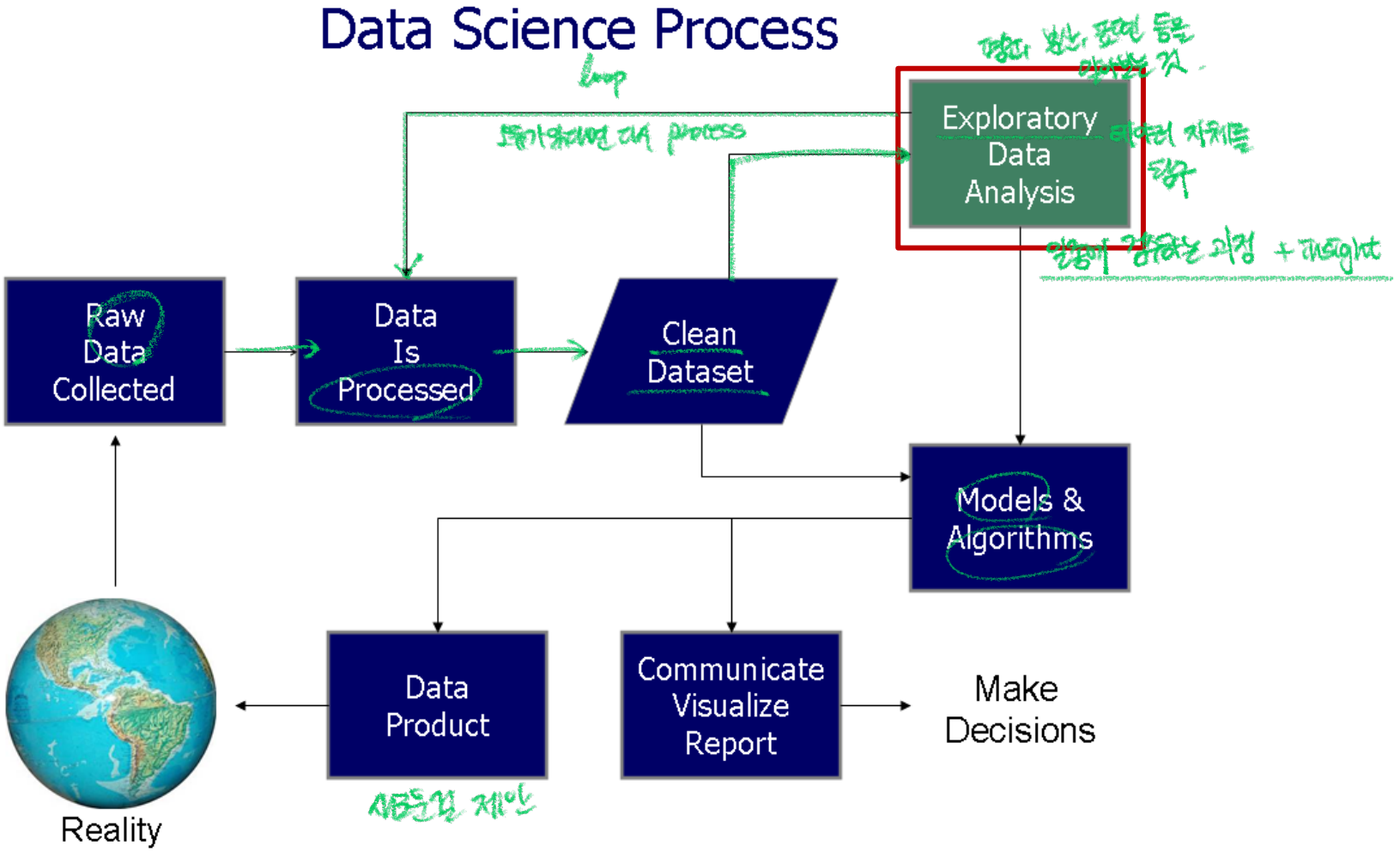
EXPLANATORY DATA ANALYSIS

Week03

Data Science Process



Explanatory Data Aanalysis

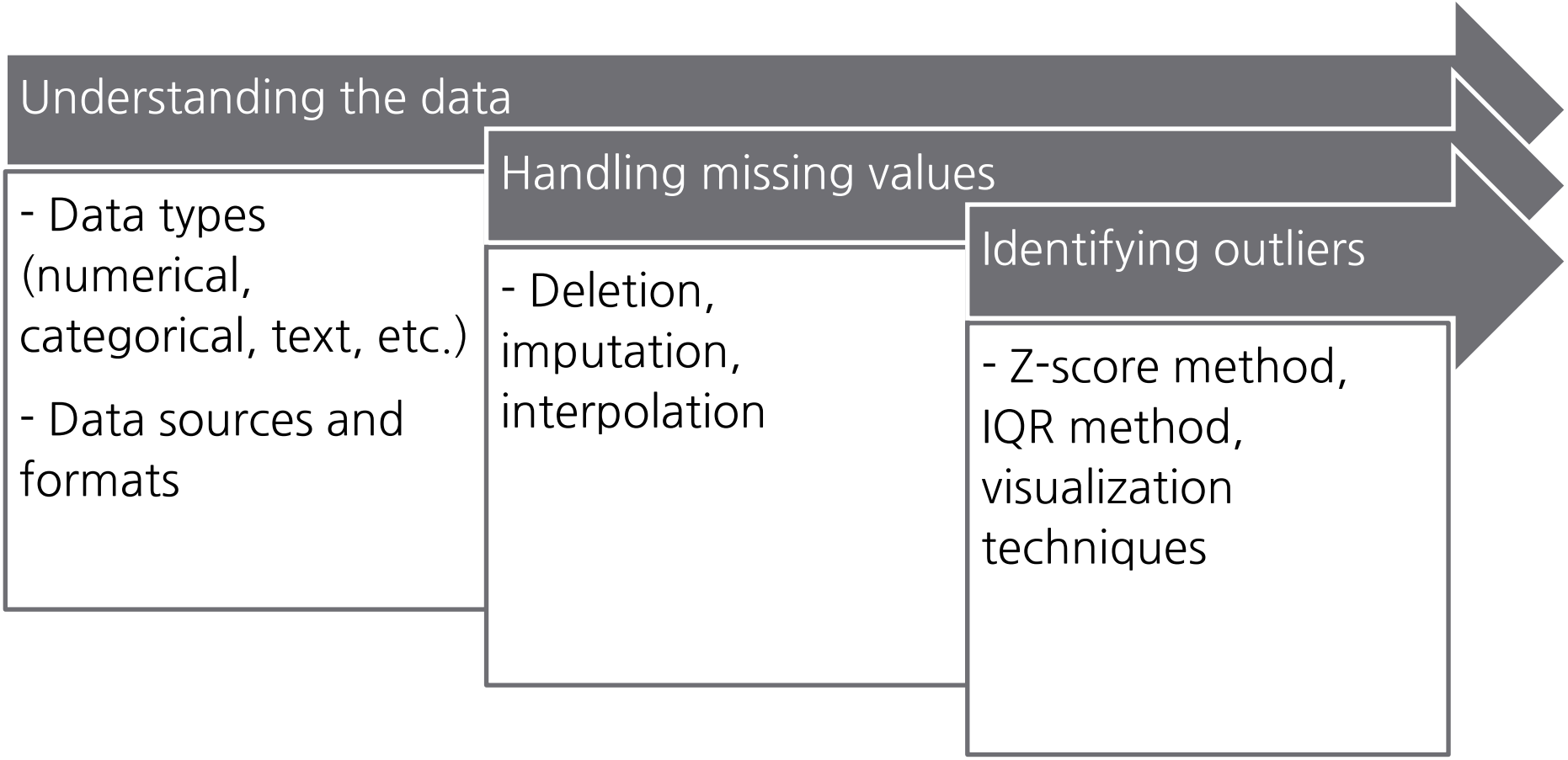
What is Exploratory Data Analysis (EDA)?

- Definition
  - EDA is an approach to analyzing data sets to summarize their main characteristics, often with visual methods
- Purpose
  - Identify patterns, detect anomalies, test hypotheses, and check assumptions with summary statistics and graphical representations
- Key Techniques
  - Data visualization, summary statistics, handling missing values, detecting outliers

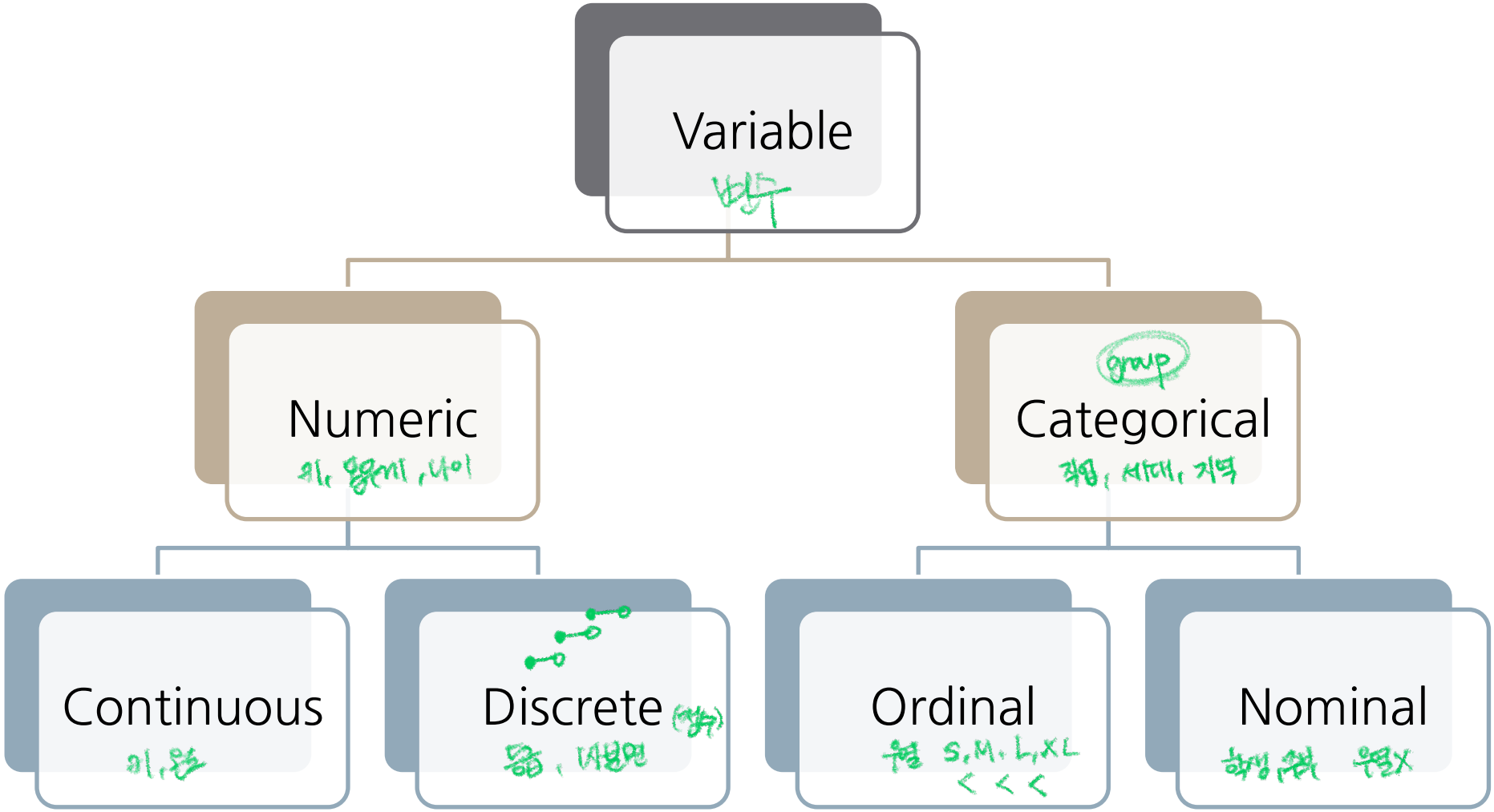
Importance of EDA

- Understanding the Data
  - Identify data distributions and relationships
  - Detect missing values and anomalies
- Data Preprocessing
  - Helps in feature selection and engineering
  - Assists in choosing appropriate machine learning models
- Insights & Decision Making
  - Provides actionable insights before model building
  - Supports better decision-making in data-driven projects.

Steps in EDA


Understanding the data
- Data types (numerical, categorical, text, etc.)
- Data sources and formats
Handling missing values
- Deletion, imputation, interpolation
Identifying outliers
- Z-score method, IQR method, visualization techniques

Understanding Data: Types of Variables

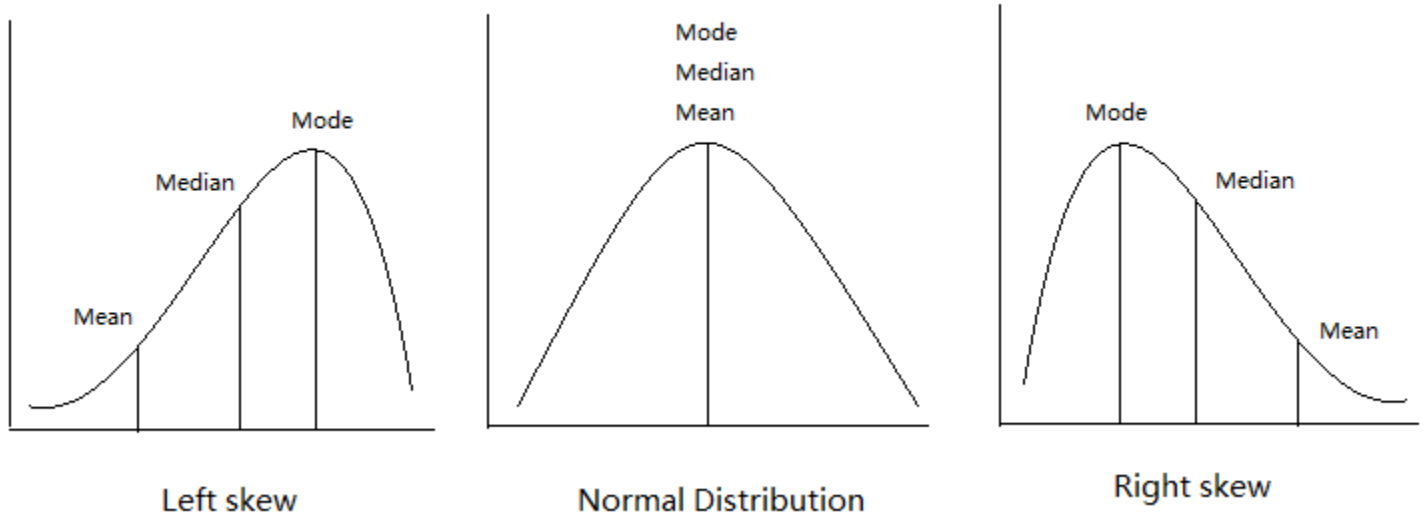


Understanding Data: Summary Statistics

- Measures of central tendency
  - Mean: The average value of a dataset

$$\mu = \frac{\sum_i x_i}{n}$$

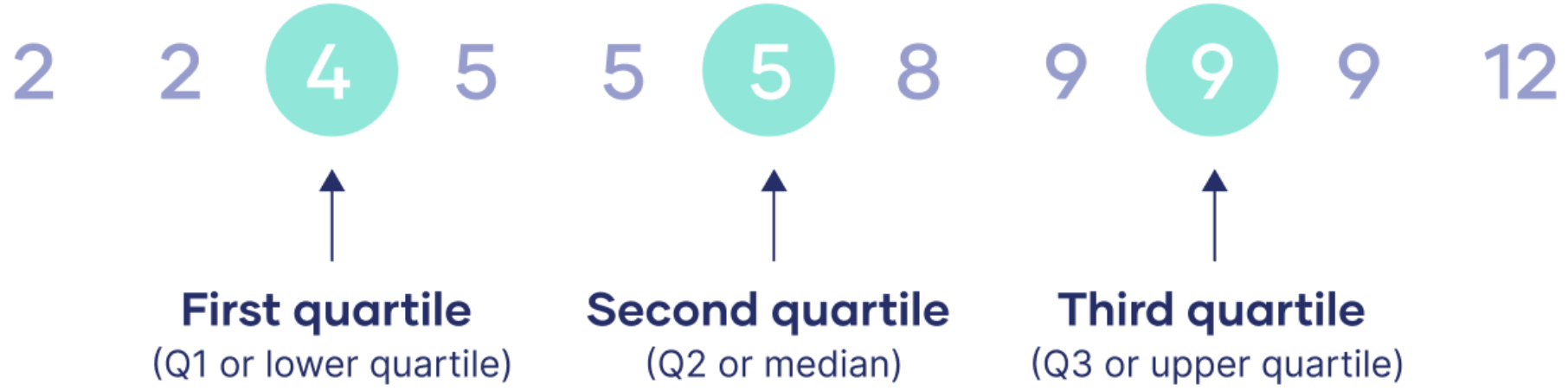
  - Median: The middle value in a sorted dataset. If the number of observations (n) is odd, the median is the middle value. If n is even, the median is the average of the two middle values
  - Mode: The most frequently occurring value in the dataset

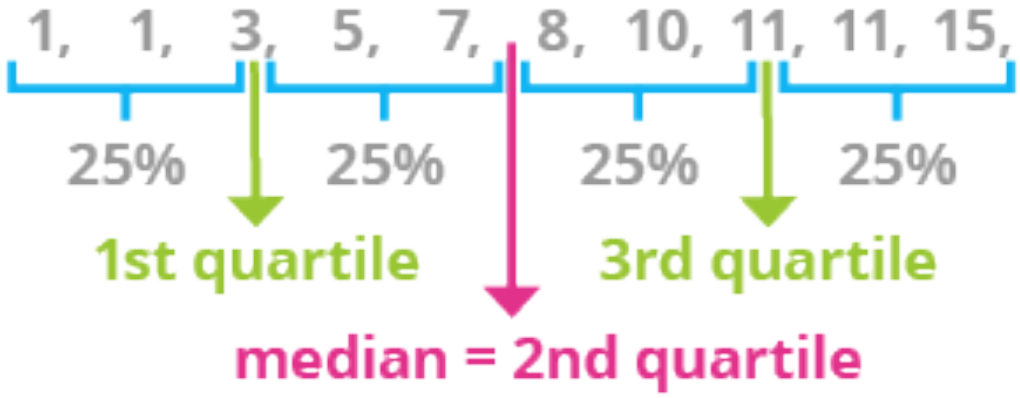


Understanding Data: Summary Statistics

- Measures of spread
  - Variance: Measures the average squared deviation from the mean
    - Population
    $$\sigma^2 = \frac{\sum_i (x_i - \mu)^2}{n}$$
    - Sample
    $$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n - 1}$$
  - Standard Deviation: The square root of the variance, indicating dispersion
    - Population
    $$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{n}}$$
    - Sample
    $$s^2 = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n - 1}}$$
  - Range: The difference between the maximum and minimum values in the dataset
  $$Range = \max(x) - \min(x)$$
  - Inter-quartile range (IQR): Measures the spread of the middle 50% of data
  $$IQR = Q_3 - Q_1$$

Understanding Data: Summary Statistics

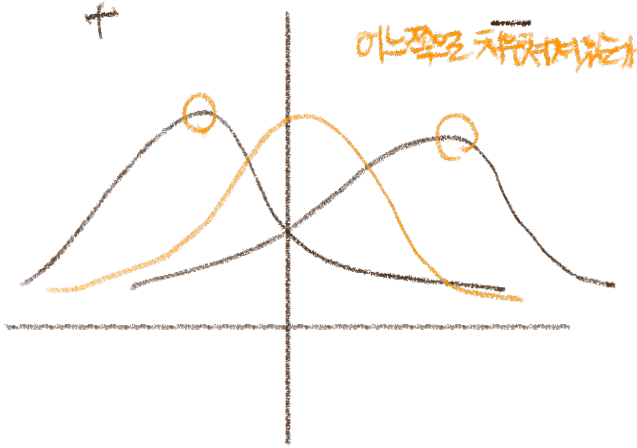




Understanding Data: Summary Statistics

- Skewness
  - Skewness measures the symmetry of a distribution
    - Population

$$\gamma_1 = \mathbb{E}\left[\left(\frac{X - \mu}{\sigma}\right)^3\right]$$

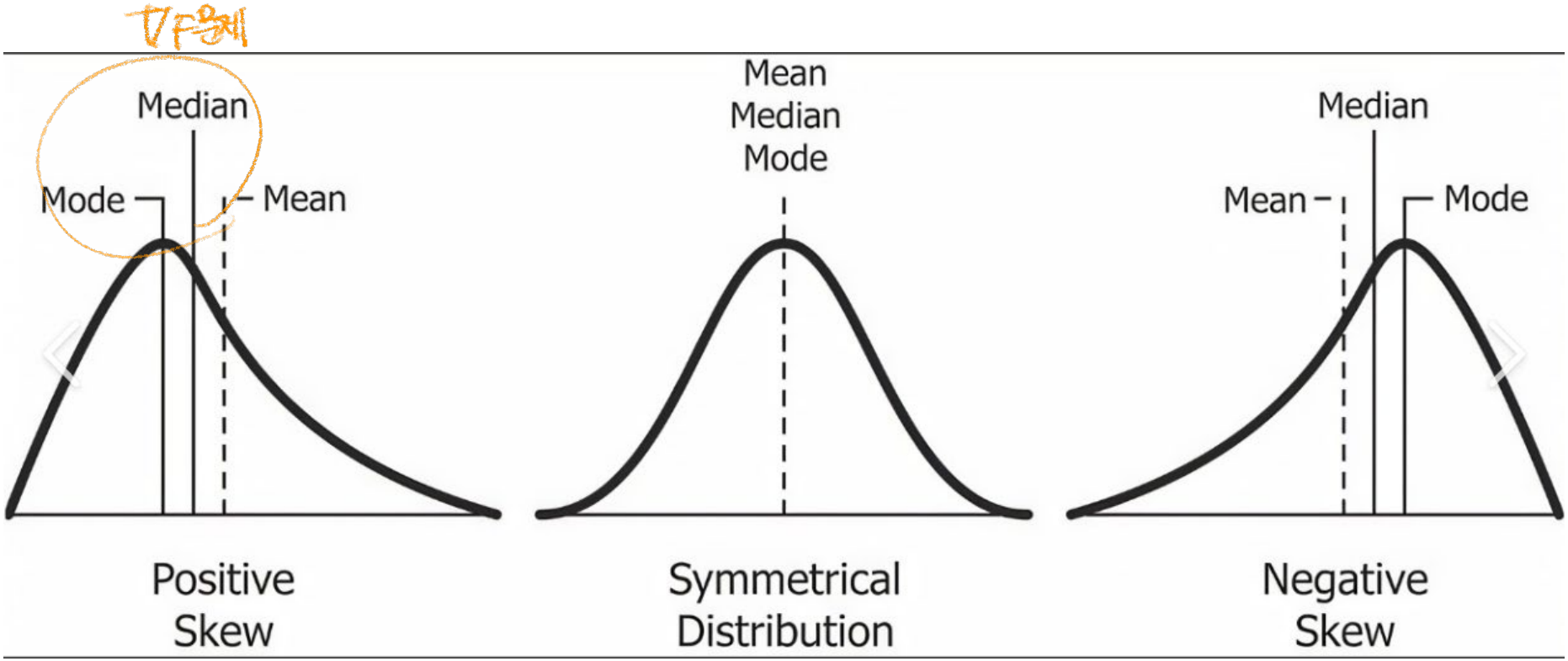


    - Sample

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

    - Positive skewness: The tail extends to the right (more values are on the left side of the mean).
    - Negative skewness: The tail extends to the left (more values are on the right side of the mean)
    - A skewness value of zero indicates a symmetric distribution

Understanding Data: Summary Statistics



Understanding Data: Summary Statistics

- Kurtosis
  - Kurtosis measures the “tailedness” or “peakedness” of a distribution compared to a normal distribution
    - Population

$$Kurt[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

    - Sample

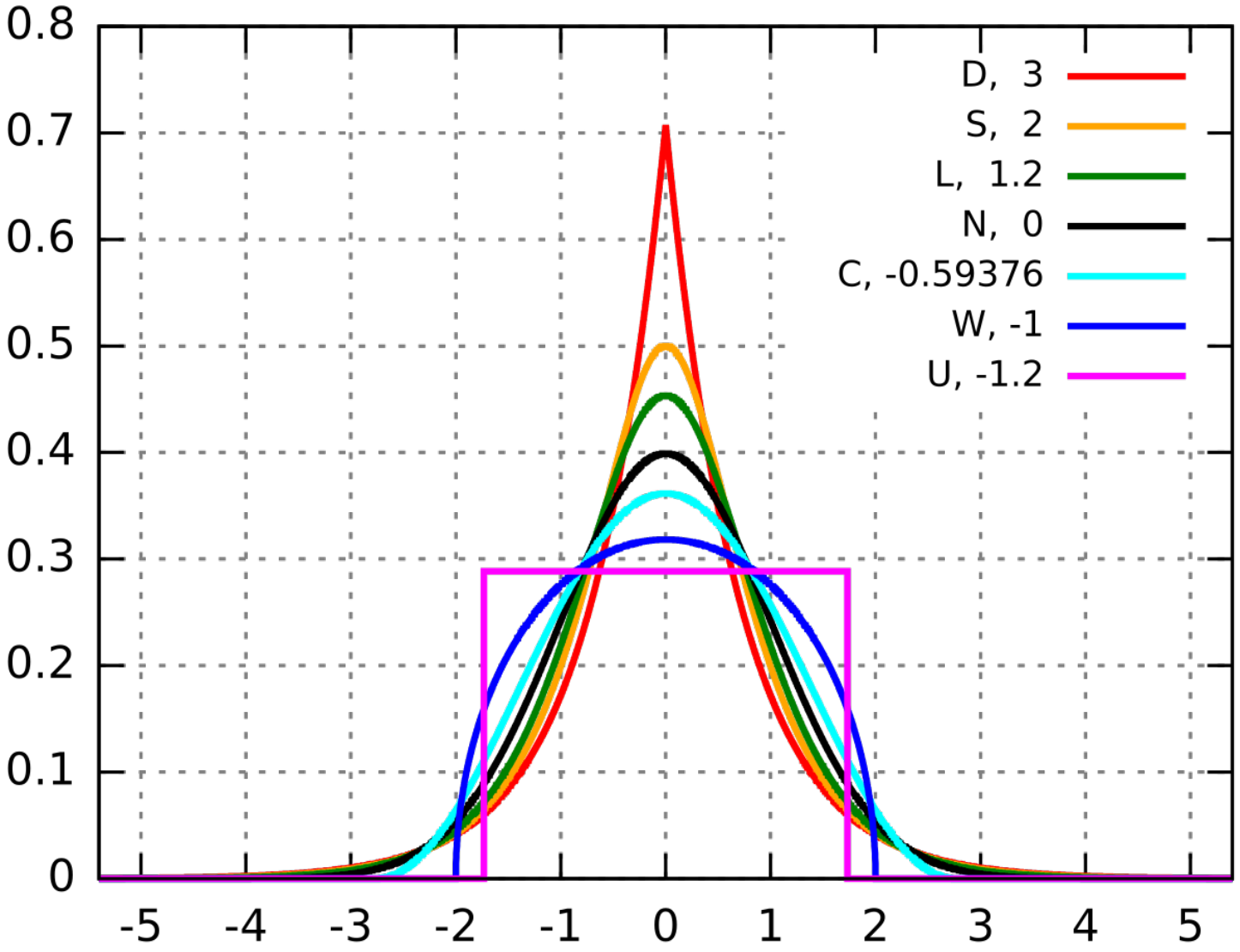
$$\kappa = \frac{m_4}{s^4}$$

    - Excess kurtosis (Fisher kurtosis)

$$Kurt[X] - 3$$

    - Positive kurtosis (leptokurtic): The distribution has heavier tails and a sharper peak than a normal distribution
    - Negative kurtosis (platykurtic): The distribution has lighter tails and a flatter peak than a normal distribution
    - A kurtosis value of 3 (or 0 for excess kurtosis) indicates a distribution similar to a normal distribution (mesokurtic)

Understanding Data: Summary Statistics



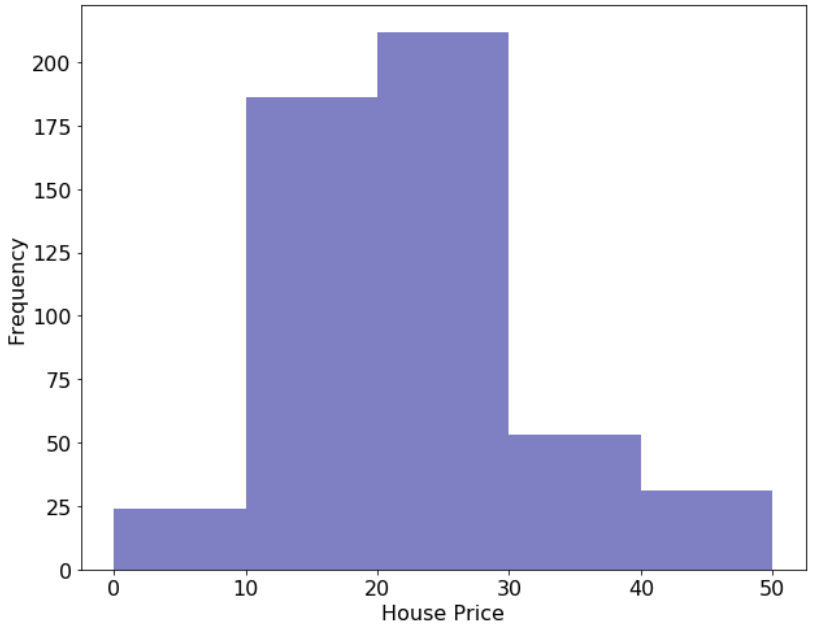
Understanding Data: Summary Statistics

- Data distribution
  - Histograms: A graphical representation of the distribution of numerical data. It divides the dataset into bins and counts the number of observations within each bin, showing how data values are distributed across different ranges

24.0,21.6,34.7,33.4,36.2,28.7,22.9,
27.1,16.5,18.9,15.0,18.9,21.7,20.4,
18.2,19.9,23.1,17.5,20.2,18.2,13.6,
19.6,15.2,14.5,15.6,13.9,16.6,14.8
… …

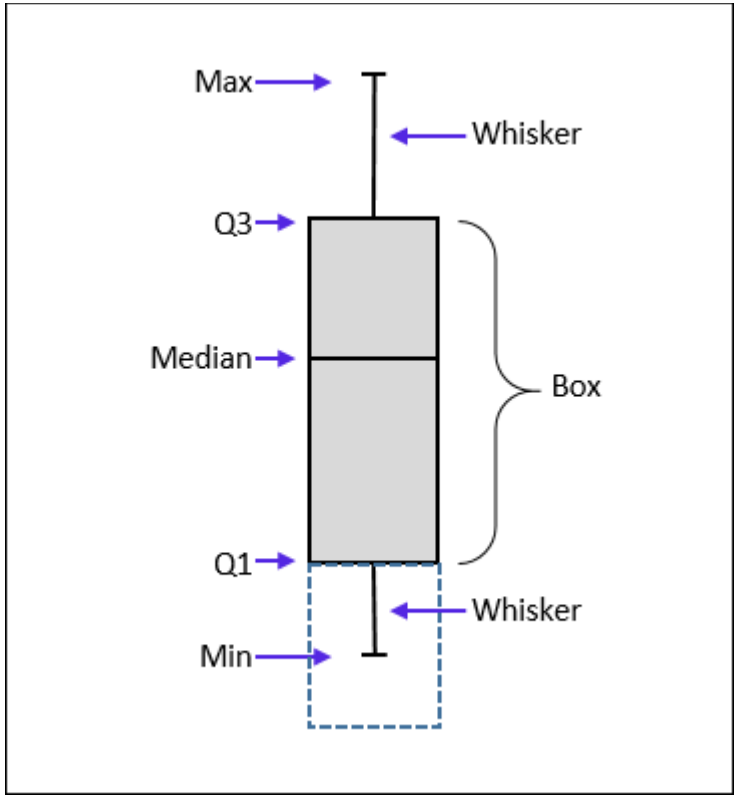
Bin	Count
0 to 10	24
10 to 20	186
20 to 30	212
30 to 40	53
40 to 50	31

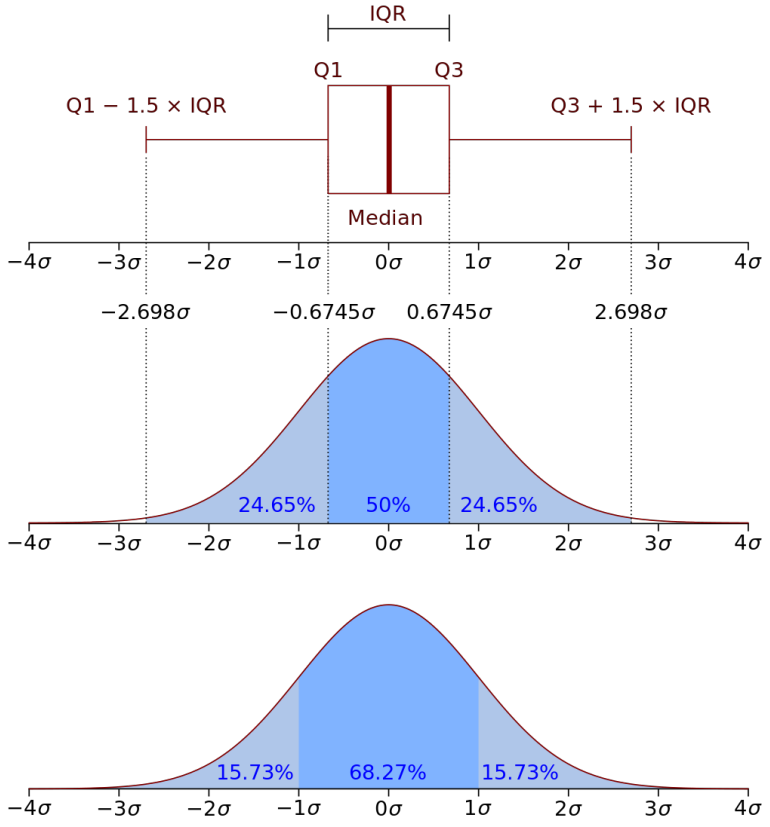
Histogram



Understanding Data: Summary Statistics

- Data distribution
  - Box plot: A visualization that summarizes the distribution of a dataset using five key summary statistics: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum. It helps identify outliers and the spread of the data





Understanding Data: Summary Statistics

- Correlations between variables
  - Pearson correlation coefficient: Measures the linear relationship between two continuous variables

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

  - Spearman rank correlation: Measures the strength and direction of the monotonic relationship between two variables

$$\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)}$$

    - d_i: the difference between the ranks of corresponding values
  - Kendall's tau: Measures the ordinal association between two variables

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

    - C: the number of concordant pairs
    - D: the number of discordant pairs

Understanding Data: Summary Statistics

- Spearman rank correlation

$$\rho = 1 - \frac{6 \sum_i d_i^2}{n(n^2 - 1)}$$

X	Y
10	40
20	30
30	50
40	20
50	60

Calculate ranks and differences in ranks

X	$Rank(X)$	Y	$Rank(Y)$	d_i	d_i^2
10	1	40	3	-2	4
20	2	30	2	0	0
30	3	50	4	-1	1
40	4	20	1	3	9
50	5	60	5	0	0

Understanding Data: Summary Statistics

- Kendall's tau

$$\tau = \frac{C - D}{\frac{1}{2}n(n-1)}$$

X	Y
10	40
20	30
30	50
40	20
50	60

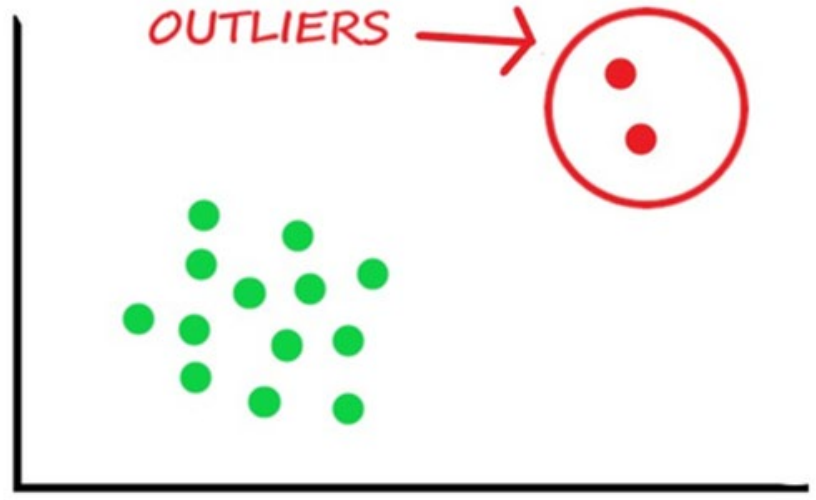
Pairs	Relationship	Pairs	Relationship
(10,40) & (20,30)	Discordant	(20,30) & (40,20)	Discordant
(10,40) & (30,50)	Concordant	(20,30) & (50,60)	Concordant
(10,40) & (40,20)	Discordant	(30,50) & (40,20)	Discordant
(10,40) & (50,60)	Concordant	(30,50) & (50,60)	Concordant
(20,30) & (30,50)	Concordant	(40,20) & (50,60)	Concordant

Handling Missing Values

- Methods to handle missing values
  - Deletion: Remove missing data points if they are few and do not significantly impact the dataset
  - Mean/Median/Mode Imputation: Replace missing values with statistical measures such as mean, median or mode
  - k-Nearest Neighbors (kNN) imputation: Use similar data points to estimate missing values
  - Regression imputation: Predict missing values based on relationships between variables
  - Multiple Imputation: Generate multiple datasets with estimated values and combine results

Identifying outliers

- Outliers
  - Outliers are data points that are significantly different from the rest of the data
  - They can be unusually high or low compared to other observations in a dataset



Identifying outliers

- Characteristics of Outliers
  - Extreme values: Outliers fall far outside the normal range of data
  - Unexpected behavior: They may arise from errors, but they can also be valid data points
  - Impact on analysis: They can distort statistical measures such as mean, standard deviation, and regression models

- Types of outliers
  - Point outliers: A single data point that deviates significantly
    - Example: A person with an extremely high income compared to a group
  - Contextual outliers: Data points that are considered outliers only in certain contexts or subgroups
    - Example: A summer day temperature of 30°C in a region with a typical summer range of 20-25°C.
  - Collective outliers: A collection of data points that, together, show unusual behavior.
    - Example: A set of consecutive days with temperatures unusually high for a region.

Identifying outliers

- Why are outliers important?
  - Affect mean and standard deviation
    - Outliers can skew the mean and inflate the standard deviation, leading to misleading conclusions
  - Modeling issues
    - Outliers can affect the performance of machine learning models (e.g., linear regression) by making predictions less accurate
  - Data Integrity
    - Outliers may indicate data entry errors, measurement mistakes, or special cases that need further investigation

Identifying outliers

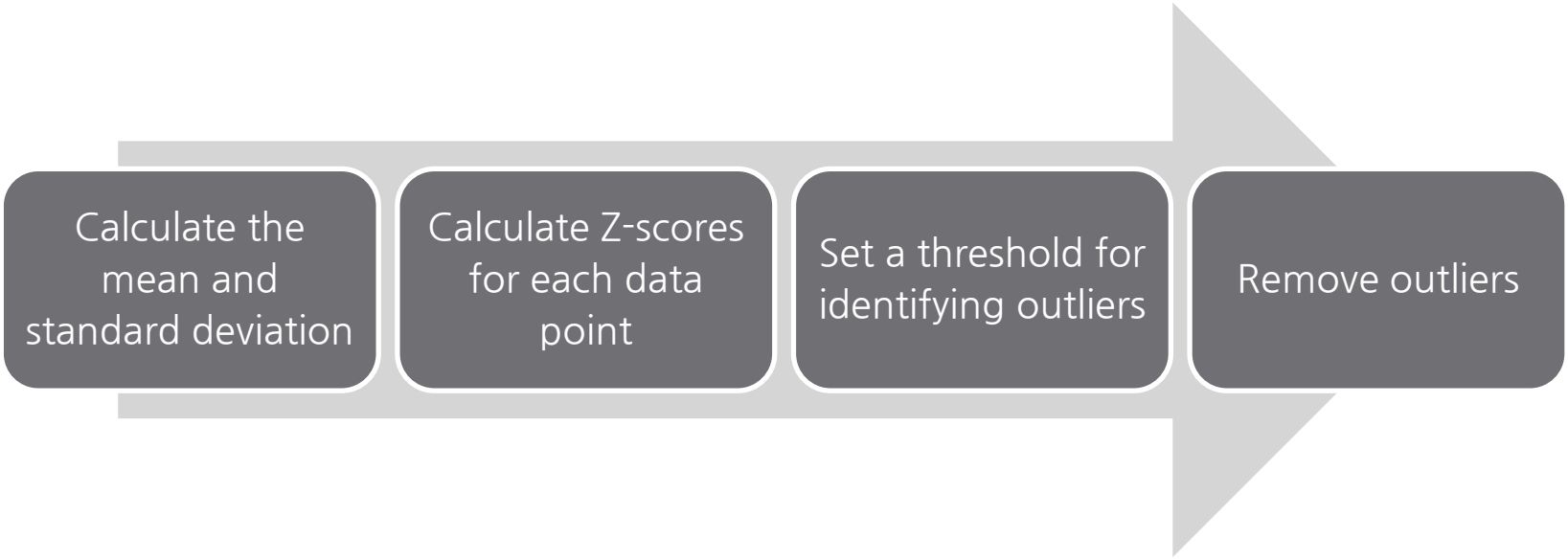
- Importance of removing outliers
  - Improved accuracy
    - Removing outliers can lead to more accurate models, ensuring that predictions are not disproportionately influenced by extreme values
  - Better data representation
    - The central tendency of the data (mean, median) becomes more representative of the general population
  - Enhanced model performance
    - Machine learning algorithms can perform better when outliers are removed, reducing noise and improving the fit of the model

Identifying outliers

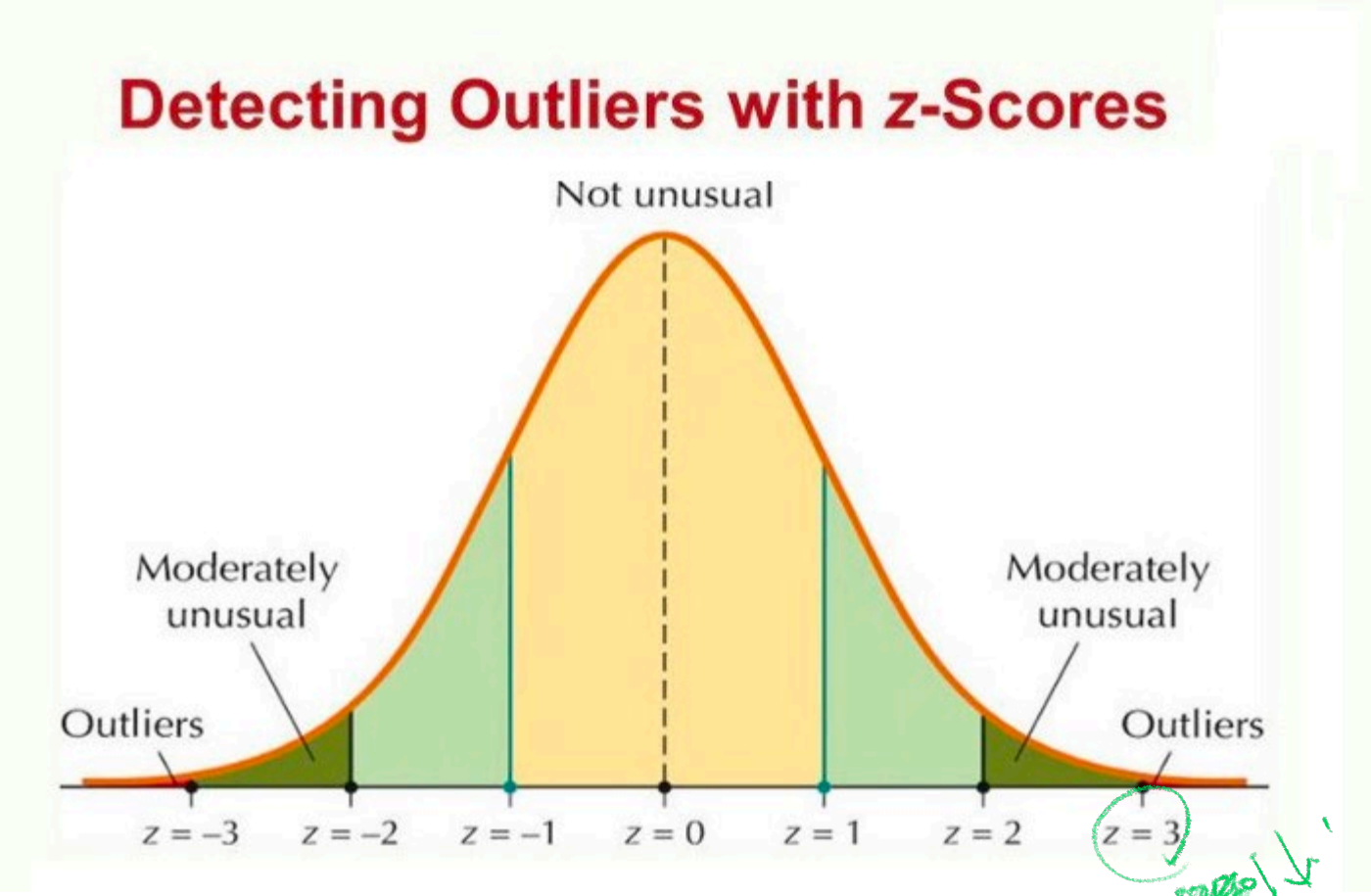
- Statistical method using Z-scores to identify outliers
  - Z-score measures how many standard deviations a data point is away from the mean. It is a standardized way of identifying outliers in normally distributed data

$$Z = \frac{X - \mu}{\sigma}$$

  - Typically, if a Z-score is greater than 3 or less than -3, the data point is considered an outlier (assuming a normal distribution)
    - This corresponds to values that are more than 3 standard deviations away from the mean



Identifying outliers



Identifying outliers

- Statistical method using IQR to identify outliers
  - Calculate the lower and upper fences for outliers using IQR
    - Lower inner fence

$$Lower\ inner\ fence = Q_1 - 1.5IQR$$

    - Upper inner fence

$$Upper\ inner\ fence = Q_3 + 1.5IQR$$

    - Lower outer fence

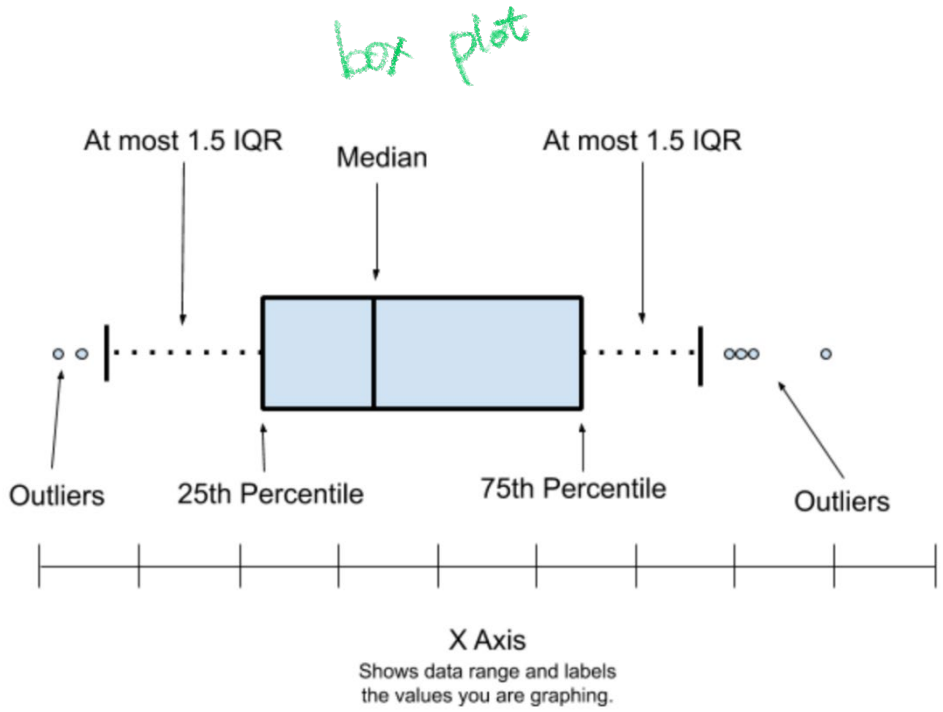
$$Lower\ outer\ fence = Q_1 - 3IQR$$

    - Upper inner fence

$$Upper\ outer\ fence = Q_3 + 3IQR$$

  - Points beyond the inner fences in either direction are mild outliers; points beyond the outer fences in either direction are extreme outliers

Identifying outliers



Visualization: Why is Visualization Important in EDA?

1. **Understanding Data Distributions**
   - Helps in identifying patterns, trends, and underlying structures in the data.
   - Enables detection of skewness, multimodal distributions, and irregularities.
2. **Identifying Outliers and Missing Data**
   - Visual tools like boxplots and scatter plots can highlight outliers.
   - Heatmaps can reveal missing values in a dataset.
3. **Discovering Relationships Between Variables**
   - Scatter plots and correlation heatmaps help identify strong or weak relationships.
   - Pair plots allow simultaneous visualization of multiple variable interactions.
4. **Enhancing Interpretability**
   - Complex numerical summaries can be difficult to interpret.
   - Visualizations simplify insights and make data-driven decisions easier.
5. **Guiding Feature Engineering**
   - Helps determine transformations like log-scaling, binning, or normalization.
   - Aids in selecting important features for modeling.

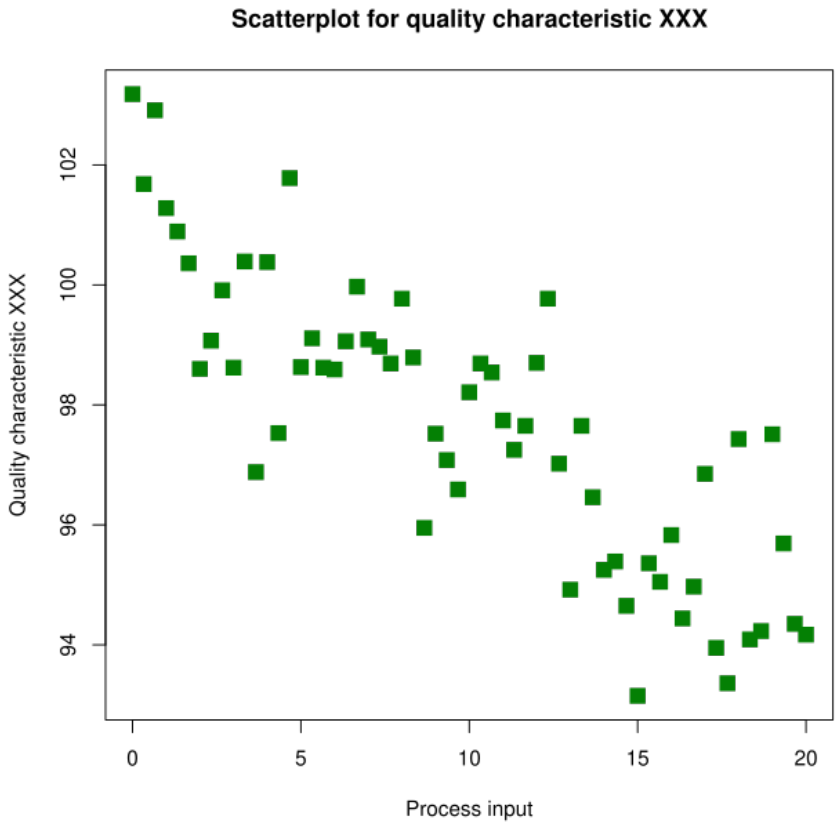
Visualization: Types of Plots

- Histograms
  - Use when: You want to analyze the distribution of a single continuous variable
  - Example: Examining the frequency of customer purchase amounts in an e-commerce dataset
  - Key insights: Shape of the distribution (normal, skewed, multimodal), outliers, and spread
- Boxplots
  - Use when: You need to compare distributions across categories or detect outliers
  - Example: Comparing salaries across different job roles in a company
  - Key insights: Median, quartiles, spread, and presence of outliers

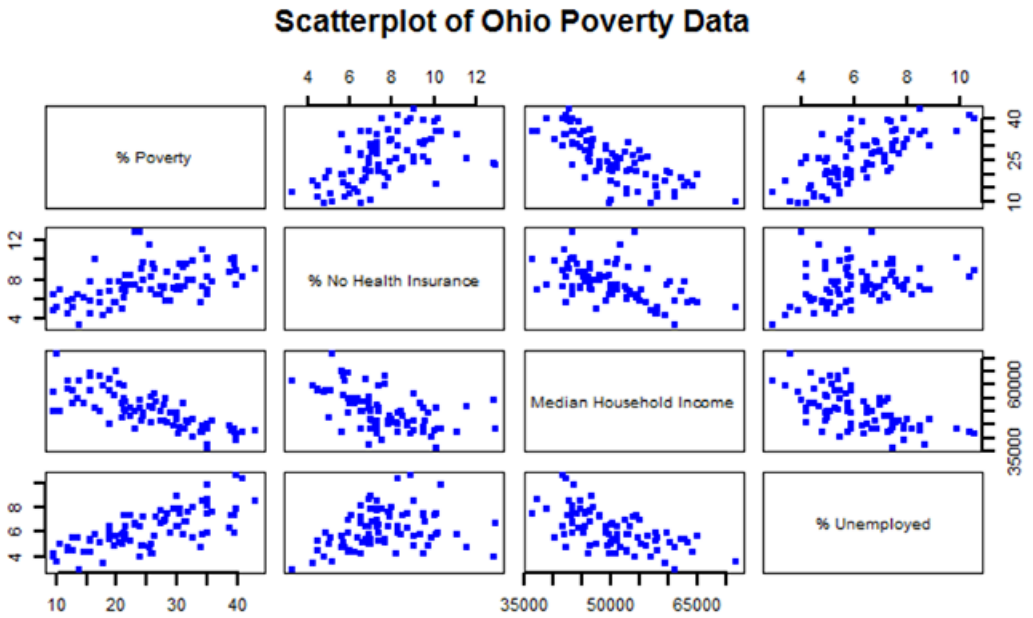
Visualization: Types of Plots

- Scatter plots
  - Use when: You want to examine the relationship between two numerical variables
  - Example: Checking if there is a correlation between advertising spend and sales revenue
  - Key insights: Strength and direction of relationships (linear, non-linear, or no correlation)
- Pair plots (Pairwise scatter plots)
  - Use when: You need to visualize relationships between multiple numerical variables in a dataset
  - Example: Analyzing relationships between weight, height, age, and cholesterol levels in a health study
  - Key insights: Trends, clusters, and correlations across multiple features

Visualization: Types of Plots



Scatter plot



Pair plot

Feature Engineering and Transformation

- Feature Engineering
  - The process of creating new features from raw data to improve model performance
  - Involves domain knowledge to extract meaningful information from raw data
  - Examples: Extracting "day of the week" from a timestamp, creating interaction terms, or encoding categorical variables.
- Feature Transformation
  - Techniques used to modify features to make them more suitable for analysis and modeling
  - Helps normalize, standardize, or reduce skewness in the data
  - Examples: Log transformation to reduce skewness, min-max scaling, and principal component analysis (PCA) for dimensionality reduction.

Feature Engineering and Transformation

- Encoding categorical variables
  - One-Hot Encoding (OHE)
    - Converts categorical variables into a series of binary columns
    - Suitable for nominal (unordered) categories
    - Example: Category=[Red, Blue, Green]

Color	X_{Red}	X_{Blue}	X_{Green}
Red	1	0	0
Blue	0	1	0
Green	0	0	1

Feature Engineering and Transformation

- Encoding categorical variables
  - Label encoding
    - Assigns a unique integer to each category without considering any ranking.
    - Suitable for nominal categorical data where categories have no natural order.
    - Can sometimes be problematic if a machine learning model incorrectly interprets the numerical values as having an order.
    - Example: Category=[Red, Blue, Green]
      Encoding: Red→0, Blue→1, Green→2
  - Ordinal encoding
    - Used for ordinal categorical data where categories have a clear order or ranking
    - Assigns numbers to categories based on their position in the order (e.g., "low" = 1, "medium" = 2, "high" = 3)
    - May introduce misleading information if the order between categories is not well-defined

Feature Engineering and Transformation

- Scaling and normalization
  - Min-max scaling: Rescales data to a fixed range [0,1]

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

  - Standardization (Z-score normalization): Transforms data to have zero mean and unit variance

$$X' = \frac{X - \mu}{\sigma}$$

  - Robust scaling: Uses median and IQR to reduce the effect of outliers

$$X' = \frac{X - \text{median}(X)}{IQR(X)}$$

Feature Engineering and Transformation

- Purposes of scaling and normalization
  - Ensuring fair comparisons between features
    - Different features in a dataset may have different units and scales (e.g., age in years vs. income in dollars). Scaling ensures that all features contribute equally to analysis and modeling
  - Improving convergence in machine learning models
    - Many machine learning algorithms (e.g., gradient descent-based models like logistic regression and neural networks) perform better when input data is normalized, leading to faster convergence and better optimization
  - Enhancing the interpretability of data
    - Some models (e.g., distance-based models like K-Nearest Neighbors and clustering algorithms) rely on numerical distances. Features with larger scales can dominate, leading to biased results. Scaling ensures that no single feature disproportionately influences outcomes
  - Reducing the impact of outliers
    - Robust scaling techniques (e.g., Robust Scaler using median and interquartile range) mitigate the influence of extreme values, improving model stability.
  - Facilitating data visualization
    - When visualizing multiple features together (e.g., in scatter plots or heatmaps), unscaled features can distort patterns. Scaling helps provide clearer insights.