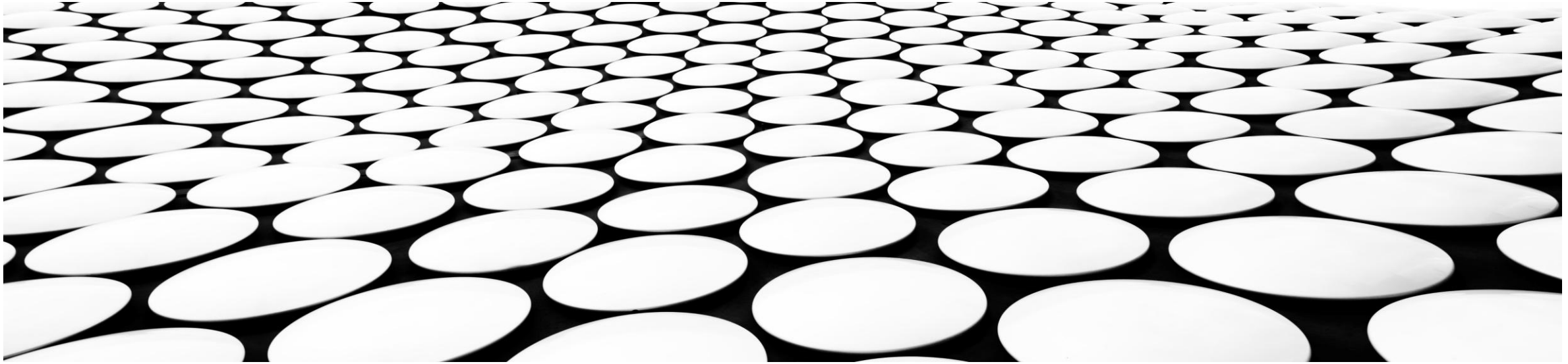
DATA MINING AND PREDICTIVE DATA ANALYTICS

DATA SUMMARIZATION AND VISUALIZATION



DATA SET

- A data set is a collection of related data, usually represented in tabular form,
  - Rows (records/samples/observations/objects/subjects/instances): represent individual objects
  - Columns (attributes/features/variables/dimensions): represent properties of the objects.
  - Cells (values): contain the actual data.

Applicant	Marital Status	Mortgage	Income ($)	Rank	Year	Risk
1	Single	Y	38,000	2	2009	Good
2	Married	Y	32,000	7	2010	Good
3	Other	N	25,000	9	2011	Good
4	Other	N	36,000	3	2009	Good
5	Other	Y	33,000	4	2010	Good
6	Other	N	24,000	10	2008	Bad
7	Married	Y	25,100	8	2010	Good
8	Married	Y	48,000	1	2007	Good
9	Married	Y	32,100	6	2009	Bad
10	Married	Y	32,200	5	2010	Good

VARIABLES

- Variables can be either qualitative or quantitative.
- A qualitative variable enables the elements to be classified or categorized according to some characteristic.
  - Examples: *status, mortgage, rank, and risk.*
  - Qualitative variables are also called categorical variables.
- A quantitative variable takes numeric values and allows arithmetic to be meaningfully performed on it.
  - Example: *income* and *year*.
  - Quantitative variables are also called numerical variables.

VARIABLES

- Data may be classified according to four levels of measurement:
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Nominal and ordinal data are categorical;
- Interval and ratio data are numerical.

VARIABLES

❑ Categorical variables

- Nominal data
  - Nominal data refer to names, labels, or categories.
  - There is no natural ordering, nor may arithmetic be carried out on nominal data.
  - Example: *status, mortgage,* and *risk*
  - *Other Example: Department (CSE, CSIT, MCA)*
- Ordinal data
  - Ordinal data can be rendered into a particular order.
  - However, arithmetic cannot be meaningfully carried out on ordinal data.
  - Example: *rank*
  - *Other Examples: Size (large, medium, small),*

VARIABLES

❑Numeric variables

- Interval data:
  - Interval data consist of quantitative data defined on an interval without a natural 0.
  - Addition and subtraction may be performed on interval data.
  - Example: *year* (Note that there is no “year 0.”)
- Ratio data:
  - Ratio data are quantitative data for which addition, subtraction, multiplication, and division may be performed.
  - A natural 0 exists for ratio data.
  - Example: *income*

VARIABLES

Interval-Scaled Data

- **Definition**: Numerical data where the **difference between values is meaningful**, but there is **no true zero**.
- **Key point**: Zero is arbitrary (it does not indicate the absence of the quantity). Ratios don't make sense.
- **Examples of Interval-Scaled Data**
  - Temperature in **Celsius or Fahrenheit** (20°C is 10° warmer than 10°C, but 0°C doesn't mean "no temperature").
  - Calendar years (e.g., 2000, 2020 → the difference of 20 years makes sense, but the year "0" is arbitrary).
  - IQ scores (an IQ of 140 is 20 points higher than 120, but not "twice as intelligent").
  - Dates in a timeline (e.g., birth years, exam dates).

VARIABLES

- **Ratio-Scaled Data**
  - **Definition**: Numerical data where both **differences and ratios are meaningful**, and there is a **true zero point** (which indicates absence of the quantity).
  - **Key point**: You can say "twice as much" or "half as much."
  - **Examples of Ratio-Scaled Data**
    - Temperature in **Kelvin** (0K = absolute absence of thermal energy).
    - Age (0 years = no age, and 20 years is twice 10 years).
    - Height and Weight (0 kg = no weight, and 60 kg is twice 30 kg).
    - Distance, Length, Speed (0 km = no distance, and 100 km is double 50 km).
    - Time duration (0 seconds = no time, 10 minutes is twice 5 minutes).

VARIABLES

Aspect	Interval Scale	Ratio Scale
Definition	Numerical scale where differences between values are meaningful, but **no true zero** exists.	Numerical scale where both differences and ratios are meaningful, with a **true zero point**.
Zero Meaning	Arbitrary zero (does not indicate absence).	Absolute zero (indicates complete absence of quantity).
Mathematical Operations	Addition, subtraction valid. Multiplication/division (ratios) **not meaningful**.	All operations valid: addition, subtraction, multiplication, division.
Examples	- Temperature in °C or °F - Calendar years - IQ scores - Dates on a timeline	- Temperature in Kelvin - Height, Weight - Age - Income/Salary - Time duration - Distance/Length
Illustration	20°C – 10°C = 10°C (difference meaningful), but 20°C is **not twice as hot** as 10°C.	20 kg – 10 kg = 10 kg (difference meaningful), and 20 kg is **twice as heavy** as 10 kg.

VARIABLES

- **Discrete Variable:**
  - Definition: Numeric variables that take **finite or countable** values.
  - They are often obtained by **counting**.
  - Cannot take values in between two numbers.
  - Examples:
    - Number of students in a class (30, 31, not 30.5)
    - Number of cars in a parking lot
    - Number of goals scored in a match
- **Graphical Representation:** Bar chart or stem-and-leaf plot.

VARIABLES

- **Continuous Variable:**
  - Definition: Numeric variables that can take **any value within a range**.
  - They are obtained by **measuring**.
  - Can include fractions and decimals.
  - Examples:
    - Height of students (e.g., 165.5 cm)
    - Weight of a person (e.g., 62.3 kg)
    - Time taken to run a race (e.g., 12.47 seconds)
  - **Graphical Representation:** Histogram or line graph.

VARIABLES

- Predictor Variable:
  - A predictor variable is a variable whose value is used to help predict the value of the response variable.
  - The predictor variables in Table are all variables, except *risk*.
- Response Variable:
  - A response variable is a variable of interest whose value is presumably determined at least in part by the set of predictor variables.
  - The response variable in Table is *risk*.

MEASURES OF CENTER, VARIABILITY, AND POSITION

- Basic Statistical Descriptions of Data
  - Measures of Central Tendency
    - Measure the location of the middle or center of a data distribution.
    - In particular, we discuss the mean, median, mode, and midrange
  - Dispersion of The Data
    - How the data are spread out in the data distribution.
    - The most common data dispersion measures are the range, quartiles, and interquartile range; the five-number summary and boxplots; and the variance and standard deviation of the data.

MEASURES OF CENTER, VARIABILITY, AND POSITION

- Mean (Arithmetic Average): The sum of all data values divided by the number of values.
  - Example: Dataset = {2, 4, 6, 8, 10} Mean = (2 + 4 + 6 + 8 + 10)/5 = 30/5 = 6
  - Advantages: Easy to compute, uses all data points.
  - Limitations: Sensitive to outliers (e.g., extreme values can distort the mean).
- Median: The middle value when the data is arranged in ascending (or descending) order.
  - Calculation: If n is odd ⟶ Median = middle value.
  - If n is even ⟶ Median = average of the two middle values.
  - Example: Dataset = {3, 5, 7, 9, 11} ⟶ Median = 7

    Dataset = {3, 5, 7, 9} ⟶ Median = (5+7)/2 = 6
  - Advantages: Not affected by outliers or skewed data.
  - Limitations: Ignores the magnitude of values (uses only positional data).

MEASURES OF CENTER, VARIABILITY, AND POSITION

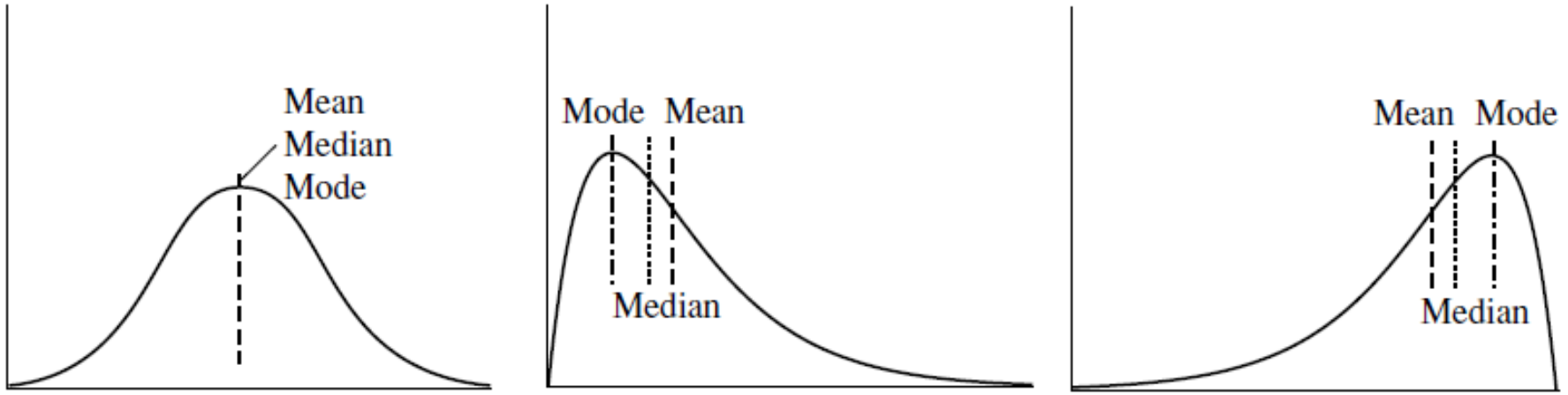
- **Mode:** The value(s) that appear most frequently in a dataset.
  - Types:
    - Unimodal (One mode), Bimodal (Two modes), Multimodal (More than two modes)
    - No mode: If all values occur with the same frequency
  - Example: Dataset = {2, 4, 4, 6, 8, 8, 8, 10} ⟶ Mode = 8
  - Advantages: Useful for categorical data (e.g., "most purchased product").
  - Limitations: May not exist or may not be unique.
- **Midrange:** It is the average of the largest and smallest values in the set
  - Applicable for numeric data.

MEASURES OF CENTER, VARIABILITY, AND POSITION

Measure	Best Used When	Sensitive to Outliers?	Data Type
Mean	Symmetric, numeric data	Yes	Interval/Ratio
Median	Skewed data or with outliers	No	Ordinal, Interval/Ratio
Mode	Most frequent item needed	No	Nominal, Ordinal, Interval

MEASURES OF CENTER, VARIABILITY, AND POSITION

- **Measuring the Dispersion of Data**
- Symmetry in data distribution
  - In a unimodal frequency curve with perfect symmetric data distribution, the mean, median, and mode are all at the same center value
    - Data in most real applications are not symmetric.
    - Positively Skewed: the mode < median.
    - Negatively Skewed: the mode > median



(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

MEASURES OF CENTER, VARIABILITY, AND POSITION

- Measures of dispersion (or variability) capture the degree to which values differ from each other and from the center.
- The measures of variability :
  - The Range
  - Quartiles
  - Interquartile Range
  - Five-number Summary
  - Boxplot
  - Z Score
  - The Variance
  - The Standard Deviation

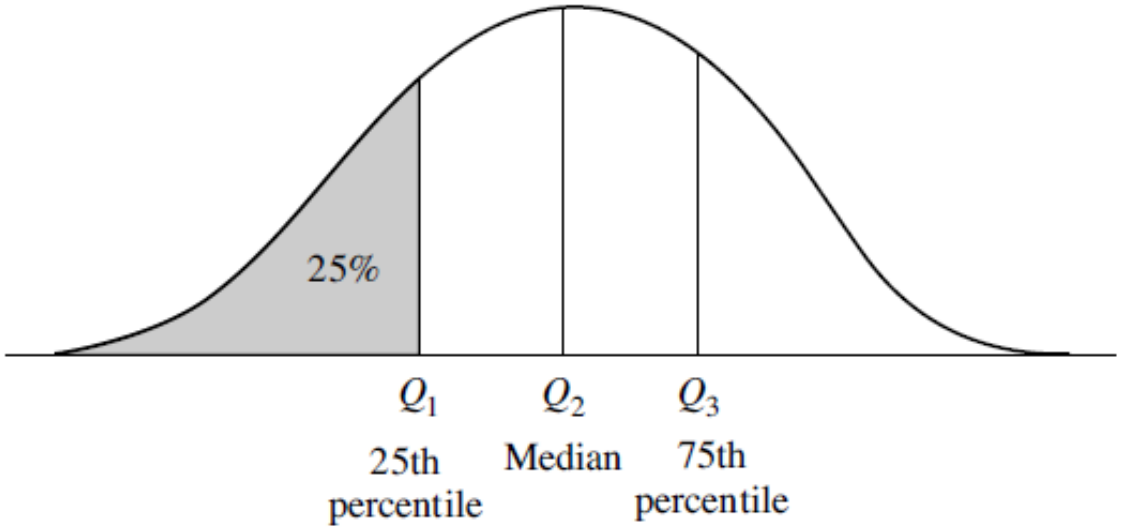
MEASURES OF CENTER, VARIABILITY, AND POSITION

- Range
  - The range is the difference between the max. and min. values in the dataset.
  - Range=Maximum Value−Minimum Value
  - **Illustration:**

    Dataset: 5, 7, 9, 10, 12
    - **Maximum = 12, Minimum = 5**
    - **Range = 12 – 5 = 7**
  - **Limitations:**
    - **Very sensitive to outliers (extreme values).**
    - **Ignores distribution of values between extremes.**

MEASURES OF CENTER, VARIABILITY, AND POSITION

- Quartiles
  - **Definition:** Quartiles divide ordered data into four equal parts.
  - Q1 (1st Quartile): 25% of data below it
  - Q2 (Median): 50% of data below it
  - Q3 (3rd Quartile): 75% of data below it
  - **Illustration:**
    - **Dataset (ordered): 5, 7, 9, 10, 12**
    - **Q1 = 7,**
    - **Median (Q2) = 9,**
    - **Q3 = 10**



MEASURES OF CENTER, VARIABILITY, AND POSITION

- Interquartile Range (IQR)
  - Definition: IQR measures the spread of the middle 50% of the data.

$$IQR = Q3 - Q1$$

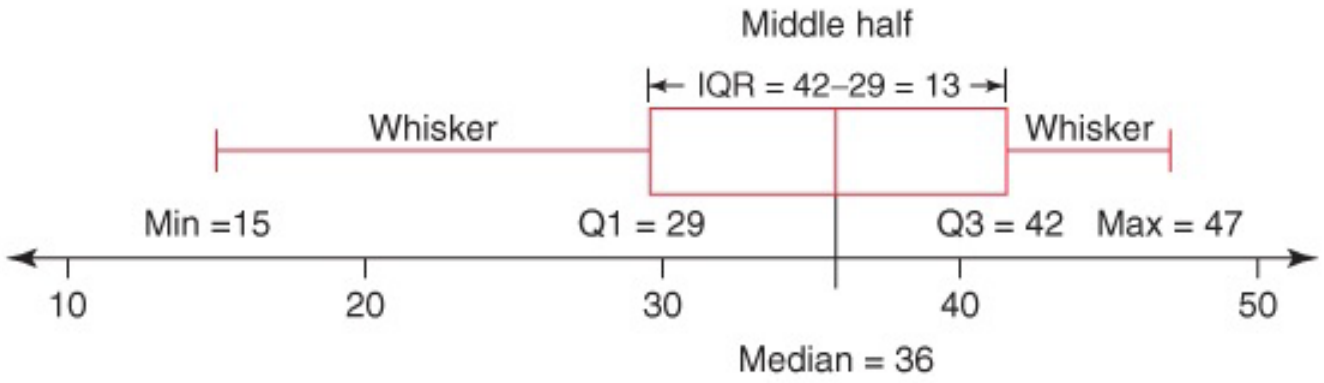
  - Example:
    - From dataset above: Q3 = 10, Q1 = 7 ⟶ IQR = 3
  - Importance:
    - Resistant to outliers (only considers central values).
    - Useful in detecting extreme values or outliers
  - NOTE: A common rule of thumb for identifying suspected outliers is to single out values falling at least (1.5 X IQR) above the third quartile or below the first quartile.

MEASURES OF CENTER, VARIABILITY, AND POSITION

- **Five-Number Summary**
  - **Definition:** A concise statistical summary of data consisting of:
    1. **Minimum**
    2. **Q1**
    3. **Median (Q2)**
    4. **Q3**
    5. **Maximum**
  - **Example: Dataset: 5,7,8,12,13,14,18,21,23,25**
    - **Median (Q2): Middle of dataset → (13 + 14)/2 = 13.5**
    - **Q1 (lower quartile): Median of lower half (5, 7, 8, 12, 13) → 8**
    - **Q3 (upper quartile): Median of upper half (14, 18, 21, 23, 25) → 21**
    - **Five-Number Summary : (5, 8, 13.5, 21, 25)**

MEASURES OF CENTER, VARIABILITY, AND POSITION

- **Box Plot (or whisker plot)**
  - A boxplot is a graphical representation of the five-number summary of a dataset.
  - Features of a Box plot
    - **Box:** Extends from Q1 to Q3 (the Interquartile Range, IQR).
    - **Median Line**: A line inside the box shows the median (Q2).
    - **Whiskers:** Lines extending from the box to the minimum and maximum values (excluding outliers).
    - **Outliers:** Data points that lie beyond $Q1-1.5\times IQR$ or $Q3+1.5\times IQR$.



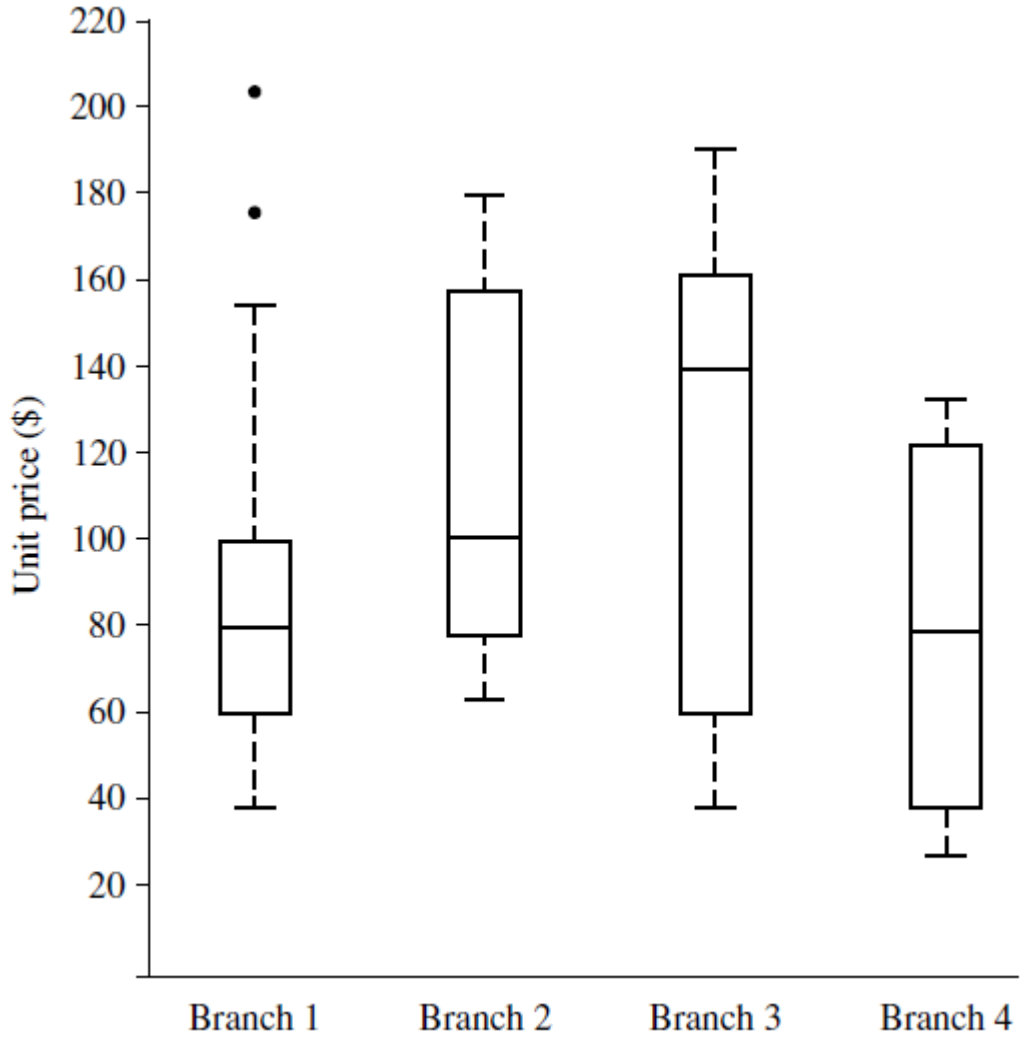
MEASURES OF CENTER, VARIABILITY, AND POSITION

Example Dataset: 5,7,8,12,13,14,18,21,23,25

- Median (Q2): Middle of dataset → (13 + 14)/2 = 13.5
- Q1 (lower quartile): Median of lower half (5, 7, 8, 12, 13) → 8
- Q3 (upper quartile): Median of upper half (14, 18, 21, 23, 25) → 21
- Five-Number Summary : (5, 8, 13.5, 21, 25)
- Determine Whiskers:
  - Lower fence = Q1 – 1.5 × IQR = 8 – 19.5 = -11.5
  - Upper fence = Q3 + 1.5 × IQR = 21 + 19.5 = 40.5
  - Since all values lie between -11.5 and 40.5 → No outlier

MEASURES OF CENTER, VARIABILITY, AND POSITION

- **Example (Box Plot)**
  - Figure shows boxplots for unit_price data for items sold at four branches of AllElectronics during a given time period.
  - Box Plot for Branch1 have two Outliers as shown.



MEASURES OF CENTER, VARIABILITY, AND POSITION

- Z-Score (or Standard Score)
  - A Z-Score measures how many standard deviations a data point is from the mean.
  - It standardizes data, making values from different distributions comparable.

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

  - Where: X = observed value
  - μ = mean of the dataset
  - σ = standard deviation

  - Key Points
    - Z=0: Value is exactly at the mean.
    - Z>0: Value lies above the mean.
    - Z<0: Value lies below the mean.
    - A higher absolute Z-score indicates a point farther from the mean.

  - NOTE: Outlier - Using Z-Scores: Data points with |Z|>3| are often considered outliers.

MEASURES OF CENTER, VARIABILITY, AND POSITION

- **Variance**
  - Definition: Variance measures the average squared deviation from the mean.

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  - Example:
    - Dataset: 2, 4, 6
    - Mean = 4
    - Deviations: $(2–4)^2=4$, $(4–4)^2=0$, $(6–4)^2=4$
    - Variance = (4+0+4)/3 = 8/3 ≈ 2.67

MEASURES OF CENTER, VARIABILITY, AND POSITION

- Standard Deviation
  - Definition: Standard deviation (SD) is the square root of variance.
  - Gives spread in the same units as the data.

$$\sigma = \sqrt{\sigma^2}$$

  - Example:
    - From above: Variance = 2.67 → SD ≈ 1.63
  - Interpretation:
    - Low SD → data close to mean
    - High SD → data widely spread

MEASURES OF CENTER, VARIABILITY, AND POSITION

Measure	Uses	Sensitivity to Outliers	Example Insight
Range	Quick spread estimate	Very sensitive	Salary range in a company
Quartiles	Position of data	Resistant	Helps compare distributions
IQR	Spread of middle 50%	Resistant	Detects variability in core data
Five-Number Summary	Quick summary	Moderate	Useful in exploratory analysis
Boxplot	Visual comparison	Moderate	Highlights outliers
Variance	Mathematical analysis	Very sensitive	Used in advanced stats, ML
Standard Deviation	Practical interpretation	Very sensitive	Risk measurement in finance