MSSE – I –Advanced Data Mining

**Submitted To: Dr. Raheel Zafar**

**Submitted By: Uzair Javed**

**Class ID: NUML-F21-12331**

**Date: 01 December 2021**

Assignment 3

Due Date: 1st December 2021

Q: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,  
trimodal, 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* diffhat is the mean of the data? What is the median?

Ans:1

**(a)**  The (arithmetic) mean of the data is: ¯x = 809/27 = 30.

The median (middle value of the ordered set, as the number of values in the set is odd) of the data is: 25.

**(b)** This data set has two values that occur with the same highest frequency and is, therefore, bimodal.The modes (values occurring with the greatest frequency) of the data are 25 and 35.

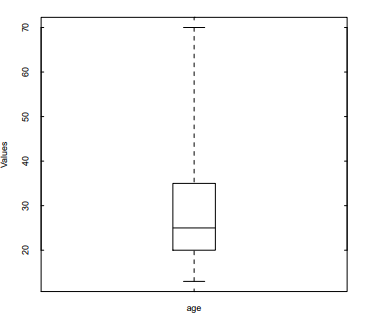
**(c)** The midrange (average of the largest and smallest values in the data set) of the data is:

(70 +13)/2 = 41.5

**(d)** The first quartile (corresponding to the 25th percentile) of the data is: 20. The third quartile

(Corresponding to the 75th percentile) of the data is: 35.

**(e)** The five number summary of a distribution consists of the minimum value, first quartile, median value, third quartile, and maximum value. It provides a good summary of the shape of the distribution and for this data is: 13, 20, 25, 35, and 70.

**(f)**

**(g)** A quantile plot is a graphical method used to show the approximate percentage of values below or equal to the independent variable in a univariate distribution. Thus, it displays quantile information for all the data, where the values measured for the independent variable are plotted against their corresponding quantile. A quantile-quantile plot however, graphs the quantiles of one univariate distribution against the corresponding quantiles of another univariate distribution. Both axes display the range of values measured for their corresponding distribution, and points are plotted that correspond to the quantile values of the two distributions. A line (y = x) can be added to the graph along with points representing where the first, second and third quantiles lie, in order to increase the graph’s informational value. Points that lie above such a line indicate a correspondingly higher value for the distribution plotted on the y-axis, than for the distribution plotted on the x-axis at the same quantile. The opposite effect is true for points lying below this line.

Q:2 In real-world data, tuples with *missing values* for some attributes are a common occurrence. Describe various methods for handling this problem.

Ans:2

The following are some of the approaches for dealing with missing values in data tuples: (a) Ignoring the tuple: When a class label is missing, this is normally done (assuming the mining task involves classification or description). Unless the tuple contains numerous characteristics with missing values, this approach is ineffective. It is especially poor when the percentage of missing values per attribute varies considerably. (b) Manually filling in the missing value: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined. (c) Using a global constant to fill in the missing value: Replace all missing attribute values by the same constant, such as a label like “Unknown,” or −∞. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common — that of “Unknown.” Hence, although this method is simple, it is not recommended. (d) Using a measure of central tendency for the attribute, such as the mean (for symtmetric numeric data), the median (for asymmetric numeric data), or the mode (for nominal data): For example, suppose that the average income of All Electronics customers is $28,000 and that the data are symmetric. Use this value to replace any missing values for income. (e) Using the attribute mean for numeric (quantitative) values or attribute mode for nominal values, for all samples belonging to the same class as the given tuple.