Name: Mayuri Shridatta Yerande
Class: D15B
Roll No: 70

EXPERIMENT - 2

AIM: Data Visualization/ Exploratory data Analysis using Matplotlib and Seaborn

TO DO:

- Create bar graph, contingency table using any 2 features.
- Plot Scatter plot, box plot, Heatmap using seaborn.
- Create histogram
- Describe what this graph and table indicates.
- Handle outlier using box plot and Interquartile range.

ABOUT DATASET:

Link of our dataset:
https://www.kaggle.com/datasets/jessemostipak/college-tuition-diversity-and-pay

Our Dataset -
College tuition, diversity, and pay

- Tuition and fees by college/university for 2018-2019, along with school type, degree length, state, in-state vs out-of-state from the Chronicle of Higher Education.
- Diversity by college/university for 2014, along with school type, degree length, state, in-state vs out-of-state from the Chronicle of Higher Education.
- Example diversity graphics from Priceonomics.
- Average net cost by income bracket from TuitionTracker.org.
- Example price trend and graduation rates from TuitionTracker.org
- Salary potential data comes from payscale.com.

The columns in our dataset incluse:
names, state, state_code, type, degree_length, room_on_board, in_state_tuition, in_state_total, out_of_state_tuition, out_of_state_total

THEORY:

Data Visualization:-

Data visualization provides a good, organized pictorial representation of the data which makes it easier to understand, observe, analyze. We will discuss how to visualize data using Python. Python provides various libraries that come with different features for visualizing data. All these libraries come with different features and can support various types of graphs.

Exploratory data analysis:-

- EDA is applied to investigate the data and summarize the key insights.
- It will give you the basic understanding of your data, its distribution, null values and much more.
- You can either explore data using graphs or through some python function.

Matplotlib library:-

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.

Seaborn Library:-

Seaborn is **a Python data visualization library based on matplotlib**. It provides a high-level interface for drawing attractive and informative statistical graphics. For a brief introduction to the ideas behind the library, you can read the introductory notes or the paper.

Bar Graph:- Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data. They are also known as bar charts. Bar graphs are one of the means of data handling in statistics

Contingency Table:- Contingency Table is one of the techniques for exploring two or even more variables. It is basically a tally of counts between two or more categorical variables.

Scatter Plot:- A scatter plot is a set of points plotted on horizontal and vertical axes. Scatter plots are important in statistics because they can show the extent of correlation, if any, between the values of observed quantities or phenomena (called variables).

Box Plot:- A Box Plot also known as Whisker plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum. In the box plot, a box is created from the first quartile to the third quartile, a vertical line is also there which goes through the box at the median.

Heat Map:- A heatmap is a two-dimensional graphical representation of data where the individual values that are contained in a matrix are represented as colors. The Seaborn package allows the creation of annotated heatmaps which can be tweaked using Matplotlib tools as per the creator's requirement.

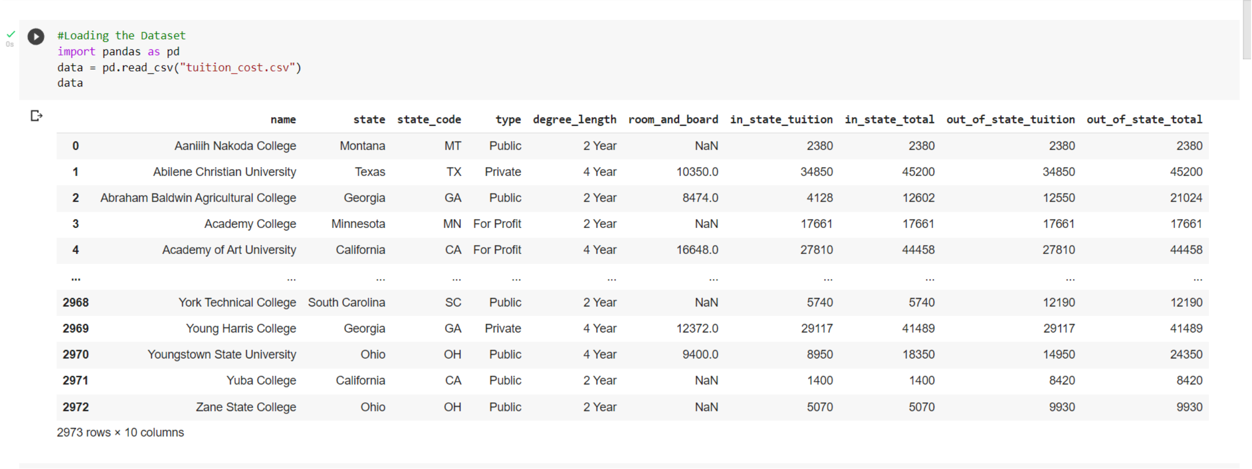
Histogram:- A histogram is a graph showing *frequency* distributions. It is a graph showing the number of observations within each given interval.

Normalized Histogram:- Histogram normalization is a technique to distribute the frequencies of the histogram over a wider range than the current range. This technique is used in image processing too. There we do histogram normalization for enhancing the contrast of poor contrast images.

Outliers:- An outlier is a value in the data set that is extremely distinct from most of the other values. Detecting Outliers: Box-Plot and Interquartile Range

SCREENSHOTS:

- **Loading Dataset**

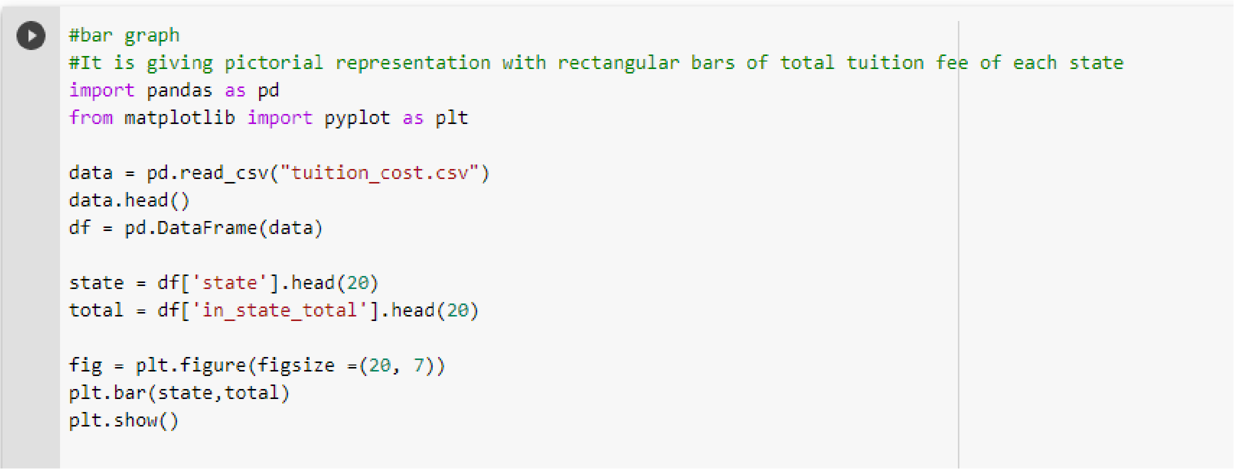


```
#Loading the Dataset
import pandas as pd
data = pd.read_csv("tuition_cost.csv")
data
```

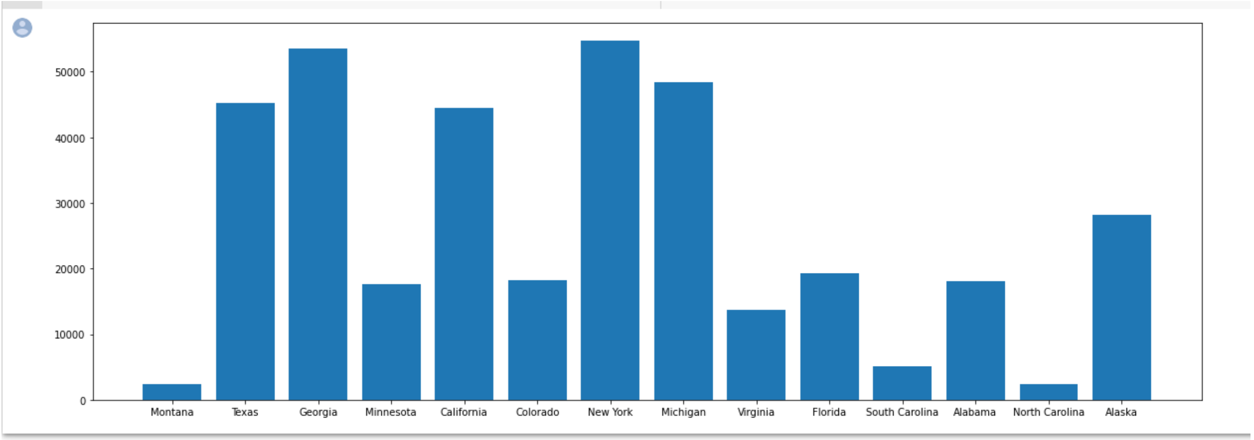
	name	state	state_code	type	degree_length	room_and_board	in_state_tuition	in_state_total	out_of_state_tuition	out_of_state_total
0	Aaniiih Nakoda College	Montana	MT	Public	2 Year	NaN	2380	2380	2380	2380
1	Abilene Christian University	Texas	TX	Private	4 Year	10350.0	34850	45200	34850	45200
2	Abraham Baldwin Agricultural College	Georgia	GA	Public	2 Year	8474.0	4128	12602	12550	21024
3	Academy College	Minnesota	MN	For Profit	2 Year	NaN	17661	17661	17661	17661
4	Academy of Art University	California	CA	For Profit	4 Year	16648.0	27810	44458	27810	44458
...	...	...	...	...	...	...	...	...	...	...
2968	York Technical College	South Carolina	SC	Public	2 Year	NaN	5740	5740	12190	12190
2969	Young Harris College	Georgia	GA	Private	4 Year	12372.0	29117	41489	29117	41489
2970	Youngstown State University	Ohio	OH	Public	4 Year	9400.0	8950	18350	14950	24350
2971	Yuba College	California	CA	Public	2 Year	NaN	1400	1400	8420	8420
2972	Zane State College	Ohio	OH	Public	2 Year	NaN	5070	5070	9930	9930

2973 rows × 10 columns

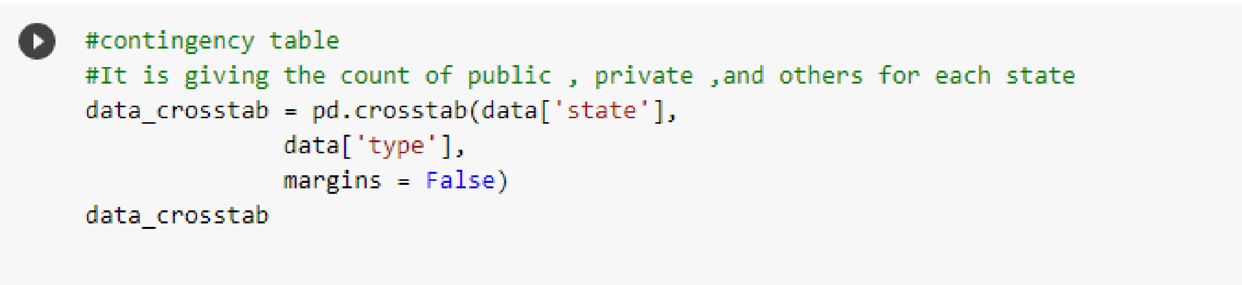
- **Bar Graph**



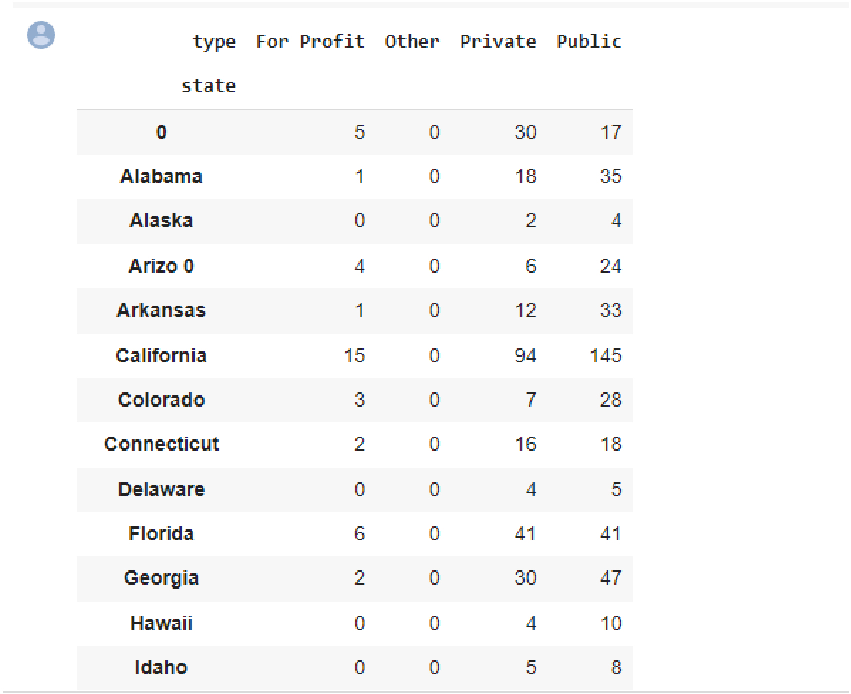
```
#bar graph
#It is giving pictorial representation with rectangular bars of total tuition fee of each state
import pandas as pd
from matplotlib import pyplot as plt

data = pd.read_csv("tuition_cost.csv")
data.head()
df = pd.DataFrame(data)

state = df['state'].head(20)
total = df['in_state_total'].head(20)

fig = plt.figure(figsize =(20, 7))
plt.bar(state,total)
plt.show()
```



- **Contingency table**

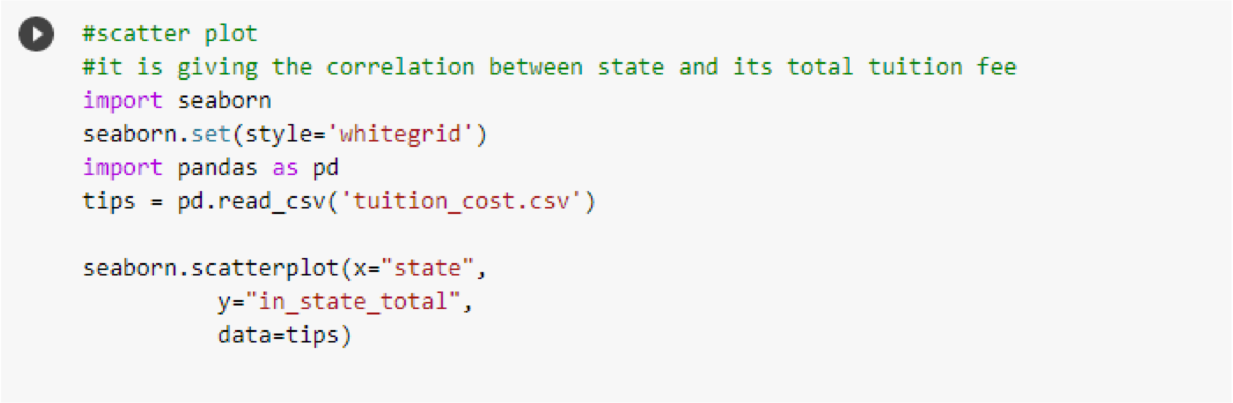


```
#contingency table
#It is giving the count of public , private ,and others for each state
data_crosstab = pd.crosstab(data['state'],
                 data['type'],
                 margins = False)
data_crosstab
```



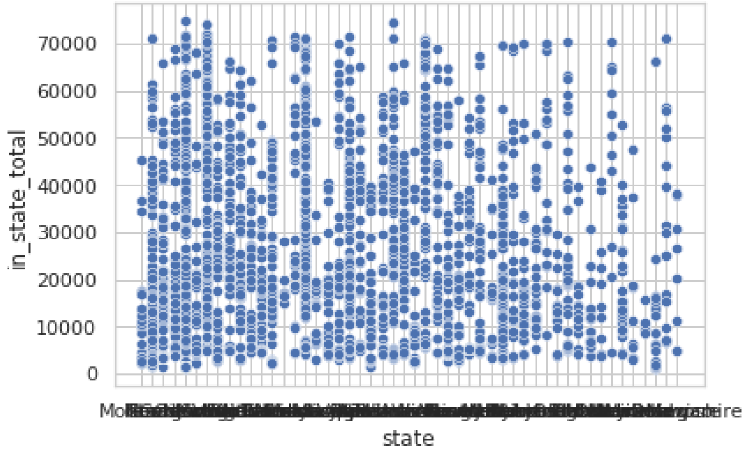
type	For Profit	Other	Private	Public
state				
0	5	0	30	17
Alabama	1	0	18	35
Alaska	0	0	2	4
Arizo 0	4	0	6	24
Arkansas	1	0	12	33
California	15	0	94	145
Colorado	3	0	7	28
Connecticut	2	0	16	18
Delaware	0	0	4	5
Florida	6	0	41	41
Georgia	2	0	30	47
Hawaii	0	0	4	10
Idaho	0	0	5	8

- **Scatter Plot**

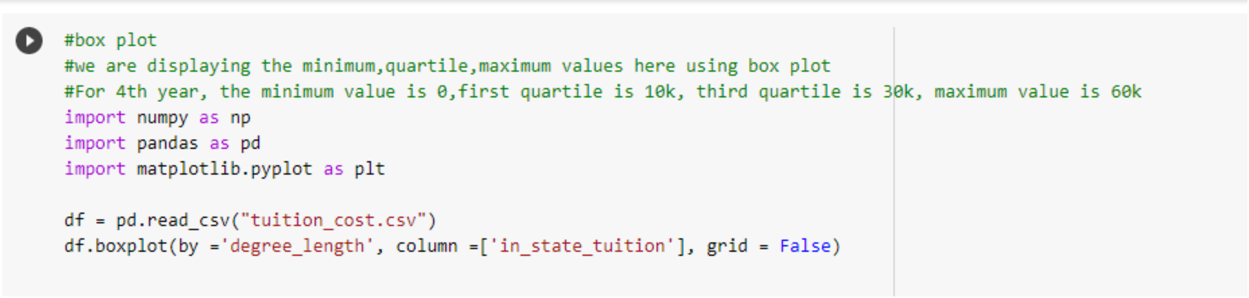


```
#scatter plot
#it is giving the correlation between state and its total tuition fee
import seaborn
seaborn.set(style='whitegrid')
import pandas as pd
tips = pd.read_csv('tuition_cost.csv')

seaborn.scatterplot(x="state",
          y="in_state_total",
          data=tips)
```

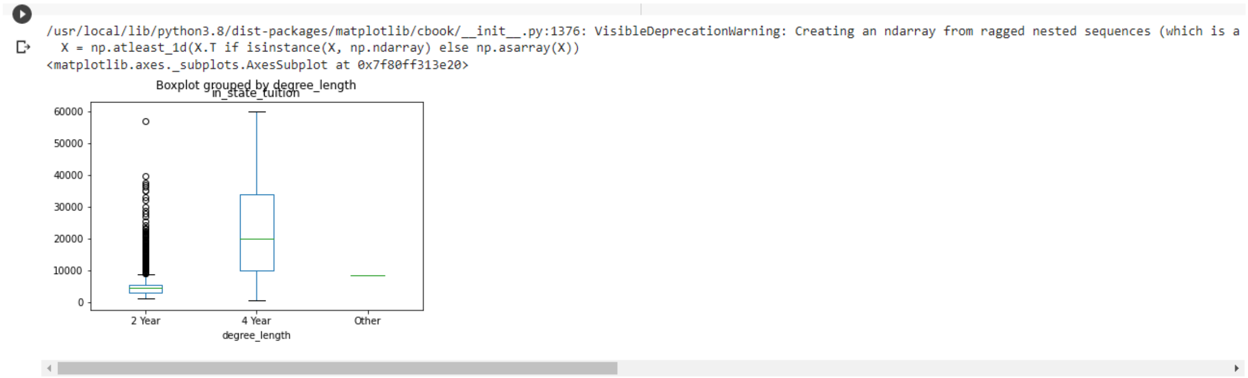
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f3307ac0160>
```



- **Box Plot**

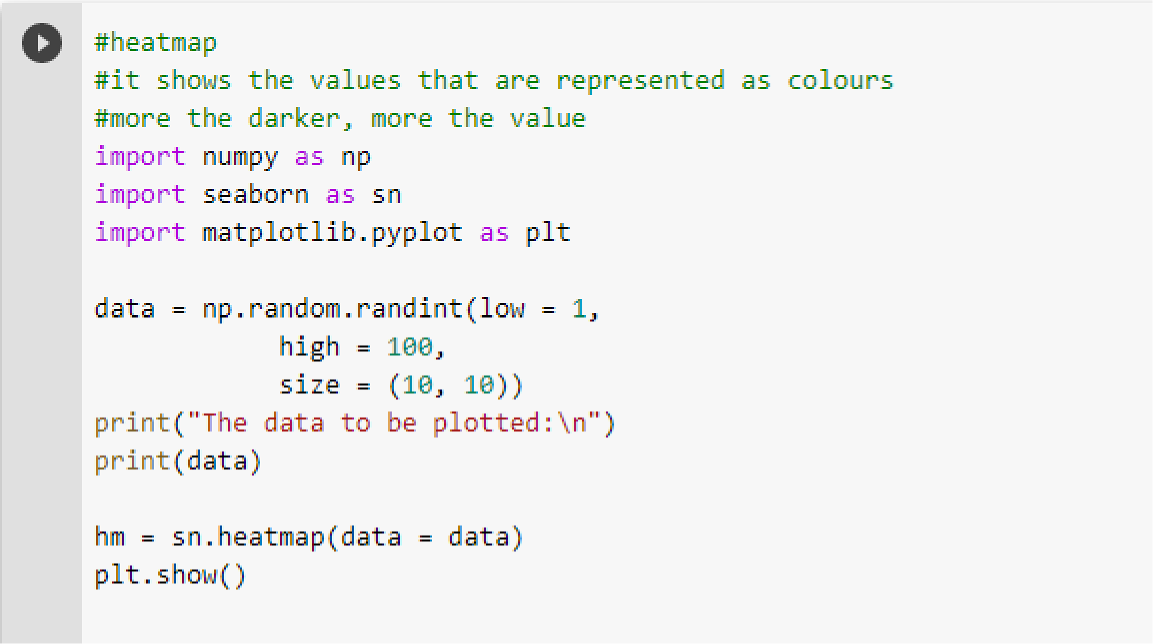


```
#box plot
#we are displaying the minimum,quartile,maximum values here using box plot
#For 4th year, the minimum value is 0,first quartile is 10k, third quartile is 30k, maximum value is 60k
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

df = pd.read_csv("tuition_cost.csv")
df.boxplot(by ='degree_length', column =['in_state_tuition'], grid = False)
```

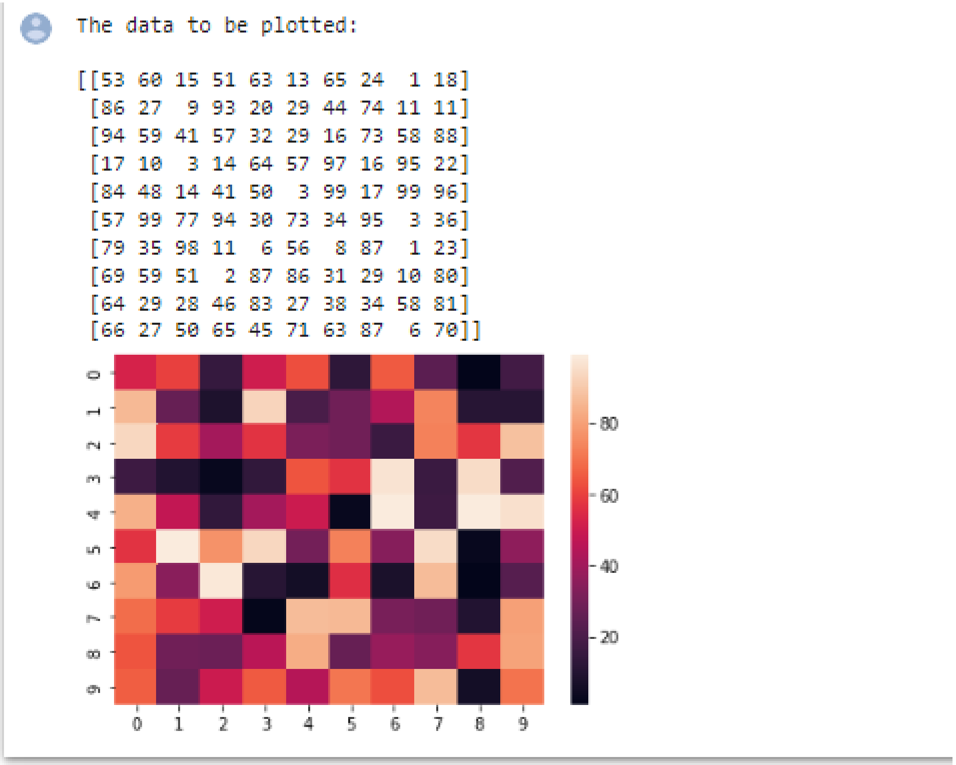


```
/usr/local/lib/python3.8/dist-packages/matplotlib/cbook/__init__.py:1376: VisibleDeprecationWarning: Creating an ndarray from ragged nested sequences (which is a
  X = np.atleast_1d(X.T if isinstance(X, np.ndarray) else np.asarray(X))
<matplotlib.axes._subplots.AxesSubplot at 0x7f80ff313e20>
```

- **HeatMap**

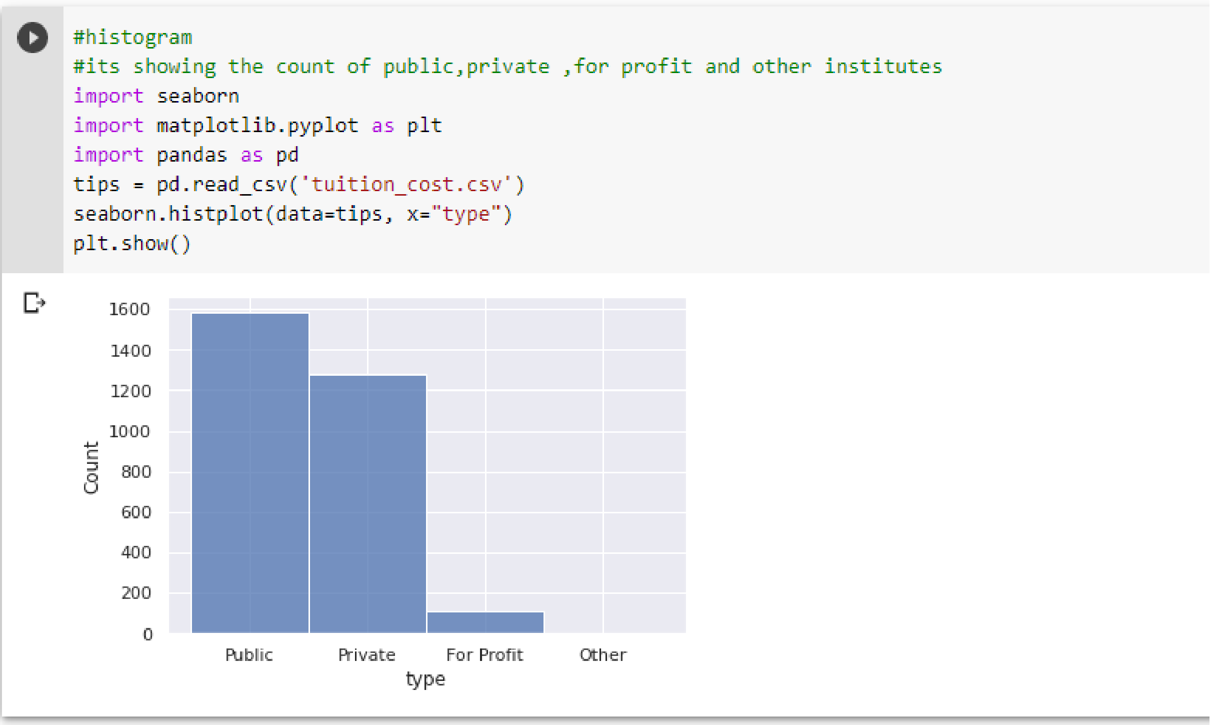


```
#heatmap
#it shows the values that are represented as colours
#more the darker, more the value
import numpy as np
import seaborn as sn
import matplotlib.pyplot as plt

data = np.random.randint(low = 1,
            high = 100,
            size = (10, 10))
print("The data to be plotted:\n")
print(data)

hm = sn.heatmap(data = data)
plt.show()
```



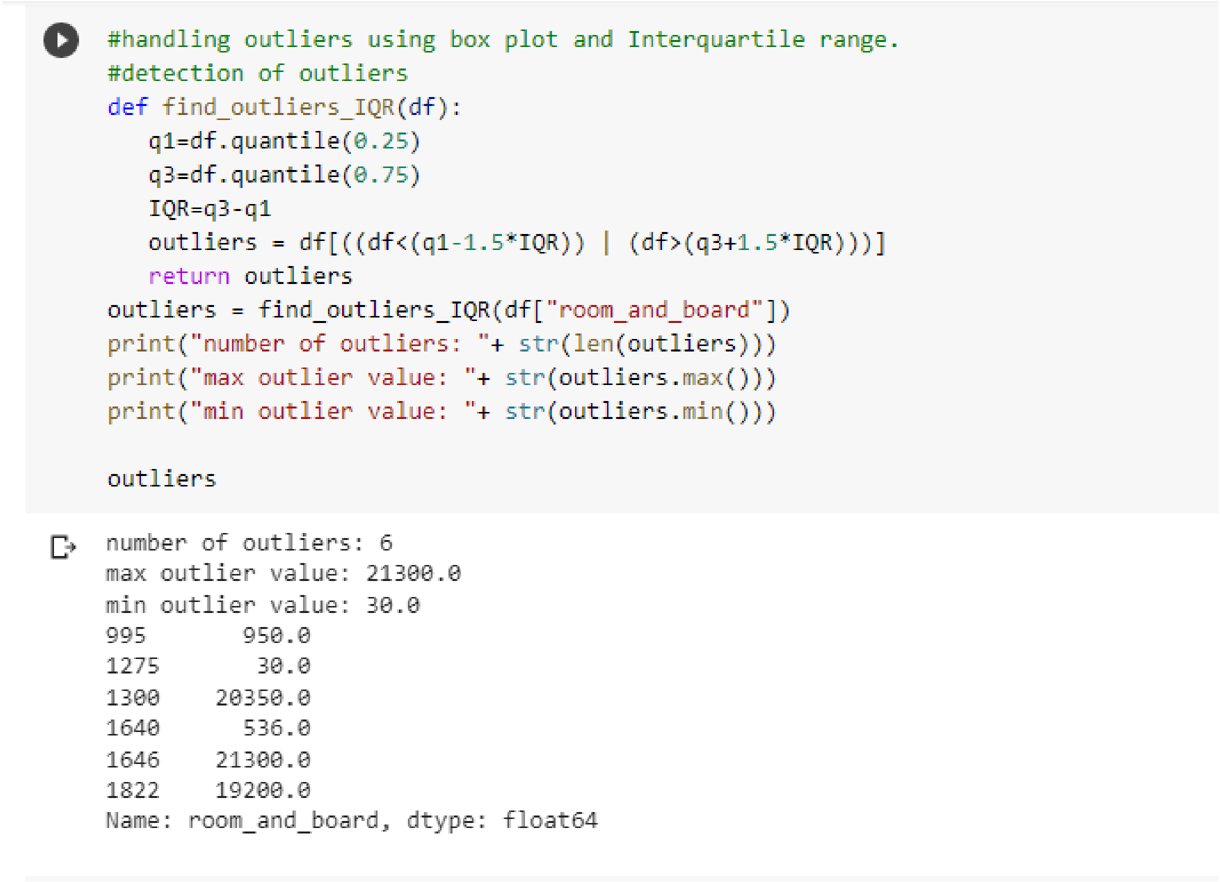
```
The data to be plotted:

[[53 60 15 51 63 13 65 24  1 18]
 [86 27  9 93 20 29 44 74 11 11]
 [94 59 41 57 32 29 16 73 58 88]
 [17 10  3 14 64 57 97 16 95 22]
 [84 48 14 41 50  3 99 17 99 96]
 [57 99 77 94 30 73 34 95  3 36]
 [79 35 98 11  6 56  8 87  1 23]
 [69 59 51  2 87 86 31 29 10 80]
 [64 29 28 46 83 27 38 34 58 81]
 [66 27 50 65 45 71 63 87  6 70]]
```

- **Histogram**



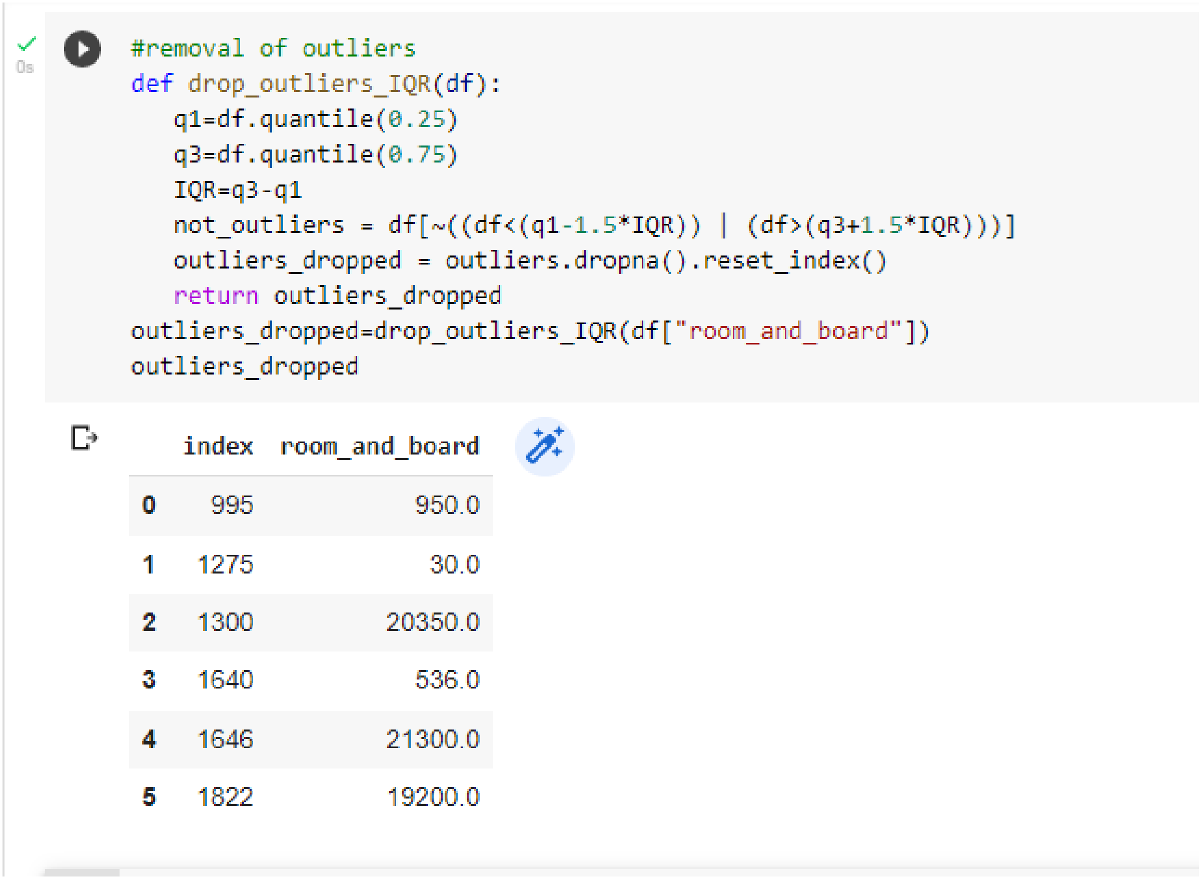
```python
#histogram
#its showing the count of public,private ,for profit and other institutes
import seaborn
import matplotlib.pyplot as plt
import pandas as pd
tips = pd.read_csv('tuition_cost.csv')
seaborn.histplot(data=tips, x="type")
plt.show()
```

- **Handling outliers using box plot and interquartile range**



```
#handling outliers using box plot and Interquartile range.
#detection of outliers
def find_outliers_IQR(df):
   q1=df.quantile(0.25)
   q3=df.quantile(0.75)
   IQR=q3-q1
   outliers = df[((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]
   return outliers
outliers = find_outliers_IQR(df["room_and_board"])
print("number of outliers: "+ str(len(outliers)))
print("max outlier value: "+ str(outliers.max()))
print("min outlier value: "+ str(outliers.min()))

outliers
```

```
number of outliers: 6
max outlier value: 21300.0
min outlier value: 30.0
995       950.0
1275       30.0
1300    20350.0
1640      536.0
1646    21300.0
1822    19200.0
Name: room_and_board, dtype: float64
```



```
#removal of outliers
def drop_outliers_IQR(df):
   q1=df.quantile(0.25)
   q3=df.quantile(0.75)
   IQR=q3-q1
   not_outliers = df[~((df<(q1-1.5*IQR)) | (df>(q3+1.5*IQR)))]
   outliers_dropped = outliers.dropna().reset_index()
   return outliers_dropped
outliers_dropped=drop_outliers_IQR(df["room_and_board"])
outliers_dropped
```

	index	room_and_board
0	995	950.0
1	1275	30.0
2	1300	20350.0
3	1640	536.0
4	1646	21300.0
5	1822	19200.0

CONCLUSION: In this practical, we studied different types of graphs according to the data present in the dataset. We used a box plot graph to find the outliers and remove it. Graphs help to make our data presentable and summarize the data in a crisp manner. We plotted the bar graph and histogram to get the data in rectangular bars, we usd the heat map to represent the data in colors, We also plotted scatter plot and made the contingency table of the required columns. Graphical representation of data helps us to analyze the data efficiently.