FLIGHT PRICE PREDICTION

Submitted by:

SHAAHIDH IRFAAN SHA

ACKNOWLEDGMENT

This includes mentioning of all the references, research papers, data sources, professionals and other resources that helped you and guided you in completion of the project.

INTRODUCTION

- ## Business Problem Framing

  Flight fares changes continuously depending on a lot of variables like duration, no of stops, type of class, etc. I have built a model using sufficient data to predict the fare by giving valuable inputs

- ## Conceptual Background of the Domain Problem

  The model's domain understanding is on basic level, journey date/time/year has been separated to find which seasonal month flight bookings are high. Have encoded ordinal and nominal datas

- ## Review of Literature

  Fare prediction machine learning model uses two main models, regression analysis and random forest. Random forest always produce better results(accuracy).

  Another main observing is, price is mostly correlated to duration and no of stops.

Analytical Problem Framing

- Mathematical/ Analytical Modeling of the Problem

  Statistical part was to use mean to find the average price of the fare, quantile range was used to find how many outliers were there and how drastically it affects the model. For the model accuracy check we are using r2 score, mean squared error and root mean squared error

- Data Sources and their formats

  I have scraped data from popular flight booking websites and combined into one excel file and I have separated into train data and test data

- Data Preprocessing Done
  1) have separated arrival time, departure time.
  2) Some journey duration did not have hrs or mins, added few lines of code to rectify it
  3) Encoded data depending on what kind of data they are
  4) Dropped all the unwanted columns

Model/s Development and Evaluation

- Identification of possible problem-solving approaches (methods)

  The approach I took was to find the relation between categorical variables, since a lot of the variables in this dataset were categorical.

  I had to use boxplots to find out about outliers and how it affects the model.

  Used label encoding and mapping for encoding the datas

- Testing of Identified Approaches (Algorithms)

  Random forest training

- Run and Evaluate selected models

```
In [153]: #splitting the dataset
          from sklearn.model_selection import train_test_split
          x_train,x_test,y_train,y_test= train_test_split(x,y,test_size=0.3,random_state=41)
```

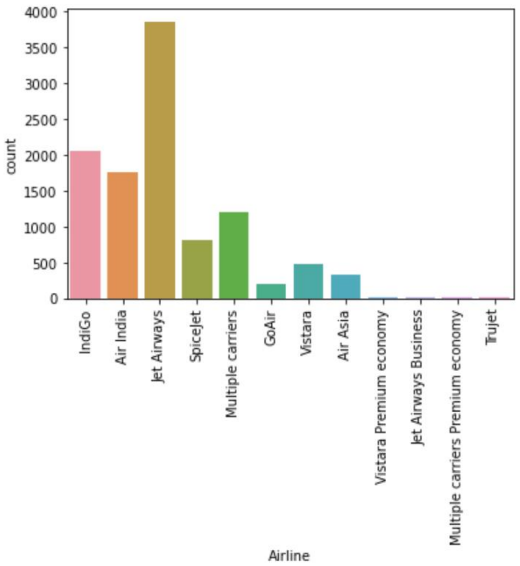
```
In [154]: from sklearn.ensemble import RandomForestRegressor
          from sklearn.metrics import mean_squared_error,r2_score

          rf= RandomForestRegressor()
          rf.fit(x_train,y_train)
          y_pred= rf.predict(x_test)

          print('Our model fits ',rf.score(x_train,y_train),' of our data')
          print('R2 score for the model is', r2_score(y_test,y_pred))
          print("MSE:",mean_squared_error(y_test,y_pred))
          print("RMSE:",np.sqrt(mean_squared_error(y_test,y_pred)))
```

```
Our model fits  0.9561997651932195  of our data
R2 score for the model is 0.7882958622139211
MSE: 4314498.712880744
RMSE: 2077.137143493598
```

Our Model gives an accuracy score of 78% which is satisfactory.

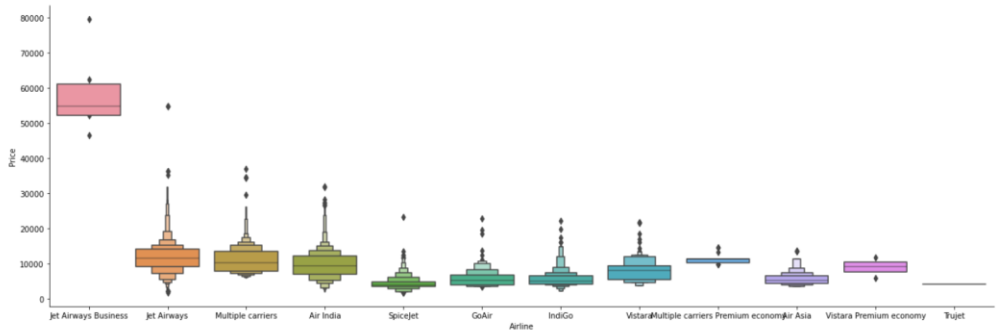
- Visualizations

  1)most booked fligts



  2)airline vs price

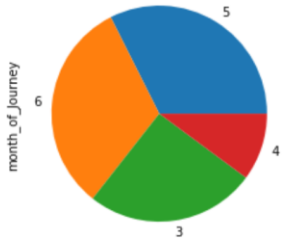
In [131]:
```
# Airline vs Price
sns.catplot(x = "Airline",y = "Price", data = df.sort_values("Price", ascending = False), kind="boxen", hei
plt.show()
```



3)which month has the highest booking

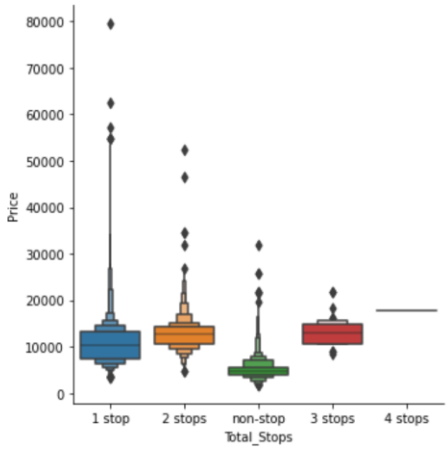
In [132]:
```
#in which month of year the flight travel is at high frquency
df['month_of_Journey'].value_counts().plot(kind='pie')
#may is the holiday month for students
```

```
<AxesSubplot:ylabel='month_of_Journey'>
```



4)does the stop increase affect the price

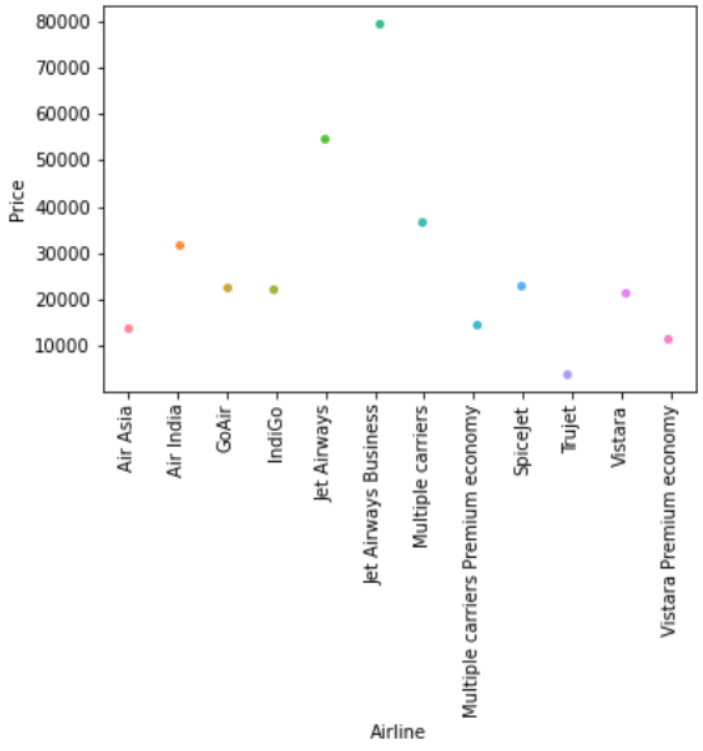
In [133]:
```
# no.of.stops vs Price
sns.catplot(y = "Price", x = "Total_Stops", data = df.sort_values("Price", ascending = False),kind='boxen')
plt.show()
```



5)which company has the highest fare

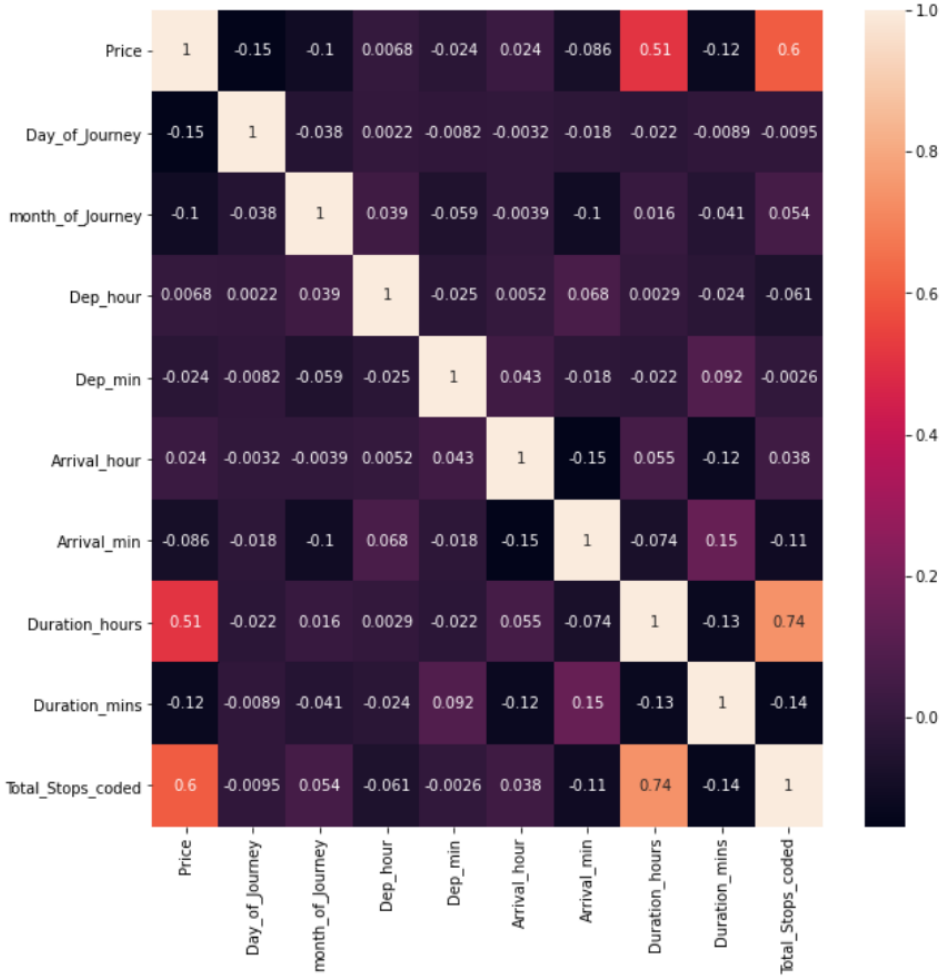
```
In [134]: #which airline charges the most expensive fair in india
          a=df.groupby('Airline').max()
          a['Airline']=a.index
          ax=sns.stripplot(x='Airline',y='Price',data=a)
          ax.set_xticklabels(labels=(a['Airline'].unique()),rotation=90)
```

```
[Text(0, 0, 'Air Asia'),
 Text(1, 0, 'Air India'),
 Text(2, 0, 'GoAir'),
 Text(3, 0, 'IndiGo'),
 Text(4, 0, 'Jet Airways'),
 Text(5, 0, 'Jet Airways Business'),
 Text(6, 0, 'Multiple carriers'),
 Text(7, 0, 'Multiple carriers Premium economy'),
 Text(8, 0, 'SpiceJet'),
 Text(9, 0, 'Trujet'),
 Text(10, 0, 'Vistara'),
 Text(11, 0, 'Vistara Premium economy')]
```



6)correlation matrix

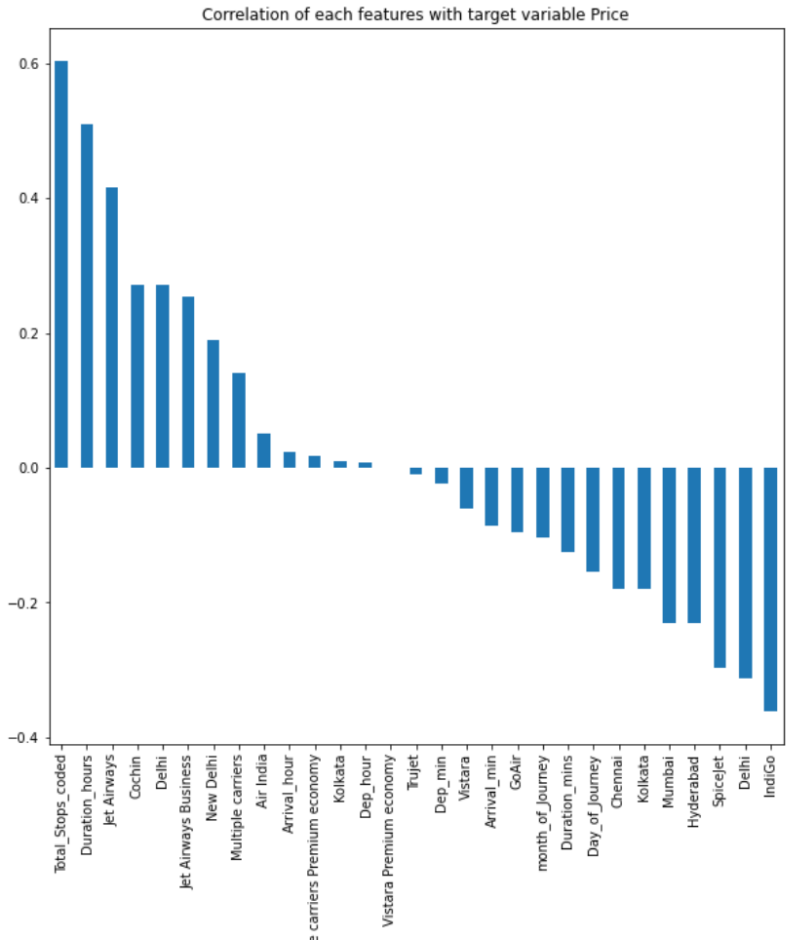
In [140]:
```
plt.figure(figsize=(10,10),facecolor='white')
ax = sns.heatmap(df.corr(),annot=True)
```



7)correlation with target variables

```
In [151]: #correlation with the target variable
          plt.figure(figsize=(10,10),facecolor='white')
          corr_data= data.corr()
          corr_data=corr_data.Price.sort_values(ascending=False).drop(['Price'])
          corr_data.plot(kind='bar',title='Correlation of each features with target variable Price')
```

```
<AxesSubplot:title={'center':'Correlation of each features with target variable Price'}>
```



- Interpretation of the Results
  1) Our random forest model gives 78% accuracy.
  2) jet airways holds the title as the most booked airlines in India
  3) jet airways business has the highest fares compared to other airline companies
  4) non-stop flights cost less than other options
  5) 5th month is the highest booked month in a year
  6) Correlation matrix proves no.of.stops to price and duration to price are highly correlated

CONCLUSION

- Key Findings and Conclusions of the Study
  1) Flight with high stop counts are in minimal disposals.
  2) May is the highest booked month since kids have their summer holiday
  3) Additional info has little to no relation to fares
  4) Duration holds the most relation to price.

- Limitations of this work and Scope for Future Work

  I was not able to predict the price with the type of class(economic, business and first class)

  Was not able to find which whether the price vary when booked in day or night