Name: Kunal Sinha Roll no: 2K17/CO/164 Batch: CO-A3

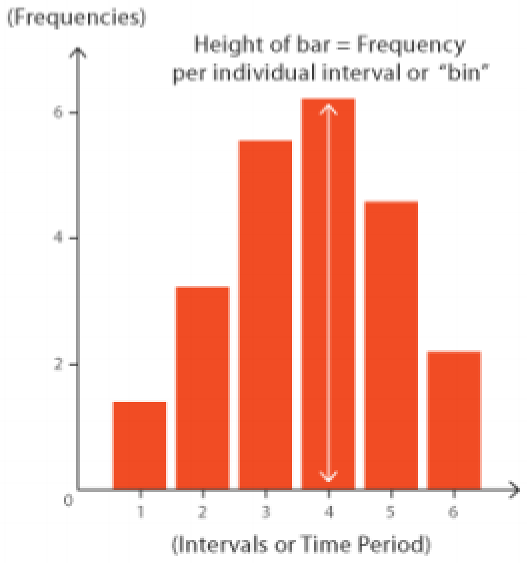
<u>Experiment 3</u>

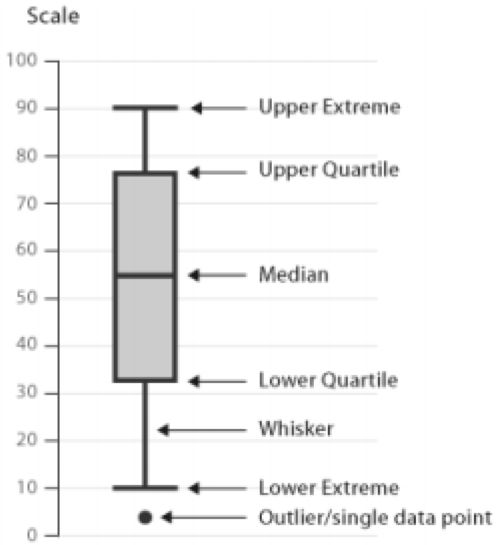
Aim: Generate Histogram and Boxplot for a sample program in WEKA.

Theory:

A histogram is a plot that lets you discover, and show, the underlying frequency distribution (shape) of a set of continuous data. This allows the inspection of the data for its underlying distribution (e.g., normal distribution), outliers, skewness, etc. A Histogram visualises the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency at each interval/bin. Histograms help give an estimate as to where values are concentrated, what the extremes are and whether there are any gaps or unusual values. They are also useful for giving a rough view of the probability distribution.

A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally. Although Box Plots may seem primitive in comparison to a Histogram or Density Plot, they have the advantage of taking up less space, which is useful when comparing distributions between many groups or datasets. Here are the types of observations one can make from viewing a Box Plot:





What the key values are, such as the average, median 25th percentile etc.

- If there are any outliers and what their values are.
- Is the data symmetrical.

- How tightly is the data grouped.
- If the data is skewed and if so, in what direction.

Two of the most commonly used variation of the Box Plot are variable-width Box Plots and notched Box Plots. A Five Number Summary includes:

- Minimum
- First Quartile
- Median (Second Quartile)
- Third Quartile
- Maximum

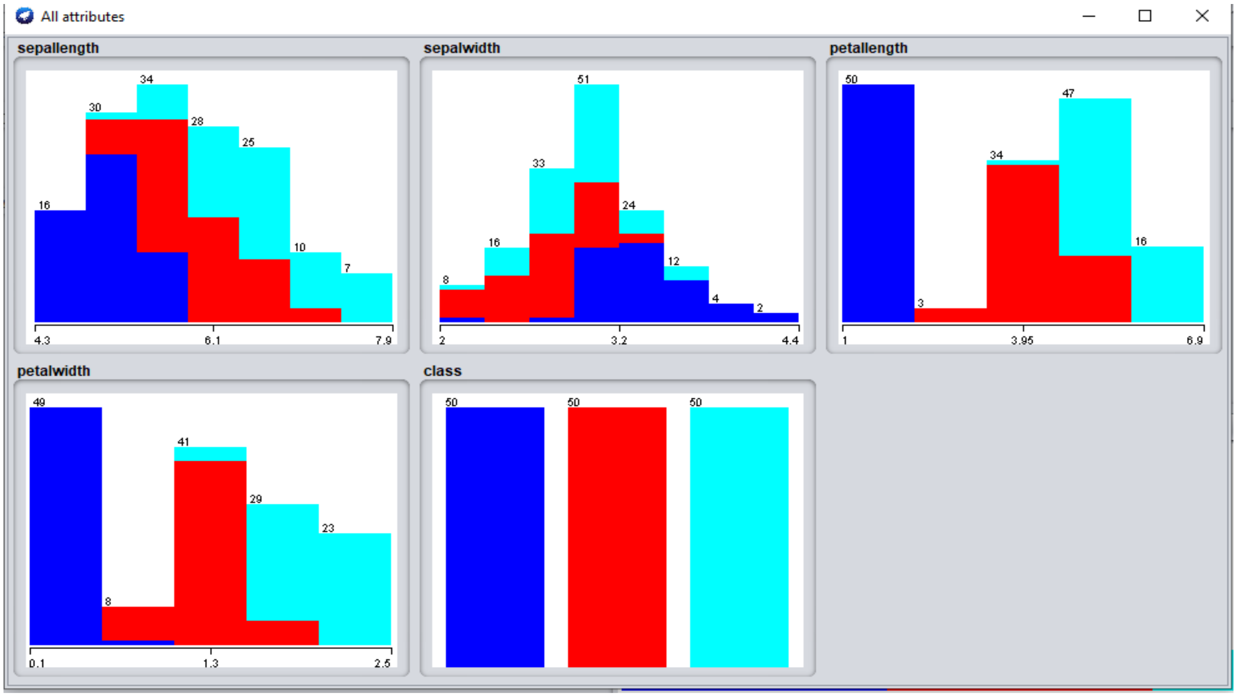
Procedure :

- Plotting histograms
  - Go to weka explorer.
  - Choose dataset in weka-3.8.3/data.
  - Above histogram, click visualize all.
- Plotting Box-Plots
  - Go to weka explorer
  - Choose dataset in weka-3.8.3/data
  - Use CPython scripting and pandas DataFrame boxplot method
  - Plot the boxplot using CPython

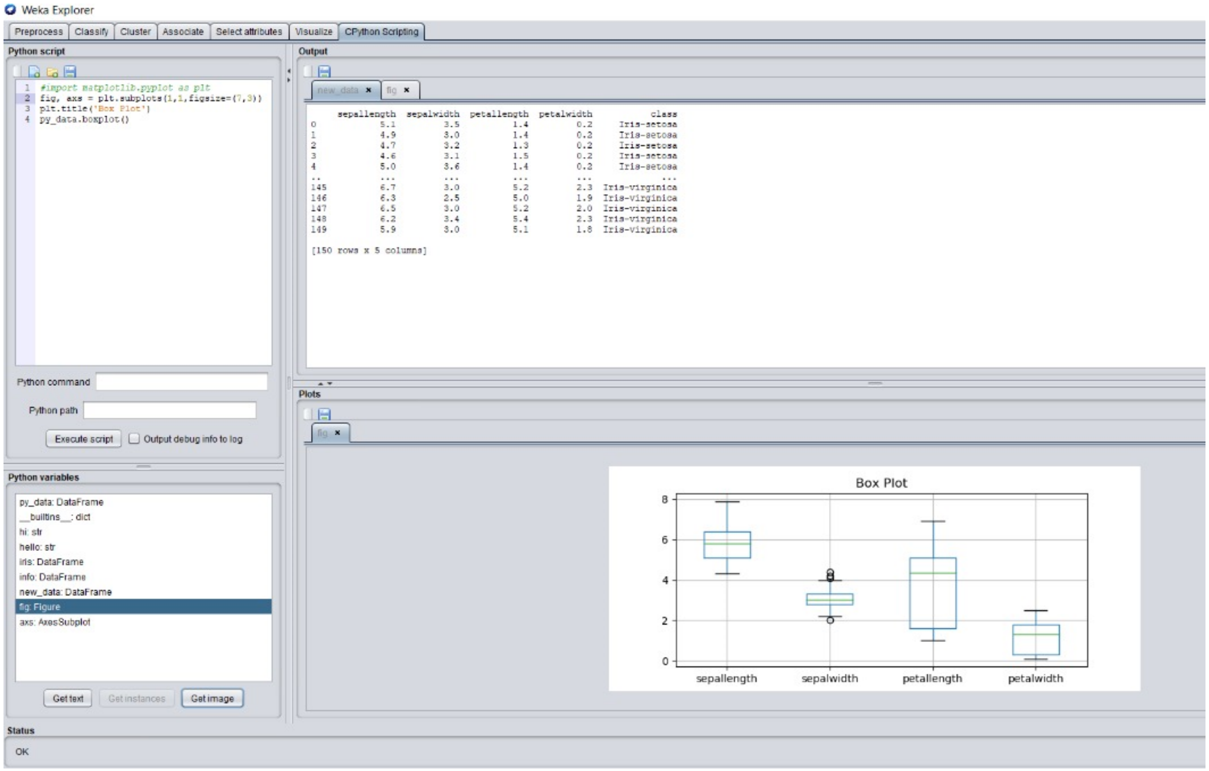
Output :

The following plots have been obtained for the dataset iris.arff

Histogram



Boxplot



```
#import matplotlib.pyplot as plt
fig, axs = plt.subplots(1,1,figsize=(7,3))
plt.title('Box Plot')
py_data.boxplot()
```

```
     sepallength  sepalwidth  petallength  petalwidth           class
0            5.1         3.5          1.4         0.2     Iris-setosa
1            4.9         3.0          1.4         0.2     Iris-setosa
2            4.7         3.2          1.3         0.2     Iris-setosa
3            4.6         3.1          1.5         0.2     Iris-setosa
4            5.0         3.6          1.4         0.2     Iris-setosa
..           ...         ...          ...         ...             ...
145          6.7         3.0          5.2         2.3  Iris-virginica
146          6.3         2.5          5.0         1.9  Iris-virginica
147          6.5         3.0          5.2         2.0  Iris-virginica
148          6.2         3.4          5.4         2.3  Iris-virginica
149          5.9         3.0          5.1         1.8  Iris-virginica

[150 rows x 5 columns]
```

Findings and Learnings :

Box plots and histograms are important tools to give an appropriate graphical representation of the raw data and hence are extensively used in data visualization. Weka Software provides a good set of functions to plot histograms and box plots with various combination of attributes. We have successfully plotted histograms and boxplots in WEKA