Flight Price Prediction

Fahad Reda

Agenda

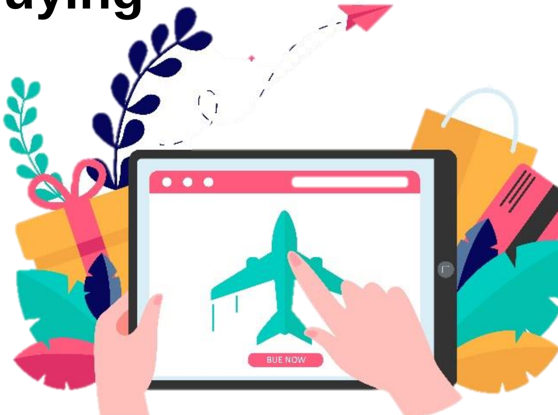
- Inspiration
- Problem Statement
- Data Source
- EDA
- Statistical Analysis
- Feature Engineering
- Models used
- Future Work and Conclusion

Inspiration

- Since we are in a Lockdown and can’t travel, so I am pretty sure that once the lockdown is lifted a lot of people are going to travel , which means the ticket prices will be higher than the usual !

Problem Statement

One of the main thing to consider while traveling is calculating the cost of the trip, where the price of the flights ticket plays an important role while preparing the budget for the trip, so this project is going to help the traveler to Predict the price of the flight ticket, as buying tickets is a very hectic process



Data Source

Dataset: Data_train.xlsx

Source of Dataset: MachineHack Hackathon

Format: xlsx

Size: 517 Kb

Shape: (10683,11)

Dataset: Test_test.xlsx

Source of Dataset: MachineHack Hackathon

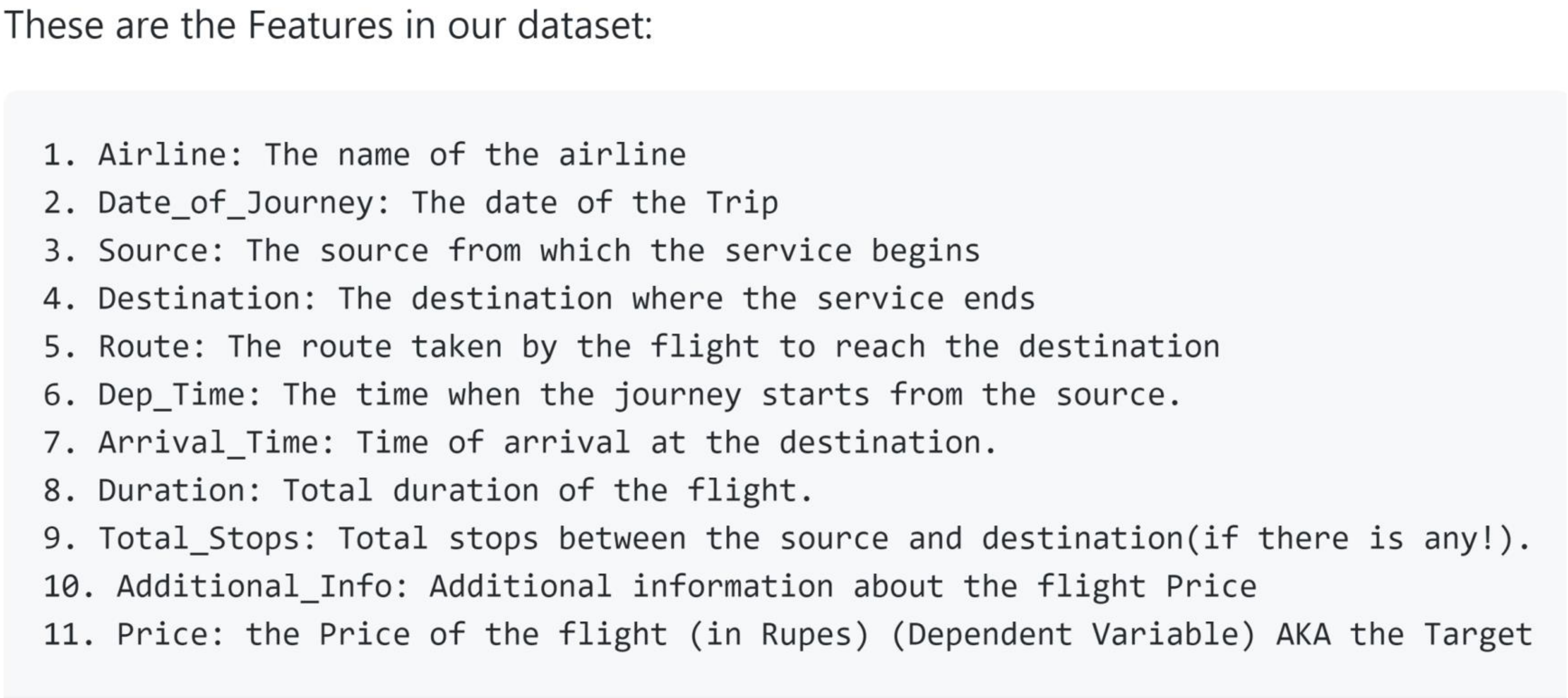
Format: xlsx

Size: 117 Kb

Shape: (2671,10)

So 80% used for training and 20% for testing

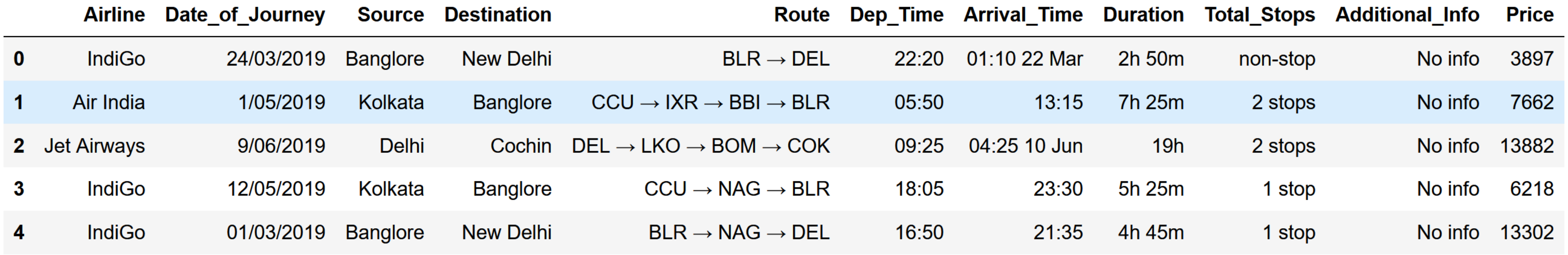
Exploratory Data Analysis(EDA)



These are the Features in our dataset:

```
1. Airline: The name of the airline
2. Date_of_Journey: The date of the Trip
3. Source: The source from which the service begins
4. Destination: The destination where the service ends
5. Route: The route taken by the flight to reach the destination
6. Dep_Time: The time when the journey starts from the source.
7. Arrival_Time: Time of arrival at the destination.
8. Duration: Total duration of the flight.
9. Total_Stops: Total stops between the source and destination(if there is any!).
10. Additional_Info: Additional information about the flight Price
11. Price: the Price of the flight (in Rupes) (Dependent Variable) AKA the Target
```

Exploratory Data Analysis(EDA)

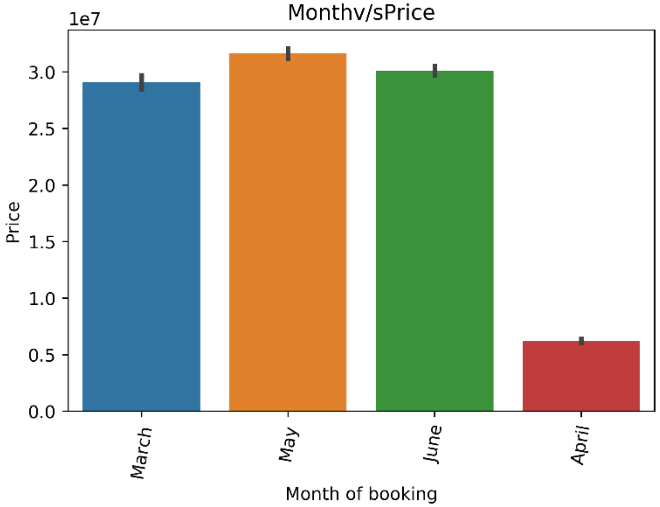


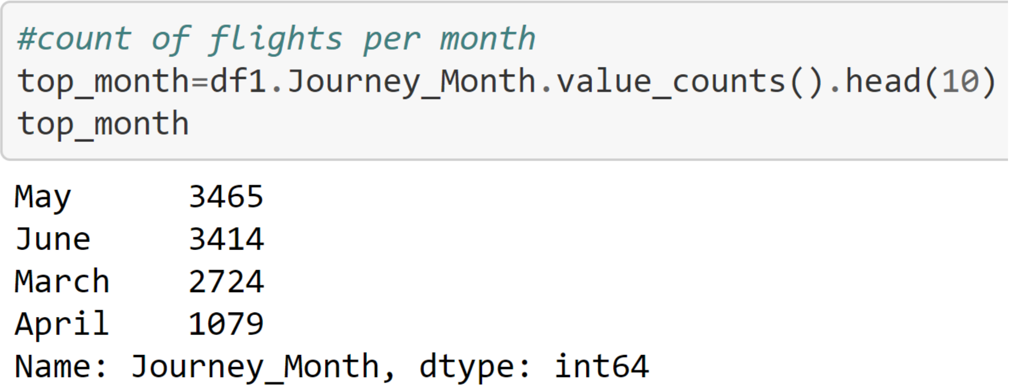
	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price
0	IndiGo	24/03/2019	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897
1	Air India	1/05/2019	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662
2	Jet Airways	9/06/2019	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882
3	IndiGo	12/05/2019	Kolkata	Banglore	CCU → NAG → BLR	18:05	23:30	5h 25m	1 stop	No info	6218
4	IndiGo	01/03/2019	Banglore	New Delhi	BLR → NAG → DEL	16:50	21:35	4h 45m	1 stop	No info	13302

Exploratory Data Analysis(EDA)

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10683 entries, 0 to 10682
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype 
---  ------           --------------  ----- 
 0   Airline          10683 non-null  object
 1   Date_of_Journey  10683 non-null  object
 2   Source           10683 non-null  object
 3   Destination      10683 non-null  object
 4   Route            10682 non-null  object
 5   Dep_Time         10683 non-null  object
 6   Arrival_Time     10683 non-null  object
 7   Duration         10683 non-null  object
 8   Total_Stops      10682 non-null  object
 9   Additional_Info  10683 non-null  object
 10  Price            10683 non-null  int64 
dtypes: int64(1), object(10)
memory usage: 918.2+ KB
```

Exploratory Data Analysis(EDA)





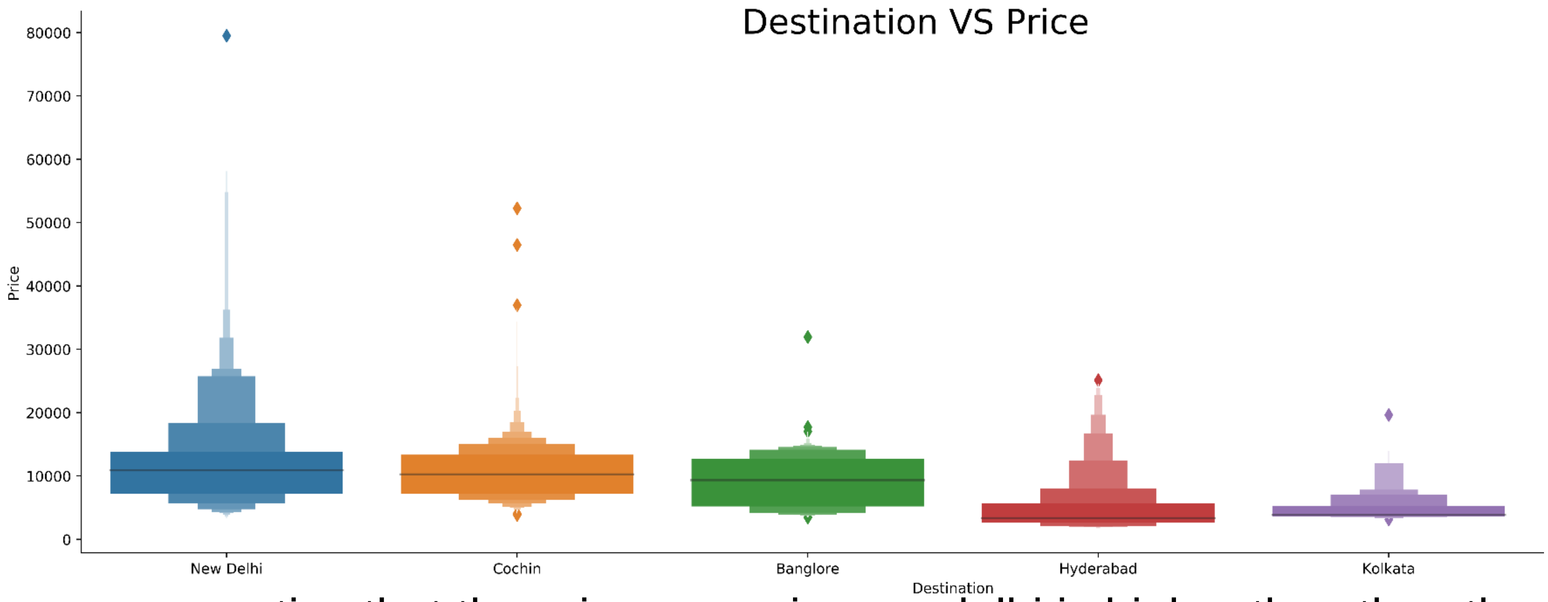
```
#count of flights per month
top_month=df1.Journey_Month.value_counts().head(10)
top_month
```

```
May      3465
June     3414
March    2724
April    1079
Name: Journey_Month, dtype: int64
```

We can notice that prices are higher in the month of May and way cheaper in April

There were 3465 flights in May and only 1079 flights in April

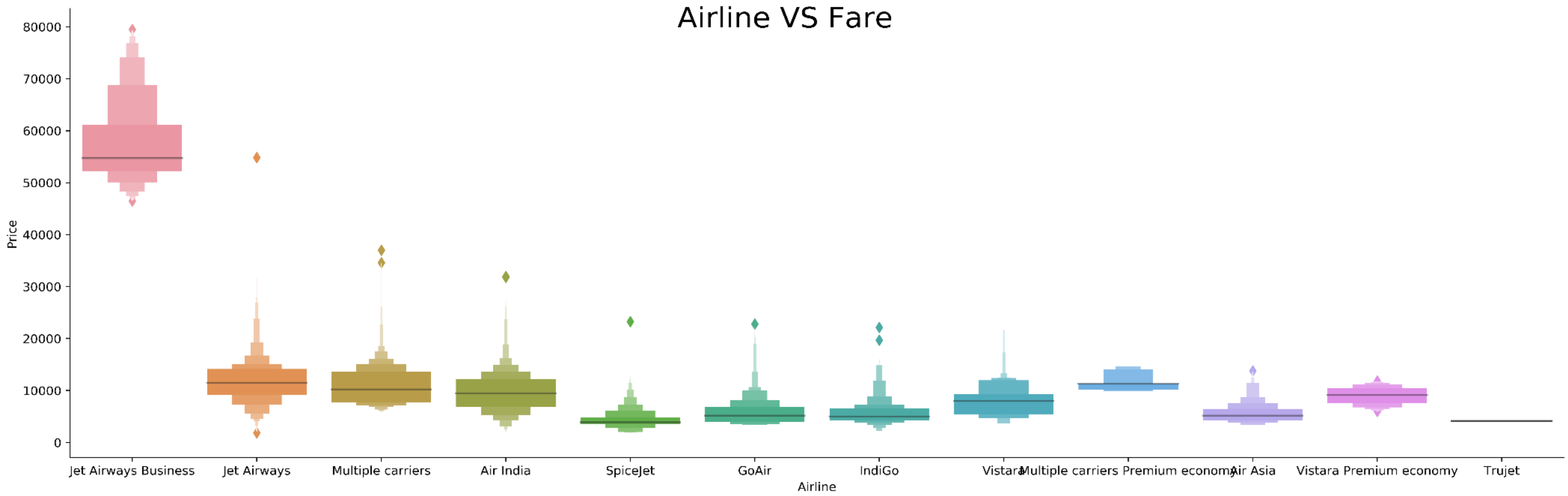
Exploratory Data Analysis(EDA)



Destination VS Price

we can notice that the price range in new delhi is higher than the other cities, and this can be due the jet fuel prices in delhi has increased in 2018 by 26.4%

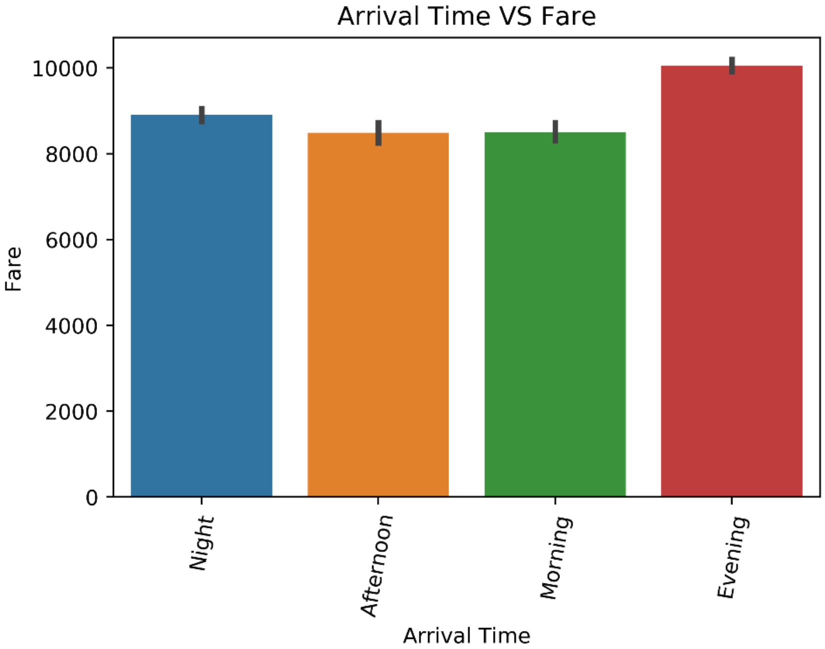
Exploratory Data Analysis(EDA)



we can notice that jet airways (both the business and the standard one) are highly priced because they are full service airlines are always expensive because of the amenities they provide

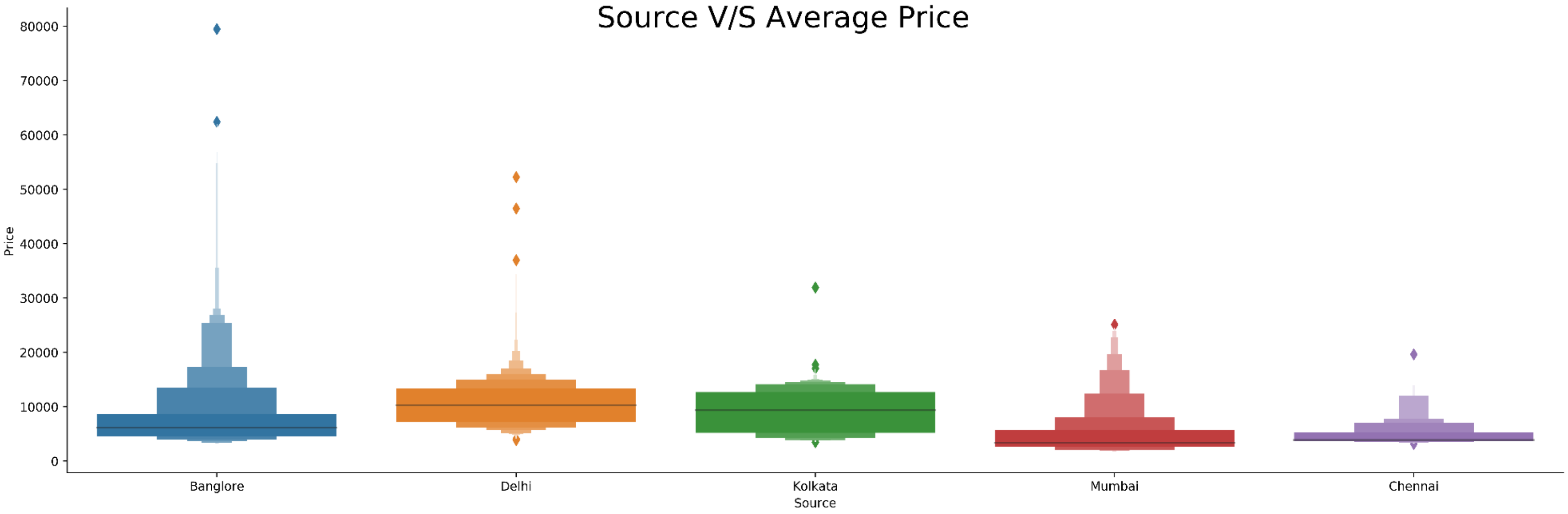
Exploratory Data Analysis(EDA)

Exploratory Data Analysis(EDA)



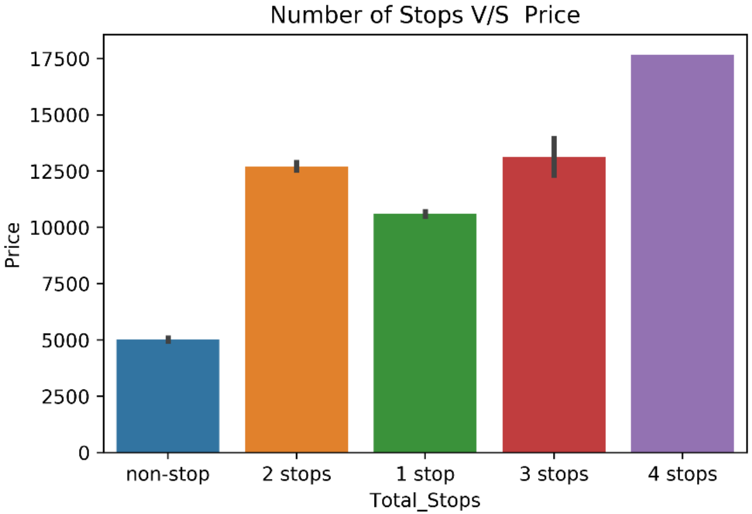
Here we can see that the flights that arrives in the evening their prices are higher than the other timings

Exploratory Data Analysis(EDA)



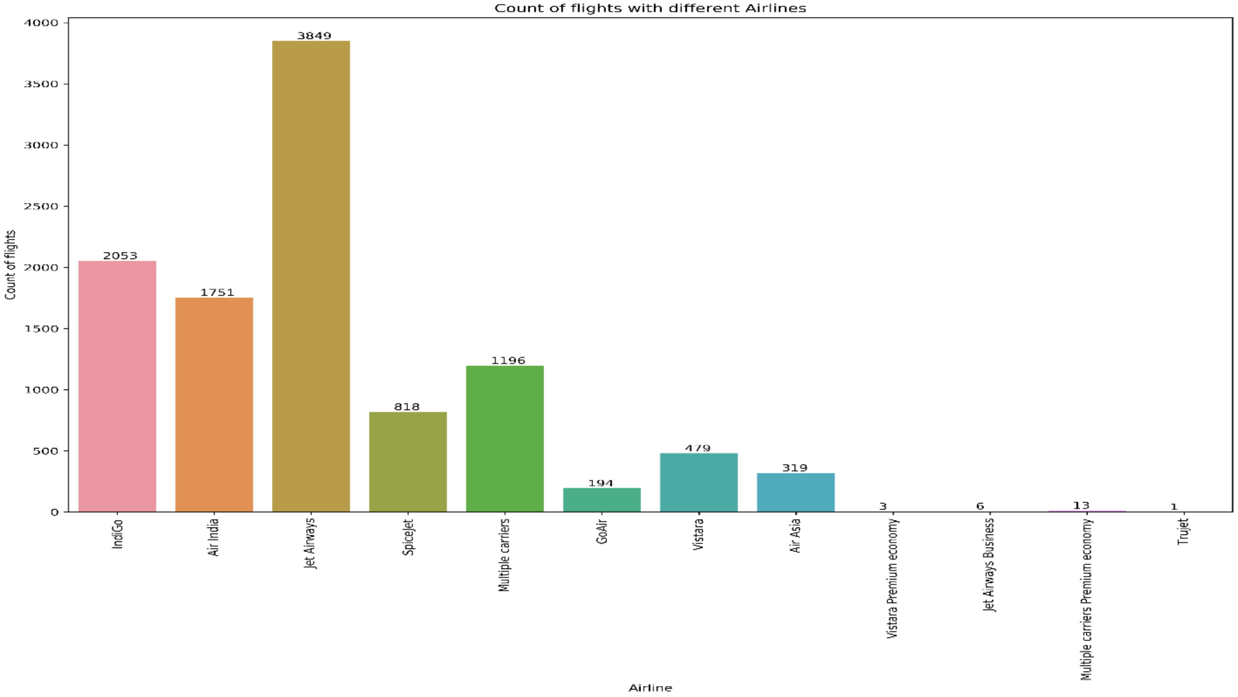
If you are traveling from Delhi and Kolkata the prices will be higher than the other cities

Exploratory Data Analysis(EDA)



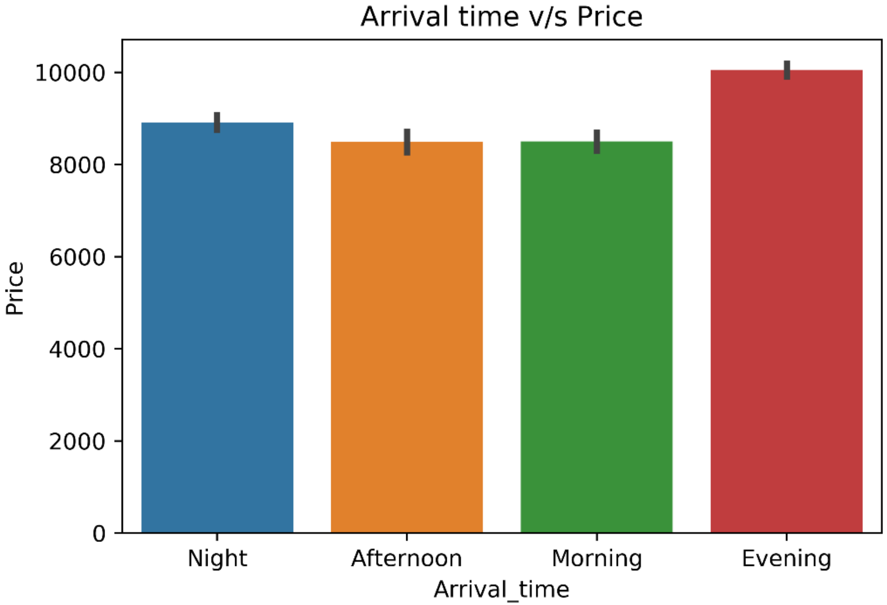
The more stops you will have on your trip, the higher the price it will get

Exploratory Data Analysis(EDA)



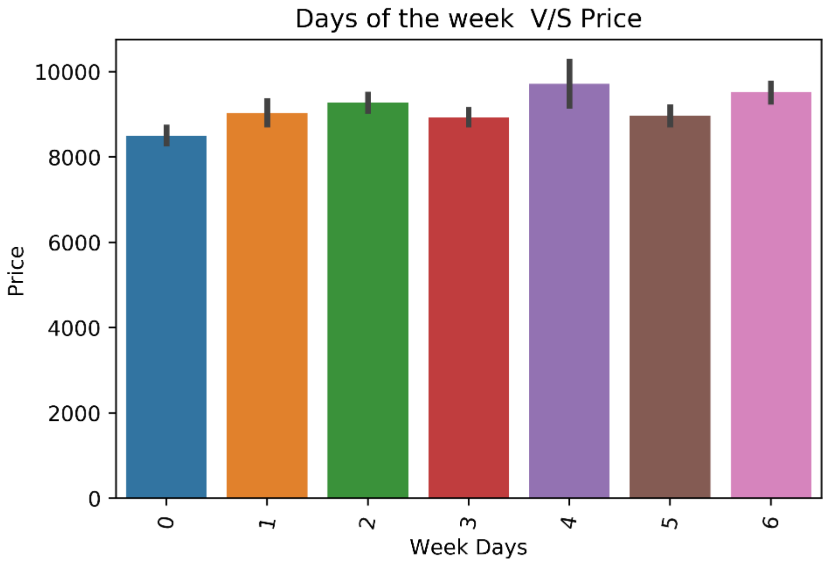
Most of the people travel using Jet airways

Exploratory Data Analysis(EDA)



we can see that flights that arrives in the evening are higher in price than the other timings

Exploratory Data Analysis(EDA)



We can see that prices are higher on Friday!

0= Monday , 1=Tuesday , 2= Wednesday , 3= Thursday , 4= Friday ,5= Saturday , 6=Sunday

Statistical Analysis

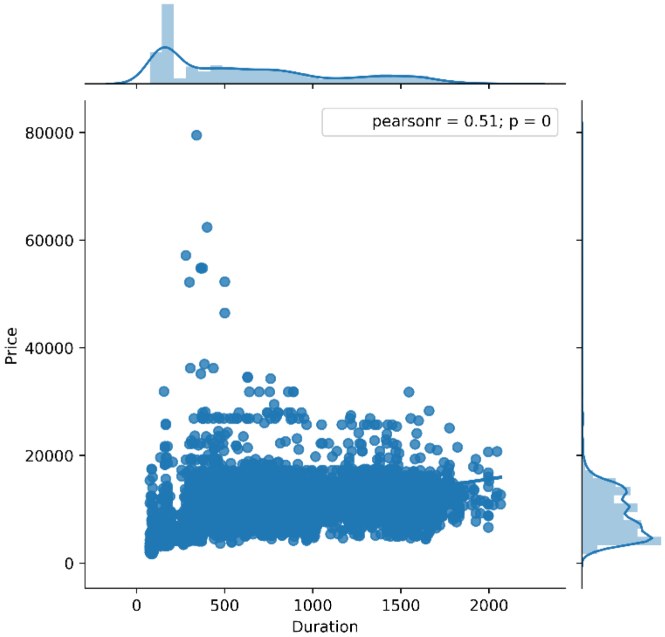
Pearson Correlation

is a measure of the strength of a linear association between two variables

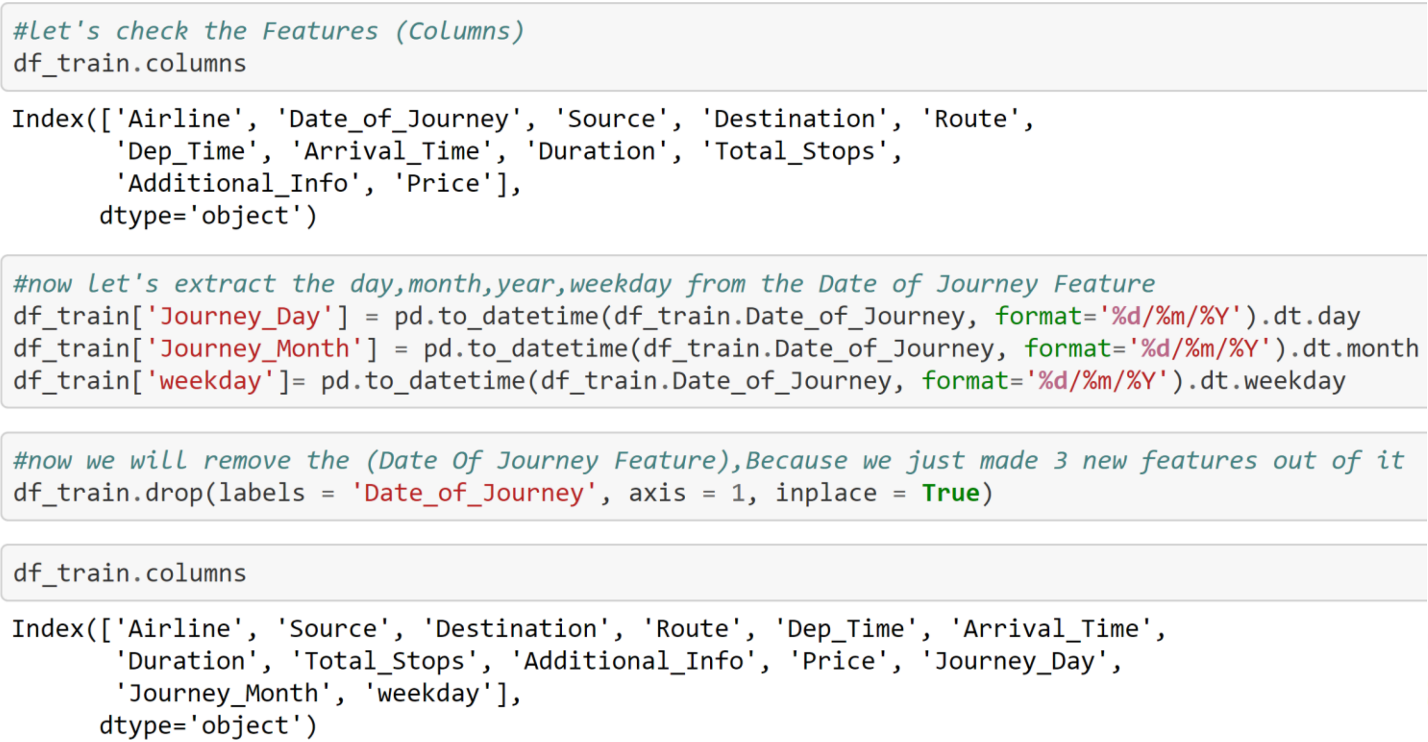
Null Hypothesis(H0): the two variables are not correlated

Alternative Hypothesis(H1): the two variables are correlated

- we can see that our p-value is greater than the 0.05, which means we accept H1 and can say that the target variable and independent variable are correlated



Feature Engineering



```
#let's check the Features (Columns)
df_train.columns
```

```
Index(['Airline', 'Date_of_Journey', 'Source', 'Destination', 'Route',
       'Dep_Time', 'Arrival_Time', 'Duration', 'Total_Stops',
       'Additional_Info', 'Price'],
      dtype='object')
```

```
#now let's extract the day,month,year,weekday from the Date of Journey Feature
df_train['Journey_Day'] = pd.to_datetime(df_train.Date_of_Journey, format='%d/%m/%Y').dt.day
df_train['Journey_Month'] = pd.to_datetime(df_train.Date_of_Journey, format='%d/%m/%Y').dt.month
df_train['weekday']= pd.to_datetime(df_train.Date_of_Journey, format='%d/%m/%Y').dt.weekday
```

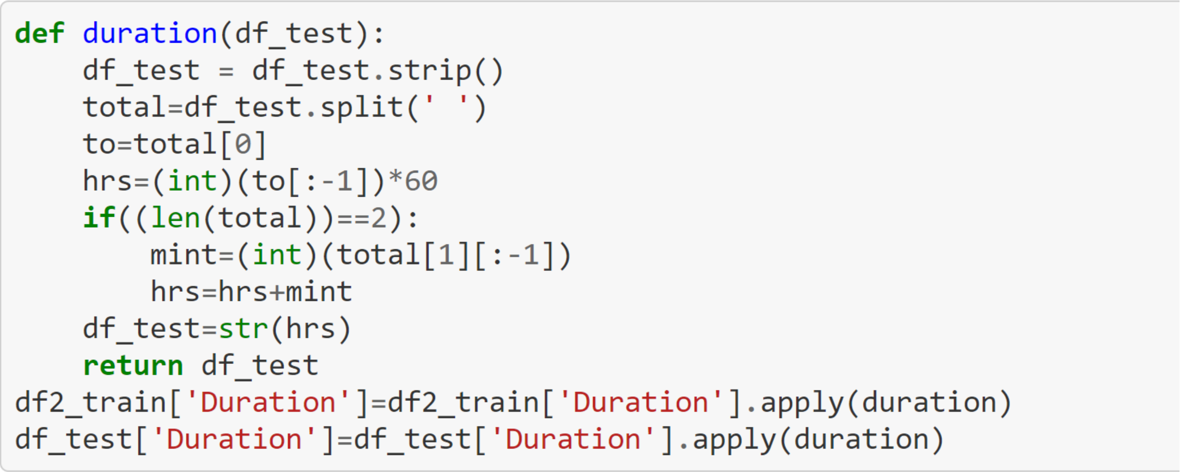
```
#now we will remove the (Date Of Journey Feature),Because we just made 3 new features out of it
df_train.drop(labels = 'Date_of_Journey', axis = 1, inplace = True)
```

```
df_train.columns
```

```
Index(['Airline', 'Source', 'Destination', 'Route', 'Dep_Time', 'Arrival_Time',
       'Duration', 'Total_Stops', 'Additional_Info', 'Price', 'Journey_Day',
       'Journey_Month', 'weekday'],
      dtype='object')
```

Feature Engineering

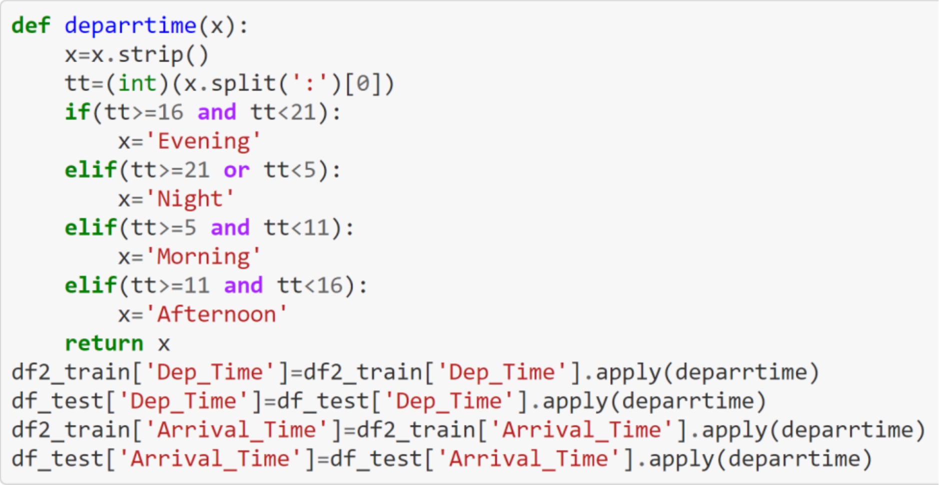
Here I had to convert the Duration into minutes



```
def duration(df_test):
    df_test = df_test.strip()
    total=df_test.split(' ')
    to=total[0]
    hrs=(int)(to[:-1])*60
    if((len(total))==2):
        mint=(int)(total[1][:-1])
        hrs=hrs+mint
    df_test=str(hrs)
    return df_test
df2_train['Duration']=df2_train['Duration'].apply(duration)
df_test['Duration']=df_test['Duration'].apply(duration)
```

Feature Engineering

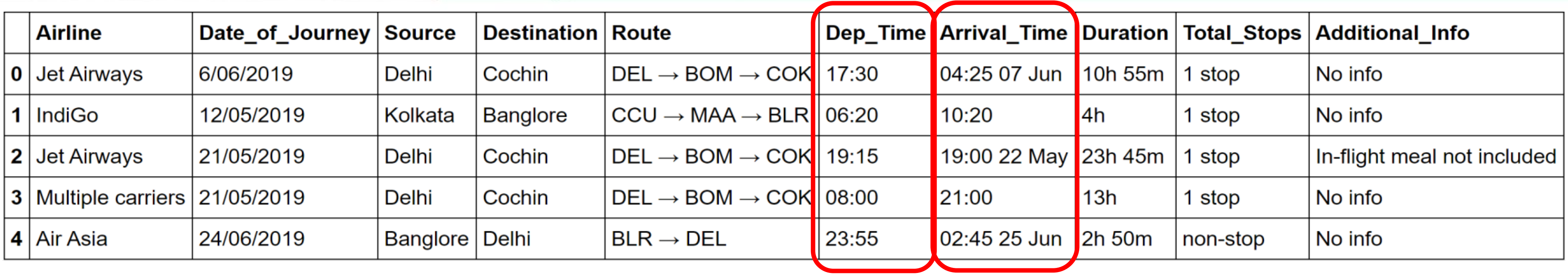
extract whether if the departure and arrival time of the flights occured at Morning , Evening , Night or Afternoon



```
def deparrtime(x):
    x=x.strip()
    tt=(int)(x.split(':')[0])
    if(tt>=16 and tt<21):
        x='Evening'
    elif(tt>=21 or tt<5):
        x='Night'
    elif(tt>=5 and tt<11):
        x='Morning'
    elif(tt>=11 and tt<16):
        x='Afternoon'
    return x
df2_train['Dep_Time']=df2_train['Dep_Time'].apply(deparrtime)
df_test['Dep_Time']=df_test['Dep_Time'].apply(deparrtime)
df2_train['Arrival_Time']=df2_train['Arrival_Time'].apply(deparrtime)
df_test['Arrival_Time']=df_test['Arrival_Time'].apply(deparrtime)
```

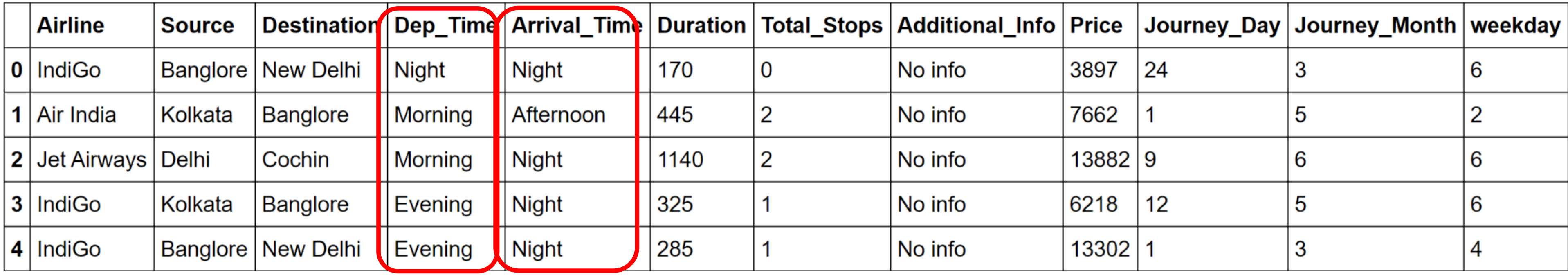
Feature Engineering

Before



	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info
0	Jet Airways	6/06/2019	Delhi	Cochin	DEL → BOM → COK	17:30	04:25 07 Jun	10h 55m	1 stop	No info
1	IndiGo	12/05/2019	Kolkata	Banglore	CCU → MAA → BLR	06:20	10:20	4h	1 stop	No info
2	Jet Airways	21/05/2019	Delhi	Cochin	DEL → BOM → COK	19:15	19:00 22 May	23h 45m	1 stop	In-flight meal not included
3	Multiple carriers	21/05/2019	Delhi	Cochin	DEL → BOM → COK	08:00	21:00	13h	1 stop	No info
4	Air Asia	24/06/2019	Banglore	Delhi	BLR → DEL	23:55	02:45 25 Jun	2h 50m	non-stop	No info

After



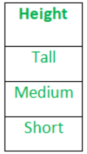
	Airline	Source	Destination	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Journey_Day	Journey_Month	weekday
0	IndiGo	Banglore	New Delhi	Night	Night	170	0	No info	3897	24	3	6
1	Air India	Kolkata	Banglore	Morning	Afternoon	445	2	No info	7662	1	5	2
2	Jet Airways	Delhi	Cochin	Morning	Night	1140	2	No info	13882	9	6	6
3	IndiGo	Kolkata	Banglore	Evening	Night	325	1	No info	6218	12	5	6
4	IndiGo	Banglore	New Delhi	Evening	Night	285	1	No info	13302	1	3	4

Feature Engineering

What is LabelEncoder? And why we use it?

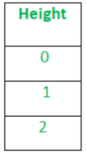
In Machine Learning Models we are required to convert the categorical features to numeric one ,so the model can read it

Before Applying LabelEncoder

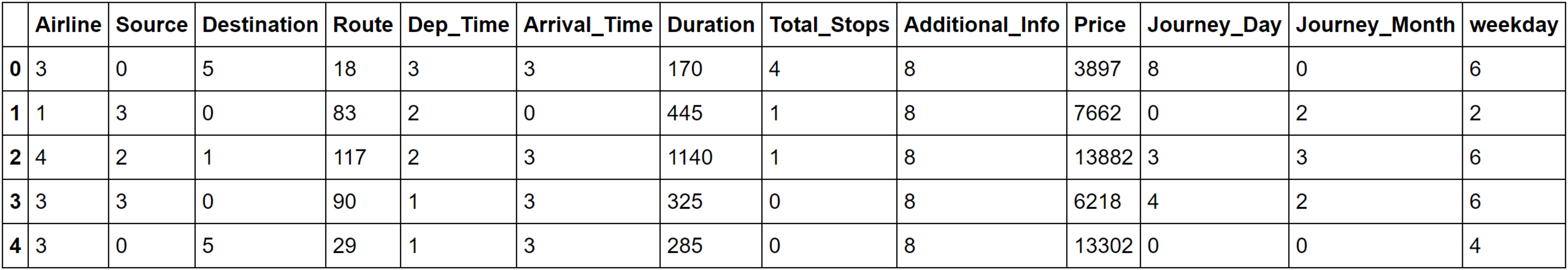


Height
Tall
Medium
Short

After Applying LabelEncoder



Height
0
1
2



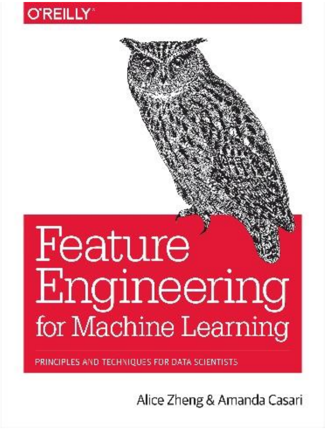
	Airline	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Journey_Day	Journey_Month	weekday
0	3	0	5	18	3	3	170	4	8	3897	8	0	6
1	1	3	0	83	2	0	445	1	8	7662	0	2	2
2	4	2	1	117	2	3	1140	1	8	13882	3	3	6
3	3	3	0	90	1	3	325	0	8	6218	4	2	6
4	3	0	5	29	1	3	285	0	8	13302	0	0	4

Models used

- Random Forest = 90.03%
- KNN= 77.05%
- XGBoost = 87.48%
- Gradient Boost= 87.59%

Challenges

1- This was by far the most challenging project, because it required a lot of feature engineering
2- Faced some issues in plotting and saving them in high quality
3- I tired to improve KNN from 75.7% to 77.05%
Took me some time to do that
4- Wanted to try Deep Learning, but couldn’t !
5- not enough materials covered in Feature engineering
Specially in General Courses like on UDEMY,so had to
Read books


O'REILLY
Feature Engineering for Machine Learning
PRINCIPLES AND TECHNIQUES FOR DATA SCIENTISTS
Alice Zheng & Amanda Casari

Sinan Ozdemir, Divya Susarla
Feature Engineering Made Easy
Identify unique features from your dataset in order to build powerful machine learning systems
Packt>

Future Work and Conclusion

Future Work:

1- Building a web app using Flask

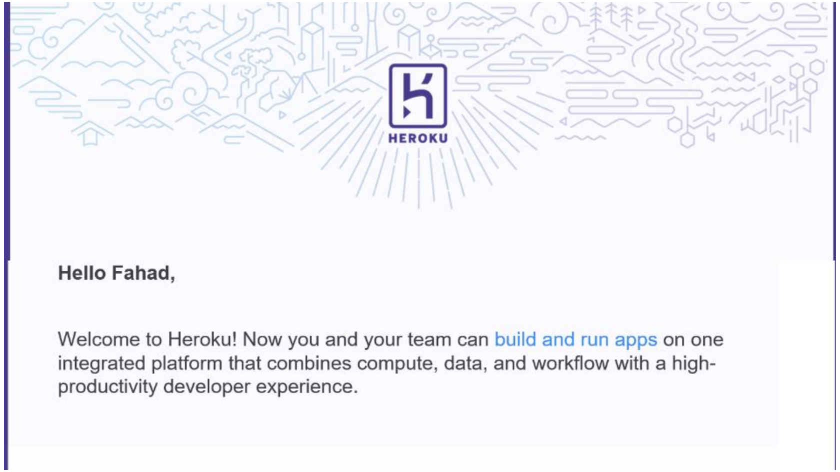
2- Will use Deep Learning and compare it with ML Models

3- will use Linear Regression and check how it performed!

Conclusion:

1- I Enjoyed working on this project as it really tested My skills in Feature engineering

2- I would like to Thank Dr.Rick and Ms.Lujain for their continuous support and help



Hello Fahad,

Welcome to Heroku! Now you and your team can build and run apps on one integrated platform that combines compute, data, and workflow with a high-productivity developer experience.

Thank you