Solution for Assignment 1 – B2

1.

Sort the number in ascending order-

2,3,4,4,5,5,6,6,6,7,8,8,8,8,9

Min=2

Max=9

Mean=89/15=5.93

Mode=8

Median=6

2.

Sorted data: 2,3,4,4,5,5,6,6,6,7,8,8,8,8,9

Assume bins=3

Bin 1: 2,3,4,4,5

Bin 2: 5,6,6,6,7

Bin 3: 8,8,8,8,9

Smoothing by bin means:

Bin 1: 3.6,3.6,3.6,3.6,3.6

Bin 2: 6,6,6,6,6

Bin 3: 8.2,8.2,8.2,8.2,8.2

Smoothing by bin medians:

Bin 1: 4,4,4,4,4

Bin 2: 6,6,6,6,6

Bin 3: 8,8,8,8,8

Smoothing by bin boundary-

Bin 1: 2,2,5,5,5

Bin 2: 5,5,5,5,7

Bin 3: 8,8,8,8,9

3.

Data pre-processing-

It contains the following major task-

I. Data Cleaning

    a. Dealing with missing values

    b. Dealing with noisy data

        1. Binning

        2. Regression

        3. Outlier Analysis

II. Data Integration

III. Data Reduction

    a. Dimensionality Reduction

    b. Numerosity Reduction

    c. Data Compression

IV. Data Transformation

4.

Regression- Data smoothing can also be done by regression. Regression is a technique that conforms data values to a function. i.e. A process that will find a particular mathematical function to represent the data. A linear regression involves finding the best line to fit two attributes or variables, so that one attribute can be used to predict the other.

Cluster Analysis- The outliers can be detected by clustering. The similar values or objects are organized into groups called as clusters. The object that fall outside the set of clusters can be considered as outliers. These outliers are treated as noise and can be eliminated further from the data mining process.