📊 What is Data Analysis?

Data Analysis is the process of: **[ICTM]**

```
🔍 Inspecting,

🧹 Cleaning,

🔄 Transforming, and

🧠 Modeling data
```

- Business: Customer segmentation, demand forecasting, fraud detection.
- Healthcare: Predicting disease outbreaks, patient diagnostics, personalized treatments.
- Finance: Credit risk assessment, stock price prediction, fraud detection.
- Science & Research: Climate change analysis, genomics, space exploration.
- Social Media & Marketing: Sentiment analysis, targeted advertising, user behavior insights.

📊 Types of Data Analysis [DaDdy PaPi]

1️⃣ Descriptive Analysis

```
Purpose: Understand what has happened in the past.

Examples: Averages, totals, counts, histograms, summary statistics.

📘 "What happened?"
```

2️⃣ Diagnostic Analysis

```
Purpose: Discover why something happened.

Methods: Drill-down, correlations, trend patterns.

📘 "Why did it happen?"
```

3️⃣ Predictive Analysis

```
Purpose: Forecast what is likely to happen in the future.

Techniques: Regression, classification, machine learning models.

📘 "What could happen?"
```

4️⃣ Prescriptive Analysis

```
Purpose: Suggest what actions to take for optimal outcomes.

Tools: Optimization algorithms, decision trees, scenario analysis.

📘 "What should we do?"
```

📚 Understanding Data

🧩 Types of Data

1️⃣ Structured Data

```
🔹 Definition: Well-organized and stored in tabular formats.

🗂️ Examples:

    Relational databases (MySQL, PostgreSQL)

    Excel spreadsheets
```

2️⃣ Unstructured Data

```
🔹 Definition: Data without a predefined model or organization.

🖼️ Examples:

    Images, videos

    Social media posts

    Emails
```

3️⃣ Semi-structured Data

```
🔹 Definition: Contains some structure, but not in a strict tabular format.

🔸 Examples:

    JSON

    XML
```

📏 Quantitative vs. Qualitative Data

🔢 Quantitative Data

```
Definition: Numerical data that can be measured and analyzed statistically.

Examples: Revenue, temperature, age
```

📝 Qualitative Data

```
Definition: Descriptive data representing categories or labels.

Examples: Gender, nationality, product category
```

🔹 Nominal Data

```
Categories without order

Examples: Colors, countries, departments
```

🔸 Ordinal Data

```
Categories with a meaningful order

Examples: Customer satisfaction (low, medium, high), education level
```

3. Data Preprocessing

Data preprocessing is a crucial step to ensure data quality and reliability before analysis.

Steps in Data Preprocessing

1. **Data Cleaning**
   Removing inaccuracies, fixing typos, and handling inconsistencies.
2. **Handling Missing Values**
   Identifying and addressing missing data points.
3. **Removing Duplicates**
   Ensuring data integrity by removing redundant records.
4. **Outlier Detection and Handling**
   Identifying extreme values that could skew analysis.
5. **Feature Scaling and Transformation**
   Normalizing or standardizing data for model compatibility.

Data Cleaning

- Fix typos (`"N/A"` vs `"NA"` vs `"null"` vs empty strings `""` convert them to `NAN`).

- Correct formatting (e.g., date formats like DD/MM/YYYY → YYYY-MM-DD).
- **Case Inconsistencies**
  Values like `"Yes"`, `"yes"`, `"YES"` should be unified.
  - **Fix:** Convert all text to lowercase or uppercase.
- **Spelling Errors and Typos**
  Common in categorical data (e.g., `"Male"`, `"male"`, `"mael"`).
  - **Fix:** Use mapping or fuzzy matching to correct.
- - **Whitespace and Formatting Issues**
    Extra spaces before/after text can cause mismatches.
    - **Fix:** Strip whitespace from strings.
- - **Incorrect Data Types**
    Numeric columns stored as strings or dates not in datetime format.
    - **Fix:** Convert columns to appropriate types.

Handling Missing Values

- **Causes:** Data entry errors, system failures, data corruption.
- **Methods to Handle Missing Data:**
  - **Deletion:** Remove rows or columns with missing values
    Example: `df.dropna()`
  - **Imputation:** Fill missing values using mean, median, mode, or interpolation
  - **Forward Fill / Backward Fill:** Use previous or next values in time series data
  - **Predictive Imputation:** Use machine learning models to estimate missing values

---

Removing Duplicates

- Duplicate records can arise from multiple data entry points or merging datasets.
- Use the following to eliminate redundant data:
  `df.drop_duplicates(inplace=True)`

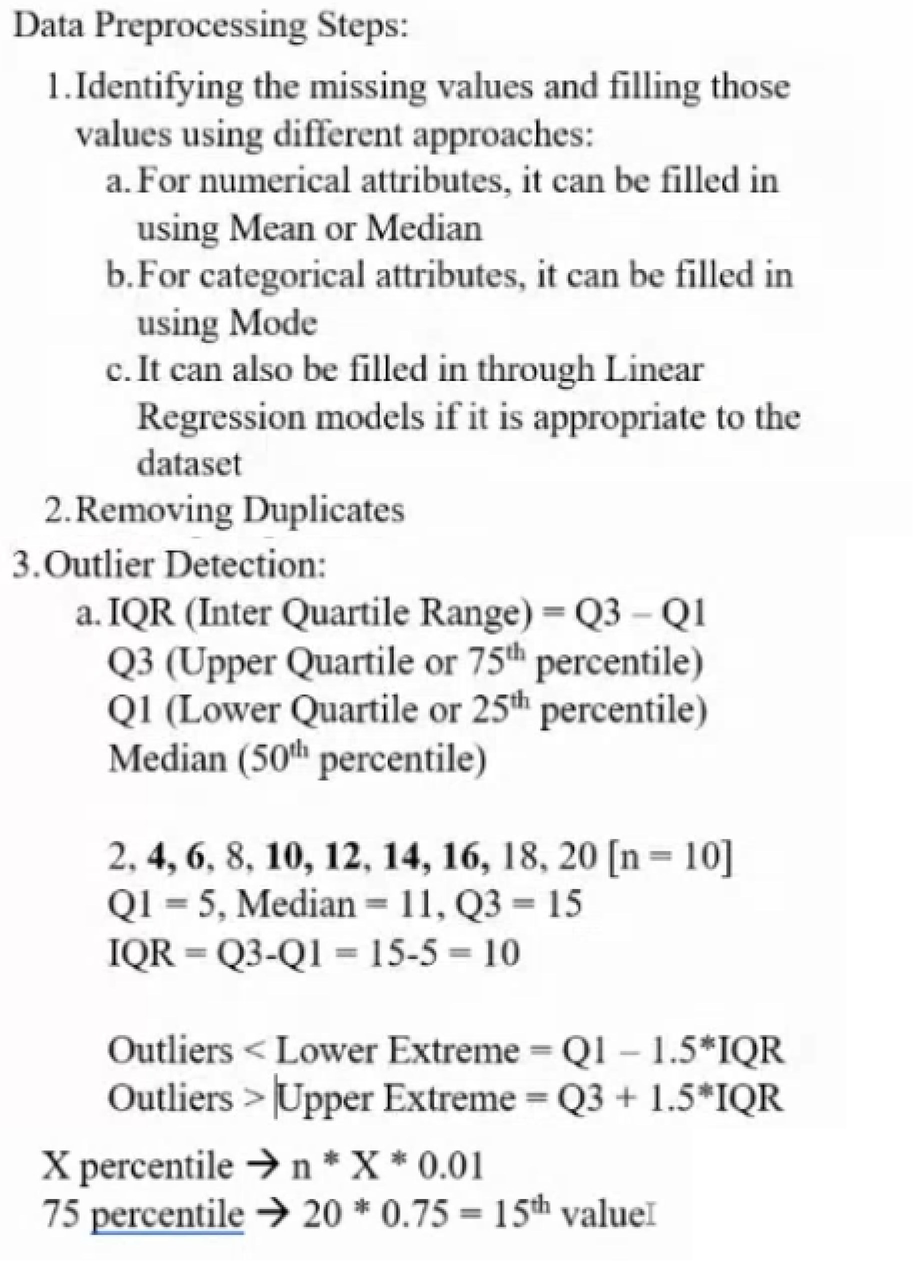
Outlier Detection and Handling

- **Why Outliers Matter:**
  Extreme values can bias statistical analysis and machine learning models.
- **Detection Methods:**
  - **Z-score method:** Identifies values that deviate significantly from the mean
  - **Interquartile Range (IQR):** Identifies values beyond 1.5 times the IQR
- **Handling Outliers:**
  - Remove extreme values
  - Cap values within a threshold
  - Apply transformations (e.g., log transformation)

Feature Scaling and Transformation

- Ensures uniformity in numerical features and improves model convergence.
- **Types of Scaling:**
    - **Standardization:**
      [(X - \text{mean}) / \text{std_dev}]
      Ensures zero mean and unit variance
    - **Normalization:**
      [(X - \text{min}) / (\text{max} - \text{min})]
      Scales values between 0 and 1
    - **Log Transformation:**
      Used for skewed data to reduce variance

Data Preprocessing Steps



Data Preprocessing Steps:

1. Identifying the missing values and filling those values using different approaches:
    a. For numerical attributes, it can be filled in using Mean or Median
    b. For categorical attributes, it can be filled in using Mode
    c. It can also be filled in through Linear Regression models if it is appropriate to the dataset
2. Removing Duplicates

3. Outlier Detection:
    a. IQR (Inter Quartile Range) = Q3 – Q1
    Q3 (Upper Quartile or 75th percentile)
    Q1 (Lower Quartile or 25th percentile)
    Median (50th percentile)

    2, **4, 6**, 8, **10, 12, 14, 16,** 18, 20 [n = 10]
    Q1 = 5, Median = 11, Q3 = 15
    IQR = Q3-Q1 = 15-5 = 10

    Outliers < Lower Extreme = Q1 – 1.5*IQR
    Outliers > Upper Extreme = Q3 + 1.5*IQR

X percentile → n * X * 0.01
75 percentile → 20 * 0.75 = 15th value

Worked Example

The ages, in years, of a number of children attending a birthday party are given below:

2, 7, 5, 4, 8, 4, 6, 5, 5, 29, 2, 5, 13,

An outlier is defined as an observation that falls more than $1.5\ \times$ the interquartile range above the upper quartile or below the lower quartile

(i) Identify any outliers within the data set.

(ii) Decide which values (if any) should be removed, justify your answer.

(i) Start by putting the data in order of size

2, 2, 4, 4, 5, 5, 5, 5, 6, 7, 8, 13, 29

LQ Median UQ

Find the interquartile range:

$Q_1 = 4, \quad Q_3 = 7.5$

$IQR = Q_3 - Q_1 = 7.5 - 4 = 3.5$

Find the upper and lower bound for the outliers:

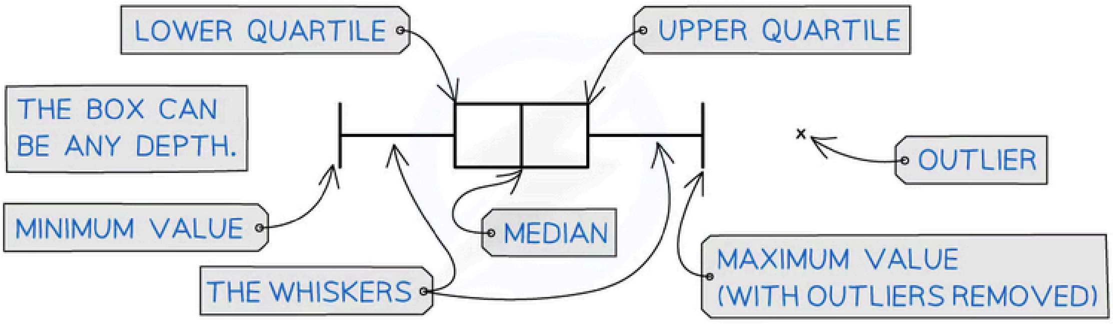
Lower bound $= Q_1 - 1.5\,(IQR)$

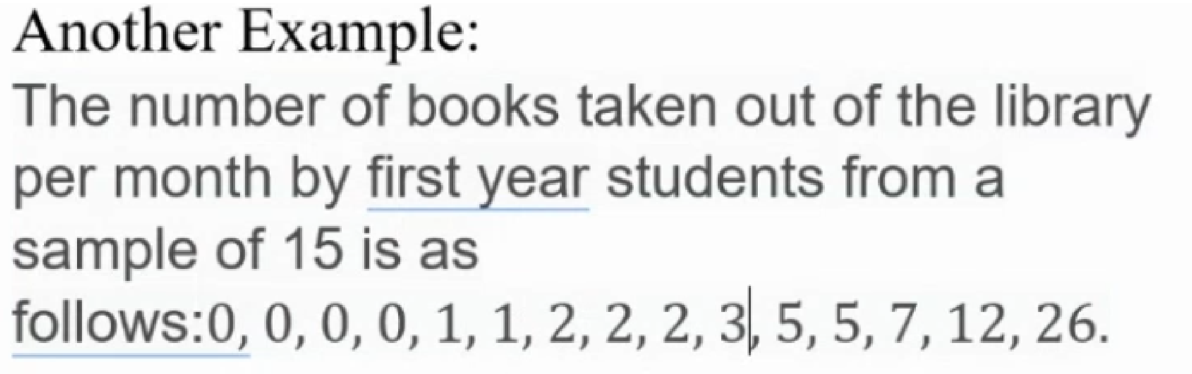
$4 - 1.5\,(3.5) = -1.25$ So there are no outliers on the lower tail.

Upper bound $= Q_3 + 1.5\,(IQR)$

$7.5 + 1.5\,(3.5) = 12.75$ So there are two outliers on the upper tail.

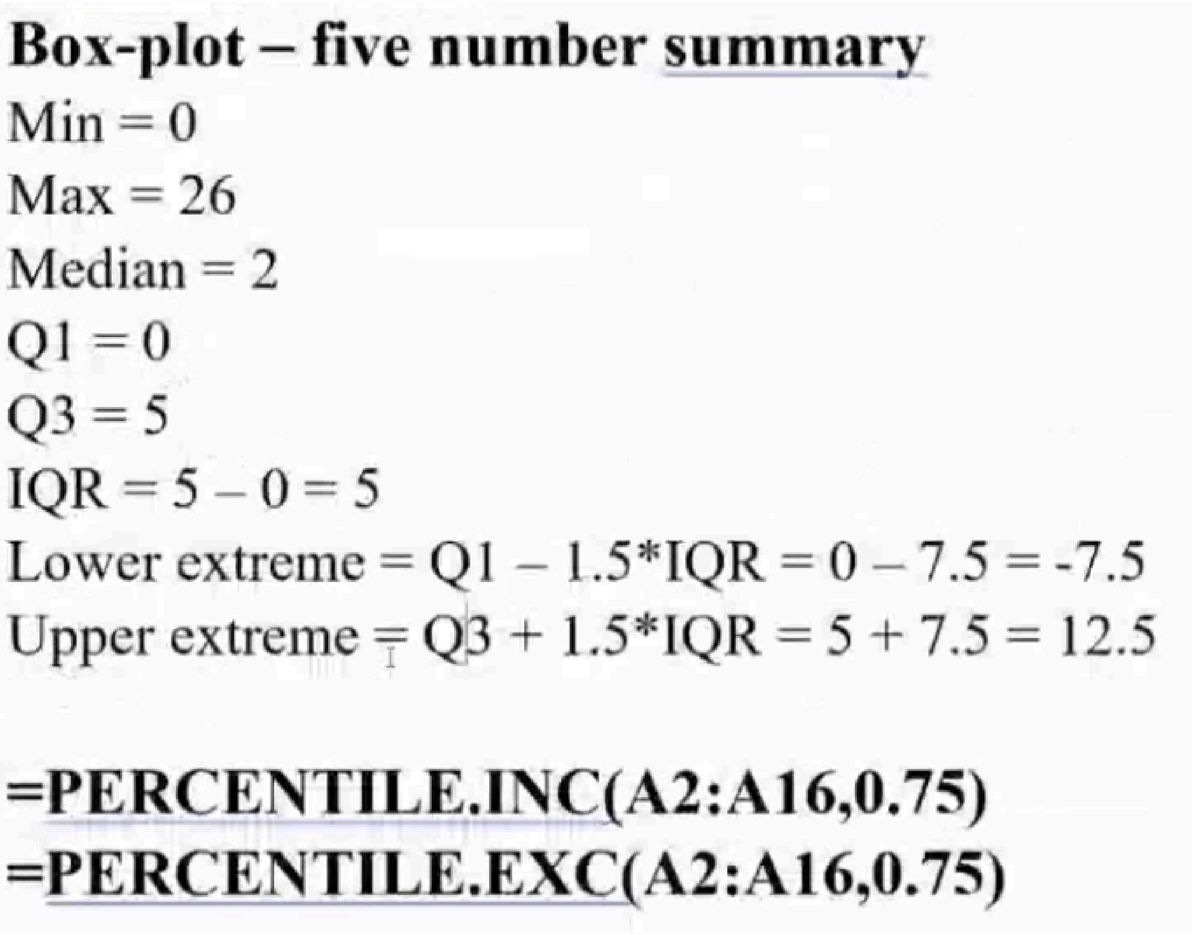
Outliers are 13 and 29





Another Example:

The number of books taken out of the library per month by first year students from a sample of 15 is as follows:0, 0, 0, 0, 1, 1, 2, 2, 2, 3, 5, 5, 7, 12, 26.



Box-plot – five number summary

Min = 0

Max = 26

Median = 2

Q1 = 0

Q3 = 5

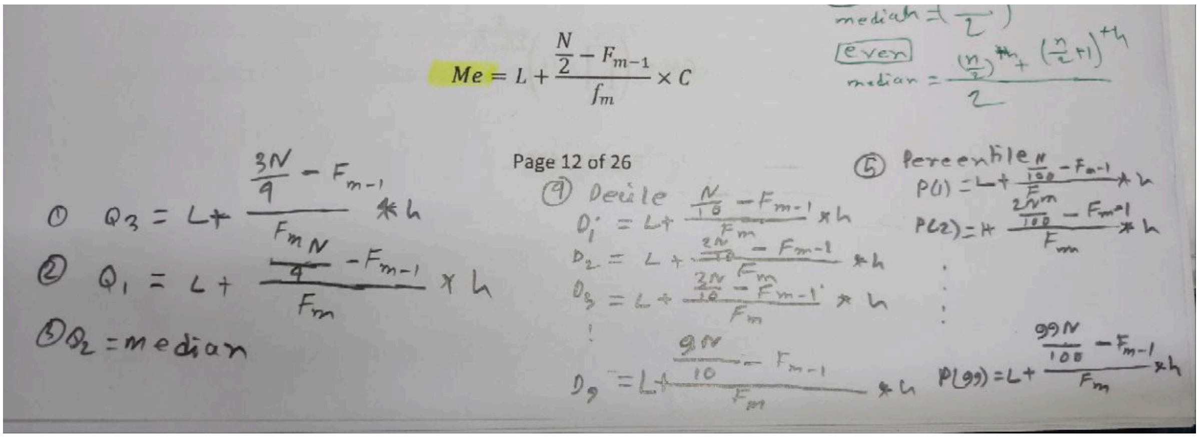
IQR = 5 – 0 = 5

Lower extreme = Q1 – 1.5*IQR = 0 – 7.5 = -7.5

Upper extreme = Q3 + 1.5*IQR = 5 + 7.5 = 12.5

=PERCENTILE.INC(A2:A16,0.75)

=PERCENTILE.EXC(A2:A16,0.75)



$$Me = L + \frac{\frac{N}{2} - F_{m-1}}{f_m} \times C$$

median $= \left(\frac{\quad}{2}\right)$

even

median $= \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

Page 12 of 26

① $Q_3 = L + \frac{\frac{3N}{4} - F_{m-1}}{F_m} * h$

② $Q_1 = L + \frac{\frac{N}{4} - F_{m-1}}{F_m} \times h$

③ Q_2 = median

④ Decile

$D_1 = L + \frac{\frac{N}{10} - F_{m-1}}{F_m} \times h$

$D_2 = L + \frac{\frac{2N}{10} - F_{m-1}}{F_m} * h$

$D_3 = L + \frac{\frac{3N}{10} - F_{m-1}}{F_m} * h$

⋮

$D_9 = L + \frac{\frac{9N}{10} - F_{m-1}}{F_m} * h$

⑤ Percentile

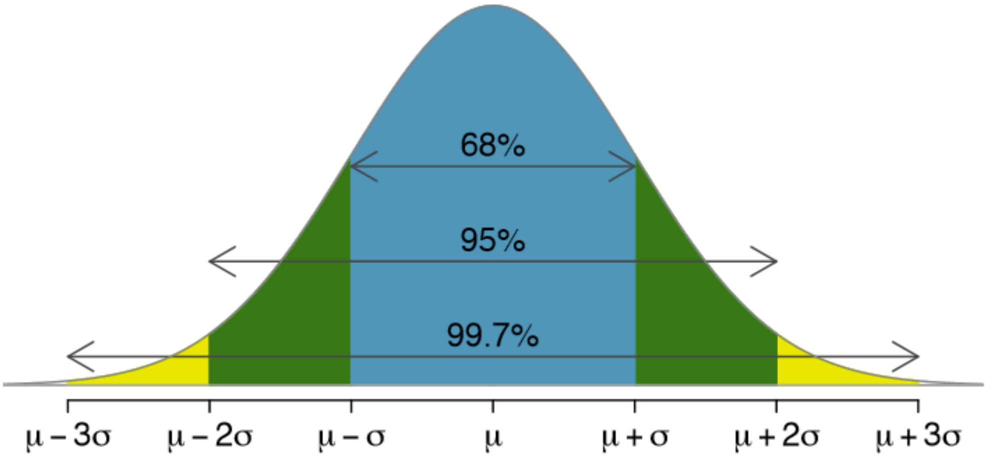
$P(1) = L + \frac{\frac{N}{100} - F_{m-1}}{F} * h$

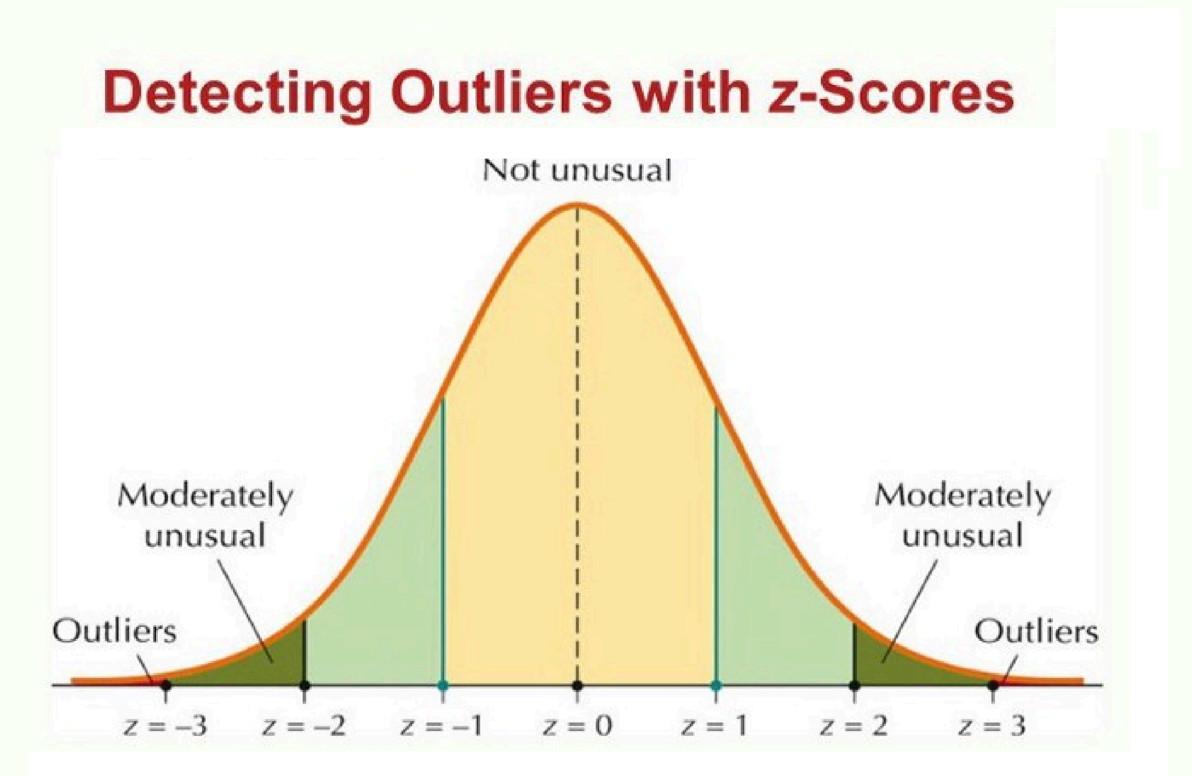
$P(2) = L + \frac{\frac{2N}{100} - F_{m-1}}{F_m} * h$

⋮

$P(99) = L + \frac{\frac{99N}{100} - F_{m-1}}{F_m} \times h$

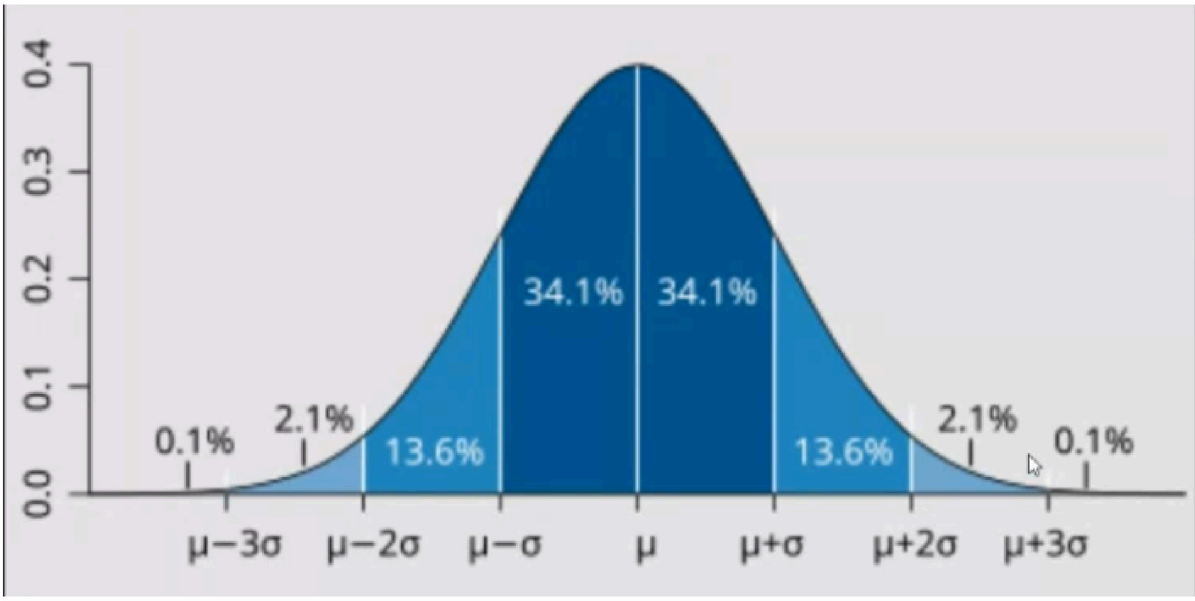
Outlier Detection

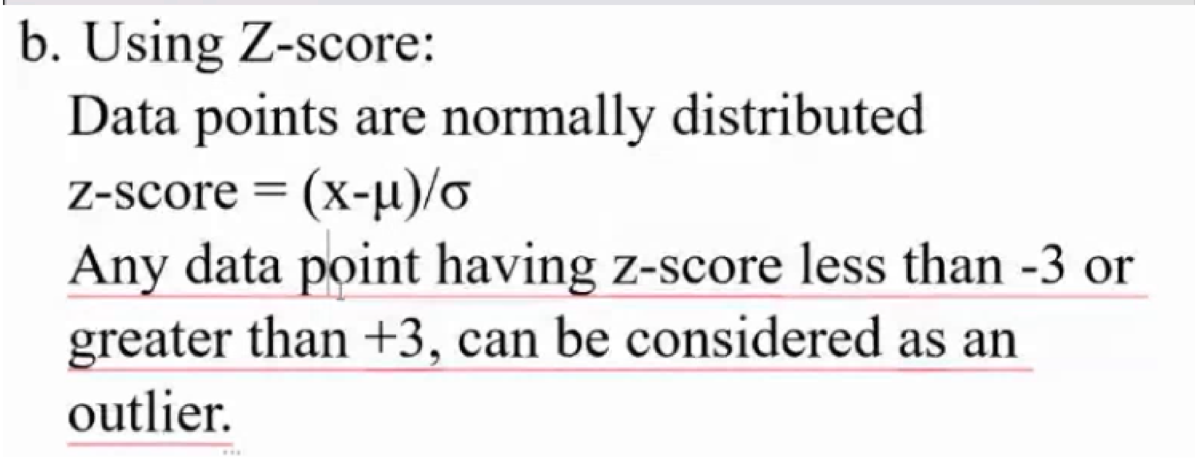




If the data follows a **normal distribution**:

- **68%** of the data falls within **±1 standard deviation (σ)** from the mean (μ)
  → That is, between **μ - σ** and **μ + σ**
- **95%** of the data falls within **±2 standard deviations**
  → Between **μ - 2σ** and **μ + 2σ**
- **99.7%** of the data falls within **±3 standard deviations**
  → Between **μ - 3σ** and **μ + 3σ**




b. Using Z-score:

Data points are normally distributed

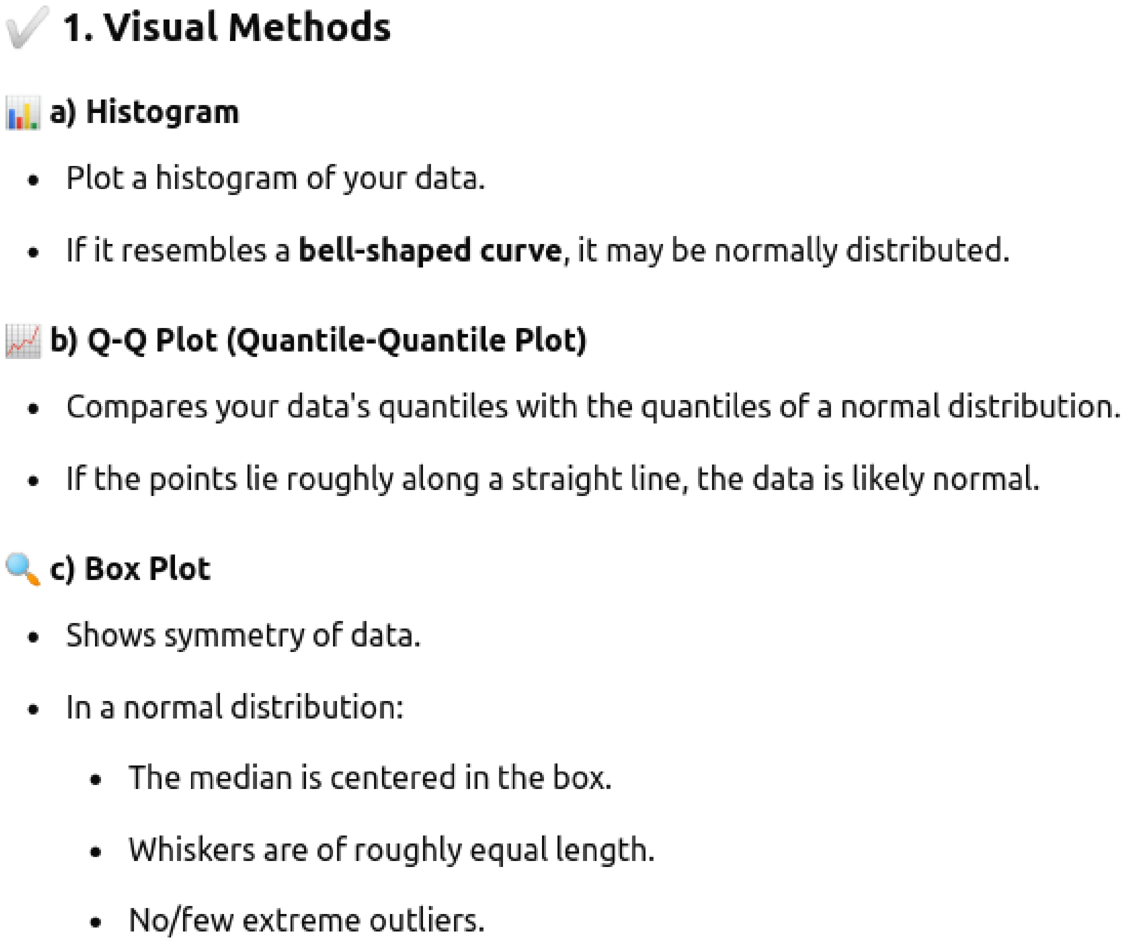
z-score = (x-μ)/σ

Any data point having z-score less than -3 or greater than +3, can be considered as an outlier.

The formula for Z-score is as follows:

$$Zscore = (x - mean)/std.deviation$$

how to find normally distributed



✔️ 1. Visual Methods

📊 a) Histogram

- Plot a histogram of your data.
- If it resembles a **bell-shaped curve**, it may be normally distributed.

📈 b) Q-Q Plot (Quantile-Quantile Plot)

- Compares your data's quantiles with the quantiles of a normal distribution.
- If the points lie roughly along a straight line, the data is likely normal.

🔍 c) Box Plot

- Shows symmetry of data.
- In a normal distribution:
  - The median is centered in the box.
  - Whiskers are of roughly equal length.
  - No/few extreme outliers.

```
z = (X – μ) / σ

Here’s what each component means:

    Z is the Z-score we’re calculating.
    X is the specific data point we want to evaluate.
    μ (mu) is the mean (average) of the dataset.
    σ (sigma) is the standard deviation, which measures how spread out the data
is.
```

Documents

- https://www.savemyexams.com/international-a-level/maths/edexcel/20/statistics-1/revision-notes/data-presentation-and-interpretation/working-with-data/outliers/
-