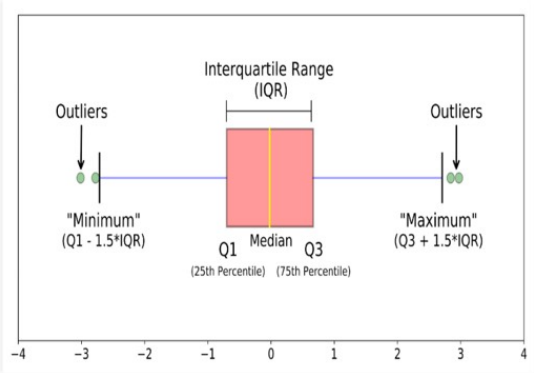
Outliers:

Detection:

→ scatter plot

→ Box-plot



Removal:

→ Z-score

→ using IQR

→ Imputation

→ Transformation ⇒ log

→ Deletion

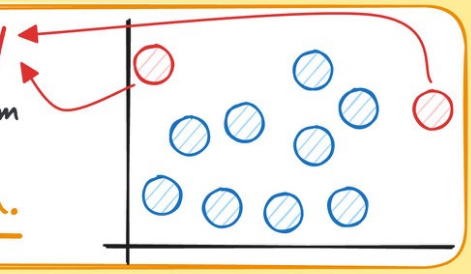
databites.tech @rfeers Josep Ferrer

Analytics Bites - Dealing with Outliers

#1 What is an outlier?

Outliers are data points that are very different from the rest of the dataset.

They are unrepresentative of the overall data.



#2 Why to detect outliers?

There are 2 main reasons:

To REMOVE them

They bias patterns and distort the learning process of ML algorithms.

Examples:
- Trai ML models.
- Forecasting.

To IDENTIFY them

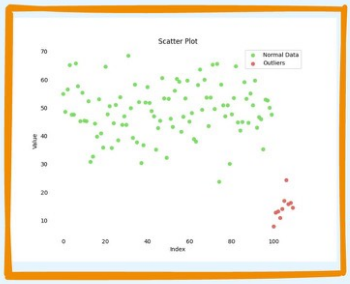
It is our primary goal to identify anomalies.

Examples:
- Fraud Detection
- Quality Control
- Anomalies

#3 How to detect them?

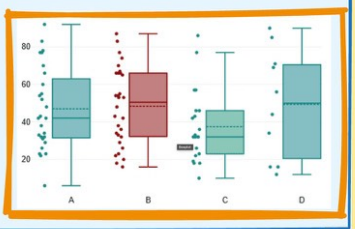
APPROACH I: GRAPHICAL

Detect outliers through data visualizations.



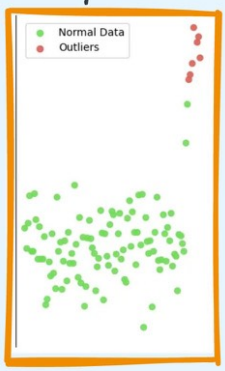
In a Scatter Plot, outliers appear as points significantly deviating from the main cluster.

In a Box Plot, outliers appear as points outside the "whiskers"

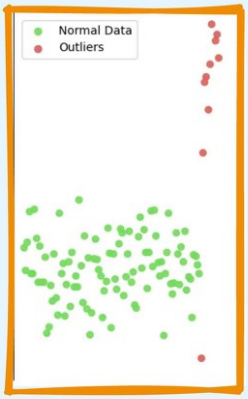


APPROACH II: GRAPHICAL

Detect outliers through statistical analysis and probability distributions.



Z-score measures how many standard deviations a data point lies away from the mean.



IQR relies on the range between the first and third quartiles of the data distribution. Any data beyond this range, are identified as outliers.

APPROACH III: ML Models

Detect outliers using ML models. It leverages the power of ML to overcome the limitations of the simpler Statistical Approaches.

Some examples are:
- Isolation Forest
- Elliptic Envelope
- LOF
- SVMs

Normal

Outliers