ATME College of Engineering

Data Visualization 21AD71

Prepared By,
Dr. Anitha DB
Associate Professor & Head
Department of CSE-Data Science
ATME College of Engineering, Mysuru

Module1 :Data Visualization and Data Exploration

- **Introduction:** Data Visualization, Importance of Data Visualization, Data Wrangling, Tools and Libraries for Visualization
- **Overview of Statistics:** Measures of Central Tendency, Measures of Dispersion, Correlation, Types of Data, Summary Statistics
- **Numpy:** Numpy Operations - Indexing, Slicing, Splitting, Iterating, Filtering, Sorting, Combining, and Reshaping
- **Pandas:** Advantages of pandas over numpy, Disadvantages of pandas, Pandas operation - Indexing, Slicing, Iterating, Filtering, Sorting and Reshaping using Pandas

Topic1: Introduction

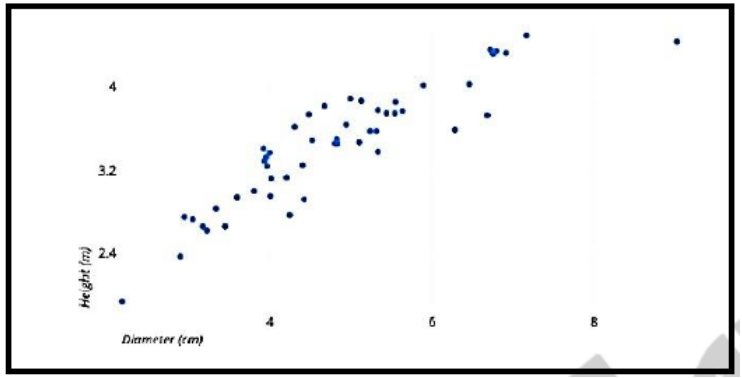
- Data Visualization,
- Importance of Data Visualization,
- Data Wrangling,
- Tools and Libraries for Visualization

Introduction to Data Visualization

- Computers and smartphones store data such as names and numbers in digital format.
- Data representation refers to the forms in which we can store, process, and transmit data.
- Effective representations can narrate story and convey fundamental discoveries to audience
- Creating representations helps to achieve a more precise, more concise, and more direct perspective of information , making it easier for anyone to understand the data.
- Representations are useful apparatus to derive insights from the data
- Representations convert data into useful information.

The Importance of Data Visualization

- Instead of just looking at data in the columns of an Excel spreadsheet, we get a better idea of what our data contains by using visualization.
- For instance, it is easy to see a pattern emerge from the numerical data that's given in the following scatter plot.
- It shows the correlation between diameter and the height of various trees.
- There is a positive correlation between diameter and height.



The Importance of Data Visualization

Visualizing data has many advantages

- Complex data can be easily understand
- A simple visual representation of outliers, target audiences, and futures market can be created
- Storytelling can be done using dashboards and animations
- Data can be explored through interactive visualizations

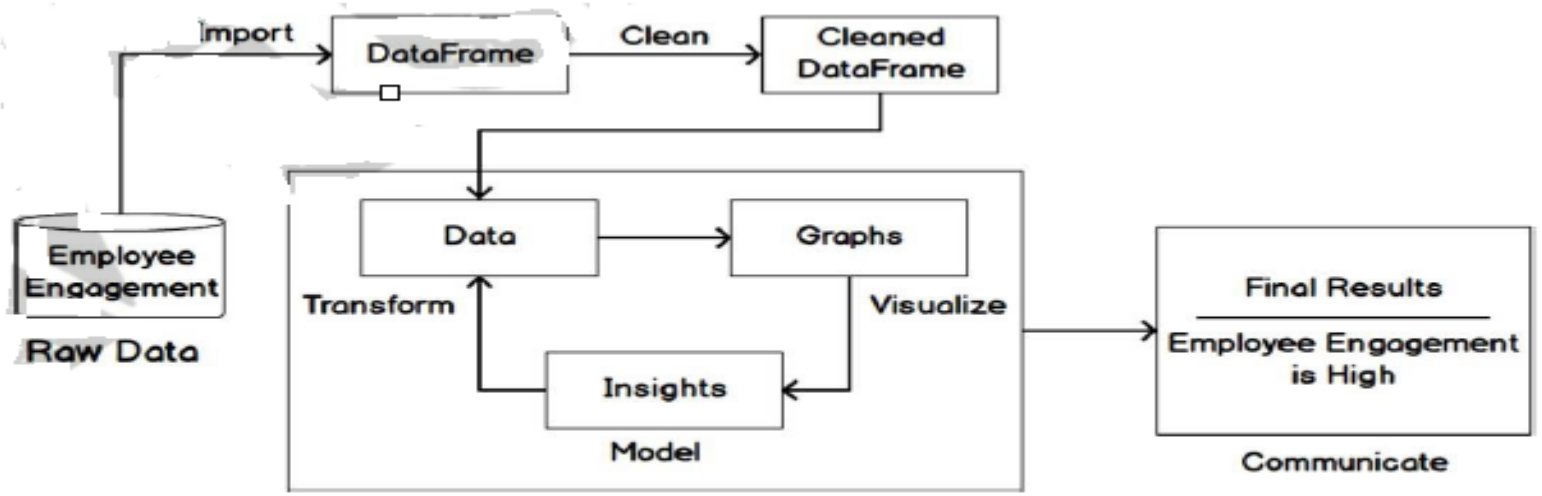
Questions

Briefly explain Data Visualization.

Why Data Visualization is Important/Significant?

Data Wrangling

- **Data wrangling** is the process of transforming raw data into a suitable representation for various tasks. It is the discipline of augmenting, cleaning, filtering, standardizing, and enriching data in a way that allows it to be used in a downstream task, which in our case is data visualization.
- Examine the following **flow diagram** of the data wrangling process to understand how precise and actionable data is prepared for business analysts to utilize.



Data wrangling process to measure employee engagement

Data Wrangling

The following steps explain the flow of the data wrangling process:

1.First, the Employee Engagement data is in its raw form.

2. Then, the data gets imported as a DataFrame and is later cleaned.

3.The cleaned data is then transformed into graphs, from which findings can be derived.

4.Finally, we analyze this data to communicate the final results.

For example, employee engagement can be measured based on raw data gathered from feedback surveys, employee tenure, exit interviews, one-on-one meetings, and so on. This data is cleaned and made into graphs based on parameters such as referrals, faith in leadership, and scope of promotions. The percentages, that is, information derived from the graphs, help us reach our result, which is to determine the measure of employee engagement.

Tools and Libraries for Visualization

- Several tools are available for creating data visualizations to suit different needs.
- Non-coding tools like **Tableau** provide an intuitive interface for exploring and understanding data.
- Alongside **Python**, **MATLAB** and **R** are also commonly used in data analytics.
- Python stands out as the industry's preferred language due to its user-friendly nature and efficiency in data manipulation and visualization.
- Its extensive library ecosystem further enhances Python's appeal, making it the optimal choice for robust data visualization tasks.

Questions:

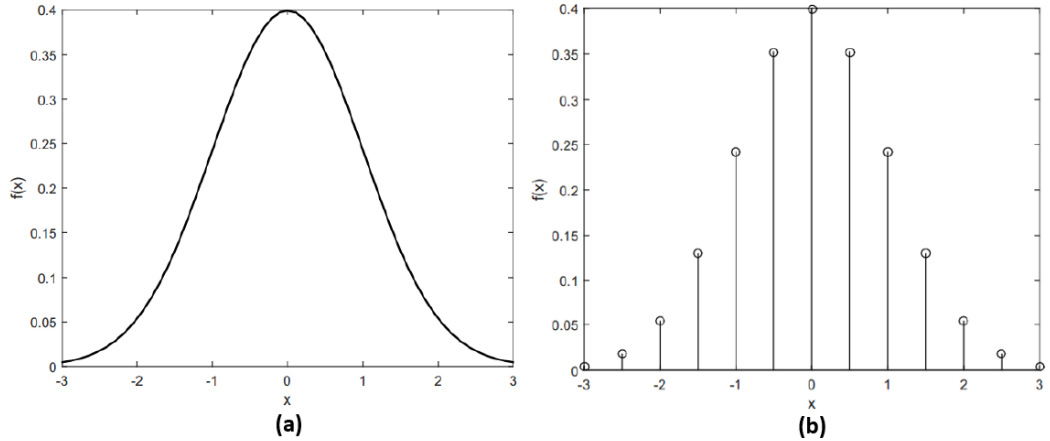
1. What is Data Wrangling?
2. Explain the data wrangling process with an example of employee engagement.
3. With a neat diagram explain the steps involved in the Data Wrangling process.

Topic 2: Overview of Statistics

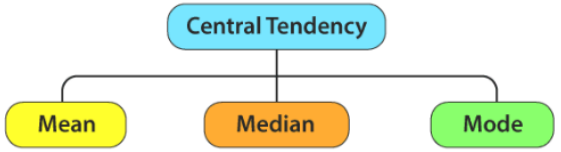
- Measures of Central Tendency,
- Measures of Dispersion,
- Correlation,
- Types of Data,
- Summary Statistics

Overview of Statistics

- **Statistics** is a combination of the analysis, collection, interpretation and representation of numerical data.
- **Probability** is a measure of the likelihood that an event will occur and is quantified as a number between 0 and 1
- A **probability distribution** is a function that provides the probability for every possible event. It is frequently used for statistical analysis.
- There are two types of probability distributions, namely continuous and discrete.



Measures of Central Tendency



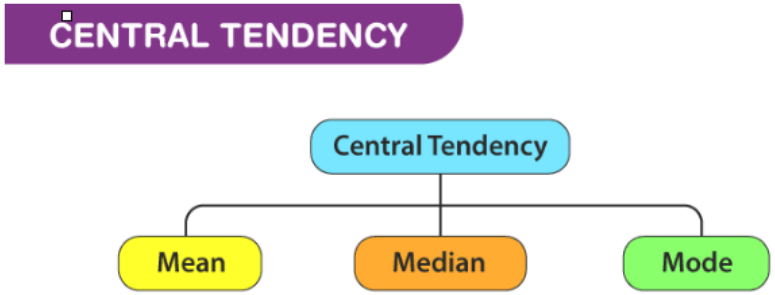
Measures of central tendency are often called **averages** and describe central or typical values for a probability distribution. Three kind of averages are Mean, Median and Mode.

Mean: The arithmetic average is computed by summing up all measurements and dividing the sum by the number of observations. The mean is calculated as follows $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

Median: The middle value in a dataset that is arranged in ascending order (from the smallest value to the largest value). If a dataset contains an even number of values, the median of the dataset is the mean of the two middle values. The median is less prone to outliers compared to the mean, where the outliers are distinct values in data

Mode: Defines the most frequently occurring value in a dataset. In some cases, a dataset may contain multiple modes, while some datasets may not have any mode at all.

Measures of Central Tendency



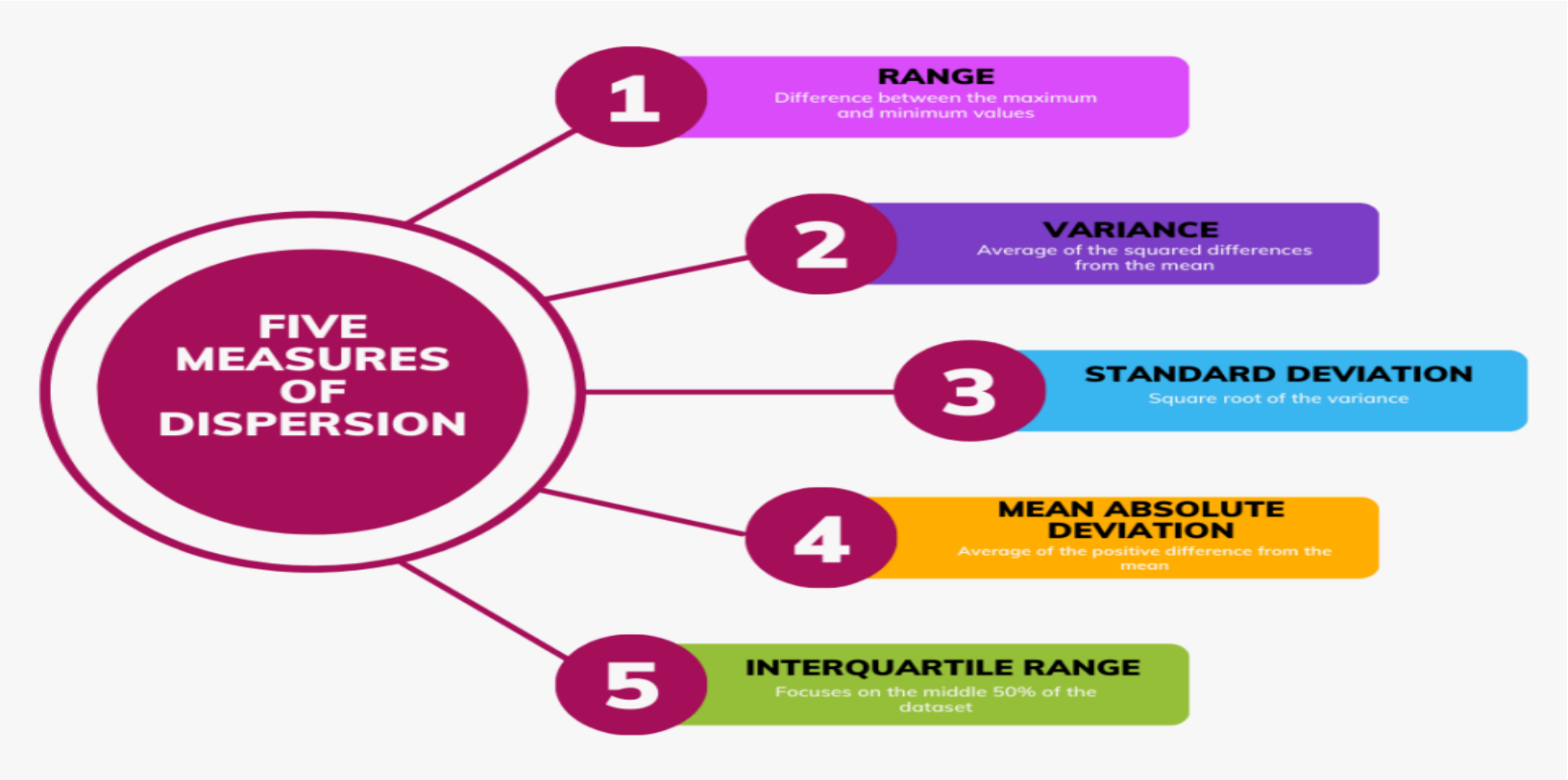
Example:A die was rolled 10 times and we got the following numbers:4,5,4,3,4,2,1,1,2,1. Find the central tendency.

Mean= (4+5+4+3+4+2+1+1+2+1)/10=2.7

Medin=middle value of ordered data=midlle value(1,1,1,2,2,3,4,4,4,5)=(2+3)/2=2.5

Mode=1 and 4

Measures of Dispersion



Measures of Dispersion

Variance: Variance is a measure of how far each data point in the set is from the mean and is calculated by taking the average of the squared differences from the mean.
Variance (σ^2) = $\Sigma(xi - \mu)^2 / N$, where μ is the mean and N is the number of data points.

Example: Consider the dataset 2, 4, 4, 4, 5.
The mean is (2+4+4+4+5)/5 = 19/5 = 3.8.
The variance would be $[(2-3.8)^2 + (4-3.8)^2 + (4-3.8)^2 + (4-3.8)^2 + (5-3.8)^2] / 5 = 1.36$.

Standard Deviation: The standard deviation is the square root of the variance and provides a more interpretable measure of dispersion.
Standard Deviation (σ) = $\sqrt{}$Variance

Example: Using the variance example above, the standard deviation would be $\sqrt{1.36} \approx 1.17$.

Measures of Dispersion

Range: The range is the simplest measure of dispersion and is calculated as the difference between the maximum and minimum values in a dataset.

Range = Maximum value - Minimum value

Example: Consider the following set of exam scores - 60, 65, 70, 75, 80.

The range would be 80 (maximum) - 60 (minimum) = 20.

Interquartile Range (IQR): Also called as **midspread** or **middle 50%,** This is the difference between the 75^{th} and 25^{th} percentiles or between the upper and lower quartiles. (the range of the middle 50% of a dataset).

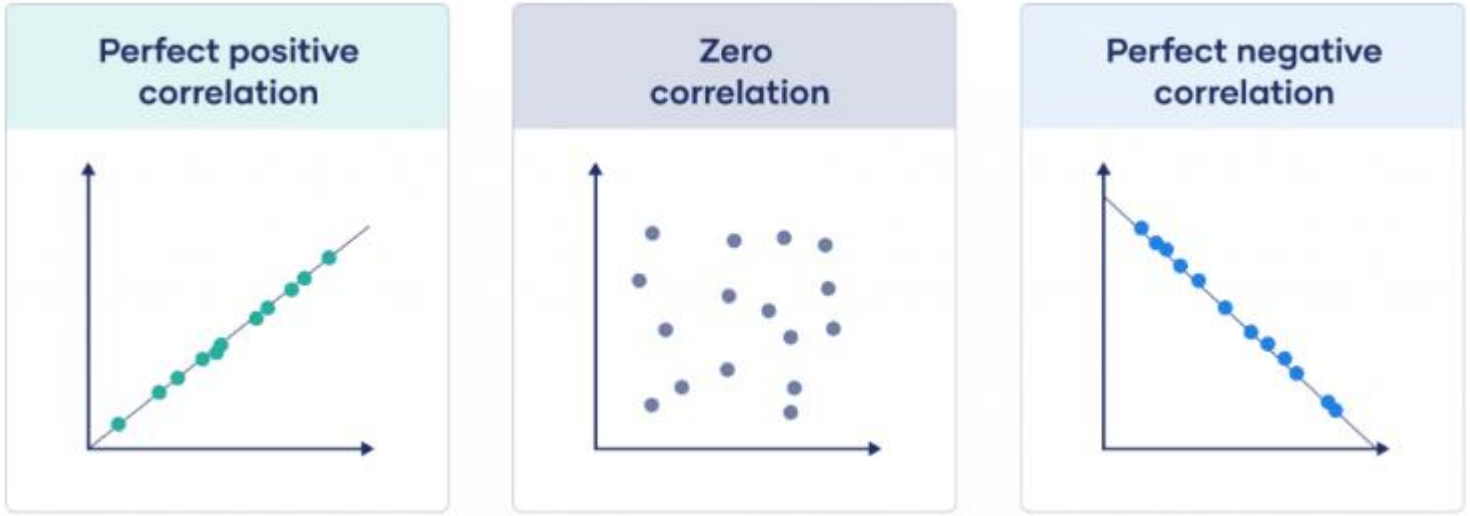
IQR = Q3 (third quartile) - Q1 (first quartile)

Example: If the dataset is 10, 15, 20, 25, 30, the first quartile (Q1) is 15, the third quartile (Q3) is 25, and the IQR would be 25 - 15 = 10.

Correlation

The correlation describes the statistical relationship between two variables:

- In a positive correlation, both variables move in the same direction.
- In a negative correlation, the variables move in opposite directions.
- In Zero correlation, the variables are not related.



Correlation

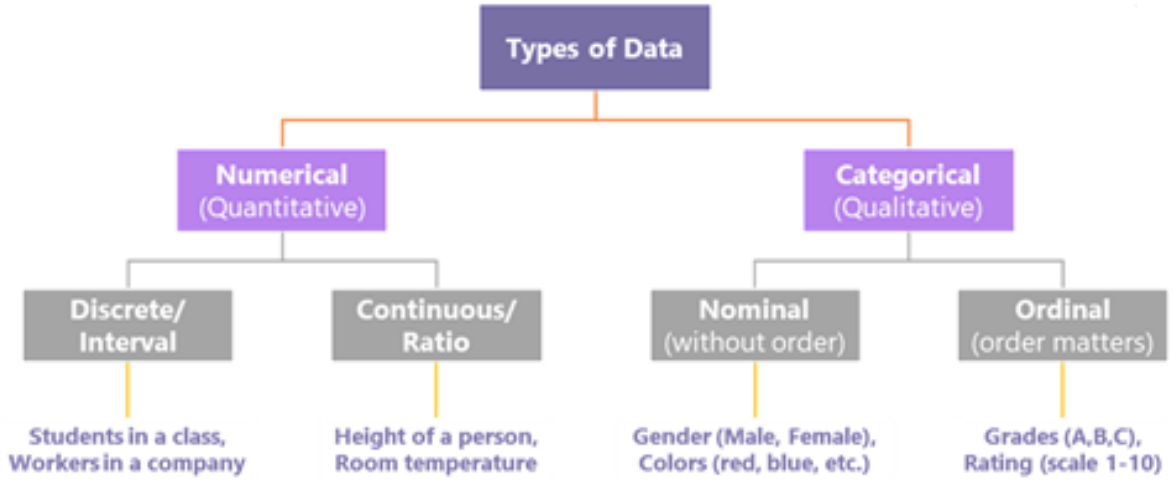
Example: Consider you want to find a decent apartment to rent that is not too expensive compared to other apartments you have found. The other apartments(all belonging to the same locality) you found on a website are priced as follows: $700, $850, $1,500 and $750 per month. Calculate some values statistical measures to help us make a decision:

Mean=$950, **Median=$800**, Standard Deviation=$322.10, Range= $800

A simple statistical analysis helped us to narrow down our choices.

Types of Data

- It is important to understand what kind of data you are dealing with so that you can select both the right statistical measure and the right visualization.
- We categorize data as **categorical/qualitative** and **numerical/quantitative**.
- **Categorical data** describes characteristics, for example, the color of an object or a persons gender.
- We can further divide the categorical data into nominal and ordinal data.
- **Numerical data** can be divided into discrete and continuous data
- Discrete data can have certain values, whereas continuous data can take any value(some times limited to a range)
- Another aspect to consider is whether the data has **temporal** domain(is it bounded to time or does it changes over time?) or spatial domain(if the data is bound to location)



Summary Statistics

In real-world applications, we often encounter enormous datasets. Therefore, summary statistics are used to summarize important aspects of data. They are necessary to communicate large amounts of information in a compact and simple way.

The following table gives an overview of which measure of central tendency is best suited to a particular type of data.

Data Type	Best measure of central tendency
Nominal	Mode
Ordinal	Median
Numerical	Mean/Median