SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

Course Name – 23ADT202 Fundamental of Data science and Analytics

II Year / IV Semester

Unit 2 – Descriptive Analytics

Averages

Averages are measures of central tendency that summarize datasets. They represent the central value that best describes the dataset. Averages provide quick insights into the typical behavior of data. They are commonly used in statistics, business, and various fields of analysis.

The **mean** is the arithmetic average of a dataset. It is calculated by summing all values and dividing by the number of values. The mean is widely used in symmetric datasets without outliers. Example: Mean of [10, 20, 30, 40, 50] is 30.

The **median** is the middle value in an ordered dataset. For an even number of data points, the median is the average of the two middle values. The median is ideal for skewed data or when outliers are present. Example: Median of [10, 20, 30, 40, 50] is 30.
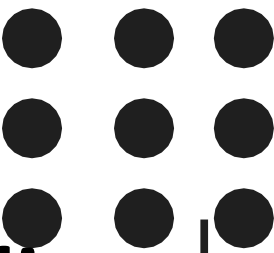
The **mode** is the most frequent value in a dataset. It can be useful for identifying trends in categorical data. A dataset can have one mode, multiple modes, or no mode at all. Example: Mode of [10, 20, 20, 30, 30, 30, 40] is 30.

Comparison of Averages

The **mean** is sensitive to outliers and may not represent skewed data well. The **median** is more robust and can better represent skewed datasets. The **mode** is primarily used for categorical data. Each average has its specific use depending on data characteristics.

Visualizing Averages

Averages can be visualized to better understand data central tendency. Graphs such as bar charts or box plots can help show the location of the mean, median, and mode. Visualization aids in identifying trends and outliers in the data.

Use the **mean** for symmetric datasets without outliers. The **median** is preferred for skewed datasets or when extreme values exist. The **mode** is best for categorical data where frequency is key. Each measure serves a different purpose depending on data type.
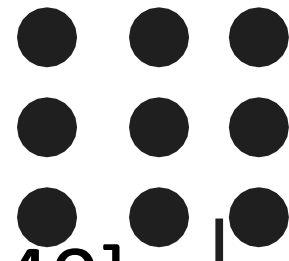
In business, use the **mean** for revenue forecasts and **median** for typical customer spending. In education, **mean** scores evaluate student performance. In healthcare, the **median** recovery time is used for hospital planning, especially when outliers are involved.

Exercise 1: Calculate the mean, median, and mode for [15, 20, 25, 30, 30, 35, 40].
Exercise 2: Identify the best measure of central tendency for a dataset with outliers, e.g., [10, 20, 30, 1000].
In conclusion, selecting the correct average depends on the dataset's nature, and each type provides valuable insights into data behavior.

THANK YOU