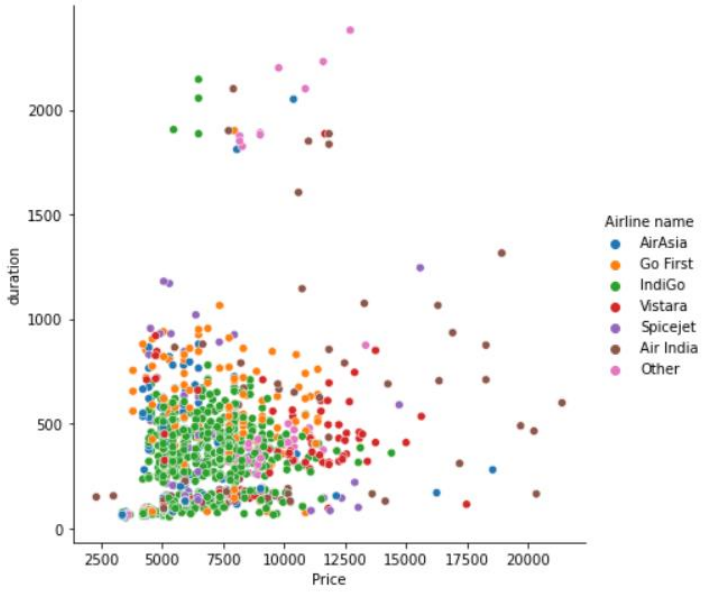
Scatter plot of flight price vs total stops of flight.



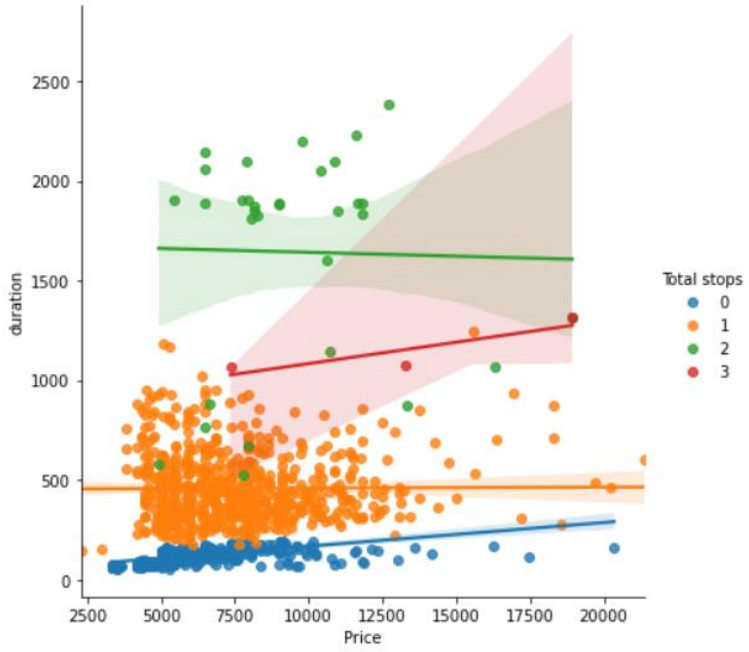
Scatter plot of duration of flight (in minutes) vs total stops of the flight.



Pair plot of duration vs price, with the hue of total stops in flight.

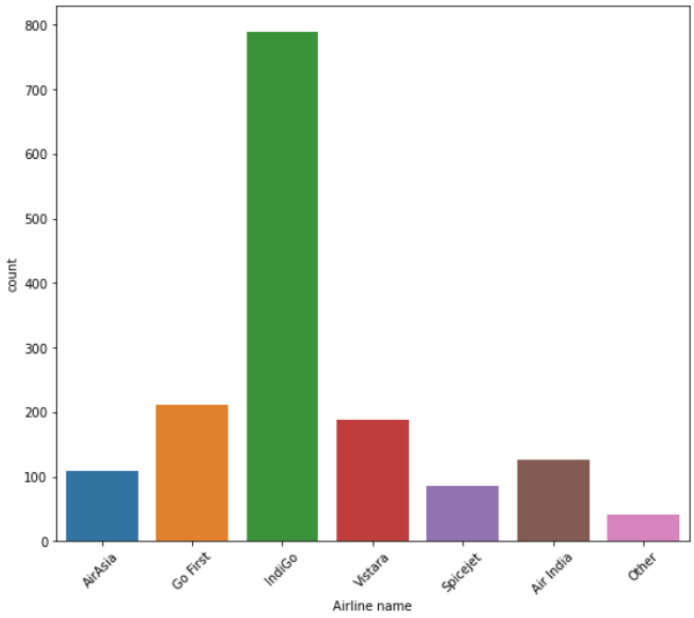


This graph shows the different pricing of airlines in a scatter plot.

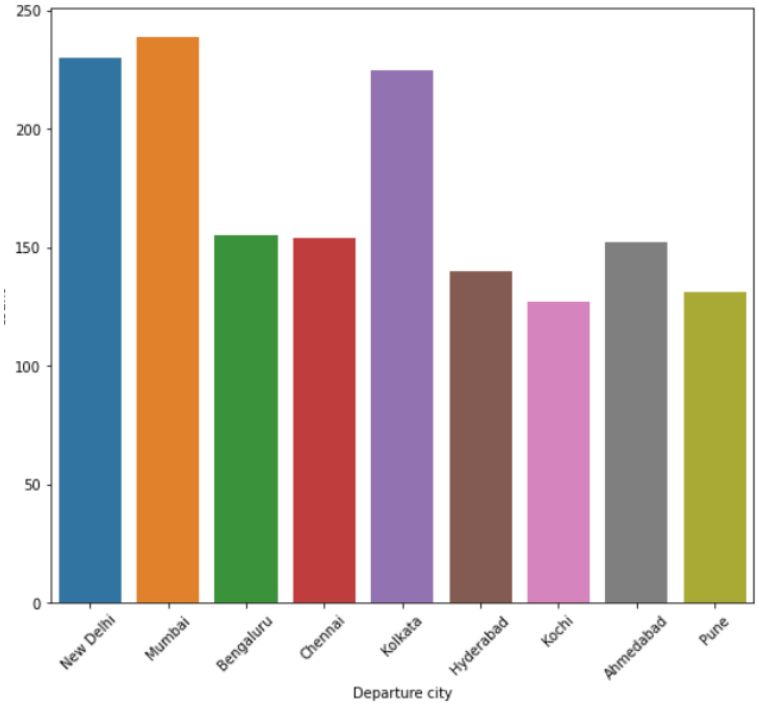


Price vs duration with a hue of total stops.

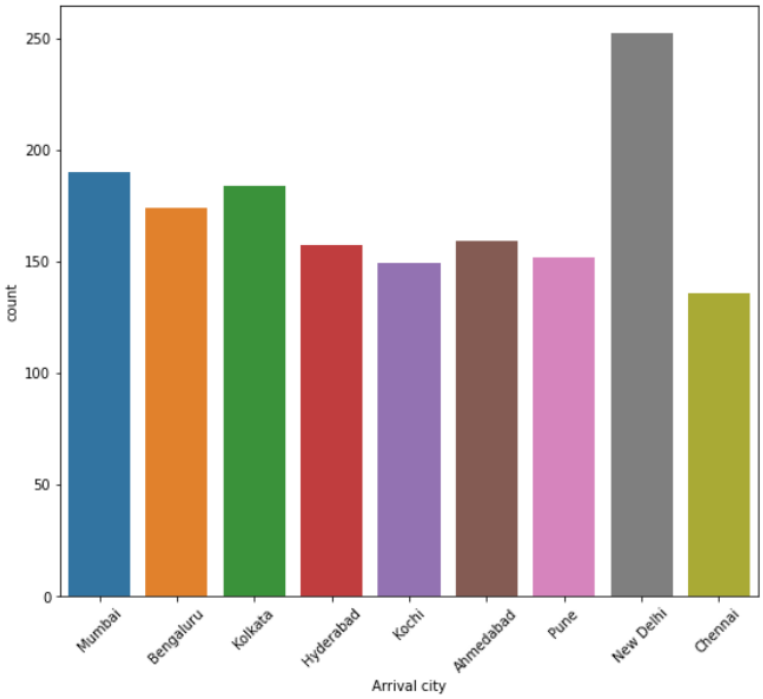
Frequency analysis of data:



Airline frequency on domestic path. We can observe that more than half of the total flights are operated by indigo airlines.

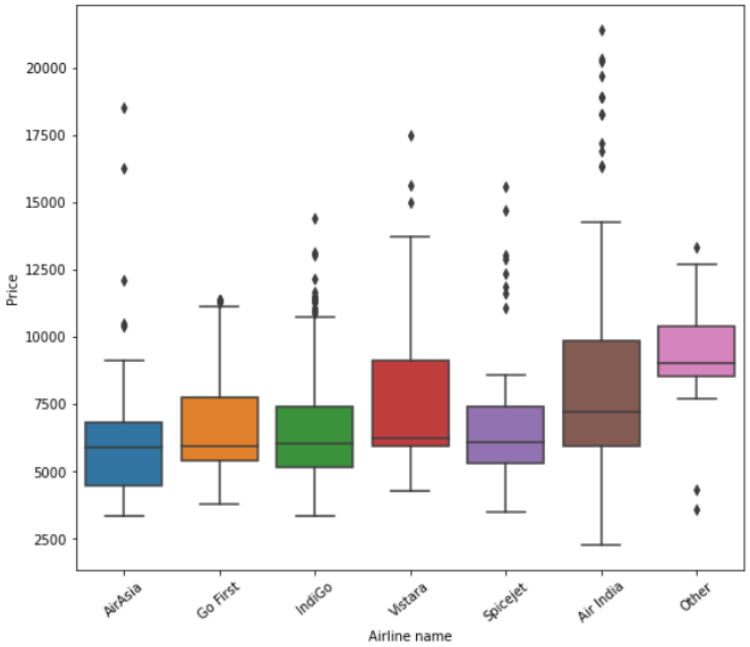


Most popular flight cities for departure in India. Mumbai is the most popular, closely followed by Delhi and Kolkata.

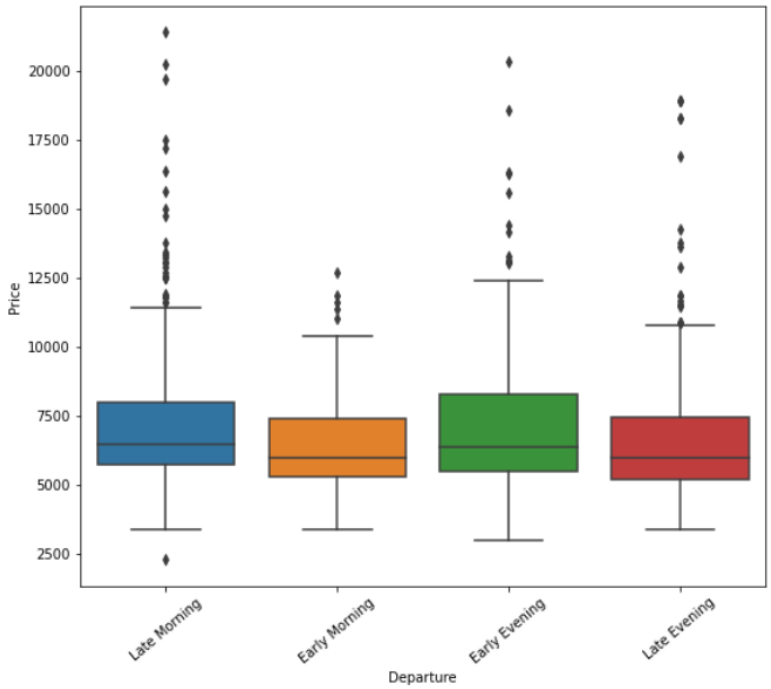


This is a graph of arrival cities in India domestic route. We can observe that Delhi is by far the most common arrival city in India, this is because Delhi is centralised and most people travel through it.

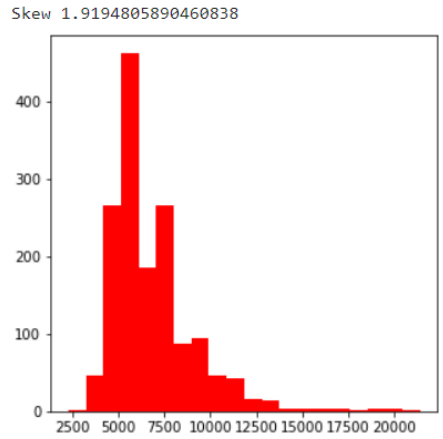
Price analysis of data:



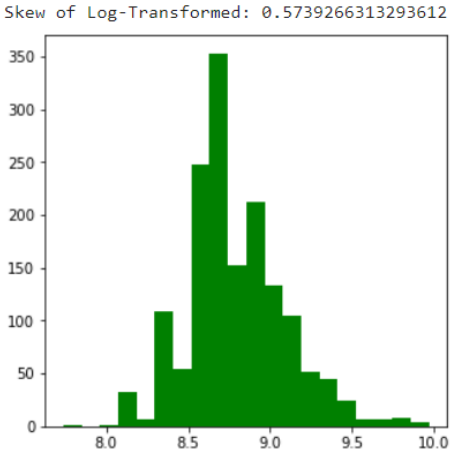
We can observe that IndiGo is not only the most frequent airline in domestic India but it is also one of the cheapest airlines out there. Most people are looking for cheaper flights and hence IndiGo is dominating the market.



Flights taking off in the early morning are the cheapest of all. Early evening slot is one of the most expensive flight take off times.



This is the skew of Price.

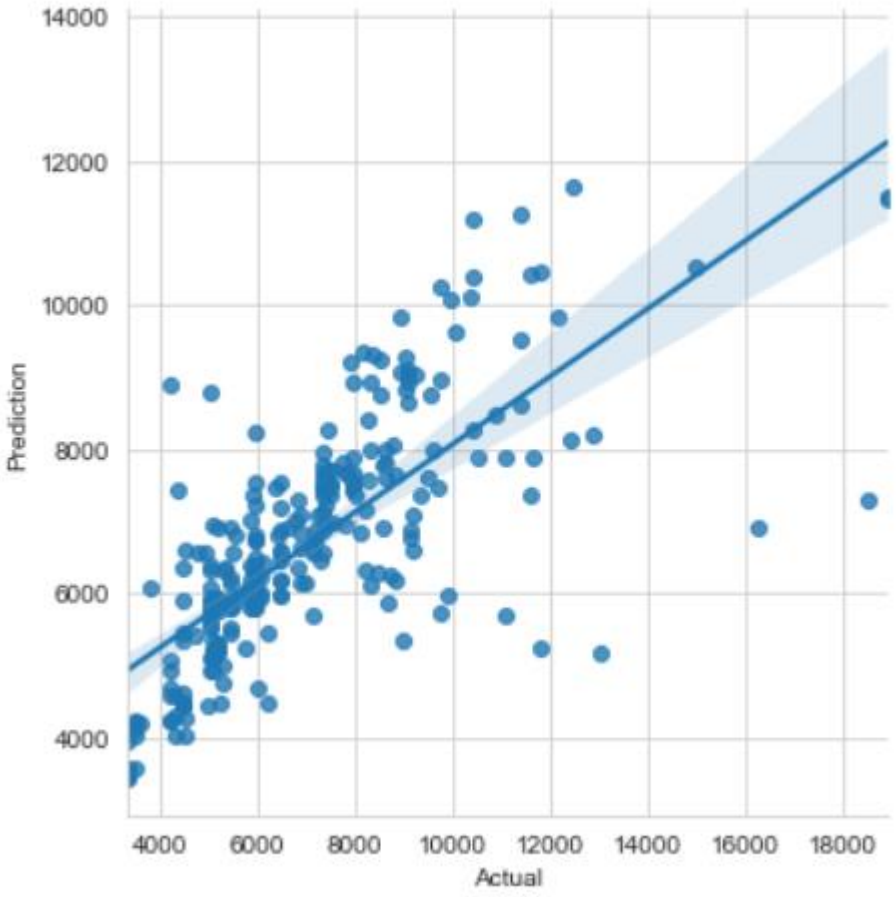


This is the skew of price after taking natural log.

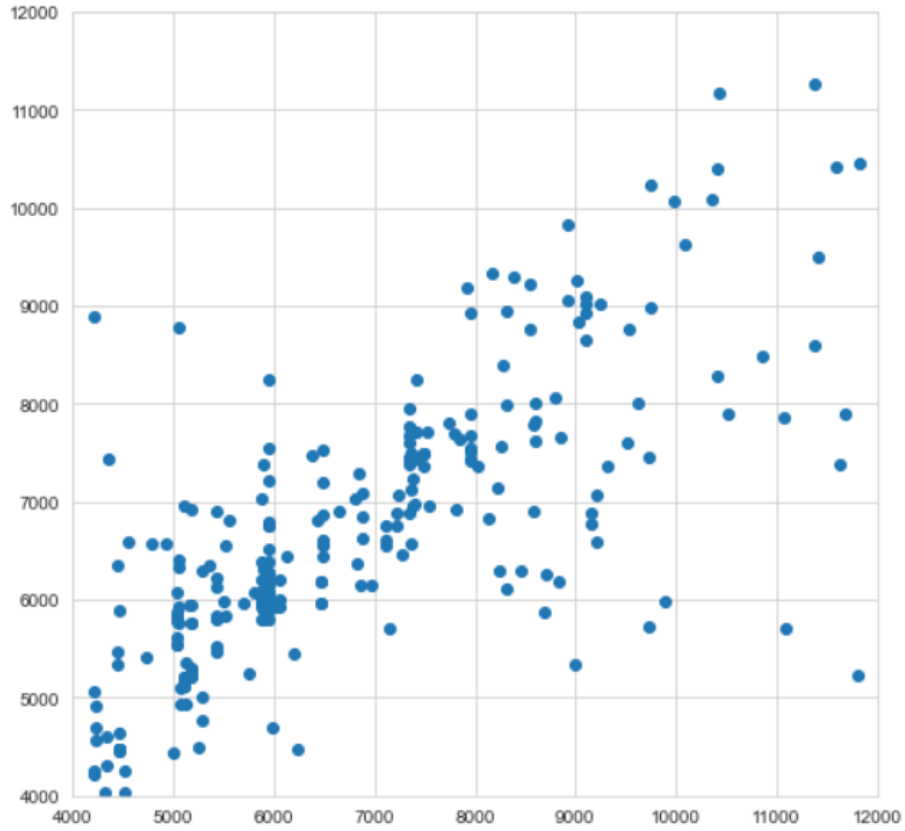
We can see that now we have a much better bell curve shape by taking log of the price variable.

- # Interpretation of the Results

Results was interpreted from the prediction vs actual price.

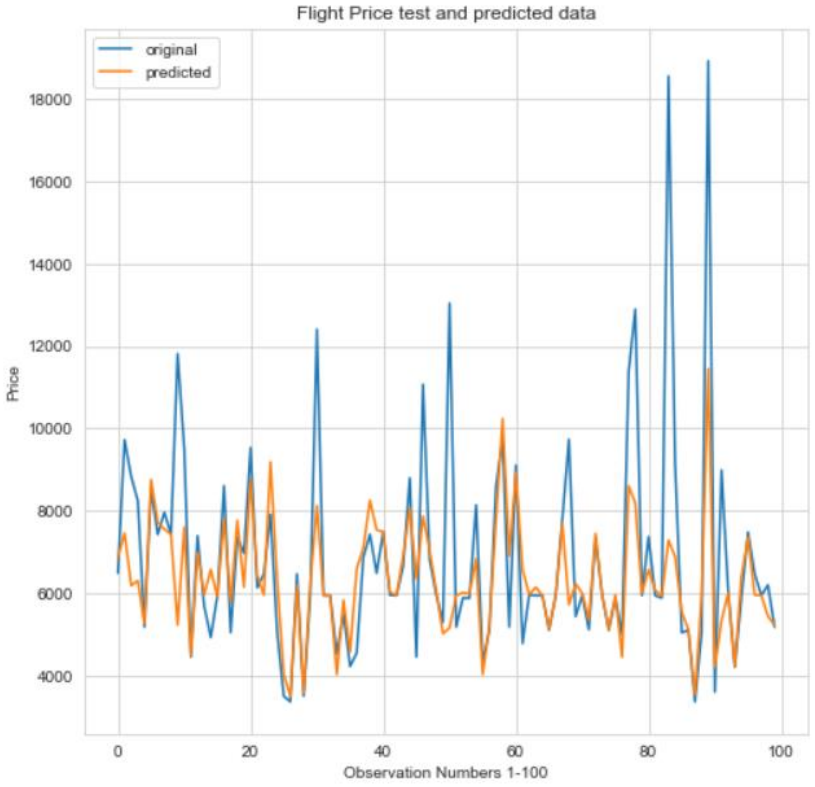


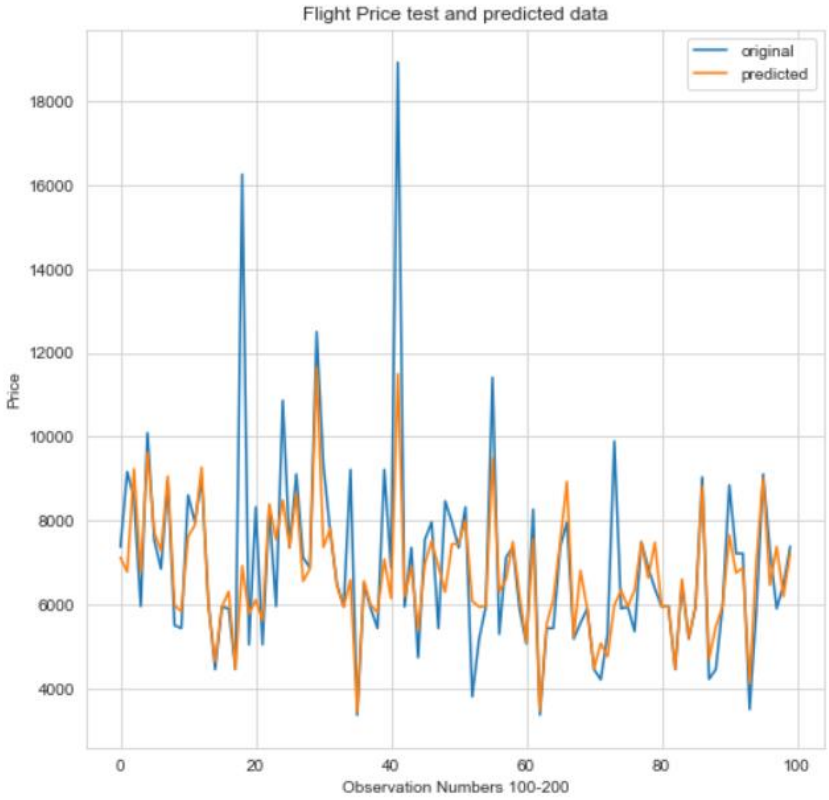
Final graph of the model predicting the values, we can see that the model is very accurate in determining the price.

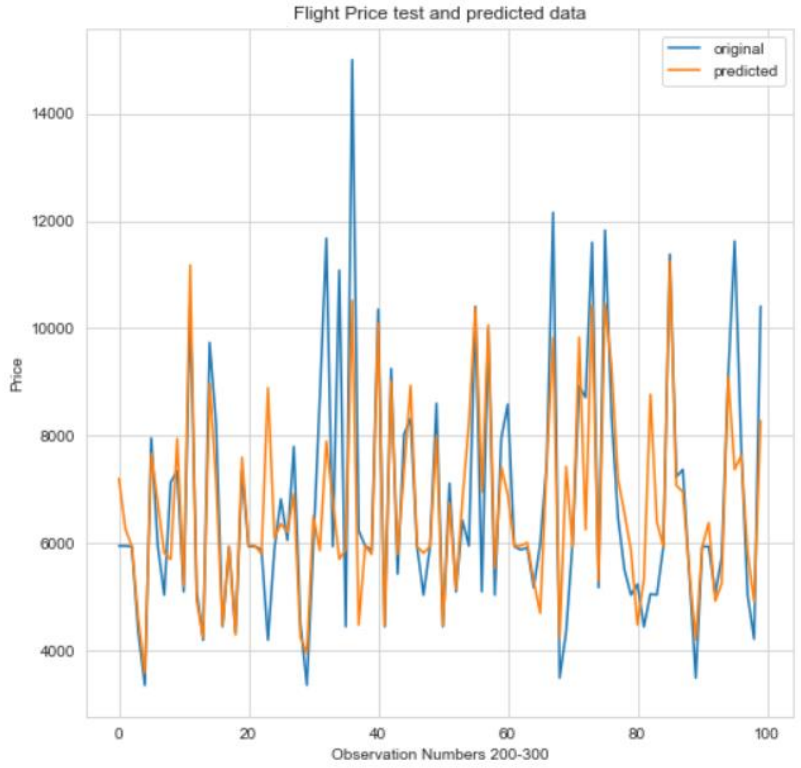


Scatter and regression plots between the predicted value and actual values.

Overlaying the predicted and actual value on a graph.







The same plot was produced for all 300 test values.

Residual plot was also created. We can see the output here.

We can observe that the model is not sure about the prices for flight which have extreme prices. This is to be expected as not many flight have extreme prices.

CONCLUSION

- ## Key Findings and Conclusions of the Study

Flight prices depend on few key factors, morning flights are cheaper than early evening flights.
The more the stops the more is the average price. Few airlines dominate the majority of domestic flight paths like IndiGo which offer the maximum number of flights at cheaper than competitors prices.

- ## Learning Outcomes of the Study in respect of Data Science

Random forest regression works best for this particular data set, hyper parameter tuning was performed and optimal parameters were found.

EDA is very powerful in understanding the data and pre-processing it before feeding it to the algorithm. Statistical methods work the best.

- # Limitations of this work and Scope for Future Work

Considering that this mode contains 72 different flight paths over 1500 data points, with the help of these graphs we can see that the model is performing well.

On average this data has 1502/72 = 21 data points for each flight path.

While the reality is that on average most of the flights are dominated by popular routes like Delhi to Bombay, Bombay to Delhi etc. among the most popular cities.

And the paths with less frequent flights like Ahmedabad to Pune has far less than 5 flight on that path.

To get a more accurate mode, much more data is required.