

# workshop-Multivariate-analysis

## › AIM:

To perform multivariant analysis on the given data set.

## › ALGORITHM:

- 1.Clean the data
- 2.Remove outliers
- 3.Apply the skew function and Kurtosis
- 4.Apply Bivariate analysis on numerical and categorical
- 5.Apply Multivariate analysis.

## › CODE

Name:A.ARUVI

REG NO:212222230014

## › Bivariate Analysis

```
import pandas as pd
import seaborn as sns
df=pd.read_csv("FlightInformation.csv")
df

df.isnull().sum()

df.info()

df['Route']=df['Route'].fillna(df['Route'].mode()[0])
df['Total_Stops']=df['Total_Stops'].fillna(df['Total_Stops'].mode()[0])
df.isnull().sum()

sns.scatterplot(x=df["Price"],y=df["Airline"],data=df)

sns.boxplot(x=df['Price'],y=df['Source'],data=df)

sns.barplot(x=df['Price'],y=df['Destination'],data=df)

sns.displot(df,x="Source",hue="Destination")
```

# Multivariate Analysis

```
df.corr()import numpy as np

import seaborn as sns
import matplotlib.pyplot as plt
df= pd.read_csv("FlightInformation.csv")
df= np.random.randint(low = 1, high = 100, size = (10,10))
print("The data to be plotted:\n")
print(df)
hm = sns.heatmap(data=df)
plt.show()
```

## OUTPUT

## Bivariate Analysis

## DATA SET

import pandas as pd  
import seaborn as sns  
df=pd.read\_csv("FlightInformation.csv")  
df

	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
...	...	...	...	...	...	...	...	...	...	...	...
10678	Air Asia	9/04/2019	Kolkata	Banglore	CCU → BLR	19:55	22:25	2h 30m	non-stop	No info	4107
10679	Air India	27/04/2019	Kolkata	Banglore	CCU → BLR	20:45	23:20	2h 35m	non-stop	No info	4145
10680	Jet Airways	27/04/2019	Banglore	Delhi	BLR → DEL	08:20	11:20	3h	non-stop	No info	7229
10681	Vistara	01/03/2019	Banglore	New Delhi	BLR → DEL	11:30	14:10	2h 40m	non-stop	No info	12648
10682	Air India	9/05/2019	Delhi	Cochin	DEL → GOI → BOM → COK	10:55	19:15	8h 20m	2 stops	No info	11753

10683 rows × 11 columns

## INFO

```
[ ] df.info()

<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
```

## ' DETECTING NULL AS PD

```
[ ] df.isnull().sum()

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
```

## ' REMOVING NULL DATA

```
[ ] df['Route']=df['Route'].fillna(df['Route'].mode()[0])
df['Total_Stops']=df['Total_Stops'].fillna(df['Total_Stops'].mode()[0])
df.isnull().sum()

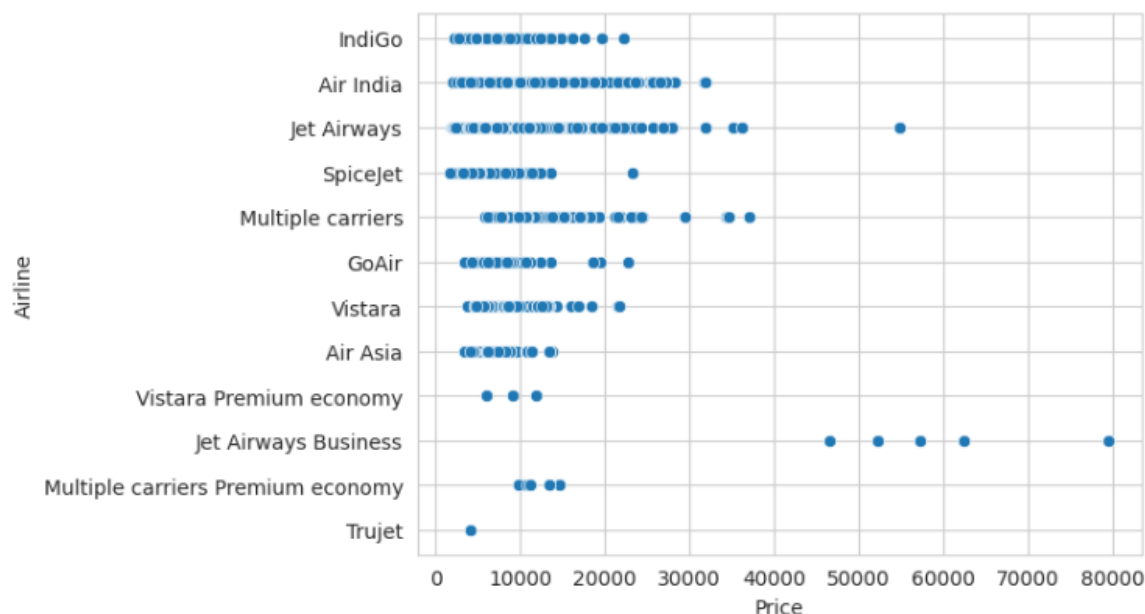
Airline                0
Date_of_Journey        0
Source                 0
Destination            0
Route                  0
Dep_Time               0
Arrival_Time           0
Duration               0
Total_Stops            0
Additional_Info        0
Price                  0
dtype: int64
```

## NUMERICAL & NUMERICAL:

### SCATTER\_PLOT

```
[20] sns.scatterplot(x=df["Price"],y=df["Airline"],data=df)
```

<Axes: xlabel='Price', ylabel='Airline'>

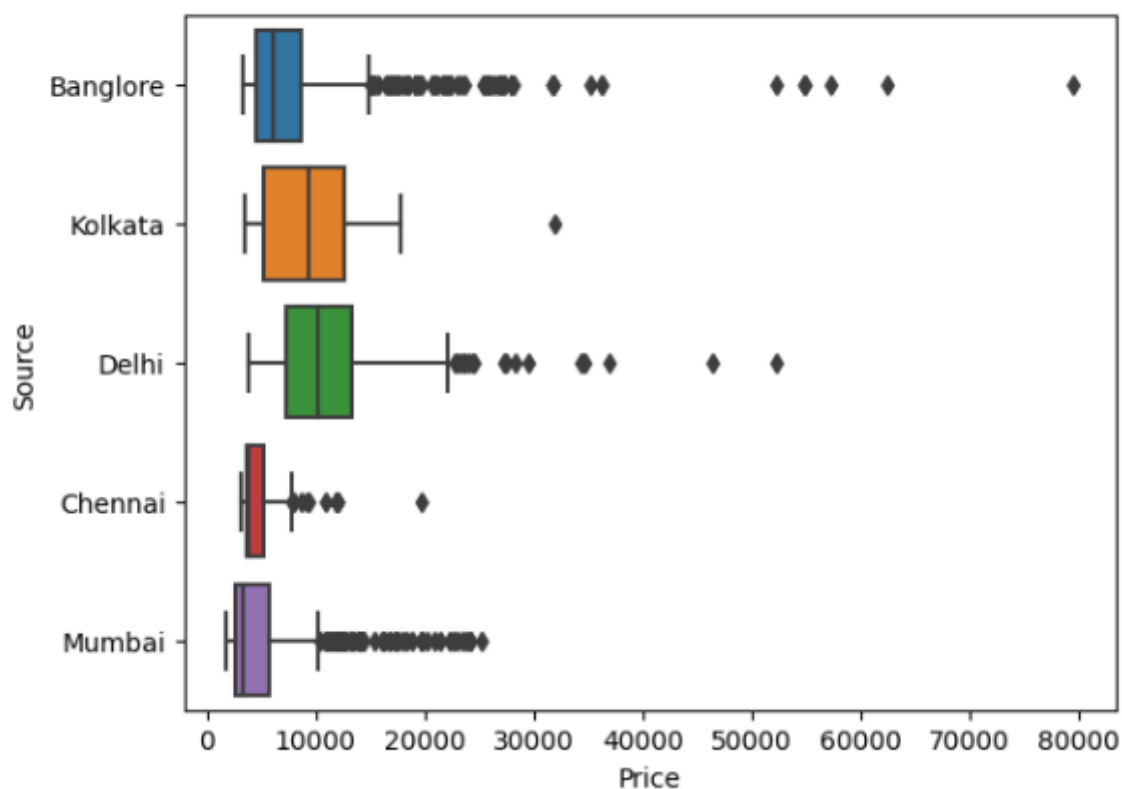


## NUMERICAL & CATEGORICAL:

### BOX PLOT

```
[13] sns.boxplot(x=df['Price'],y=df['Source'],data=df)
```

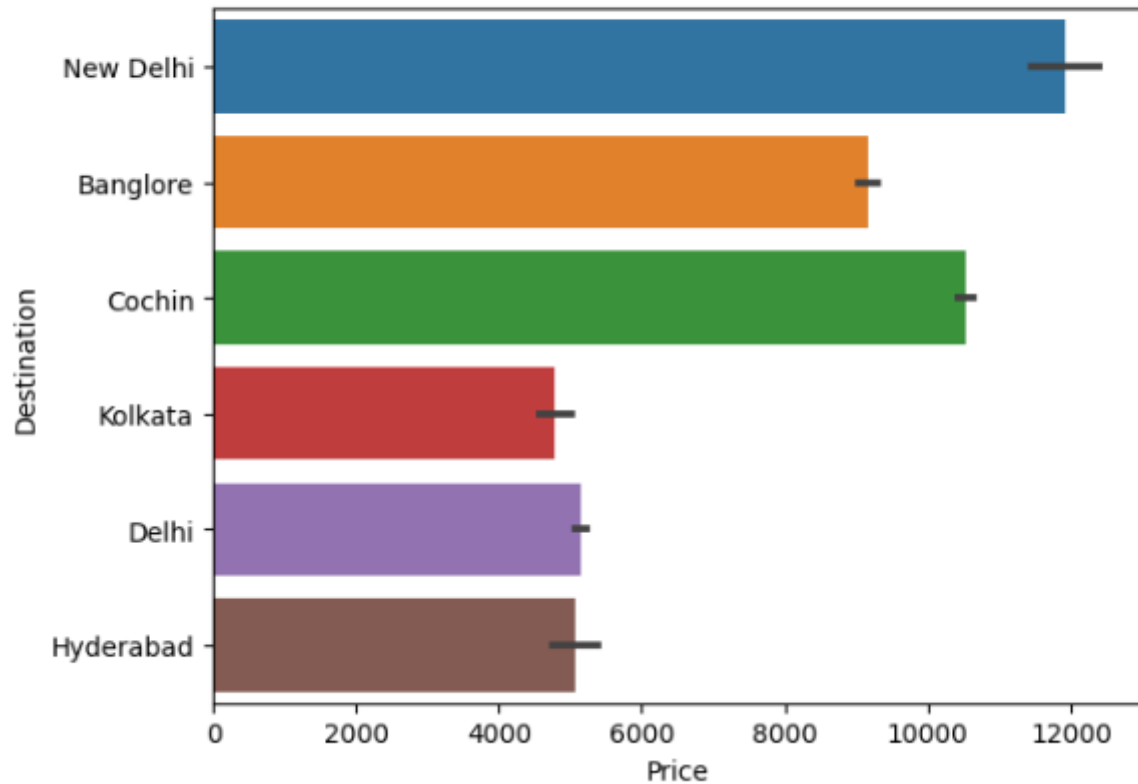
<Axes: xlabel='Price', ylabel='Source'>



## ' BAR PLOT

```
[11] sns.barplot(x=df['Price'],y=df['Destination'],data=df)
```

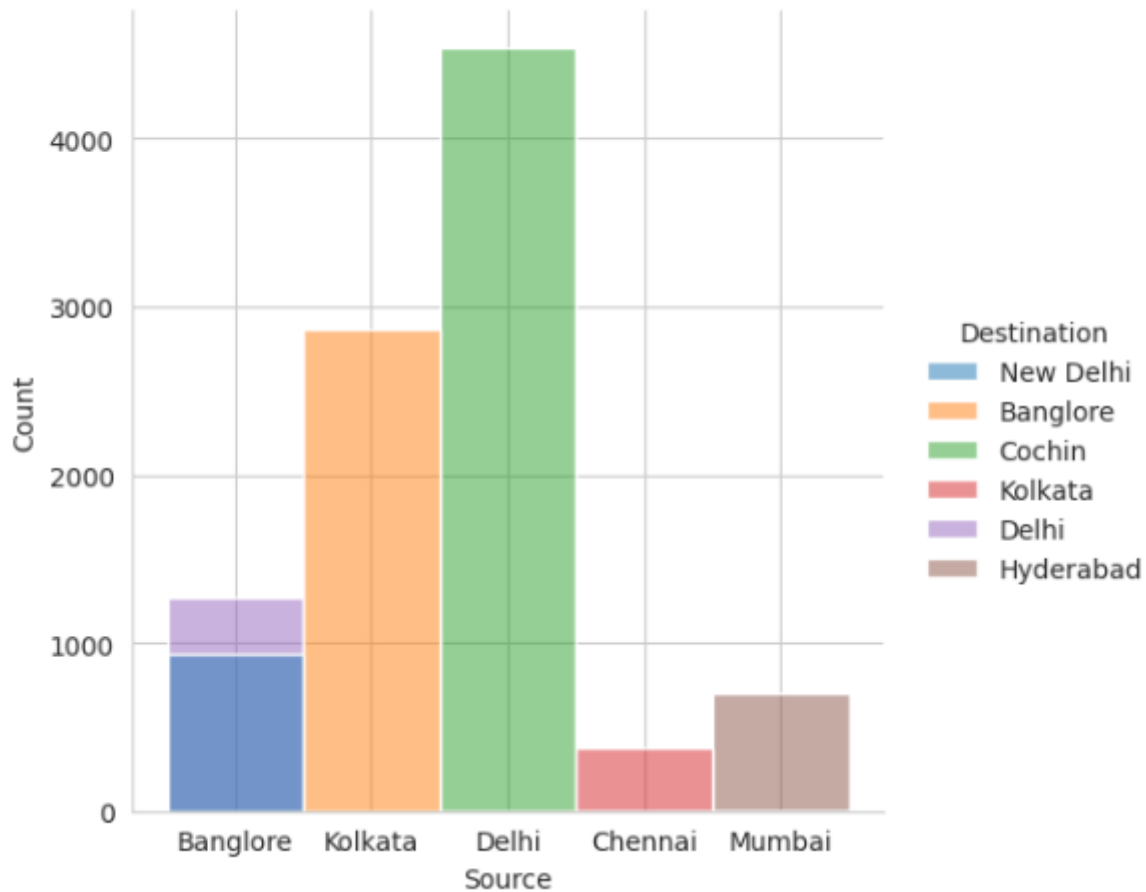
```
<Axes: xlabel='Price', ylabel='Destination'>
```



## ' DIST PLOT

```
[29] sns.displot(df,x="Source",hue="Destination")
```

```
<seaborn.axisgrid.FacetGrid at 0x7f8f33c4af40>
```



## ' MULTIVARIANT ANALYSIS

## ' CORRELATION

```
[39] df.corr()
```

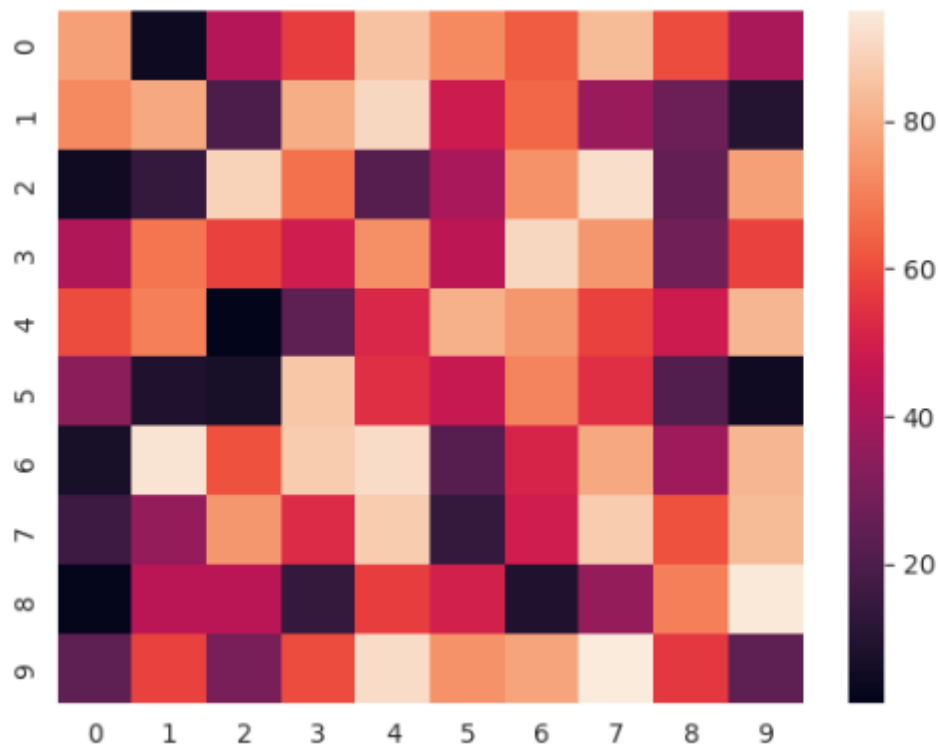
Price	
Price	1.0

## ' HEAT MAP

```
[41] import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
df= pd.read_csv("FlightInformation.csv")
df= np.random.randint(low = 1, high = 100, size = (10,10))
print("The data to be plotted:\n")
print(df)
hm = sns.heatmap(data=df)
plt.show()
```

The data to be plotted:

```
[[77  4 43 57 85 72 63 83 60 41]
 [72 79 20 80 90 48 65 37 27 10]
 [ 5 14 89 67 22 40 74 92 25 77]
 [42 68 58 49 73 45 90 75 28 58]
 [60 70  1 24 52 81 75 58 48 82]
 [34  9  7 86 54 47 71 54 21  5]
 [ 7 93 61 87 91 22 51 79 38 82]
 [16 36 75 53 87 14 49 87 61 83]
 [ 2 44 44 14 57 50  9 36 70 94]
 [24 58 30 60 91 74 78 95 56 24]]
```



## ' RESULT

Thus the Bivariate and Multivariate analysis is observed for the given data set.