**Assignment – 1(b)**

**Note: This is a build from scratch assignment where you will have to do everything on your own right from importing the libraries, reading datasets etc. Please follow the instructions give below and complete the assignment. The marks corresponding to each task is mentioned.**

1. Iris Dataset (20 points)

It is always recommended to familiarize yourself with the data you intend to use for data mining purposes. The Iris dataset in particular has a rich history, having been introduced in 1936 by Sir Ronald Fisher, often considered one of the fathers of modern statistical theory.

You can find the csv file as iris.csv in Assignment-1(b) folder.

For any description of dataset visit the link: <http://archive.ics.uci.edu/ml/datasets/Iris>

* 1. **Summary statistics**

Print the first 5 elements of your DataFrame using the command head(). How many features are there and what are their types (e.g., numeric, nominal)? Compute and display summary statistics for each feature available in the dataset. These must include the minimum value, maximum value, mean, range, standard deviation, variance, count, and 25:50:75% percentiles.

* 1. **Data Visualization**

*1.2.1 Histograms*

To illustrate the feature distributions, create a histogram for each feature in the dataset. You may plot each histogram individually or combine them all into a single plot. When generating histograms for this assignment, use the default number of bins. Recall that a histogram provides a graphical representation of the distribution of the data.

*1.2.2 Boxplots*

To further assess the data, create a boxplot for each feature in the dataset. All of the boxplots will be combined into a single plot. Recall that a boxplot provides a graphical representation of the location and variation of the data through their quartiles; they are especially useful for comparing distributions and identifying outliers.

2. Pen-Based Handwritten Digits Dataset (20 points)

Repeat the same process described in Section 1 but this time using this dataset, which we will refer to as Digits. Note that Digits is a much larger dataset than Iris, both with respect to the number of instances and the number of features.

The dataset is present in Assignment-1(b) folder as pendigits.csv

For any description of dataset visit the link: <http://archive.ics.uci.edu/ml/datasets/Pen-Based+Recognition+of+Handwritten+Digits>

1. Conceptual Questions (20 points)

*Note: You have to answer these question in markdown cells in your jupyter notebook*

1. *Consider the histograms you generated for the Iris dataset.*
2. How do the shapes of the histograms for petal length and petal width differ from those for sepal length and sepal width?
3. Now consider just the petal width histogram. Is there a particular value of petal length (which ranges from 1.0 to 6.9) where the distribution of petal lengths (as illustrated by the histogram) could be best segmented into two parts?
4. Now consider the boxplots you generated for the Iris dataset. There should be four boxplots, one for each feature. Based upon these boxplots, is there a pair of features that appear to have significantly different medians?
5. Recall that the degree of overlap between variabilities is an important initial indicator of the likelihood that differences in means or medians are meaningful. Also, based solely upon the box plots, which feature appears to explain the greatest amount of the data?
6. *Lastly, consider the boxplots you generated for the Digits dataset.*
7. Do you observe any outliers? If so, for what features?
8. Now consider the corresponding histograms. What sort of distribution do the second and forth features display?
9. With that in mind, explain the outliers, or lack thereof, in terms of what you observe from the histograms.