

Experiment 4

Aim: To Explore data visualization techniques.

I-OBJECTIVE

➤ **To understand basic concepts of data visualization and techniques**

➤ **To Explore data visualization techniques**

II-THEORY

1. What is Data Visualization and Why is It Important?

Data visualization is the practice of translating information into a visual context, such as a map or graph, to make data easier for the human brain to understand. Its primary goal is to identify patterns, trends, and outliers in large datasets. In an era of "Big Data," visualization tools are indispensable for converting raw numbers into actionable insights.

Data visualization is essential for understanding and communicating information effectively. Key reasons why it's important are:

- **Deciphers Complexity:** Converts dense, multifaceted datasets into intuitive formats like charts and graphs.
- **Identifies Trends:** Highlights correlations and shifts that remain hidden in traditional spreadsheets.
- **Accelerates Efficiency:** Enables stakeholders to grasp key takeaways at a glance, bypassing the need for manual data processing.
- **Enhances Communication:** Provides a universal language to share technical findings with non-technical audiences.
- **Narrative Strength:** Uses data to tell a cohesive story, guiding viewers toward logical conclusions and informed decisions.

2. Uses of data visualization.

Data visualization is used across various industries to improve decision-making and drive results.

- **Business Analytics** Companies use interactive dashboards to monitor real-time performance and track Key Performance Indicators (KPIs). By visualizing sales trends and customer behavior, stakeholders can move away from gut feelings and make objective, data-driven decisions that increase profitability.
- **Healthcare** Visualization is life-saving in medicine, used to map the spread of infectious diseases and analyze patient outcomes. It allows hospital administrators to manage complex logistics, such as bed occupancy and staff allocation, through clear, easy-to-read graphical interfaces.
- **Sports & Athletics** Coaches and scouts use visual data to break down player mechanics, heatmaps of movement on the field, and overall team efficiency. These insights allow for the optimization of training regimens and the development of high-level game strategies based on statistical probability.

- **Retail and E-commerce** Retailers visualize purchasing patterns and inventory turnover to predict future demand and prevent stockouts. By seeing which products are trending in specific regions, businesses can fine-tune their marketing efforts and personalize the shopping experience for different customer segments.

3. Types of data visualization.

1. Bar Charts

Bar charts are used to compare values across different categories using rectangular bars. X-axis shows categories while Y-axis represents values. Common types include horizontal, stacked and grouped bar charts.

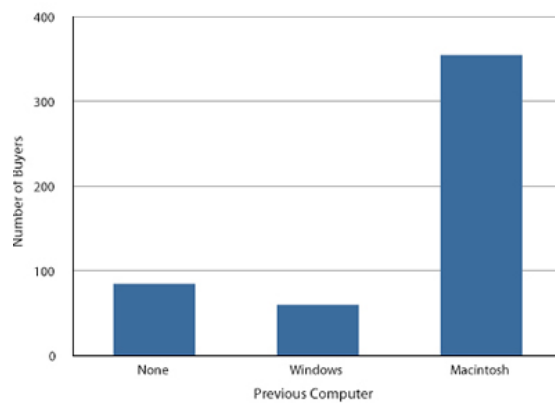


Fig 1: Bar Chart

When to Use:

- To compare different categories
- To rank values from highest to lowest
- To show relationships between multiple variables

2. Line Charts

Line charts show how values change over time by connecting data points with lines. They help visualize trends like increases, decreases or stability.

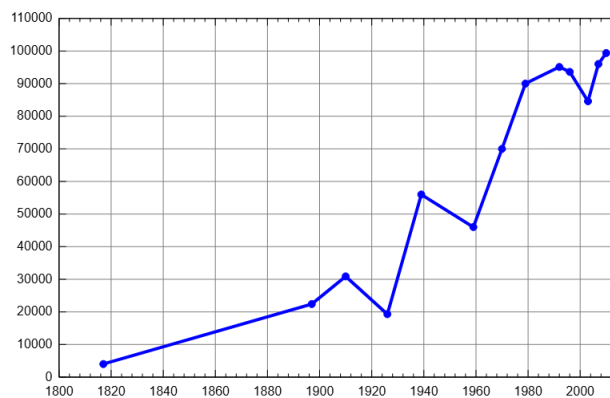


Fig 2: Line Chart

When to Use:

- To track changes over time
- To compare trends across multiple data series
- For time series analysis

3. Pie Charts

Pie charts are round charts divided into slices, where each slice shows a part of the whole. The size of each slice represents its percentage.

Number of Students

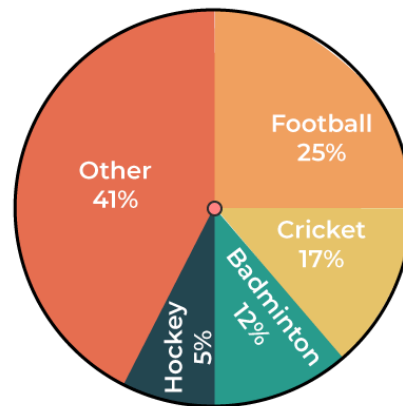


Fig 3: Pie Chart

When to Use:

- To show how different parts contribute to a whole
- To highlight a dominant category

4. Scatter Chart (Plots)

Scatter charts use dots to show the relationship between two numerical variables. X-axis shows the independent variable and Y-axis shows the dependent variable.

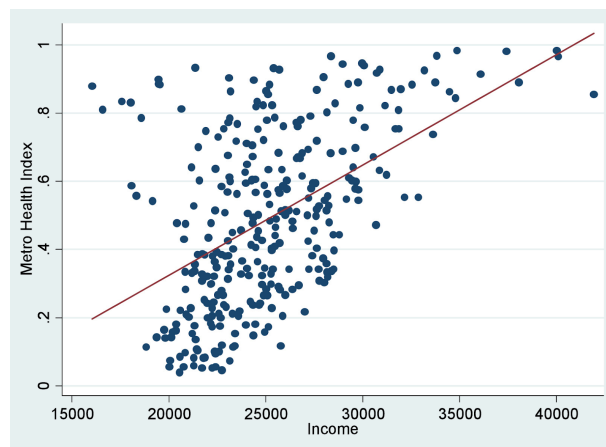


Fig 4: Scatter plot

When to Use:

- To observe relationships between two variables
- To detect patterns, clusters or outliers in data

5. Histogram

A histogram displays the distribution of numerical data by grouping values into intervals (bins) and showing their frequency as bars. It helps reveal the shape, spread and patterns in the data.

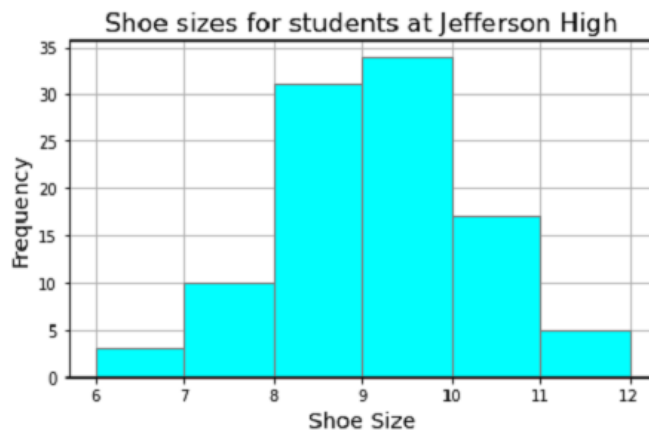


Fig 5: Histogram

When to Use:

- To visualize the distribution of numerical data
- To explore patterns, trends and outliers

6. Heatmap

A heatmap displays data in a matrix format using color to represent values. It's ideal for spotting patterns, correlations and variations in large datasets.

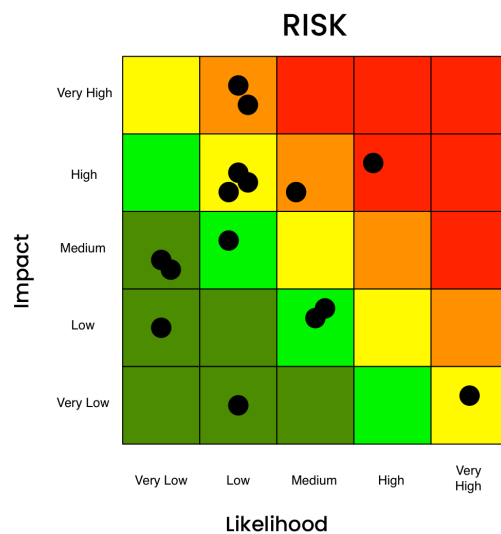


Fig 6: Heatmap

When to Use:

- To identify clusters or groupings in data
- To visualize correlations between variables
- For risk analysis in fields like finance or network security

7. Area Chart

An area chart shows trends over time by filling the space beneath a line. It's ideal for visualizing time-series data and highlighting changes across periods.

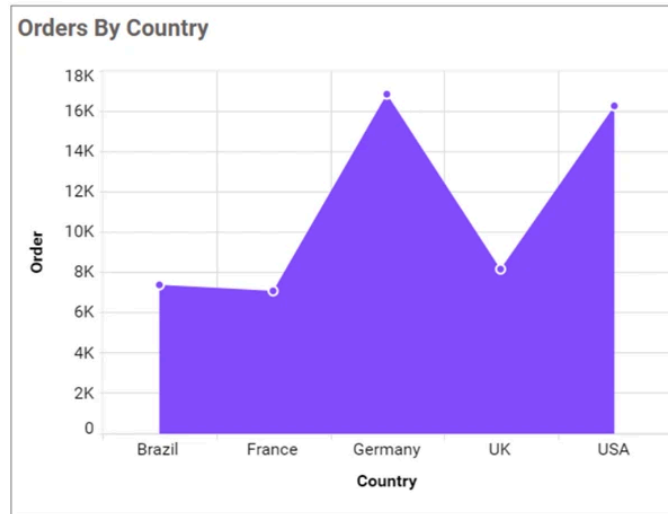


Fig 7: Area Chart

When to Use:

- To track changes or trends over time
- To compare multiple data series
- To highlight patterns like seasonality or cycles

8. Box Plot (Box-and-Whisker Plot)

A box plot displays distribution of numerical data, showing median, quartiles and outliers. It's useful for understanding variability and detecting unusual values.

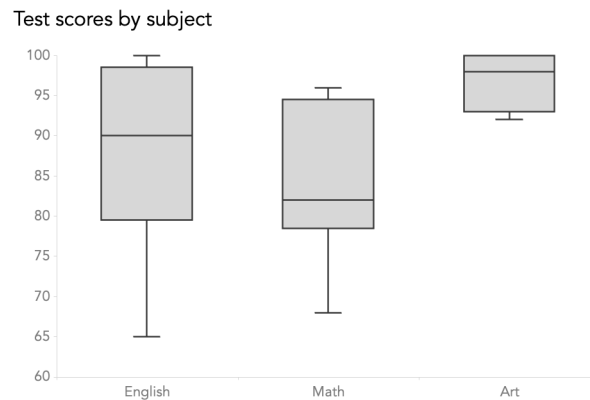


Fig 8: Box Plot

When to Use:

- To detect outliers in data
- To compare distributions across groups
- To visualize spread and variability

4. Univariate, Bivariate and Multivariate data and its analysis.

a) Univariate Data

Univariate analysis is the most fundamental form of statistics, focusing exclusively on a **single variable** to describe its characteristics. Since it does not look at relationships between different factors, its primary purpose is to summarize the data and find its underlying distribution.

- **Central Tendency:** It identifies the "typical" value within a dataset using the **mean**, **median**, and **mode**.
- **Data Dispersion:** It measures how spread out the numbers are using metrics like **range**, **variance**, and **standard deviation**, which help determine if data points are clustered together or widely scattered.
- **Pattern Detection:** This analysis helps identify **skewness** and **outliers** that could potentially distort the overall interpretation.
- **Visual Representation:** Frequent tools include **histograms** to show frequency, **box plots** to show quartiles, and **density plots** to visualize the "shape" of the information.

Example: Recording the heights (in cm) of a group of students. The goal is simply to find the average height or the tallest/shortest student, rather than comparing height to age or weight.

b) Bivariate data

Bivariate analysis examines the connection between **two different variables** within the same observation. The primary objective is to determine if a relationship exists and how one variable might influence or correlate with the other.

- **Relationship Detection:** It identifies the nature of the link—whether it is **positive**, **negative**, or **non-existent**.
- **Correlation Strength:** It utilizes the **correlation coefficient** to mathematically measure the direction and intensity of the relationship.
- **Trend Visualization:** **Scatter plots** are the most common tool for this analysis, often paired with **regression lines** to illustrate the mathematical trend.
- **Analytical Foundation:** This serves as a critical stepping stone, providing the basic insights needed before advancing to more complex, multi-layered models.

Example: Analyzing the link between **daily temperature** and **ice cream sales**. By plotting these two variables, you can visually and statistically confirm if hotter days lead to higher revenue.

c) Multivariate data

Multivariate analysis examines **three or more variables** simultaneously to understand how they collectively influence a specific outcome. In the real world, results are rarely driven by a single factor; this method allows data scientists to account for the "noise" and overlapping influences in complex systems.

- **Holistic Insight:** This approach is vital when an outcome—such as consumer spending or disease progression—is driven by several independent factors at once.
- **Advanced Statistical Techniques:** It employs sophisticated methods like **Multiple Regression**, **MANOVA** (Multivariate Analysis of Variance), and **Clustering** to group similar data points based on multiple attributes.
- **Complexity Management:** It uses "Dimensionality Reduction" techniques, such as **Principal Component Analysis (PCA)**, to simplify massive datasets while retaining the most important information.
- **High-Dimensional Visualization:** Since humans can't easily visualize more than three dimensions, this analysis relies on **heatmaps**, **pair plots**, and **3D scatter plots** to reveal hidden structures.

Example: An advertiser evaluating **click-through rates**. Instead of just looking at the ad type, they analyze how **ad format**, **user gender**, and **time of day** interact to drive the highest engagement.

Dataset - Employee Attribution Dataset

This dataset consists of 31 columns and 1,470 rows, providing a comprehensive snapshot of a corporate workforce.

- **Context:** The Employee dataset contains historical data regarding demographics, compensation, and professional engagement metrics for individual employees within an organization.
- **Objective:** The primary objective is to monitor and analyze workforce dynamics, salary structures, and the factors influencing employee performance and retention.
- **Identification Attributes:** Each entry is identified by an employee index, with categorical attributes such as **Department**, **JobRole**, and **EducationField** defining the organizational structure.
- **Core Target Variable:** The **MonthlyIncome** serves as a central metric, representing the primary financial compensation assigned to each employee.
- **Career Indicators:** Attributes such as **TotalWorkingYears**, **JobLevel**, and **YearsAtCompany** provide critical insights into the professional experience and tenure of the staff.
- **Engagement & Environment:** Metrics including **JobSatisfaction**, **WorkLifeBalance**, and **EnvironmentSatisfaction** help evaluate the qualitative aspects of the workplace and their potential impact on employee attrition.

III-IMPLEMENTATION

```
import pandas as pd
import matplotlib.pyplot as plt

df_emp = pd.read_csv('empattds.csv')

plt.figure(figsize=(12, 6))
plt.plot(df_emp.index[:100], df_emp['MonthlyIncome'][:100], label='Monthly Income', color='blue')

plt.xlabel('Employee Index')
plt.ylabel('Monthly Income ($)')
plt.title('Employee Monthly Income Distribution')
plt.grid(True)
plt.legend()
plt.show()
```



Fig 3: Line Chart Python

Insights:

1. Emphasis on Continuity and Progression

While bar charts treat data as discrete, independent categories, a line chart connects individual data points to highlight a sequence. In the context of HR data, this creates a visual "career timeline." By sorting observations—such as by **Total Working Years**—the chart effectively illustrates the trajectory of **Monthly Income**, showing how experience directly correlates to earnings growth.

2. Trend Identification and Noise Reduction

With a large dataset exceeding 1,400 observations, individual data points can create significant "noise." A line chart smooths these variations into a cohesive "silhouette" of the workforce. This allows for the immediate identification of:

- **Secular Trends:** Determining if the general direction of compensation is moving upward as expected.
- **Volatility:** Identifying specific experience levels or roles where income fluctuations are most frequent or sharp.

3. Efficiency in Multi-Variable Comparison

The dataset contains multiple overlapping financial variables, such as **DailyRate**, **MonthlyRate**, and **MonthlyIncome**. Attempting to display these using grouped bar charts would result in thousands of bars, leading to visual congestion. A line chart allows for the overlay of multiple variables on a single axis, permitting a clear comparison of different pay structures without sacrificing readability.

4. Optimization of Informational Density

Line charts offer a high "information-to-space" ratio, which is essential when dealing with high-volume datasets.

- **Scalability:** Unlike bar charts, which require specific widths for legibility and often necessitate horizontal scrolling, a line chart can compress 1,470 observations into a single, comprehensive view.
- **Logic of Composition:** Since metrics like salary are not "parts of a whole" (100%), tools like pie charts are statistically inappropriate. The line chart remains the most efficient use of "visual real estate" for identifying the big-picture compensation structure.

IV- CONCLUSION

Data visualization functions as the essential link between complex statistical datasets and practical application. By converting abstract numerical values into graphical narratives, it allows for a more intuitive understanding of data. Through the strategic use of tools like line charts, histograms, and scatter plots, analysts can execute **univariate, bivariate, and multivariate analyses**. These methods are vital for identifying latent patterns, systemic correlations, and statistical outliers that are often undetectable in standard tabular formats.

In this experiment, we evaluated various data visualization methodologies and applied them to a comprehensive workforce dataset containing 1,470 employee records. By implementing a line chart via Python's Matplotlib library, we successfully visualized the distribution of **Monthly Income** across the organization to identify structural compensation trends. This specific technique was selected because analyzing workforce demographics requires a clear representation of progression; the line chart effectively connects individual data points to illustrate the relationship between professional variables and financial compensation. Through this practical application, we demonstrated how selecting the appropriate visual tool is essential for accurately interpreting organizational health, identifying pay equity patterns, and communicating complex human resource metrics.

VI- POST LAB QUESTION/ANSWER

1. Compare and contrast the use of heatmaps and treemaps for visualizing hierarchical data. Discuss their strength, limit and scenario where one would be more effective than the other.

Feature	Treemap	Heatmap
Primary Goal	Displaying proportions within a hierarchy.	Displaying variance and patterns across a matrix.
Spatial Logic	Containment (Parent encloses Child).	Intersection (Row crosses Column).
Data Dimensions	2 Quantitative (Size + Color) + Hierarchy.	1 Quantitative (Color) + 2 Categorical Axes.
Best For	"How much does X contribute to the whole?"	"Where are the hotspots or outliers?"
Weakness	Precision comparison of values (Area is hard to judge).	Showing hierarchical depth (requires axis grouping).

Treemap:

A treemap displays hierarchical data as a set of nested rectangles. Each branch of the tree is given a rectangle, which is then tiled with smaller rectangles representing sub-branches.

- **How it works:**
 - Size: Represents the primary quantitative value (e.g., Revenue).
 - Color: Represents a secondary metric (e.g., Profit Margin or Growth).
 - Nesting: Represents the parent-child relationship (e.g., Region → Country → City).
- **Strengths:**
 - Space-Efficiency: It displays thousands of items simultaneously within a limited screen space.
 - Part-to-Whole: You can instantly see that "Electronics" makes up 40% of the total, and within Electronics, "Phones" are the dominant contributor.
 - Drill-Down: It is intuitive for interactive dashboards (clicking a "Region" box zooms into "Countries").
- **Limitations:**
 - Human Perception: Humans are bad at comparing the area of non-aligned rectangles. It is hard to tell if one box is 10% larger than another if they have different aspect ratios.
 - Labeling: Small rectangles become impossible to label legibly.

Heatmap:

A heatmap is a graphical representation of data where individual values contained in a matrix are represented as colors. While not inherently hierarchical, it can visualize hierarchy by grouping axes (e.g., sorting the Y-axis by Category, then Sub-Category).

- **How it works:**
 - X and Y Axes: Categorical or continuous variables (e.g., Days of the Week vs. Hours of the Day).
 - Color Intensity: Represents the magnitude of the value at that intersection.
- **Strengths:**
 - Pattern Recognition: It is unbeatable for spotting clusters, outliers, or gaps in dense data.
 - Standardization: Because every cell is the same size, it is much easier to scan a row or column to compare values directly.
 - High Density: You can plot thousands of rows (e.g., gene expression data) to see macro trends.
- **Limitations:**
 - No "Part-to-Whole": A heatmap doesn't tell you that Row A contributes 50% to the total. It only tells you the value of Row A relative to Row B.
 - Abstract Hierarchy: You lose the visual sense of "containment." You have to rely on axis labels to understand the structure.

Scenario 1: Game Analytics (Resource Management)

- **The Data:** we want to see which items players are hoarding in their inventory. The hierarchy is Item Type (Weapon, Potion) → Rarity (Common, Legendary) → Specific Item.
- **Best to use: Treemap.**
 - We need to see that "Weapons" take up 60% of inventory space (Size), and within that, "Legendary Swords" are being hoarded. The "part-to-whole" nature is crucial here.

Scenario 2: Server Log Analysis (Time Series)

- **The Data:** we want to identify when your game servers experience the highest latency. The dimensions are Day of Week and Hour of Day.
 - **Best to use: Heatmap.**
 - We are looking for a "hotspot" (e.g., Saturday at 8 PM is bright red). A treemap would be confusing here because "Saturday" doesn't "contain" 8 PM; they are intersecting variables.
2. **Explain how improper scaling and access manipulation in line charts can lead to misleading interpretation, provide a realworld example of how this would distort decision making in data science.**

Improper scaling and axis manipulation are among the most common ways to distort truth in data visualization. By altering the visual frame of reference, a chart can force the viewer to perceive a trend, volatility, or correlation that does not statistically exist.

1. The Truncated Y-Axis

The most pervasive manipulation is starting the Y-axis at a value other than zero.

- **The Mechanism:** If a metric moves from 100 to 105, a chart starting at 0 shows a modest 5% rise. A chart starting at 95 shows a line shooting up at a 45-degree angle, visually suggesting a massive 100% increase in the "visible" space.
- **The Distortion:** It exaggerates volatility. Minor, insignificant fluctuations look like boom-and-bust cycles. This is often used in cable news or financial reporting to create drama where there is none.

2. Dual Y-Axes

This involves plotting two different variables with different scales on the same chart (e.g., "Ice Cream Sales" on the left axis and "Shark Attacks" on the right axis).

- **The Mechanism:** By adjusting the scales of the two axes independently, you can force the two lines to overlap perfectly or cross at specific points.
- **The Distortion:** It implies causality. Because the lines move together, the brain assumes A causes B. In reality, you likely just tweaked the scales until the lines matched. This is notorious for creating "spurious correlations."

3. Inconsistent Time Intervals

Plotting data points at uneven intervals (e.g., Jan 1, Jan 2, Mar 1, Dec 1) but spacing them equally on the X-axis destroys the concept of "rate of change." A slow, gradual decline over 6 months will look identical to a sharp drop over 2 days if the axis spacing isn't proportional to time.

Real-World Data Science Scenario:

Imagine someone is leading a team developing a Reinforcement Learning (PPO) agent for a game.

Your team has a baseline model with a win rate of 52.0%. A junior data scientist trains a new experimental model that achieves 52.3%.

The Distortion:

To present their results, they create a line chart of the "Win Rate over Training Steps."

1. **Truncated Axis:** They set the Y-axis range to [51.5%, 52.5%].
2. **Visual Result:** The chart shows a steep, rocket-ship trajectory upwards. The gap between the baseline and the new model looks like a massive chasm.

The Reality:

If plotted on a scale of [0%, 100%], the line is effectively flat. The 0.3% gain is likely statistical noise or within the margin of error for a stochastic algorithm like PPO.