"Univariate, Bivariate, and Multivariate Data Visualizations: A Detailed Exploration"

Project Report submitted in Partial fulfillment of the requirement for the

Award of the degree of

MASTER OF BUSINESS ADMINISTRATION

of

BANGALORE UNIVERSITY

By

R Harshavardhan

Reg. No: P03EH23M015018

Under the guidance of

Dr. Arpit Yadav

IFIM COLLEGE (AUTONOMOUS)

8P & 9P, KIADB Industrial Area,

Electronics City Phase I, Bengaluru -

560100

DATA VISUALIZATION ASSIGNMENT

What is Data Visualization?

Data visualization is the graphical representation of information and data. By using visual elements like charts, graphs, and maps, it transforms complex datasets into easily understandable formats. Think of it as turning raw numbers into compelling stories that can be quickly grasped and acted upon.

Impact of Data Visualization in Business:

Data visualization has become a powerful tool in the business world, transforming how companies operate and make decisions. Here's how:

Improved Decision-Making:

Visualizing data helps identify trends, patterns, and outliers that might be missed in raw data tables. This allows businesses to make more informed, data-driven decisions, leading to better outcomes.

Enhanced Communication:

Complex data can be difficult to convey through text alone. Visualizations simplify complex information, making it easier for everyone, regardless of their technical expertise, to understand and engage with data. This improves communication across departments and stakeholders.

Increased Efficiency:

Interactive dashboards powered by data visualization tools allow businesses to track key performance indicators (KPIs) in real time. This enables quicker identification of issues, allowing for prompt problem-solving and resource optimization.

Better Storytelling:

Data visualizations can transform dry statistics into compelling narratives. By showcasing insights in a visually appealing way, businesses can engage audiences, build stronger arguments, and influence decision-making.

Importance of Data Visualization:

In today's data-driven world, data visualization is no longer a luxury, it's a necessity.

Clarity and Understanding: It converts complex data into easily digestible visuals, making it accessible to a wider audience.

Pattern Recognition: Visual representations highlight patterns and trends that are often hidden in raw data.

Data Exploration: Interactive visualizations allow for exploration and analysis, revealing insights that might have gone unnoticed.

Enhanced Memory and Recall: Visuals are processed by the brain faster and more effectively than text, leading to better information retention.

Improved Collaboration: Visualizations facilitate data sharing and discussion, promoting collaboration and cross-functional understanding.

In conclusion, data visualization is a powerful tool that empowers businesses to unlock the hidden potential of their data. By transforming information into compelling visual stories, it drives better decision-making, improves communication, increases efficiency, and ultimately leads to better business outcomes.

Introduction to Data Visualisation Types

Data visualization is the process of creating graphical representations of data to help people understand and make decisions from complex information. With the massive growth of data in various fields, data visualization has become an essential tool for businesses, scientists, researchers, and analysts to communicate their findings effectively.

There are several types of data visualization, each designed to help convey different types of data insights. Understanding these types is crucial for selecting the most suitable visualization for your data.

Understanding Data Visualization Types: Univariate, Bivariate, and Multivariate

Data visualization helps us understand complex datasets by representing them visually. This makes patterns, trends, and outliers easier to identify.

Here's a breakdown of the three main types:

1. Univariate Analysis

Univariate analysis is a statistical method that examines a single variable at a time to understand its characteristics, distribution, and patterns. It provides a concise summary of the data and helps identify potential outliers, trends, and central tendencies.

Definition: Univariate analysis focuses on analysing a single variable. It explores the distribution, frequency, and central tendency of that variable alone.

Impact:

- Understanding the characteristics of a single variable.
- Identifying outliers and potential errors in data.
- Summarizing data using descriptive statistics like mean, median, and standard deviation.

Uses of Univariate Analysis

1. Understanding Data Distribution

- **Purpose:** Provides insights into how values are spread within a dataset.
- **Example:** Histograms or frequency tables can reveal if data follows a normal, skewed, or uniform distribution.

2. Detecting Outliers

- **Purpose:** Identifies unusual or extreme values that may affect analysis.
- **Example:** Boxplots highlight data points that fall outside the interquartile range (IQR).

3. Measuring Central Tendency

- **Purpose:** Summarizes data with measures like mean, median, and mode.
- **Example:** Knowing the average income of a population helps to gauge overall economic status.

4. Measuring Variability

- **Purpose:** Assesses data spread with metrics like range, variance, and standard deviation.
- **Example:** Understanding the variability in student test scores helps evaluate consistency in performance.

5. Supporting Data Cleaning

- **Purpose:** Helps identify missing values, incorrect entries, or inconsistencies in the dataset.
- **Example:** A column with many zero values in sales data could indicate data entry errors.

6. Basis for Hypothesis Testing

- **Purpose:** Provides foundational knowledge for further statistical tests.
- **Example:** If a variable is skewed, you may choose a non-parametric test instead of a parametric one.

7. Guiding Data Visualization

- **Purpose:** Enhances the interpretation of data with appropriate visual tools.
- **Example:** Bar charts for categorical variables and histograms for numerical variables effectively convey key patterns.

8. Simplifying Data Exploration

- **Purpose:** Offers a straightforward approach for gaining initial insights into the dataset.
- **Example:** Univariate analysis helps determine key features for predictive models.

9. Decision-Making in Business

- **Purpose:** Supports informed decisions by highlighting key metrics.
- **Example:** Analysing customer age distribution helps target specific age groups for marketing campaigns.

10. Preparation for Multivariate Analysis

- **Purpose:** Ensures that variables are understood independently before exploring relationships.
- **Example:** Analysing individual sales trends before evaluating their correlation with marketing expenses.

Here are some common graphs used for univariate analysis:

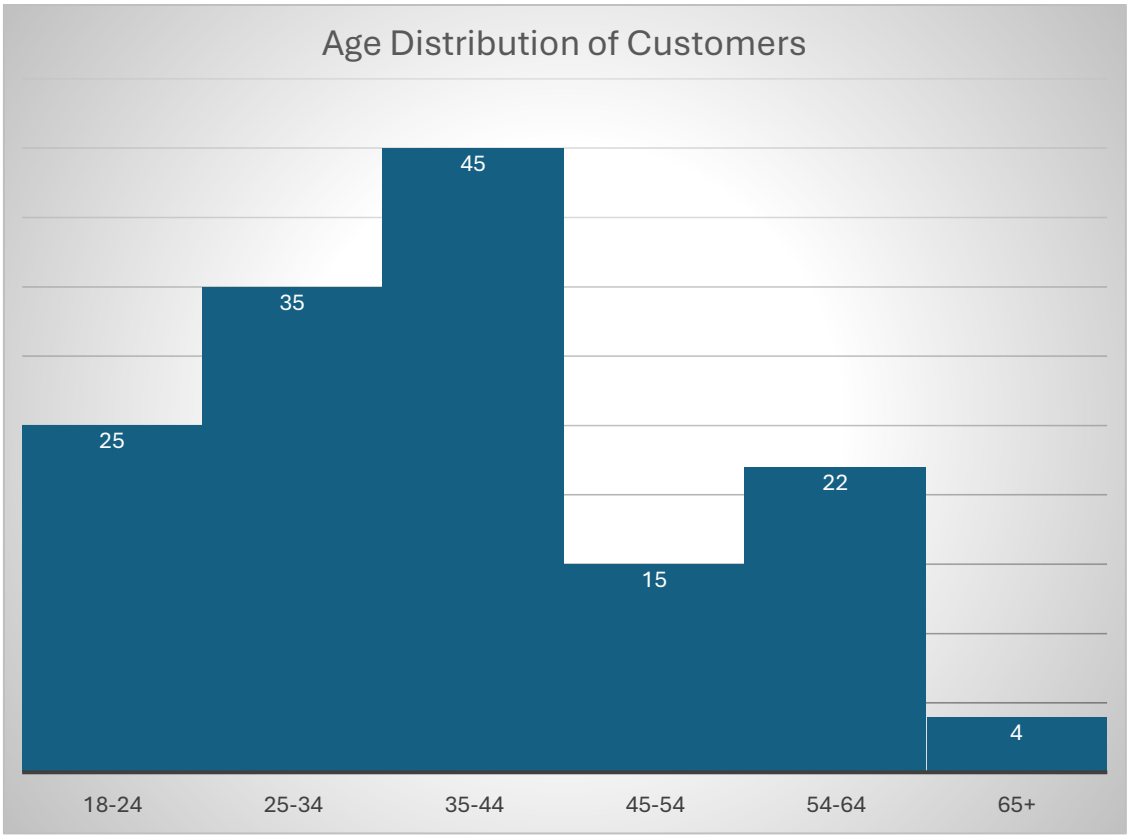
1. Histogram

When to use: To visualize the distribution of a numerical variable.

Type of data: Numerical

Industry use case: A marketing team analysing the age distribution of their customers to target specific demographics.

Age Group	Customers
18-24	25
25-34	35
35-44	45
45-54	15
54-64	22
65+	4



HISTOGRAM CHART

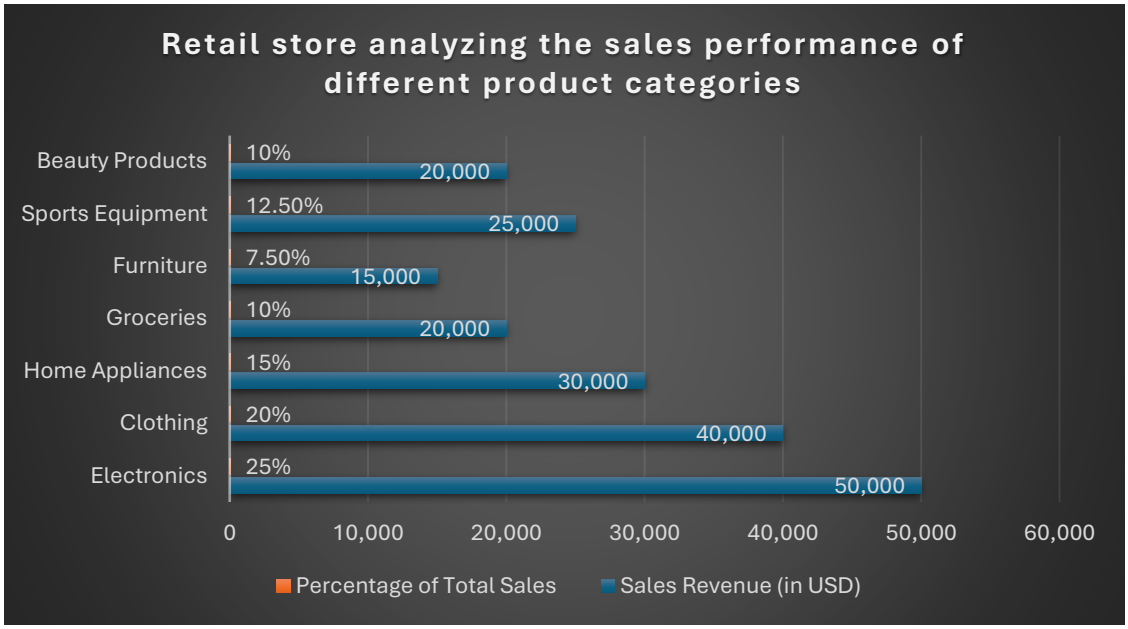
2. Bar Chart

When to use: To compare the frequencies or proportions of different categories of a categorical variable.

Type of data: Categorical

Industry use case: A retail store analysing the sales performance of different product categories to identify bestsellers.

Product Category	Sales Revenue (in USD)	Percentage of Total Sales
Electronics	50,000	25%
Clothing	40,000	20%
Home Appliances	30,000	15%
Groceries	20,000	10%
Furniture	15,000	7.50%
Sports Equipment	25,000	12.50%
Beauty Products	20,000	10%



BAR CHART

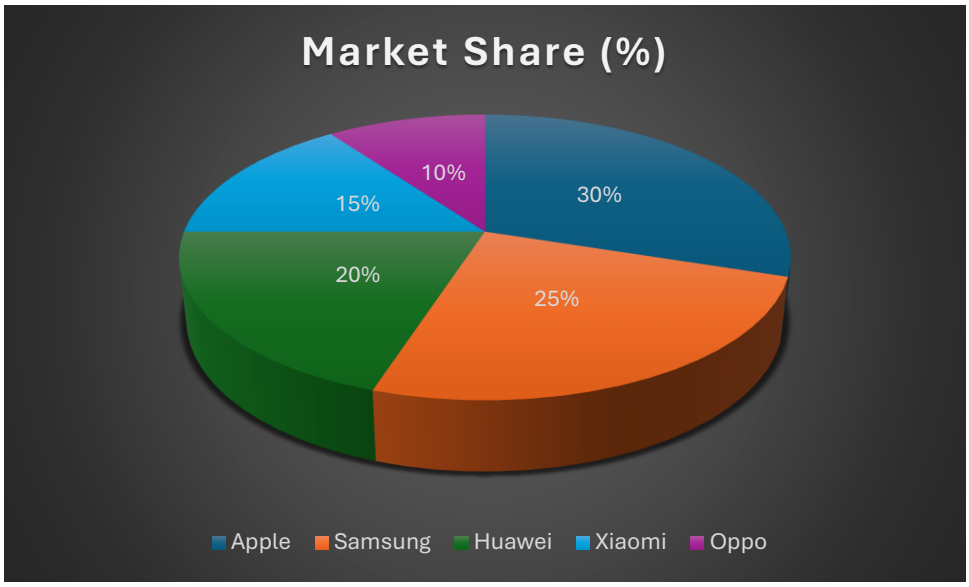
3. Pie Chart

When to use: To show the proportions of a whole represented by different categories of a categorical variable.

Type of data: Categorical

Industry use case: A company visualizing the market share of its competitors in a specific industry.

Company Name	Market Share (%)
Apple	30%
Samsung	25%
Huawei	20%
Xiaomi	15%
Oppo	10%



PIE CHART

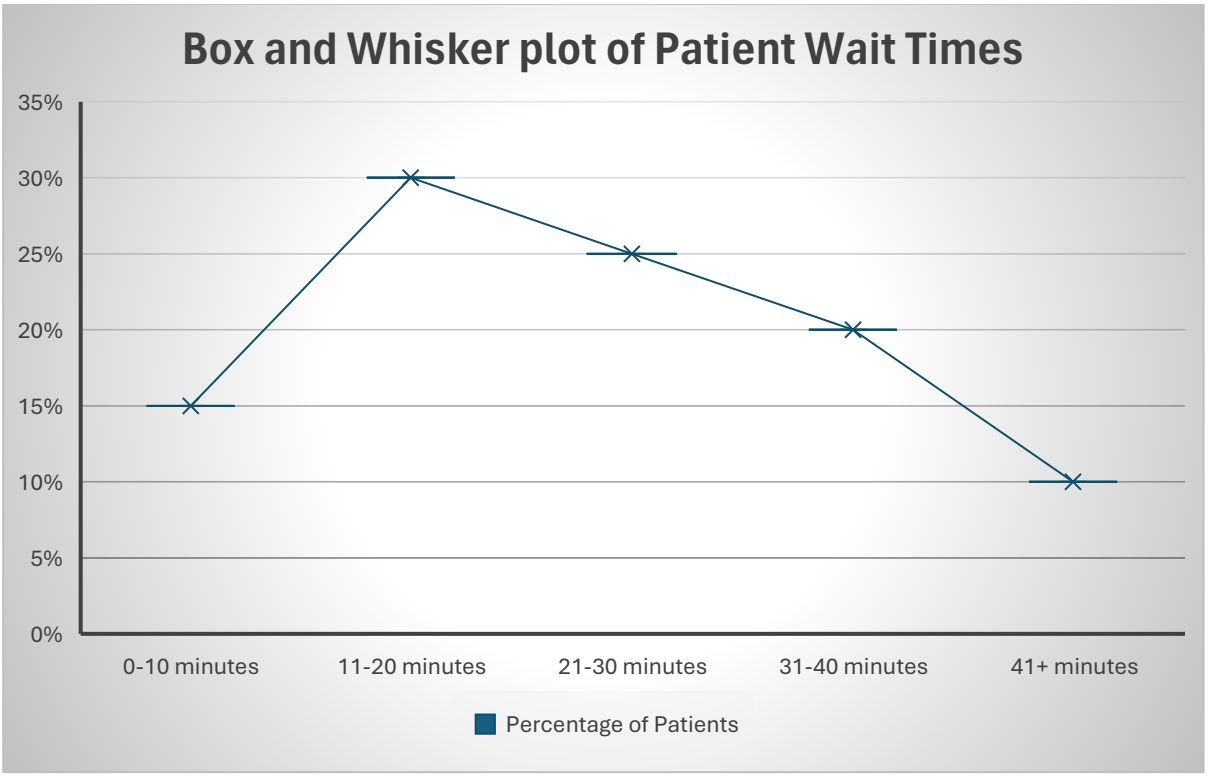
4. Box Plot

When to use: To visualize the distribution of a numerical variable, including its quartiles, median, outliers, and potential skewness.

Type of data: Numerical

Industry use case: A healthcare provider analysing the distribution of patient wait times to identify areas for improvement.

Wait Time Range (Minutes)	Percentage of Patients
0-10 minutes	**15%**
11-20 minutes	**30%**
21-30 minutes	**25%**
31-40 minutes	**20%**
41+ minutes	**10%**



BOX AND WHISKER PLOT

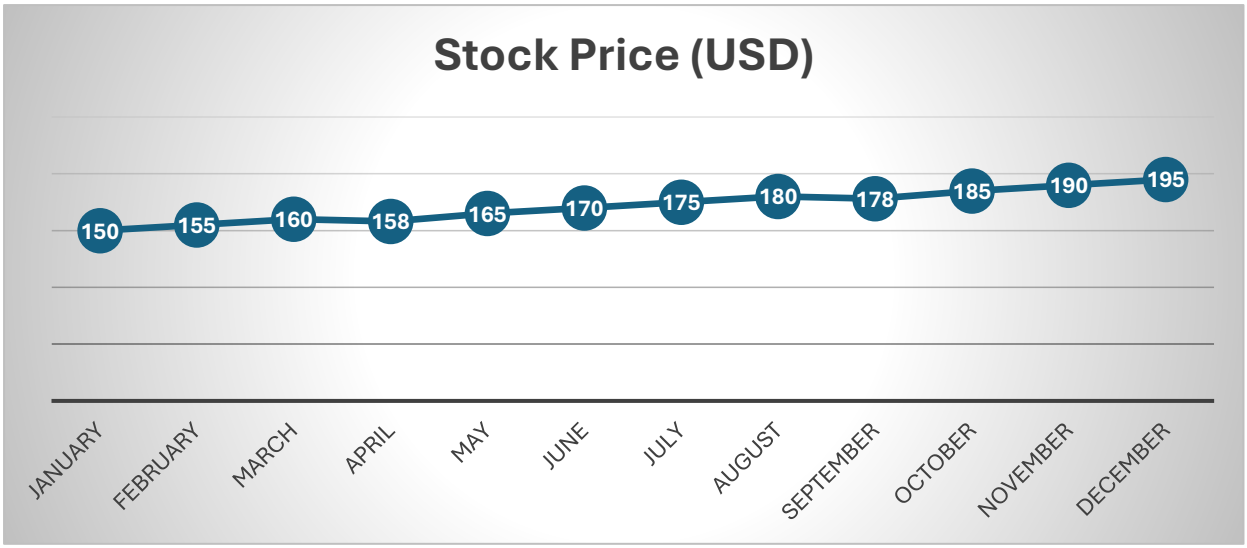
5. Line Chart

When to use: To show the trend of a numerical variable over time or another continuous variable.

Type of data: Numerical

Industry use case: A financial analyst charting the stock price of a company over several months.

Month	Stock Price (USD)
January	150
February	155
March	160
April	158
May	165
June	170
July	175
August	180
September	178
October	185
November	190
December	195



LINE CHART

6. Frequency Table

When to use: To list the frequency of each value or range of values of a numerical or categorical variable.

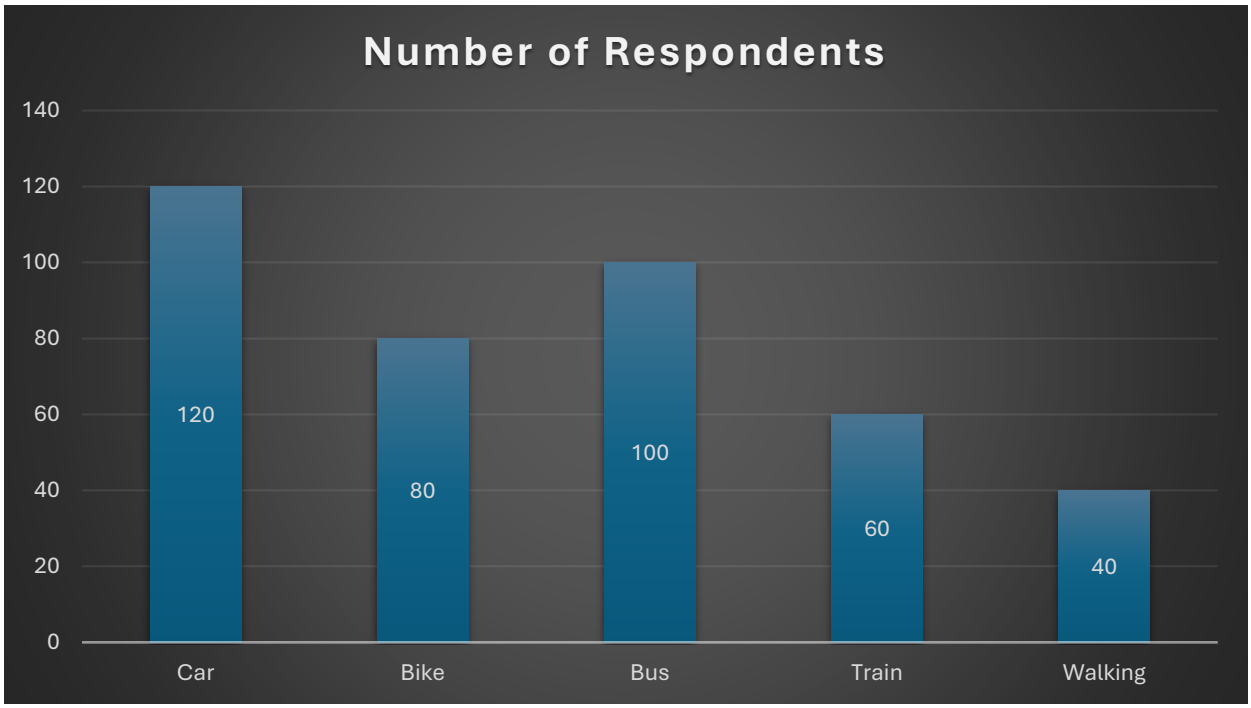
Type of data: Both Numerical and Categorical

Industry use case: A market research firm counting the number of respondents who chose each answer option in a survey.

"What is your preferred mode of transportation for daily commute?"

Options and Responses:

Answer Option	Number of Respondents
Car	120
Bike	80
Bus	100
Train	60
Walking	40



FREQUENCY TABLE

2.Bivariate Analysis

Bivariate analysis is a statistical method used to explore the relationship between two variables. It helps us understand if and how these variables are related, providing insights into potential dependencies, correlations, or patterns.

Definition: Bivariate analysis examines the relationship between two variables. It helps determine if there's a correlation, association, or causal link between them.

Impact:

- Identifying relationships and patterns between variables.
- Understanding how changes in one variable may affect another.
- Predicting future outcomes based on the relationship between variables.

Uses of Bivariate Analysis

1. Identifying Relationships

- **Purpose**: To evaluate whether two variables are related (positively, negatively, or not at all).
- **Example**: In real estate, bivariate analysis can examine how property prices are related to their proximity to public transportation.

2. Predictive Insights

- **Purpose**: Helps in building predictive models by understanding how one variable can predict another.
- **Example**: In finance, analysing the relationship between a company's earnings and its stock price helps forecast future stock performance.

3. Decision-Making

- **Purpose**: Provides evidence-based insights for decision-making.
- **Example**: In marketing, comparing ad spend (independent variable) to sales (dependent variable) guides budgeting strategies.

4. Detecting Patterns

- **Purpose**: Identifies trends or anomalies in data.
- **Example**: In HR, evaluating the correlation between employee engagement scores and turnover rates can reveal key trends affecting retention.

5. Testing Hypotheses

- **Purpose**: Tests assumptions about relationships between variables using statistical techniques like correlation or regression.
- **Example**: In academia, researchers test if study hours positively impact students' exam scores.

6. Business Strategy Development

- **Purpose**: Guides strategies by linking operational metrics to performance outcomes.
- **Example**: In e-commerce, analysing the relationship between website traffic and conversion rates helps optimize user experience strategies.

7. Risk Assessment

- **Purpose**: Identifies risk factors by examining correlations.
- **Example**: In insurance, analysing the relationship between policyholder age and claim frequency helps in risk profiling.

Methods Used in Bivariate Analysis:

- **Correlation Analysis**: Measures the strength of the relationship between two numerical variables.
- **Regression Analysis**: Examines how one variable affects another.
- **Chi-Square Test**: Used for relationships between categorical variables.
- **T-Tests/ANOVA**: Compares means of numerical variables across categories.

Here are some common graphs used for bivariate analysis:

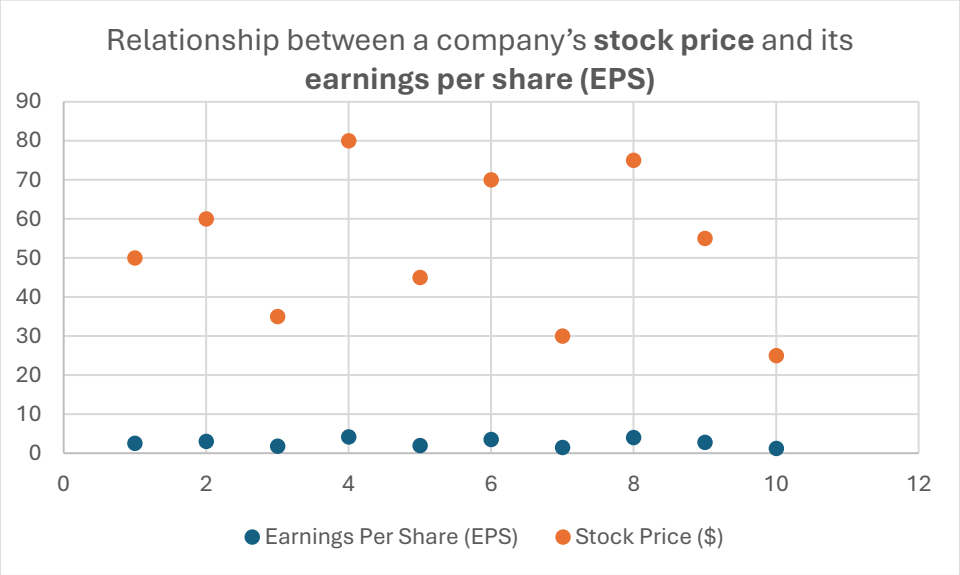
1. Scatter Plots

When to use: To explore the relationship between two numerical variables.

Variable types: Numerical vs. Numerical

Industry example: In finance, a scatter plot could show the relationship between a company's stock price and its earnings per share.

Company	Earnings Per Share (EPS)	Stock Price ($)
A	2.5	50
B	3	60
C	1.8	35
D	4.2	80
E	2	45
F	3.5	70
G	1.5	30
H	4	75
I	2.8	55
J	1.2	25



SCATTER PLOT

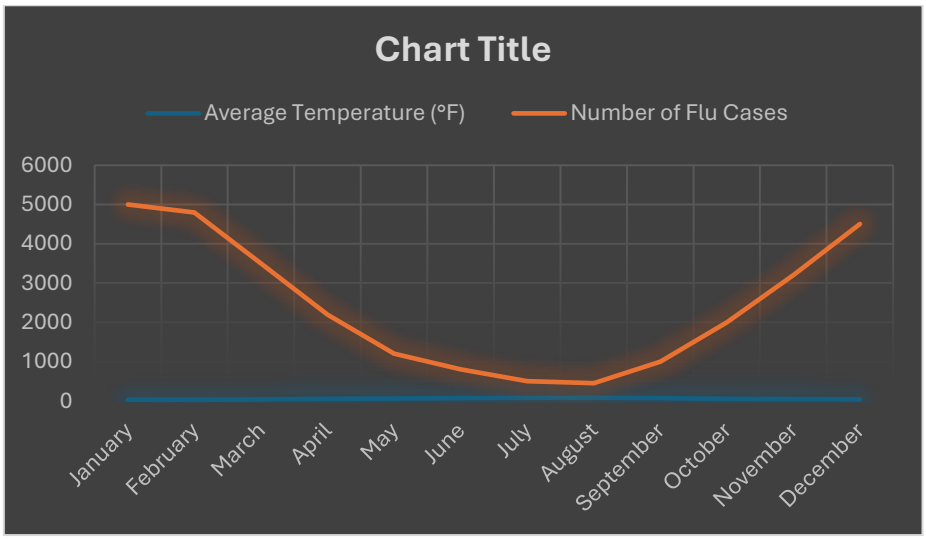
2. Line Charts

When to use: To visualize trends and changes in the relationship between two numerical variables over time.

Variable types: Numerical vs. Numerical (with time as one variable)

Industry example: In healthcare, a line chart could depict the correlation between the number of flu cases and the average temperature throughout the year.

Month	Average Temperature (°F)	Number of Flu Cases
January	30	5,000
February	32	4,800
March	40	3,500
April	50	2,200
May	60	1,200
June	70	800
July	75	500
August	73	450
September	65	1,000
October	55	2,000
November	45	3,200
December	35	4,500



LINE CHART

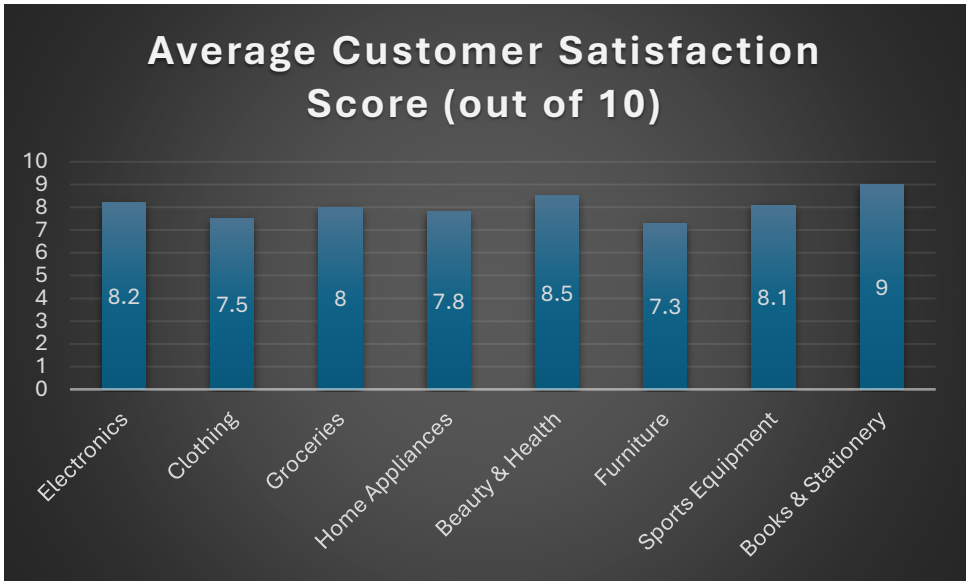
3. Bar Charts

When to use: To compare the frequencies or averages of a numerical variable across different categories of a categorical variable.

Variable types: Numerical vs. Categorical

Industry example: In marketing, a bar chart could showcase the average customer satisfaction scores for different product categories.

Product Category	**Average Customer Satisfaction Score (out of 10)**
Electronics	8.2
Clothing	7.5
Groceries	8
Home Appliances	7.8
Beauty & Health	8.5
Furniture	7.3
Sports Equipment	8.1
Books & Stationery	9



BAR CHART

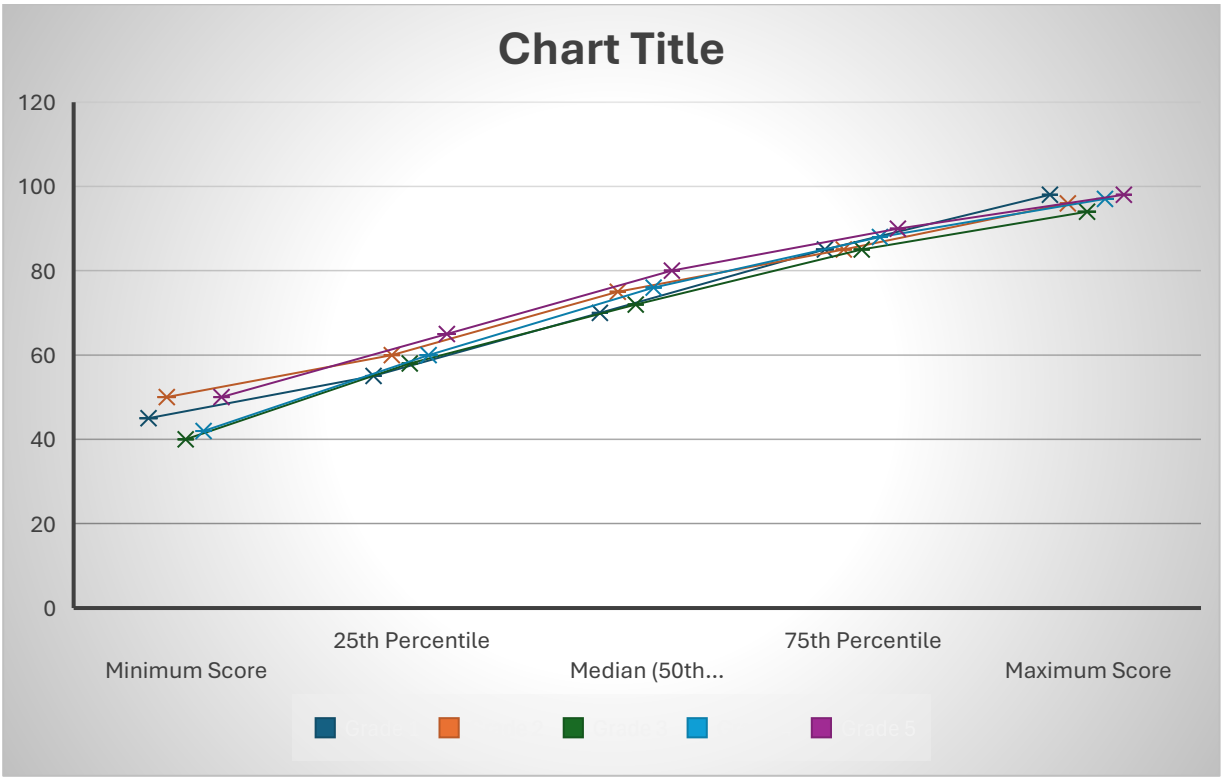
4. Boxplots

When to use: To compare the distribution of a numerical variable across different categories of a categorical variable.

Variable types: Numerical vs. Categorical

Industry example: In education, a boxplot could illustrate the distribution of exam scores for students in different grade levels.

Grade Level	Minimum Score	25th Percentile	Median (50th Percentile)	75th Percentile	Maximum Score
Grade 1	45	55	70	85	98
Grade 2	50	60	75	85	96
Grade 3	40	58	72	85	94
Grade 4	42	60	76	88	97
Grade 5	50	65	80	90	98



BOX PLOT

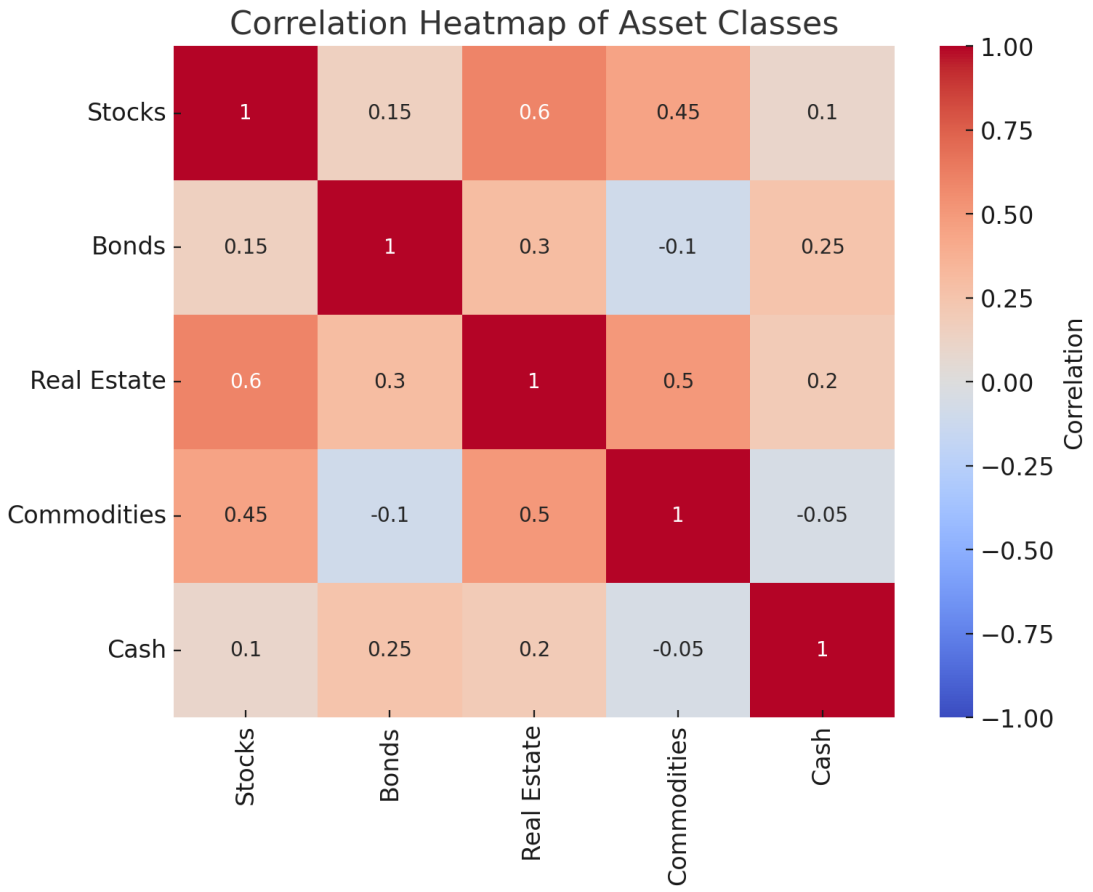
5. Heatmaps

When to use: To visualize the correlation or strength of relationship between multiple pairs of numerical variables.

Variable types: Numerical vs. Numerical (multiple pairs)

Industry example: In finance, a heatmap can display the correlation matrix between different asset classes, helping investors understand their interdependencies.

Asset Class	Stocks	Bonds	Real Estate	Commodities	Cash
Stocks	1	0.15	0.6	0.45	0.1
Bonds	0.15	1	0.3	-0.1	0.25
Real Estate	0.6	0.3	1	0.5	0.2
Commodities	0.45	-0.1	0.5	1	-0.05
Cash	0.1	0.25	0.2	-0.05	1



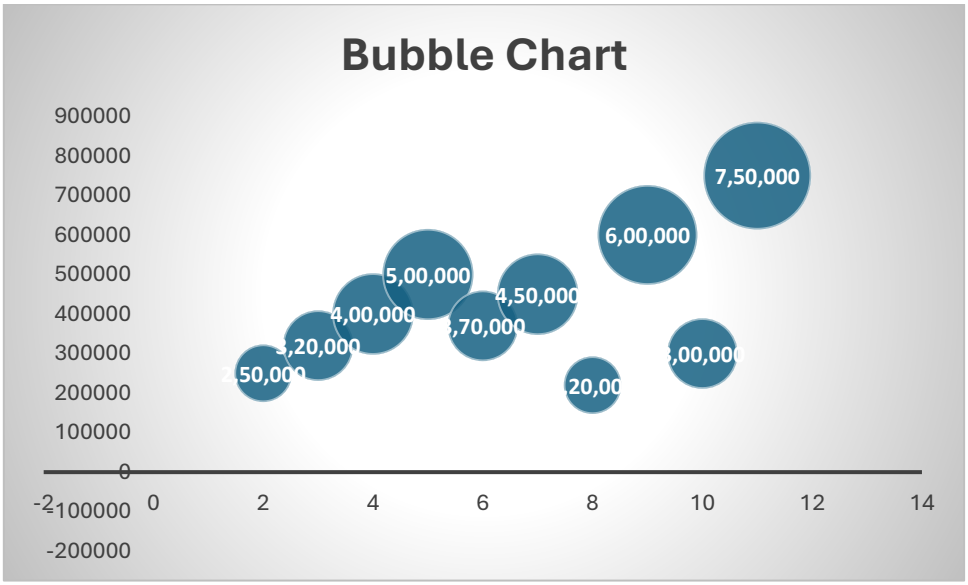
6. Bubble Charts

When to use: To display three variables simultaneously: two numerical variables are plotted on the X and Y axes, while a third numerical variable is represented by the size of the bubbles.

Variable types: Numerical vs. Numerical vs. Numerical

Industry example: In real estate, a bubble chart might show the relationship between house prices (Y-axis) and lot size (X-axis), with the size of the bubbles representing the number of bedrooms.

Lot Size (sq ft)	House Price ($)	Number of Bedrooms
5,000	2,50,000	2
6,500	3,20,000	3
8,000	4,00,000	4
10,000	5,00,000	5
7,500	3,70,000	3
9,000	4,50,000	4
4,000	2,20,000	2
12,000	6,00,000	6
6,000	3,00,000	3
15,000	7,50,000	7



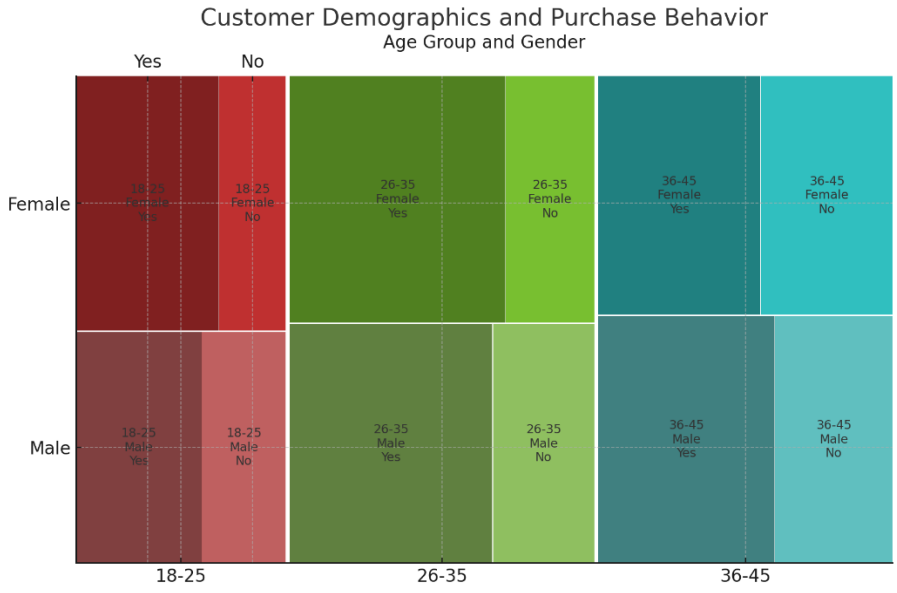
7. Mosaic Plots

When to use: To visualize the frequencies or proportions of observations across two categorical variables.

Variable types: Categorical vs. Categorical

Industry example: In marketing, a mosaic plot could show the distribution of customer demographics (e.g., age group and gender) based on their purchase behavior.

Age Group	Gender	Purchased	Count
18-25	Male	Yes	120
18-25	Male	No	80
18-25	Female	Yes	150
18-25	Female	No	70
26-35	Male	Yes	200
26-35	Male	No	100
26-35	Female	Yes	220
26-35	Female	No	90
36-45	Male	Yes	180
36-45	Male	No	120
36-45	Female	Yes	160
36-45	Female	No	130



3. Multivariate Analysis

Multivariate analysis is a set of statistical techniques used to analyze data with three or more variables. It goes beyond simple relationships between two variables (bivariate analysis) to explore complex interactions and patterns within datasets.

Definition: Multivariate analysis involves examining more than two variables simultaneously. It allows for a more complex understanding of relationships and dependencies within the data.

Impact:

- Uncovering complex interactions between multiple variables.
- Identifying hidden patterns and trends.
- Building more accurate predictive models.

Uses of Multivariate Analysis

Multivariate analysis examines the relationships between three or more variables simultaneously. It is crucial for understanding complex data interactions and making informed decisions across various fields.

1. Understanding Complex Relationships

- **Purpose**: To analyze how multiple variables interact with each other.
- **Example**: In healthcare, studying how age, diet, and exercise collectively influence heart disease risk.

2. Predictive Modelling

- **Purpose**: Builds models that predict outcomes using several predictors.
- **Example**: In marketing, using customer age, income, and purchase history to predict future buying behavior.

3. Market Segmentation

- **Purpose**: Identifies distinct customer groups based on multiple characteristics.
- **Example**: In retail, clustering customers by shopping habits, spending levels, and product preferences.

4. Risk Assessment and Management

- **Purpose**: Evaluates risk by Analysing various contributing factors.
- **Example**: In finance, assessing investment risk by Analysing market trends, interest rates, and economic indicators.

5. Product and Process Optimization

- **Purpose**: Improves product quality and operational efficiency by Analysing multiple process variables.
- **Example**: In manufacturing, adjusting temperature, pressure, and material composition to optimize product durability.

6. Policy and Decision-Making

- **Purpose**: Supports evidence-based policy decisions by Analysing complex data sets.
- **Example**: In public health, studying how income, education, and access to healthcare influence disease spread.

Common Methods Used in Multivariate Analysis:

- **Multiple Regression Analysis**: Evaluates how multiple independent variables affect a dependent variable.
- **Factor Analysis**: Identifies underlying variables (factors) that explain data correlations.
- **Cluster Analysis**: Groups similar data points based on shared attributes.
- **Principal Component Analysis (PCA)**: Reduces data complexity by transforming variables into principal components.
- **Discriminant Analysis**: Classifies data into categories based on predictor variables.
- **MANOVA (Multivariate Analysis of Variance)**: Tests differences between groups on multiple dependent variables.

Here are some common graphs used for multivariate analysis:

1. Scatterplot Matrix

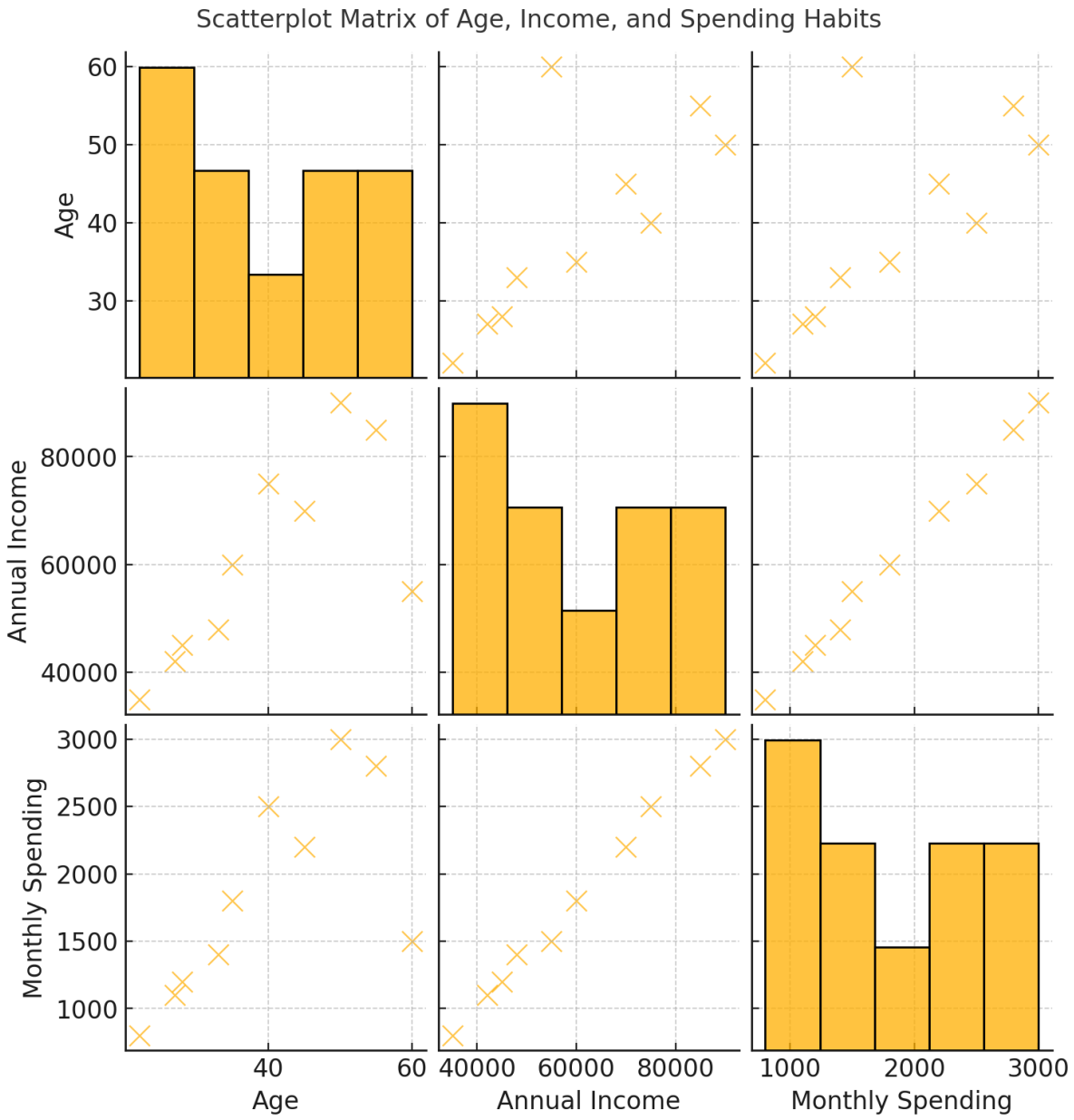
When to use:

- To visualize the relationship between multiple numerical variables.
- To identify potential correlations and outliers.

Column Types: Numerical, Numerical

Example Use Case: A marketing team analysing customer data might create a scatterplot matrix to see how factors like age, income, and spending habits are interrelated.

e (Years)	**Annual Income ($)**	**Monthly Spending ($)**
22	35,000	800
28	45,000	1,200
35	60,000	1,800
40	75,000	2,500
50	90,000	3,000
60	55,000	1,500
45	70,000	2,200
33	48,000	1,400
27	42,000	1,100
55	85,000	2,800



Scatterplot Matrix

2. Parallel Coordinates Plot

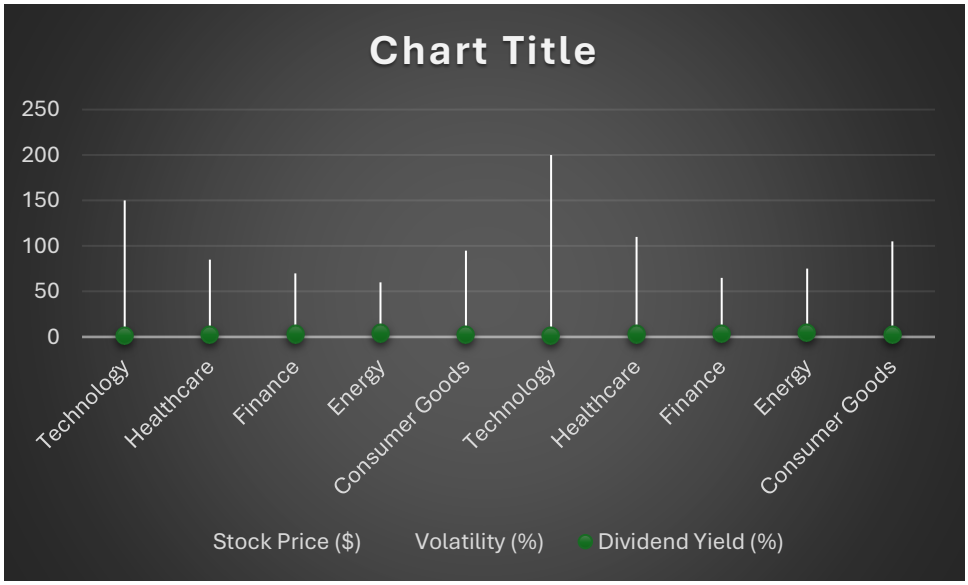
When to use:

- To visualize and compare multiple numerical variables across different observations.
- To identify clusters and patterns in high-dimensional data.

Column Types: Numerical, Numerical

Example Use Case: A financial analyst might use a parallel coordinates plot to analyze stock performance across different sectors, considering factors like price, volatility, and dividend yield.

Sector	Stock Price ($)	Volatility (%)	Dividend Yield (%)
Technology	150	20	1.2
Healthcare	85	15	2.5
Finance	70	25	3
Energy	60	30	4
Consumer Goods	95	18	2
Technology	200	22	0.8
Healthcare	110	12	3
Finance	65	28	3.5
Energy	75	35	4.5
Consumer Goods	105	16	1.8



3. Radar Chart (Spider Chart)

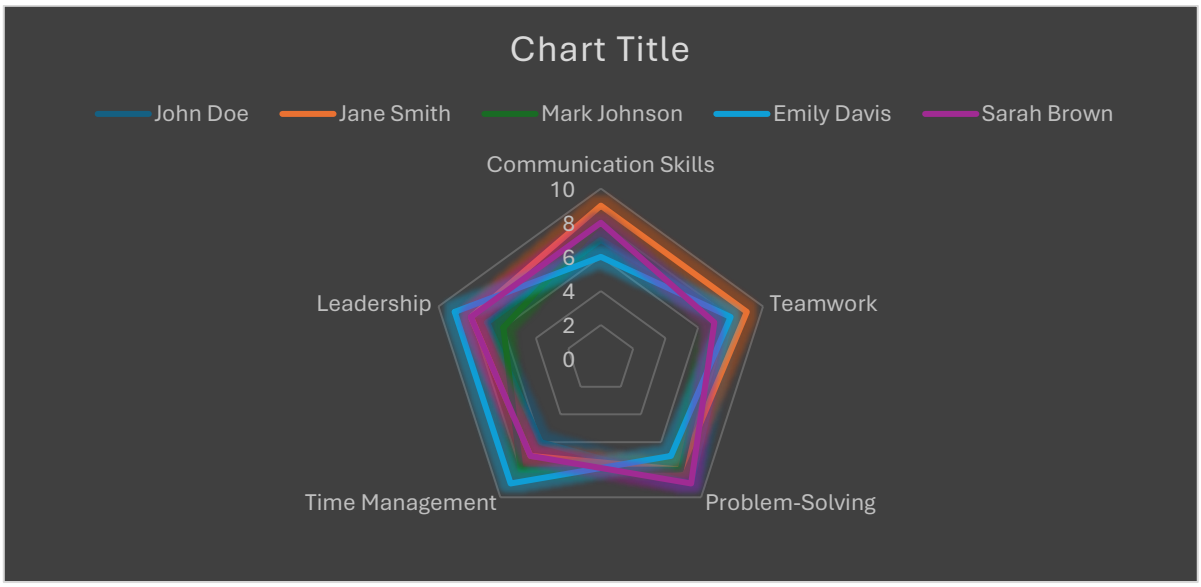
When to use:

- To compare the relative performance of different entities across multiple variables.
- To highlight strengths and weaknesses of each entity.

Column Types: Numerical, Categorical (for entity labels)

Example Use Case: An HR department might use a radar chart to compare the performance of different employees based on criteria like communication skills, teamwork, and problem-solving.

ame	Communication Skills	Teamwork	Problem-Solving	Time Management	Leadership
John Doe	7	8	9	6	7
Jane Smith	9	9	8	7	8
Mark Johnson	8	7	8	8	6
Emily Davis	6	8	7	9	9
Sarah Brown	8	7	9	7	8



4. Biplot

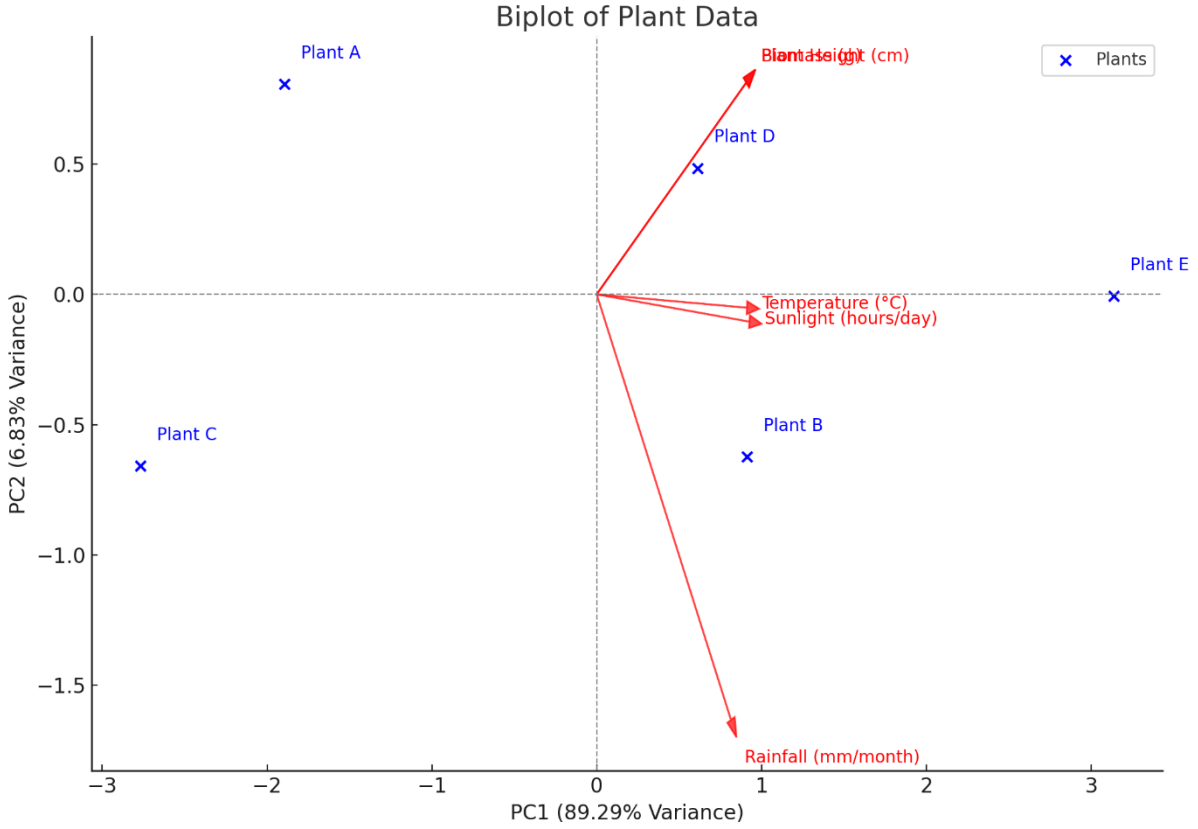
When to use:

- To visualize both the relationships between variables and the relationship between variables and observations.

Column Types: Numerical, Numerical

Example Use Case: A biologist studying plant growth might use a biplot to see how different environmental factors (e.g., sunlight, rainfall, temperature) contribute to plant height and biomass.

Plant Species	Sunlight (hours/day)	Rainfall (mm/month)	Temperature (°C)	Plant Height (cm)	Biomass (g)
Plant A	6	80	22	35	25
Plant B	8	120	25	40	30
Plant C	5	100	20	30	20
Plant D	7	110	23	45	35
Plant E	9	130	27	50	40



5. Heatmap

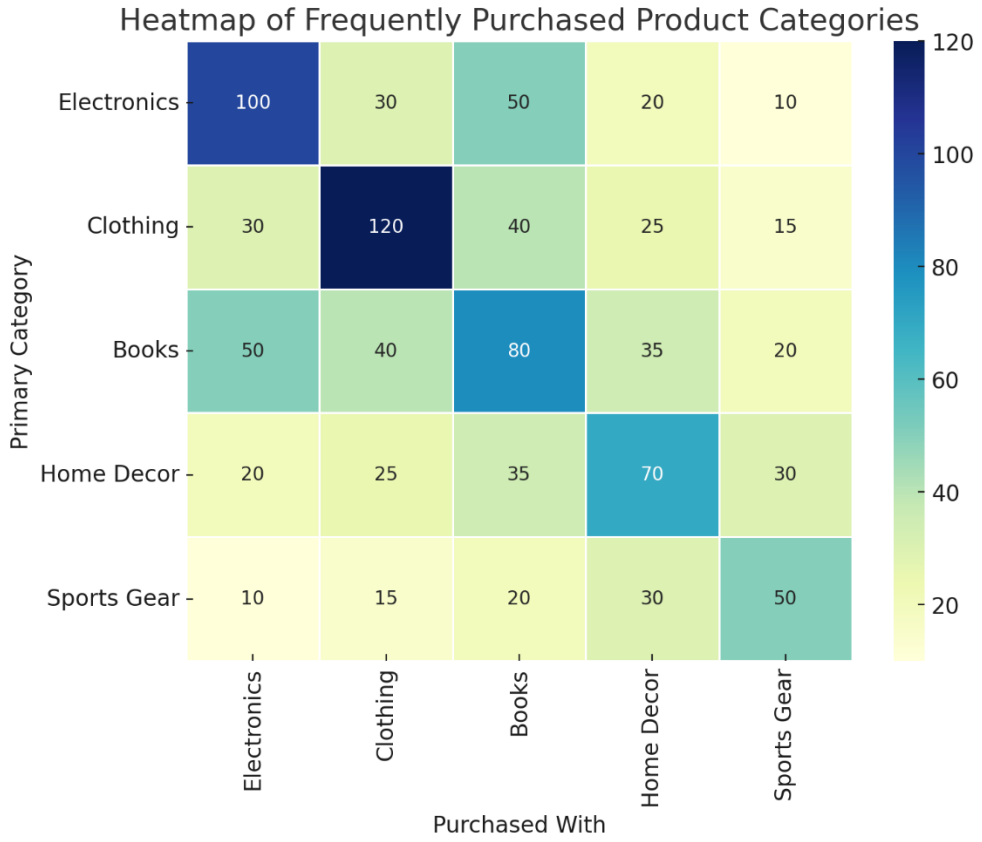
When to use:

- To visualize the correlation matrix between multiple variables.
- To identify strong positive or negative relationships between variables.

Column Types: Numerical, Numerical

Example Use Case: An e-commerce company might use a heatmap to identify which product categories are frequently purchased together.

	Electronics	**Clothing**	**Books**	**Home Decor**	**Sports Gear**
Electronics	100	30	50	20	10
Clothing	30	120	40	25	15
Books	50	40	80	35	20
Home Decor	20	25	35	70	30
Sports Gear	10	15	20	30	50



Conclusion

Data visualization has emerged as an indispensable tool in today's data-driven world, playing a critical role in enhancing decision-making, communication, and efficiency across industries. This project highlights the transformative power of visual analytics through the implementation of Power BI dashboards for Shopperr.ai. By converting complex datasets into interactive and easily interpretable visuals, it becomes possible to uncover trends, detect anomalies, and derive actionable insights that directly impact business outcomes.

The integration of various visualization types, such as histograms, scatter plots, and line charts, demonstrates how univariate, bivariate, and multivariate data analyses are effectively communicated. These tools facilitate not only a deeper understanding of the data but also a means to make informed strategic decisions. Moreover, dashboards enable real-time monitoring of key performance indicators (KPIs), improving agility and operational responsiveness.

Beyond business applications, data visualization fosters enhanced collaboration across teams by presenting insights in a universally comprehensible format. This eliminates barriers posed by technical data, ensuring all stakeholders can actively contribute to discussions and decisions. Additionally, visual analytics aids in storytelling, allowing businesses to present compelling narratives that resonate with clients, investors, and other stakeholders.

In conclusion, the project underscores the necessity of adopting advanced visualization tools like Power BI in modern organizations. As businesses continue to generate and rely on vast amounts of data, leveraging the power of data visualization will remain critical for staying competitive, driving growth, and fostering innovation in a dynamic global landscape.