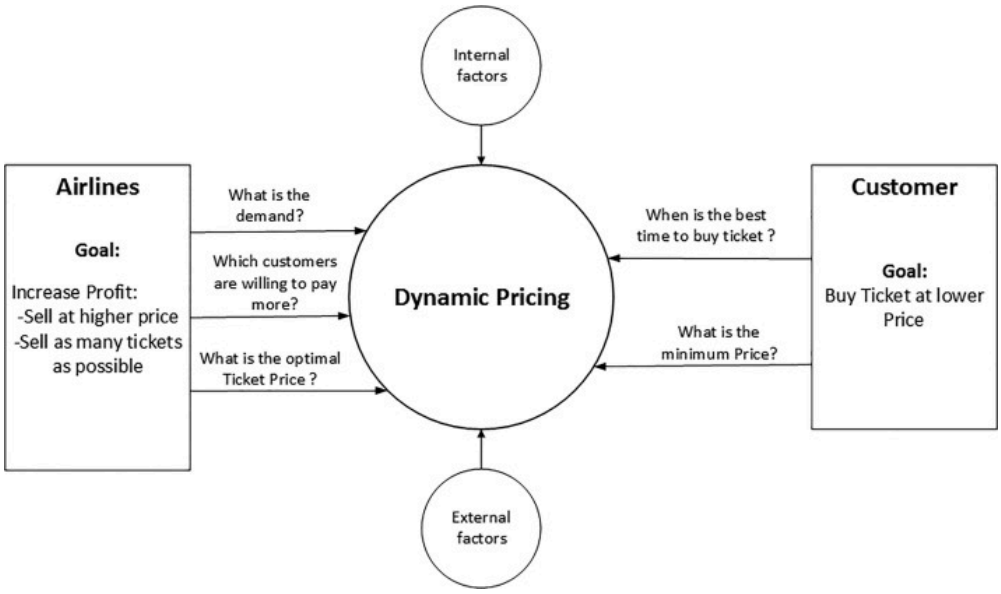
AIRLINE PRICE PREDICTION

ABSTRACT

Nowadays, the airline industry is using some of the most sophisticated strategies to assign the prices dynamically in order to maximize revenue and increase their profits. They use various kinds of algorithms and computation techniques that are largely based on the demand prediction, the commercial and marketing factors. It has become a challenging task for passengers to travel economically due to the dynamic change in prices. In this project, Machine Learning algorithms like Linear Regression, Support Vector Machine and several others are used to make a prediction model.



1.DATASET AND FEATURES

The dataset used in the project is provided by the MachineHack Hackathon. The training set has 10600 records and the test set has 2671 records. The data ranges from airline prices of Indian airline companies spanning over four months from March 2019 to June 2019. The data source contains information of around 127 different routes operated by 11 different airlines. The features selected to

use in our model include Date of journey, Route, Departure time, Arrival time, Duration, Total Stops.

According to the dataset, the most common airlines are Jet Airways, Indigo and Air India, which are also moderately priced all throughout the four months. As shown in Fig(2), the mean price of the airlines mostly ranges from 3000 to less than 2000.

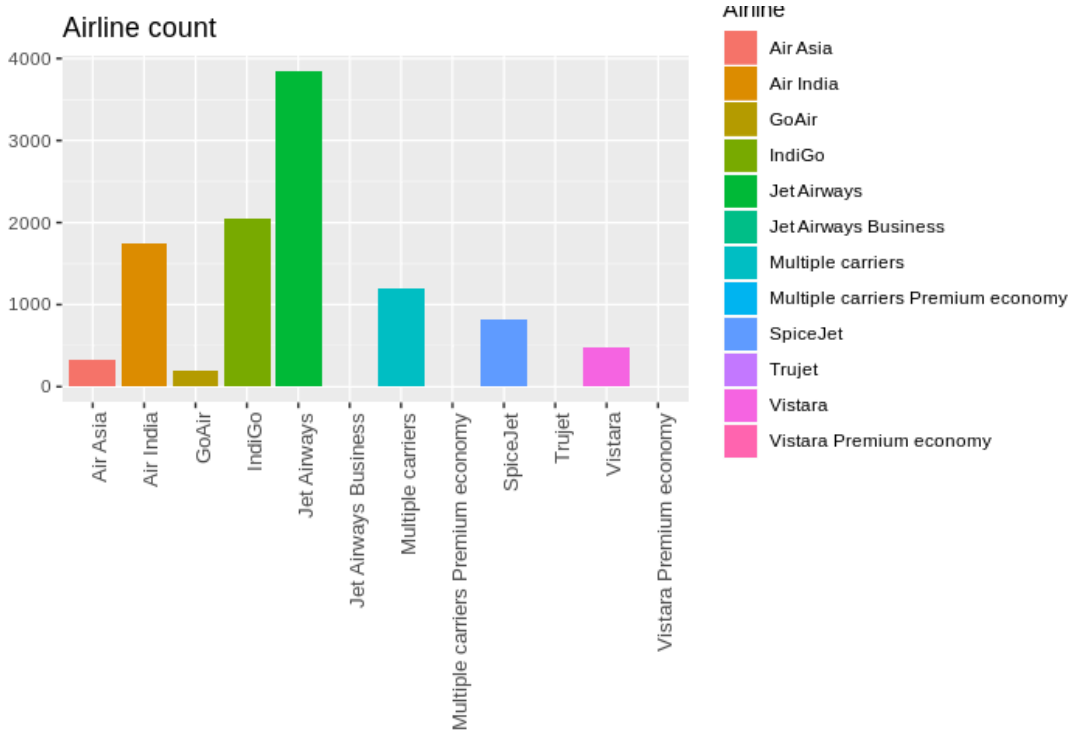


Figure1: overview of the data set

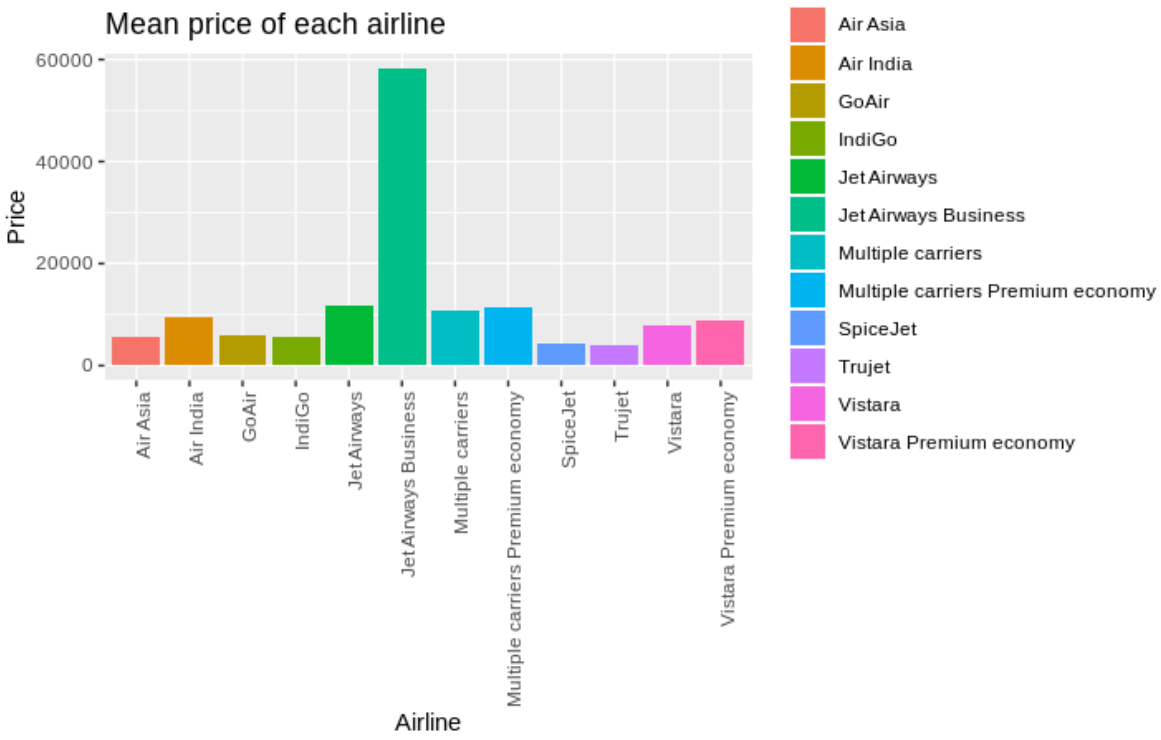


Figure 2: Mean Price of each Airline

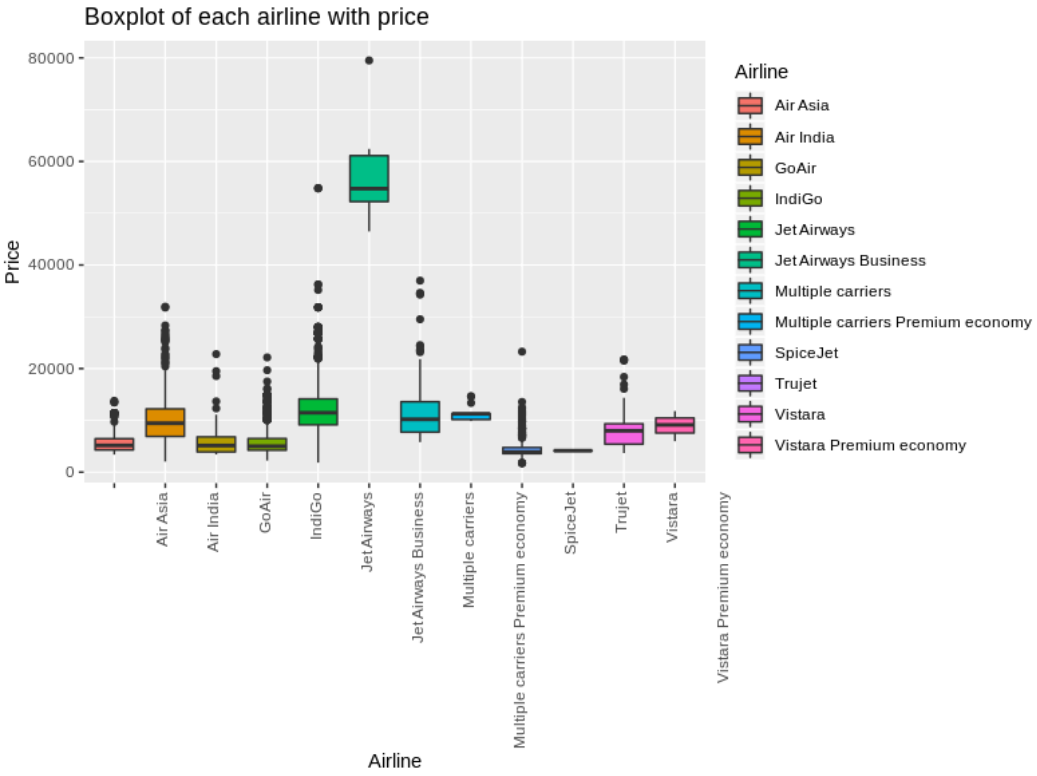


Figure 3: BoxPlot of each Airline vs price

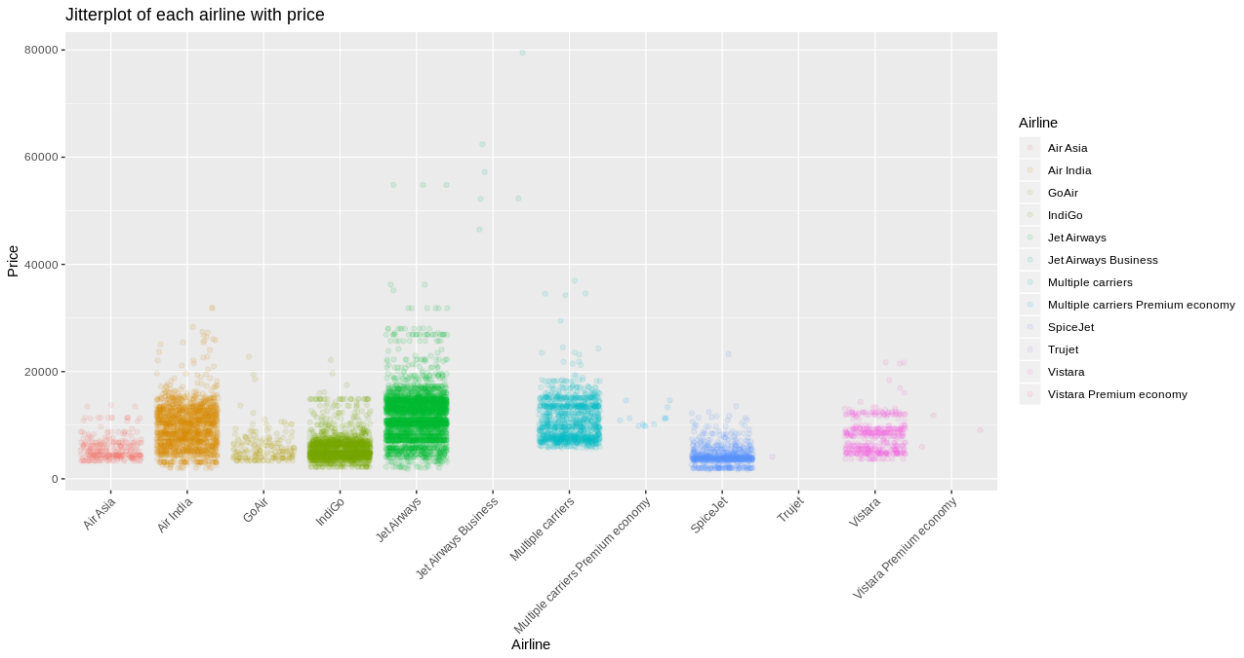


Figure 4: JitterPlot of each Airline vs price

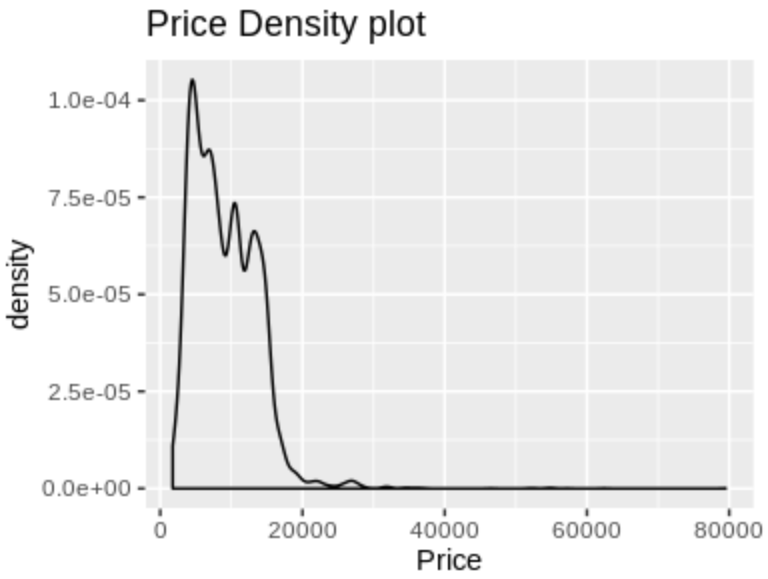


Figure 4: Price Density Plot

From the above graphs, it is understood that the maximum number of flights is in Jet Airways and the pricing is moderate. While on the other hand, Jet Airways Business price goes up to 80000Rs. It is seen from the Boxplot graph that the average price of JetAirways business is around 55000Rs. As seen in Fig 1, the count of Jet Airways Business and Vistara Premium Economy and TruJet is very negligible, it was dropped. Indigo, though the majority of the flights are moderately priced throughout the four months, touches the second-highest price of around 56000Rs.

2. METHODOLOGY

2.1 Data Cleaning

Removing the Na values and duplicates

Since the dataset has a large number of records, it is essential to find out the missing values so as to prevent any errors during analysis.

To check the Na values is.na() function is used. Since the number of Na values is negligible in this data set, we chose to omit it. It can be done using the function na.omit()

Removing the less significant features

From the above analysis of the dataset, it is evident that airlines such as Jet Airways Business Multiple Carriers Premium economy, Trujet, Vistara Premium Economy have a very negligible number of records and hence we choose to drop it.

2.2 Importing Packages

R packages are a collection of R functions, compiled code and sample data. They are stored under a directory called "library" in the R environment.

Packages used are :

Tidyr - Tools to help create tidy data, where each column is a variable, each row is an observation, and each cell contains a single value.

Caret - The caret package (short for Classification And REgression Training) is a set of functions that attempt to streamline the process for creating predictive models.

Lubridate -It is used to work with date-time objects and also contains mathematical operations that can be performed with date-time objects. It introduces three new time span classes,durations, periods and intervals.

Rpart - The rpart code builds classification or regression models of a very general structure using a two-stage procedure

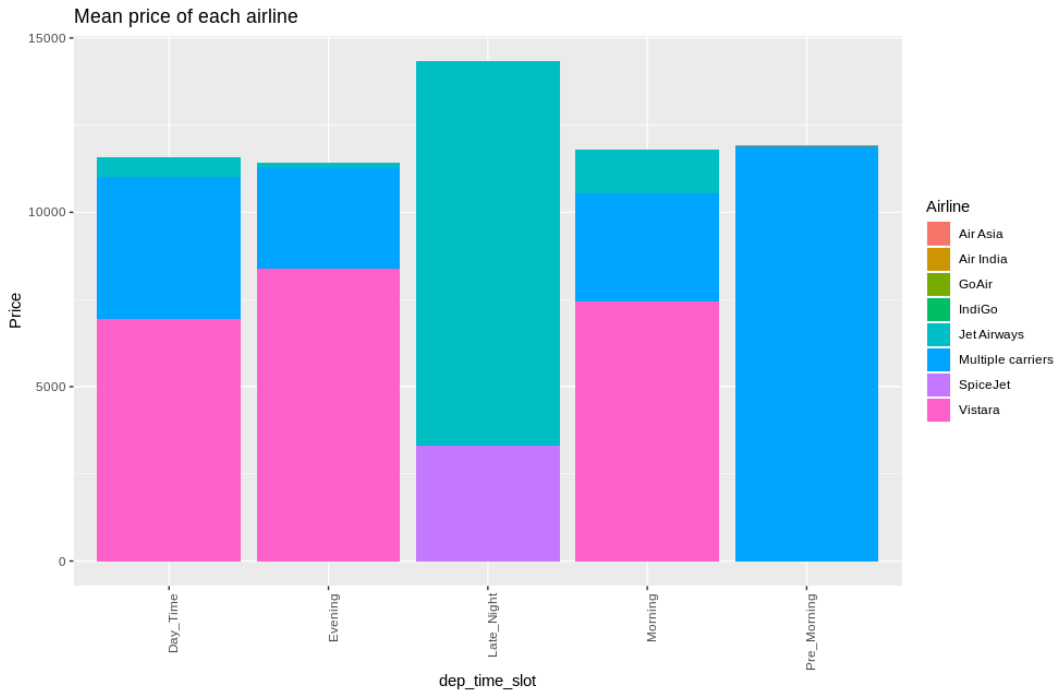
Ggplot2 - ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics.

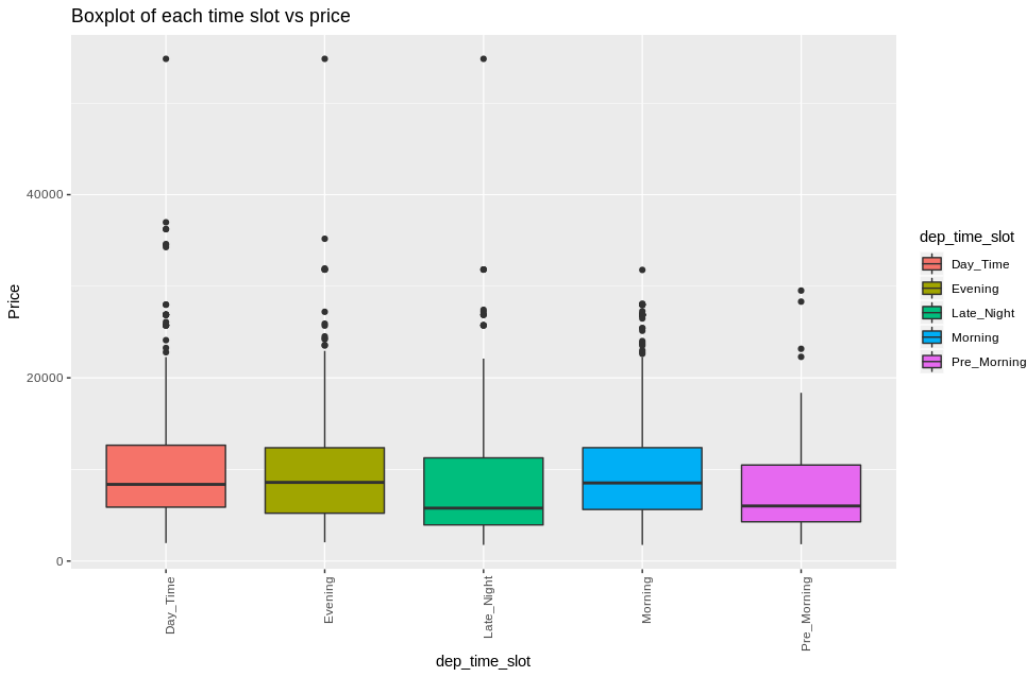
2.3 Data Transformation

The data available in the real world will be of different data types. In R, there are only three data types: Numeric, Character and Logical. In order to work with the data, you might need to convert it into numeric or logical. Before we proceed with applying different learning algorithms on our dataset we need to ensure that the data is of the same format and the required type to work with. In the given dataset, firstly, we ensured that the data in each column in the same format. It can be done manually by analysing the different values present in the column. The unique() function can be used to display the unique values in the column.

In the dataset, the columns Data_of_journey and Departure time is in chr. It is necessary to convert them into date-time objects in order to use them. A new column departure is created which is of POSIXlt %d-%m-%Y %H:%M format.

The column is further split into departure hour and classified into time slots (early Morning, Morning, evening, late night)

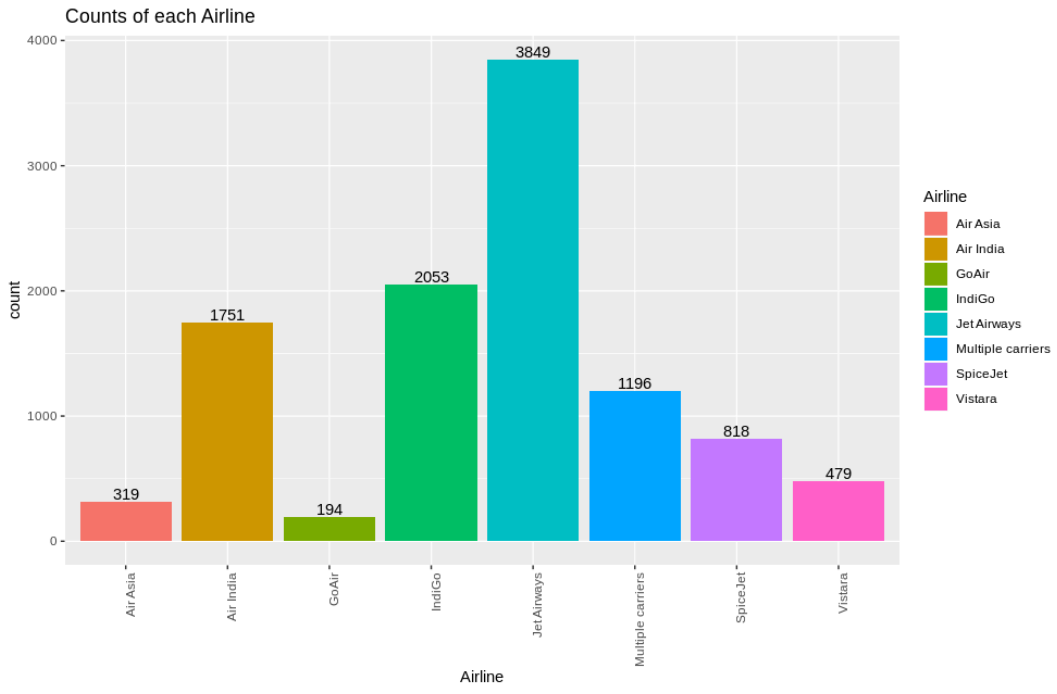


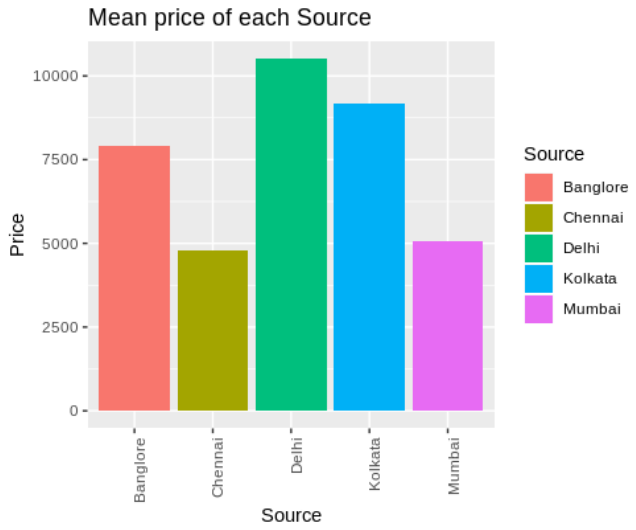


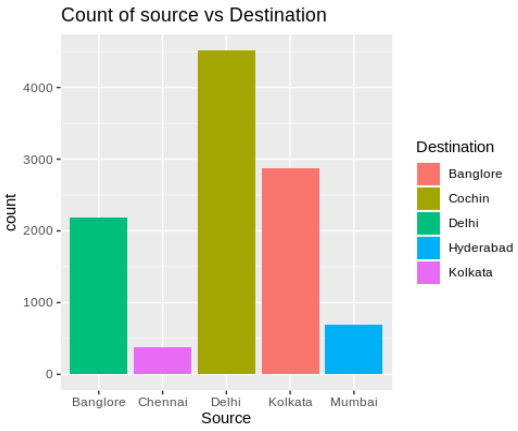
It is inferred from the above graph that evening and morning flights have a higher average price compared to other time slots.

Final Summary of the dataset after processing it.

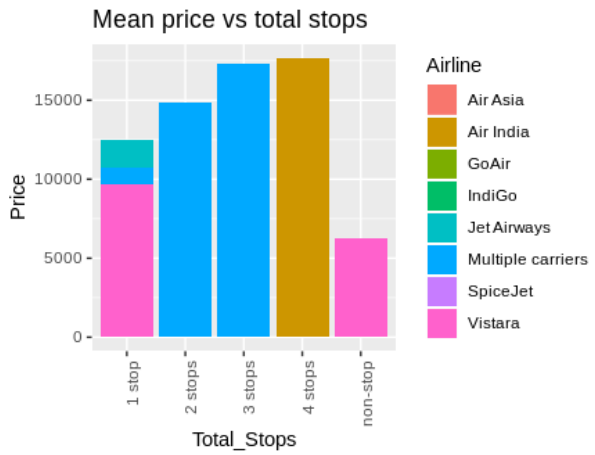
3. DATA VISUALIZATION

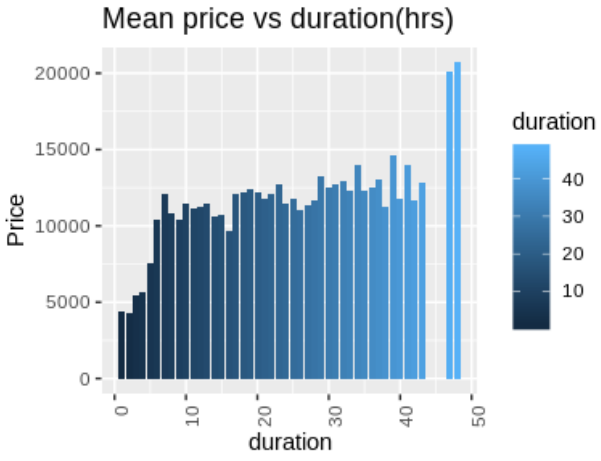


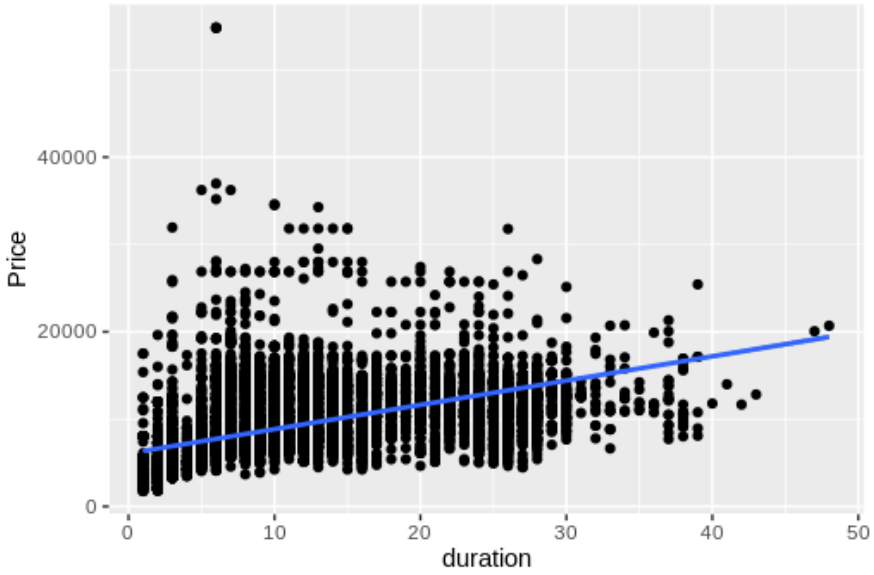




From the above graph, it is clear that Bangalore flights go only to Delhi, Chennai flights go only to Kolkata, flights from Delhi, Kolkata and Mumbai go to Cochin, Bangalore and Hyderabad respectively. There are no flights originating from Cochin and Hyderabad.







Based on our analysis of the above visualizations, features like Additional Info, departure and Route are all dropped. It is understood from the graph that with an increase in the duration of the flight, the price also increases and with more number of stops, the price is also high.

Creation of dummy variables(one-hot encoding): A new column of dummy variables were created for each and every value of airlines, the date of departure, Source, Destination, Total_stops. The package fastDummies was used to create dummy columns

The final dataset consists of 64 columns.

```
> head(train3)
# A tibble: 6 x 64
  Price duration `Airline_Air In… Airline_IndiGo `Airline_Jet Ai… Airline_SpiceJet `Airline_Air As… Airline_Vistara
  <dbl>    <dbl>            <int>          <int>            <int>            <int>            <int>           <int>
1 29528       13                0              0                0                0                0               0
2 23170       15                0              0                0                0                0               0
3 14752       20                1              0                0                0                0               0
4 12599       23                1              0                0                0                0               0
5 16000        2                0              1                0                0                0               0
6 28322       28                1              0                0                0                0               0
# … with 56 more variables: Airline_GoAir <int>, Source_Banglore <int>, Source_Mumbai <int>, Source_Chennai <int>,
#   Source_Kolkata <int>, Destination_Delhi <int>, Destination_Hyderabad <int>, Destination_Kolkata <int>,
#   Destination_Banglore <int>, `Total_Stops_non-stop` <int>, `Total_Stops_2 stops` <int>, `Total_Stops_3 stops` <int>,
#   `Total_Stops_4 stops` <int>, dep_time_slot_Evening <int>, dep_time_slot_Late_Night <int>,
#   dep_time_slot_Morning <int>, dep_time_slot_Pre_Morning <int>, dep_day_03Mar <int>, dep_day_06Mar <int>,
#   dep_day_09Mar <int>, dep_day_12Mar <int>, dep_day_15Mar <int>, dep_day_18Mar <int>, dep_day_21Mar <int>,
#   dep_day_24Mar <int>, dep_day_27Mar <int>, dep_day_01Apr <int>, dep_day_03Apr <int>, dep_day_06Apr <int>,
#   dep_day_09Apr <int>, dep_day_12Apr <int>, dep_day_15Apr <int>, dep_day_18Apr <int>, dep_day_21Apr <int>,
#   dep_day_24Apr <int>, dep_day_27Apr <int>, dep_day_01May <int>, dep_day_03May <int>, dep_day_06May <int>,
#   dep_day_09May <int>, dep_day_12May <int>, dep_day_15May <int>, dep_day_18May <int>, dep_day_21May <int>,
#   dep_day_24May <int>, dep_day_27May <int>, dep_day_01Jun <int>, dep_day_03Jun <int>, dep_day_06Jun <int>,
#   dep_day_09Jun <int>, dep_day_12Jun <int>, dep_day_15Jun <int>, dep_day_18Jun <int>, dep_day_21Jun <int>,
#   dep_day_24Jun <int>, dep_day_27Jun <int>
```

A similar methodology of data analysis and processing is done on the test set.

4. MODELLING

Linear Regression

On performing linear regression on the test set after training on the update train set, the prices are predicted. Linear regression is used to predict the value of a continuous variable Y based on one or more input predictor variables X.

Support Vector machines

Support vector machine(SVM) are supervised learning models with associated learning algorithms that analyze data used for regression and classification analysis. In this method, each data item is plotted as a point in n-dimensional space (n is the number of features), with the value of each feature being the value of a particular coordinate.

Cross validation

It is a resampling procedure used to evaluate a model. In K fold cross-validation, the given test data is split into k number of groups. Here, we have taken the value of K as 10.

Ridge Regression

It is a technique used for analysing data which has multicollinearity. Multicollinearity occurs when variables that are independent in a regression model are correlated. If the degree of correlation is high enough, the model's accuracy will be affected. In order to solve this, ridge regression adds a degree of bias to the regression estimates. It performs L2 regularization which adds a penalty to the square of coefficients of parameters

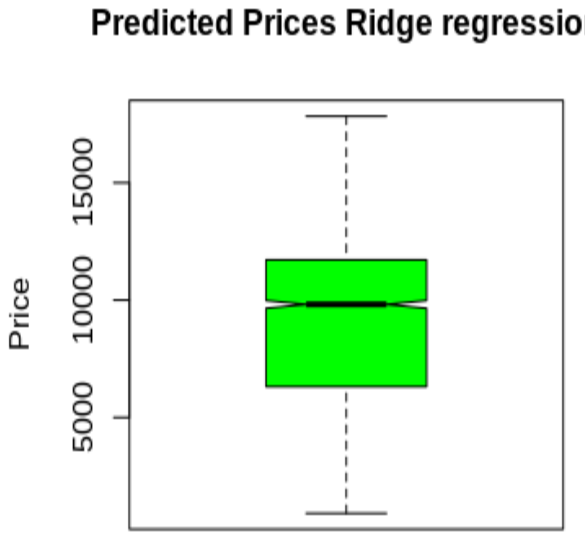
Lasso Regression

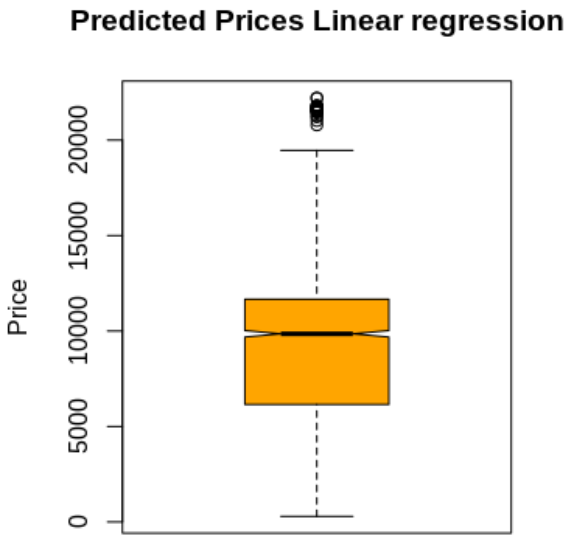
LASSO(Least Absolute Shrinkage and Selection Operator) is a method that uses shrinkage ie the data is shrunk to a central point. It is also used for models with a high degree of multicollinearity.
It performs L1 regularization which adds a penalty to the absolute value of the magnitude of coefficients.

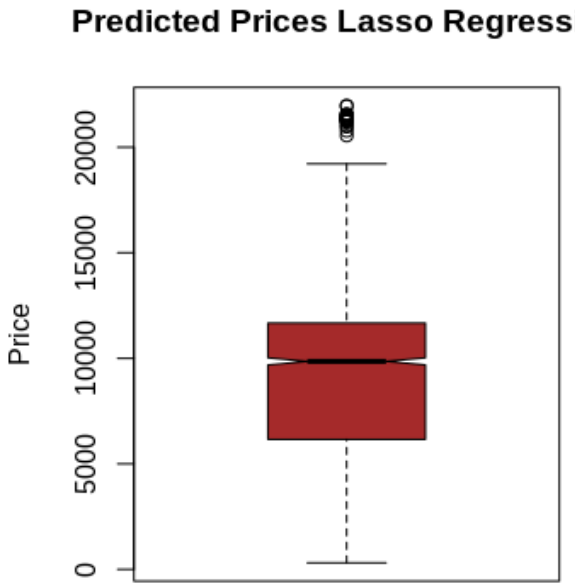
Elastic Net regression

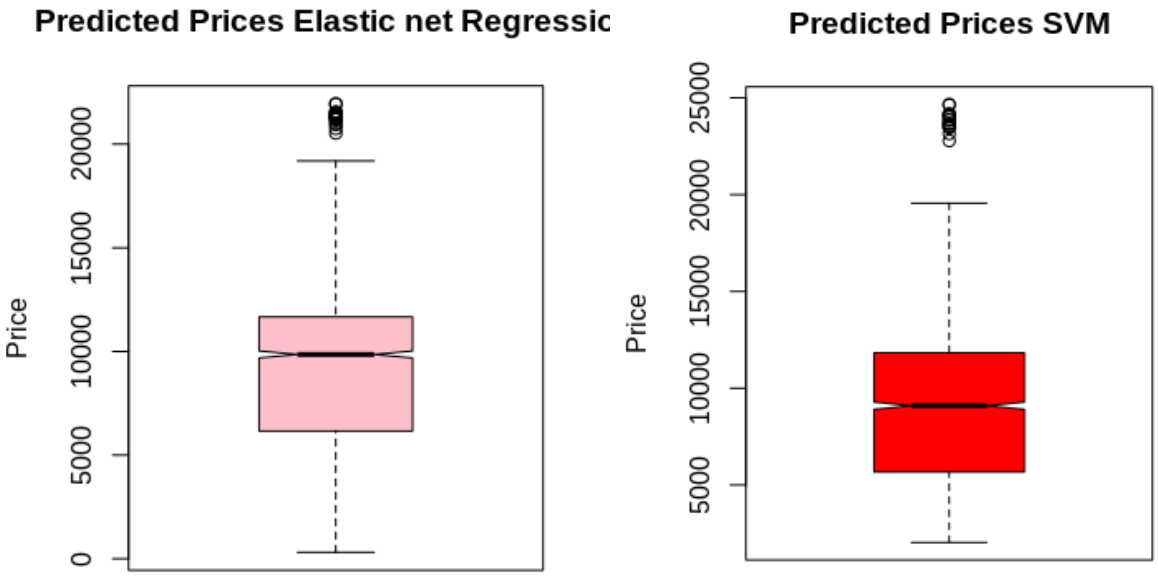
It is a combination of lasso and ridge regression. It adds both L1 and L2 regularization.

From the below box plots of the prices predicted by different machine learning models, we can infer that the prices on an average range from 6000rs to 13000Rs. The prices predicted by SVM is on a lower level, with the mean price falling below 10,000 Rs. Comparing Ridge, Lasso with elastic net Regression









5. CONCLUSION

According to the analysis done, with the availability of proper datasets, it is possible to predict the airline tickets which would help customers travel economically. There is a significant market for reliable pricing prediction models which would assist buyers while planning their travel. Currently, there are no reliable models which provide nearly accurate estimates to the customers. With further development, we can help customers cope up with the dynamic price changes.

In this analysis, we studied the prices of airlines data spanning over four months, after performing exploratory analysis, we arrived at our conclusions on the important factors that significantly affect the airline prices. We then used different machine learning models to predict airline prices for the month of October. There is much scope for additional cost reductions that can be discovered to obtain results close to optimal results and help customers save more money.

6. REFERENCES

1. Wang, Tianyi, et al. "A Framework for Airfare Price Prediction: A Machine Learning Approach." *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*. 2019.
2. Abdella, Juhar Ahmed, et al. "Airline ticket price and demand prediction: A survey." *Journal of King Saud University-Computer and Information Sciences* (2019).
3. Tziridis, K., et al. "Airfare prices prediction using machine learning techniques." *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017.
4. https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/
5. https://www.r-bloggers.com/explore-your-dataset-in-r/
6. https://cran.r-project.org/web/packages/jtools/vignettes/summ.html
7. https://www.machinelearningplus.com/machine-learning/complete-introduction-linear-regression-r/
8. http://rpubs.com/kdomijan/325930
9. https://www.analyticsvidhya.com/blog/2018/05/improve-model-performance-cross-validation-in-python-r/
10. https://dzone.com/articles/doing-residual-analysis-post-regression-in-r