

*FLIGHT PRICE PREDICTION*

Submitted by:

Vivek kumar singh

**ACKNOWLEDGMENT**

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). The 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. The framework achieves a high

prediction accuracy with 0.869 adjusted R squared score on

the testing dataset.

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) make sit 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). The 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. The framework achieves a high prediction accuracy with 0.88 adjusted R squared score on the testing dataset.

**INTRODUCTION**

* Business Problem Framing

With the explosive growth of the Internet and ecommerce, air passengers nowadays can check airfare and availability of any airlines around the world easily. When satisfying with an airfare, these customers can purchase their desired tickets online through official airline or agent websites. To help the customers to buy the most inexpensive airfare, there have been a number of prediction models to predict the airfare prices. Social media today is an integral part of people’s daily routines and the livelihood of some. As a result, it is abundant in user opinions. The analysis of brand specific opinions can inform companies on the level of satisfaction within consumers. Airline Ticket prices change very dynamically and significantly for the same flight day by day. It is very difficult a customer to purchase an air ticket in the lowest price, since the price changes dynamically. Sentimate analysis (Reviews) also helps to customers in decision making on selection of best airlines on analyzing other customer’s opinion.The purpose of this study is to better analyze the features that affect airfare and to develop and tune models to predict the airfare well in advance.

* Conceptual Background of the Domain Problem

Using 27 million flight bookings for 2 years from a major international airline company, we built a Next Likely Destination model to ascertain customers’ next flight booking. The resulting model achieves an 88% predictive accuracy using historical data. A unique aspect of the model is the incorporation of self-competence, where the model defers when it cannot reasonably make a recommendation. We then compare the performance of the Next Likely Destination model in a real-life consumer study with 35,000 actual airline customers. In the user study, the model obtains a 51% predictive accuracy. What happened? The Individual Behaviour Framework theory provides insights into possibly explaining this inconsistency in evaluation outcomes. Research results indicate that algorithmic approaches in competitive industries must account for shifting customer preferences, changes to the travel environment, and confounding business effects rather than relying solely on historical data.

* Review of Literature

To begin, we need information on aircraft business and mass transit in order to develop the airline ticket pricing model just at

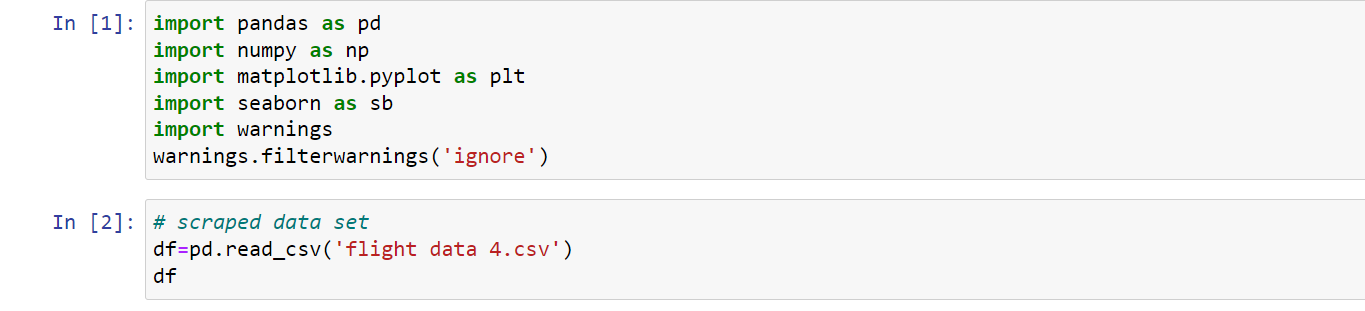
market level. As a result, we have one datasets. Further we divided it into training and testing . The dataset contains 1505 items with parameters such as airline, source, destination, route, time of departure, estimated time of arrival, length, maximum stop, and price. The testing data contains the 0.25% of total dataset with the following attributes: Airlines, Departure city ,Arrival city, Departure time ,arrival time , stops ,total time duration and target variable is price.

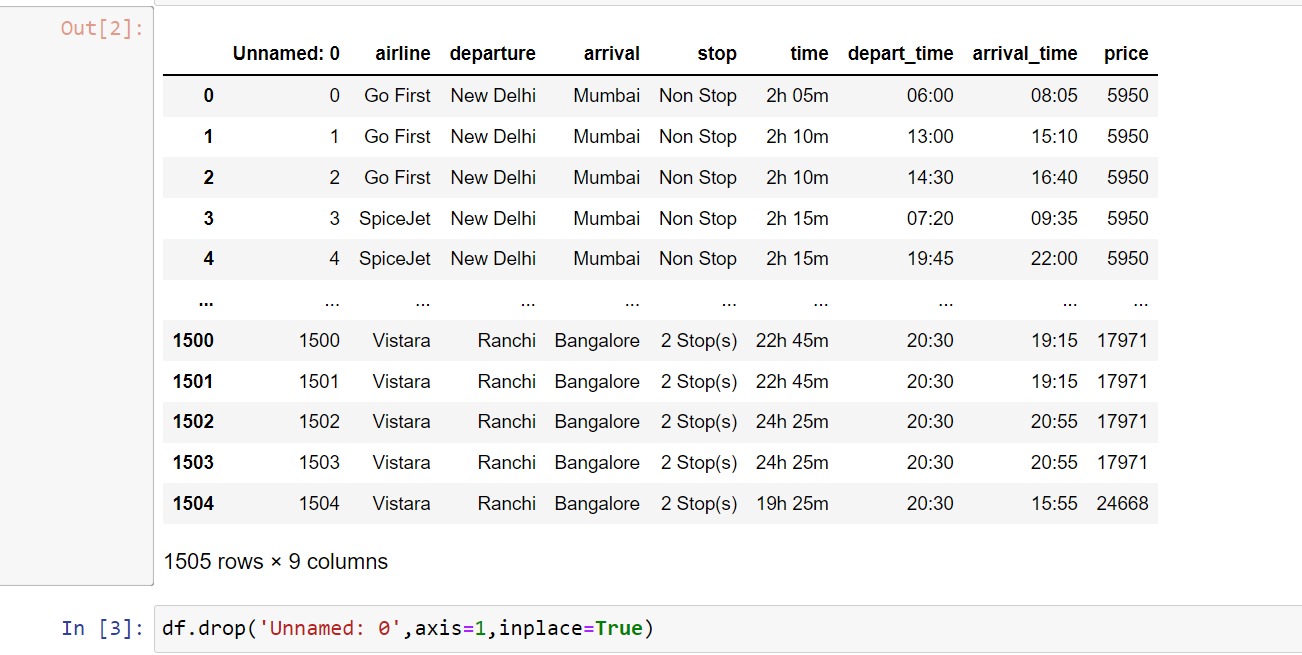
**Analytical Problem Framing**

* Mathematical/ Analytical Modeling of the Problem
* I am working with the used flight price dataset that contains various features and information about the flight business and its ticket price. Using the scraped data in form of *‘read\_csv’* function provided by the Pandas package, which can import the data into our python environment.
* After importing the data, I have used the ‘head’ function to get a glimpse of our dataset. Data has been scraped from website Yatra.com
* Hence the reason for continuous research in this sector. This project simply examines a dataset, which consists 1500+ observations and 10 features that contribute to the flight ticket price . Dataset was cleaned and transformed and some explorations were done on it to answer some basic questions that anybody would like to ask about project.
* Our dataset is ready in the right form with the right variables to be used in the algorithms, which results in improved model accuracy. Different ensemble algorithms were used on the dataset in this project. The overall result of this project shows that the most important variables that determine the price of airline tickets.
* **Keywords**: price, Analysis, encoding, Ensemble Algorithms and Feature Engineering

.

* **Data Sources and their formats**
* The dataset has been scraped from website Yatra.com this data is only for academic use, not for any commercial.
* The dataset describe data related to Flight Business with 1500+ records.
* The dataset is in csv. Format which contains train and test data.
* This dataset is to use simply examines data, which consists 1500+ observations and 9 features for model predication.
* The dataset is in both numerical as well as categorical data.





Here, I am dropping the unwanted columns which is not relevant.

* **Data Pre-processing Done**

In order to get a better understanding of the data, we plotted a distribution plot of the data. We noticed that the dataset had not many outliers. So, we intervein into that we keep the data as it is.

There are more categorical columns in data set. We did some of presentation.

Visualization and model building is done here. We also used feature selection technique in the project. Variance inflation factor is also used in model for multicollinearity problem.

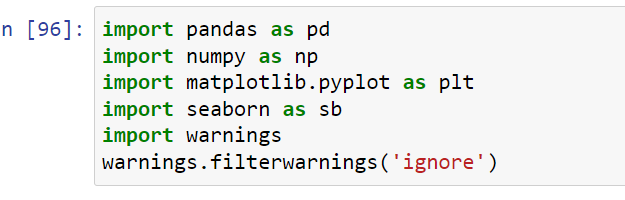
And then we scale data and data is ready for training.

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

The problem of this project is regression not classification so keeping the assumption should done under scale.

* Hardware and Software Requirements and Tools Used

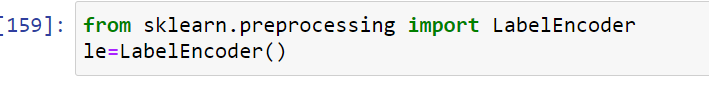
Tools imported for project are :-

1. 

Pandas is used for data import in data fame, data cleaning, data visualization, data normalization, data fill, merge and join.

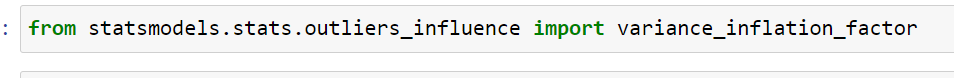
Numpy is used for number in python.

Matplotlib and seaborn is for visualization .

2. 

Label encoder is used to convert objective categorical columns into numeric value.

3.



Variance inflation factor is used to check multicollinearity problem

4.

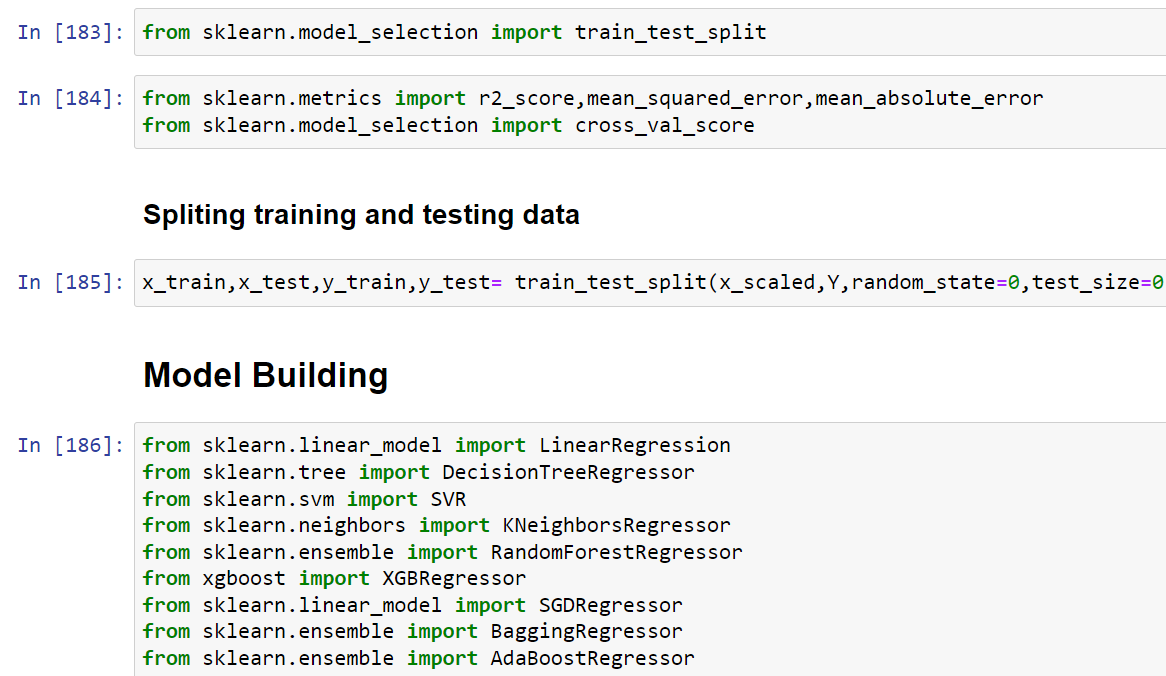


This is used for both feature selection as well as model training

5.



This package is used for scaling the data.

6. 

Training, splitting data into train test and importing important Algorithum for prediction.

**Model/s Development and Evaluation**

* Identification of possible problem-solving approaches (methods)

First part of problem solving is to scrap data from www.yatra.com website which we already done. Next part of problem solving is building machine learning model to predict flight price. This problem can be solve using regression-based machine learning algorithm like linear regression. For that purpose, first task is to convert categorical variable into numerical features. Once data encoding is done then data is scaled using standard scalar. Final model is built over this scaled data. For building ML model before implementing regression algorithm, data is split in training & test data using train\_test\_split from model\_selection module of sklearn library. After that model is train with various regression algorithm and 5-fold cross validation is performed. Further Hyperparameter tuning performed to build more accurate model out of best model.

* Testing of Identified Approaches (Algorithms)

Web Scraping Strategy employed in this project as follow:

1. Selenium will be used for web scraping data from www.yatra.com

2. Flights on route of New Delhi to Mumbai in duration of 23 Jan 2022 to 4 Feb 2022.

3. Data is scrap in three parts: ▪ Economy class flight price extraction ▪ Business class flight price extraction ▪ Premium Economy class price extraction

4. Selecting features to be scrap from website.

5. In next part web scraping code executed for above mention details.

6. Exporting final data in csv file.

The different regression algorithm used in this project to build ML model are as below:

❖ Linear Regression

❖ Random Forest Regressor

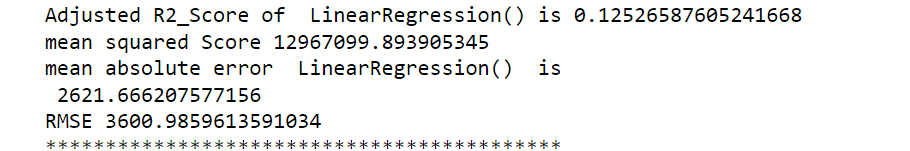
❖ Decision Tree Regressor

❖ XGB Regressor

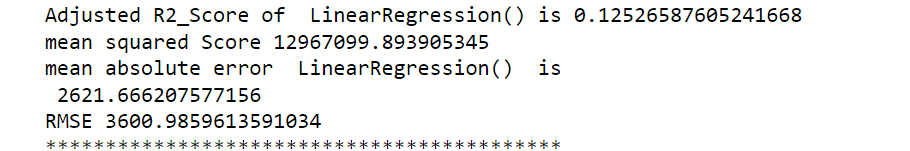
❖ Extra Tree Regressor

* Bagging Regressor
* Gradient Boosting Regressor
* Extra Tree Regressor and etc.
* Run and evaluate selected models

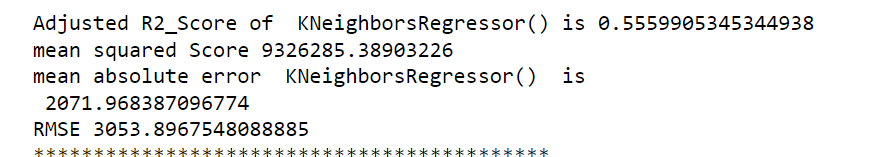
Linear regression

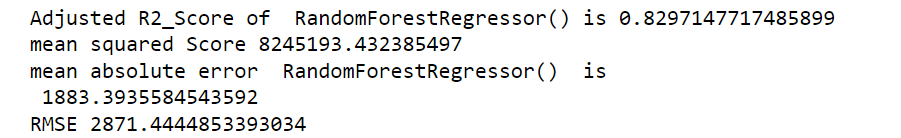


Decision tree regressor

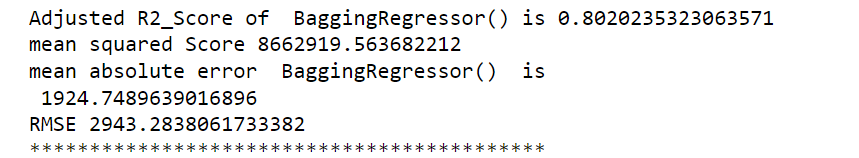


K-neighbours regressor



Random forest 

Bagging regressor



* Key Metrics for success in solving problem under consideration

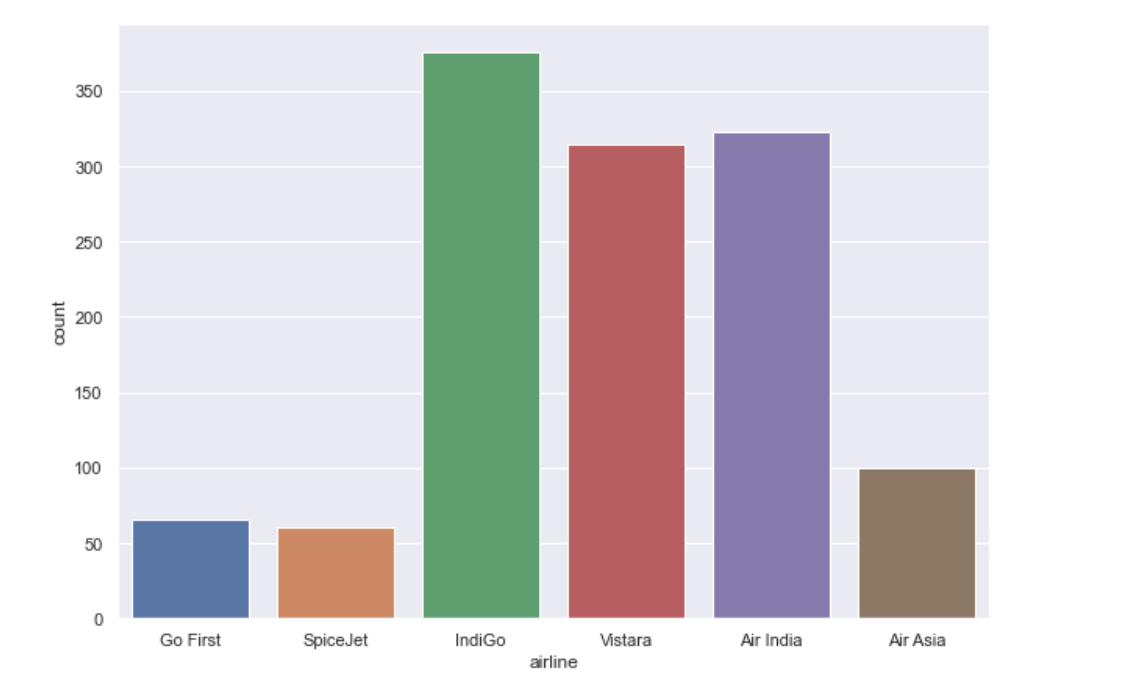
Following metrics used for evaluation:

1. Mean absolute error which gives magnitude of difference between the prediction of an observation and the true value of that observation.

2. Root mean square error is one of the most commonly used measures for evaluating the quality of predictions.

3. Adjusted R2 score which tells us how accurate our model predict result.

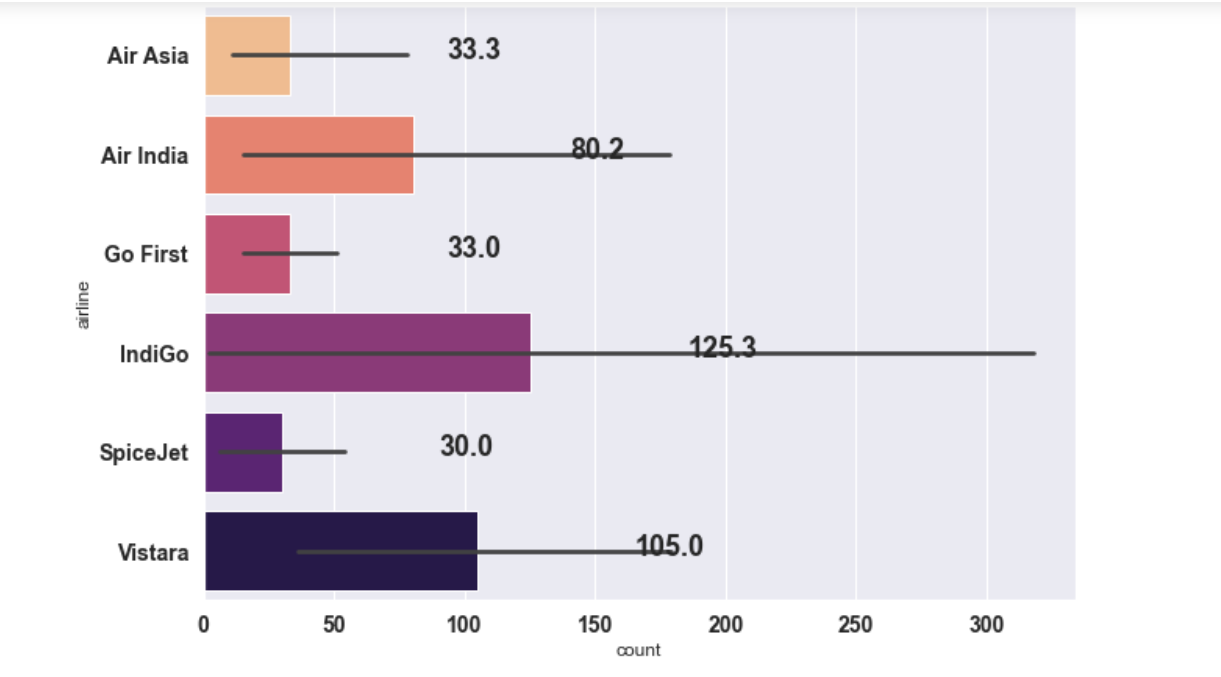
* Visualizations



Here we visualize airline company

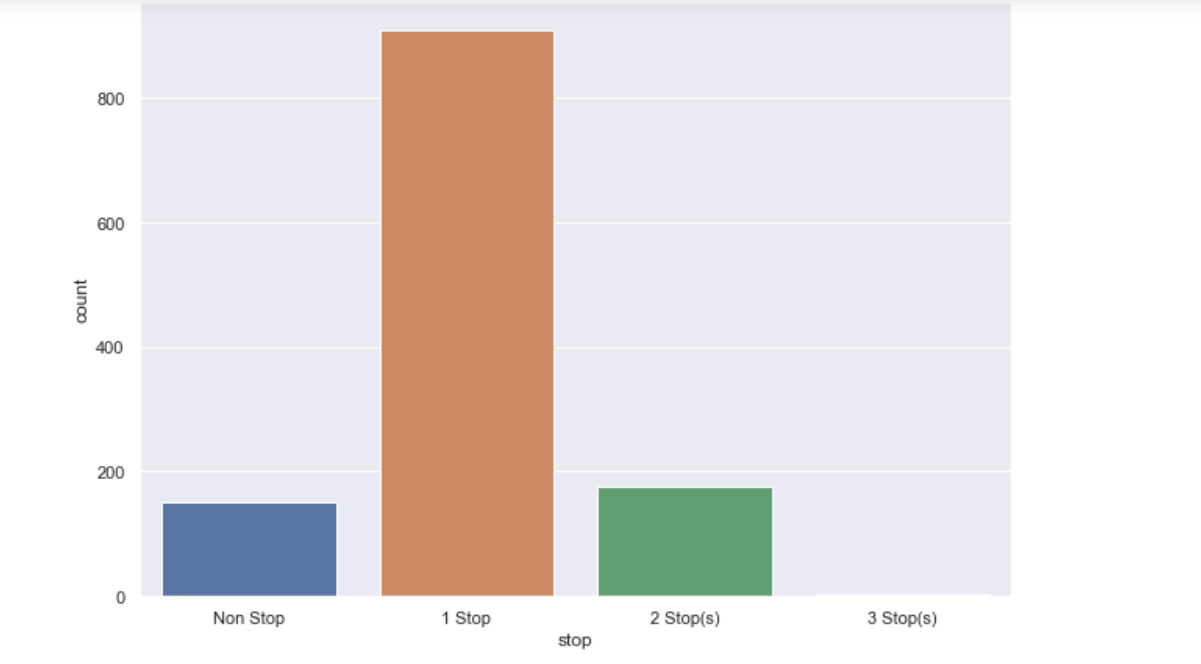
1.indigo is highest among all.

2.Air India and Vistra are second, although there owner are same.



As per count we see indigo has 125 data in data set, which is highest.

Vistara and Air Asia are the second in data set.

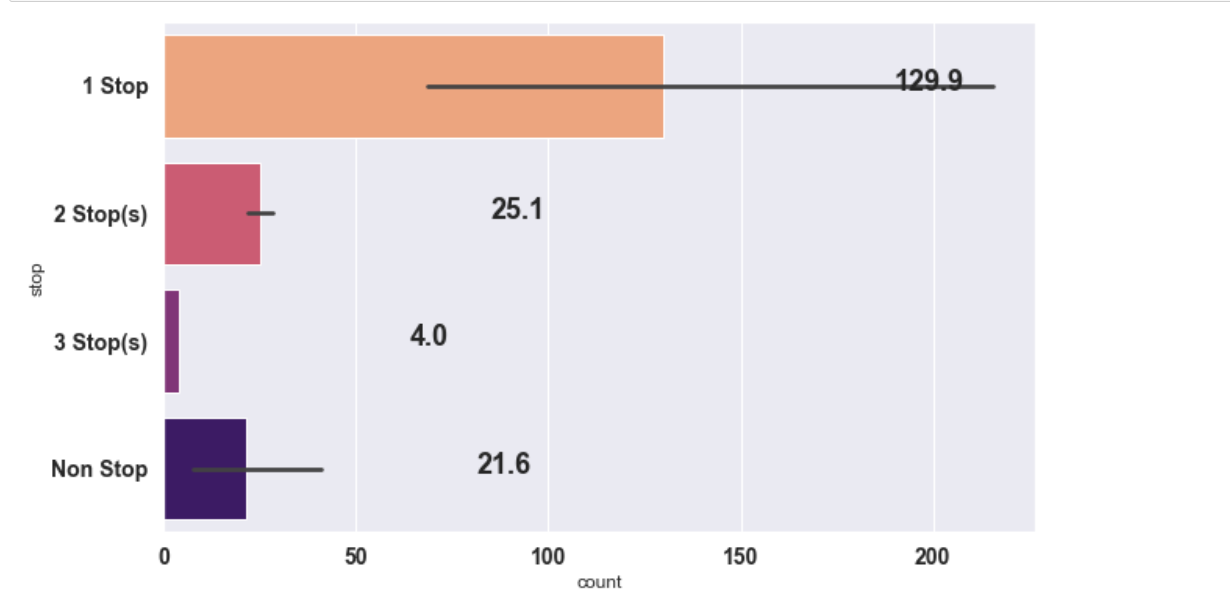


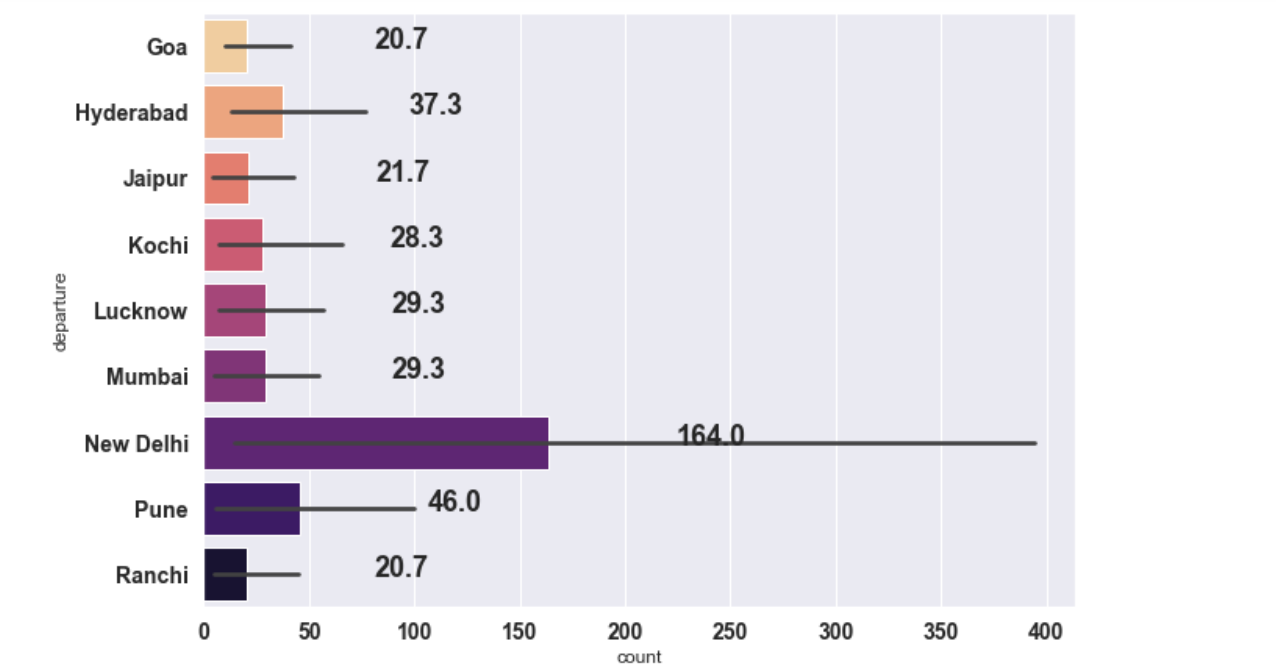
Flights make stops at time of journey: -

1.single stops are more among all other.

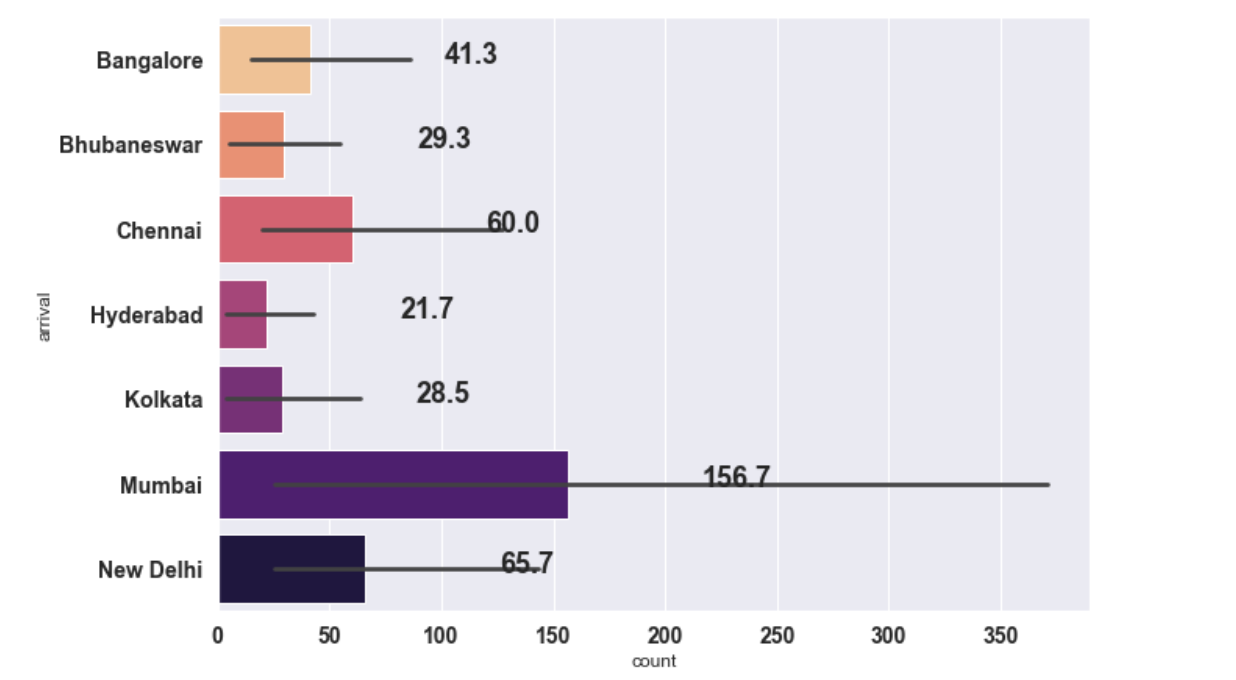
2.non stops and 2 stops has some of data

3. 3-stops has very less data in data set , may be 2 -3 data in whole dataset.

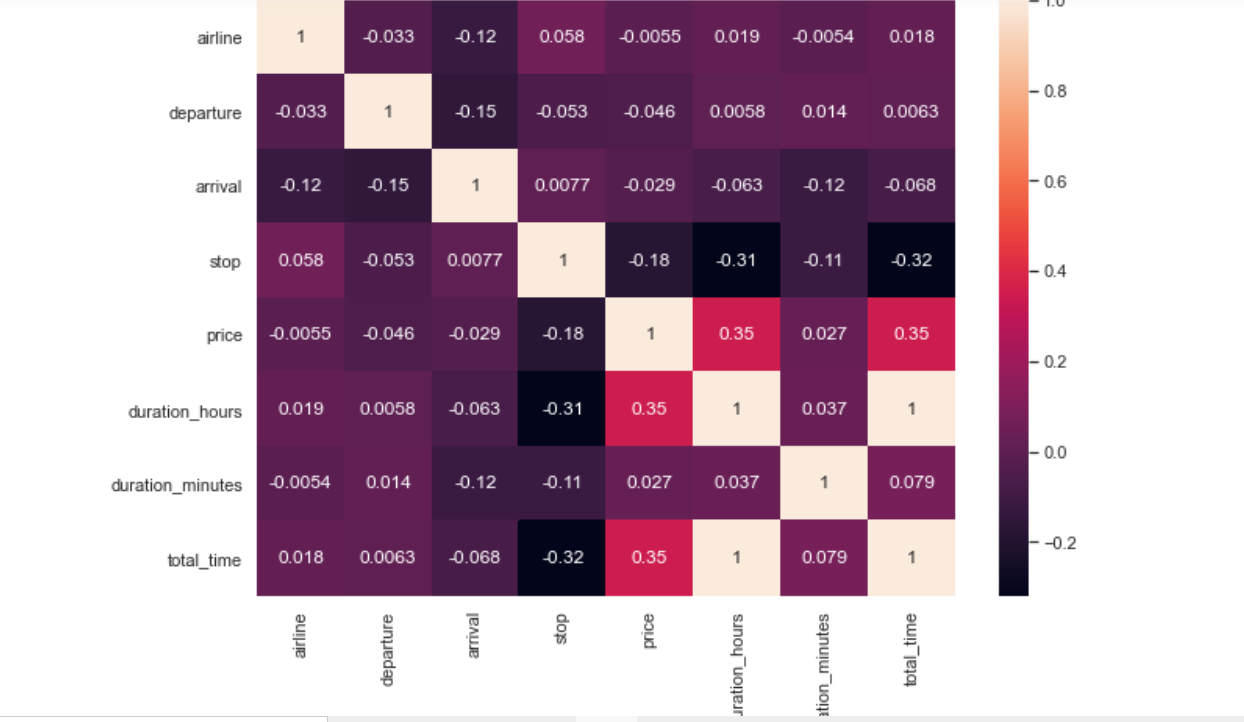




This is departure of flight as per data we see

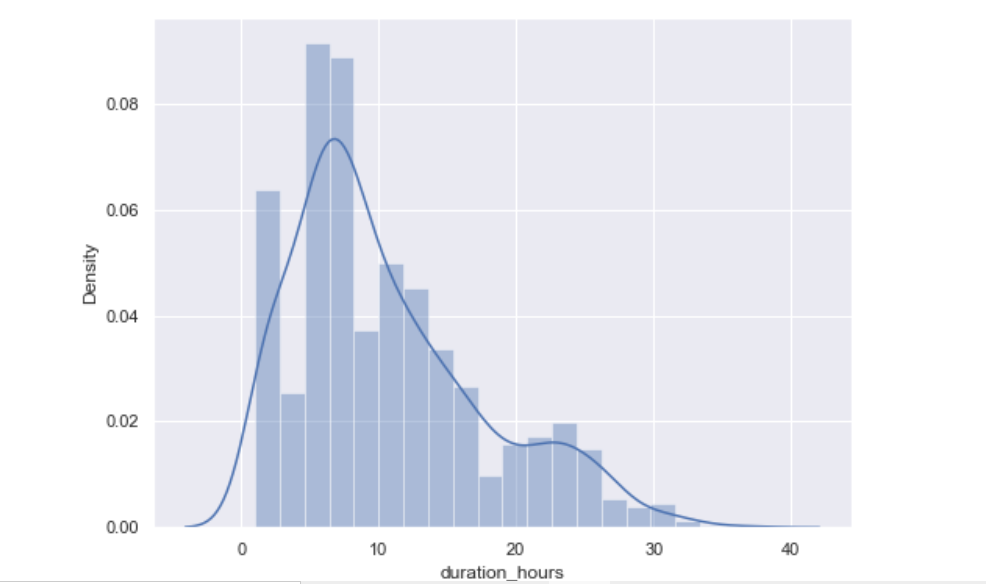


This visualization shows the arrival of city as per data



This is the correlation graph we see here it shows the relation

Between the features and between features and target variable



Some more visualization is done for continuous feature.

We are keeping the data as it is, we see that data is very little skewed. So, we have no need for extra operations.

**CONCLUSION**

* Learning Outcomes of the Study in respect of Data Science

Decision tree Regressor giving us maximum adjusted R2 Score, so Decision tree Regressor is selected as best model.

Adjusted r2score is 0.88

* mean squared Score 12614347.905085126
* mean absolute error DecisionTreeRegressor is 2154.07446
* mean absolute error DecisionTreeRegressor is 2154.074462365
* Limitations of this work and Scope for Future Work

In this study we focus on flights on route of New Delhi to Mumbai, more route can incorporate in this project to extend it beyond present investigation

* This investigation focus on short timeframe (14 days prior flights take off) which can be extended variation over larger period.
* Time series analysis can be performed over this model.