FLIP ROBO

FLIGHT PRICE PREDICTION

Submitted by:
Bishwajit Bhattacharya

ACKNOWLEDGMENT

I taken data bye web scrap yatra.com, I filtered for single person economy. The data is date from 23jan 2022 to 30jan 2022. Source and destination is like Delhi to Mumbai, Delhi to Bangalore, Delhi to Chennai, Mumbai to Bangalore, Mumbai to Chennai.

INTRODUCTION

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)

The another issue of this situation is because of covid situation lots of flight either cancelled or got delay. Which is unexpectedly the prices vary.

Analytical Problem Framing

I use selenium for web scrap data , I went to yatra.com site and filter for single person economy and enter source destination and press enter I went to the link and take data like airline name , departure time arrival time , stops and price. After this first I create a data frame and then put data on one excel. Again I just change the date of site and again do the same process. After 30th jan 2022 I changes source and destination.

```
flight_name=driver.find_elements_by_xpath("//span[@class='i-b text ellipsis']")
flights= []
for i in flight_name:
    a = i.get_attribute('innerText')
    print(a)
    flights.append(a)
print(flights)
```

This is to flight data.

```
dept_time=driver.find_elements_by_xpath("//div[@class='i-b col-4 no-wrap text-right dtime col-3']")
dept= []
for i in dept_time:
    a = i.get_attribute('innerText')
    print(a)
    dept.append(a)
print(dept)
```

This is for departure time

```
arr_time=driver.find_elements_by_xpath("//div[@class='i-b pdd-0 text-left atime col-5']")
arr= []
for i in arr_time:
    a = i.get_attribute('innerText')
    print(a)
    arr.append(a)
print(arr)
```

This is for arrival time.

```
duration_time=driver.find_elements_by_xpath("//div[@class='stop-cont pl-13']")
duration= []
for i in duration_time:
    a = i.get_attribute('innerText')
    print(a)
    duration.append(a)
```

```
print(duration)
```

This is for duration

```
price_details=driver.find_elements_by_xpath("//div[@class='i-b tipsy fare-summary-tooltip fs-18']")
price= []
for i in price_details:
    a = i.get_attribute('innerText')
    print(a)
    price.append(a)
print(price)
```

This is to get price details

```
Flight=pd.DataFrame({})
Flight['Airline Name']=flights
Flight['Dept Time']=dept
Flight['Arr Time']=arr
Flight['Duration']=duration
Flight['Price']=price
Flight.head()
```

This is for crating data frame

```
df=Flight
df.index = df.index+1
df.to_csv('Flightdetails1.csv', index=False)
```

This is how I put data on excel

After getting all the data I put all together on one single excel.

After this I put this one single data frame for analysis data.

Model/s Development and Evaluation

First, we need to put al file to a single data frame

```
df = pd.DataFrame(pd.read_excel("Flightdetails.xlsx"))
```

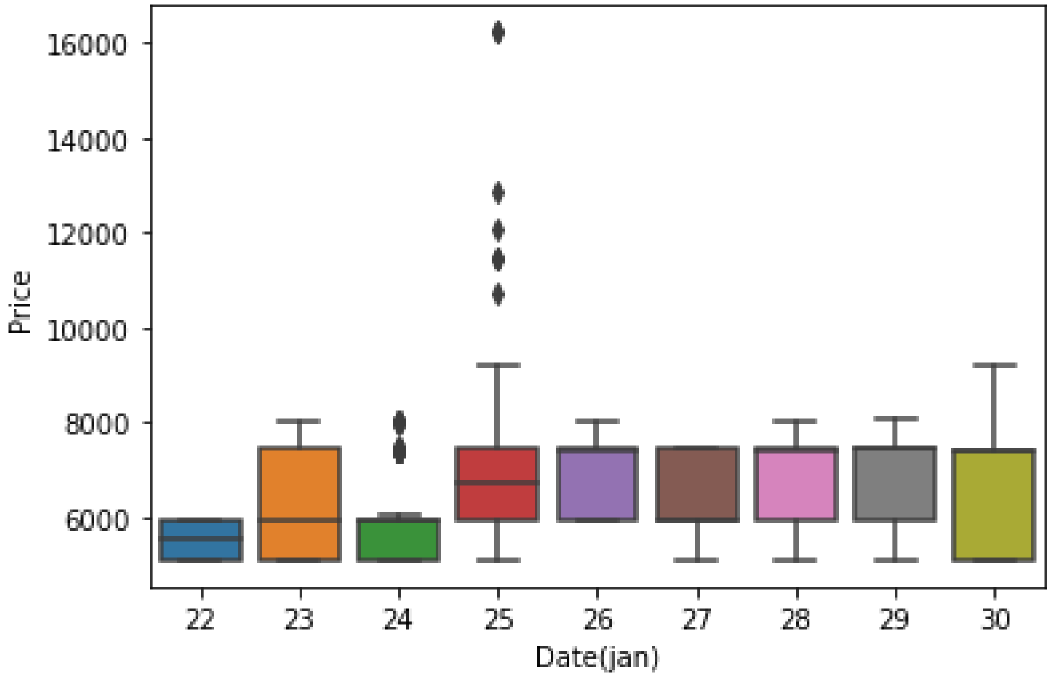
We have total 1513 rows and 9 columns, price columns is the target columns

Then we need to check the null value, since there is no null value, we can go ahead. As target column is int data type we will process **Regression Technique.**

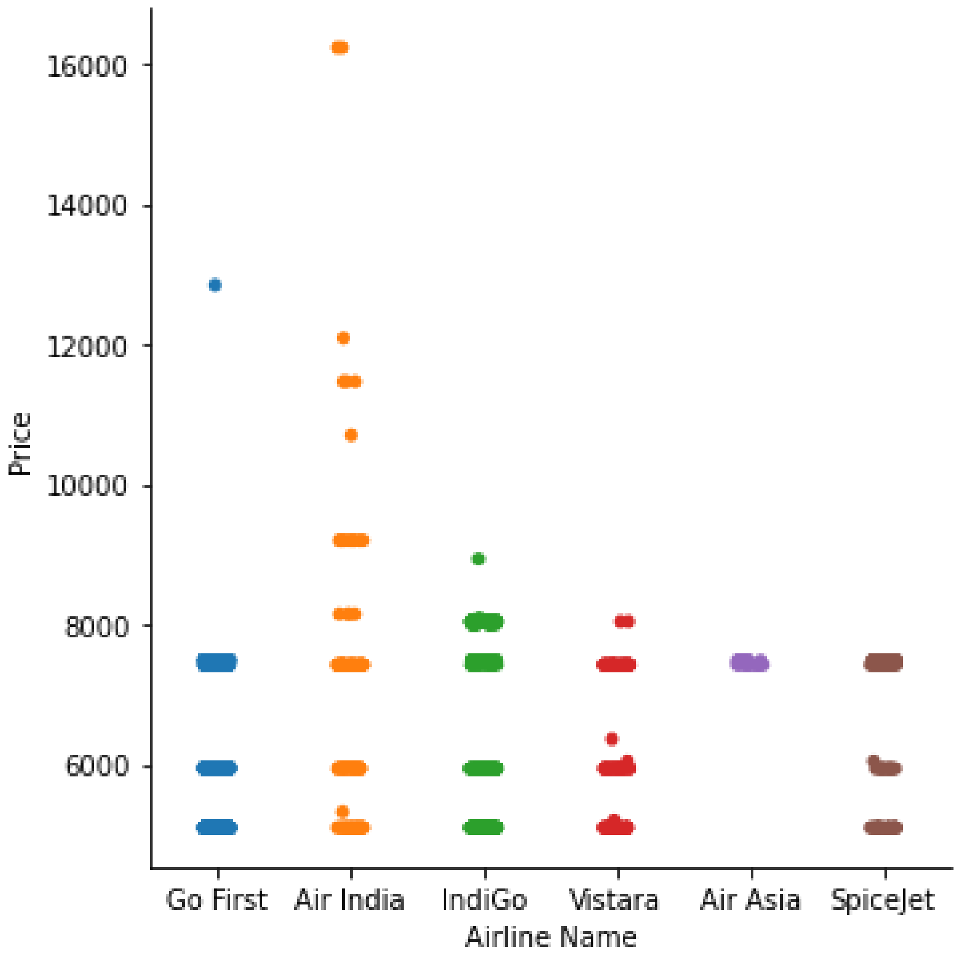
- **Visualization:- Since there is different types of data type for proper visualization we need to encode data types first**

```
df['Airline Name']=lencode.fit_transform(df['Airline Name'])
df['Source']=lencode.fit_transform(df['Source'])
df['Destination']=lencode.fit_transform(df['Destination'])
df['Duration']=lencode.fit_transform(df['Duration'])
df['Dept Time']=lencode.fit_transform(df['Dept Time'])
```

Date Vs Price

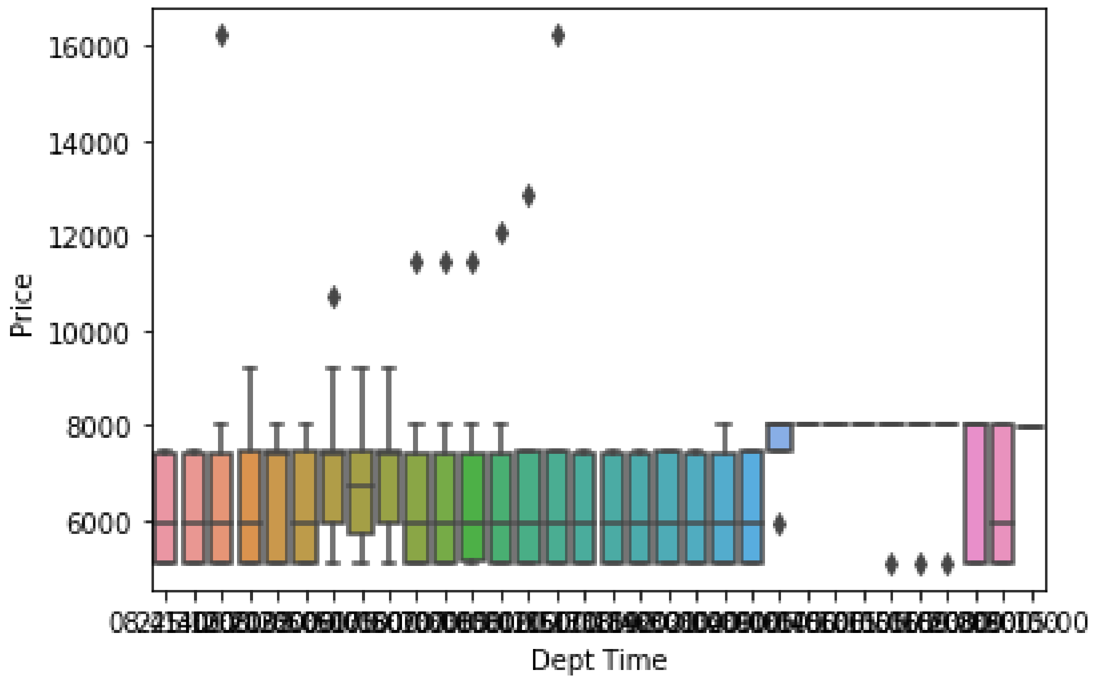


The visualization proof that price is vary on date, price might be increase on holydays.
Airline name vs Price

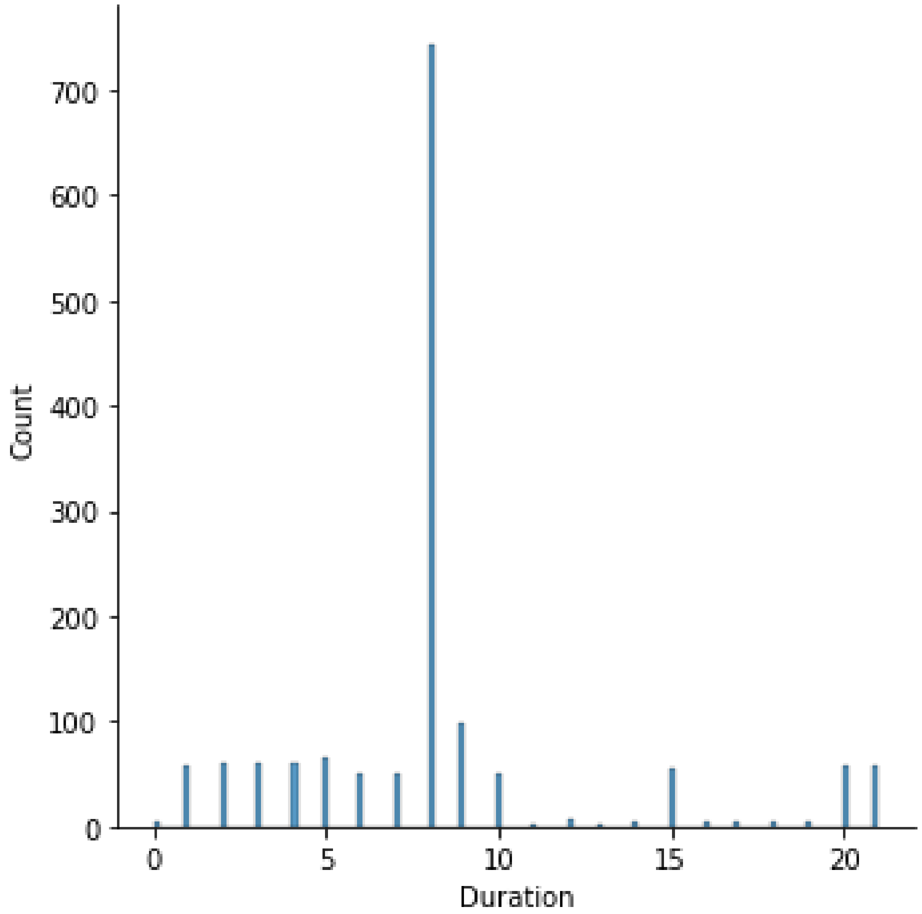


Price is vary on different airline as we can see Spicejet , Go First, Air Asia has cheaper price than others.

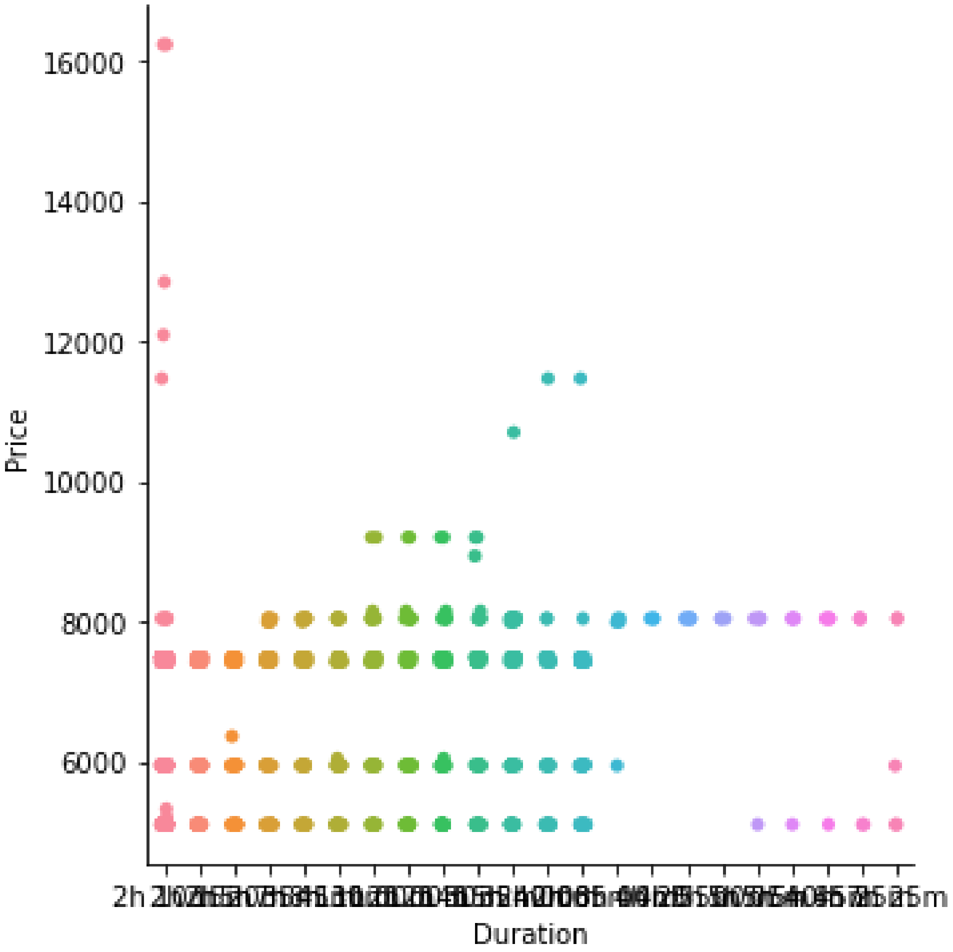
Departure Time vs Price



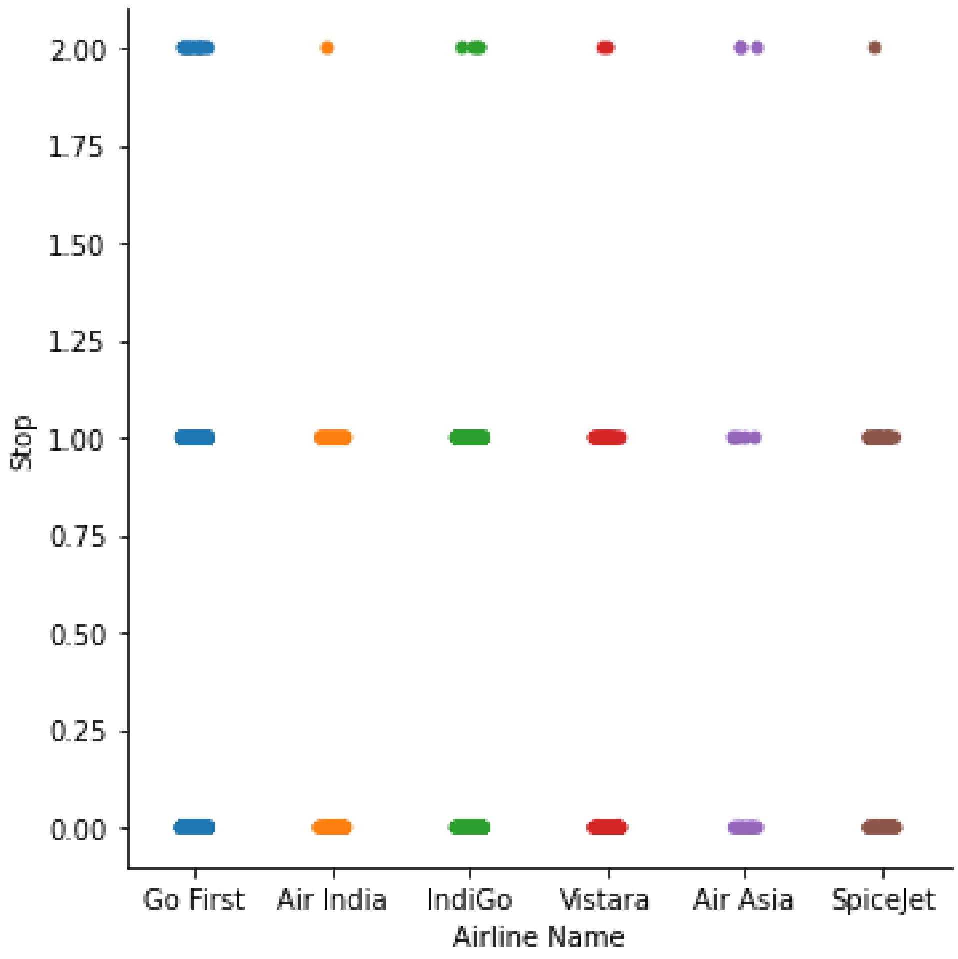
As we can see in night time price is little higher side.
Now some interesting plot bye considering this pandemic situation.



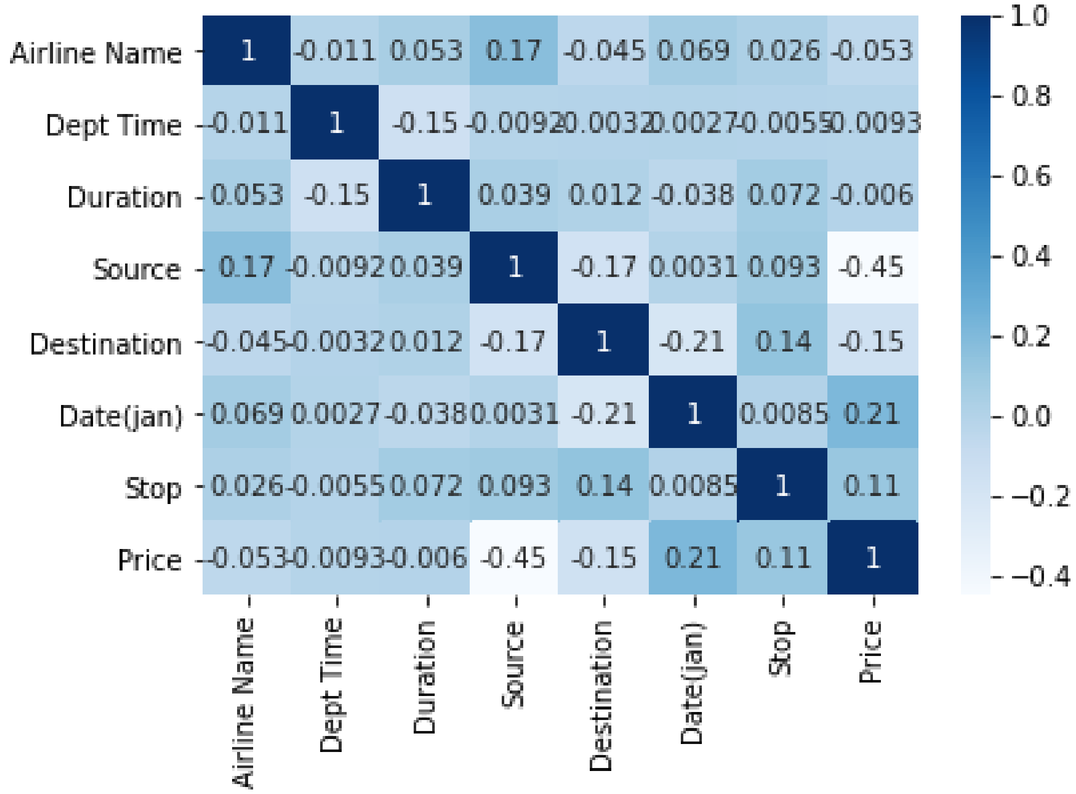
If you can see the duration plot you can see the delay of airline part some time its delay almost 1 day which is the reason of price vary.
Which we can see on next plot.



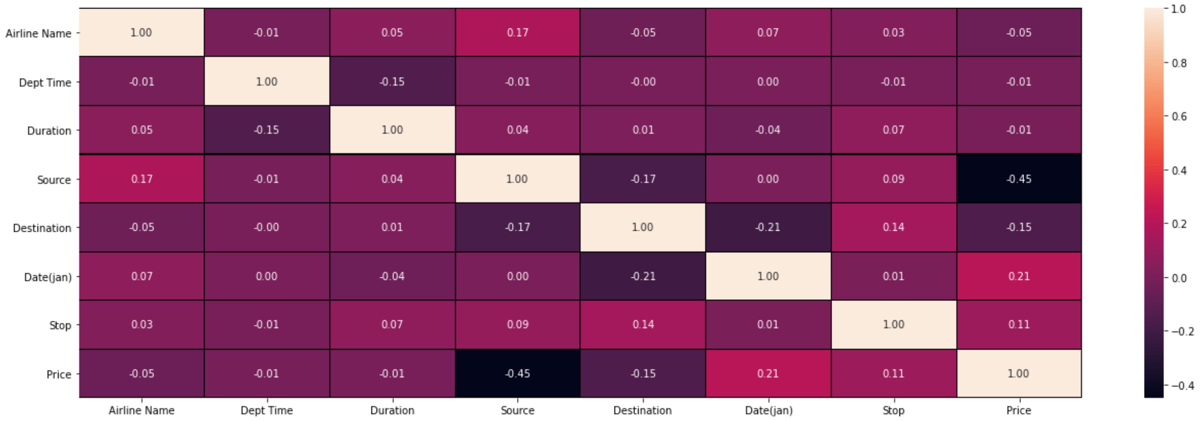
Now next plot we will se airline name vs stop.



This delay things has lots of effect on my result also. The price different and duration different is too much which can not give higher percentage model.



	Airline Name	Dept Time	Duration	Source	Destination	Date(jan)	Stop	Price
Airline Name	1	-0.011	0.053	0.17	-0.045	0.069	0.026	-0.053
Dept Time	-0.011	1	-0.15	-0.0092	-0.0032	0.0027	-0.0055	-0.0093
Duration	0.053	-0.15	1	0.039	0.012	-0.038	0.072	-0.006
Source	0.17	-0.0092	0.039	1	-0.17	0.0031	0.093	-0.45
Destination	-0.045	-0.0032	0.012	-0.17	1	-0.21	0.14	-0.15
Date(jan)	0.069	0.0027	-0.038	0.0031	-0.21	1	0.0085	0.21
Stop	0.026	-0.0055	0.072	0.093	0.14	0.0085	1	0.11
Price	-0.053	-0.0093	-0.006	-0.45	-0.15	0.21	0.11	1



	Airline Name	Dept Time	Duration	Source	Destination	Date(jan)	Stop	Price
Airline Name	1.00	-0.01	0.05	0.17	-0.05	0.07	0.03	-0.05
Dept Time	-0.01	1.00	-0.15	-0.01	-0.00	0.00	-0.01	-0.01
Duration	0.05	-0.15	1.00	0.04	0.01	-0.04	0.07	-0.01
Source	0.17	-0.01	0.04	1.00	-0.17	0.00	0.09	-0.45
Destination	-0.05	-0.00	0.01	-0.17	1.00	-0.21	0.14	-0.15
Date(jan)	0.07	0.00	-0.04	0.00	-0.21	1.00	0.01	0.21
Stop	0.03	-0.01	0.07	0.09	0.14	0.01	1.00	0.11
Price	-0.05	-0.01	-0.01	-0.45	-0.15	0.21	0.11	1.00

If you see the heatmap you will understand the effect,

I only got 31% results for Lasso , linear, and Ridge. Svr Technique I got only 24% and enr technique 31%.

After ensemble technique of best model give 62% results.

CONCLUSION

- All though model only give 62% but bye considering this delaying of flight the result is quite good.
- We can get better result if remove this stop.
- Also, for getting emergency time flight price I done some assumption which is also effect results.

So, this model can get better output on different date. Where this delay time will not present.

You can get total project on https://github.com/bishwa2017/Flight-Price-Prediction

Thank you