FLIP ROBO

# Flight Price Prediction


Submitted by:

Sumair Dhir

# ACKNOWLEDGMENT

I express my sincere gratitude to Flip Robo Technologies for giving me the opportunity to work on this project on Flight Price Prediction using machine learning algorithms. I acknowledge my indebtedness to the authors of the paper titled: “Airline ticket price and demand prediction: A survey” and the online article titled: “Trying to Predict Airfares When The Unpredictable Happens” for providing me with invaluable insights and knowledge of the various factors that determine the price of Flight tickets.

# 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 –

- 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)

Therefore, a predictive model to accurately predict Air fares is required to be made.

## Conceptual Background of the Domain Problem

Predictive modelling, Regression algorithms are some of the machine learning techniques used for predicting Flight Ticket prices. Identifying various relevant attributes like Airline Brand, flight duration, source and destination etc are crucial for working on the project as they determine the valuation of air fare.

## Review of Literature

A Research paper titled: “Airline ticket price and demand prediction: A survey” by Juhar Ahmed Abdella and online article titled: “Trying to Predict Airfares When The Unpredictable Happens” were reviewed and studied to gain insights into all the attributes that contribute to the pricing of flight tickets.

It is learnt that deterministic features like Airline Brand, flight number, departure dates, number of intermediate stops, week day of departure, number of competitors on route and aggregate features – which are based on collected historical data on minimum price, mean price, number of quotes on non-stop,1-stop

and multi-stoppage flights are some the most important factors that determine the pricing of Flight Tickets.

- Airline ticket price and demand prediction: A survey - ScienceDirect
- Flight Price Predictor | American Express GBT (amexglobalbusinesstravel.com)

## Motivation for the Problem Undertaken

With airfares fluctuating frequently, knowing when to buy and when to wait for a better deal to come along is tricky. The fluctuation in prices is frequent and one has limited time to book the cheapest ticket as the prices keep varying due to constant manipulation by Airline companies. Therefore, it is necessary to work on a predictive model based on deterministic and aggregate feature data that would predict with good accuracy the most optimal Air fare for a particular destination, route and schedule.

# Analytical Problem Framing

## Mathematical/ Analytical Modeling of the Problem

Various Regression analysis techniques were used to build predictive models to understand the relationships that exist between Flight ticket price and Deterministic and Aggregate features of Air travel. The Regression analysis models were used to predict the Flight ticket price value for changes in Air travel deterministic and aggregate attributes. Regression modelling techniques were used in this Problem since Air Ticket Price data distribution is continuous in nature.

In order to forecast Flight Ticket price, predictive models such as ridge regression Model, Random Forest Regression model, Decision tree Regression Model, Support Vector Machine Regression model and Extreme Gradient Boost Regression model were used to describe how the values of Flight Ticket Price depended on the independent variables of various Air Fare attributes.

## Data Sources and their formats

The Dataset was compiled by scraping Data for various Air Fare attributes and Price from https://www.yatra.com/ and https://www.easemytrip.com/

The data was converted into a Pandas Dataframe under various Feature and Label columns and saved as a .csv file.

```
In [3]: DF.head(50)
```

Out[3]:

| | Unnamed: 0 | Airline | Flight Number | Date of Departure | From | To | Duration | Total Stops | Price |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | Air Asia | I5-764 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,395 |
| 1 | 1 | Air Asia | I5-482 | Fri, Feb 11 | New Delhi | Mumbai | 2h 15m | Non Stop | 2,395 |
| 2 | 2 | SpiceJet | SG-8701 | Fri, Feb 11 | New Delhi | Mumbai | 2h 15m | Non Stop | 2,407 |
| 3 | 3 | SpiceJet | SG-8157 | Fri, Feb 11 | New Delhi | Mumbai | 2h 20m | Non Stop | 2,407 |
| 4 | 4 | Vistara | UK-927 | Fri, Feb 11 | New Delhi | Mumbai | 2h 05m | Non Stop | 2,410 |
| 5 | 5 | Vistara | UK-975 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,410 |
| 6 | 6 | Vistara | UK-993 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,410 |
| 7 | 7 | Vistara | UK-951 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,410 |
| 8 | 8 | Vistara | UK-933 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,410 |
| 9 | 9 | Vistara | UK-985 | Fri, Feb 11 | New Delhi | Mumbai | 2h 10m | Non Stop | 2,410 |
| 10 | 10 | IndiGo | 6E-2331/905 | Sun, Mar 13 | New Delhi | Goa | 7h 10m | 1 Stop | 3,288 |
| 11 | 11 | IndiGo | 6E-967/6433 | Sun, Mar 13 | New Delhi | Goa | 5h 50m | 1 Stop | 3,626 |

## Dataset Description

**The Independent Feature columns are**:

- Airline: The name of the airline.
- Flight Number: Number of Flight
- Date of Departure: The date of the journey
- From: The source from which the service begins
- To: The destination where the service ends
- Duration: Total duration of the flight
- Total Stops: Total stops between the source and destination.

**Target / Label Column:**

- Price: The Price of the Ticket

## Data Preprocessing Done

- Duplicate data elements in various columns: ‘Airline’,’From’,’To’, which had their starting letters in upper case and lower case were converted to data elements starting with uppercase letters.
- Data in column ‘Price’ was converted to int64 data type.
- Columns: Unnamed: 0(just a series of numbers) was dropped since it doesn't contribute to building a good model for predicting the target variable values.
- The Date format of certain data elements in ‘Date of Departure’ was changed to match the general Date format of majority of the data elements of the column.

### Feature Engineering:

- To better understand the relationships between Flight price and Air Fare attributes, ‘Day’,’Date’ and ‘Month’ columnswere created based on data of existing column: ‘Date of Departure’.
- The values in Column: ‘Duration’ was converted from Hours-Minutes format to minute format and the data type was converted to int64.

## Data Inputs- Logic- Output Relationships

- The Dataset consists mainly of Int and Object data type variables. The relationships between the independent variables and dependent variable were analyzed.

# Hardware and Software Requirements and Tools Used

## Hardware Used:

- Processor: Intel Core i3
- Physical Memory: 12.0GB (2400MHz)

## Software Used:

- Windows 10 Operating System
- Anaconda Package and Environment Manager: Anaconda is a distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment. The distribution includes data-science packages suitable for Windows and provides a host of tools and environment for conducting Data Analytical and Scientific works. Anaconda provides all the necessary Python packages and libraries for Machine learning projects.
- Jupyter Notebook: The Jupyter Notebook is an open-source web application that allows data scientists to create and share documents that integrate live code, equations, computational output, visualizations, and other multimedia resources, along with explanatory text in a single document.
- Python3: It is open source, interpreted, high level language and provides great approach for object-oriented programming. It is one of the best languages used for Data Analytics and Data science projects/application. Python provides numerous libraries to deal with mathematics, statistics, and scientific function.

- Python Libraries used:
    - Pandas: For carrying out Data Analysis, Data Manipulation, Data Cleaning etc
    - Numpy: For performing a variety of operations on the datasets.
    - matplotlib.pyplot, Seaborn: For visualizing Data and various relationships between Feature and Label Columns
    - Scipy: For performing operations on the datasets
    - Statsmodels: For performing statistical analysis

- sklearn for Modelling Machine learning algorithms, Data Encoding, Evaluation metrics, Data Transformation,Data Scaling, Component analysis,Feature selection etc.

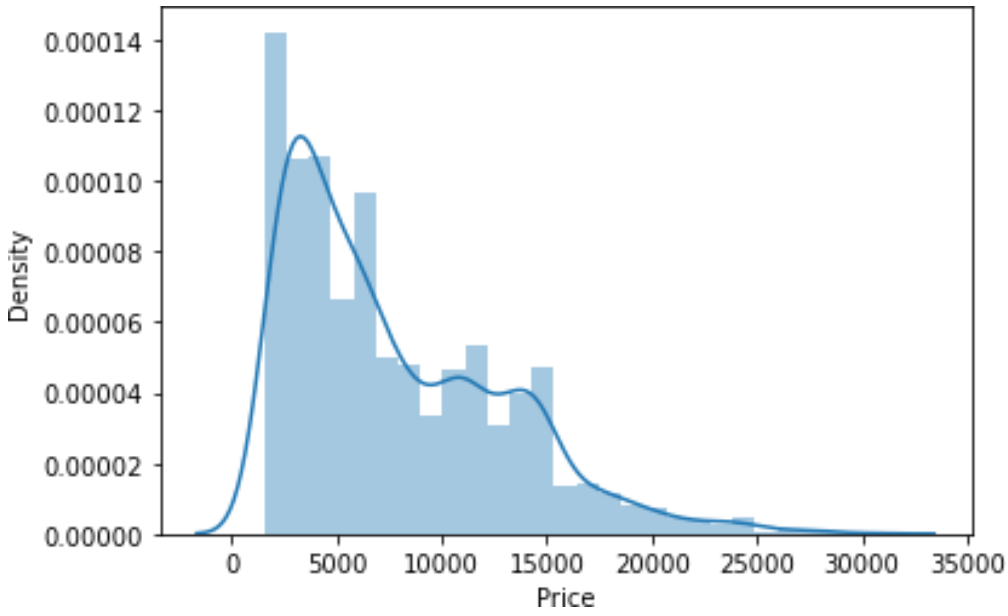
# Exploratory Data Analysis

### Visualizations

Barplots,Distplots,Boxplots,Countplots,lineplots were used to visualise the data of all the columns and their relationships with Target variable.

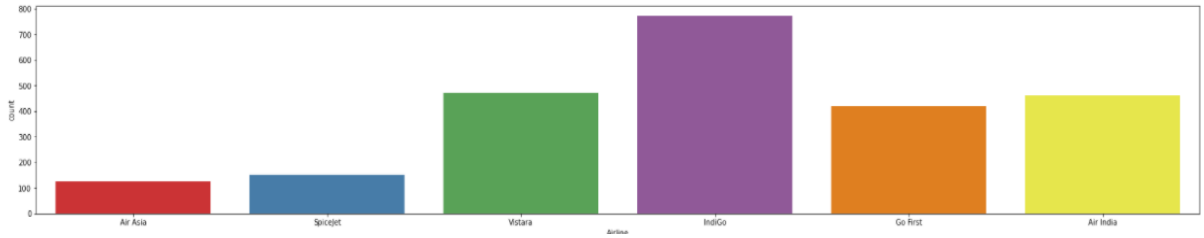
## Univariate Analysis

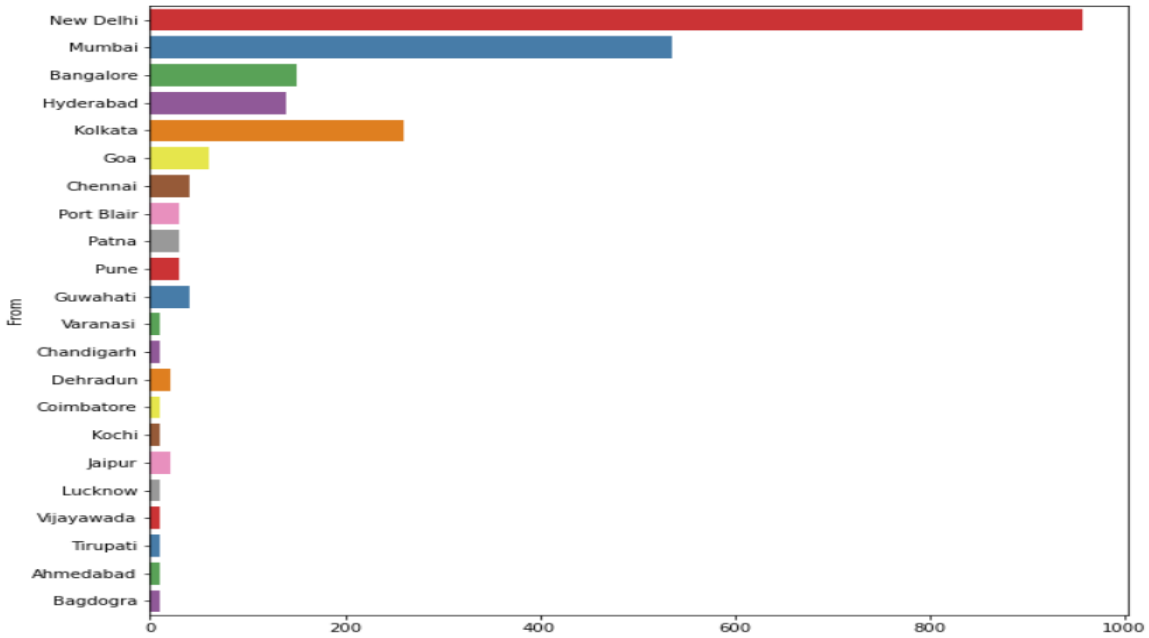
## Analyzing the Target Variable

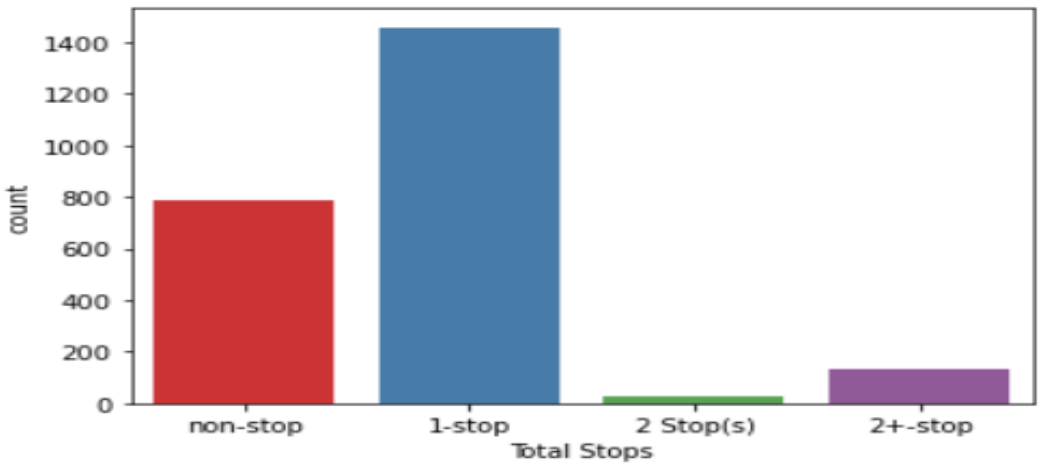


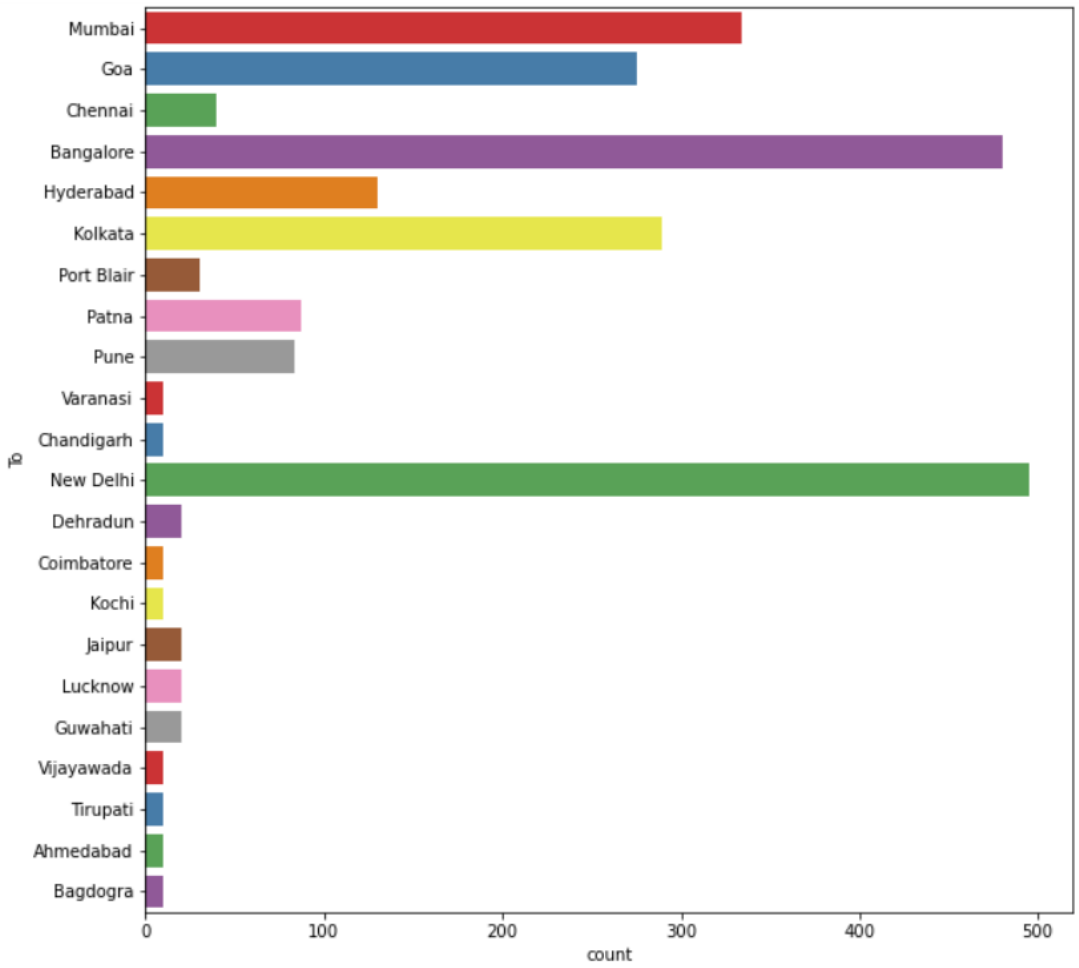

From the graph above it is observed that the Price data forms a continuous distribution with mean of 6511.87 and tails of from 15000 mark and the distribution is skewed.

# Analyzing the Feature Columns





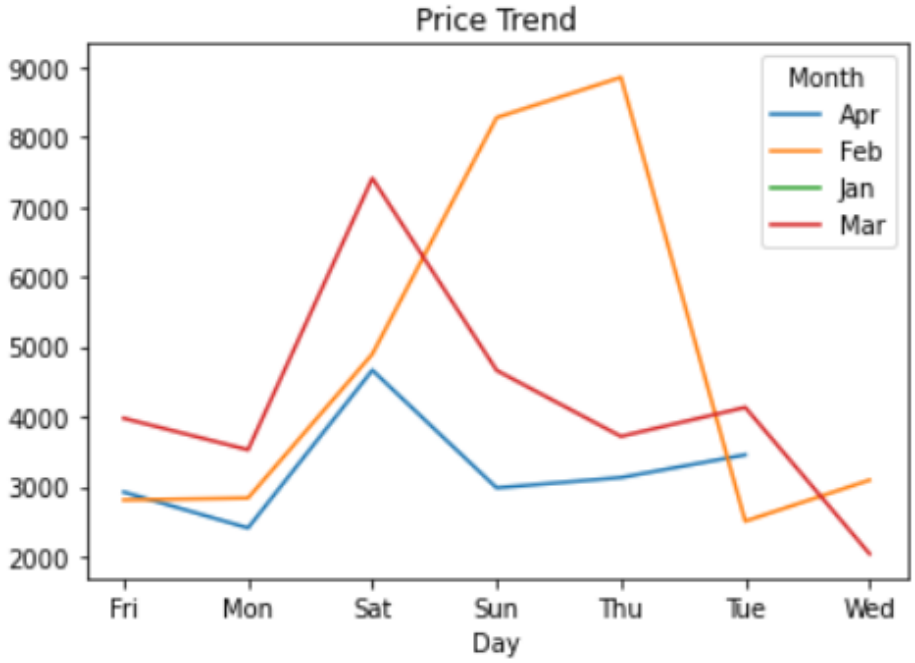

Following observations are made from graphs above:

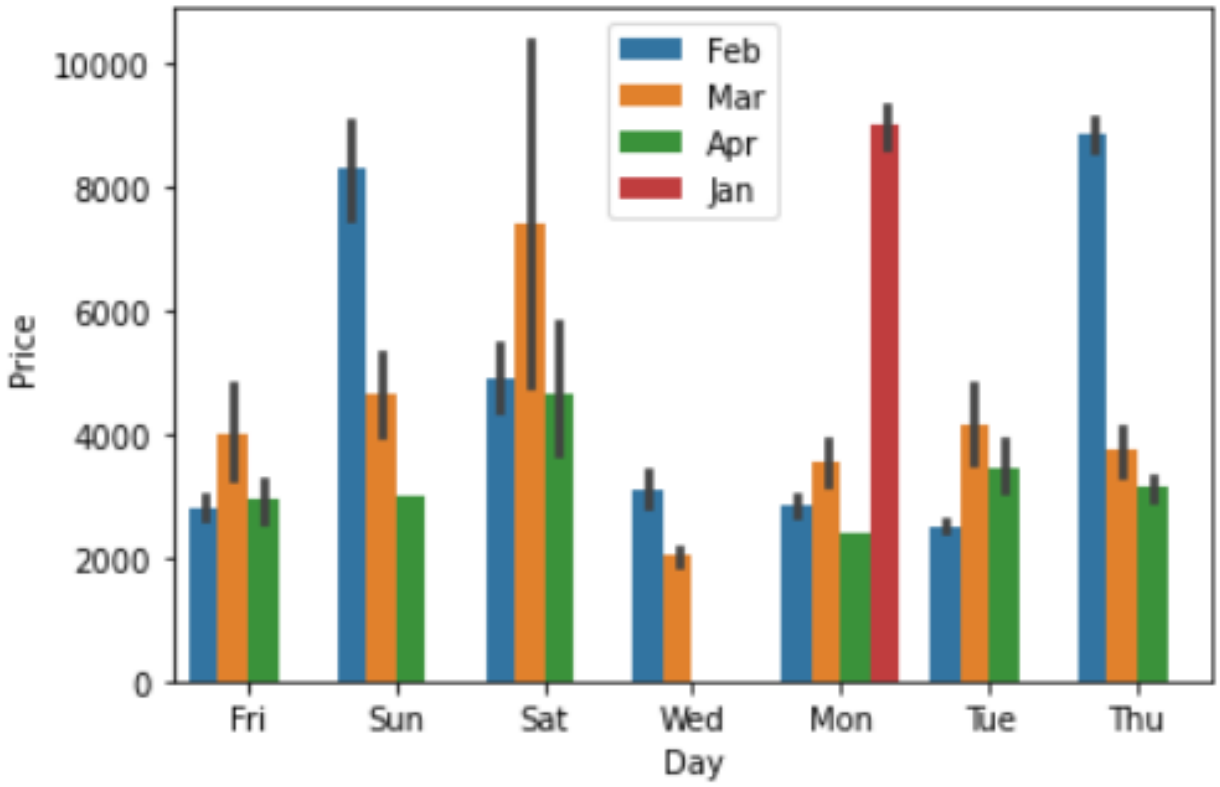
- IndiGo has the highest number of flights followed by Vistara and Air India
- Highest number of flights are from Delhi followed by Mumbai, Kolkata,Bangalore and Hyderabad
- New Delhi is the most popular destination followed by Bangalore, Mumbai, Kolkata, and Goa
- Highest number of flights have only 1 stop between source and destination while 2nd highest number of flights are non-stop.

## Bivariate Analysis

## Interpreting Relationship between Dependent Variable and Independent Variable Columns

### Analyzing Relationship between Day, Month columns and Price

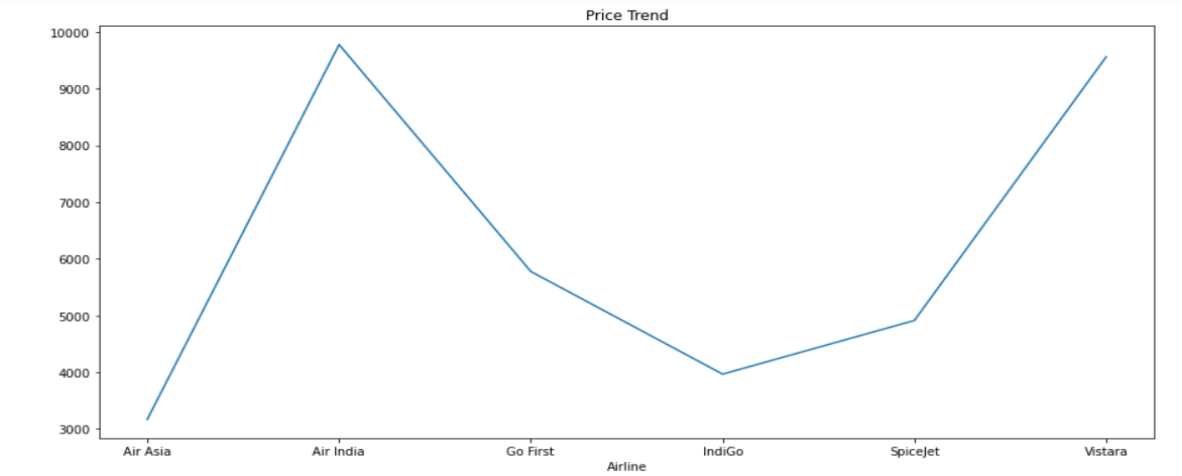


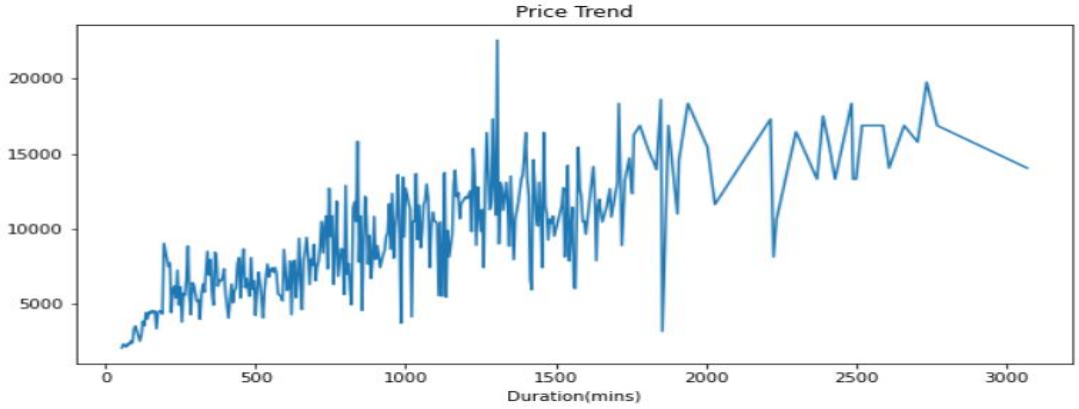



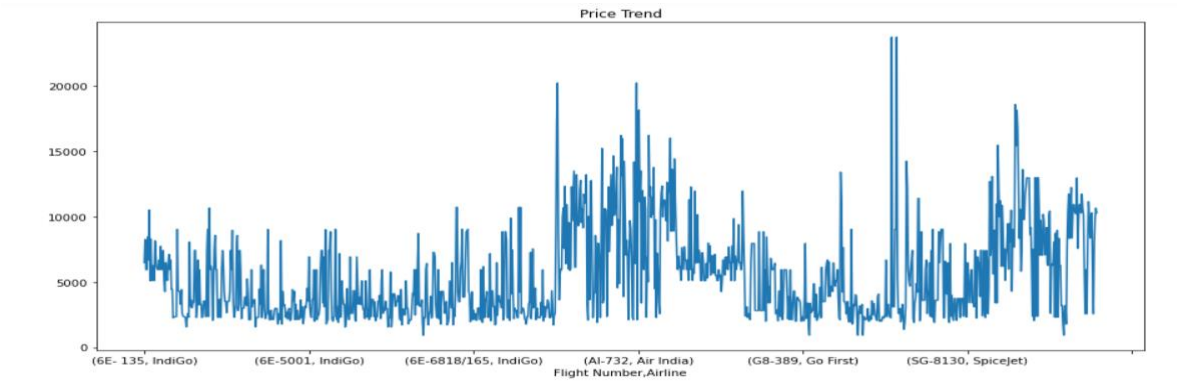
Following observations are made from graphs above:

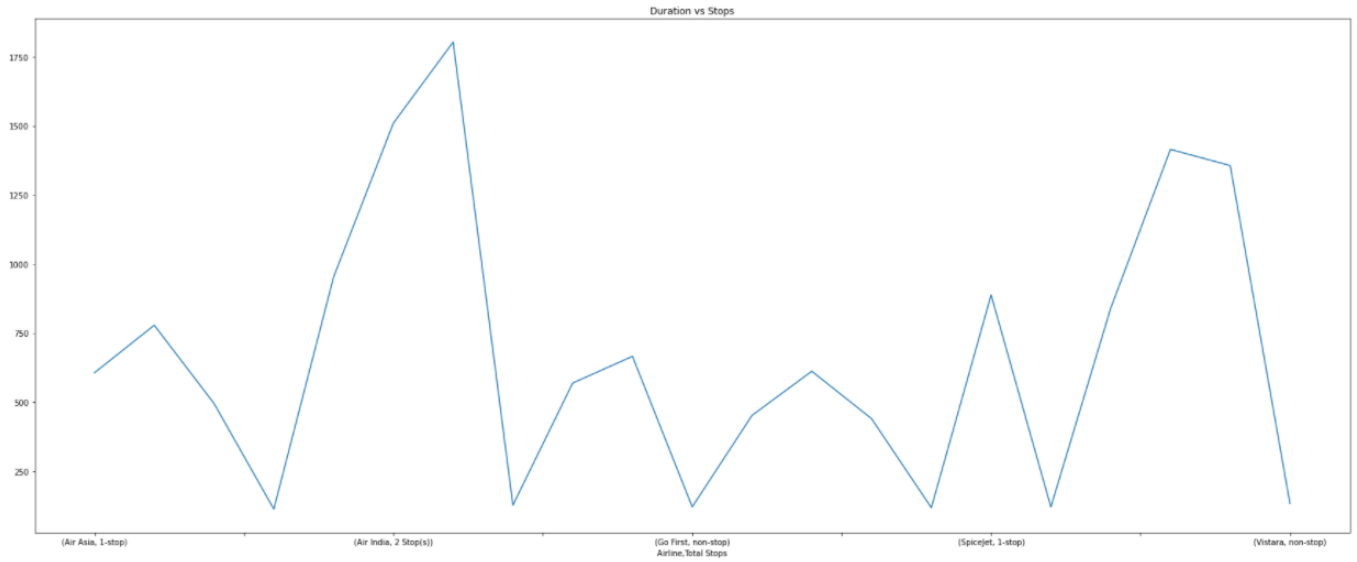
- From above graphs it can be observed that on an average, there is a steady decline in Flight price from January to April, with the prices being lowest in March.
- Flight Ticket prices are the highest on Thursdays, Mondays and during the weekend on an average.

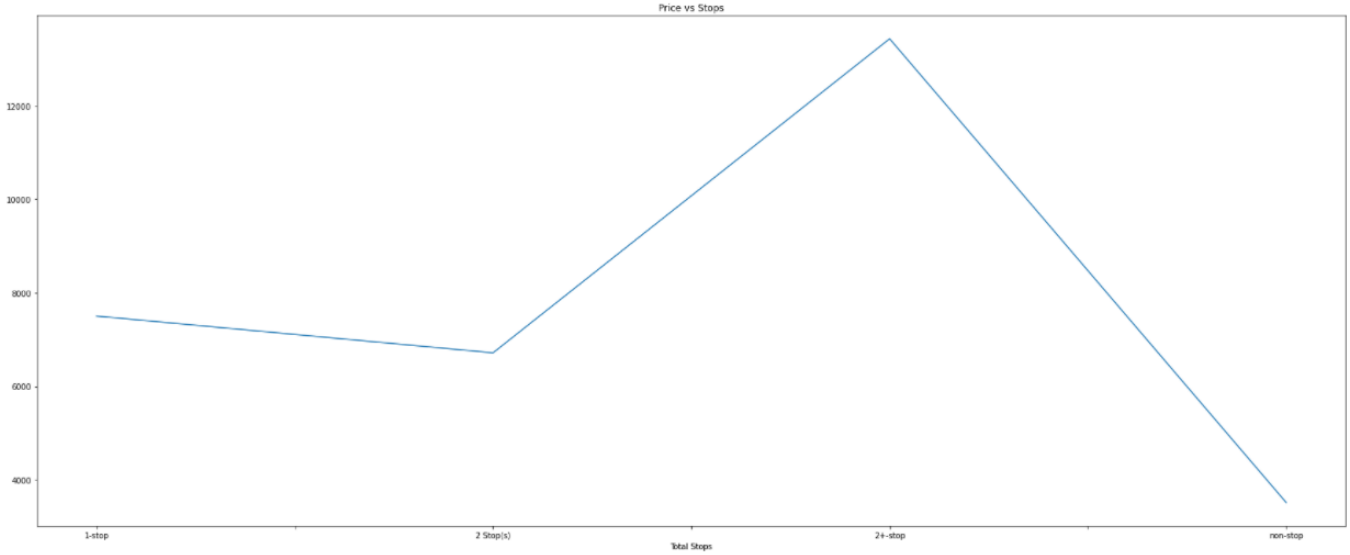
## Analyzing Relationship between Airlines, Flight Duration and Price

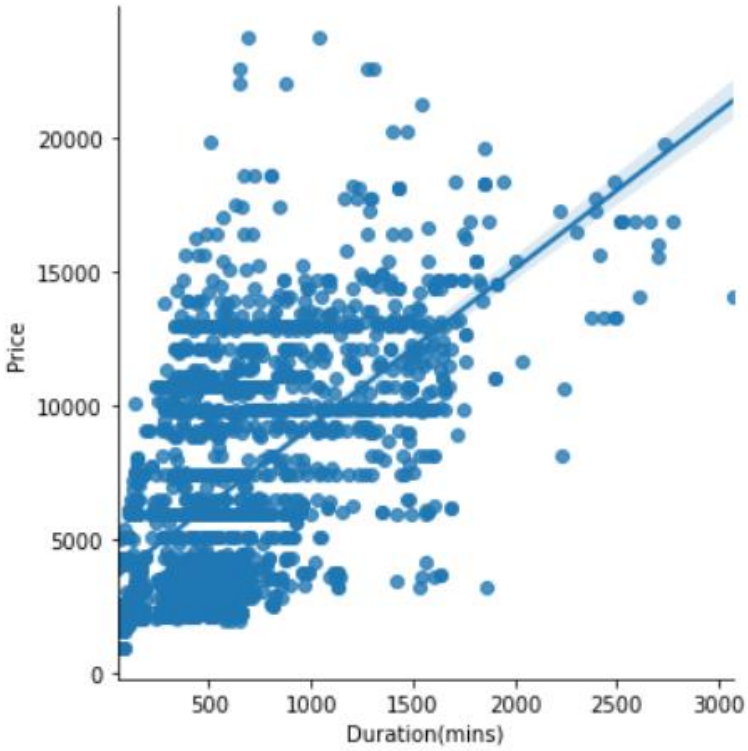



Price Trend
20000
15000
10000
5000
0
(6E- 135, IndiGo)
(6E-5001, IndiGo)
(6E-6818/165, IndiGo)
(AI-732, Air India)
(G8-389, Go First)
(SG-8130, SpiceJet)
Flight Number,Airline


Duration vs Stops
1750
1500
1250
1000
750
500
250
(Air Asia, 1-stop)
(Air India, 2 Stop(s))
(Go First, non-stop)
(Spicejet, 1-stop)
(Vistara, non-stop)
Airline,Total Stops

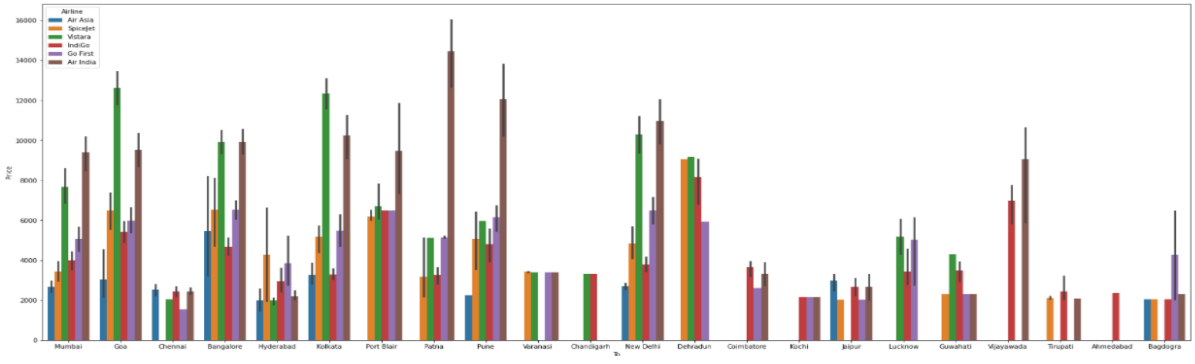

Price vs Stops
12000
10000
8000
6000
4000
1-stop
2 Stop(s)
2+-stop
non-stop
Total Stops




Following Observation is made from graphs above:

- Air Asia, IndiGo and SpiceJet offer air tickets at the most affordable prices on average, whereas Vistara, Air India are the most expensive on average. It can be observed that Number of Stops impact the travel time of Airlines.
- It can be observed that Number of Stops impact the Air Ticket Pricing of Airlines.
- There is a linear relationship between Price and flight duration.

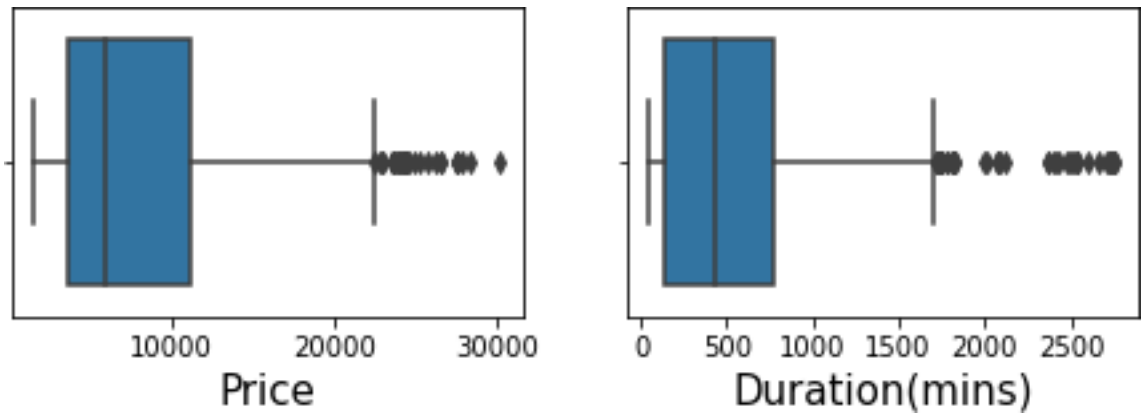
## Multivariate Analysis




Following Observations are made from graphs above:

- Goa, Mumbai, Pune, Bangalore, Kolkata, Port Blair, New Delhi are the most expensive destinations while, Kochi, Coimbatore, Jammu, Chennai, Hyderabad, Indore, Tirupati are the most affordable destinations
- Indigo, Air Asia and SpiceJet provide most affordable Air tickets to the destinations.

## Checking for Outliers




There are considerable outliers in the columns.

Outliers were Removed using Z score method which resulted in a total data loss of 1.00%, which is within acceptable range.

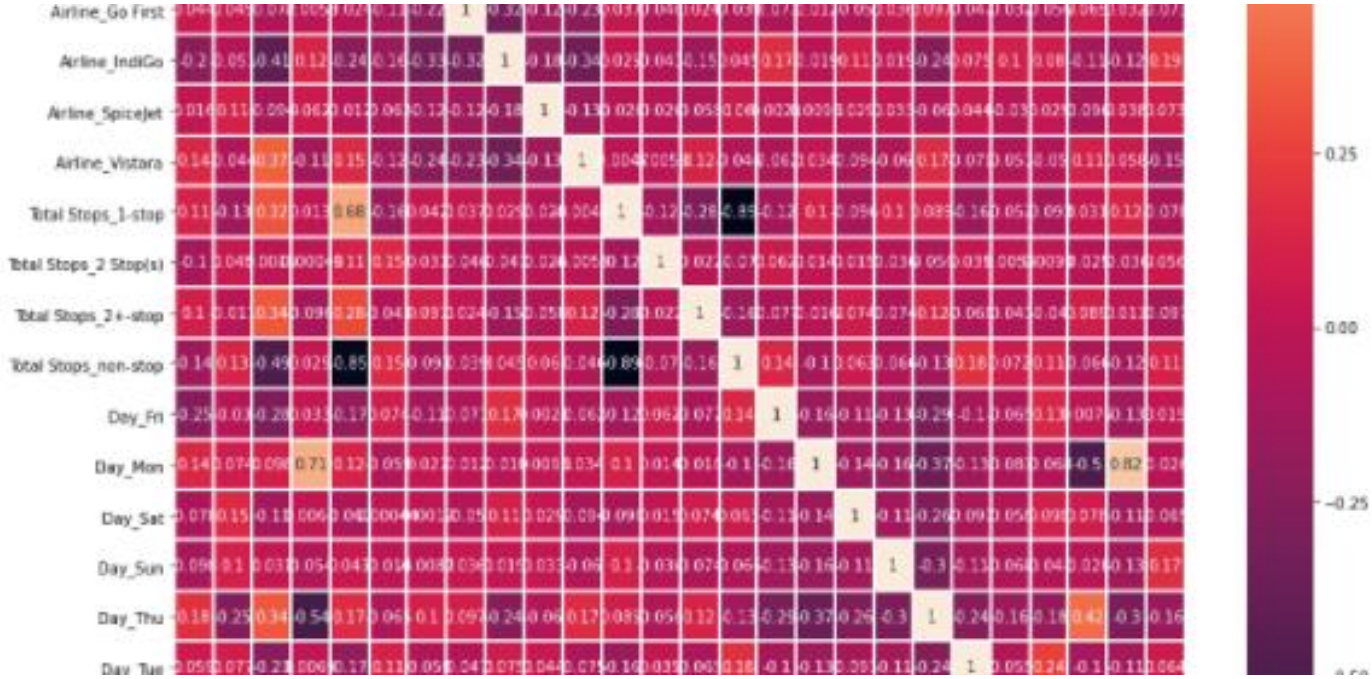
### Data Normalization

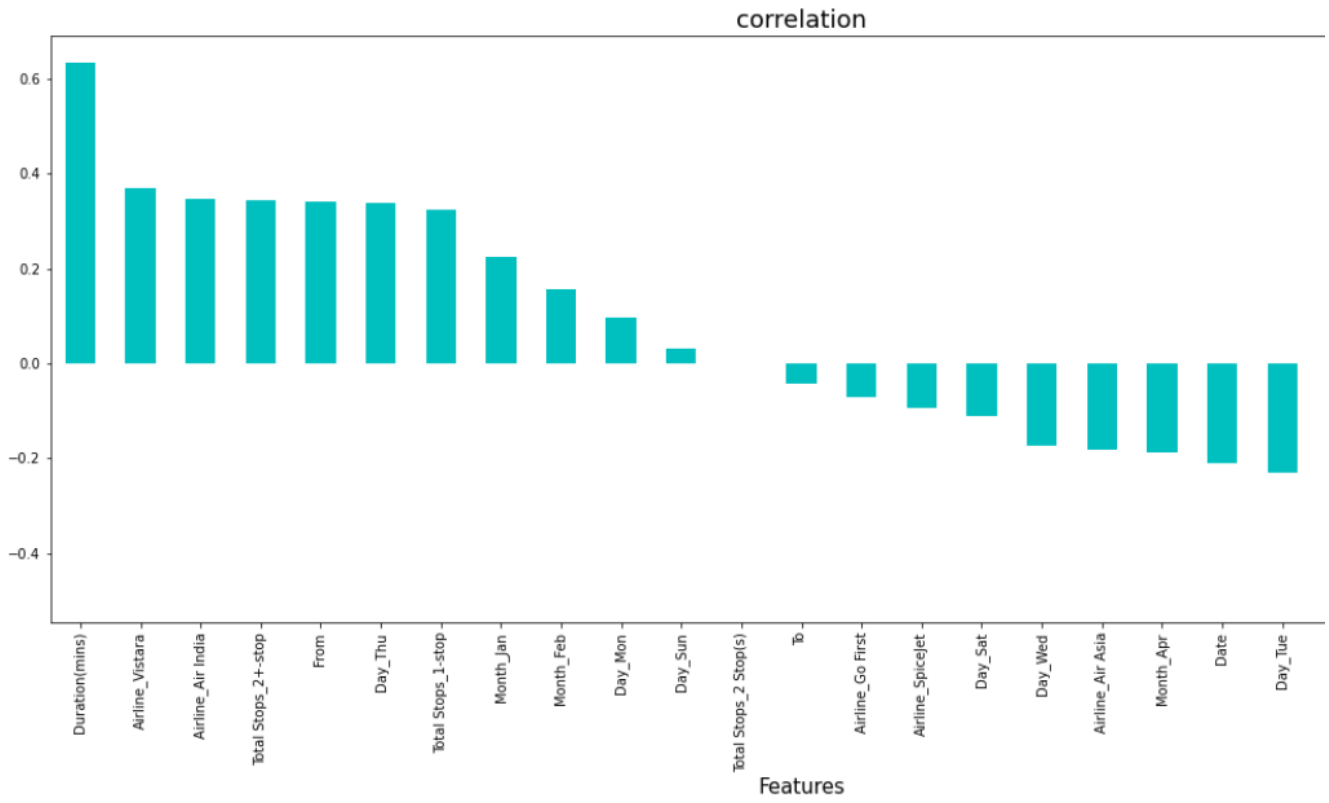
Data in Column ‘Duration(mins) was normalized using Power Transformer technique.

## Encoding Categorical Columns

Categorical Columns were encoded using Label Encoding technique and get_dummies () technique.

## Finding Correlation between Feature and Target columns






It is observed that Duration(mins), Airline_Vistara, Airline_Air India, Total Stops_2+stop, From and Day_Thur have the highest positive correlation with Price, while Total Stops_non-stop, Airline_IndiGo, Day_Fri, Month_Mar have the highest negative correlation with Price

# Model/s Development and Evaluation

## Feature Selection

Features were first checked for presence of multicollinearity and then based on respective ANOVA f-score values, the feature columns were
selected that would best predict the Target variable, to train and test machine learning models.

MultiCollinearity exists amongst many columns, Based on ANOVA F scores, columns scoring the lowest will be dropped.

Using SelectKBest and f_classif for measuring the respective ANOVA f-score values of the columns, the best features were selected. Using StandardScaler, the features were scaled by resizing the distribution values so that mean of the observed values in each feature column is 0 and standard deviation is 1. From sklearn.model_selection's train_test_split, the data was divided into train and test data. Training data comprised 75% of total data where as test data comprised 25% based on the best random state that would result in best model accuracy.

The model algorithms used were as follows:

- Ridge: Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. Since the features have multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be

far away from the actual values. Ridge shrinks the parameters. Therefore, it is used to prevent multicollinearity.

- DecisionTreeRegressor: Decision Tree solves the problem of machine learning by transforming the data into a tree representation. Each internal node of the tree representation denotes an attribute and each leaf node denotes a class label. A decision tree does not require normalization of data. A decision tree does not require normalization of data.
- XGBRegressor: XGBoost uses decision trees as base learners; combining many weak learners to make a strong learner. As a result it is referred to as an ensemble learning method since it uses the output of many models in the final prediction. It uses the power of parallel processing,supports regularization, and works well in small to medium dataset.
- RandomForestRegressor: A random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. A random forest produces good predictions that can be understood easily. It reduces overfitting and can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm.
- Support Vector Regressor: SVR works on the principle of SVM with few minor differences. Given data points, it tries to find the curve. But since it is a regression algorithm instead of using the curve as a decision boundary it uses the curve to find the match between the vector and position of the curve. Support Vectors helps in determining the closest match between the data points and the function which is used to represent them. SVR is robust to the outliers. SVR performs lower computation compared to other regression techniques.

# Regression Model Building

```
from sklearn.model_selection import train_test_split
```

```
from sklearn.metrics import r2_score
```

### Finding the Best Random State

```
from sklearn.ensemble import RandomForestRegressor
maxAcc = 0
maxRS=0
for i in range(1,100):
    x_train,x_test,y_train,y_test = train_test_split(scaled_x_be
    modRF =  RandomForestRegressor()
    modRF.fit(x_train,y_train)
    pred = modRF.predict(x_test)
    acc  = r2_score(y_test,pred)
    if acc>maxAcc:
        maxAcc=acc
        maxRS=i
print(f"Best Accuracy is: {maxAcc} on random_state: {maxRS}")
```

```
Best Accuracy is: 0.9008463834516044 on random_state: 24
```

Best random state was determined to be 24

## Training The Models

```
x_train,x_test,y_train,y_test = train_test_split(scaled_x_best,y,test_size = .25, random_state =24)
```

```
from sklearn.model_selection import GridSearchCV
from sklearn.linear_model import Ridge
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from xgboost import XGBRegressor
from sklearn.svm import SVR
```

```
from sklearn.metrics import r2_score,mean_squared_error
from sklearn.model_selection import ShuffleSplit,cross_val_score
```

```
rf = RandomForestRegressor()
dt = DecisionTreeRegressor()
xg = XGBRegressor()
SV= SVR()
r=Ridge()
```

## Analyzing Accuracy of The Models

Mean Squared Error and Root Mean Squared Error metrics were used to evaluate the Model performance. The advantage of MSE and RMSE being that it is easier to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced than smaller error, hence the model can now focus more on the larger errors.

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

Using cross-validation, there are high chances that we can detect over-fitting with ease. Model Cross Validation scores were then obtained for assessing how the statistical analysis generalises to an independent data set. The models were evaluated by training several models on subsets of the available input data and evaluating them on the complementary subset of the data.

```
models=[rf,dt,xg,SV,r]
```

```
for m in models:
    m.fit(x_train,y_train)
    m_pred=m.predict(x_test)
    R2=r2_score(y_test,m_pred)
    MSE=mean_squared_error(y_test,m_pred)
    RMSE=np.sqrt(MSE)
    CVS=cross_val_score(m,scaled_x_best,y,cv=ShuffleSplit(5)).mean()
    print(m,"results :")
    print("R2 Score :",R2)
    print("Cross Validation Score :",CVS)
    print("Mean Squared Error :",MSE)
    print("Root Mean Squared Error :",RMSE)
    print("\n")
```

RandomForestRegressor() results :
R2 Score : 0.9002717909950755
Cross Validation Score : 0.8894369514201905
Mean Squared Error : 1669352.4628051412
Root Mean Squared Error : 1292.0342343781535

DecisionTreeRegressor() results :
R2 Score : 0.7682890662292263
Cross Validation Score : 0.8428986854410832
Mean Squared Error : 3878613.902812792
Root Mean Squared Error : 1969.4196868145682

---

```
XGBRegressor(base_score=0.5, booster='gbtree', colsample_bylevel=1,
             colsample_bynode=1, colsample_bytree=1, enable_categorical=False,
             gamma=0, gpu_id=-1, importance_type=None,
             interaction_constraints='', learning_rate=0.300000012,
             max_delta_step=0, max_depth=6, min_child_weight=1, missing=nan,
             monotone_constraints='()', n_estimators=100, n_jobs=4,
             num_parallel_tree=1, predictor='auto', random_state=0, reg_alpha=0,
             reg_lambda=1, scale_pos_weight=1, subsample=1, tree_method='exact',
             validate_parameters=1, verbosity=None) results :
R2 Score : 0.8845522754052964
Cross Validation Score : 0.8745158019881878
Mean Squared Error : 1932481.7451389474
Root Mean Squared Error : 1390.137311613118


SVR() results :
R2 Score : 0.013957462725666003
Cross Validation Score : 0.01135375807756973
Mean Squared Error : 16505385.531873526
Root Mean Squared Error : 4062.6820613818068


Ridge() results :
R2 Score : 0.7015433734957995
Cross Validation Score : 0.6880095530325813
Mean Squared Error : 4995871.373471663
Root Mean Squared Error : 2235.1445978888396
```

---

# Interpretation of the Results

Based on comparing Accuracy Score results with Cross Validation results, it is determined that Random Forest Regressor is the best model. It also has the lowest Root Mean Squared Error score.

## Hyper Parameter Tuning

GridSearchCV was used for Hyper Parameter Tuning of the Random Forest Regressor model.

**Hyper Parameter Tuning**

```
from sklearn.model_selection import GridSearchCV
```

```
parameter = {'n_estimators':[30,60,80],'max_depth': [40,50,80],'min_samples_leaf':[5,10,20],'min_samples_split':[2,5,10],'criteri
```

```
GridCV = GridSearchCV(RandomForestRegressor(),parameter,cv=ShuffleSplit(5),n_jobs = -1,verbose = 1)
```

```
GridCV.fit(x_train,y_train)
```

```
Fitting 5 folds for each of 486 candidates, totalling 2430 fits
```

```
GridSearchCV(cv=ShuffleSplit(n_splits=5, random_state=None, test_size=None, train_size=None),
             estimator=RandomForestRegressor(), n_jobs=-1,
             param_grid={'criterion': ['mse', 'mae'], 'max_depth': [40, 50, 80],
                         'max_features': ['auto', 'sqrt', 'log2'],
                         'min_samples_leaf': [5, 10, 20],
                         'min_samples_split': [2, 5, 10],
                         'n_estimators': [30, 60, 80]},
             verbose=1)
```

```
GridCV.best_params_
```

```
{'criterion': 'mse',
 'max_depth': 40,
 'max_features': 'auto',
 'min_samples_leaf': 5,
 'min_samples_split': 2,
 'n_estimators': 30}
```

```
Best_mod = RandomForestRegressor(n_estimators = 30,criterion = 'mse', max_depth= 40, max_features = 'auto',min_samples_leaf = 5,

Best_mod.fit(x_train,y_train)
```

```
RandomForestRegressor(max_depth=40, min_samples_leaf=5, n_estimators=30)
```

```
rfpred = Best_mod.predict(x_test)
acc = r2_score(y_test,rfpred)
print(acc*100)
```

```
87.15436311972498
```

Based on the input parameter values and after fitting the train datasets The Random Forest Regressor model was further tuned based on the parameter values yielded from GridsearchCV. The Random Forest Regressor model displayed an accuracy of 87.15%

This model was then tested using a scaled Test Dataset. The model performed with good amount of accuracy.

```
Prediction_accuracy = pd.DataFrame({'Predictions': mod.predict(scaled_x_best), 'Actual Values': y})
Prediction_accuracy.head(30)
```

| | Predictions | Actual Values |
|---|---|---|
| 0 | 2466.304560 | 2395 |
| 1 | 2780.334913 | 2395 |
| 2 | 2812.148971 | 2407 |
| 3 | 3065.649254 | 2407 |
| 4 | 2489.049192 | 2410 |
| 5 | 2448.446162 | 2410 |
| 6 | 2448.446162 | 2410 |
| 7 | 2448.446162 | 2410 |
| 8 | 2448.446162 | 2410 |
| 9 | 2448.446162 | 2410 |
| 10 | 3464.614197 | 3288 |
| 11 | 3390.629409 | 3626 |
| 12 | 3419.819193 | 3626 |
| 13 | 3493.567357 | 3626 |
| 14 | 3493.567357 | 3626 |
| 15 | 4262.644013 | 3882 |

# SUMMARY

In summary, based on the visualizations of the feature-column relations hips, it is determined that, Features like Source, month, Duration, Total Stops, Airline, Date are some of the most important features to predict the label values. Random Forest Regressor Performed the best out of all the models that were tested. It also worked well with the outlier handling.

# CONCLUSION

## Key Findings and Conclusions of the Study and LearningOutcomes with respect to Data Science

Based on the in-depth analysis of the Flight Price Prediction Project, The Exploratory analysis of the datasets, and the analysis of the Outputs of the models the following observations are made:

• Air Fare attributes like Date, Month, Duration, Total Stops etc play a bigrole in influencing the used Flight price.

• Airline Brand also has a very important role in determining the used Flight Ticket price.

• Various plots like Barplots, Countplots and Lineplots helped in visualising the Feature-label relationships which corroborate the importance of Air Fare features and attributes for estimating Flight Ticket Prices.

• Due to the Training dataset being very small, only very small amount of the outliers was removed to ensure proper training of the models.

• Therefore, Random Forest Regressor, which uses averaging to improve the predictive accuracy and controls over-fitting. performed well despite having to work on small dataset and produced good predictions that can be understood easily.

## Learning Outcomes of the Study in respect of DataScience

Data cleaning was a very important step in removing plenty of anomalous data from the huge dataset that was provided. Visualizing data helped identify outliers and the relationships between target and feature columns as well as analyzing the strength of correlation that exists between them.

## Limitations of this work and Scope for Future Work

A small dataset to work with posed a challenge in building highly accurate models. This project also relied heavily on historical data and was unable to account for various other factors that influence demand and ticket pricing like pandemic status affecting demand, government regulations on air travel, shifting in routes, weather conditions, etc.

Most airline companies also do not publicly make available their ticket pricing strategies, which makes gathering price and air fare related data sets using web scraping the only means to build a dataset for building predicting models.

Availability of more features and a larger dataset would help build better models.