FLIP ROBO

NAME OF THE PROJECT

Flight Price Prediction

Submitted by:

Akif Perwez

ACKNOWLEDGMENT

I would like to thank our SME (Sapna Verma) for his expert advice and encouragement throughout this project.

INTRODUCTION

- Business Problem Framing

  In this project, we have to make Flight price valuation model using new machine learning models from new data. Because with the change in market due to covid 19 impact, we have to check how price are drastically change when search for same date.

- Conceptual Background of the Domain Problem
  1. Firstly, we will prepare our own dataset using web scraping.
  2. After that we will check whether the project is a regression type or a classification type.
  3. We will also check whether our dataset is balanced or imbalanced. If it is an imbalanced one, we will apply sampling techniques to balance the dataset.
  4. Then we will do model building and check its accuracy.
  5. Our main motto is to build a model with good accuracy and for that we will also go for hyper parameter tuning.

- Review of Literature

  I am summarizing my research done on the topic.

- I have created my own dataset using web scraping and imported important libraries for my project.
- I have created the data frame.
- I have analysed my data by checking its shape, number of columns, presence of null values if any and checking the datatypes.
- Then I have done some data cleaning steps, e.g. Checking the value counts of the target variable, dropping some irrelevant columns from the dataset, checking correlation between the dependant and independent variables using heat map, visualizing data using distribution plots, detecting and removing skewness in my data if any, outliers detection using boxplots and removing them, balancing dataset using.

- Motivation for the Problem Undertaken

  I have been working towards many project, but this project give me immense pleasure to do. We all know how different flight website are used in the market and the price are quite different for all of them. My work is to make best model and suggest our client to use this which help them to make their business more and more profit. After all we engineers can provide best solution to the clients to improve their business. In this Project I used my potential to scrap data from the different website like Yatra, Ixigo, and MakemyTrip, etc which help to get the actual price of the market

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  If you look at data science, we are actually using mathematical models to model (and hopefully through the model to explain some of the things that we have seen) business circumstances, environment etc. and through these model, we can get more insights such as the outcomes of our decision undertaken, what should we do next or how shall we do it to improve the odds. So mathematical models are important, selecting the right one to answer the business question can bring tremendous value to the organization. Here I am using Decision Tree Regressor with accuracy 100% after hyper parameter tuning.

- Data Sources and their formats

  Data Source: The read_xlsx function of the pandas library is used to read the content of an Excel file into the python environment as a pandas Data Frame. The function can read the files from the OS by using proper path to the file. Data description: Pandas describe() is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values

- Data Preprocessing Done
  - I have checked for null values and there are null values present.
  - I have label encoded the object type columns in the dataset.
  - I have checked the correlation between dependant and independent variables using heat map. I have seen most of the independent variables are correlated with each other and the target variable is positively correlated with a very few independent variable.
  - I have checked outliers using boxplots, outliers are present in price column.
  - I also have checked for skewness in my data, but the skewness present is very negligible, so I don’t consider it.
  - I have splitted the dependant and independent variables into x and y.
  - I have scaled the data using StandardScaler method and made my data ready for model building.

- Hardware and Software Requirements and Tools Used
  - Hardware requirements: Processor: Intel(R) Core (TM) i3- 4030U CPU @ 1.90GHz RAM: 3.98 GB System type: 64-bit operating system, x64-based processor.
  - Software requirements:
    Python: One of the most used programming languages Tools used: Jupyter notebook: Jupyter is a free, opensource, interactive web tool known as a computational notebook where I have written my python codes. NumPy: NumPy is an open-source numerical Python library. NumPy contains a multi-dimensional array and matrix data structures. Pandas: Pandas is an open source Python package that is most widely used for data science/data analysis and machine learning tasks. It is built on top of another package named Numpy, which provides support for multi-dimensional arrays. Matplotlib: It is a Python library used for plotting graphs with the help of other libraries like Numpy and Pandas. It is a powerful tool for visualizing data in Python. Seaborn: It is also a Python library used for plotting graphs with the help of Matplotlib, Pandas, and Numpy.

    Scikit-learn:
    It is probably the most useful library for machine learning in Python. The sklearn library contains a lot of efficient tools for machine learning and statistical modeling including classification, regression, and clustering and dimensionality reduction.
    Scipy.stats: This module contains a large number of probability distributions as well as a growing library of statistical functions

Model/s Development and Evaluation

1. To check the correlation among the data, I have used heat map to visualize it.

2. To get a clear view of the columns visually, I have used distribution plots.
3. For checking outliers, I have used boxplots.
4. For scaling the data, I have used StandardScaler method.
5. For training and testing the data, I have imported train_test_split library from scikit-learn.
6. For model building, I have used 6 Regressor models(Linear Regression,DecisionTreeRegressor(),KNeighborsRegressor(),AdaBo ostRegr essor(),GradientBoostingRegressor(), RandomForestRegressor()),out of which AdaBoostRegressor model is the best model for my dataset
7. For better accuracy of the model, I have used hyper parameter tuning

- Testing of Identified Approaches (Algorithms)

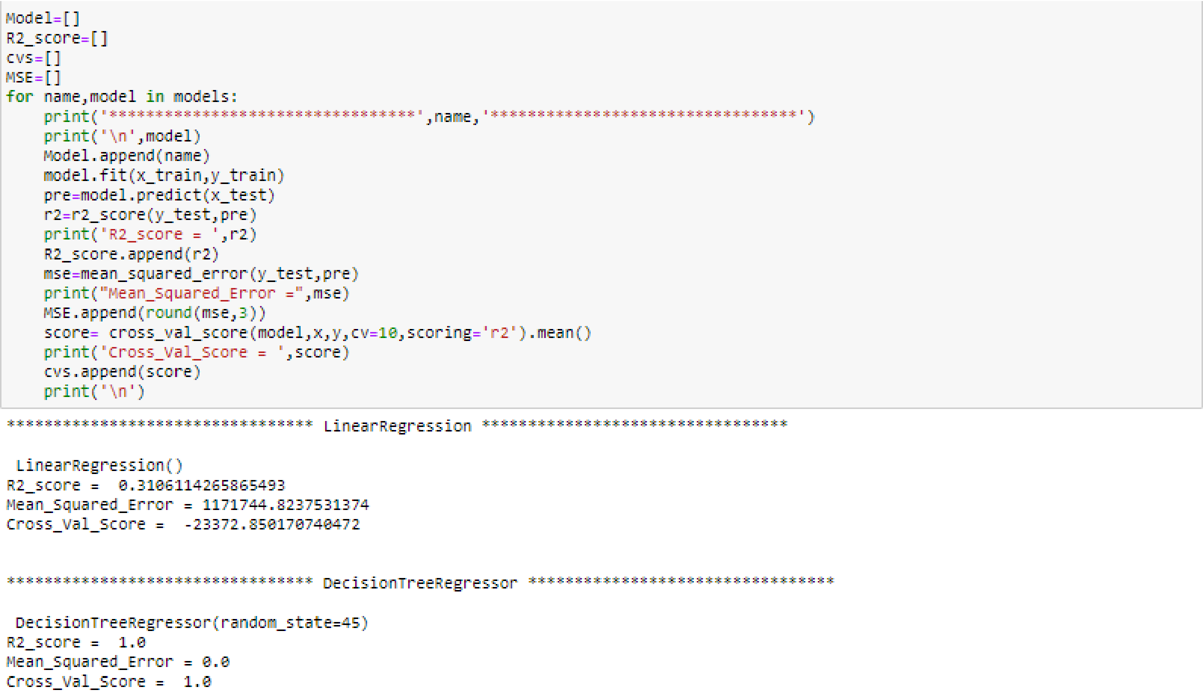
I have used 5 algorithms for testing.

- **DecisionTreeRegressor**-
  Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.
- **KNeighborsRegressor**-
  In k-NN regression, the output is the property value for the object. This value is the average of the values of k nearest neighbors.
- **AdaBoostRegressor**-
  An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.
- **LinearRegression**-
  Linear regression is a linear approach for modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables).
- **GradientBoostingRegressor**-

Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning.Gradient boosting algorithm can be used for predicting not only continuous target variable (as a Regressor) but also categorical target variable (as a Classifier).

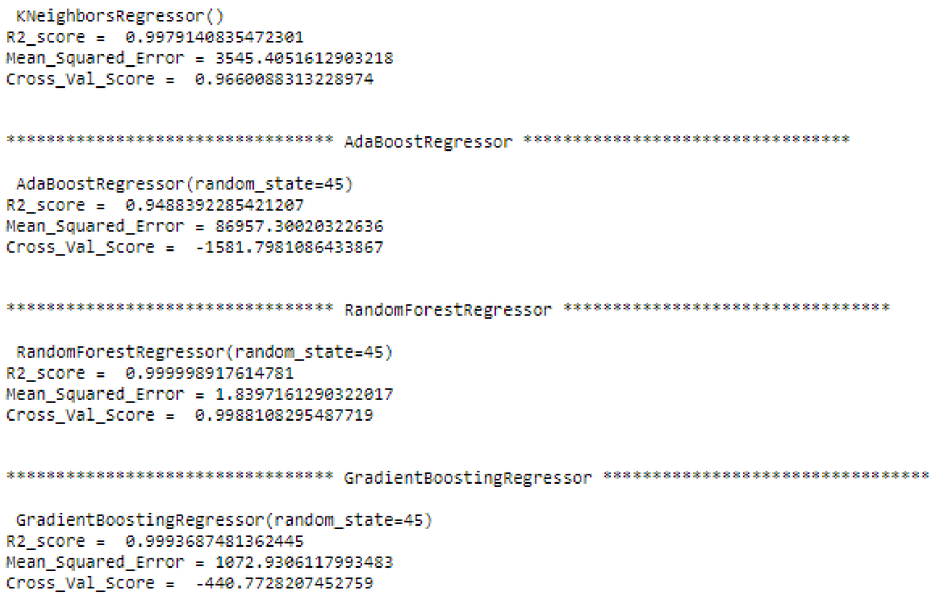
- **kneighborsRegressor-**
  K-Nearest Neighbors or KNN is a supervised machine learning algorithm and it can be used for classification and regression problems. KNN utilizes the entire dataset. Based on k neighbors value and distance calculation method (Minkowski, Euclidean, etc.), the model predicts the elements.

- Run and Evaluate selected models



```
Model=[]
R2_score=[]
cvs=[]
MSE=[]
for name,model in models:
    print('*********************************',name,'*********************************')
    print('\n',model)
    Model.append(name)
    model.fit(x_train,y_train)
    pre=model.predict(x_test)
    r2=r2_score(y_test,pre)
    print('R2_score = ',r2)
    R2_score.append(r2)
    mse=mean_squared_error(y_test,pre)
    print("Mean_Squared_Error =",mse)
    MSE.append(round(mse,3))
    score= cross_val_score(model,x,y,cv=10,scoring='r2').mean()
    print('Cross_Val_Score = ',score)
    cvs.append(score)
    print('\n')
```

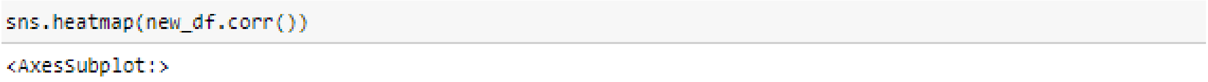
```
********************************* LinearRegression *********************************

 LinearRegression()
R2_score =  0.3106114265865493
Mean_Squared_Error = 1171744.8237531374
Cross_Val_Score =  -23372.850170740472


********************************* DecisionTreeRegressor *********************************

 DecisionTreeRegressor(random_state=45)
R2_score =  1.0
Mean_Squared_Error = 0.0
Cross_Val_Score =  1.0
```



```
 KNeighborsRegressor()
R2_score =  0.9979140835472301
Mean_Squared_Error = 3545.4051612903218
Cross_Val_Score =  0.9660088313228974


********************************** AdaBoostRegressor **********************************

 AdaBoostRegressor(random_state=45)
R2_score =  0.9488392285421207
Mean_Squared_Error = 86957.30020322636
Cross_Val_Score =  -1581.7981086433867


********************************** RandomForestRegressor **********************************

 RandomForestRegressor(random_state=45)
R2_score =  0.999998917614781
Mean_Squared_Error = 1.8397161290322017
Cross_Val_Score =  0.9988108295487719


********************************** GradientBoostingRegressor **********************************

 GradientBoostingRegressor(random_state=45)
R2_score =  0.9993687481362445
Mean_Squared_Error = 1072.9306117993483
Cross_Val_Score =  -440.7728207452759
```

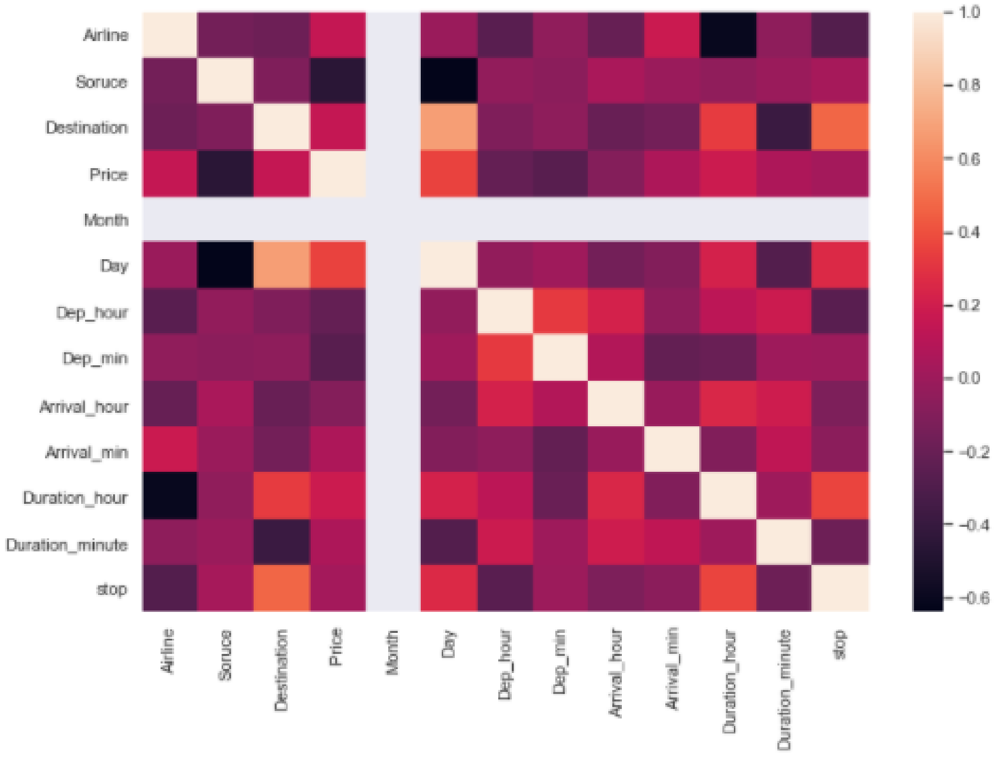
- # Visualizations

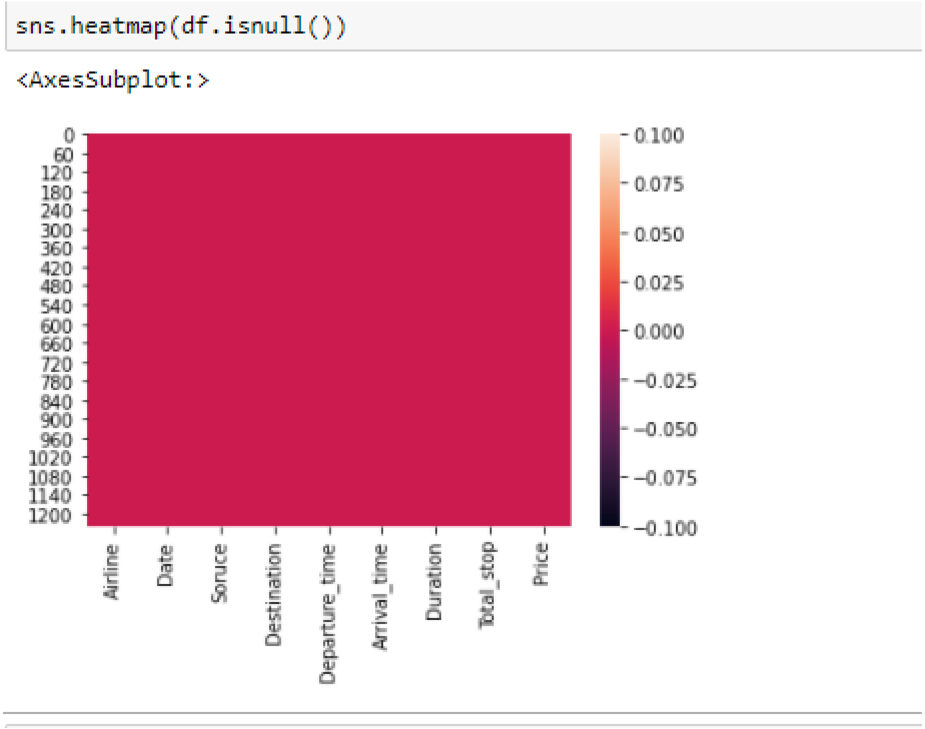
Correlation matrix using heatmap-checking correlation between dependant and independent variables.



```
sns.heatmap(new_df.corr())
```

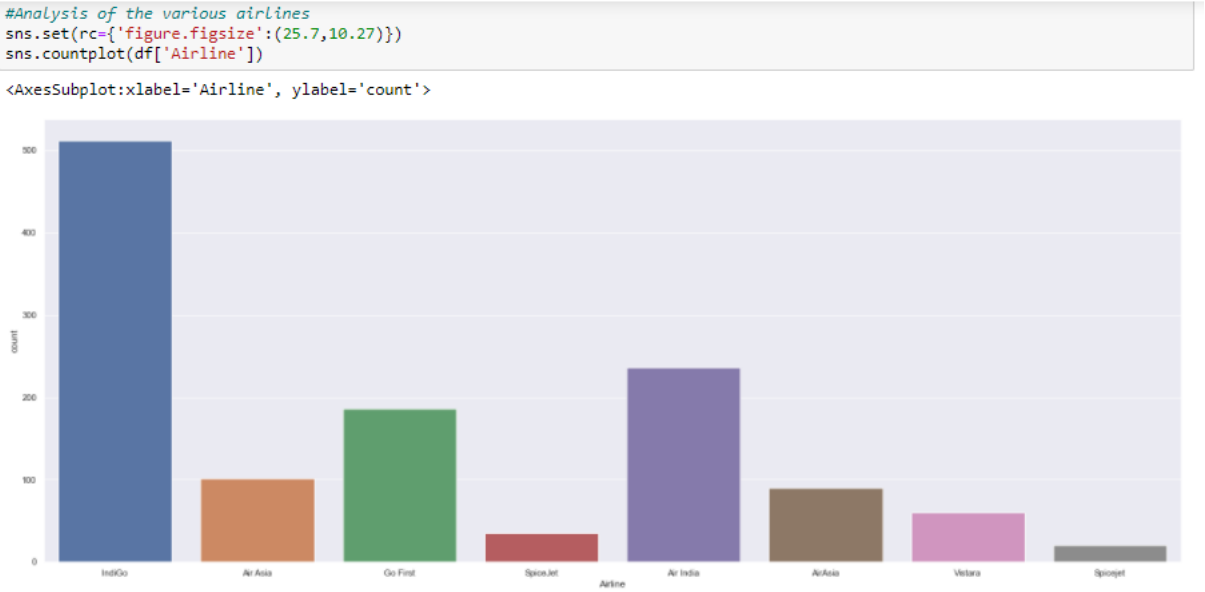
```
<AxesSubplot:>
```





```
sns.heatmap(df.isnull())
```

```
<AxesSubplot:>
```



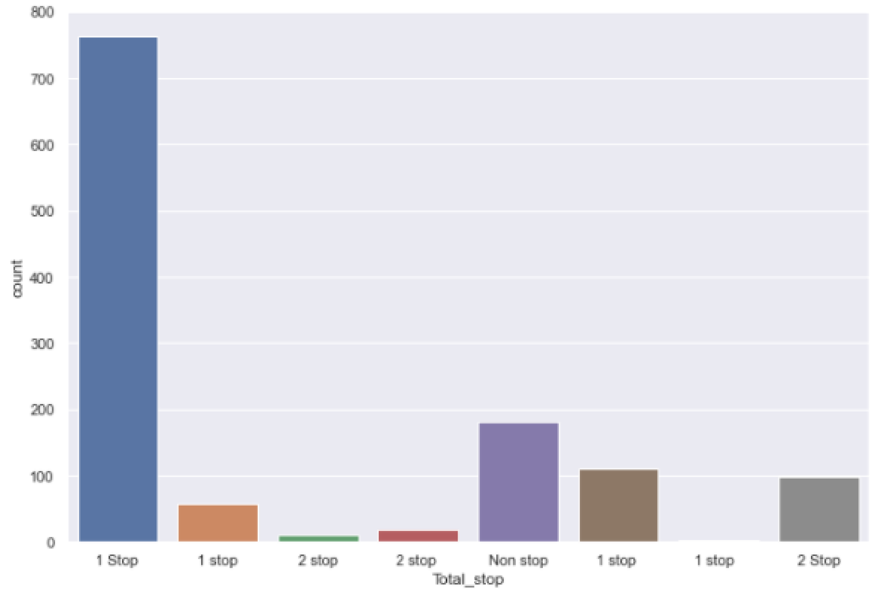
```
#Analysis of the various airlines
sns.set(rc={'figure.figsize':(25.7,10.27)})
sns.countplot(df['Airline'])
```

```
<AxesSubplot:xlabel='Airline', ylabel='count'>
```

As shown in the above plot, Indigo Travelled more in numbers in compare to other flights

```
sns.set(rc={'figure.figsize':(10.7,7.27)})
sns.countplot(df['Total_stop'])
```

```
<AxesSubplot:xlabel='Total_stop', ylabel='count'>
```

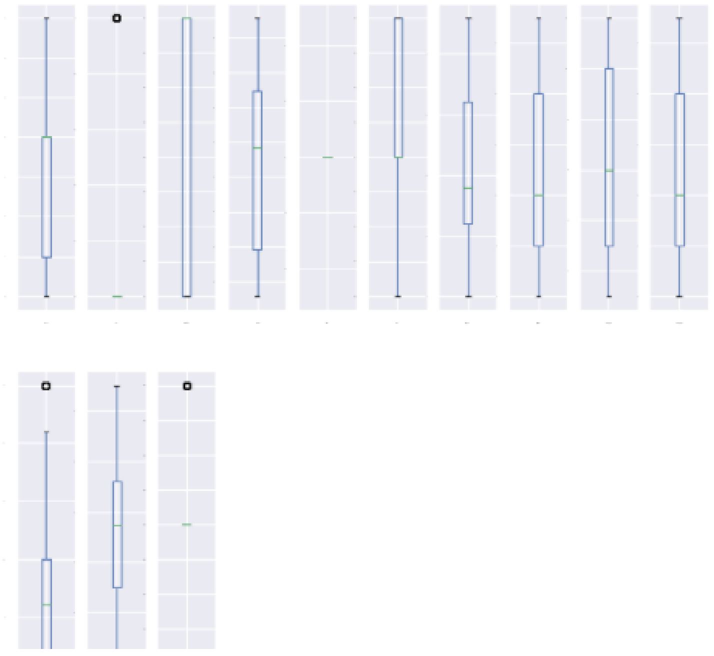


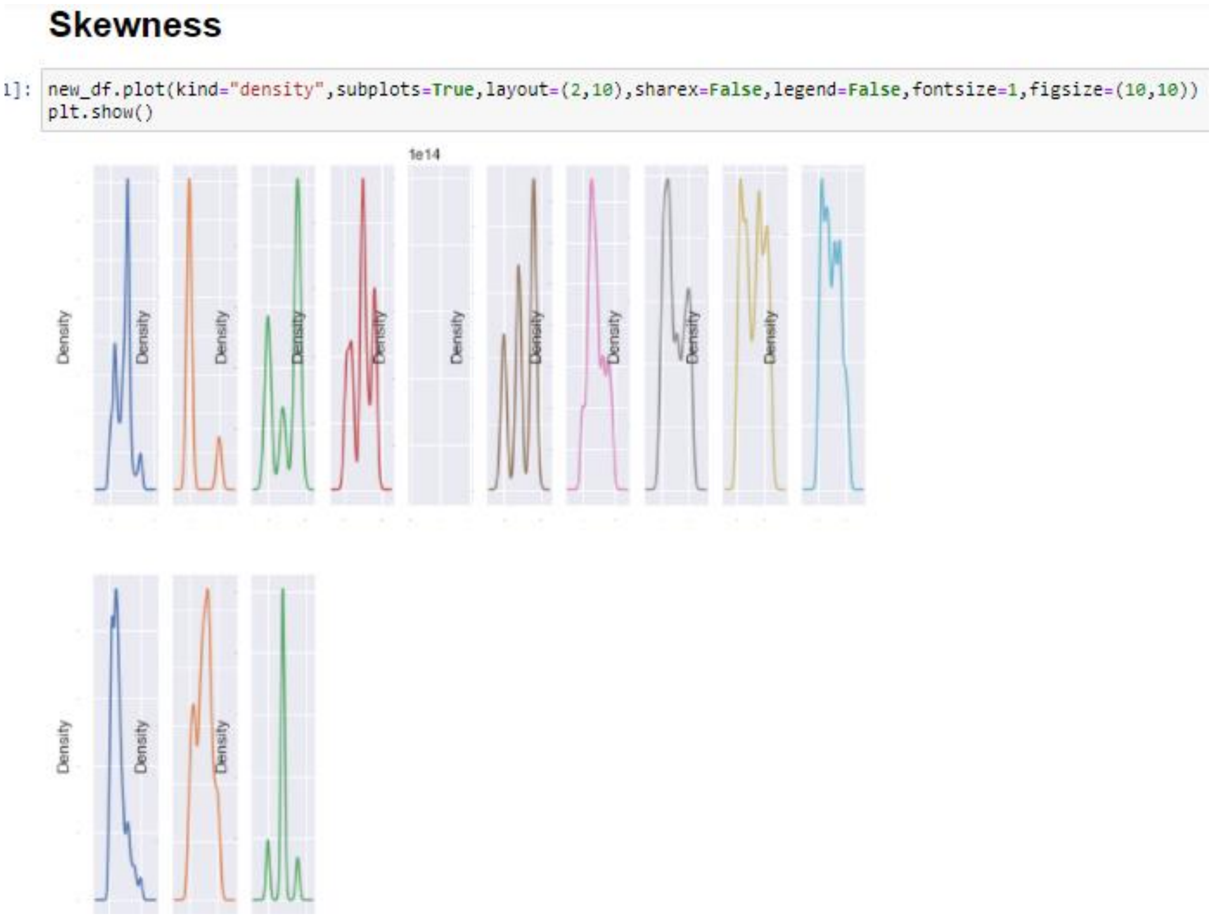
There are more flights which has more than 1 stop

OUTLIERS

3]:

```
new_df.plot(kind="box",subplots=True,layout=(2,10),sharex=False,legend=False,fontsize=1,figsize=(10,10))
plt.show()
```





Skewness

```
new_df.plot(kind="density",subplots=True,layout=(2,10),sharex=False,legend=False,fontsize=1,figsize=(10,10))
plt.show()
```

- Interpretation of the Results

In the visualization part, I have seen how my data looks like using heatmap, boxplot, distribution plots, histogram etc. In the pre-processing part, I have cleaned my data using many methods like interpolate, LabelEncoder etc. In the modelling part, I have designed our model using algorithm like DecisionTreeRegressor. The accuracy , Mean Absolute Error, Mean Squared Error, Root Mean Absolute Error are achieved for the model.

CONCLUSION

- Key Findings and Conclusions of the Study

The key findings are we have to study the data very clearly so that we are able to decide which data are relevant for our findings. The techniques that I have used are heatmap, interpolate,LabelEncoder etc. The conclusion of our study is we have to achieve a model with good accuracy and R2-score

- ## Learning Outcomes of the Study in respect of Data Science

  We will develop relevant programming abilities. We will demonstrate proficiency with statistical analysis of data. We will develop the ability to build and assess data-based models. We will execute statistical analysis with professional statistical software. The best algorithm for this project according to my work is DecisionTreeRegressor because the accuracy that I have achieved is quite satisfactory than the other model.

Limitations of this work and Scope for Future Work

The scope for future work is to collect as many data as we can so that the model can be built more efficiently.