

Advanced Data Visualization Lecture Notes (Lecture 1 - 2 Hours)

I. Introduction to Data Types (30 minutes)

A. Overview of Data Types

1. Binary Data

- Definition: Data with **two possible values** (e.g., Yes/No, True/False).
- Examples: Gender (Male/Female), Employment Status (Employed/Unemployed).

2. Categorical Data (Nominal)

- Definition: Data with **distinct categories without a specific order**.
- Examples: Blood Type (A, B, AB, O), Brand Names (Nike, Adidas).

3. Ordinal Data

- Definition: **Categorical data** with a **meaningful order** but **without a consistent scale**.
- Examples: Survey ratings (Poor, Fair, Good, Excellent), Education Levels (High School, Bachelor's, Master's).

4. Scale Data (Continuous)

- Encompasses **Interval** and **Ratio data**.
- Examples: Temperature, Age, Salary.

5. Interval Data

- Definition: **Continuous data** with **equal intervals** between values but **no true zero point**.
- Examples: Temperature in Celsius or Fahrenheit, Dates.

6. Ratio Data

- Definition: **Continuous data** with **equal intervals** and a **true zero point**.
- Examples: Height, Weight, Distance, Sales revenue.

B. Summary and Examples

- Discuss real-world datasets and identify the data types present.
 - Group activity: Categorize data from a provided dataset.
-

II. Measures of Central Tendency and Dispersion (30 minutes)

A. Measures of Central Tendency

1. Mean (Average)

- Definition: Sum of all values divided by the number of values.
- Appropriate for: **Interval** and **Ratio** data.
- Example: Average income, average temperature.

2. Median

- Definition: Middle value when data is ordered.
- Appropriate for: Ordinal, Interval, and Ratio data.
- Example: Median household income, median age.

3. Mode

- Definition: Most frequently occurring value.
- Appropriate for: Nominal, Ordinal, Interval, and Ratio data.
- Example: Most common blood type, most frequent rating in a survey.

B. Measures of Dispersion

1. Range

- Definition: Difference between the highest and lowest values.
- Appropriate for: Interval and Ratio data.
- Example: Temperature range, range of ages in a class.

2. Variance

- Definition: Measure of how much values differ from the mean.
- Appropriate for: Interval and Ratio data.
- Example: Variance in test scores, variance in income.

3. Standard Deviation

- Definition: Square root of the variance.
- Appropriate for: Interval and Ratio data.
- Example: Standard deviation of heights, standard deviation of sales revenue.

C. Summary and Practical Examples

- Discuss the importance of understanding central tendency and dispersion.
 - Group activity: Calculate these measures using sample data.
-

III. Appropriate Visualization Techniques for Each Data Type (60 minutes)

A. Visualizing Binary Data

1. Bar Chart

- Simple and effective for showing proportions of two categories.
- Example: Employment status, survey responses (Yes/No).

2. Pie Chart

- Useful for showing the composition of a binary dataset.
- Example: Market share of two competing products.

B. Visualizing Categorical Data (Nominal)

1. Bar Chart

- Display frequency of each category.

- Example: Distribution of blood types, number of students per major.
- 2. **Pie Chart**
 - Show **percentage of each category** in the whole.
 - Example: Market share distribution, customer preferences.

C. Visualizing Ordinal Data

1. **Bar Chart**
 - Display **ordered categories** with the **frequency of each**.
 - Example: Survey ratings, levels of education.
2. **Box Plot**
 - Show **distribution and identify outliers**.
 - Example: Student performance ratings, customer satisfaction levels.

D. Visualizing Scale Data (Interval and Ratio)

1. **Histogram**
 - Show the **distribution of continuous data**.
 - Example: Distribution of ages, distribution of income.
2. **Box Plot**
 - Display **distribution, median, quartiles, and outliers**.
 - Example: Salary distribution, test scores.
3. **Scatter Plot**
 - Show **relationship between two continuous variables**.
 - Example: Relationship between height and weight, age and income.
4. **Line Chart**
 - Show **trends over time**.
 - Example: Stock prices over time, temperature changes over a year.

E. Advanced Techniques

1. **Heatmap**
 - Show the **intensity of data** at intersections of categories.
 - Example: Correlation matrix, activity levels over time.
2. **Bubble Chart**
 - Display **three dimensions of data (x, y, size)**.
 - Example: Sales data with region, product, and revenue.
3. **Violin Plot**
 - Combine box plot and density plot for richer data visualization.
 - Example: Distribution of exam scores across different classes.

F. Summary and Interactive Session

- Review the visualizations and their appropriate use cases.
- Hands-on activity: Create visualizations using sample datasets.

Measures of Central Tendency and Dispersion

I. Measures of Central Tendency

A. Mean (Average)

Definition: The mean is the sum of all values divided by the number of values. It provides a measure of the central value of a dataset.

Calculation:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

- x_i is each individual value in the dataset.
- n is the number of values.

Example:

Dataset: 5, 8, 12, 20, 25

$$\bar{x} = \frac{5 + 8 + 12 + 20 + 25}{5} = \frac{70}{5} = 14$$

B. Median

Definition: The median is the middle value in an ordered dataset. If the dataset has an even number of observations, the median is the average of the two middle numbers.

Calculation:

1. Order the dataset from smallest to largest.
2. If the number of values (n) is odd, the median is the middle value.
3. If n is even, the median is the average of the two middle values.

Example:

Example (Odd number of values):

Dataset: 5, 8, 12, 20, 25

Median = 12

Example (Even number of values):

Dataset: 5, 8, 12, 20, 25, 30

Median = $\frac{12+20}{2} = 16$

C. Mode

Definition: The mode is the value that appears most frequently in a dataset. A dataset can have more than one mode (bimodal, multimodal) or no mode if no number repeats.

Calculation: Identify the value(s) that occur most frequently in the dataset.

Example:

Example:

Dataset: 5, 8, 12, 12, 20, 25

Mode = 12

Dataset: 5, 8, 8, 12, 12, 20, 25

Modes = 8 and 12 (bimodal)

II. Measures of Dispersion

A. Range

Definition: The range is the difference between the highest and lowest values in a dataset. It provides a measure of how spread out the values are.

Calculation:

$$\text{Range} = \text{Maximum Value} - \text{Minimum Value}$$

Example: Dataset: 5, 8, 12, 20, 25 Range = $25 - 5 = 20$

B. Variance

Definition: Variance measures the average squared deviation of each value from the mean. It indicates how spread out the values are around the mean.

Calculation:

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

- x_i is each individual value.
- \bar{x} is the mean.
- n is the number of values.

Example:

Dataset: 5, 8, 12, 20, 25

Mean = 14

$$\begin{aligned}\sigma^2 &= \frac{(5 - 14)^2 + (8 - 14)^2 + (12 - 14)^2 + (20 - 14)^2 + (25 - 14)^2}{5} \\ &= \frac{81 + 36 + 4 + 36 + 121}{5} = \frac{278}{5} = 55.6\end{aligned}$$

A. Standard Deviation

Definition: Standard deviation is the square root of the variance. It provides a measure of the average distance of each value from the mean.

Calculation:

$$\sigma = \sqrt{\text{Variance}}$$

Example:

Variance = 55.6

$$\sigma = \sqrt{55.6} \approx 7.45$$
