Day-41 (100 days of ML)

Outliers

1. **What are Outliers?**
   - Outliers are a data point that is essentially a statistical anomaly, a data point that significantly deviates from other observations in a dataset.
   - Outliers can arise due to measurement errors, natural variation, or rare events, and they can have a disproportionate impact on statistical analyses and machine learning models if not appropriately handled.

2. **When is it dangerous?**
   - Outliers are dangerous when they are data entry errors or represent genuine extreme values that, if not handled properly, can distort statistical measures (like the mean), skew models, ruin visualizations, and lead to incorrect conclusions in research or analyses. However, not all outliers are "bad"—they can also reveal important new information or anomalies, so their true danger depends on the context of the data and the goal of the analysis.

3. **Effect of Outliers on ML Algorithm**
   - Distortion of statistical metrics
   - Reduced model accuracy
   - Model overfitting
   - Prolonged training times

   Algorithm-Specific Sensitivity:
   - Sensitive Algorithms: Algorithms like Linear Regression, Logistic Regression, K-Means Clustering, and DBSCAN are generally more sensitive to outliers as they rely on distance metrics or assume a specific data distribution.
   - Robust Algorithms: Some algorithms, such as Tree-based models (e.g., Random Forest, Gradient Boosting), Support Vector Machines (SVMs) with certain kernel functions, and robust statistical methods, can be more robust to outliers as they are less affected by extreme values or have mechanisms to mitigate their influence.

4. **How to treat Outliers?**
   - Removal (Trimming):
     If outliers are due to data entry errors or don't represent the underlying process, you can remove the rows containing them. However, be cautious as this can lead to loss of information.

   - Transformation:
     Apply a mathematical function, such as a logarithmic transformation, to reduce the spread of the data and lessen the impact of outliers.

   - Imputation (Missing value):
     Replace outliers with a more central value, such as the median (which is resistant to outliers) or the mean.

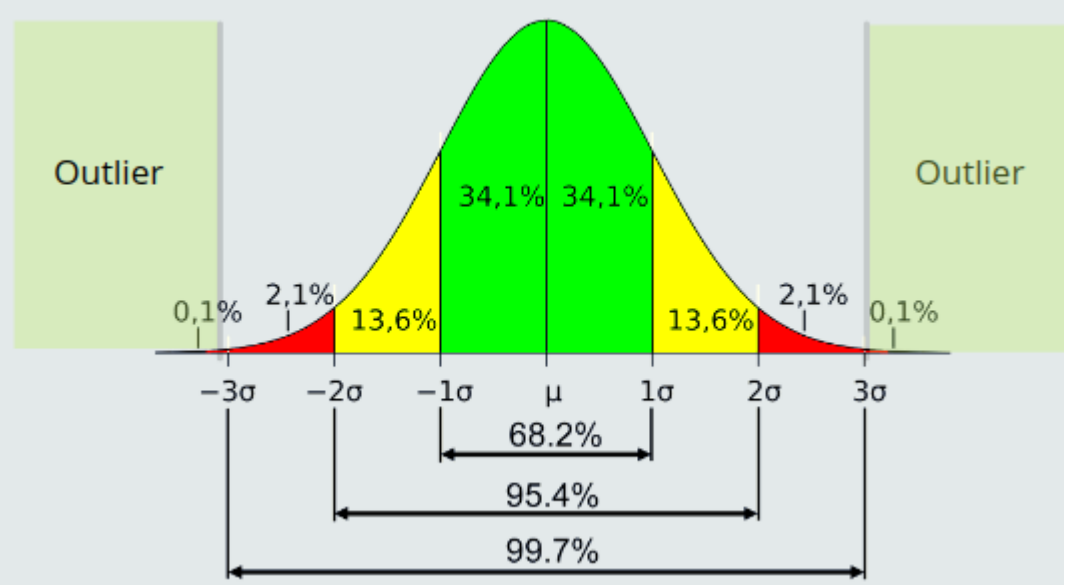
- Capping/Winsorizing:
  Modify outlier values by replacing them with a specified maximum or minimum value within the acceptable range.

- Robust Estimation:
  Use statistical methods that are inherently less sensitive to extreme values, such as using the median and IQR instead of the mean and standard deviation for analysis.

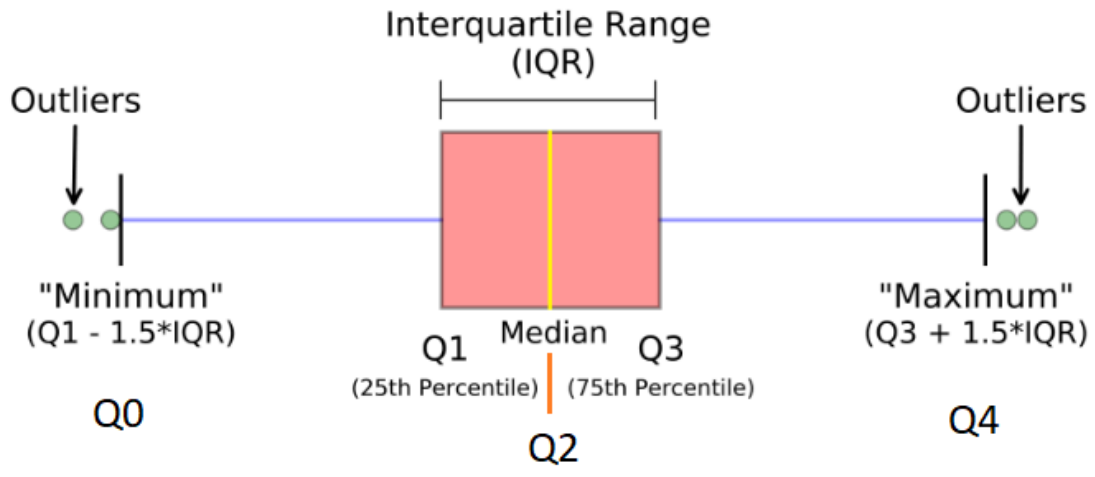
- Separate Analysis:
  Sometimes, outliers hold valuable insights and should be analyzed separately to uncover unique patterns or extreme but important cases.

5. How to detect Outliers ?

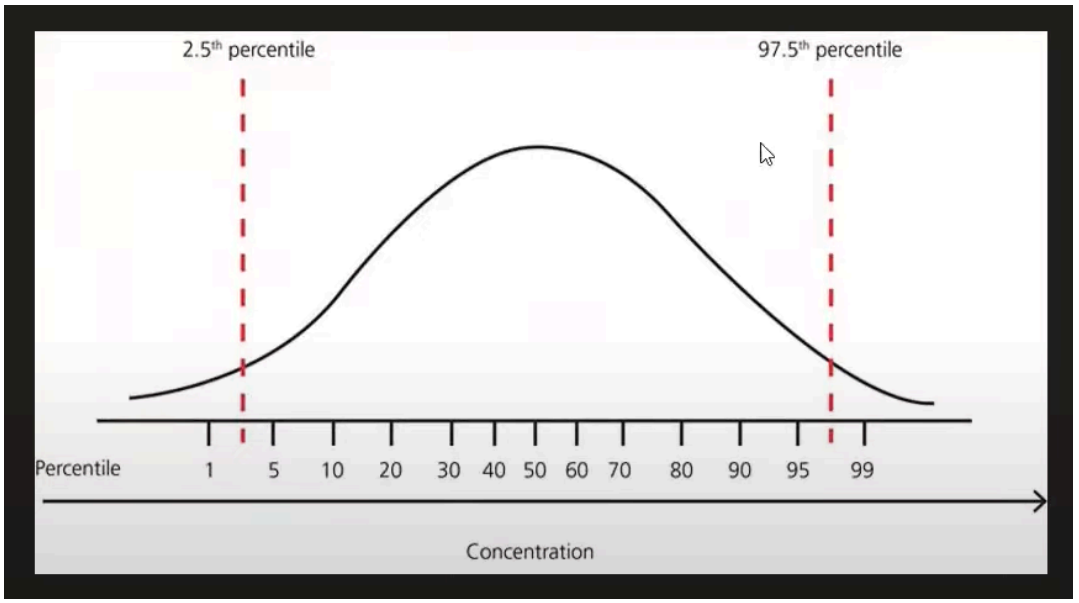
- Normal Distribution



- Skewed Distribution



- Other distribution



6. Techniques for Outliers detection and removal

- Z-score Method
  Calculates a Z-score for each data point, representing its distance from the mean in terms of standard deviations. Data points with a Z-score above a certain threshold (e.g., > 3) are flagged as outliers.

- Interquartile Range (IQR) Method
  Calculates the difference between the third quartile (Q3) and the first quartile (Q1). Data points below Q1 - 1.5IQR or above Q3 + 1.5IQR are considered outliers.

- Visualization
  - Box Plots: Visually display outliers as points falling outside the box plot's whiskers, which are set at the 1.5*IQR bounds.
  - Scatter Plots: Help identify outliers by showing any data points that fall outside the general pattern of the other points.

Removal/Treatment Techniques

- Trimming/Deletion
  Remove outlier observations from the dataset, best used when outliers are due to errors and few in number.

- Capping/Winsorisation
  Replace extreme outlier values with a specified maximum or minimum value (e.g., the 1.5*IQR boundary) to limit their impact.

- Imputation
  Replace outlier values with estimated values, such as the mean, median, or values predicted by a model.
  Robust Estimation