MrBriit

```
In [97]: !pip install chart_studio
```

```
Requirement already satisfied: chart_studio in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (1.1.0)
Requirement already satisfied: requests in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from chart_studio) (2.24.0)
Requirement already satisfied: plotly in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (fr
om chart_studio) (4.8.1)
Requirement already satisfied: retrying>=1.3.3 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
kages (from chart_studio) (1.3.3)
Requirement already satisfied: six in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (from
chart_studio) (1.14.0)
Requirement already satisfied: certifi>=2017.4.17 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-
packages (from requests->chart_studio) (2019.11.28)
Requirement already satisfied: idna<3,>=2.5 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from requests->chart_studio) (2.8)
Requirement already satisfied: urllib3!=1.25.0,!=1.25.1,<1.26,>=1.21.1 in /Users/mybeast/opt/anaconda
3/lib/python3.7/site-packages (from requests->chart_studio) (1.25.8)
Requirement already satisfied: chardet<4,>=3.0.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
ackages (from requests->chart_studio) (3.0.4)
```

```
In [98]: !pip install cufflinks
```

```
Requirement already satisfied: cufflinks in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(0.17.3)
Requirement already satisfied: six>=1.9.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from cufflinks) (1.14.0)
Requirement already satisfied: ipywidgets>=7.0.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
ackages (from cufflinks) (7.5.1)
Requirement already satisfied: colorlover>=0.2.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
ackages (from cufflinks) (0.3.0)
Requirement already satisfied: setuptools>=34.4.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-
packages (from cufflinks) (46.0.0.post20200309)
Requirement already satisfied: numpy>=1.9.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from cufflinks) (1.18.1)
Requirement already satisfied: plotly>=4.1.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packa
ges (from cufflinks) (4.8.1)
Requirement already satisfied: ipython>=5.3.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pack
ages (from cufflinks) (7.12.0)
Requirement already satisfied: pandas>=0.19.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pack
ages (from cufflinks) (1.0.5)
Requirement already satisfied: ipykernel>=4.5.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pa
ckages (from ipywidgets>=7.0.0->cufflinks) (5.1.4)
Requirement already satisfied: widgetsnbextension~=3.5.0 in /Users/mybeast/opt/anaconda3/lib/python3.
7/site-packages (from ipywidgets>=7.0.0->cufflinks) (3.5.1)
Requirement already satisfied: traitlets>=4.3.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pa
ckages (from ipywidgets>=7.0.0->cufflinks) (4.3.3)
Requirement already satisfied: nbformat>=4.2.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
kages (from ipywidgets>=7.0.0->cufflinks) (5.0.4)
Requirement already satisfied: retrying>=1.3.3 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
kages (from plotly>=4.1.1->cufflinks) (1.3.3)
Requirement already satisfied: jedi>=0.10 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from ipython>=5.3.0->cufflinks) (0.14.1)
Requirement already satisfied: backcall in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from ipython>=5.3.0->cufflinks) (0.1.0)
Requirement already satisfied: prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0 in /Users/mybeast/opt/ana
conda3/lib/python3.7/site-packages (from ipython>=5.3.0->cufflinks) (3.0.3)
Requirement already satisfied: pickleshare in /Users/mybeast/opt/anaconda3/lib/python3.7/site-package
s (from ipython>=5.3.0->cufflinks) (0.7.5)
Requirement already satisfied: pexpect; sys_platform != "win32" in /Users/mybeast/opt/anaconda3/lib/p
ython3.7/site-packages (from ipython>=5.3.0->cufflinks) (4.8.0)
Requirement already satisfied: appnope; sys_platform == "darwin" in /Users/mybeast/opt/anaconda3/lib/
python3.7/site-packages (from ipython>=5.3.0->cufflinks) (0.1.0)
Requirement already satisfied: pygments in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from ipython>=5.3.0->cufflinks) (2.5.2)
Requirement already satisfied: decorator in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from ipython>=5.3.0->cufflinks) (4.4.1)
Requirement already satisfied: python-dateutil>=2.6.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/s
ite-packages (from pandas>=0.19.2->cufflinks) (2.8.1)
Requirement already satisfied: pytz>=2017.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from pandas>=0.19.2->cufflinks) (2019.3)
Requirement already satisfied: jupyter-client in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pack
ages (from ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (5.3.4)
Requirement already satisfied: tornado>=4.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (6.0.3)
Requirement already satisfied: notebook>=4.4.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
kages (from widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (6.0.3)
Requirement already satisfied: ipython-genutils in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pa
ckages (from traitlets>=4.3.1->ipywidgets>=7.0.0->cufflinks) (0.2.0)
Requirement already satisfied: jsonschema!=2.5.0,>=2.4 in /Users/mybeast/opt/anaconda3/lib/python3.7/
site-packages (from nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (3.2.0)
Requirement already satisfied: parso>=0.5.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from jedi>=0.10->ipython>=5.3.0->cufflinks) (0.5.2)
Requirement already satisfied: wcwidth in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (f
rom prompt-toolkit!=3.0.0,!=3.0.1,<3.1.0,>=2.0.0->ipython>=5.3.0->cufflinks) (0.1.8)
Requirement already satisfied: ptyprocess>=0.5 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
kages (from pexpect; sys_platform != "win32"->ipython>=5.3.0->cufflinks) (0.6.0)
Requirement already satisfied: pyzmq>=13 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from jupyter-client->ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (18.1.1)
Requirement already satisfied: jupyter-core>=4.6.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site
-packages (from jupyter-client->ipykernel>=4.5.1->ipywidgets>=7.0.0->cufflinks) (4.6.1)
Requirement already satisfied: Send2Trash in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (1.5.0)
Requirement already satisfied: terminado>=0.8.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pa
ckages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.8.3)
Requirement already satisfied: jinja2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (fr
om notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (2.11.1)
Requirement already satisfied: prometheus-client in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
ackages (from notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.7.1)
Requirement already satisfied: attrs>=17.4.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packa
ges (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (19.3.0)
Requirement already satisfied: importlib-metadata; python_version < "3.8" in /Users/mybeast/opt/anaco
nda3/lib/python3.7/site-packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->c
ufflinks) (1.5.0)
Requirement already satisfied: pyrsistent>=0.14.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-
packages (from jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidgets>=7.0.0->cufflinks) (0.15.7)
Requirement already satisfied: testpath in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.4.4)
Requirement already satisfied: entrypoints>=0.2.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-
packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(0.3)
Requirement already satisfied: pandocfilters>=1.4.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/sit
e-packages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(1.4.2)
Requirement already satisfied: bleach in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (fr
om nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (3.1.0)
Requirement already satisfied: defusedxml in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (0.6.0)
Requirement already satisfied: mistune<2,>=0.8.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
ackages (from nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(0.8.4)
Requirement already satisfied: MarkupSafe>=0.23 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pa
ckages (from jinja2->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks) (1.1.
1)
Requirement already satisfied: zipp>=0.5 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
(from importlib-metadata; python_version < "3.8"->jsonschema!=2.5.0,>=2.4->nbformat>=4.2.0->ipywidget
s>=7.0.0->cufflinks) (2.2.0)
Requirement already satisfied: webencodings in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
es (from bleach->nbconvert->notebook>=4.4.1->widgetsnbextension~=3.5.0->ipywidgets>=7.0.0->cufflinks)
(0.5.1)
```

```
In [99]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         from chart_studio.plotly import plot,iplot
         import cufflinks as cf
         import seaborn as sns

         %matplotlib inline
```

```
In [100]: df=pd.read_excel("Data_Train.xlsx")
```

```
In [101]: df.head()
```

Out[101]:

	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

Automatic EDA

```
In [102]: import pandas_profiling as pf
```

```
In [103]: pf.ProfileReport(df)
```

Pandas Profiling Report

Overview Variables Interactions Correlations Missing values Sample Duplicate rows

Overview

Overview Reproduction Warnings 2

Dataset statistics

Number of variables	11
Number of observations	10683
Missing cells	2
Missing cells (%)	< 0.1%
Duplicate rows	220
Duplicate rows (%)	2.1%
Total size in memory	918.2 KiB
Average record size in memory	88.0 B

Variable types

CAT	10
NUM	1

Variables

Airline
Categorical

Distinct count 12

Jet Airways	3849
IndiGo	2053
Air India	1752

Out[103]:

Manul EDA

```
In [104]: df.shape
```

```
Out[104]: (10683, 11)
```

```
In [105]: df.dtypes #checking the data types
```

```
Out[105]: Airline            object
          Date_of_Journey    object
          Source             object
          Destination        object
          Route              object
          Dep_Time           object
          Arrival_Time       object
          Duration           object
          Total_Stops        object
          Additional_Info    object
          Price               int64
          dtype: object
```

```
In [106]: df.isna().sum() #Checking null values
```

```
Out[106]: Airline            0
          Date_of_Journey    0
          Source             0
          Destination        0
          Route              1
          Dep_Time           0
          Arrival_Time       0
          Duration           0
          Total_Stops        1
          Additional_Info    0
          Price              0
          dtype: int64
```

```
In [107]: #Remove the NaN values from the dataset
          df.dropna(how='any',inplace=True)
          df.isnull().sum()
```

```
Out[107]: Airline            0
          Date_of_Journey    0
          Source             0
          Destination        0
          Route              0
          Dep_Time           0
          Arrival_Time       0
          Duration           0
          Total_Stops        0
          Additional_Info    0
          Price              0
          dtype: int64
```

Feature Engineering

Let's convert **Date_of_Journey** to its appropriate format as **datetime** with regards to **day** and **month**

```
In [108]: df.head(3)
```

Out[108]:

	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

Date_of_journey

```
In [109]: df['Date_of_Journey']=pd.to_datetime(df['Date_of_Journey'])
          df['Day_of_Journey']=(df['Date_of_Journey']).dt.day
          df['Month_of_Journey']=(df['Date_of_Journey']).dt.month
```

```
In [110]: df.head(3)
```

Out[110]:

	Airline	Date_of_Journey	Source	Destination	Route	Dep_Time	Arrival_Time	Duration	Total_Stops	Additional_Info	Price	Day_of_
0	IndiGo	2019-03-24	Banglore	New Delhi	BLR → DEL	22:20	01:10 22 Mar	2h 50m	non-stop	No info	3897	
1	Air India	2019-01-05	Kolkata	Banglore	CCU → IXR → BBI → BLR	05:50	13:15	7h 25m	2 stops	No info	7662	
2	Jet Airways	2019-09-06	Delhi	Cochin	DEL → LKO → BOM → COK	09:25	04:25 10 Jun	19h	2 stops	No info	13882	

```
In [111]: #we can now drop the Date_of_Journey
          df.drop(["Date_of_Journey"],axis=1,inplace=True)
```

Dep_time

```
In [112]: #convert to datetime
          df['Dep_hr']=pd.to_datetime(df['Dep_Time']).dt.hour
          df['Dep_min']=pd.to_datetime(df['Dep_Time']).dt.minute
```

```
In [113]: #we can now drop the 'Dep_Time'

          df.drop(["Dep_Time"],axis=1,inplace=True)
```

Arrival_time

```
In [114]: df['Arrival_hr']=pd.to_datetime(df['Arrival_Time']).dt.hour
          df['Arrival_min']=pd.to_datetime(df['Arrival_Time']).dt.minute
```

```
In [115]: #we can now drop the 'Arrival_Time'

          df.drop(["Arrival_Time"],axis=1,inplace=True)
```

Duration Time

```
In [116]: duration=df["Duration"].str.split(' ',expand=True) #split duration datapoints based space ' '
          duration[1].fillna('00m',inplace=True)  #fill all 'NAN' with '00m'
          df['duration_hr']=duration[0].apply(lambda x: x[:-1]) #select the item at index o and leave the last o
          ne in this case the 'h')
          df['duration_min']=duration[1].apply(lambda x: x[:-1]) #select the item at index 1 and leave the last
           one in this case the 'm')
```

```
In [117]: #we can now drop the 'Duration'

          df.drop(["Duration"],axis=1,inplace=True)
```

```
In [118]: df.head(3)
```

Out[118]:

	Airline	Source	Destination	Route	Total_Stops	Additional_Info	Price	Day_of_Journey	Month_of_Journey	Dep_hr	Dep_min	Arriva
0	IndiGo	Banglore	New Delhi	BLR → DEL	non-stop	No info	3897	24	3	22	20	
1	Air India	Kolkata	Banglore	CCU → IXR → BBI → BLR	2 stops	No info	7662	5	1	5	50	
2	Jet Airways	Delhi	Cochin	DEL → LKO → BOM → COK	2 stops	No info	13882	6	9	9	25	

```
In [119]: cf.set_config_file(theme='ggplot',sharing='public',offline=True)
```

Count of Airlines

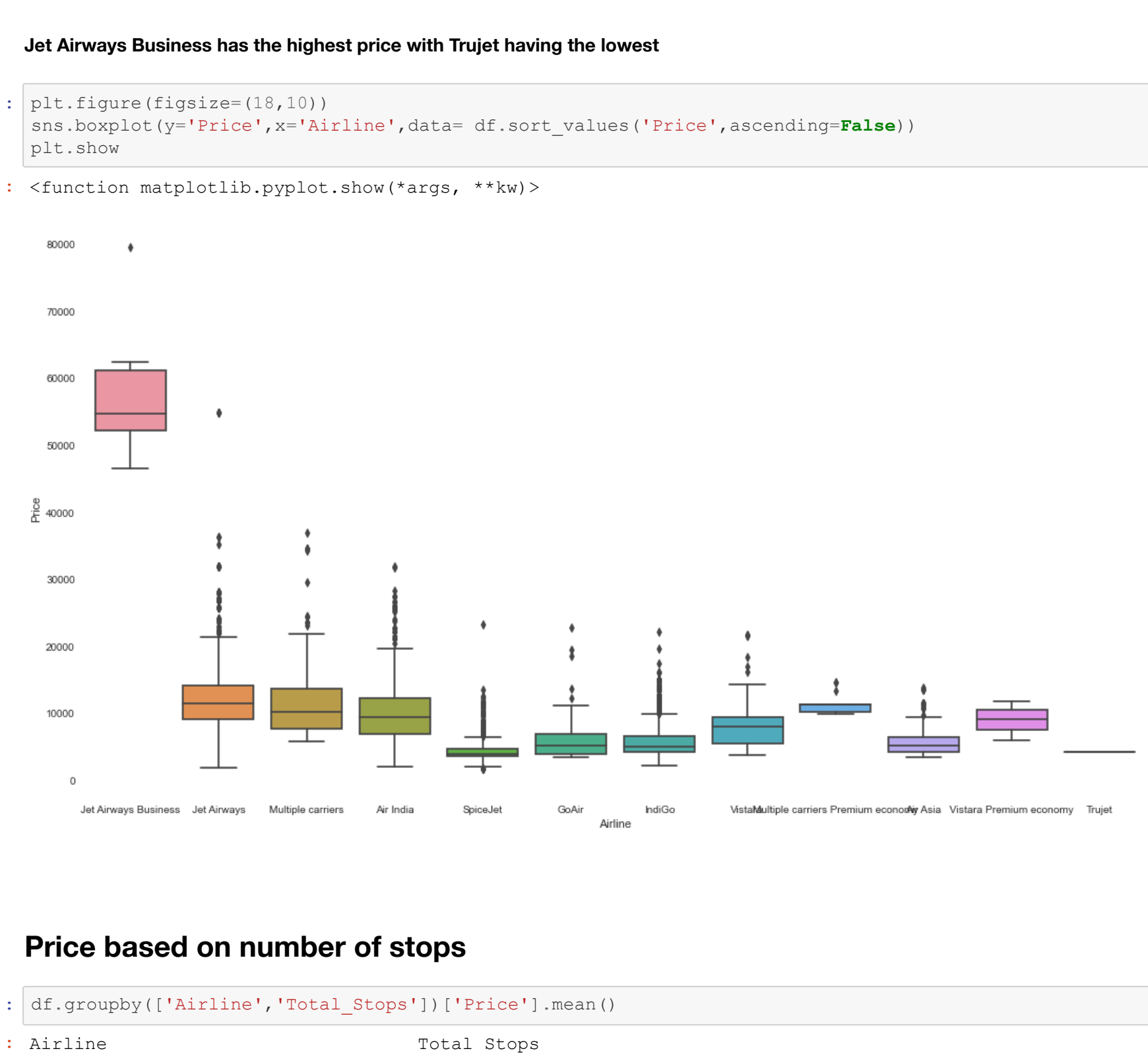
```
In [120]: df['Airline'].value_counts()
```

```
Out[120]: Jet Airways                          3849
          IndiGo                               2053
          Air India                            1751
          Multiple carriers                    1196
          SpiceJet                              818
          Vistara                               479
          Air Asia                              319
          GoAir                                 194
          Multiple carriers Premium economy      13
          Jet Airways Business                    6
          Vistara Premium economy                 3
          Trujet                                  1
          Name: Airline, dtype: int64
```

Airline vs Price

```
In [121]: Airprices=df.groupby('Airline')['Price'].mean().sort_values(ascending=False)
          plt.figure(figsize=(15,10))
          sns.barplot(Airprices.index,Airprices.values)
          plt.xticks(rotation=270)
```

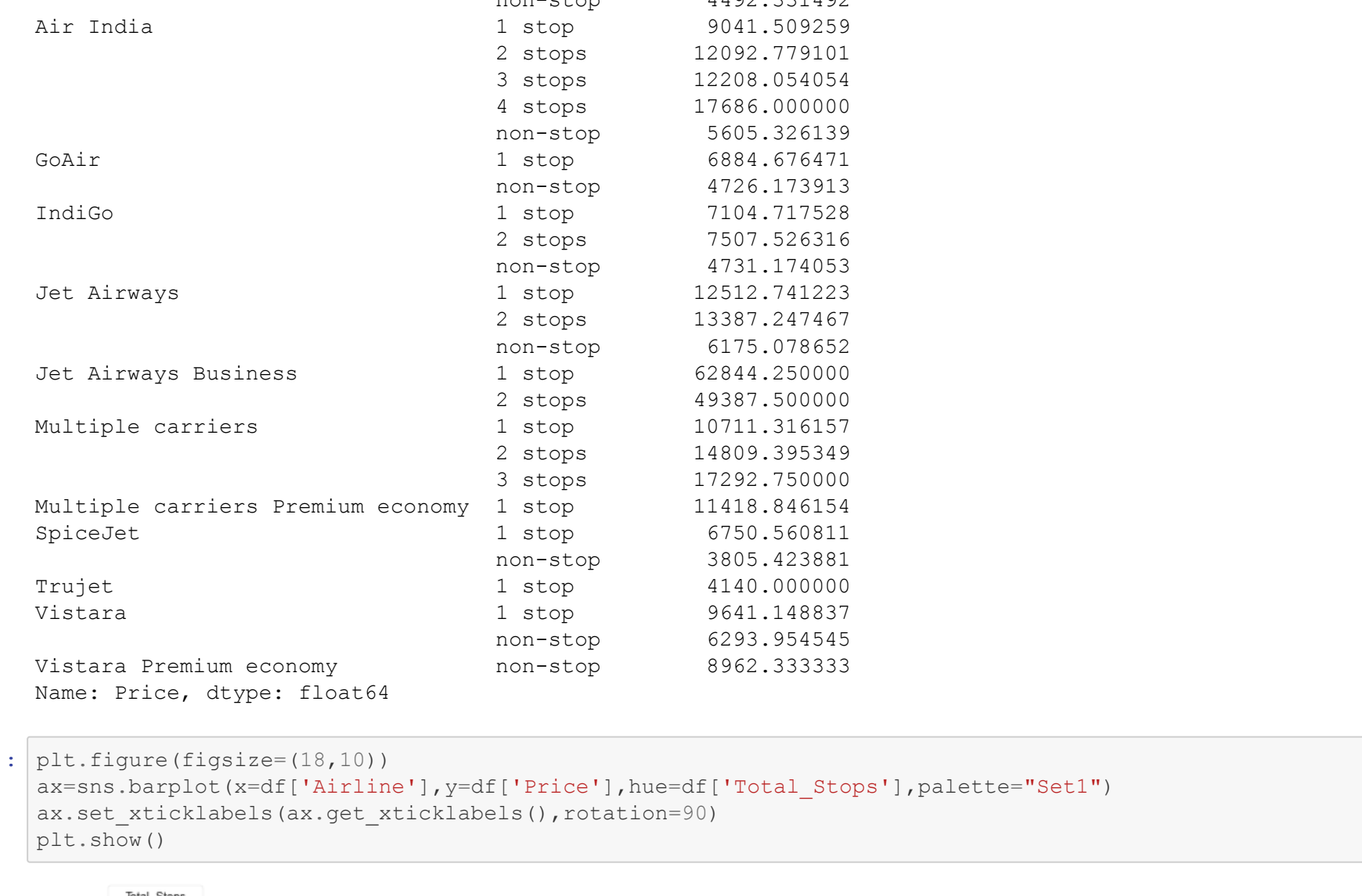
```
Out[121]: (array([ 0,  1,  2,  3,  4,  5,  6,  7,  8,  9, 10, 11]),
           <a list of 12 Text major ticklabel objects>)
```



Jet Airways Business has the highest price with Trujet having the lowest

```
In [122]: plt.figure(figsize=(18,10))
          sns.boxplot(y='Price',x='Airline',data= df.sort_values('Price',ascending=False))
          plt.show
```

```
Out[122]: <function matplotlib.pyplot.show(*args, **kw)>
```

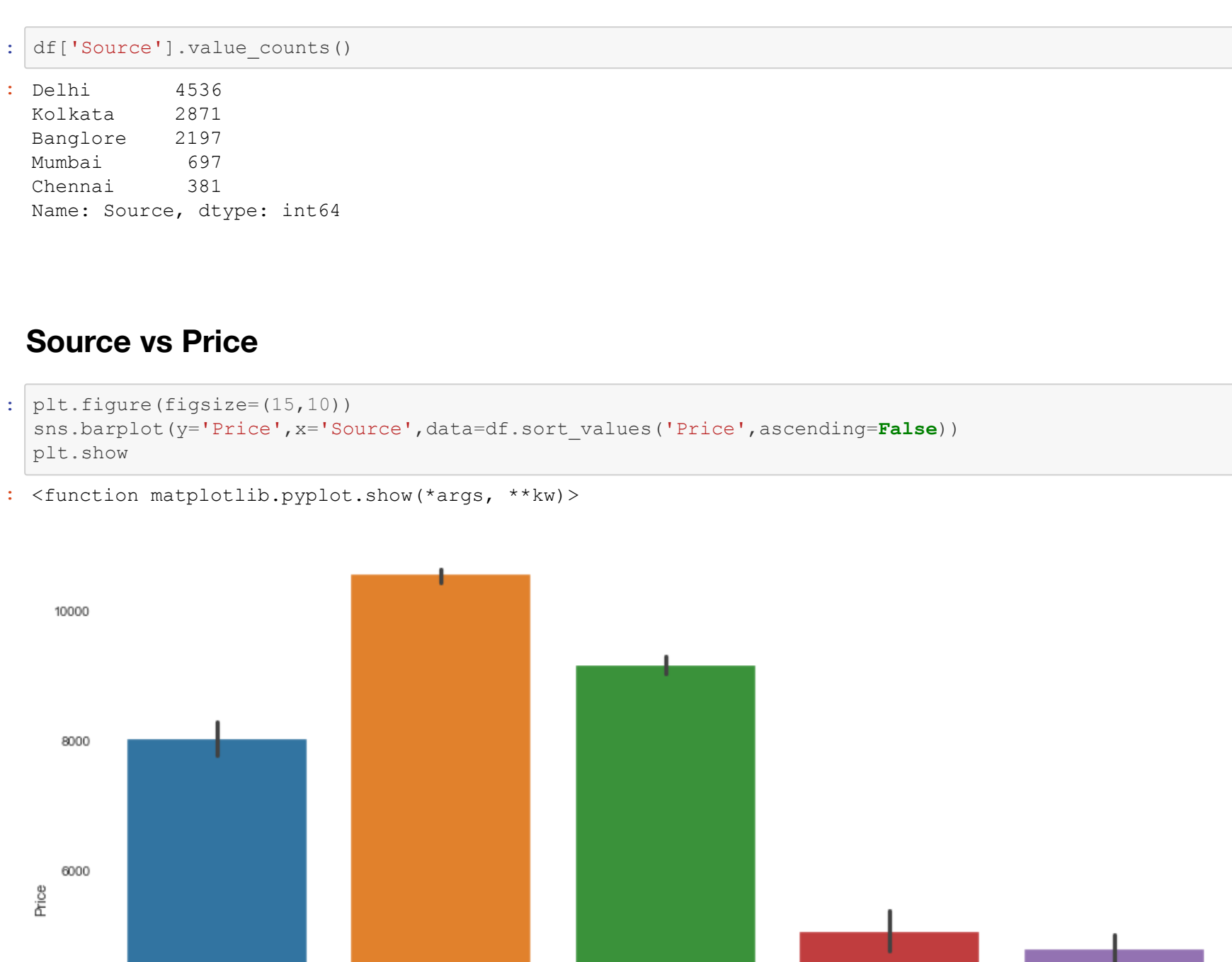


Price based on number of stops

```
In [123]: df.groupby(['Airline','Total_Stops'])['Price'].mean()
```

```
Out[123]: Airline                            Total_Stops
          Air Asia                           1 stop          7078.325581
                                             2 stops         6341.888889
                                             non-stop        4492.331492
          Air India                          1 stop          9041.509259
                                             2 stops        12092.779101
                                             3 stops        12208.054054
                                             4 stops        17686.000000
                                             non-stop        5605.326139
          GoAir                              1 stop          6884.676471
                                             non-stop        4726.173913
          IndiGo                             1 stop          7104.717528
                                             2 stops         7507.526316
                                             non-stop        4731.174053
          Jet Airways                        1 stop         12512.741223
                                             2 stops        13387.247467
                                             non-stop        6175.078652
          Jet Airways Business               1 stop         62844.250000
                                             2 stops        49387.500000
          Multiple carriers                  1 stop         10711.316157
                                             2 stops        14809.395349
                                             3 stops        17292.750000
          Multiple carriers Premium economy  1 stop         11418.846154
          SpiceJet                           1 stop          6750.560811
                                             non-stop        3805.423881
          Trujet                             1 stop          4140.000000
          Vistara                            1 stop          9641.148837
                                             non-stop        6293.954545
          Vistara Premium economy            non-stop        8962.333333
          Name: Price, dtype: float64
```

```
In [124]: plt.figure(figsize=(18,10))
          ax=sns.barplot(x=df['Airline'],y=df['Price'],hue=df['Total_Stops'],palette="Set1")
          ax.set_xticklabels(ax.get_xticklabels(),rotation=90)
          plt.show()
```



One stop and two stpos Jet Airways Business is having the highest price

```
In [125]: df['Source'].value_counts()
```

```
Out[125]: Delhi       4536
          Kolkata     2871
          Banglore    2197
          Mumbai       697
          Chennai      381
          Name: Source, dtype: int64
```

Source vs Price

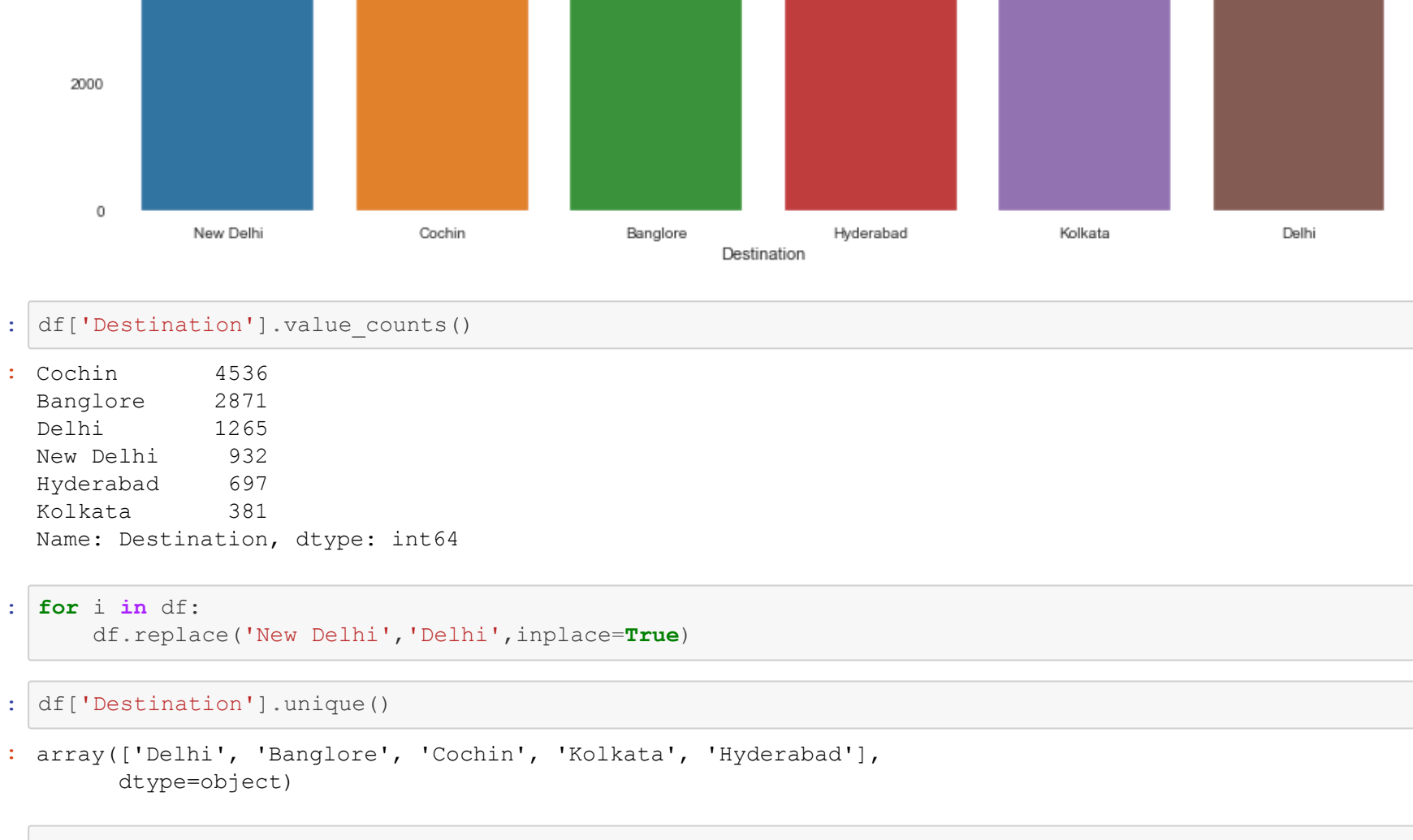
```
In [126]: plt.figure(figsize=(15,10))
          sns.barplot(y='Price',x='Source',data=df.sort_values('Price',ascending=False))
          plt.show
```

```
Out[126]: <function matplotlib.pyplot.show(*args, **kw)>
```

Destination vs Price

```
In [127]: plt.figure(figsize=(15,10))
          sns.barplot(y='Price',x='Destination',data=df.sort_values('Price',ascending=False))
          plt.show
```

```
Out[127]: <function matplotlib.pyplot.show(*args, **kw)>
```




```
In [128]: df['Destination'].value_counts()
```

```
Out[128]: Cochin       4536
          Banglore     2871
          Delhi        1265
          New Delhi     932
          Hyderabad     697
          Kolkata       381
          Name: Destination, dtype: int64
```

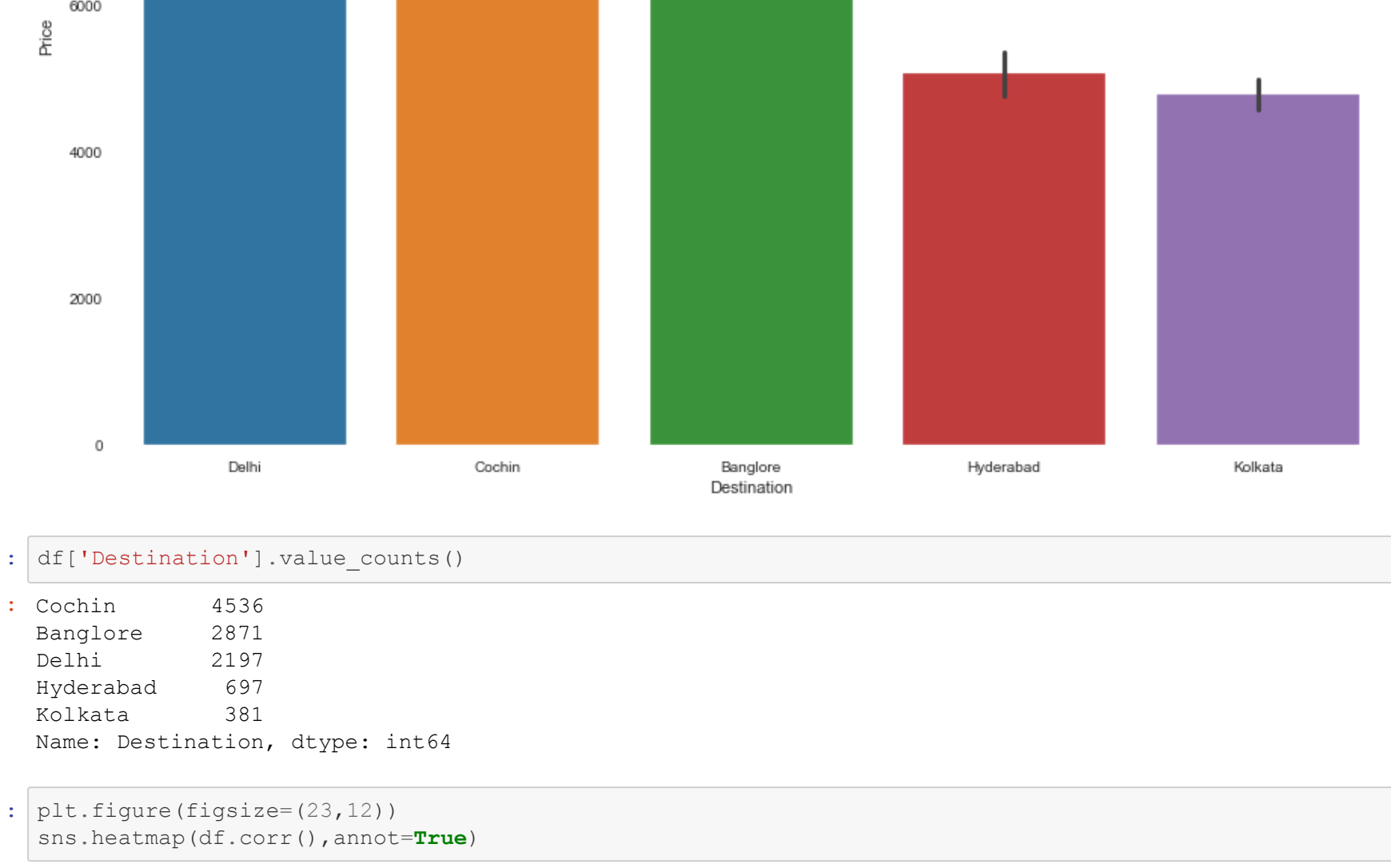
```
In [129]: for i in df:
              df.replace('New Delhi','Delhi',inplace=True)
```

```
In [130]: df['Destination'].unique()
```

```
Out[130]: array(['Delhi', 'Banglore', 'Cochin', 'Kolkata', 'Hyderabad'],
                dtype=object)
```

```
In [131]: plt.figure(figsize=(15,10))
          sns.barplot(y='Price',x='Destination',data=df.sort_values('Price',ascending=False))
          plt.show
```

```
Out[131]: <function matplotlib.pyplot.show(*args, **kw)>
```

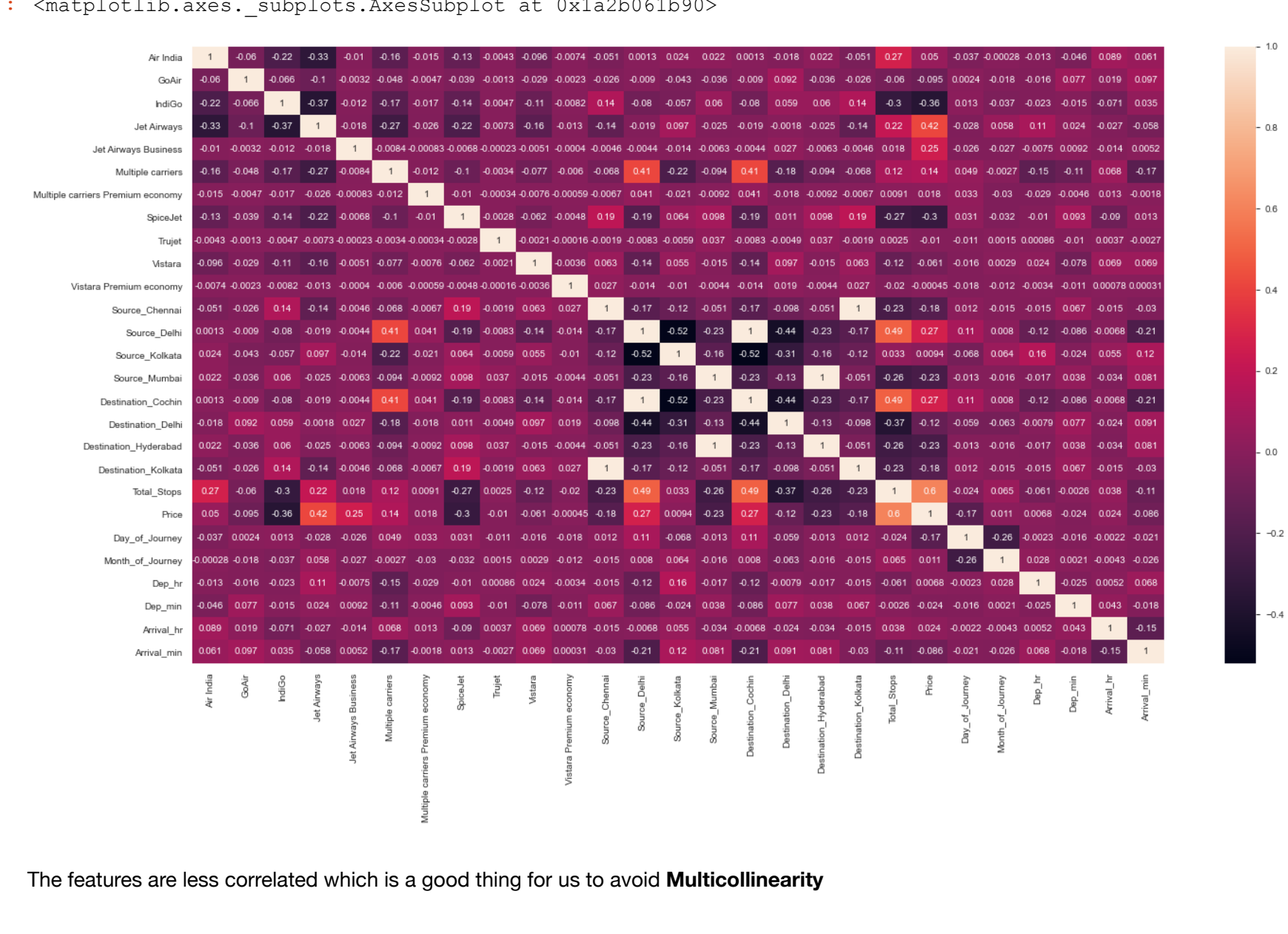



```
In [132]: df['Destination'].value_counts()
```

```
Out[132]: Cochin       4536
          Banglore     2871
          Delhi        2197
          Hyderabad     697
          Kolkata       381
          Name: Destination, dtype: int64
```

```
In [149]: plt.figure(figsize=(23,12))
          sns.heatmap(df.corr(),annot=True)
```

```
Out[149]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2b061b90>
```



The features are less correlated which is a good thing for us to avoid **Multicollinearity**

Apply Onehot Encoding on the categorical variables

```
In [133]: #Handling Categorical Values
          df['Total_Stops']=df['Total_Stops'].map({'non-stop':0,'2 stops':2,'1 stop':1,'3 stops':3,'4 stops'
          :4})
          air_dummy=pd.get_dummies(df['Airline'],drop_first=True)
          # print(df['Source'].value_counts())
          # print(df['Destination'].value_counts())

          source_dest_dummy=pd.get_dummies(df[['Source','Destination']],drop_first=True)
          df=pd.concat([air_dummy,source_dest_dummy,df],axis=1)
```

```
In [134]: df.drop(['Airline','Source','Destination'],inplace=True,axis=1)
```

```
In [135]: df.head()
```

Out[135]:

	Air India	GoAir	IndiGo	Jet Airways	Jet Airways Business	Multiple carriers	Multiple carriers Premium economy	SpiceJet	Trujet	Vistara	...	Additional_Info	Price	Day_of_Journey	Mor
0	0	0	1	0	0	0	0	0	0	0	...	No info	3897	24	
1	1	0	0	0	0	0	0	0	0	0	...	No info	7662	5	
2	0	0	0	1	0	0	0	0	0	0	...	No info	13882	6	
3	0	0	1	0	0	0	0	0	0	0	...	No info	6218	5	
4	0	0	1	0	0	0	0	0	0	0	...	No info	13302	3	

5 rows × 31 columns

```
In [136]: df.shape
Out[136]: (10682, 31)
```

Test Data

```
In [137]: df_test=pd.read_excel("Test_set.xlsx")
```

```
In [138]: df_test.head()
```

Out[138]:

	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

Let's Perform that same steps as we did for train data earlier

```
In [139]: df_test['Date_of_Journey']=pd.to_datetime(df_test['Date_of_Journey'])
          df_test['Day_of_Journey']=(df_test['Date_of_Journey']).dt.day
          df_test['Month_of_Journey']=(df_test['Date_of_Journey']).dt.month

          #Dep_time
          df_test['Dep_hr']=pd.to_datetime(df_test['Dep_Time']).dt.hour
          df_test['Dep_min']=pd.to_datetime(df_test['Dep_Time']).dt.minute

          #Arrival_time
          df_test['Arrival_hr']=pd.to_datetime(df_test['Arrival_Time']).dt.hour
          df_test['Arrival_min']=pd.to_datetime(df_test['Arrival_Time']).dt.minute

          #Splitting duration  time

          a=df_test['Duration'].str.split(' ',expand=True)
          a[1].fillna('00m',inplace=True)
          df_test['dur_hr']=a[0].apply(lambda x: x[:-1])
          df_test['dur_min']=a[1].apply(lambda x: x[:-1])

          #dropping the data
          df_test.drop(['Date_of_Journey','Duration','Arrival_Time','Dep_Time'],inplace=True,axis=1)

          #Handling Categorical Values
          df_test['Total_Stops']=df_test['Total_Stops'].map({'non-stop':0, '2 stops':2, '1 stop':1, '3 stops':3,
           '4 stops':4})

          air_dummy=pd.get_dummies(df_test['Airline'],drop_first=True)
          source_dest_dummy=pd.get_dummies(df_test[['Source','Destination']],drop_first=True)
          df_test=pd.concat([air_dummy,source_dest_dummy,df_test],axis=1)
```

```
In [140]: df_test.drop(['Airline','Source','Destination','Additional_Info',"Route"],inplace=True,axis=1)
```

```
In [141]: print('train_shape',df.shape)
          print('test_shape',df_test.shape)

          train_shape (10682, 31)
          test_shape (2671, 28)
```

```
In [142]: df.head()
```

Out[142]:

	Air India	GoAir	IndiGo	Jet Airways	Jet Airways Business	Multiple carriers	Multiple carriers Premium economy	SpiceJet	Trujet	Vistara	...	Additional_Info	Price	Day_of_Journey	Mor
0	0	0	1	0	0	0	0	0	0	0	...	No info	3897	24	
1	1	0	0	0	0	0	0	0	0	0	...	No info	7662	5	
2	0	0	0	1	0	0	0	0	0	0	...	No info	13882	6	
3	0	0	1	0	0	0	0	0	0	0	...	No info	6218	5	
4	0	0	1	0	0	0	0	0	0	0	...	No info	13302	3	

5 rows × 31 columns

```
In [144]: x=df.drop(['Route', 'Price','Additional_Info'],axis=1)
          y=df['Price']
```

```
In [154]: x.head(3)
```

Out[154]:

	Air India	GoAir	IndiGo	Jet Airways	Jet Airways Business	Multiple carriers	Multiple carriers Premium economy	SpiceJet	Trujet	Vistara	...	Destination_Kolkata	Total_Stops	Day_of_Jou
0	0	0	1	0	0	0	0	0	0	0	...	0	0	
1	1	0	0	0	0	0	0	0	0	0	...	0	2	
2	0	0	0	1	0	0	0	0	0	0	...	0	2	

3 rows × 28 columns

Model Building and Hyperparameter Tuning

ExtraTreesRegressor

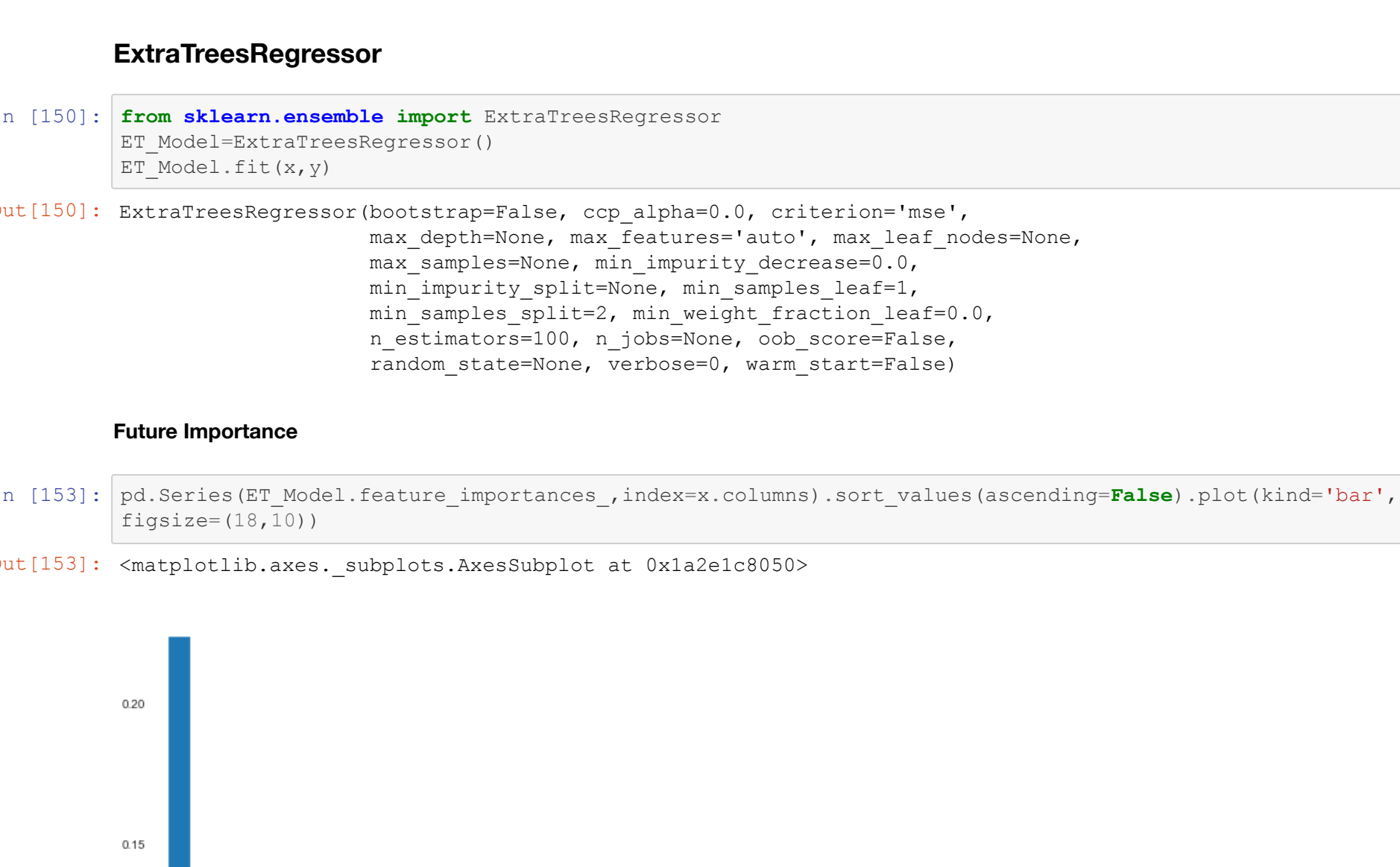
```
In [150]: from sklearn.ensemble import ExtraTreesRegressor
          ET_Model=ExtraTreesRegressor()
          ET_Model.fit(x,y)

Out[150]: ExtraTreesRegressor(bootstrap=False, ccp_alpha=0.0, criterion='mse',
                              max_depth=None, max_features='auto', max_leaf_nodes=None,
                              max_samples=None, min_impurity_decrease=0.0,
                              min_impurity_split=None, min_samples_leaf=1,
                              min_samples_split=2, min_weight_fraction_leaf=0.0,
                              n_estimators=100, n_jobs=None, oob_score=False,
                              random_state=None, verbose=0, warm_start=False)
```

Future Importance

```
In [153]: pd.Series(ET_Model.feature_importances_,index=x.columns).sort_values(ascending=False).plot(kind='bar',
          figsize=(18,10))

Out[153]: <matplotlib.axes._subplots.AxesSubplot at 0x1a2e1c8050>
```




```
In [156]: #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 = 0.2, random_state = 50)
```

```
In [157]: #Preparing Extra Tree Regression
          from sklearn.ensemble import  ExtraTreesRegressor
          ET_Model=ExtraTreesRegressor(n_estimators = 120)
          ET_Model.fit(X_train,y_train)
          y_predict=ET_Model.predict(X_test)


          from sklearn.metrics import r2_score
          r2_score(y_test,y_predict)

Out[157]: 0.7682238280133812
```

```
In [160]: from sklearn.ensemble import RandomForestRegressor
          RF_Model=RandomForestRegressor()
          RF_Model.fit(X_train,y_train)
          y_predict=RF_Model.predict(X_test)
          r2_score(y_test,y_predict)

Out[160]: 0.792954708381232
```

Hyperparameter

```
In [178]: from sklearn.model_selection import RandomizedSearchCV

          n_estimators = [int(x) for x in np.linspace(start = 80, stop = 1500, num = 10)]
          max_features = ['auto', 'sqrt']
          max_depth = [int(x) for x in np.linspace(6, 45, num = 5)]
          min_samples_split = [2, 5, 10, 15, 100]
          min_samples_leaf = [1, 2, 5, 10]

          # create random grid

          rand_grid={'n_estimators': n_estimators,
                     'max_features': max_features,
                     'max_depth': max_depth,
                     'min_samples_split': min_samples_split,
                     'min_samples_leaf': min_samples_leaf}

          rf=RandomForestRegressor()

          rCV=RandomizedSearchCV(estimator=rf,param_distributions=rand_grid,scoring='neg_mean_squared_error',n_i
          ter=10,cv=3,random_state=42, n_jobs = 1)
```

```
In [179]: rCV.fit(X_train,y_train)

Out[179]: RandomizedSearchCV(cv=3, error_score=nan,
                             estimator=RandomForestRegressor(bootstrap=True,
                                                             ccp_alpha=0.0,
                                                             criterion='mse',
                                                             max_depth=None,
                                                             max_features='auto',
                                                             max_leaf_nodes=None,
                                                             max_samples=None,
                                                             min_impurity_decrease=0.0,
                                                             min_impurity_split=None,
                                                             min_samples_leaf=1,
                                                             min_samples_split=2,
                                                             min_weight_fraction_leaf=0.0,
                                                             n_estimators=100,
                                                             n_jobs=None, oob_score=Fals...
                             iid='deprecated', n_iter=10, n_jobs=1,
                             param_distributions={'max_depth': [6, 15, 25, 35, 45],
                                                  'max_features': ['auto', 'sqrt'],
                                                  'min_samples_leaf': [1, 2, 5, 10],
                                                  'min_samples_split': [2, 5, 10, 15,
                                                                        100],
                                                  'n_estimators': [80, 237, 395, 553, 711,
                                                                   868, 1026, 1184, 1342,
                                                                   1500]},
                             pre_dispatch='2*n_jobs', random_state=42, refit=True,
                             return_train_score=False, scoring='neg_mean_squared_error',
                             verbose=0)
```

```
In [180]: rf_pred=rCV.predict(X_test)
          rf_pred

Out[180]: array([12500.2967405 ,  4286.76073494,  6136.50519908, ...,
                 11657.36838275,  9659.32444484,  8655.77083065])
```

```
In [181]: from sklearn.metrics import mean_absolute_error,mean_squared_error
          print('MAE',mean_absolute_error(y_test,rf_pred))
          print('MSE',mean_squared_error(y_test,rf_pred))

          MAE 1156.8110505706231
          MSE 3974736.1005828027
```

```
In [182]: r2_score(y_test,rf_pred)

Out[182]: 0.8064284149245494
```

```
In [189]: !pip install catboost

          Collecting catboost
            Downloading catboost-0.23.2-cp37-none-macosx_10_6_intel.macosx_10_9_intel.macosx_10_9_x86_64.macosx
          _10_10_intel.macosx_10_10_x86_64.whl (10.9 MB)
               |[redacted]| 10.9 MB 140 kB/s eta 0:00:01
          Requirement already satisfied: scipy in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (fro
          m catboost) (1.4.1)
          Requirement already satisfied: plotly in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (fr
          om catboost) (4.8.1)
          Requirement already satisfied: six in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages (from
          catboost) (1.14.0)
          Collecting graphviz
            Downloading graphviz-0.14.1-py2.py3-none-any.whl (18 kB)
          Requirement already satisfied: matplotlib in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
          (from catboost) (3.2.2)
          Requirement already satisfied: pandas>=0.24.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pack
          ages (from catboost) (1.0.5)
          Requirement already satisfied: numpy>=1.16.0 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packa
          ges (from catboost) (1.18.1)
          Requirement already satisfied: retrying>=1.3.3 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-pac
          kages (from plotly->catboost) (1.3.3)
          Requirement already satisfied: kiwisolver>=1.0.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-p
          ackages (from matplotlib->catboost) (1.1.0)
          Requirement already satisfied: pyparsing!=2.0.4,!=2.1.2,!=2.1.6,>=2.0.1 in /Users/mybeast/opt/anacond
          a3/lib/python3.7/site-packages (from matplotlib->catboost) (2.4.6)
          Requirement already satisfied: python-dateutil>=2.1 in /Users/mybeast/opt/anaconda3/lib/python3.7/sit
          e-packages (from matplotlib->catboost) (2.8.1)
          Requirement already satisfied: cycler>=0.10 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
          es (from matplotlib->catboost) (0.10.0)
          Requirement already satisfied: pytz>=2017.2 in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packag
          es (from pandas>=0.24.0->catboost) (2019.3)
          Requirement already satisfied: setuptools in /Users/mybeast/opt/anaconda3/lib/python3.7/site-packages
          (from kiwisolver>=1.0.1->matplotlib->catboost) (46.0.0.post20200309)
          Installing collected packages: graphviz, catboost
          Successfully installed catboost-0.23.2 graphviz-0.14.1
```

```
In [190]:
```