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

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 3 – PREDICTIVE MODELING AND MACHINE LEARNING

Topic 5: Detecting outliers

Detecting Outliers

•Outliers are those data points that are really far from the rest of your data points. In other words an outlier is a value or data point that differs substantially from the rest of the data.

Reasons for outliers in data

•Errors during data entry or a faulty measuring device (a faulty sensor may result in extreme readings).

•Natural occurrence

Box plots

•Box plots are a visual method to identify outliers. Box plots is one of the many ways to visualize data distribution.

•Box plot plots the q1 (25th percentile), q2 (50th percentile or median) and q3 (75th percentile) of the data along with (q1–1.5*(q3-q1)) and (q3+1.5*(q3-q1)).

•Outliers, if any, are plotted as points above and below the plot.

Detecting Outliers

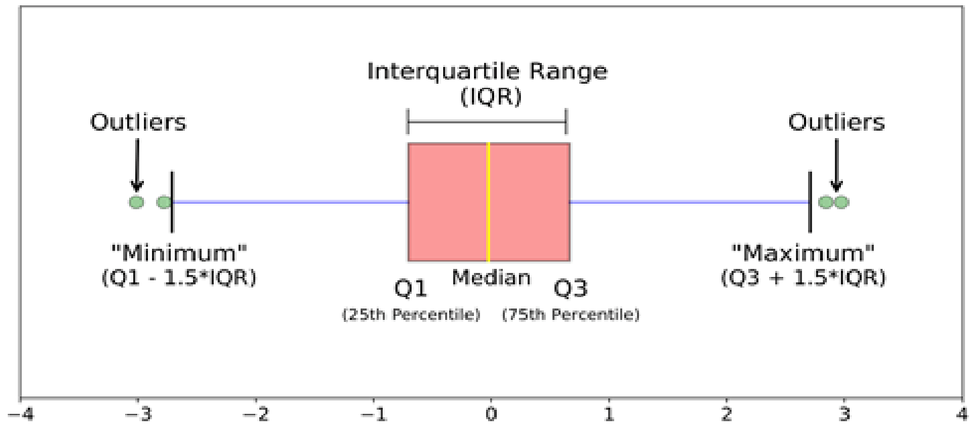
IQR method

IQR method is used by box plot to highlight outliers. IQR stands for interquartile range, which is the difference between q3 (75th percentile) and q1 (25th percentile).

The IQR method computes lower bound and upper bound to identify outliers.

Lower Bound = q1–1.5*IQR

Upper Bound = q3+1.5*IQR



IQR to detect outliers

Detecting Outliers

Z-score method

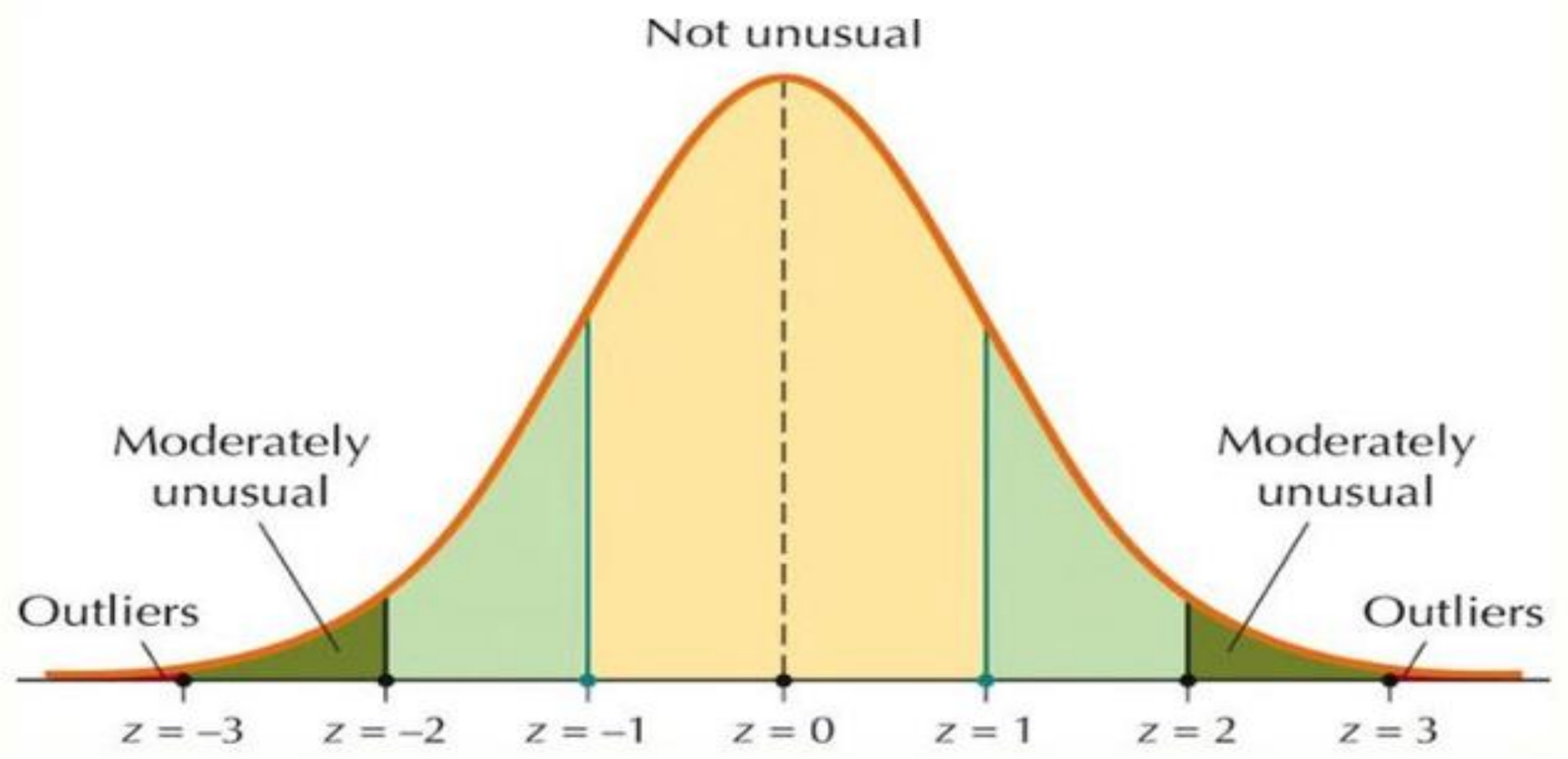
•Z-score method is another method for detecting outliers. This method is generally used when a variable' distribution looks close to Gaussian.

•Z-score is the number of standard deviations a value of a variable is away from the variable' mean. Z-Score = (X-mean) / Standard deviation

•when the values of a variable are converted to Z-scores, then the distribution of the variable is called standard normal distribution with mean=0 and standard deviation=1.

•The Z-score method requires a cut-off specified by the user, to identify outliers. The widely used lower end cut-off is -3 and the upper end cut-off is +3.

Detecting Outliers



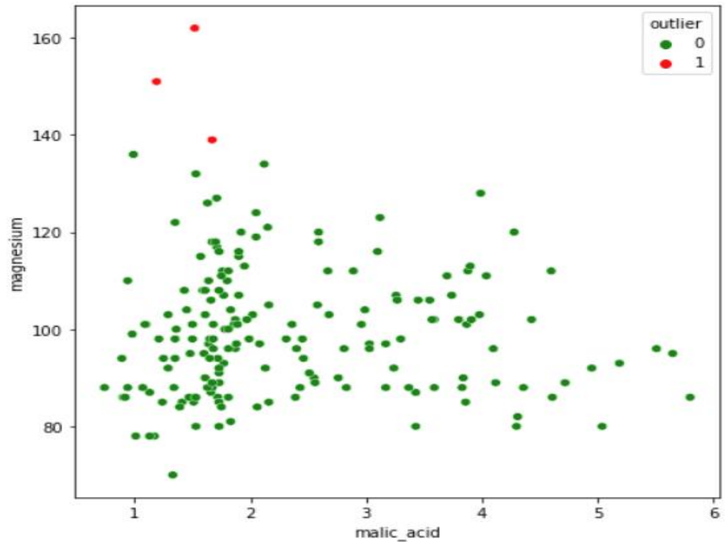
Detecting Outliers

Distance from the mean' method (Multivariate method)

•Unlike the previous methods, this method considers multiple variables in a data set to detect outliers.

•This method calculates the Euclidean distance of the data points from their mean and converts the distances into absolute z-scores.

•Any z-score greater than the pre-specified cut-off is considered to be an outlier.



THANK YOU