Unit-1: Fundamentals of Data Analytics

1.2 Understanding the Data

1.2.1 Quantitative Data: Discrete and Continuous

1.2.2 Qualitative Data :Non-numerical (Normal and Ordinal)

→ **Quantitative data** is numerical and can be counted or measured, while qualitative data is descriptive and non-numerical, focusing on qualities and characteristics. **Quantitative data** answers "how much" or "how many," such as a person's height or the number of products sold, whereas qualitative data answers "why" or "how," such as a customer's feedback on a product or the texture of a fabric.

→ **Quantitative data:** Data that can be counted, measured, and expressed in numbers. It is objective and universal.

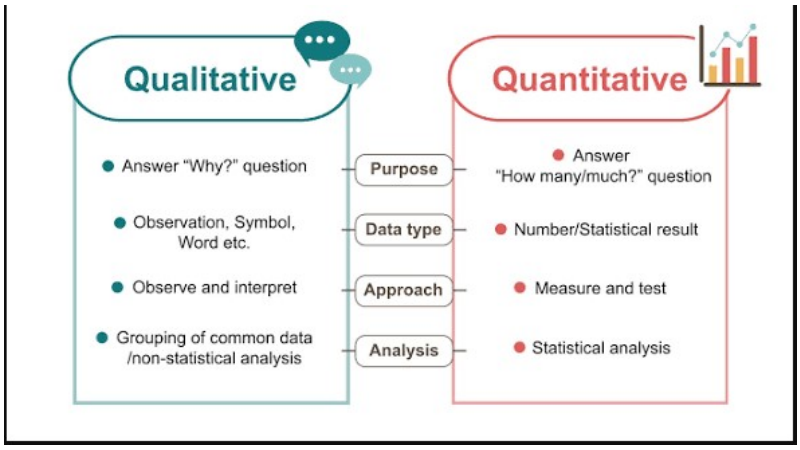
Examples:

- A customer's age or the number of products they purchased.
- The temperature of a room.
- A person's height or weight.
- Test scores in a school.
- The number of clicks on a website.

→ **Qualitative data:** Descriptive data that cannot be measured numerically, but can be observed or described based on qualities and characteristics. It is subjective and explores "why" and "how".

Examples:

- Customer feedback describing a product as "user-friendly" or "difficult to use".
- Observations about a person's appearance, like hair color or clothing.
- The smell of a flower or the taste of food.
- The reason a person signed up for an email newsletter, such as "to learn about local events".
- Descriptions from interviews, such as a patient discussing their quality of life after a new medication



Key Differences at a Glance

Aspect	Quantitative Data	Qualitative Data
Nature	Numerical, measurable	Descriptive, narrative, language-based
Focus	Quantity ("how much," "how many")	Quality ("why," "how")
Analysis	Statistical and mathematical	Thematic, content, or narrative analysis
Objectivity	More objective and conclusive	More subjective and open to interpretation
Collection Methods	Surveys (closed-ended), experiments	Interviews (open-ended), focus groups, observations

→ Quantitative Data: Discrete and Continuous

1. Discrete Data

Discrete data can only take on specific, distinct, and separate values. These values are typically whole numbers and cannot be meaningfully broken down into fractions or decimals. This type of data often arises from counting items.

Key Characteristics:

- Countable: You can count individual units.
- Fixed values: The data points have clear gaps between them.
- Usually integers: Values are generally whole numbers.

Examples:

- The number of students in a classroom (you can't have 22.5 students).
- The number of cars in a parking lot.
- The results of rolling a die (only values 1, 2, 3, 4, 5, or 6 are possible).
- The number of items purchased by a customer.

2. Continuous Data

Continuous data can take any value within a given range and can be infinitely broken down into smaller, finer parts, including fractions and decimals. This type of data often arises from measuring physical quantities.

Key Characteristics:

- Measurable: The data is obtained through measurement, not counting.
- Infinite possible values: There are an endless range of possible values between any two points.
- Can be divided: Values can be expressed with high precision (e.g., 1.573 or 4.5).

Examples:

- Height of a person (e.g., 175.5 cm, 175.55 cm, etc.).
- Weight of an object (e.g., 7.25 kg).

- Temperature of a room (e.g., 21.5°C).
- Time taken to complete a race (e.g., 10.45 seconds).

Feature	Discrete Data	Continuous Data
Type of values	Specific, fixed values	Any value in a range
Nature	Countable	Measurable
Examples	Number of books, Test Scores	Weight, distance, time, temperature
Can be fractional?	No	Yes

→ Qualitative Data: Non-Numerical (Nominal and Ordinal)

1. Nominal Data

Nominal data is used to label or categorize variables without any quantitative value or order. The data points fall into distinct, mutually exclusive categories. "Nominal" means "name," indicating that the data is merely used for naming or identifying categories.

Key Characteristics:

- Categories only: Data is divided into groups.
- No inherent order: The sequence of the categories doesn't matter.
- No mathematical operations: You cannot add, subtract, or average nominal data.

Examples:

- Gender: Male, Female, Non-binary.
- Hair Color: Black, Brown, Blonde, Red.
- Nationality: American, French, Japanese.
- Type of car: Sedan, SUV, Truck.
- Political affiliation: Democrat, Republican, Independent.

2. Ordinal Data

Ordinal data involves categories that have a meaningful, natural order or rank. While there is an order, the difference between the ranks is not necessarily equal or measurable in a precise numerical sense. When used qualitatively, it describes a scale of quality or satisfaction.

Key Characteristics:

- Ordered categories: Data can be ranked.
- Unequal intervals: The "distance" between ranks is not quantifiable.

Examples:

- Satisfaction levels: Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied. (We know "Very Satisfied" is better than "Satisfied," but we can't say exactly how much better).
- Education Level: High School Diploma, Bachelor's, Master's, Ph.D.
- Performance rating: Excellent, Good, Fair, Poor.
- Economic Status: Low Income, Middle Income, High Income.

Feature	Nominal Data	Ordinal Data
Order/Ranking	No	Yes

Type of values	Categories only	Ordered categories
Examples	Gender, Blood group, Colors	Satisfaction level, Education Ranks, Survey ratings
Measurement	Labeling & Counting	Ranking without measurable intervals

Common Use Cases of Qualitative Data

- Understanding **customer behavior**
- Analyzing **market trends**
- Studying **user experiences**
- Organizing **feedback and opinions** for improvement

Summary Chart

Data Type	**Sub-type**	**Characteristics**	**Examples**
Quantitative	Discrete	Countable, whole numbers	Number of pets or students
	Continuous	Measurable, includes decimals	Height, weight, temperature
Qualitative	Nominal	Categorical, no order	Gender, nationality, colors
	Ordinal	Categorical with order, no fixed scale	Satisfaction levels, class ranks

1.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is a crucial step in the data science process. It involves a set of techniques used to summarize, visualize, and understand the main characteristics of a dataset. The primary goal of EDA is to uncover patterns, spot anomalies, test hypotheses, and check assumptions with the help of statistical graphics and other data visualization methods. By performing EDA, data scientists can gain insights that guide feature engineering, model selection, and more effective data preparation.

Goal: To understand the data and prepare it for further analysis.

Objectives:

- Detecting patterns and trends in the data.
- Identifying missing values.
- Spotting outliers and anomalies.
- Choosing the right analysis techniques.
- Summarizing data through statistics and visualizations.

1.1.1 Types of Exploratory Data Analysis

EDA is not a single technique but a collection of methods that can be broadly classified based on the number of variables being analyzed at once.

- **Univariate Analysis**: The analysis of a single variable.
- **Bivariate Analysis**: The analysis of the relationship between two variables.
- **Multivariate Analysis**: The analysis of the relationship between three or more variables.

1.1.2 Univariate Analysis

Univariate analysis is the simplest form of EDA, focusing on one variable at a time. The purpose is to describe the variable's central tendency (e.g., mean, median), spread (e.g., standard deviation, range), and distribution.

Methods for Univariate Analysis:

→ For Quantitative Variables (e.g., age, height):

Descriptive Statistics: Calculating the mean, median, mode, standard deviation, and variance.

Visualization: Using histograms, box plots, and density plots to show the distribution, identify

skewness, and detect potential outliers.

→ For Categorical Variables (e.g., gender, city):

Frequency Tables: Counting the occurrences of each category.

Visualization: Using bar charts or pie charts to display the proportion of each category.

1.1.3 Bivariate Analysis

Bivariate analysis explores the relationship between two variables. This helps to identify correlations, associations, and dependencies.

Methods for Bivariate Analysis:

→ For Two Quantitative Variables:

Scatter Plots: The most common tool to visualize the relationship. The pattern of the points can reveal a positive, negative, or no correlation.

Correlation Coefficient: A numerical value (like Pearson's r) that quantifies the strength and direction of a linear relationship between the variables.

→ For Two Categorical Variables:

Cross-Tabulation (Contingency Tables): Shows the frequency distribution of the variables in a table format.

Stacked Bar Charts: A visualization that shows the frequency of each category of one variable broken down by the categories of another.

For One Quantitative and One Categorical Variable:

Box Plots: A box plot for each category can show how the quantitative variable's distribution differs across the groups.

T-tests or ANOVA: Statistical tests to determine if the mean of the quantitative variable is significantly different across the categories.

1.1.4 Multivariate Analysis

Multivariate analysis examines the relationships among three or more variables simultaneously. This is essential for understanding complex interactions in a dataset.

Methods for Multivariate Analysis:

Pair Plots: A grid of scatter plots showing every possible bivariate relationship between all variables in the dataset.

Heat maps: A visual representation of a matrix, typically used to show the correlation matrix of all numerical variables. Colors are used to indicate the strength of the correlation.

3-D Scatter Plots: A visualization tool for showing the relationship between three variables.

Principal Component Analysis (PCA): A dimensionality reduction technique that can help

visualize high-dimensional data in 2D or 3D space while preserving most of its variance.

1.1.5 Handling Missing Data and Outliers

An important part of EDA is identifying and addressing common data quality issues like missing data and outliers, as they can significantly impact model performance.

Handling Missing Data:

Identification: Use functions to count the number of missing values (null or NaN) per variable.

Strategies:

- Deletion: Remove rows or columns with a high percentage of missing values.
- Imputation: Fill in missing values using a replacement strategy. Common methods include:
- Replacing with the mean, median or mode of the variable.
- Using a predictive model (e.g., k-Nearest Neighbors) to estimate the missing values.

Handling Outliers:

Identification: Outliers are data points that are significantly different from other observations. They can be identified using:

- Box Plots: Points beyond the "whiskers" of the box plot are potential outliers.
- Z-score: Measures how many standard deviations a data point is from the mean. A Z-score greater than 3 or less than -3 is often considered an outlier.

Strategies:

- Deletion: Remove the outlier if it's clearly a data entry error.
- Transformation: Apply a log or square root transformation to the variable to reduce the impact of the outlier.
- Keep them: Sometimes, outliers are meaningful and should be kept, especially in fields like fraud detection. The decision depends on the context.

Summary Table

Step	**Purpose**	**Tools/Techniques**
Univariate	Understand one variable	Histograms, Bar charts, Summary stats
Bivariate	Explore relationship between two	Scatter plots, Correlation, Box plots
Multivariate	Understand complex relationships	Heat maps, PCA, Multivariate regression

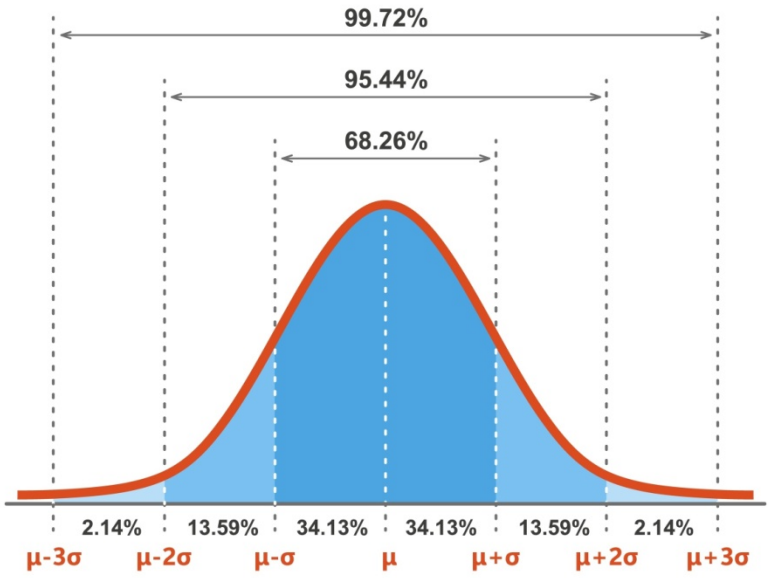
Missing Data	Handle gaps in data	Imputation, Deletion, Prediction
Outliers	Detect and treat unusual values	Box plot, Z-score, IQR

1.3 Spread of Data:

When we collect data (like the heights of students in a class or test scores), the numbers aren't all the same. They "spread out." Understanding *how* they spread out is a big part of statistics. This guide will walk you through the most common shapes this spread can take.

1.3.1 The Normal Distribution (The "Bell Curve")

The Normal Distribution also known as the Gaussian distribution or bell curve is the most famous and common shape for data. It's also called a "Bell Curve" because it looks like a bell.



Key Features:

- **Symmetrical:** The left side is a perfect mirror image of the right side.

- **Center:** The most frequent data points are all in the middle. In a perfect normal distribution, the **Mean (average)**, **Median (middle value)**, and **Mode (most frequent value)** are all the same.
- **Tails:** The data becomes less and less frequent as you move away from the center, forming two "tails" that get closer and closer to zero.

Simple Example: Imagine measuring the height of all adult men in a large city. Most men will be around the average height. A few will be very tall, and a few will be very short. If you plot this data, it will look very much like a normal distribution.

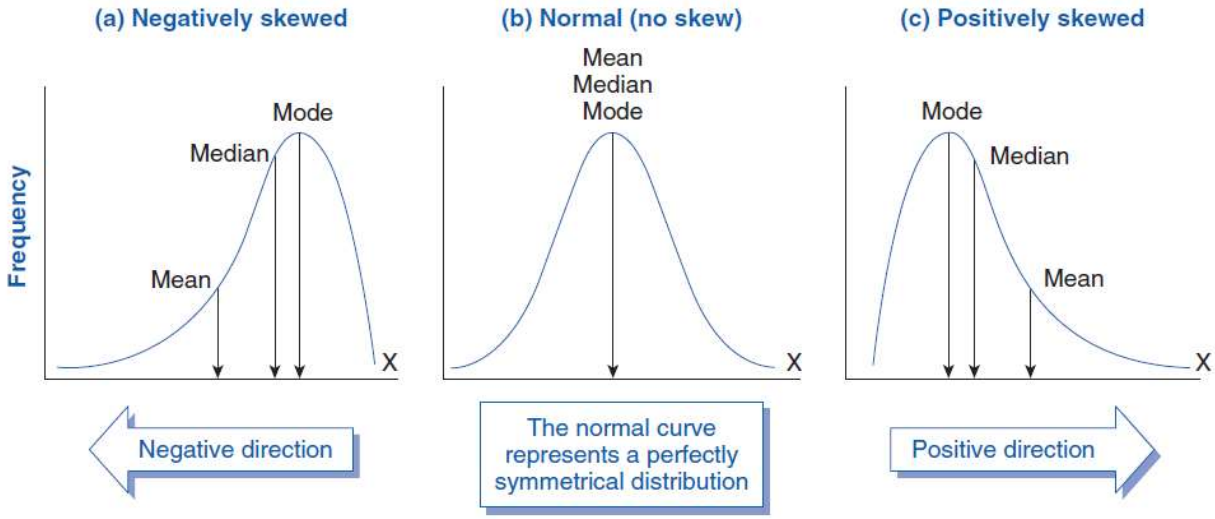
1.3.2 Skewed Distribution (The "Lopsided" Curve)

A skewed distribution is **not symmetrical**. It's lopsided because a long "tail" of less frequent, extreme values is pulling it to one side.

There are two types of skew:

1. Positive Skew (or "Right-Skewed")

- **What it is:** The "tail" of low-frequency, high values is on the **right**.
- **Looks like:** A "hump" of data on the left, with a long tail stretching out to the right.
- **How it happens:** Most data points are low, but there are a few very high values (outliers) that pull the mean (average) to the right.
- **Rule:** Mean > Median > Mode



Simple Example: Household income. Most households earn an average income (the hump on the left), but a few households earn *extremely* high incomes. These few high values create a long tail to the right.

2. Negative Skew (or "Left-Skewed")

- **What it is:** The "tail" of low-frequency, low values are on the **left**.
- **Looks like:** A "hump" of data on the right, with a long tail stretching out to the left.
- **How it happens:** Most data points are high, but there are a few very low values (outliers) that pull the mean (average) to the left.
- **Rule:** Mean < Median < Mode

Simple Example: Scores on a very easy test. Most students will get high scores (the hump on the right), but a few students who didn't study will get very low scores, creating a tail to the left.

1.3.3 Skewness and Kurtosis (The "Measurements")

"Skewness" and "Kurtosis" are just numbers that *measure* the shape of your data's distribution.

Skewness: Measures Asymmetry

Skewness is a number that tells you *how much* and in *which direction* your data is skewed.

- **Skewness = 0:** Perfectly symmetrical (like a Normal Distribution).
- **Skewness > 0 (Positive):** The data is **Positively Skewed** (tail to the right).
- **Skewness < 0 (Negative):** The data is **Negatively Skewed** (tail to the left).

The further the number is from 0, the more skewed the data is.

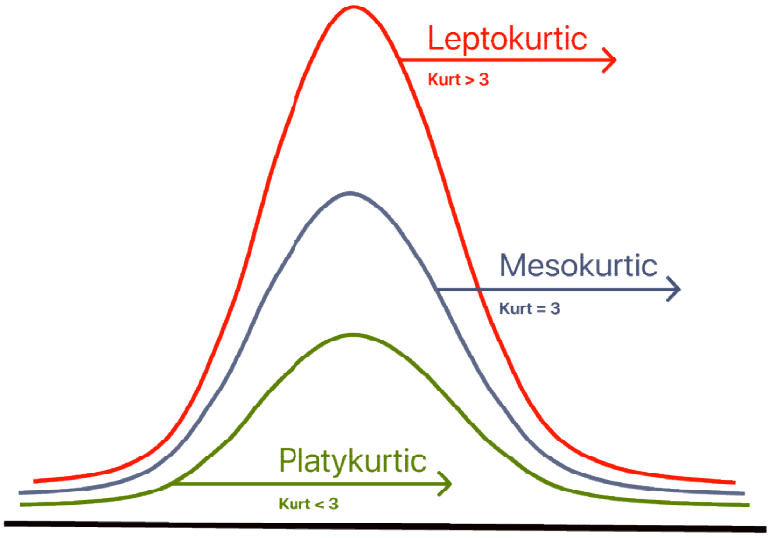
Kurtosis: Measures "Peakiness" or "Tailedness"

Kurtosis is a number that tells you how "peaky" your distribution is and how "fat" its tails are, compared to a normal distribution.

[Image comparing kurtosis types (leptokurtic, mesokurtic, platykurtic)]

- **Leptokurtic (High Kurtosis):**
  - **Looks like:** A sharp, skinny peak and "fat" tails.
  - **What it means:** There are more extreme values (outliers) in the tails than in a normal distribution.
- **Platykurtic (Low Kurtosis):**
  - **Looks like:** A flat, wide peak and "thin" tails.
  - **What it means:** There are fewer extreme values (outliers) in the tails than in a normal distribution.
- **Mesokurtic (Medium Kurtosis):**
  - **Looks like:** A normal "bell curve."

- **What it means:** The peak and tails are "normal." The kurtosis of a normal distribution is used as the baseline (often set to 0 or 3, depending on the formula).



In short: Skewness measures side-to-side lopsidedness, and Kurtosis measures the peakiness and tail weight.

Summary Table

Concept	Description	Shape /Example
Normal Distribution	Symmetrical bell-shaped curve; mean = median = mode	Heights, IQ scores
Right Skewed Distribution	Tail to the right; mean > median	Income, waiting times
Left Skewed Distribution	Tail to the left; mean < median	Retirement age, easy exams
Skewness	Measures asymmetry of distribution	0=symmetric; >0=right skew; <0= left skew
Kurtosis	Measures tail weight and peak sharpness	>0=sharp peak;<0=flat peak

Key Takeaways

- **Normal distribution** serves as the foundation for statistical analysis.
- **Skewness** and **kurtosis** quantify **deviations from normality**.
- Real-world data often deviates due to external factors—recognizing these deviations helps improve data modeling accuracy.