BMS INSTITUTE OF TECHNOLOGY AND MANAGEMENT

(An Autonomous Institution, Affiliated to VTU Belagavi)

Department of MCA

(Accredited by NBA, New Delhi)

CASE STUDY - FIRST INTERNAL EXAMINATION, JULY 2023

Data Analytics – Outlier Detection and Treatment

- One of the most important steps as part of data pre-processing is detecting and treating the outliers as they can negatively affect the statistical analysis.
- An Outlier is an observation in a given dataset that lies far from the rest of the observations.
- In other words, outliers are extreme observations that are very dissimilar to the rest of the population.
- An outlier may occur due to the variability in the data, or due to experimental error/human error. They may indicate an experimental error or heavy skewness in the data.
- In general, two types of outliers can be considered:
  Outliers are generally classified into two types: **Univariate** and **Multivariate**.

  Univariate Outliers – A univariate outlier is a case with an extreme value that falls outside the expected population values for a **single variable. Example Marks scored by students.**

  Multivariate Outliers – A multivariate outlier is a combination of unusual scores on at least two variables. **Example income and age**

  Both types of outliers can influence the outcome of statistical analyses.

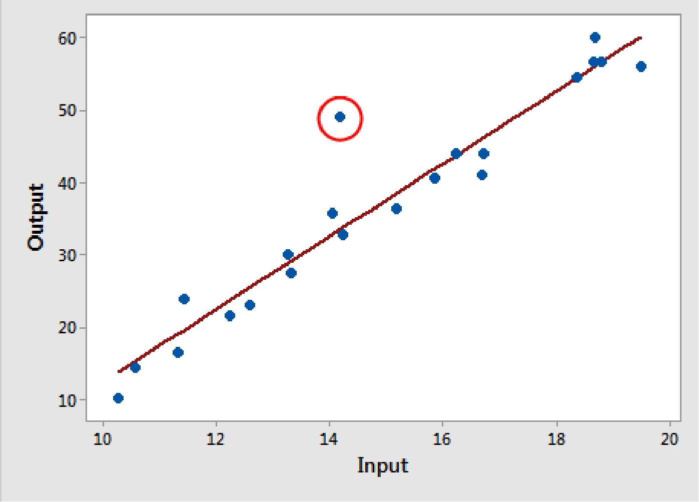


Figure: Univariate Outliers

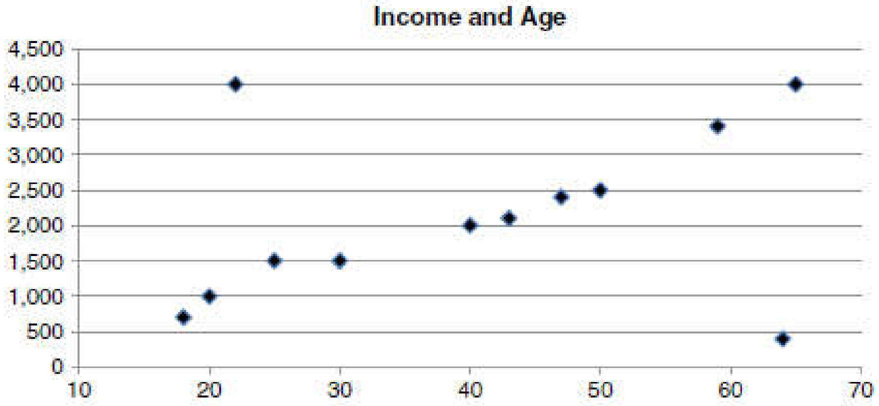


Figure: Multivariate Outliers

- Two important steps in dealing with outliers are detection and treatment. A first obvious check for outliers is to calculate the minimum and maximum values for each of the data elements.

Detecting Outliers:

- If our dataset is small, we can detect the outlier by just looking at the dataset. But what if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques.
- Univariate outliers can be detected using Histograms, Boxplots, Z-Scores while multivariate outliers can be detected by fitting regression lines and inspecting the observations with large errors (using, for example, a residual plot).
- Below are the techniques of detecting univariate outliers
  1. **Histograms**
  2. **Boxplots**
  3. **Z-score**

- A histogram is a graphical representation of data points organized into user-specified ranges. Similar in appearance to a bar graph, the histogram condenses a data series into an easily interpreted visual by taking many data points and grouping them into logical ranges or bins. The figure shown below presents an example of a distribution for age whereby the circled areas clearly represent outliers.

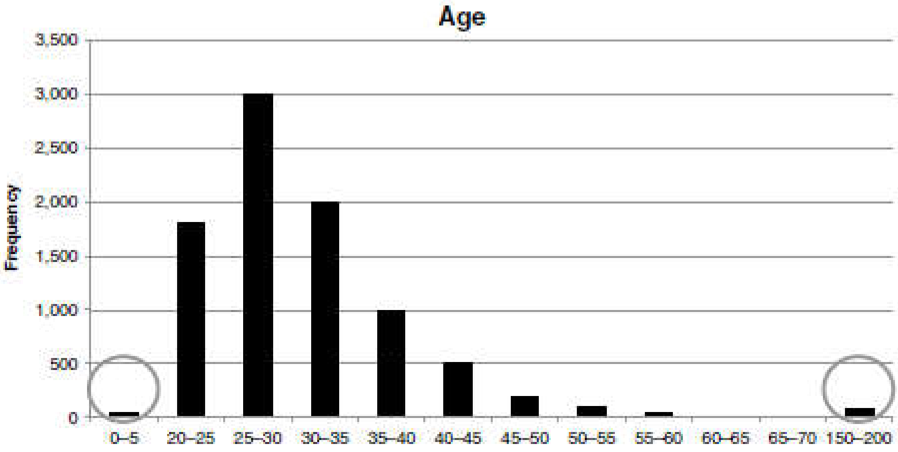
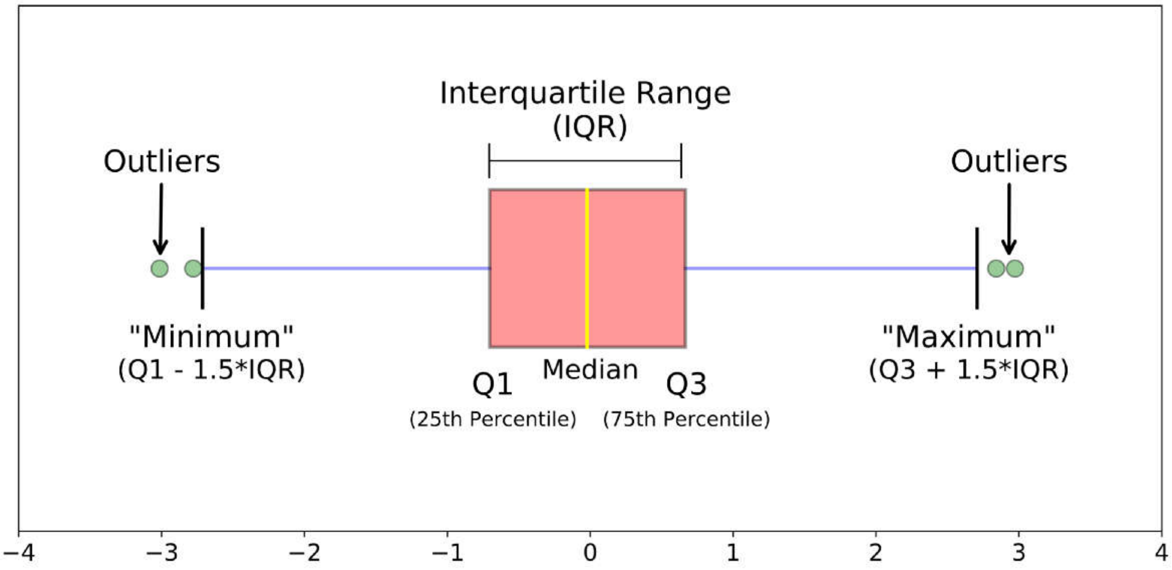


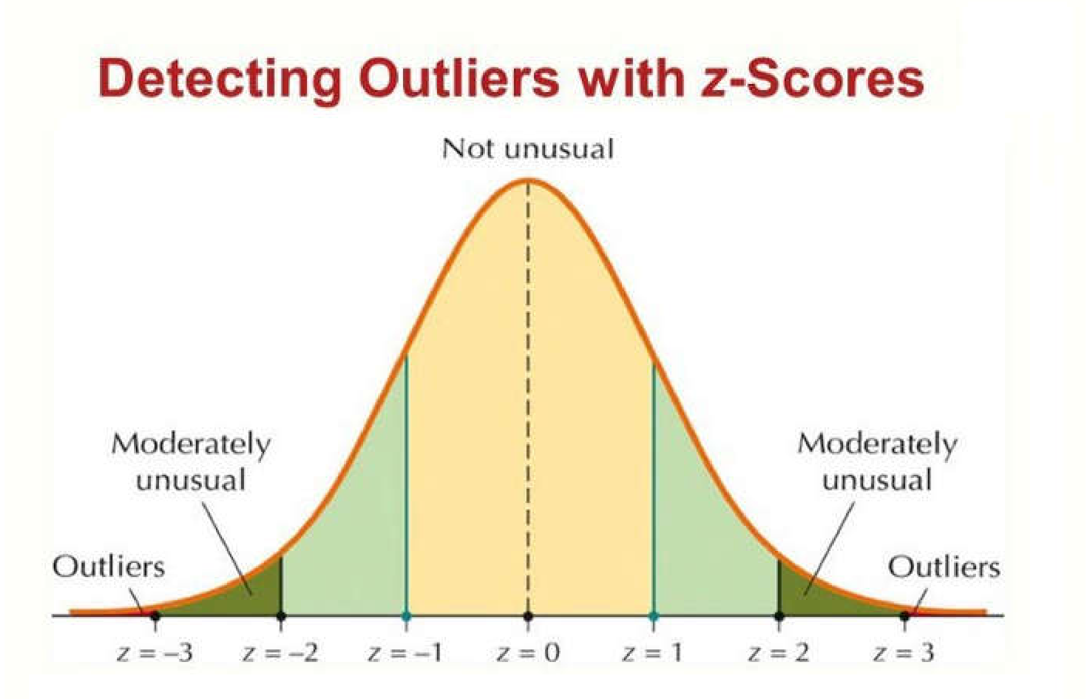
Figure: Histograms for Outlier Detection

- Another useful visual mechanism are ***Box plots***. Box plot is a data visualization plotting function. It shows the min, max, median, first quartile and third quartile. First quartile (25 percent of the observations have a lower value), the median (50 percent of the observations have a lower value), and the third quartile (75 percent of the observations have a lower value). All three quartiles are represented as a box. The minimum and maximum values are then also added unless they are too far away from the edges of the box.
- Outliers in Box Plots are then quantified as more than (1.5 * IQR) where Inter quartile Range IQR = Q3 − Q1.



- Another way is to calculate *Z-**scores***, measuring how many standard deviations an observation lies away from the mean, as follows:

$$z_i=(x_i-\mu)/\sigma$$



where **μ** represents the Mean of the variable and σ its standard deviation.

- A practical rule of thumb then defines outliers when the absolute value of the *z*-score |z| is bigger than 3. Note that the *z* score relies on the normal distribution.
- Some analytical techniques (e.g., decision trees, neural networks, Support Vector Machines (SVMs)) are fairly robust with respect to outliers.
  Others (e.g., linear/logistic regression) are more sensitive to outliers.

Treating Outliers:

- Various schemes exist to deal with outliers.
- It highly depends on whether the outlier represents a valid or invalid observation.
- For invalid observations (e.g., age is 300 years), one could treat the outlier as a missing value using any of the schemes discussed.
- For valid observations (e.g., income is $1 million), popular schemes like trimming the outlier or truncation/capping are used.