SESSIONS 5

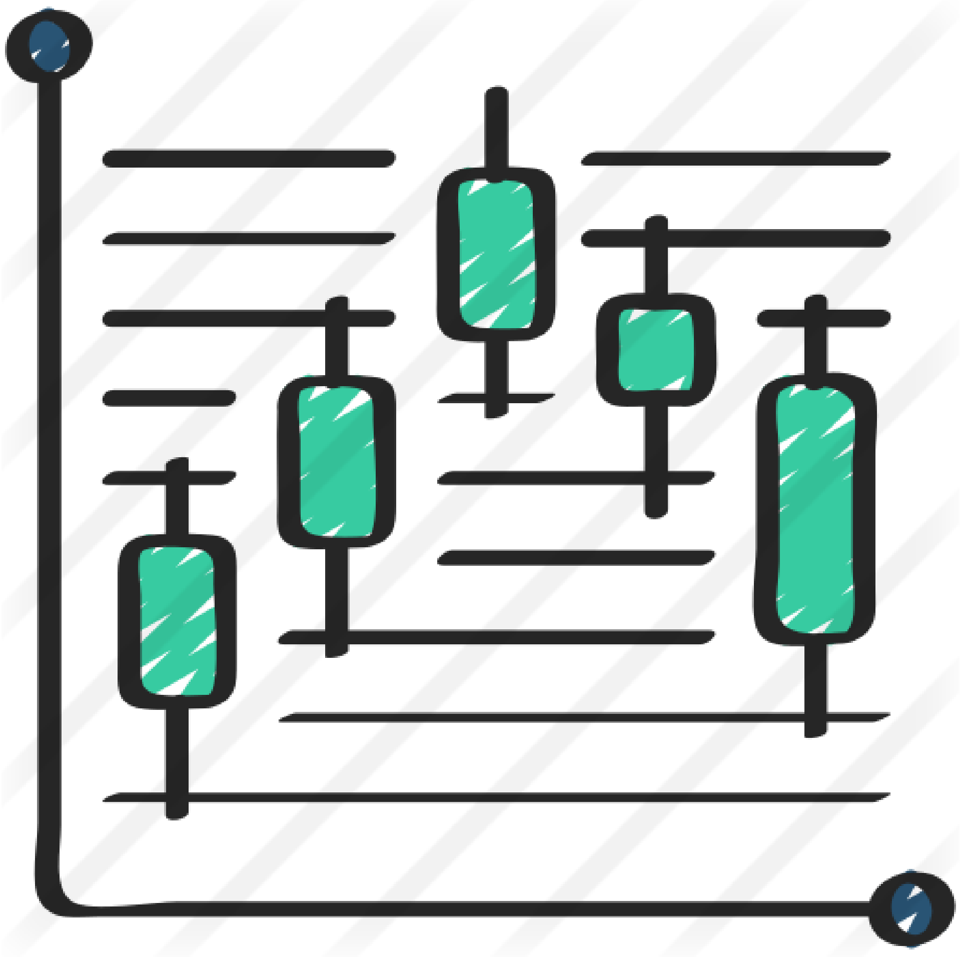
Boxplot

Data Science Program

Purwadhika
Startup and Coding School

Outline

- What is Boxplot?
- Element of a Boxplot
- Information from Boxplot
- Create Boxplot using Matplotlib, Seaborn, and Pandas

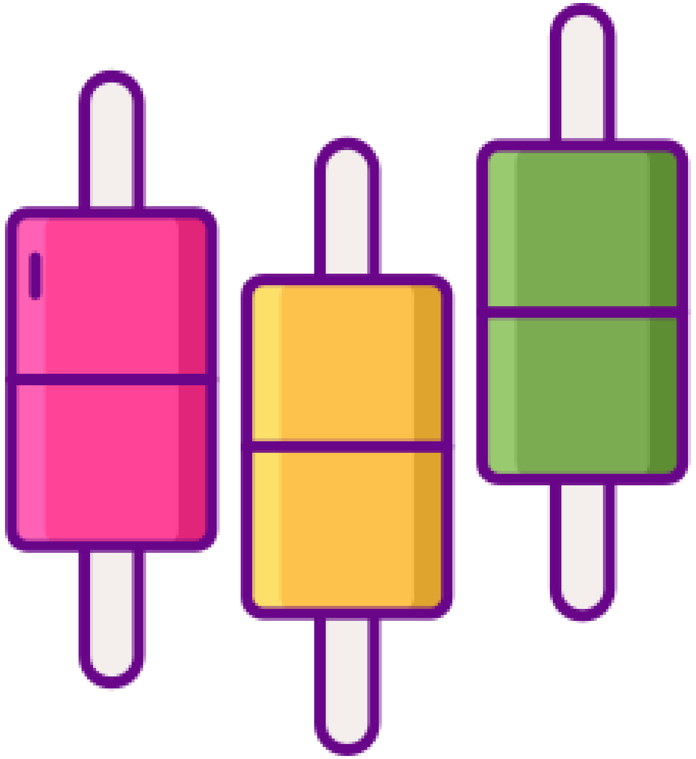


What is Boxplot?

Purwadhika
Startup and Coding School

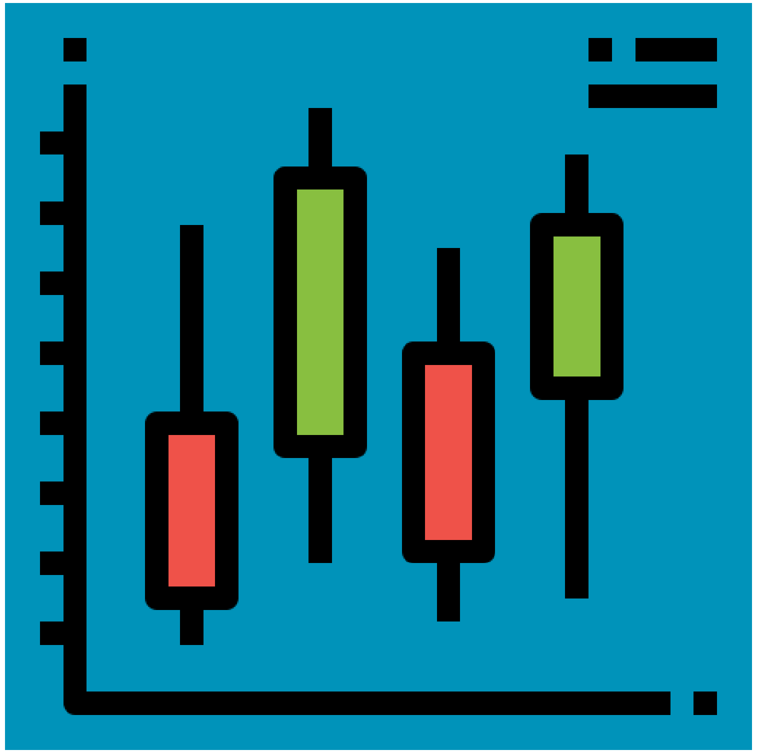
What is boxplot?

- A box plot (also known as box and whisker plot) is a type of chart often used in explanatory data analysis to visually show the distribution of numerical data and Skewness through displaying the data quartiles (or percentiles) and averages.
- Box plots show the five-number summary of a set of data: including the minimum score, first (lower) quartile, median, third (upper) quartile, and maximum score.



What is boxplot?

- In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles.
- Box plots may also have lines extending from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram.
- Outliers may be plotted as individual points.

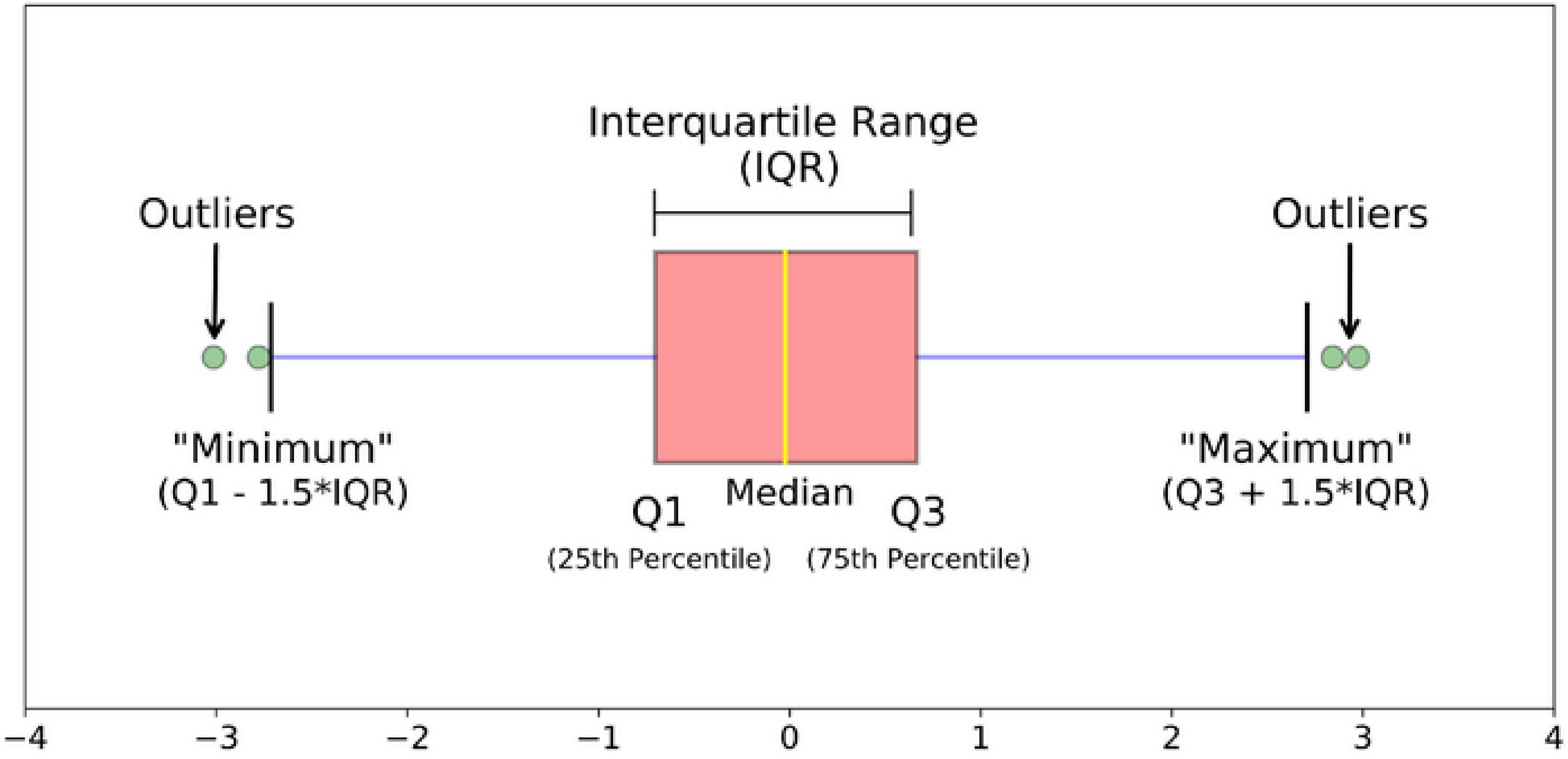


Purwadhika
Startup and Coding School

Element of a Boxplot

Purwadhika
Startup and Coding School

Element of a Boxplot



Purwadhika
Startup and Coding School

Element of a Boxplot

A boxplot is a standardized way of displaying the dataset based on a five-number summary:

1. **Minimum**: the lowest data point excluding any outliers.
2. **Maximum**: the largest data point excluding any outliers.
3. **Median** (Q2 / 50th Percentile): the middle value of the dataset.
4. **First quartile** (Q1 / 25th Percentile): also known as the lower quartile qn(0.25), is the median of the lower half of the dataset.
5. **Third quartile** (Q3 / 75th Percentile): also known as the upper quartile qn(0.75), is the median of the upper half of the dataset.

An important element used to construct the box plot by determining the minimum and maximum data values feasible, but is not part of the aforementioned five-number summary, is the interquartile range or IQR denoted below:

- **Interquartile Range** (IQR) is the distance between the upper and lower quartile.

  IQR = Q3 - Q1

Information from Boxplot

Information from Boxplot

1. **Box plots are useful as they show the average score of a data set**

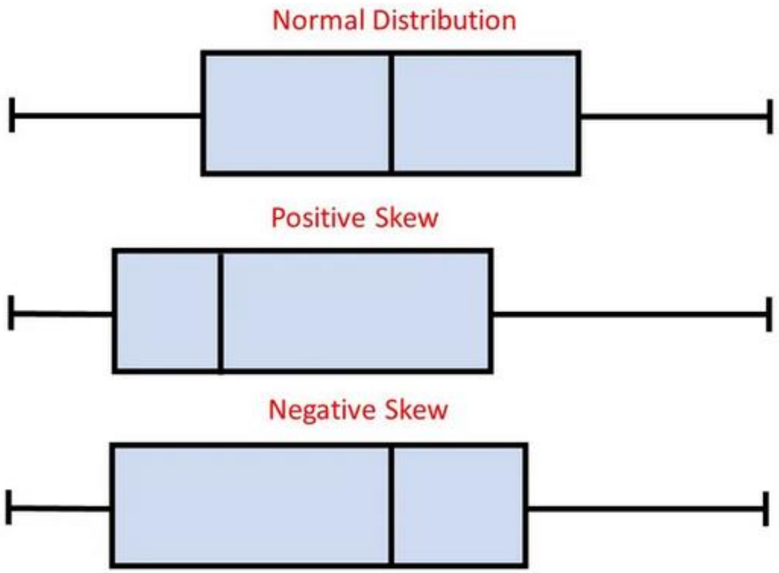
   The median is the average value from a set of data and is shown by the line that divides the box into two parts. Half the scores are greater than or equal to this value, and half are less.

2. **Box plots are useful as they show the skewness of a data set**

   The box plot shape will show if a statistical data set is normally distributed or skewed.

Information from Boxplot

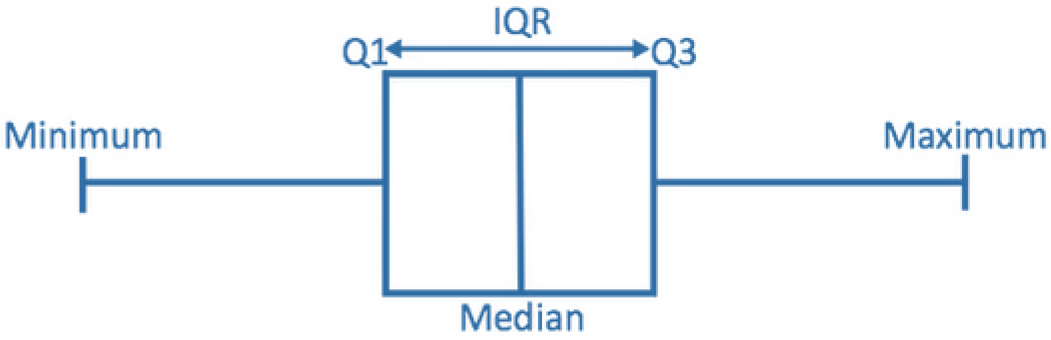
- When the median is in the middle of the box, and the whiskers are about the same on both sides of the box, then the distribution is **symmetric**.
- When the median is closer to the bottom of the box, and if the whisker is shorter on the lower end of the box, then the distribution is positively skewed (**skewed right**).
- When the median is closer to the top of the box, and if the whisker is shorter on the upper end of the box, then the distribution is negatively skewed (**skewed left**).



Information from Boxplot

3. Box plots are useful as they show the dispersion of a data set

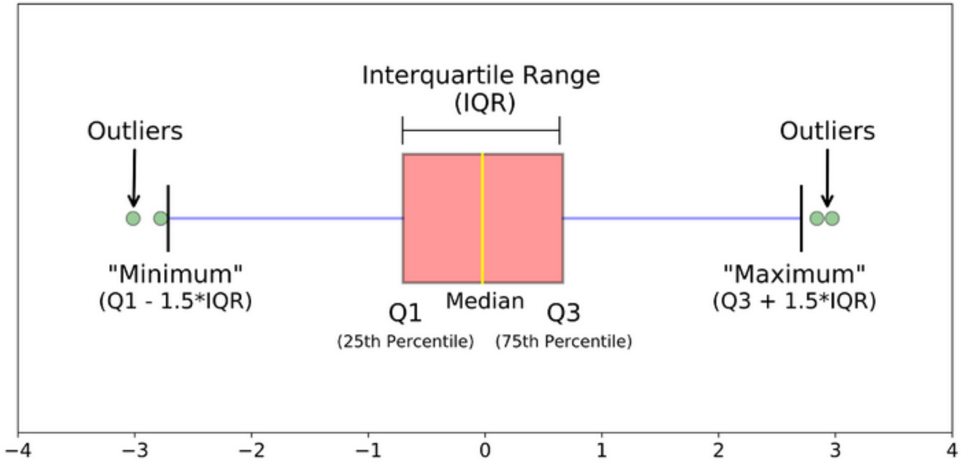
- In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed.
- The smallest value and largest value are found at the end of the ‘whiskers’ and are useful for providing a visual indicator regarding the spread of scores (e.g. the range).
- The interquartile range (IQR) is the box plot showing the middle 50% of scores and can be calculated by subtracting the lower quartile from the upper quartile (e.g. Q3−Q1).



Information from Boxplot

4. Box plots are useful as they show outliers within a data set

- An outlier is an observation that is numerically distant from the rest of the data.
- When reviewing a box plot, an outlier is defined as a data point that is located outside the whiskers of the box plot.
- For example, outside 1.5 times the interquartile range above the upper quartile and below the lower quartile (Q1 - 1.5 * IQR or Q3 + 1.5 * IQR).

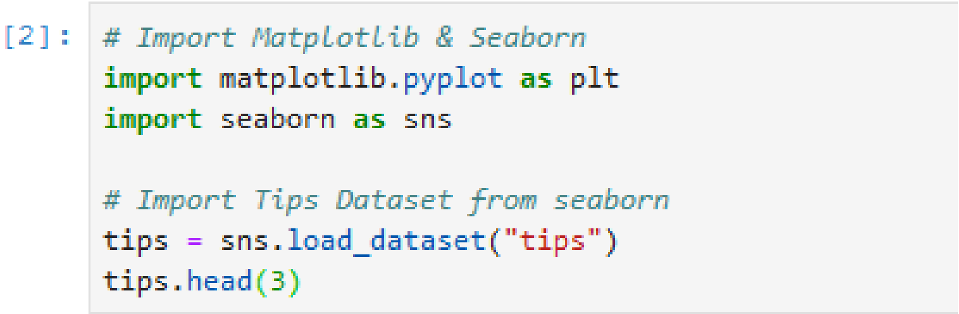


Create Boxplot using Matplotlib

Purwadhika
Startup and Coding School

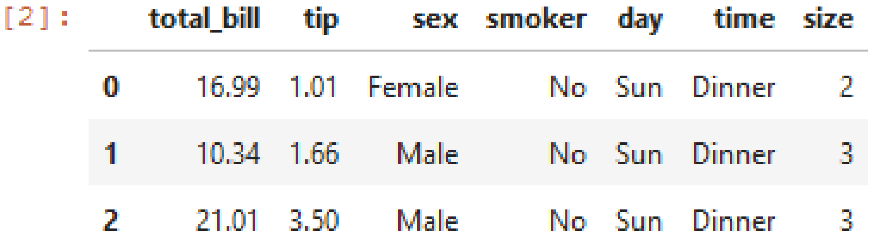
Create Boxplot using Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.



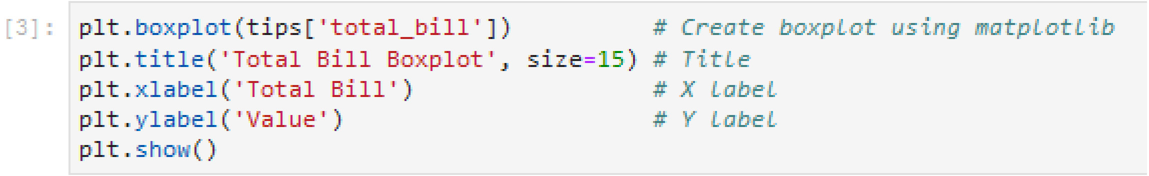
```
[2]: # Import Matplotlib & Seaborn
     import matplotlib.pyplot as plt
     import seaborn as sns

     # Import Tips Dataset from seaborn
     tips = sns.load_dataset("tips")
     tips.head(3)
```

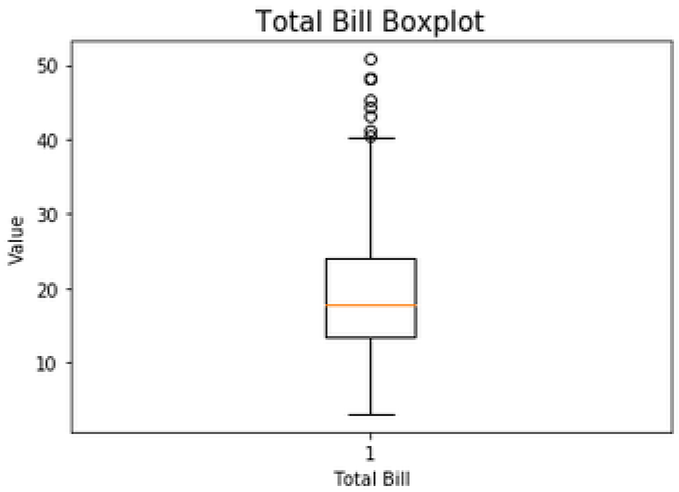


[2]:

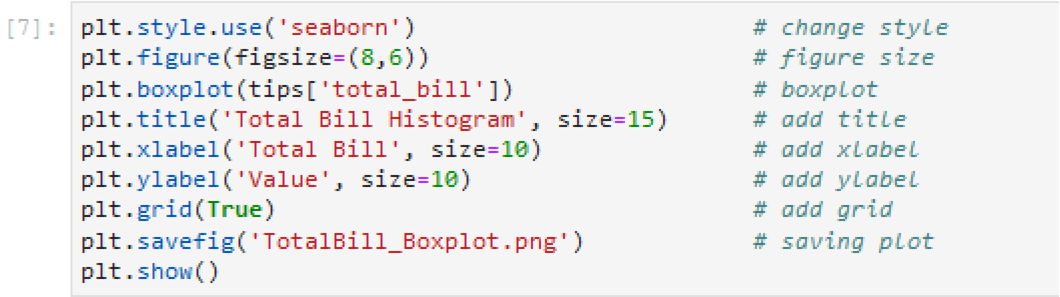
	total_bill	tip	sex	smoker	day	time	size
0	16.99	1.01	Female	No	Sun	Dinner	2
1	10.34	1.66	Male	No	Sun	Dinner	3
2	21.01	3.50	Male	No	Sun	Dinner	3



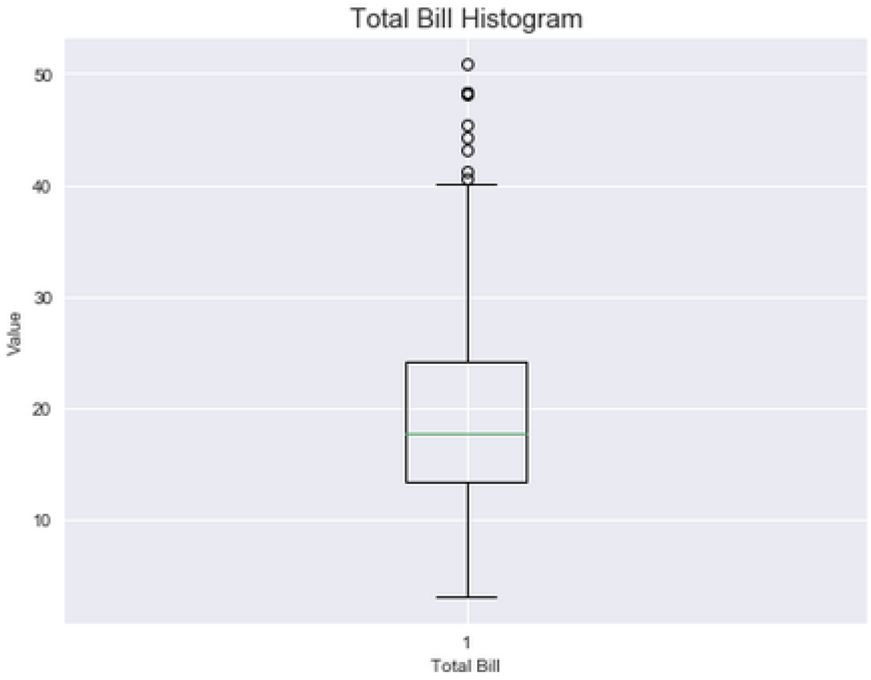
```
[3]: plt.boxplot(tips['total_bill'])          # Create boxplot using matplotlib
     plt.title('Total Bill Boxplot', size=15) # Title
     plt.xlabel('Total Bill')                 # X label
     plt.ylabel('Value')                      # Y label
     plt.show()
```



Create Boxplot using Matplotlib



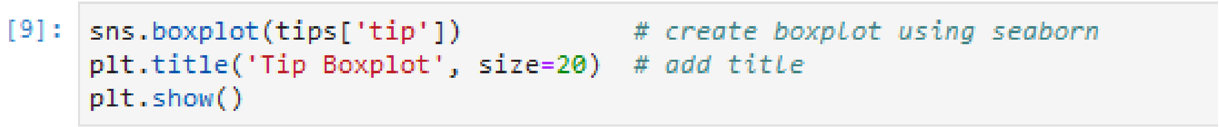
```
[7]:  plt.style.use('seaborn')                          # change style
      plt.figure(figsize=(8,6))                          # figure size
      plt.boxplot(tips['total_bill'])                    # boxplot
      plt.title('Total Bill Histogram', size=15)         # add title
      plt.xlabel('Total Bill', size=10)                  # add xlabel
      plt.ylabel('Value', size=10)                       # add ylabel
      plt.grid(True)                                     # add grid
      plt.savefig('TotalBill_Boxplot.png')               # saving plot
      plt.show()
```



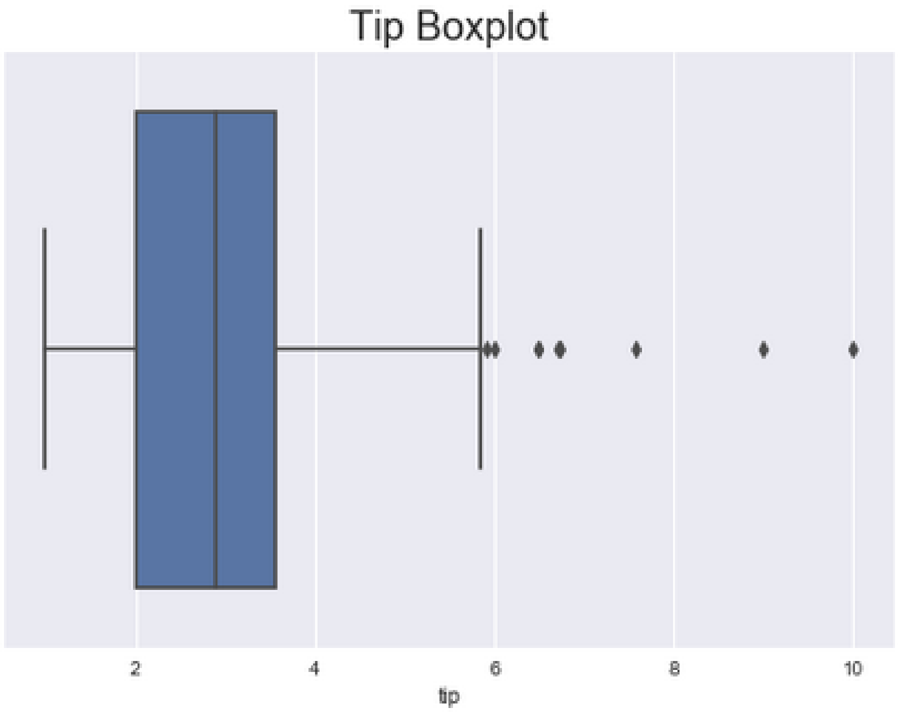
Create Boxplot using Seaborn

Purwadhika
Startup and Coding School

Create Boxplot using Seaborn



```
[9]: sns.boxplot(tips['tip'])              # create boxplot using seaborn
     plt.title('Tip Boxplot', size=20)   # add title
     plt.show()
```



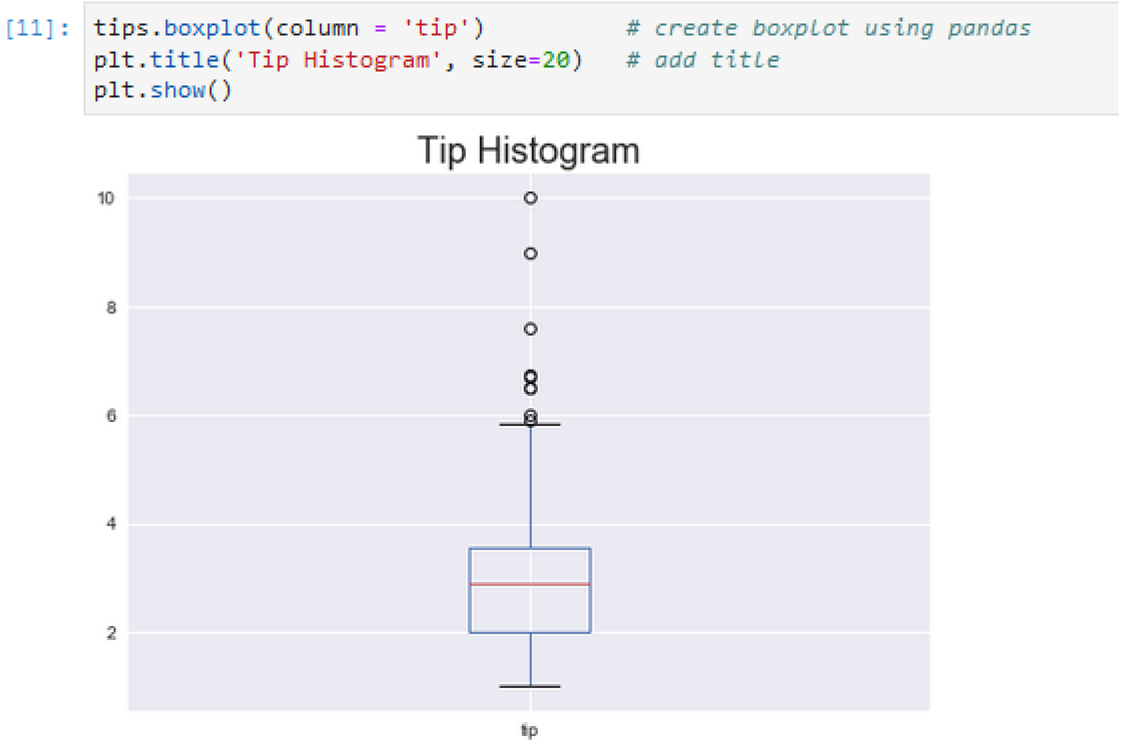
Seaborn is a Python data visualization library based on matplotlib.

It provides a high-level interface for drawing attractive and informative statistical graphics.

Create Boxplot using Pandas

Purwadhika
Startup and Coding School

Create Boxplot using Pandas



```
[11]: tips.boxplot(column = 'tip')            # create boxplot using pandas
      plt.title('Tip Histogram', size=20)    # add title
      plt.show()
```

Pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language.

Reference

- Wikipedia, "Box plot", https://en.wikipedia.org/wiki/Box_plot
- Michael Galarnyk, "Understanding Boxplots", https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51
- Saul McLeod, "What does a box plot tell you?", https://www.simplypsychology.org/boxplots.html
- Seaborn, "seaborn.boxplot", http://seaborn.pydata.org/generated/seaborn.boxplot.html?highlight=box
- Mohit Gupta_OMG, "Understanding different Box Plot with visualization", https://www.geeksforgeeks.org/understanding-different-box-plot-with-visualization/