**Flight Fare Prediction Analysis**

**By:**

**Ayush Pandey**

**Abstract:**

Nowadays, airline ticket prices can vary dynamically and significantly for the same flight on different dates and can be sometime hard to guess the prices.

We might have often heard travellers saying that flight ticket prices are so unpredictable. So this project is to analyze the different factors that affect the prices of the flight and to predict the price based on various parameters so that a person can get an idea of the prices and can decide whether to buy a ticket of that airline or not.

**About Data:**

The dataset consists of 10,683 records and 11 different features containing the data like the name of the airline, the date of journey, the source from which the service of the flight begins, the destination where the service ends, route taken by the flight to reach the destination, the time when the journey starts from the source, time of arrival at the destination, total duration of the flight, total stops between the source and its destination, more information about the flight, and the price of the ticket from the month of March to June 2019.

**Approach:**

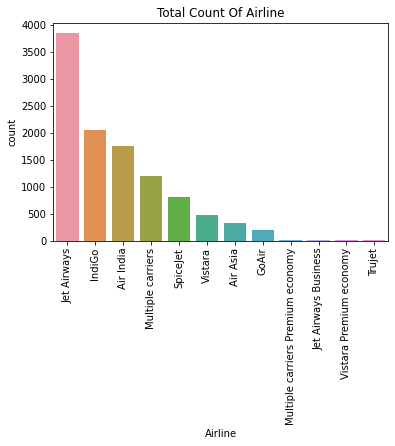
* Various python libraries such as NumPy, Pandas, Matplotlib, Seaborn are used for the purpose of mathematical calculations, extraction and visualization of the data.
* The extracted dataset was found to have null values which are then effectively handled to achieve accurate analysis.
* Various different regression models was implemented to get the best prediction of the flight price and from that Random Forest Regression technique was selected as it had the best R-Squared score (measures the strength of the relationship between the model and the dependent variable on a convenient 0 – 100% scale).

**ANALYSIS AND VISUALIZATION**

1. **Type of Airlines along with their count:**

**Purpose:**

* To get the total count of each airline from the month March to June.



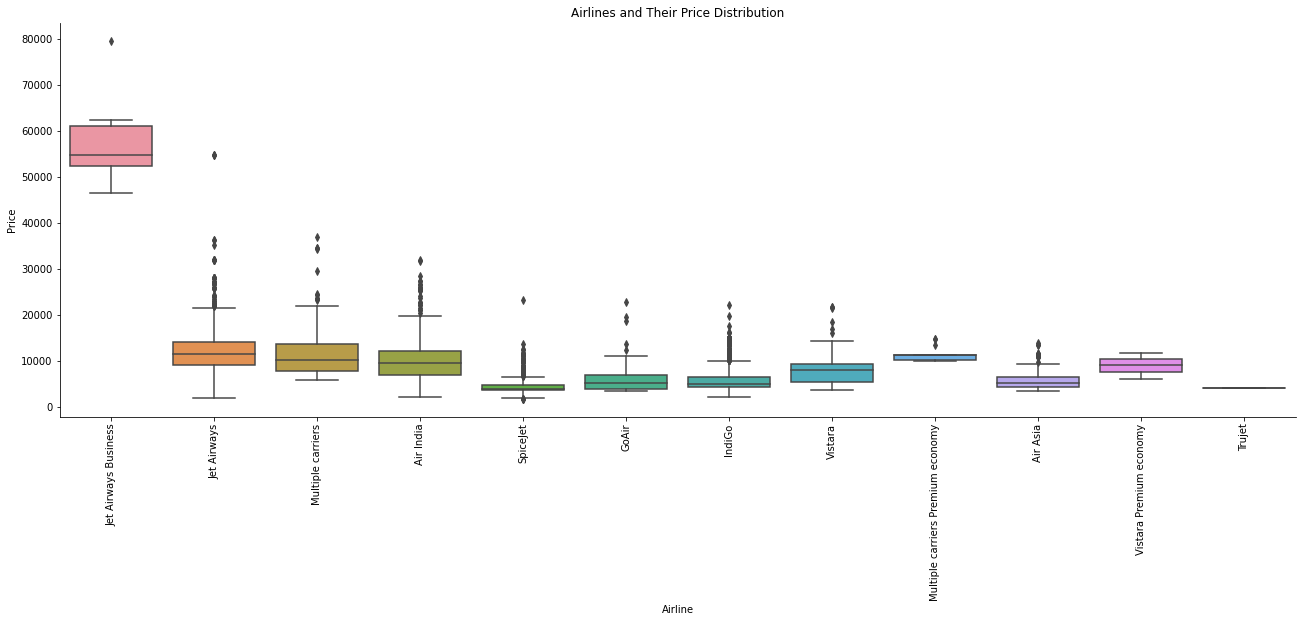
**Inference:**

* Jet Airways is the most preferred airline with the highest row count of 3,849 followed by Indigo (2,053) and AirIndia (1,751).
* Count for Vistara Premium economy, Trujet, Multiple carries premium economy and Jet airways business is quite low.

1. **Graphical representation between the type of Airlines and Price:**

**Purpose:**

* To plot a graph between type of Airlines and Price, to make some useful inferences.



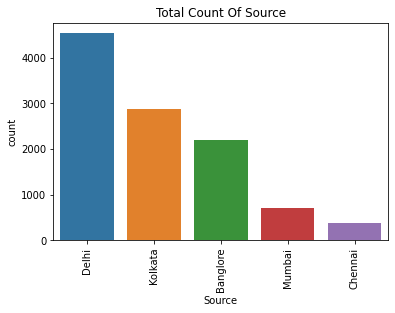
**Inference:**

* Jet airways, Multiple carriers and Air India are full-service airlines and are always highly priced due to various facility they provide and other airlines have almost similar median with minimal fluctuations.
* Through this visualization the travellers can see the highest to lowest price and can know the difference of prices in each airline.

1. **Total number of flights that take off from each city:**

**Purpose:**

* To get the total count of flight that take off from each city.



**Inference:**

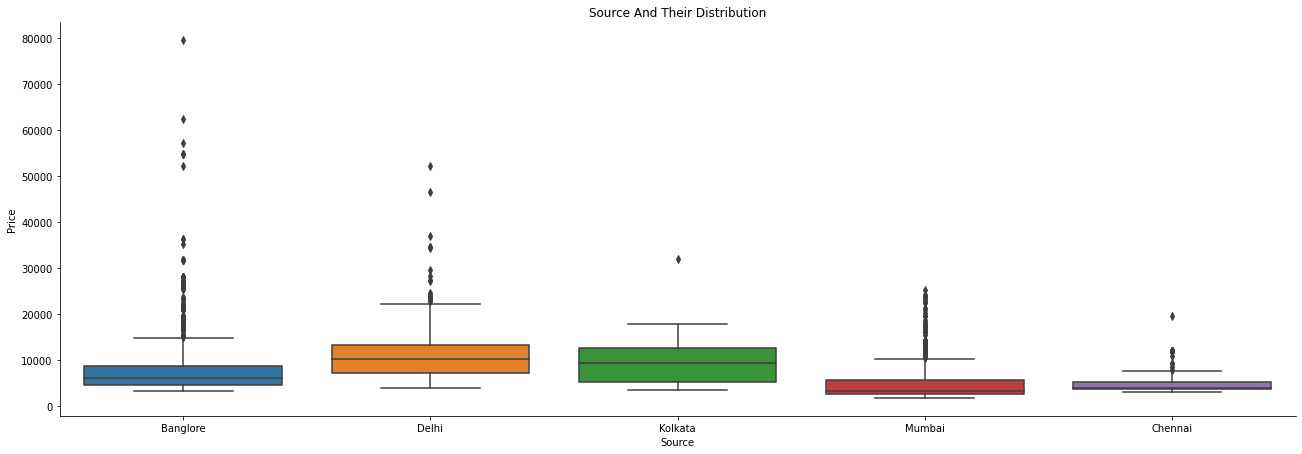
* Majority of the flights take off from Delhi and from March to June total 4,536 flights have take off followed by Kolkata where more than 2,800 (2,871) flights have take off.

* Chennai has the minimum count of flight take off where from month March to June only 381 flights have take off.

1. **Boxplot representation between the Flight Source and Price:**

**Purpose:**

* To plot a boxplot between the Flight Source and Price and to obtain useful inferences.



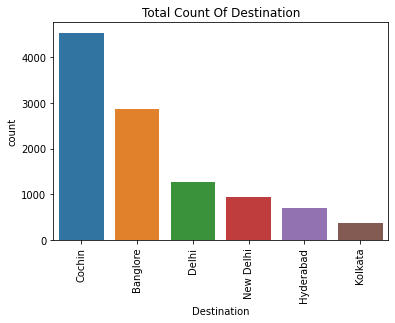
**Inference:**

* From the above visualization, the highest average value of flight fare is of Delhi followed by Kolkata.
* Outliners are present but we will not remove them as prices are affected by date of journey, source and total stops.

1. **Number of flights going to each city:**

**Purpose:**

* To count the number of flights going to each city.



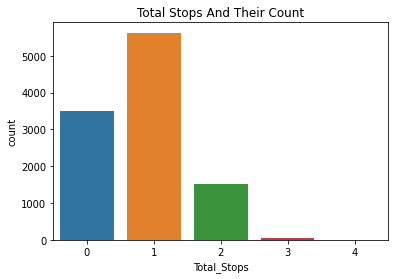
**Inference:**

* Maximum flight land is in Cochin where more than 4,500 flight have landed followed by Bangalore where more than 2,800 (2,871) flight have landed in the month from March to June.
* Kolkata has the least rumber of receiving the flight where only 381 flights have landed in the month from March to June.

1. **Graphical representation of Flight Stopage along with their count:**

**Purpose:**

* To plot a countplot on the number of stops and to obtain some useful inferences.



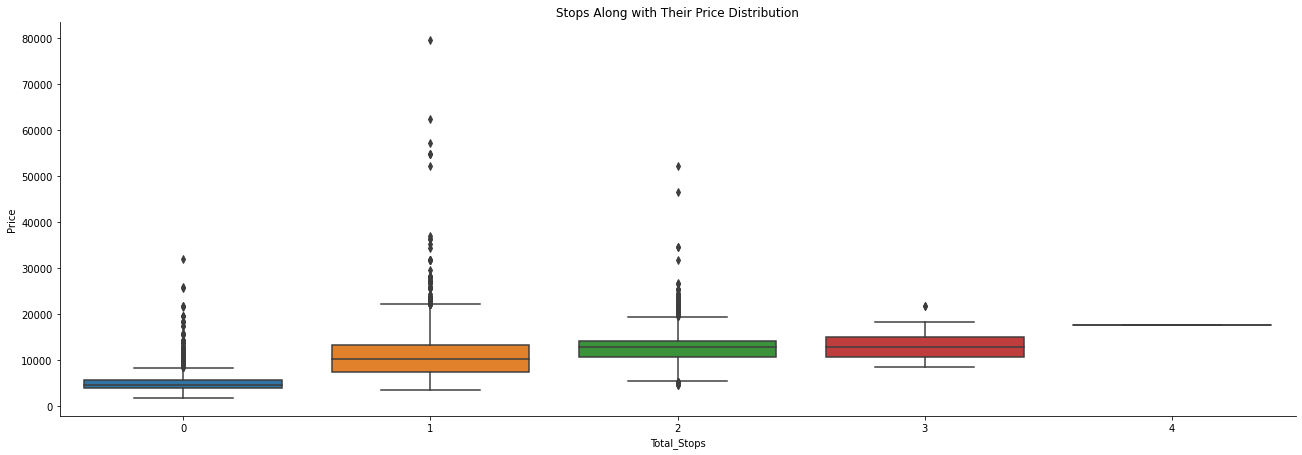
**Inference:**

* Majority of the flights have one stop and from March to June total 5,625 flights have one stop between their start source and end source followed by Non-Stop flight with more than 3,490 flights which have no stopage between it’s source and destinaton.
* Flights with 4 stopage between its start souce and end source is 1 from the month of March to June followed by 3 stops with 45 flights which have 3 stopage between start source and destination source.

1. **Boxplot representation between Total Stops and Price:**

**Purpose:**

* To plot a boxplot between the Total Stops and Price and to obtain useful inferences.



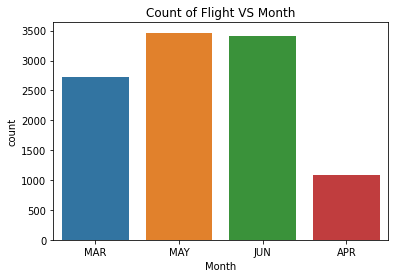
**Inference:**

* Airlines with 1 or 2 stops have many outliners and hence the price varies. On the contrary price for airline with 4 stops is not fluctuating.
* We will not remove the outliners as the flight prices are affected by the number of stops.

1. **Graphical result analysis between Month and count of flight**

**Purpose:**

* To visualize the graph between the count of flight for each month.



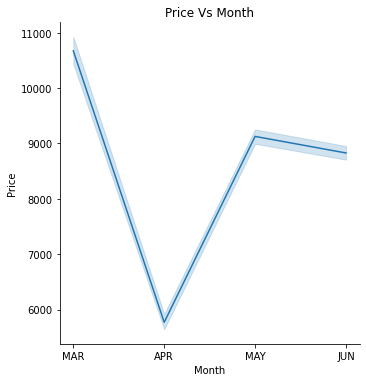
**Inference:**

* Flights in May and June are quite high with more than 3,400 flights which shows people travel more during holiday month.
* April has the lowest number of flights with less than 1,080 (1079) flights which shows people travel less in April.

1. **Graphical result analysis between Month and Price**

**Purpose:**

* To plot a graph between the Month and Price to make some useful inferences.



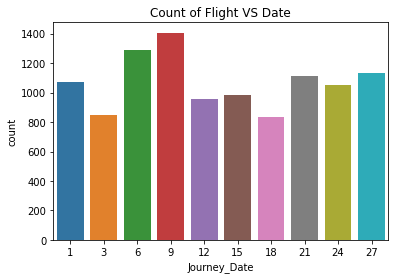
**Inference:**

* The prices are higher in the month of March as people travel more, the company increases the expenses.
* Some airline company increase the price for the business purpose, so this analysis can help passengers to see which month is best to travel with affordable prices as from month to month price changes.

1. **Graph between date of journey and their count:**

**Purpose:**

* To plot a countplot on the number of flights that take off on that particular date.
* The journey day plays an important role in price as the flight fare for weekdays might be lesser and for weekends the price might be higher based on the offers and airlines chosen to board the flight.



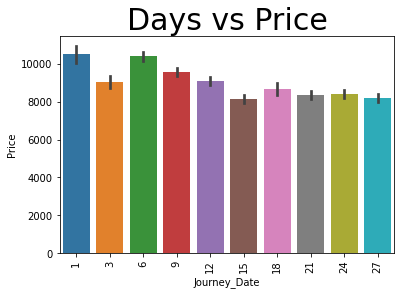
**Inference:**

* From the above visulaization, the maximum count of flight is on 9th with 1,406 flights followed by 6th with 1,287 flights.
* The distribution of the number of flight is almost similar for all dates.

1. **Graphical result analysis between the Journey Date and Price:**

**Purpose:**

* To plot a graph between Date and Price and to make some useful inferences.



**Inference:**

* From the above visualization, flight prices at the start of the month are a bit high as compared to the end date and also the prices increases as the date of booking is near to the date of journey.
* The journey date plays an important role in price as the flight charges for weekdays might be lesser and for weekends the price might be higher based on the offers and airlines chosen to board the flight.

1. **Relationship between the Journey date, Journey Month and Departure Hour:**

**Purpose:**

* To check whether the columns Journey date, Journey Month and Departure Hour are dependent or independent with the help of one-way Anova test.
* Null Hypothesis-> The three columns Journey date,Journey Month and Departure Hour are co-reated that is dependent.
* Alternate Hypothesis-> The three columns Journey date,Journey Month and Departure Hour are independent of each other.
* Level of Significance (Is the probability of rejecting the null hypothesis when it is true.) as 0.05. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

**Inference:**

* The p value is 5.39e-90 and since the p-value (5.39e-90) of our experiment is less than the level of significance(0.05) so we will reject the null hypothesis and accept the alternate hypothesis that is the three columns Journey date,Journey Month and Departure Hour are independent of each other.

1. **Relationship between the Arrival Hour, Arrival Minute and Duration Hour:**

**Purpose:**

* To check whether the columns Arrival Hour, Arrival Minute and Duration Hour are dependent or independent with the help of one-way Anova test.
* Null Hypothesis->The three columns Arrival Hour,Arrival Minute and Duration Hour are co-reated that is dependent.
* Alternate Hypothesis->The three columns Arrival Hour,Arrival Minute and Duration Hour are independent of each other.
* Level of Significance (Is the probability of rejecting the null hypothesis when it is true.) as 0.05. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.

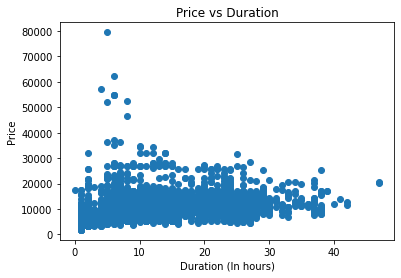
**Inference:**

* The p value is 0.0 and since the p-value (0.0) of our experiment is less than the level of significance(0.05) so we will reject the null hypothesis and accept the alternate hypothesis that is the three columns Arrival Hour,Arrival Minute and Duration Hour are independent of each other.

1. **Pearson correlation between the Price and the total duration:**

**Purpose:**

* To find out the association between the two variables( Price and Total Duration), pearson correlation coefficient is used to find out how strong is the relationship between the two variables.
* Null Hypothesis->The independent column that is Duration Hour is not correlated with the target column that is Price.
* Alternate Hypothesis->The independent column that is Duration Hour is correlated with the target column that is Price.
* Level of Significance (Is the probability of rejecting the null hypothesis when it is true.) as 0.05. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.



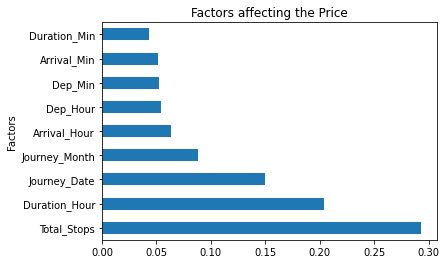
**Inference:**

* The p-value is 0.0 an since the p-value(0.0) is less than the level of significance(0.05) we reject the Null Hypothesis and accept the Alternate Hypothesis that is the target column (Price) and the independent column (Duration\_Hour) are correlated as pearson coefficient is 0.51 which means moderately correlated.This shows that the time duration plays an important role in affecting the price of the flight.

1. **Pearson correlation between the Price and the Total Stops:**

**Purpose:**

* To find out the association between the two variables( Price and Total Stops), pearson correlation coefficients is used to find out how strong is the relationship between the two variables.
* Null Hypothesis->The independent column that is Total Stops is not correlated with the target column that is Price.
* Alternate Hypothesis->The independent column that is Total Stops is correlated with the target column that is Price.
* Level of Significance (Is the probability of rejecting the null hypothesis when it is true.) as 0.05. A significance level of 0.05 indicates a 5% risk of concluding that a difference exists when there is no actual difference.



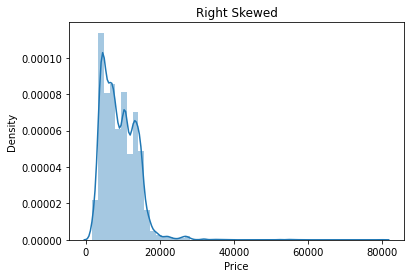
**Inference:**

* The p-value is 0.0 an since the p-value (0.0) is less than the level of significance(0.05) we reject the Null Hypothesis and accept the Alternate Hypothesis that is the target column (Price) and the independent column (Total\_Stops) are correlated as pearson coefficient is 0.604 which means moderately correlated.This shows that the ‘stops’ plays an important role in affecting the price of the flight.
* And from the above visualization it is confirmed that the price of the flight are dependent on the total number of stops.

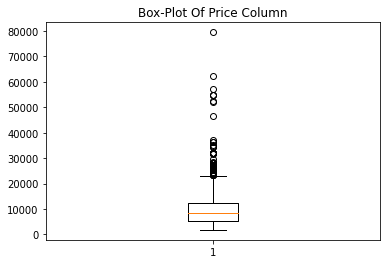
1. **Distribution of the Price column:**

**Purpose:**

* To obtain the maximum, minimum price of the flight along with its distribution.



**Figure 1**



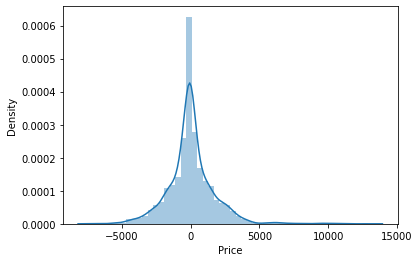
**Inference:**

* Maximum price flight is Rs 79,512 and minimum price flight is Rs 1,759. Majority of the flights have price range between 1,759-20,000 and the number of flights having price greater than 20,0000 are quite less.
* From the above visualization (Figure 1) , it is clear that the Price range is skewed toward right.

1. **Random Forest Regression Technique for Prediction:**

**Purpose:**

* Random forest regression is a supervised learning algorithm. The benefit of the random forest is, it very well may be utilized for both characterization and relapse issue which structure most of current machine learning framework. Random forest forms numerous decision trees, whats more, adds them together to get an increasingly exact and stable expectation. Random Forest has nearly the equivalent parameters as a decision tree or a stowing classifier model. It is very simple to discover the significance of each element on the expectation when contrasted with others in this calculation.



**Inference:**

* Random Forest model has the best accuracy in predicting the price of the flight with the help of different features with an R-Squared score (It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model) of above 85% (85.44%).
* From the above visualization, the difference in the actual price and the predicted price of flight is forming a bell curve as most of the values are present in the range of -5000 to +5000.
* Mean Absolute error for this model is around 1,119 and root mean square error is 1,790. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) are metrics used to evaluate a Regression Model. These metrics tell us how accurate our predictions are and, what is the amount of deviation from the actual values.RMSE is the **R**oot of the **M**ean of the **S**quare of **E**rrors and MAE is the **M**ean of **A**bsolute value of **E**rrors.

**Conclusion:**

* Jet Airways is the most preferred airline with the highest number of row count,followed by Indigo and AirIndia.
* Majority of the flights take off from Delhi.
* Maximum flights land in Cochin.
* Majority of the flights have 1 stop, flight with 3 and 4 stops are quite low.
* There is no specific dates when the flight travel, the distribution is almost the same for all dates.
* People tend to travel more in the month of May and June as they are the holiday months.
* People travel less in the month of April.
* Majority of the flight have travel time for 2-6 hrs as all the flights are domestic flights.
* Maximum price flight is Rs 79,512 and minimum price flight is Rs 1,759. Majority of the flights have price range between 1,759-20,000 and the number of flights having price greater than 20,0000 are quite less.
* Prices are higher in the month of March.
* As the number of stops and duration of flight increases the price of the flight increases as they are corelated.
* Random Forest model has the best accuracy in predicting the price of the flight with the help of different features with an R-Squared score of above 85% (85.44%).

**Reference:**

* Dataset for Flight Price Pridiction:

<https://www.kaggle.com/nikhilmittal/flight-fare-prediction-mh>