Experiment 2

Aim: Explore the descriptive statistics on the given dataset.

I-OBJECTIVE

- **To understand basic concepts of Statistics**
- **To Explore the descriptive statistics on the given dataset**

II-THEORY

1. Difference between Mathematics and Statistics.

Aspect	Mathematics	Statistics
Fundamental Nature	The study of quantity, structure, space, and change. It is built on absolute certainty and rigorous proof.	The science of learning from data. It deals specifically with uncertainty, variability, and risk.
Logic & Reasoning	Deductive: Starts with axioms (established truths) and follows strict logic to reach a certain conclusion.	Inductive: Starts with observations (data) and works backward to find patterns or generalize about a population.
Handling Uncertainty	Generally seeks to eliminate uncertainty through precise formulas and equations.	Embraces uncertainty; it measures and quantifies "how likely" an event is to be true.
Core Objectives	To discover and define universal truths that exist regardless of real-world application.	To transform raw information into actionable insights and decision-making tools.
Context	Often context-free. 2 + 2 = 4 is true whether you are counting apples or planets.	Highly context-dependent. Numbers mean nothing in statistics without knowing the source and the "why" behind the data.
Primary Tools	Calculus, Algebra, Geometry, Topology, and Number Theory.	Probability distributions, Hypothesis testing, Regression, and Sampling.

2. Different types of data.

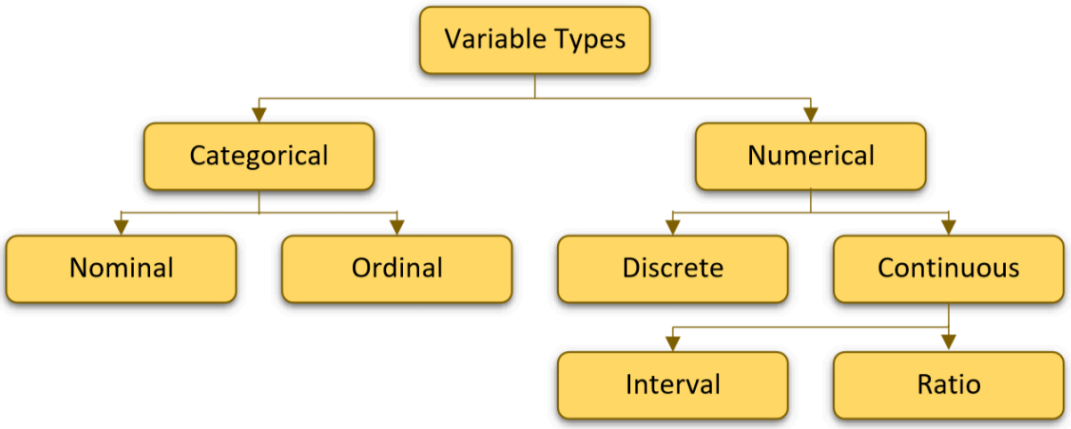

Fig 1: Different types of data

1. Qualitative Data (Categorical)

Qualitative data describes qualities, attributes, or characteristics. It is used to label or categorize observations without using meaningful numerical values.

A. Nominal Data

Nominal data represents discrete units with no inherent order or ranking. It simply names a category.

- **Characteristics:** No "greater than" or "less than" relationship.
- **Examples:** Gender, Marital Status, Blood Type, or Country of Origin.
- **Analysis & Visualization:** Usually summarized with **Frequency Tables** or **Bar Charts**. Common statistical tests include the **Chi-Square Test** and **Fisher's Exact Test**.

B. Ordinal Data

Ordinal data consists of categories that follow a clear natural order or rank; however, the mathematical distance between these ranks is not defined or equal.

- **Characteristics:** You can say one value is "higher" than another, but you cannot calculate the exact difference.
- **Examples:** Education Level (High School < Bachelor's < Master's), Customer Satisfaction (Poor, Fair, Good, Excellent), or Job Seniority.
- **Analysis & Visualization:** Represented using **Bar Charts** or **Pie Charts**. Analyzed using non-parametric tests like the **Mann-Whitney U Test** or **Wilcoxon Signed-Rank Test**.

2. Quantitative Data (Numerical)

Quantitative data represents measurable quantities and is expressed in numbers. It allows for mathematical operations like addition, subtraction, and averaging.

A. Discrete Data

Discrete data consists of distinct, countable values. These are typically "whole numbers" where there are no possible values between two points.

- **Characteristics:** Values are "counted" rather than "measured."
- **Examples:** The number of employees in a department, the number of cars in a parking lot, or the number of children in a family.

B. Continuous Data

Continuous data represents measurements and can take any value within a given range (including fractions and decimals).

- **Characteristics:** Values are "measured" and can be infinitely subdivided.
- **Examples:** Height, Weight, Temperature, or Salary (when considering cents).
- **Sub-types:** Continuous data is often further divided into **Interval** (no true zero, like Celsius) and **Ratio** (has a true zero, like Weight).

3. Phases of Descriptive Analysis

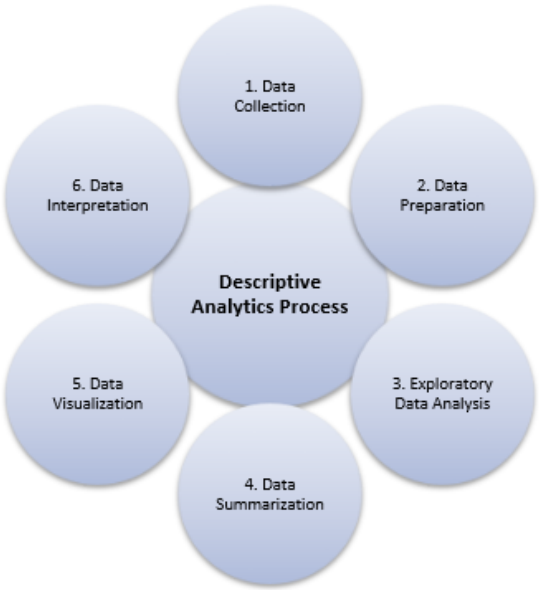

Fig 2: Phases of Descriptive Analysis

1. Data Collection

This foundational phase involves gathering raw data from diverse and relevant sources.

- **Sources:** Databases, APIs, web scraping, surveys, IoT sensors, or social media feeds.
- **Goal:** To assemble a comprehensive dataset that aligns with the specific objectives of the research or business problem.
- **Example:** Exporting a year’s worth of transaction logs from an e-commerce SQL database.

2. Data Preparation

Raw data is rarely "analysis-ready." This step focuses on refining the data to ensure quality and consistency.

- **Key Actions:** Handling missing values, removing duplicates, normalizing formats (e.g., date styles), and detecting outliers.
- **Significance:** Prevents "garbage in, garbage out" by ensuring the analysis is built on a reliable foundation.

3. Exploratory Data Analysis

In this stage, statistical techniques are used to explore the dataset's characteristics and uncover hidden patterns.

- **Techniques:** Calculating measures of central tendency (mean, median, mode), dispersion (variance, standard deviation), and identifying correlations.
- **Goal:** To understand the distribution of data and pinpoint significant variables.

4. Data Summarization

Once analyzed, the complex data points are condensed into a structured and logical summary.

- **Key Actions:** Organizing insights into categories, creating pivot tables, and drafting executive summaries.
- **Goal:** To provide stakeholders with a high-level view of the most important findings without overwhelming them with technical details.

5. Data Visualization

This step translates numerical findings into a visual language that is easier for the human brain to process.

- **Tools:** Bar charts for comparisons, line graphs for trends, histograms for distributions, and interactive dashboards.
- **Impact:** Highlights trends and anomalies that might remain hidden in a standard spreadsheet.

6. Data Interpretation

Data without context is just numbers. This phase assigns meaning to the results based on real-world factors.

- **Key Actions:** Linking data trends to business goals, identifying the "why" behind the numbers, and drawing logical conclusions.
- **Example:** Determining that a dip in user engagement was caused by a specific website UI update rather than market competition.

4. Classification of Descriptive Analysis.

Descriptive analysis is mainly classified into:

1. Measures of Central Tendency
2. Measures of Dispersion or Variability

1. Measures of Central Tendency: Central tendency represents the "center" or "typical" value of a dataset. It aims to summarize a large group of numbers with a single, representative figure.

a) Mean:

The mean is the most common measure, calculated by summing all observations and dividing by the total count.

- **Formula:**

$$\textit{Mean} = \sum x/n$$

Equation 2.1: Mean

- ***Variables:***
  - *x = Each individual value in the dataset*
  - *n = Total number of observations*
- **Best Use:** For symmetric data without extreme values.
- **Limitation:** It is highly sensitive to **outliers**. A single massive value can "pull" the mean away from the true center.

b) Median

The median is the literal "middle" point of a dataset when sorted in order.

- **Calculation:**
  - **Odd number of values:** The exact middle number.
  - **Even number of values:** The average of the two middle numbers.
- **Best Use:** For skewed distributions (like salaries or housing prices) because it ignores extreme outliers.

c) Mode

The mode is the value that appears most often in a dataset.

- **Characteristics:** It is the only measure of central tendency applicable to **Qualitative (Categorical) Data**.
- **Variations:** A dataset can be **Unimodal** (one mode), **Bimodal** (two), or **Multimodal** (multiple).

2. Measures of Dispersion (Variability)

While central tendency finds the center, dispersion measures how "spread out" the data is around that center. Two datasets could have the same mean but look completely different based on their spread.

a) Range

The simplest measure of spread, representing the distance between the two extremes.

- **Formula:**

$$\textit{Range} = \textit{Maximum value} - \textit{Minimum value}$$

Equation 2.2 : Range

- **Limitation:** It only considers the two most extreme points and ignores how the rest of the data is distributed in between.

b) Variance

Variance quantifies how much the data points fluctuate around the mean by looking at the average of the squared differences.

- **Formula:**

$$\textit{Variance} = \Sigma(x - \bar{x})^2 / n$$

Equation 2.3 : Variance

- ***Variables:***
  - x = *Each individual value in the dataset*
  - $\bar{x}$ = *Mean (average) of all values*
  - $(x - \bar{x})$ = *Deviation of each value from the mean*
  - $(x - \bar{x})^2$ = *Squared deviation*
  - n = *Total number of observations*
- **Concept:** Because it squares the differences, it penalizes larger deviations more heavily. However, its units are also squared (e.g., "dollars squared"), making it hard to interpret directly.

c) Standard Deviation (SD)

The standard deviation is the square root of the variance, bringing the measure back to the original units of the data.

- **Formula:**

$$\textit{Standard Deviation} = \sqrt{\textit{Variance}}$$

Equation 2.4 : Standard Deviation

- **Best Use:** This is the gold standard for measuring variability.
- **Interpretation:** A **low SD** indicates data points are clustered tightly around the mean; a **high SD** indicates the data is spread widely across a large range of values.

DATASET USED - Employee Attrition Dataset

The **Employee Attrition Dataset** contains comprehensive information regarding employee demographics, job roles, financial compensation, and satisfaction levels within an organization.

- **Dataset Size**

  Number of Rows: 1470

  Number of Columns: 31

- **Identification & Demographics:** It includes individual attributes such as **Age**, **Gender**, **Marital Status**, and **Education**, providing a baseline profile of the workforce.
- **Central Target Variable: Attrition** is the primary variable of interest, indicating whether an employee has left the company (Yes/No).
- **Job & Role Indicators:** Attributes like **Department**, **Job Role**, **Job Level**, and **Business Travel** help evaluate how different positions and work requirements impact retention.
- **Satisfaction & Work-Life Factors:** The dataset captures employee sentiment through **Job Satisfaction**, **Environment Satisfaction**, **Relationship Satisfaction**, and **Work-Life Balance** scores.
- **Financial & Incentive Metrics:** Economic factors such as **Monthly Income**, **Hourly Rate**, **Percent Salary Hike**, and **Stock Option Level** are included to measure the correlation between compensation and loyalty.
- **Tenure & Experience:** Historical career data, including **Total Working Years**, **Years at Company**, **Years in Current Role**, and **Years with Current Manager**, helps track how career progression and management stability influence the decision to stay.

III-IMPLEMENTATION

Python Code:

```
import numpy as np
import pandas as pd

df = pd.read_csv("empattds.csv")

df_numeric = df.select_dtypes(include=[np.number])

results = []
for col in df_numeric.columns:
    modes = df_numeric[col].mode()
    mode_val = modes[0] if not modes.empty else np.nan
```

```
    stats = {
        "Column": col,
        "Mean": df_numeric[col].mean(),
        "Median": df_numeric[col].median(),
        "Mode": mode_val,
        "Variance": df_numeric[col].var(),
        "Standard Deviation": df_numeric[col].std(),
        "Range": df_numeric[col].max() - df_numeric[col].min(),
    }
    results.append(stats)

results_df = pd.DataFrame(results)

numeric_cols_stats = [
    "Mean",
    "Median",
    "Mode",
    "Variance",
    "Standard Deviation",
    "Range",
]
results_df[numeric_cols_stats] = results_df[numeric_cols_stats].round(2)

print(results_df)

results_df.to_csv("empattds_formatted_stats.csv", index=False)
```

```
(venv) [aditya@archlinux ads2]$ python main.py
                      Column      Mean   Median  Mode     Variance  Standard Deviation  Range
0                        Age     36.92     36.0    35        83.46                9.14     42
1                  DailyRate    802.49    802.0   691    162819.59              403.51   1397
2           DistanceFromHome      9.19      7.0     2        65.72                8.11     28
3                 HourlyRate     65.89     66.0    66       413.29               20.33     70
4              MonthlyIncome   6502.93   4919.0  2342  22164857.07             4707.96  18990
5                MonthlyRate  14313.10  14235.5  4223  50662878.17             7117.79  24905
6         NumCompaniesWorked      2.69      2.0     1         6.24                2.50      9
7          PercentSalaryHike     15.21     14.0    11        13.40                3.66     14
8           StockOptionLevel      0.79      1.0     0         0.73                0.85      3
9          TotalWorkingYears     11.28     10.0    10        60.54                7.78     40
10    TrainingTimesLastYear      2.80      3.0     2         1.66                1.29      6
11            YearsAtCompany      7.01      5.0     5        37.53                6.13     40
12       YearsInCurrentRole      4.23      3.0     2        13.13                3.62     18
13  YearsSinceLastPromotion      2.19      1.0     0        10.38                3.22     15
14     YearsWithCurrManager      4.12      3.0     2        12.73                3.57     17
(venv) [aditya@archlinux ads2]$
```

Fig 3: Output using Python

Insights:

The data reveals a mature workforce with an **average age of 37**, though the wide **range (42 years)** indicates a multi-generational environment spanning from entry-level youths to seasoned professionals.

While the **average tenure is 7 years**, the **median (5.0)** suggests that a significant portion of the staff is relatively new or in the early stages of their career cycle within the company.

1		Age	DailyRate	DistanceFrom	HourlyRate	MonthlyIncom	MonthlyRate	NumCompani	PercentSalary	StockOptionL	TotalWorking	TrainingTimes	YearsAtComp	YearsInCurre	YearsSinceLa	YearsWithCurrManag
38		50	869	3	86	2683	3810	1	14	0	3	2	3	2	0	2
39		35	890	2	97	2014	9687	1	13	0	2	3	2	2	2	2
40		36	852	5	82	3419	13072	9	14	1	6	3	1	1	0	0
41		33	1141	1	42	5376	3193	2	19	2	10	3	5	3	1	3
42		35	464	4	75	1951	10910	1	12	1	1	3	1	0	0	0
43		27	1240	2	33	2341	19715	1	13	1	1	6	1	0	0	0
44		26	1357	25	48	2293	10558	1	12	0	1	2	1	0	0	1
45		27	994	8	37	8726	2975	1	15	0	9	0	9	8	1	7
46		30	721	1	58	4011	10781	1	23	0	12	2	12	8	3	7
47		41	1360	12	49	19545	16280	1	12	0	23	0	22	15	15	8
48		34	1065	23	72	4568	10034	0	20	0	10	2	9	5	8	7
49		37	408	19	73	3022	10227	4	21	0	8	1	1	0	0	0
50		46	1211	5	98	5772	20445	4	21	0	14	4	9	6	0	8
1472	Mean	36.92380952	802.4857143	9.192517007	65.89115646	6502.931293	14313.1034	2.693197279	15.20952381	0.793877551	11.27959184	2.799319728	7.008163265	4.229251701	2.187755102	4.123129252
1473	Median	36	802	7	66	4919	14235.5	2	14	1	10	3	5	3	1	3
1474	Mode	35	691	2	66	2342	9150	1	11	0	10	2	5	2	0	2
1475	SD	9.135373489	403.5090999	8.106864436	20.32942759	4707.956783	7117.786044	2.498009006	3.659937717	0.852076667	7.780781676	1.289270621	6.126525152	3.623137035	3.222430279	3.568136121
1476	Variance	70.6335034	163299.7083	67.79676871	406.6156463	21254023.01	42048370.88	5.958333333	15.56972789	0.579931972	46.3494898	2.24829932	32.76445578	12.49659864	8.284863946	13.67687075

Fig 4: Output using Excel

Insights:

The excel data shows a stable organizational profile (avg. age 37) with solid tenure (avg. 7 years), but heavily impacted by income inequality, as seen in the massive variance of Monthly Income. While training engagement is high (avg. 2.8), the huge range in Distance from Home (28 miles) and the gap between mean (6,502) and median (4,919) pay highlights a sharp divide in the employee experience. This suggests that financial stratification and commute stress are the primary drivers of attrition risk.

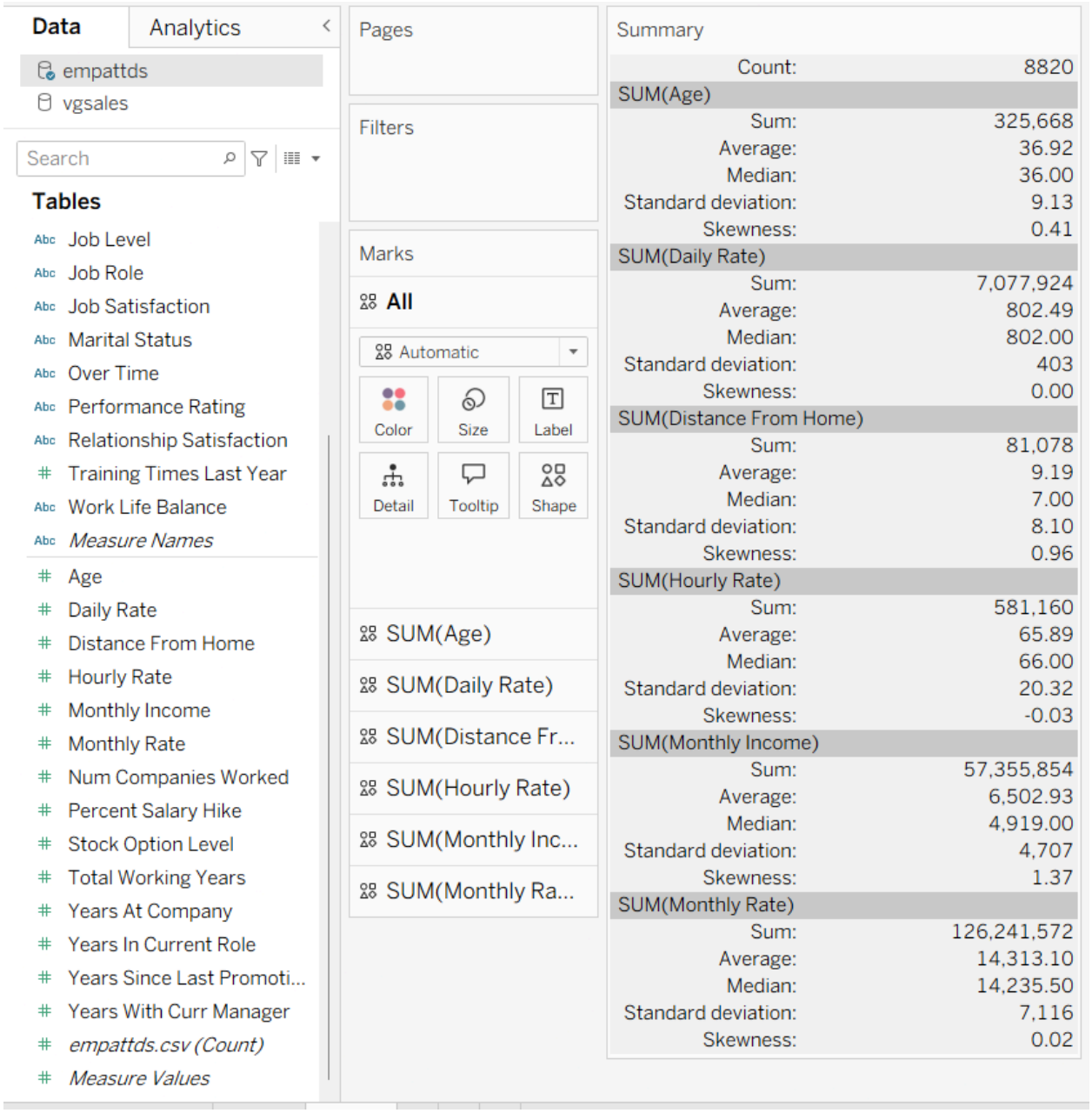

Fig 6: Output using Tableau

Insights:

Using Tableau, we understand that within the organization, while the average employee tenure is 7 years, the data reveals a profound **"compensation-commute" divide**. Huge variances in **Monthly Income** and **Distance from Home** show that the employee experience is heavily skewed, with extreme internal economic inequality directly impacting retention potential.

IV- CONCLUSION

Descriptive statistics serve as the essential foundation for data analysis by transforming raw, unorganized employee data into a meaningful and interpretable summary. By utilizing measures of central tendency—such as the mean monthly income of $6,503 and the median of $4,919—and measures of dispersion like the monthly income variance of $22,164,857, we can pinpoint the "typical" compensation of an employee and understand how much financial reward fluctuates. This phase of analysis is crucial for identifying patterns, such as the right-skewed distribution of $1.37, which reveals that a select group of high-level earners pulls the average upward while the majority of the workforce earns below the mean.

A multi-tool descriptive analysis was performed using Python, Excel, and Tableau to explore the organizational landscape of employee attrition. Through Python, the calculation of statistical metrics was automated across thirty-one key demographic and job-related variables, while Excel facilitated a granular examination of specific organizational outliers, notably the impact of high-tenure veterans, with one individual achieving a record-breaking $40 years at the company. Furthermore, Tableau provided a high-level visual summary that revealed a near-perfect symmetry in daily rates (skewness of $0), confirming that while pay is stratified, work-value assessment is distributed evenly across the workforce. Collectively, these tools demonstrated that while the organization sees a common tenure around $5 years (the mode), long-term retention remains concentrated in a small fraction of loyal, high-earning elite roles.

VI- POST LAB QUESTION/ANSWER

1. Differentiate between Descriptive and Inferential Statistics.

Descriptive Statistics	Inferential Statistics
Describes and summarizes raw data in a meaningful way.	Draws conclusions and makes inferences about the population using sample data.
Helps in organizing, analyzing, and presenting data clearly.	Helps in comparing data, making predictions, and testing hypotheses.
Focuses on describing a given dataset or situation.	Focuses on explaining the probability or chance of an event occurring.
Limited to the data available (sample or population).	Extends results from the sample to the entire population.
Deals with already known and observed data.	Attempts to reach conclusions about unknown population characteristics.
Examples: mean, median, mode, range, variance, charts, graphs.	Examples: confidence intervals, hypothesis testing, regression, p-values.
Used mainly for summarizing trends and patterns.	Used for predicting trends and generalizing results.
Achieved through tables, charts, and graphical representation.	Achieved using probability theory and statistical models.

2. Explain the types of Variables in Statistics.

In statistics, a variable is any characteristic or attribute that can take different values among individuals, objects, or observations. Variables are mainly classified into two broad types:

1. Qualitative Variables (Categorical)

Qualitative variables describe the "qualities" or attributes of an observation. They represent data that can be sorted into groups but cannot be measured numerically in a meaningful way.

- **Nominal Variables:** These represent discrete categories with no inherent rank or order. One category is not "higher" or "lower" than another; they are simply different.
    - *Examples:* Eye color, marital status, or department name (HR vs. Sales).
- **Ordinal Variables:** These categories follow a natural progression or ranking system. While the order matters, the mathematical distance between the categories is inconsistent or unknown.

    - *Examples:* Economic status (Low, Middle, High), survey responses (Agree, Neutral, Disagree), or military rank.

2. Quantitative Variables (Numerical)

Quantitative variables represent "quantities" and are expressed in numbers. These variables allow for mathematical operations like addition or averaging.

- **Discrete Variables:** These consist of distinct, separate values that are typically obtained by **counting**. You cannot have a fraction of a discrete unit.
    - *Examples:* The number of children in a family, the number of defects in a product, or goals scored in a match.
- **Continuous Variables:** These represent measurements that can take on any value within a given range, including decimals and fractions. They are typically obtained by **measuring**.
    - *Examples:* Exact weight, atmospheric temperature, or the time it takes to complete a task.