Data Analysis/Exploration

Notes by Mannan Ul Haq (BDS-3C)

Data analysis is the process of examining and understanding data to gain insights, make informed decisions, and solve problems.

Data Quality

Data quality is a measure of how good or reliable data is. It considers factors like accuracy, completeness, consistency, reliability, and up-to-date. High-quality data is essential for digital businesses.

Dimensions of Data Quality

1. **Accuracy** (Is the information correct?)
    - Data should be error-free and reflect real-world scenarios.
    - Errors can lead to significant problems, such as unauthorized access to bank accounts.
2. **Completeness** (How comprehensive is the information?)
    - Completeness measures how exhaustive a dataset is.
    - It ensures that all required values are available, making the information usable.
3. **Consistency/Reliability** (Does the information contradict other trusted resources?)
    - Consistency refers to data uniformity across networks and applications.
    - Data in different locations should not conflict with each other.
4. **Relevance/Timeliness** (Is the information needed, and is it up-to-date?)
    - Relevance checks if data fulfills its intended purpose.
    - Timeliness ensures data is available when required, preventing wrong decisions.
5. **Interpretability** (How easy is it to understand the data?)

- Interpretability reflects how easily data can be understood.

Measuring these data quality dimensions helps organizations identify and resolve data errors, ensuring that their data is fit for its intended purpose. High-quality data is the cornerstone of effective digital businesses.

Exploring the Dataset

Once you have the data, the next step is to explore it. This preliminary investigation helps you understand the specific characteristics of the data. It can answer questions like:

- **Is the data balanced**, or are there more instances of one category than another (class balance)?
- Do certain attributes vary together, indicating potential **correlations**?
- How are the values of data attributes spread out **(dispersion)**?
- Skewness
- Are there **missing values** in the dataset?
- Are there any extreme values **(outliers)** that need attention?

Class Balance

Class Balance or **Data Balance** refers to the distribution of data points among different categories or classes within a dataset.

Imagine you have a dataset of customer reviews for a product, and you want to classify these reviews as either **"positive"** or **"negative"**. If you have **90% positive** reviews and only **10% negative** reviews, you have an imbalance because one class (positive) dominates the dataset, while the other class (negative) is underrepresented.

Here's a simple explanation:

- **Balanced Data**: When you have roughly an equal number of data points for each class or category, it's called balanced data. For example, if you have 50 positive reviews and 50 negative reviews in your dataset, it's balanced.
- **Imbalanced Data**: When one class has significantly more data points than the other class, it's called imbalanced data. For example, if you have 90 positive reviews and

only 10 negative reviews, it's imbalanced.

Why is Class Balance Important?

Class balance is essential because it can impact the performance of machine learning models. In an imbalanced dataset, models may become biased towards the majority class (the one with more data points) because they see more examples of it. This can lead to poor predictions for the minority class.

Attributes Correlation

Correlation is a statistical measure that helps us understand the relationship or association between two or more variables in a dataset. It tells us how these variables change in relation to each other. Correlation is often used to determine whether there's a connection between variables and, if so, the strength and direction of that connection.

Here are some key points about correlation:

1. **Positive Correlation**: When two variables have a positive correlation, it means that as one variable increases, the other tends to increase as well.
2. **Negative Correlation**: Conversely, a negative correlation indicates that as one variable increases, the other tends to decrease. They move in opposite directions.
3. **No Correlation**: If there's no apparent pattern or relationship between two variables, they are said to have no correlation. Changes in one variable do not have a consistent effect on the other.

Correlation Coefficient:

To quantify the strength and direction of the correlation between two variables, we use a number called the **correlation coefficient**. This number ranges between -1 and 1.

- A correlation coefficient of **1** indicates a perfect positive correlation.
- A correlation coefficient of **-1** indicates a perfect negative correlation.
- A correlation coefficient close to **0** suggests little to no correlation.

Descriptive Statistics

In data analysis, statistics are important for summarizing and understanding the data. Depending on whether the data attributes are discrete or continuous, different statistics are calculated:

Measures of Central Tendency

In data analysis, measures of central tendency help us understand the central or typical value within a dataset. They provide insights into where the data tends to cluster. There are three primary measures of central tendency: the **mean**, the **median**, and the **mode**.

Mean:

The **mean** is often referred to as the average. It's calculated by adding up all the values in a dataset and then dividing by the number of data points. Here's the advantage and limitation of using the mean:

Advantage of the Mean:

- The mean can be used for both continuous and discrete numeric data.

Limitations of the Mean:

- The mean cannot be calculated for categorical data because the values cannot be summed.
- The mean is not robust against outliers, meaning that a single large value (an outlier) can significantly skew the average.

Median:

The **median** is the middle value in a dataset when it's ordered from smallest to largest. The median is less affected by outliers and skewed data, making it a preferred measure of central tendency when the data distribution is not symmetrical.

2 2 5 6 7 8 9

Mode:

The **mode** is the value that occurs most frequently in a dataset.

2 2 5 6 7 8 9

Advantage of the Mode:

- Unlike the mean and median, which are mainly used for numeric data, the mode can be found for both numerical and categorical (non-numeric) data.

Limitation of the Mode:

- In some distributions, the mode may not reflect the center of the distribution very well, especially if the data is multimodal (has multiple modes) or if all values occur with similar frequencies.

Measures of Dispersion

Measures of dispersion provide insights into how data values are spread out or vary within a dataset. Common measures of dispersion include the range, variance, standard deviation.

Range:

The **range** is the simplest measure of dispersion. It's calculated by subtracting the minimum value from the maximum value in the dataset. While it's easy to compute, it's sensitive to extreme values (outliers) and may not provide a complete picture of data variability.

Variance and Standard Deviation:

Variance is a measure that quantifies how far each data point is from the mean.

Standard deviation is the square root of the variance and provides a measure of dispersion in the same units as the data.

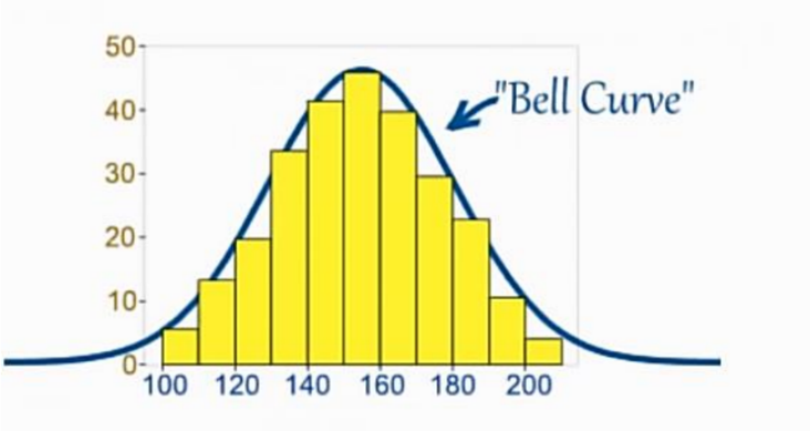
These measures give us a more detailed understanding of how data points deviate from the mean. Higher variance and standard deviation values indicate greater data spread.

Data Distribution

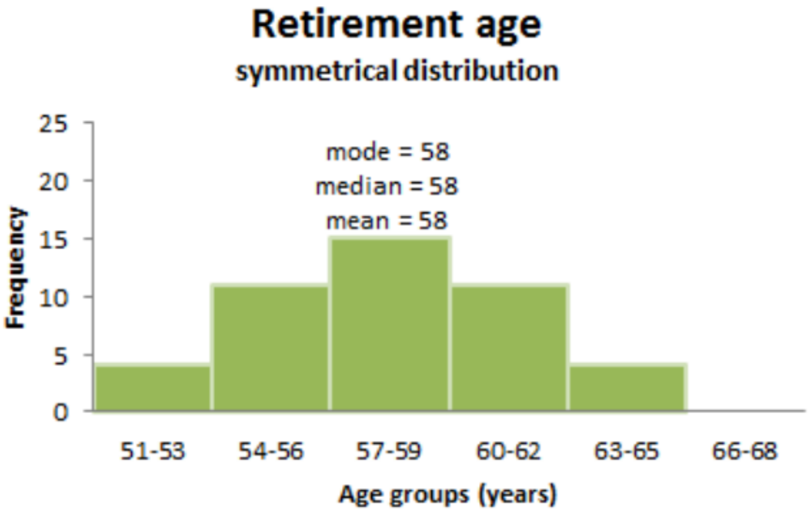
The shape of a data distribution can significantly affect the choice of measures of central tendency:

Normal or Symmetrical Distribution:

A **normal distribution** is a specific type of distribution where data tends to cluster around a central value with no bias to the left or right. It is often represented as a bell curve.



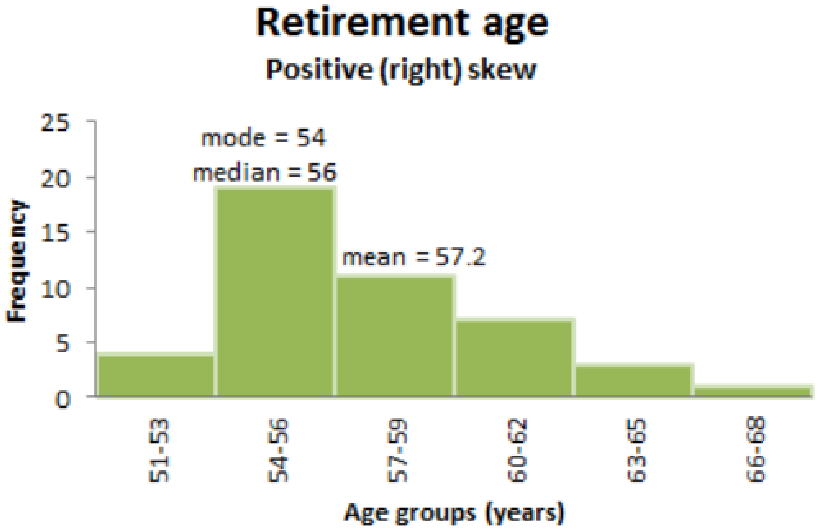
In a normal or symmetrical distribution, the mean, median, and mode are all centered at the same point. They are approximately equal, making any of these measures a suitable representation of central tendency.



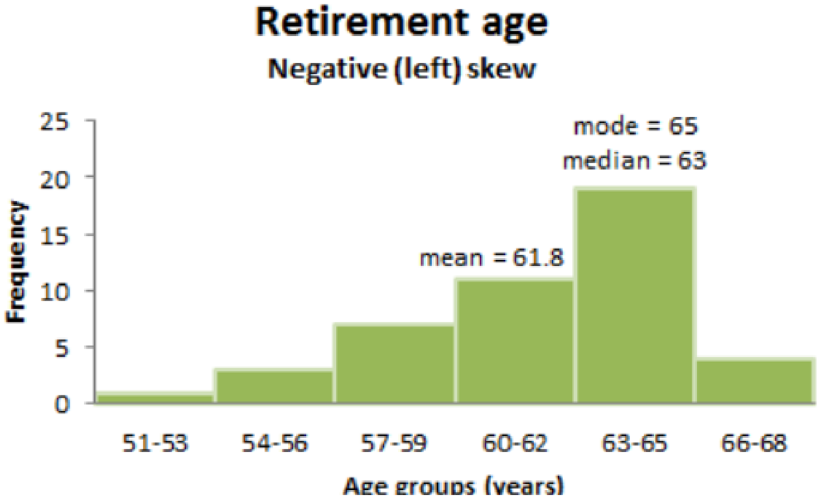
Skewness:

Skewness measures the degree of asymmetry in a distribution. When a distribution is skewed, it deviates from the symmetrical bell curve of a normal distribution. In skewed distributions:

- **Positive Skewness (Right Skewed)**: The tail on the right side is longer, and the mean is pulled toward the right tail. In such cases, the median is often preferred as a measure of central tendency because it's less affected by extreme values.

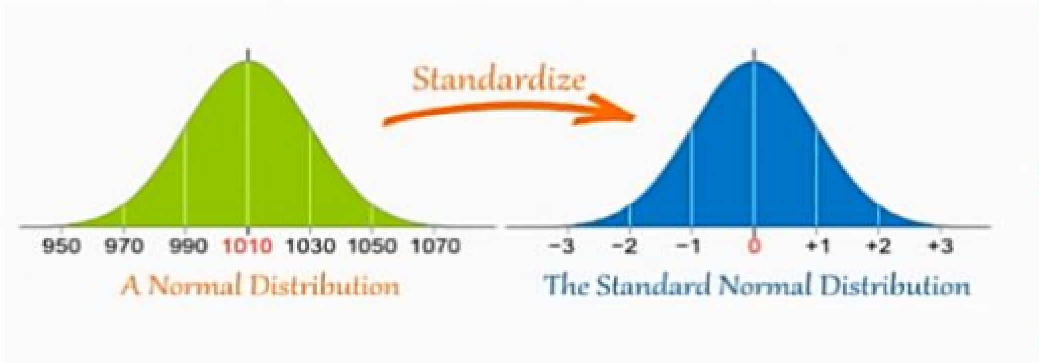


- **Negative Skewness (Left Skewed)**: The tail on the left side is longer, and the mean is pulled toward the left tail. Again, the median is often a better choice in this situation.



Standard Normal Distribution

The **Standard Normal Distribution**, also known as the **Z-Distribution**, is a specific type of probability distribution. It is a continuous probability distribution that is symmetrically shaped like a bell curve. This distribution has a mean (average) of 0 and a standard deviation of 1.



Z-Score:

A Z-Score, also known as a standard score, measures how many standard deviations a particular data point is away from the mean of a distribution. It's a way to standardize data and compare it to the standard normal distribution.

Z = x * μ / σ

x = data point

μ= mean

σ = standard deviation

The Z-Score tells you how many standard deviations a data point is above or below the mean. A positive Z-Score indicates that the data point is above the mean, while a negative Z-Score indicates that it's below the mean.

The Z-Score is useful for comparing data points from different distributions, identifying outliers, and making statistical inferences. In particular, it helps determine how extreme or unusual a data point is in the context of its distribution.