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

Outliers

What Are Outliers?

Outliers are data points that deviate significantly from the overall pattern of a dataset. They can heavily distort statistical analyses and modeling results. Identifying them is essential to ensure accurate predictions. For example, in [10, 12, 14, 16, 100], 100 is an outlier.

Types of Outliers - Univariate Outliers

Univariate outliers occur when a data point is extreme in one variable. For example, in a set of heights [160, 165, 170, 250], the value 250 is an outlier. These outliers are easy to spot in simple datasets with one variable. Statistical methods like Z-scores can help detect them.

Types of Outliers - Multivariate Outliers

Multivariate outliers are extreme values when considering multiple variables together. An example is a person with height = 200 cm and weight = 30 kg, which may be an outlier in a dataset of normal human proportions. Identifying these requires multidimensional analysis techniques.

Types of Outliers - Contextual Outliers

Contextual outliers are unusual data points within a specific context. For example, a temperature of 30°C is normal in summer but unusual in winter. These outliers depend on the environmental context and are typically identified using time series analysis or other contextual methods.

Causes of Outliers - Measurement Errors

Measurement errors can lead to outliers, such as incorrect data entry or faulty sensors. For example, entering a value of 1000 instead of 100 can produce an outlier. These errors are typically easy to detect and correct once identified.

Causes of Outliers - Natural Variability

Outliers can also occur naturally in datasets due to extreme but valid values. For instance, exceptional athletes in sports data might appear as outliers. These outliers are often valid but may need special treatment depending on the context of the analysis.

Methods to Detect Outliers - Visual Inspection

Visual inspection helps identify outliers using graphs like boxplots and scatter plots. Boxplots can reveal univariate outliers, while scatter plots help spot multivariate outliers. This approach is simple but can be subjective and limited with large datasets.

Methods to Detect Outliers - Statistical & ML Approaches

Statistical methods like Z-scores and Interquartile Range (IQR) are commonly used for detecting outliers. Machine learning methods such as Isolation Forest and DBSCAN are more advanced and can detect outliers in complex datasets by isolating points or identifying noise.

Handling Outliers

Outliers can be handled in several ways: they can be removed, transformed (e.g., with logarithms), imputed with mean or median values, or handled using models that are less sensitive to outliers. The approach depends on the cause of the outlier and the dataset's nature.

THANK YOU