FLIGHT PRICE PREDICTION

Problem Statement :-

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.

Data Collection and cleaning

Collection Of Data:-

We have collected data from different websites and then we have put them into test and train data,

So that we can predict and match the data more accurately and comfortably.

As we load the training data we can visualize that there are 11 column s taken into consideration for almost 11000 data rows.

The columns are-

Airline, Date of Journey, Source, Destination, Route, Departure Time,

Arrival Time, Duration, Total Stops, Additional Info, Price

Cleaning the Data:-

Null Values-We don’t find as such null values to show ,we have dropped the minimal null values.

Data type Evaluation-

We can see that the **‘Date of Journey’ , ’Departure Time’ , ’Arrival Time’ , ’Duration’** column is object data type, and these are number or we can say time related columns. so we have to convert it to get the prediction right.

So as we can see that these columns have something common they all are time related columns so we will split the columns into there respective group of time ,i.e. hours, minutes , seconds.

After doing the above we will have sum new columns and some old columns, We will drop the old columns as the new columns are the output of these old columns , so if we keep old columns then it will be a duplicate entry and it will not help us to reach the desired result.

After doing all that we will have these columns:-

Date of journey- day_of_depature,month_of_depature.

Departure Time- dep_hour,dep_minute.

Arrival Time- arrival_hour,arrival_minute.

Duration- Duration_hour, Duration_min.

We can see that **‘Route’ and ‘Total Stops’** columns have common relation between them which shows that if we take both into consideration it should not give good result ,so we are dropping **Route.**

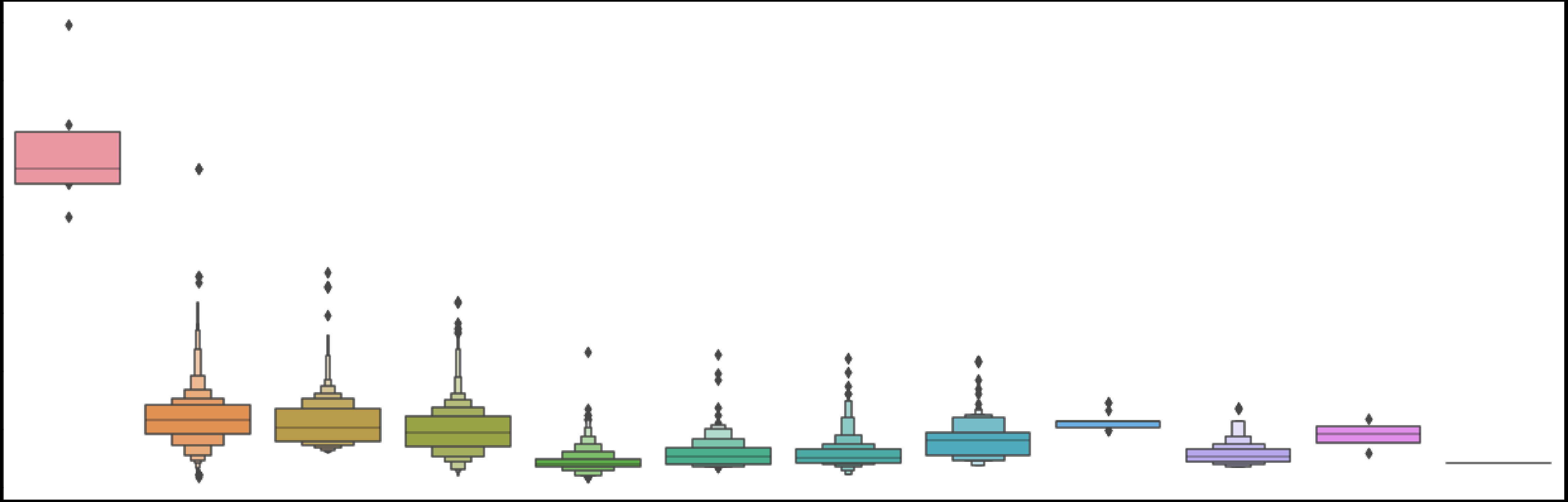
We can also see in **‘Additional Info’** most of the data shows no info ,which shows us that these data’s are some what like null data so we think we should drop this column.

Exploratory Data Analysis

In this data set we can see some relation with feature and target columns

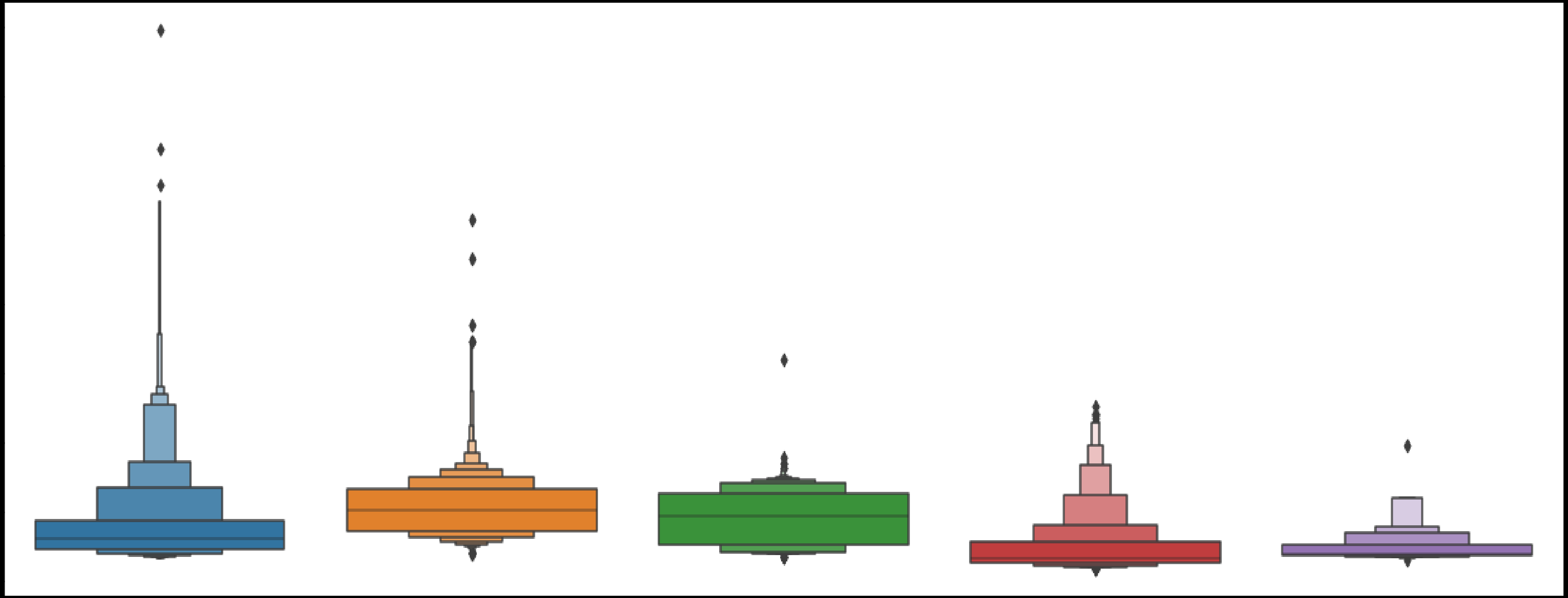
Lets explore that:-

- We have visualized the relation between **Airline companies** and **Price**



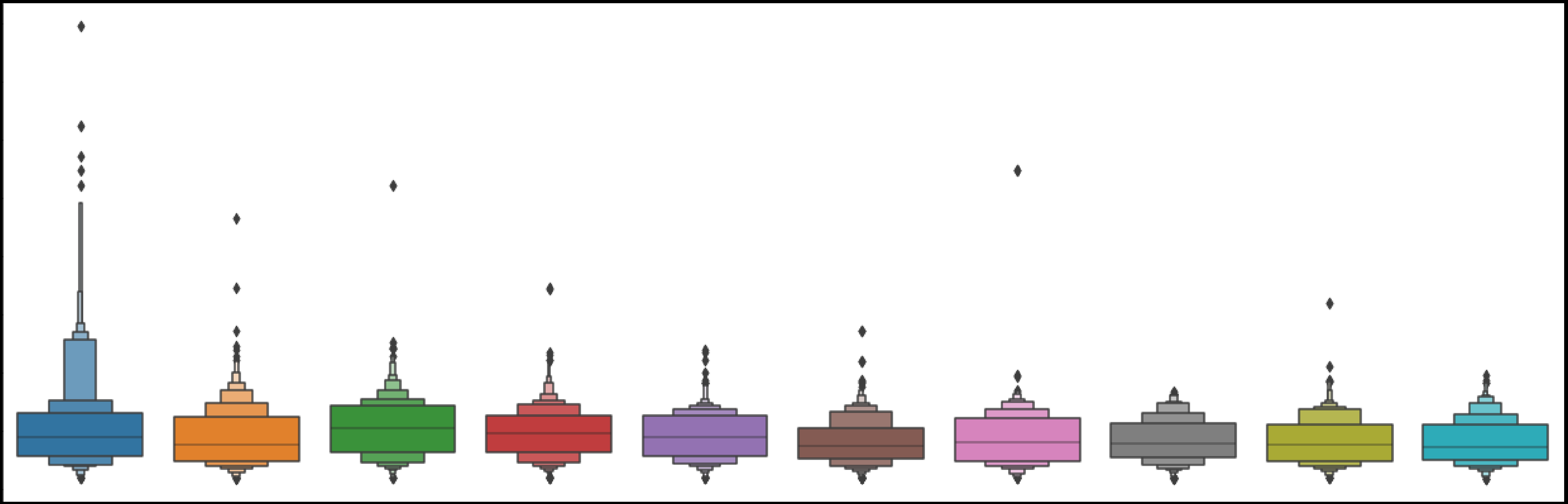
we can see that **jet airways** are having more price than others the median is some how similar.

- We have visualized the relation between **Source** and **Price**



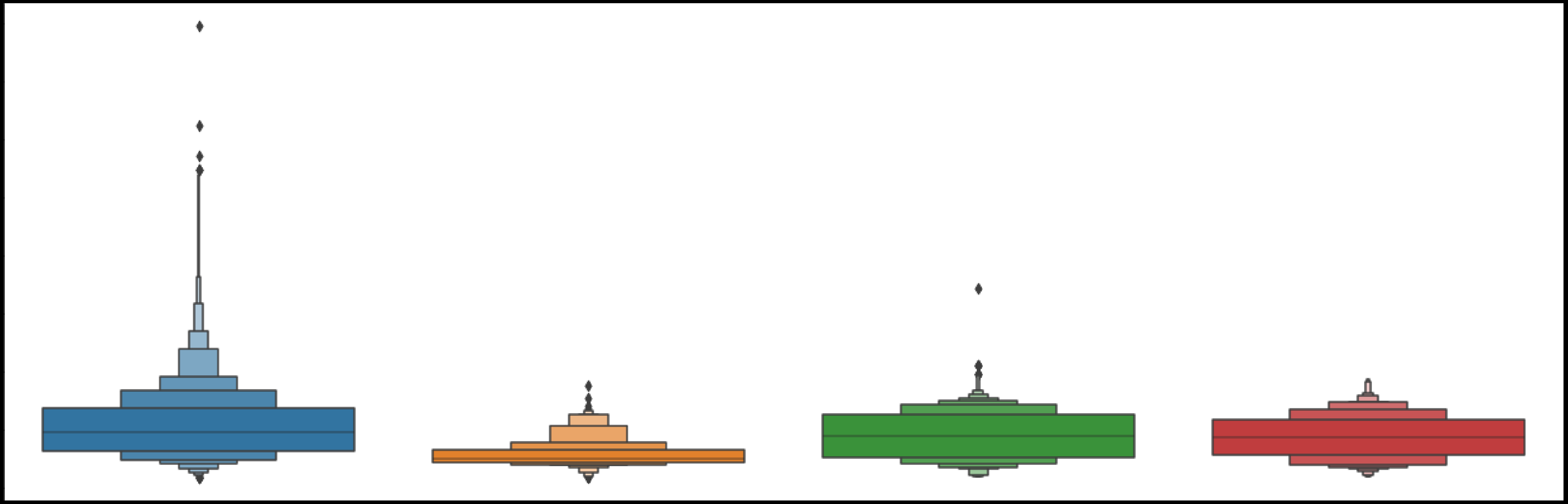
we can see **Bangalore** has the most expensive tickets median is different but it shows fluctuating median

- We have visualized the relation between **day_of_depature** and **Price**



we can see that 1st day of month shows the most price as we can see that this date shows most demand of tickets

- We have visualized the relation between **month_of_depature** and **Price**



We can see 3rd month has the most price people explore more in 3rd month

Data Pre-processing

As we can see that all the columns are categorical except target column so we are going use the encoder to encode the remaining object columns.

- As Airline is Nominal Categorical data we will perform OneHotEncoding.
- As Source is Nominal Categorical data we will perform OneHotEncoding.
- As Destination is Nominal Categorical data we will perform OneHotEncoding.
- We have used label encoder for remaining object columns
- So , after encoding **we have these columns**:-

'Total_Stops', 'Price', 'day_of_depature', 'month_of_depature', 'dep_hour', 'dep_minute’,
'arrival_hour', 'arrival_minute', 'Duration_hour', 'Duration_min', 'Airline_Air India’,
'Airline_GoAir','Airline_IndiGo', 'Airline_Jet Airways', 'Airline_Jet Airways Business’,
'Airline_Multiple carriers’,'Airline_Multiple carriers Premium economy', 'Airline_SpiceJet',
'Airline_Trujet', 'Airline_Vistara', 'Airline_Vistara Premium economy','Source_Chennai’,
'Source_Delhi', 'Source_Kolkata', 'Source_Mumbai','Destination_Cochin', 'Destination_Delhi’,
'Destination_Hyderabad','Destination_Kolkata', 'Destination_New Delhi’

Correlation

we can see that hour duration has the most and jet airways also has some positive relation , total stop have negative relation.

Model Building

- We will not perform any skewness removal techniques as all the features are categorical columns and skewness removal will not effect the prediction in a positive way.
- Now we will Standardize the data so that the scale of the feature remain at a standard rate with **Standard Scaler.**
- We have few model building techniques , **linear Regression model,**

Random Rainforest Model, KNN model, Decision tree Regressor.

- We have also done **regularization and hyper parameter tunings**.
- We can see that Random rainforest gives the best accuracy.

Conclusion

From the above prediction and research we came to the conclusion:-

- we can see that hour duration has the most and jet airways also has some positive relation , total stop have negative relation , which indicate date the more the stop the less the price and vice – versa .

- We can see Bangalore has the most expensive tickets which indicate that the price of ticket also vary with the type of city,
- We can also see tickets have relation with early days of the month which indicate salaried person use flights in these days more, as the salaried person are more willing to pay on early days of the month mostly.