**Module 2 assignment 1**

**1.**

A = [77,86,78,79,70]

Mean = (77+86+78+79+70)/5 = 78

Mode = 78

Median = no mode

Standard deviation = sqrt((xi-U)^2/N) = 5.09

**2.**

Chart, line chart

Description automatically generated

**3.**

Data can be defined as a systematic record of a particular quantity. It is the different values of that quantity represented together in a set. It is a collection of facts and figures to be used for a specific purpose such as a survey or analysis. When arranged in an organized form, can be called information. The source of data ( primary data, secondary data) is also an important factor.

**Types of Data**

Data may be qualitative or quantitative. Once you know the difference between them, you can know how to use them.

**Qualitative Data:** They represent some characteristics or attributes. They depict descriptions that may be observed but cannot be computed or calculated. For example, data on attributes such as intelligence, honesty, wisdom, cleanliness, and creativity collected using the students of your class a sample would be classified as qualitative. They are more exploratory than conclusive in nature.

**Quantitative Data:** These can be measured and not simply observed. They can be numerically represented and calculations can be performed on them. For example, data on the number of students playing different sports from your class gives an estimate of how many of the total students play which sport. This information is numerical and can be classified as quantitative.

**4.**

**Data Cleaning**

The data cleaning process detects and removes the errors and inconsistencies present in the data and improves its quality. Data quality problems occur due to misspellings during data entry, missing values or any other invalid data. Basically, “dirty” data is transformed into clean data. “Dirty” data does not produce the accurate and good results. Garbage data gives garbage out. So it becomes very important to handle this data. Professionals spend a lot of their time on this step.

**Reasons for “dirty” or “unclean” data**

Dummy values

Absence of data

Violation of business rules

Data integration problems

Contradicting data

Inappropriate use of address line

Reused primary keys

Non-unique identifiers

What to do to clean data?

Handle Missing Values

Handle Noise and Outliers

Remove Unwanted data

**Handle Missing Values**

Missing values cannot be looked over in a data set. They must be handled. Also, a lot of models do not accept missing values. There are several techniques to handle missing data, choosing the right one is of utmost importance. The choice of technique to deal with missing data depends on the problem domain and the goal of data mining process. The different ways to handle missing data are:

**Ignore the data row:** This method is suggested for records where maximum amount of data is missing, rendering the record meaningless. This method is usually avoided where only less attribute values are missing. If all the rows with missing values are ignored i.e. removed, it will result in poor performance.

Fill the missing values manually: This is a very time consuming method and hence infeasible for almost all scenarios.

Use a global constant to fill in for missing values: A global constant like “NA” or 0 can be used to fill all the missing data. This method is used when missing values are difficult to be predicted.

Use attribute mean or median: Mean or median of the attribute is used to fill the missing value.

Use forward fill or backward fill method: In this, either the previous value or the next value is used to fill the missing value. A mean of the previous and succession values may also be used.

Use a data-mining algorithm to predict the most probable value

**Handle Noise and Outliers**

Noise in data may be introduced due to fault in data collection, error during data entering or due to data transmission errors, etc. Unknown encoding (Example : Marital Status — Q), out of range values (Example : Age — -10), Inconsistent Data (Example : DoB — 4th Oct 1999, Age — 50), inconsistent formats (Example : DoJ — 13th Jan 2000, DoL — 10/10/2016), etc. are different types of noise and outliers.

Noise can be handled using binning. In this technique, sorted data is placed into bins or buckets. Bins can be created by equal-width (distance) or equal-depth (frequency) partitioning. On these bins, smoothing can be applied. Smoothing can be by bin mean, bin median or bin boundaries.

Outliers can be smoothed by using binning and then smoothing it. They can be detected using visual analysis or boxplots. Clustering can be used identify groups of outlier data.The detected outliers may be smoothed or removed.

**5.**

A z-score, or standard score, is used for standardizing scores on the same scale by dividing a score's deviation by the standard deviation in a data set. The result is a standard score. It measures the number of standard deviations that a given data point is from the mean. A z-score can be negative or positive.

**6.**

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide an accessible way to see and understand trends, outliers, and patterns in data.