**Assignment No. 1 - Preprocessing**

**Course Name: Data Mining and Warehousing Course Code:** **DJ19CEC501**

**Class**: TY BTech **Sem:** V

**Data: 19/10/2023 Marks: 15**

**Note: Marks will be based on presentation, conceptually correct and uniqueness.**

|  |  |
| --- | --- |
| **QNo** | **Questions** |
| 1 | Suppose that the data for analysis includes the attribute *age*. The *age* values for the data  tuples are (in increasing order)  13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.  (a) What is the *mean* of the data? What is the *median*?  (b) What is the *mode* of the data? Comment on the data’s modality (i.e., bimodal, tri-modal, etc.).  (c) What is the *midrange* of the data?  (d) Can you find (roughly) the first quartile (*Q*1) and the third quartile (*Q*3) of the data?  (e) Give the *five-number summary* of the data.  (f) Show a *boxplot* of the data.  (g) How is a *quantile–quantile plot* different from a *quantile plot*? |
|  |  |
| 2. | Suppose that the values for a given set of data are grouped into intervals. The intervals and corresponding frequencies are as follows:    Compute an *approximate median* value for the data. |
|  |  |
| 3. | Suppose that a hospital tested the age and body fat data for 18 randomly selected adults  with the following results:    (a) Calculate the mean, median, and standard deviation of *age* and *%fat*.  (b) Draw the boxplots for *age* and *%fat*. |
|  |  |
| 4. | Given two objects represented by the tuples (22, 1, 42, 10) and (20, 0, 36, 8):  (a) Compute the *Euclidean distance* between the two objects.  (b) Compute the *Manhattan distance* between the two objects.  (c) Compute the *Minkowski distance* between the two objects, using *q* D 3.  (d) Compute the *supremum distance* between the two objects. |
|  |  |
| 5 | It is important to define or select similarity measures in data analysis. However, there is no commonly accepted subjective similarity measure. Results can vary depending on the similarity measures used. Nonetheless, seemingly different similarity measures may be equivalent after some transformation. Suppose we have the following 2-D data set:    (a) Consider the data as 2-D data points. Given a new data point, ***x*** = .1.4, 1.6/ as a query, rank the database points based on similarity with the query using Euclidean distance, Manhattan distance, supremum distance, and cosine similarity.  (b) Normalize the data set to make the norm of each data point equal to 1. Use Euclidean distance on the transformed data to rank the data points. |