ÉCOLE SUPÉRIEURE EN SCIENCES ET TECHNOLOGIES DE L'INFORMATIQUE ET DU NUMÉRIQUE

# FUNDAMENTALS OF DATA SCIENCE AND DATA MINING

## CHAPTER 4:

## DATA VISUALIZATION

Dr. Chemseddine Berbague

2024-2025

# CONTENT

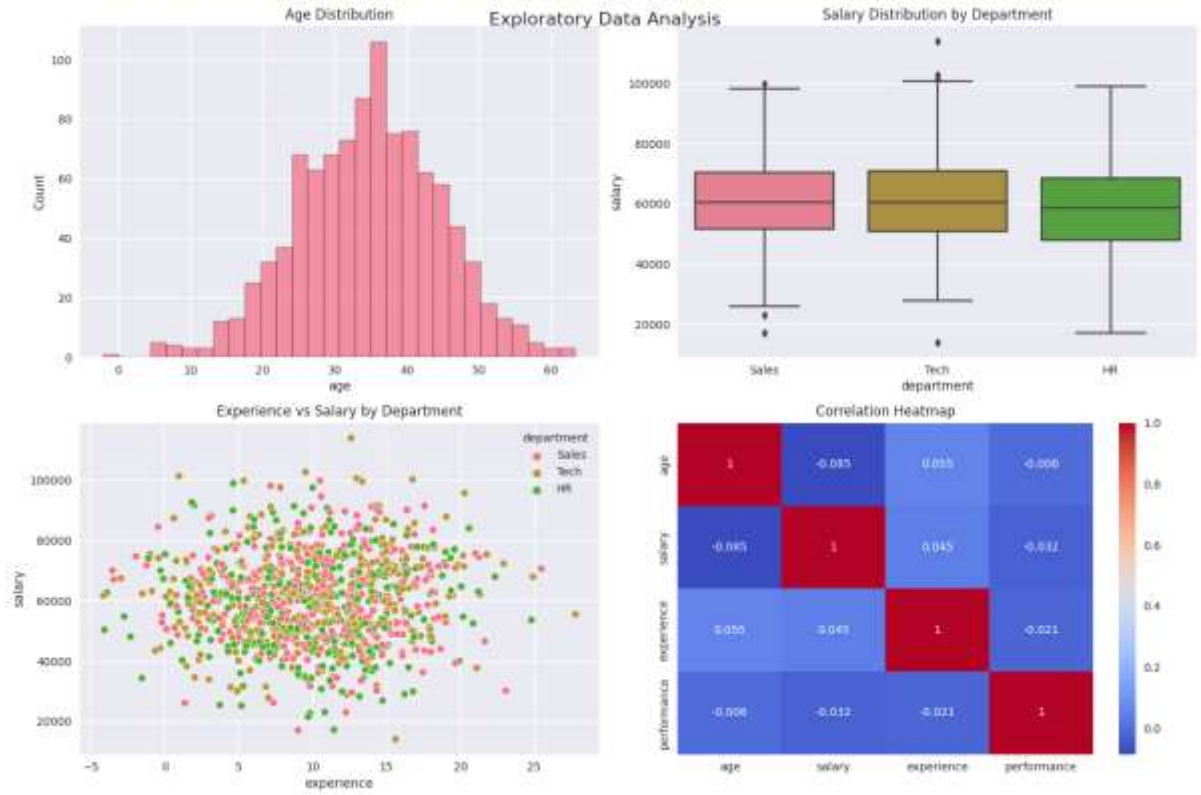
- Introduction to data visualization
- Functions of data visualization
- Principals of data visualization
  - Type-driven data visualization
  - Purpose-driven data visualization
  - Advanced data visualization
- Tools of data visualization
- Case study
- Annex of different visualizations

# INTRODUCTION

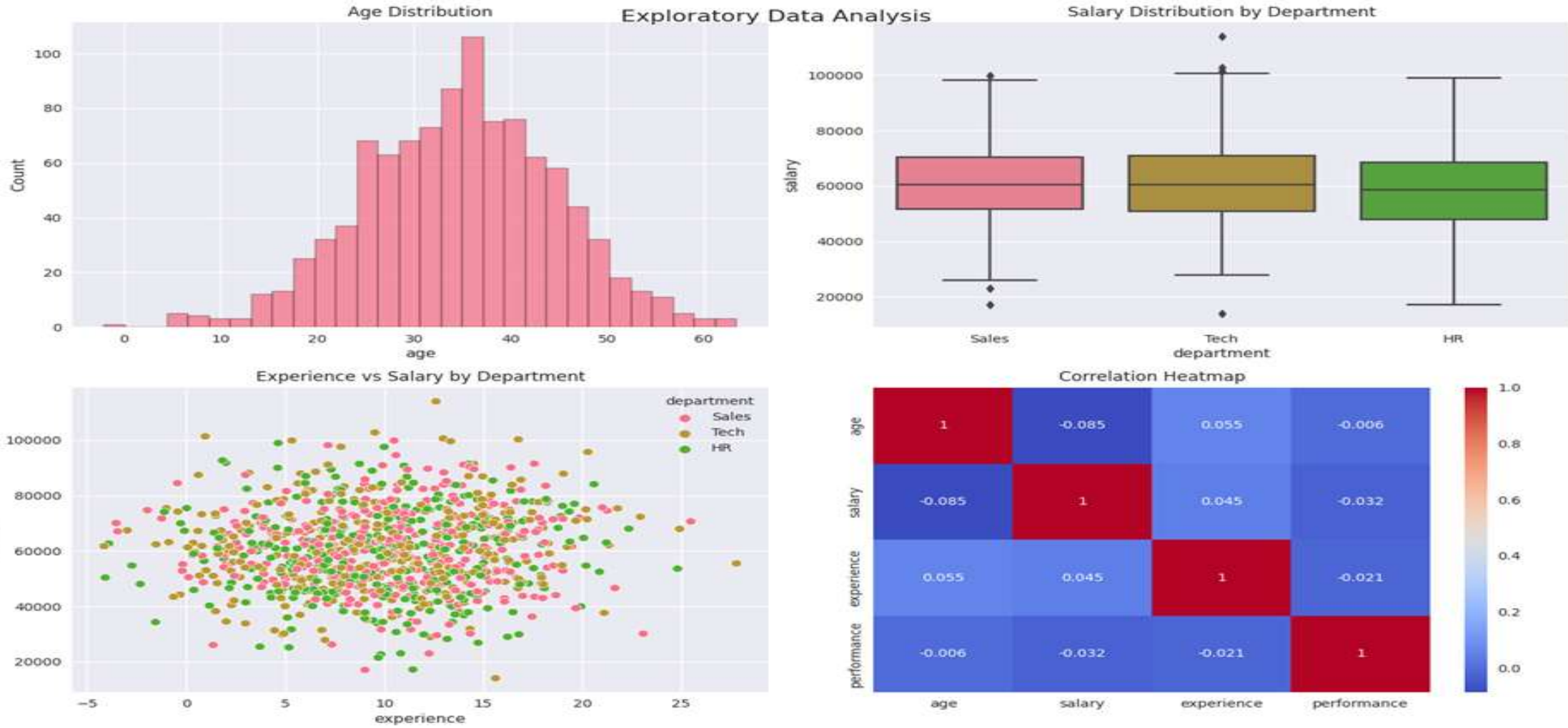
- **Data visualization** is a critical tool for transforming complex datasets into clear, understandable, and actionable insights. It allows data scientists to explore, interpret, and communicate the findings from data analysis through visual means, enabling better decision-making and storytelling.
- Data visualization serves several essential functions in data science:
  - **Exploratory Analysis**: Visualization helps data scientists uncover hidden patterns, trends, and relationships within large datasets. This exploratory phase is crucial in understanding the data before applying advanced analytical techniques.
  - **Model Evaluation**: By visualizing model performance metrics, such as error distributions or feature importance, data scientists can quickly assess model effectiveness and diagnose issues.
  - **Insight Communication**: Visualization bridges the gap between technical analysis and business decision-making, allowing data scientists to present complex results in an easily understandable format.

# FUNCTIONS OF DATA VISUALIZATION: EXPLORATORY ANALYSIS

- Exploratory Analysis:
  - Histograms for distribution analysis
  - Box plots for comparing groups
  - Scatter plots for relationships
  - Correlation heatmaps

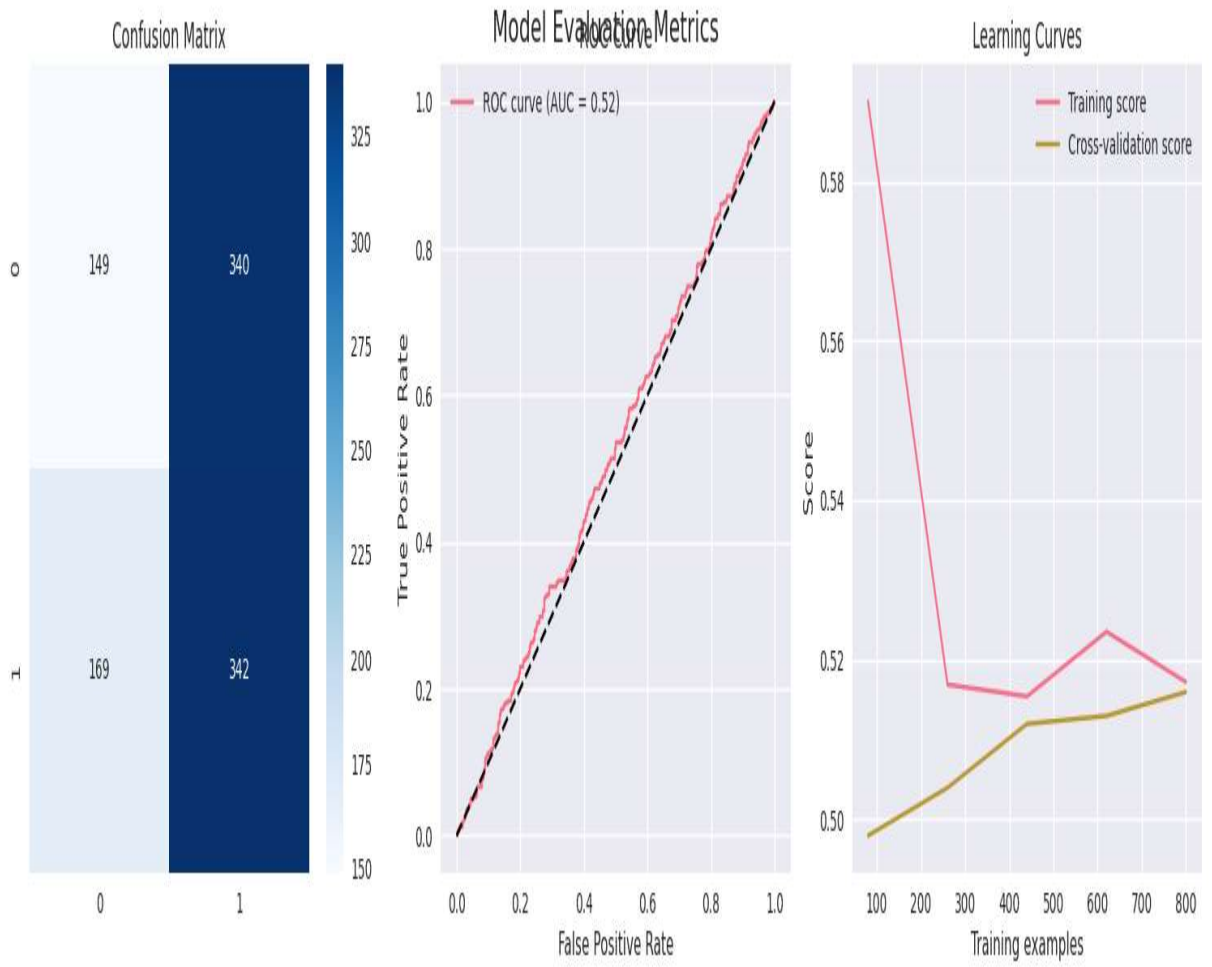


# FUNCTIONS OF DATA VISUALIZATION: EXPLORATORY ANALYSIS

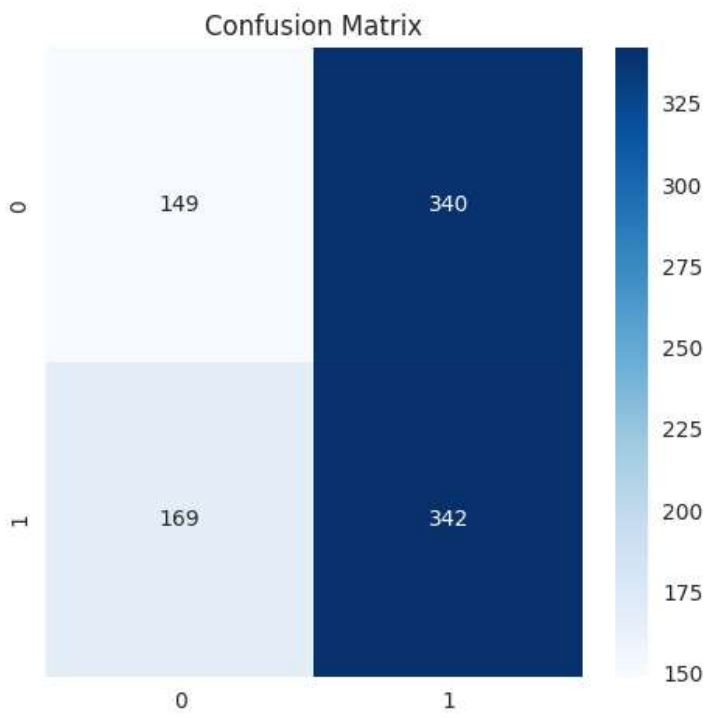


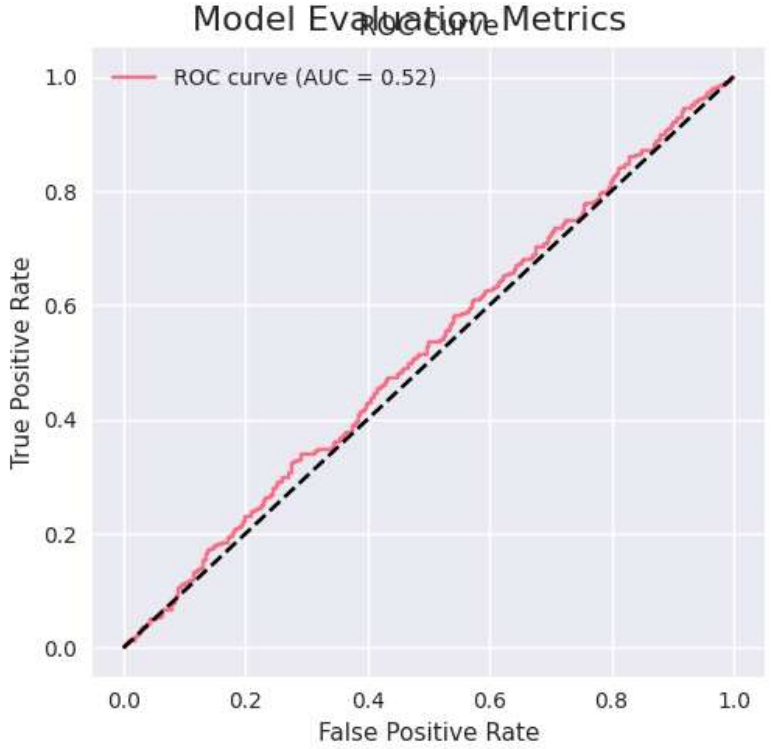
# FUNCTIONS OF DATA VISUALIZATION: MODEL EVALUATION

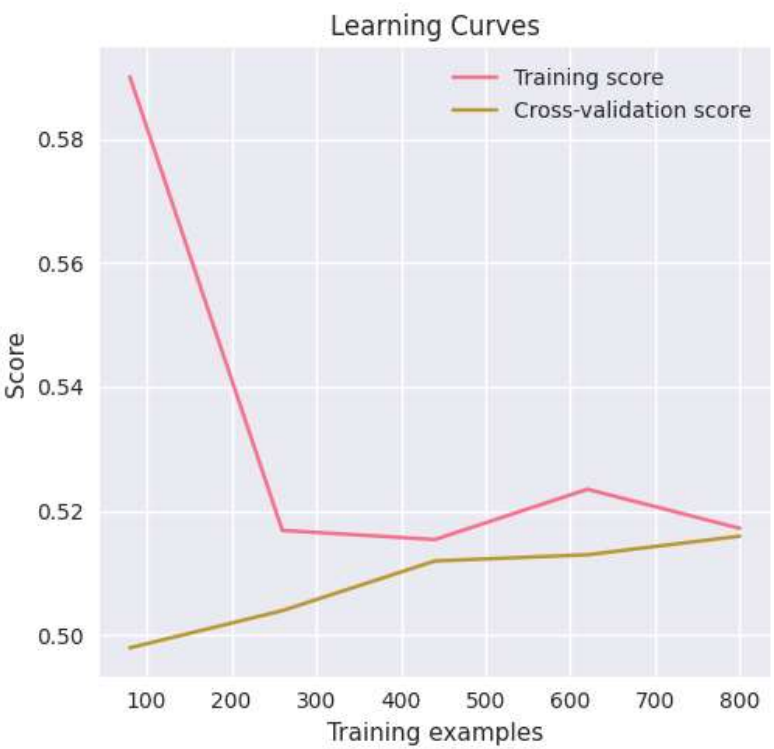
- Model Evaluation:
  - Confusion matrix heatmap
  - ROC curve with AUC
  - Learning curves



# FUNCTIONS OF DATA VISUALIZATION: MODEL EVALUATION

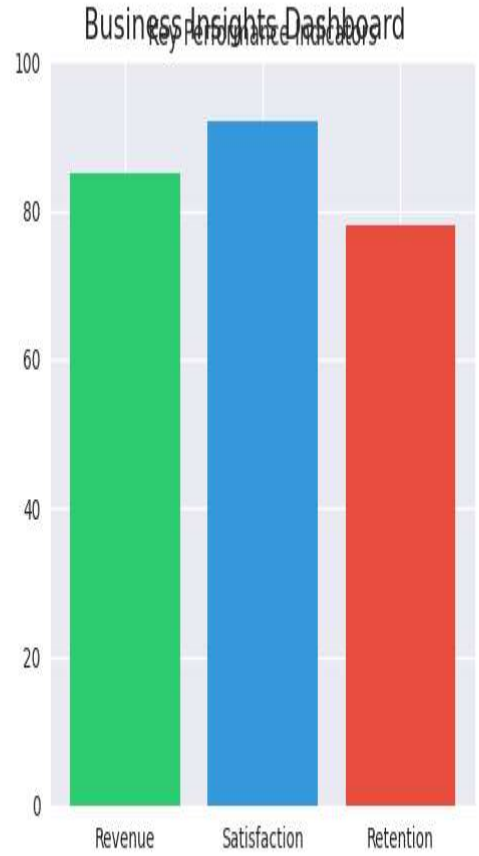


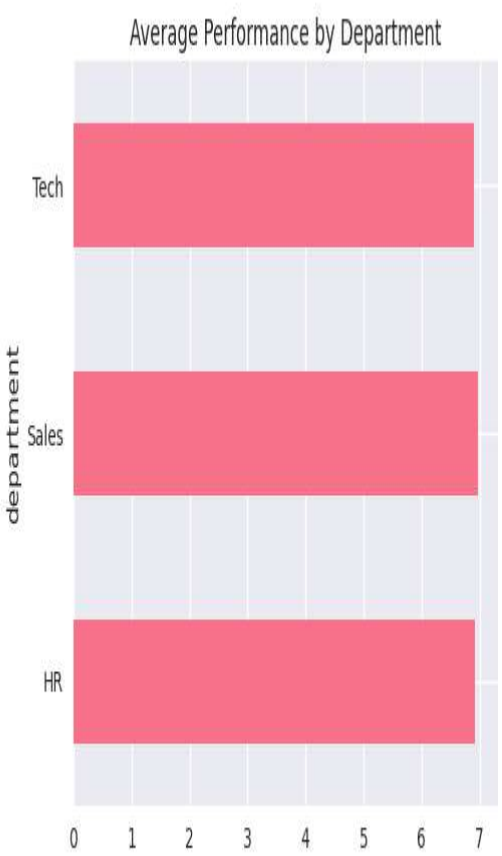



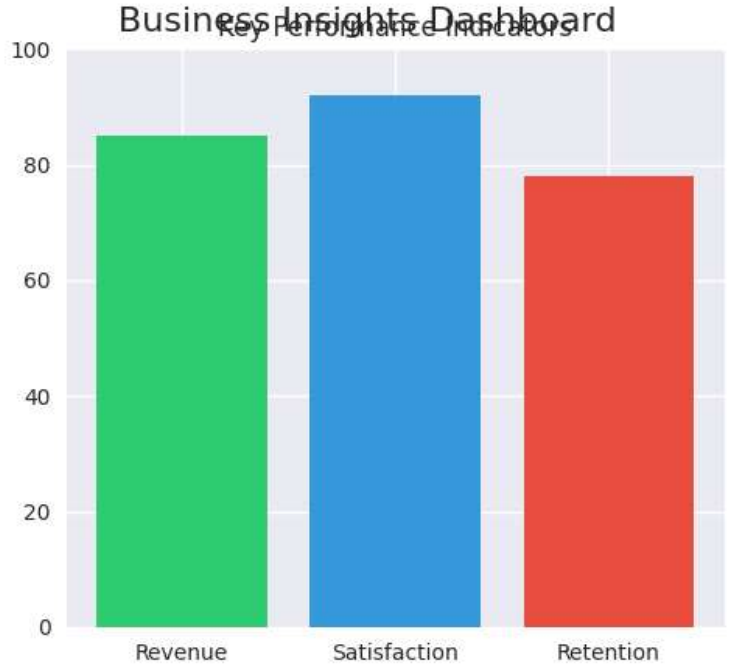
# FUNCTIONS OF DATA VISUALIZATION: MODEL EVALUATION

- Insight Communication:
  - Time series trends
  - KPI dashboard
  - Department performance comparison

# FUNCTIONS OF DATA VISUALIZATION: MODEL EVALUATION

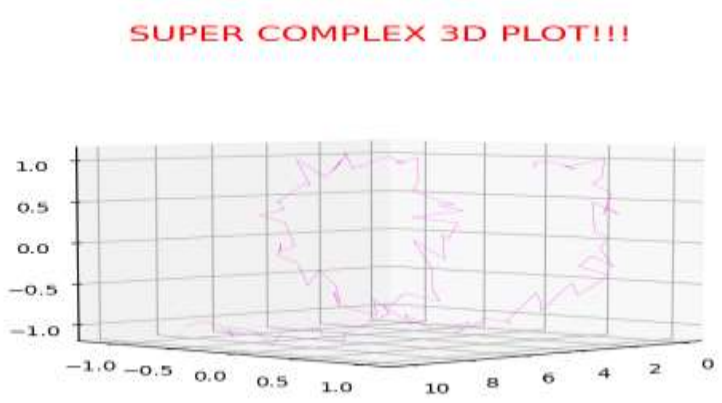

# DATA VISUALIZATION: PRINCIPALS

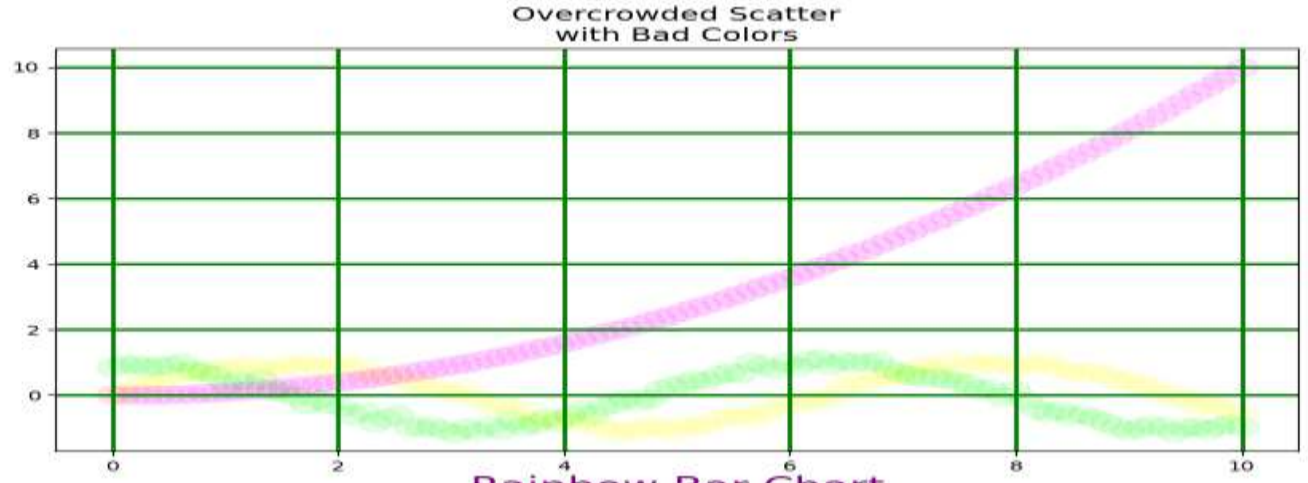
- **1. Clarity and Simplicity**
  - **Principle**: The visualization should clearly and simply convey its intended message without unnecessary elements.
  - **Purpose**: To ensure that the audience quickly grasps the main insights without being distracted.
  - **Example**: Use a bar chart to compare categories rather than an overly complex 3D chart, as it enhances clarity.
- **2. Choosing the Right Visualization Type**
  - **Principle**: Match the type of visualization with the nature of the data and the purpose of the analysis.
  - **Purpose**: Different chart types serve distinct purposes. Choosing the right type enhances the data's message.
  - **Examples**:Use **line charts** for trends over time.
    - Use **scatter plots** to show relationships between variables.
    - Use **heatmaps** to reveal patterns in large datasets.
- **3. Accuracy and Integrity**
  - **Principle**: Visualizations should accurately represent the data without distorting its meaning.
  - **Purpose**: Misleading visualizations can lead to incorrect conclusions; accuracy builds trust.
  - **Example**: Use a consistent scale on the y-axis to prevent exaggeration of small differences.
- **4. Focus on the Key Message**
  - **Principle**: Highlight the most critical data points or trends to guide the audience's attention.
  - **Purpose**: The audience should easily identify the primary insight.
  - **Example**: Emphasize important data points with contrasting colors or labels to make them stand out.

# DATA VISUALIZATION: PRINCIPALS

- **5. Effective Use of Color, Size, and Visual Attributes**
  - **Principle**: Use colors, sizes, and other attributes thoughtfully to enhance comprehension without causing distraction.
  - **Purpose**: Color and size should reinforce the data's meaning, not overwhelm or mislead.
  - **Example**: Use a single color gradient to show a range of values rather than multiple colors, which can be confusing.
- **6. Consistency and Readability**
  - **Principle**: Ensure a consistent design and clear labeling across visualizations.
  - **Purpose**: Consistency aids readability and helps the audience interpret visuals faster.
  - **Example**: Maintain consistent font sizes, chart styles, and label formats across a series of visualizations.
- **7. Encourage Exploration (when appropriate)**
  - **Principle**: In interactive or complex visualizations, allow the audience to explore the data at their own pace.
  - **Purpose**: Engaging visuals can deepen understanding, especially in dashboards or analytical reports.
  - **Example**: Use filters, zooming, or hover-over details in interactive charts to help the audience explore the data.
- **8. Avoid Common Pitfalls (Bias, Distortion, and Overloading)**
  - **Principle**: Avoid biases, data distortions, and information overload by simplifying visuals and avoiding misleading elements.
  - **Purpose**: Ensures that the visualization is ethical and communicates an honest representation of the data.
  - **Example**: Avoid using pie charts with too many slices or complex 3D graphics that distort perception.

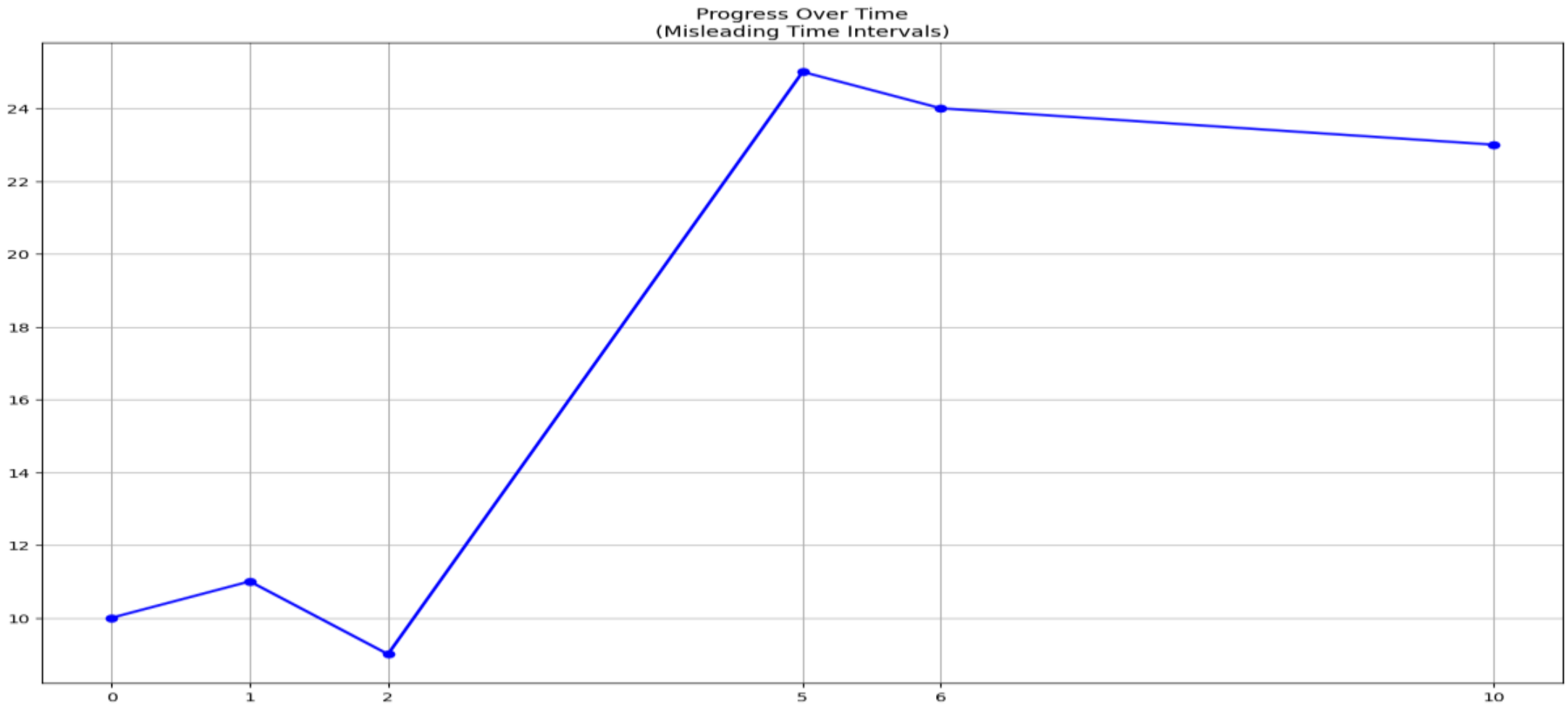
# DATA VISUALIZATION: PRINCIPALS



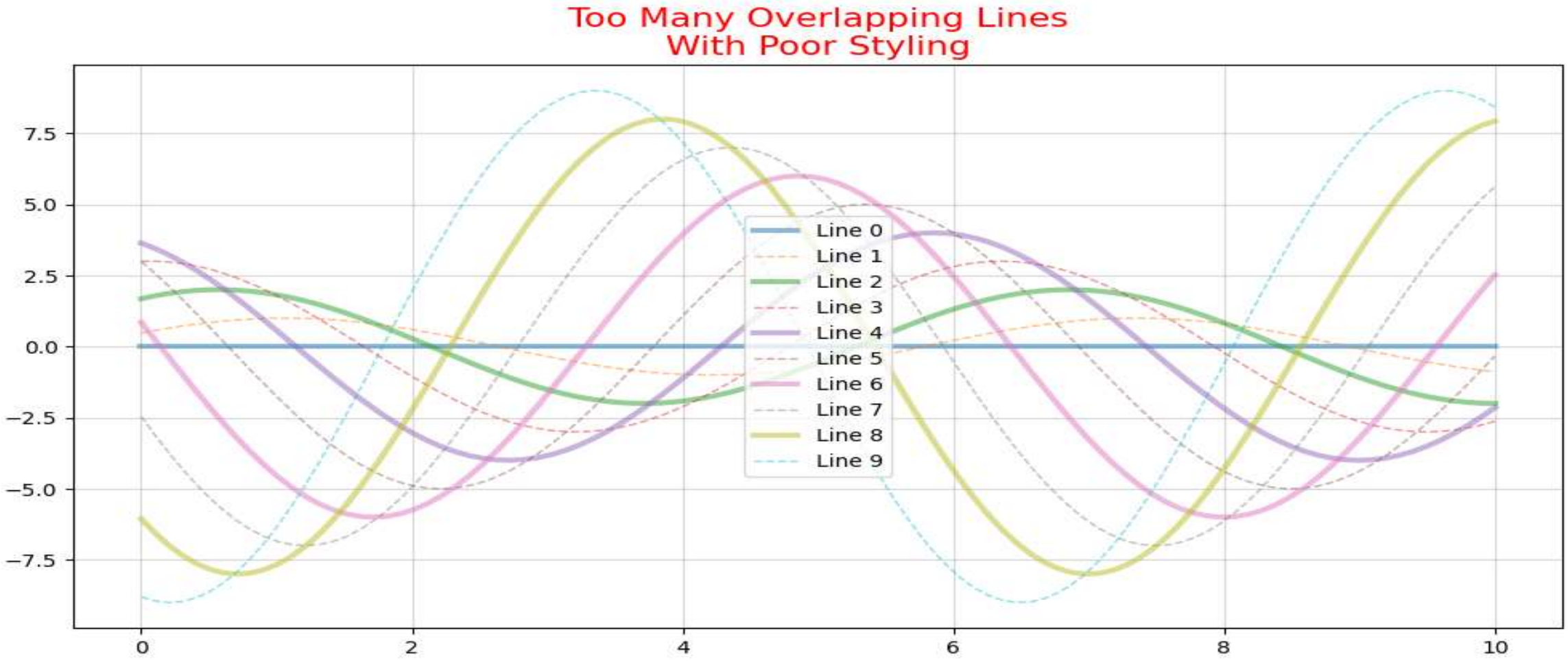





# DATA VISUALIZATION: PRINCIPALS



# DATA VISUALIZATION: PRINCIPALS

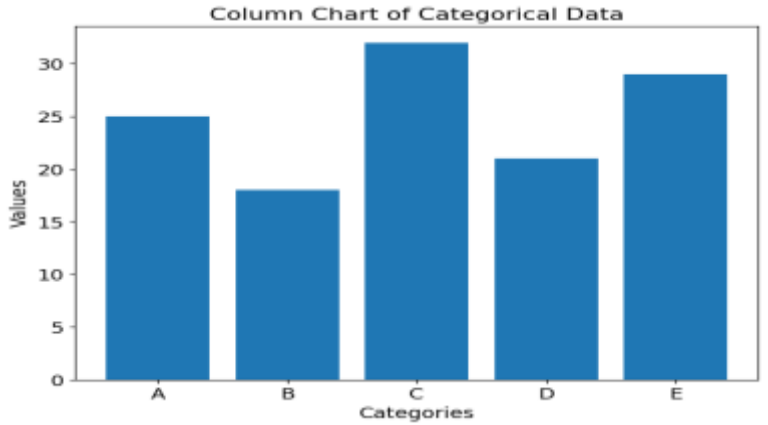


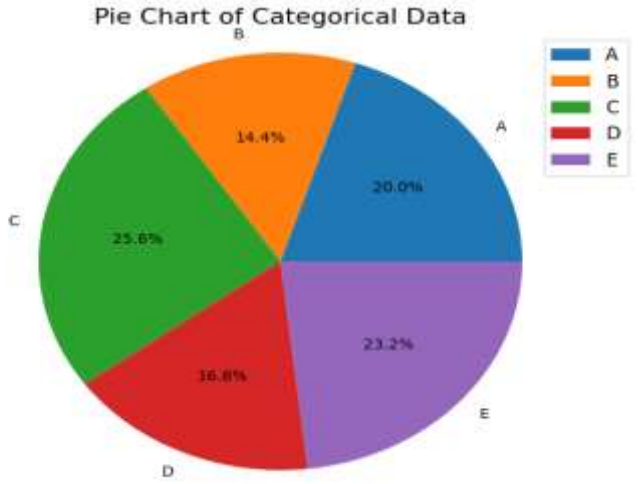
# SELECTING DATA VISUALIZATION TYPE

- Selecting the appropriate chart or graph depends heavily on:
  - the nature of the data
  - , and the message we aim to convey (purpose)
- Different visualization types excel at highlighting specific aspects of data, such as comparisons, trends, distributions, and relationships.

# TYPE–DRIVEN DATA VISUALIZATION CHOICES

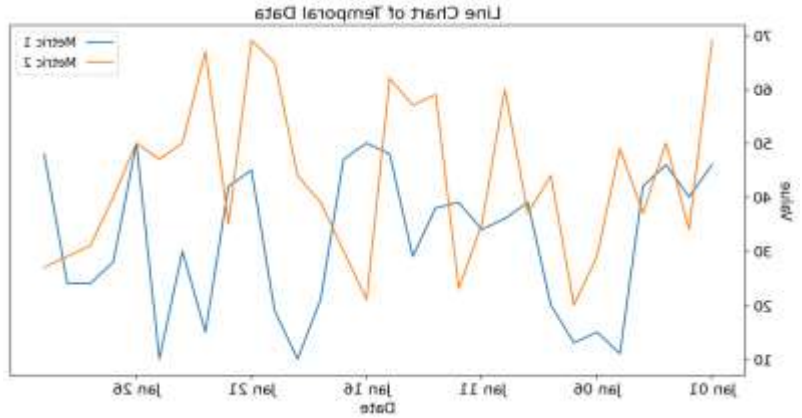
- **Categorical Data (e.g., product types, regions):**
  - Often used to compare quantities across distinct categories.
  - Best Visualizations: Bar charts, column charts, and pie charts (for proportions).
  - These provide a clear view of differences or shares between categories.

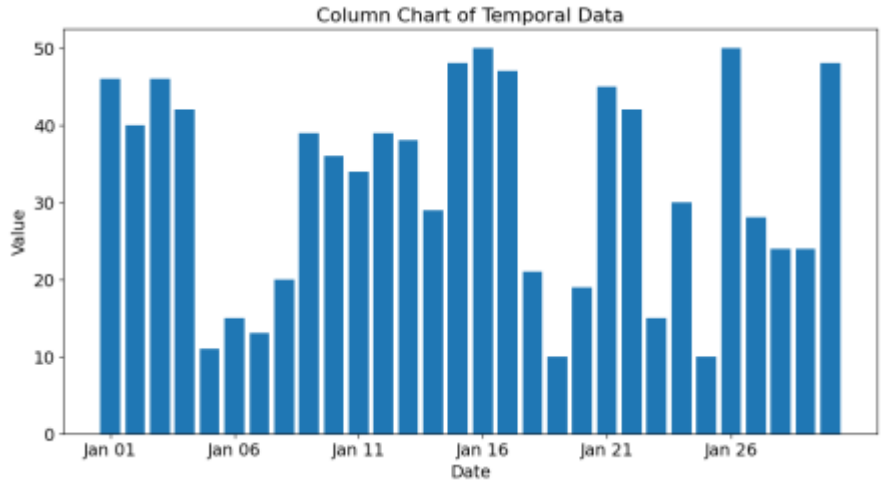


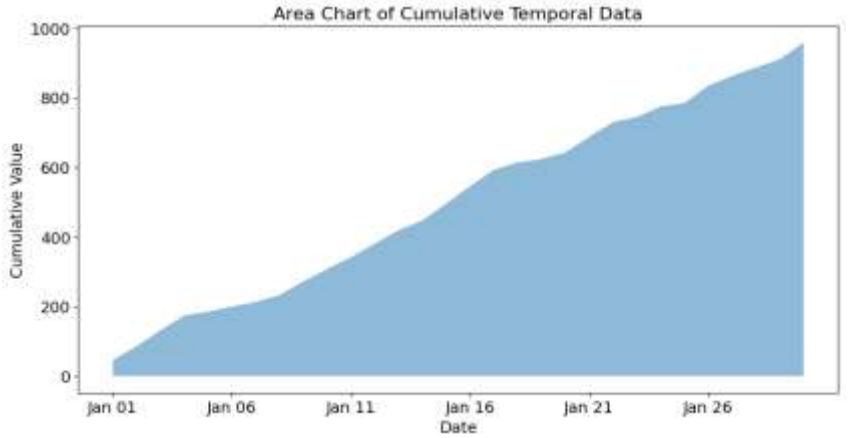



# TYPE–DRIVEN DATA VISUALIZATION CHOICES

- **Time-Series Data (e.g., monthly sales, stock prices):**
  - Shows how a variable changes over a specified period.
  - Best Visualizations: Line charts for continuous time tracking, area charts for cumulative totals, and column charts for discrete time intervals.
  - These visualizations help reveal trends, cycles, and seasonality.

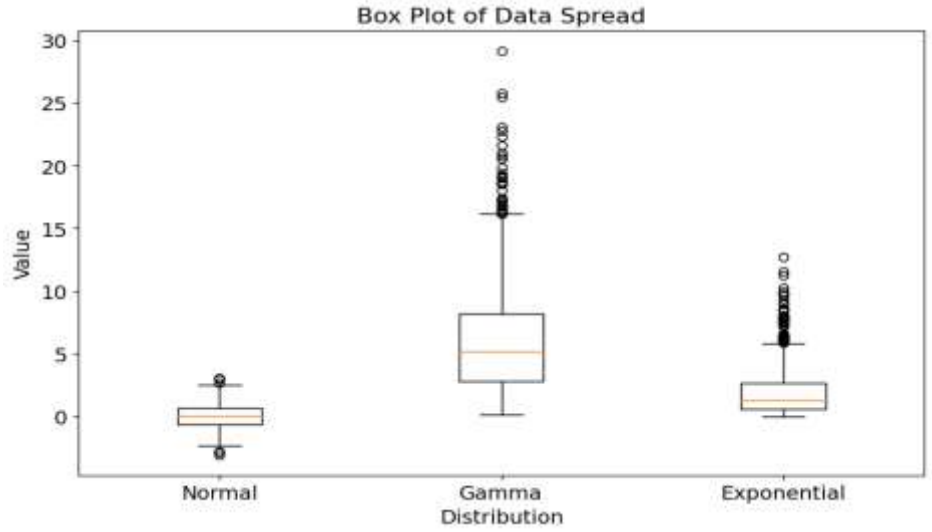


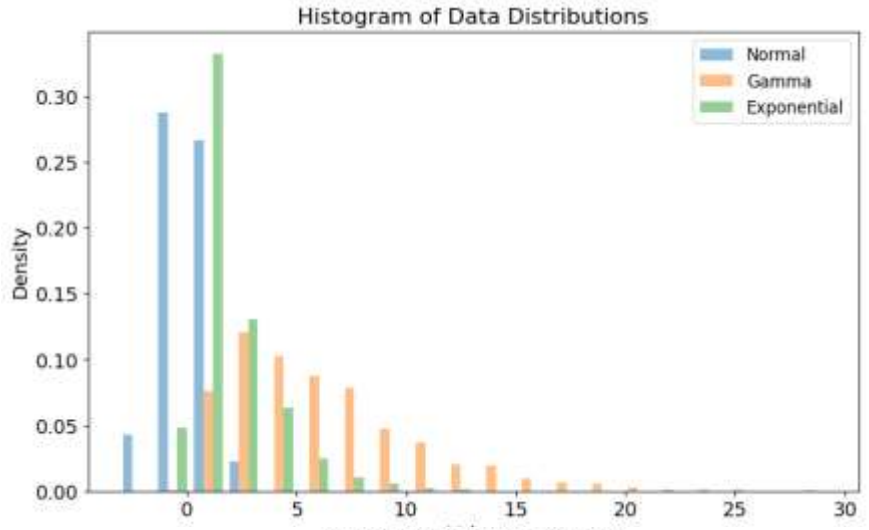


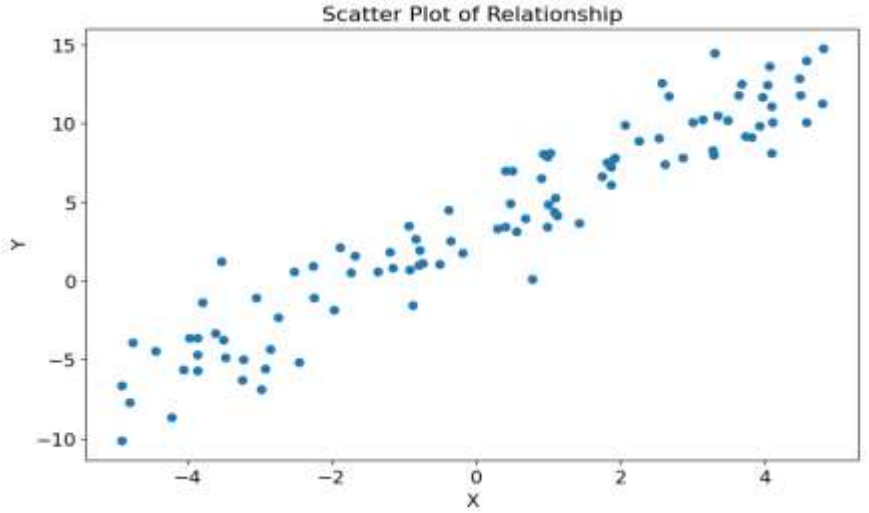



# TYPE –DRIVEN DATA VISUALIZATION CHOICES

- **Quantitative Data (e.g., temperature, revenue):**
  - Continuous or discrete numerical values that need to show ranges, distributions, or precise values.
  - Best Visualizations: Histograms for distributions, scatter plots for relationships, and box plots for data spread and outliers.
  - Quantitative visualizations are essential for exploring variability and statistical properties.

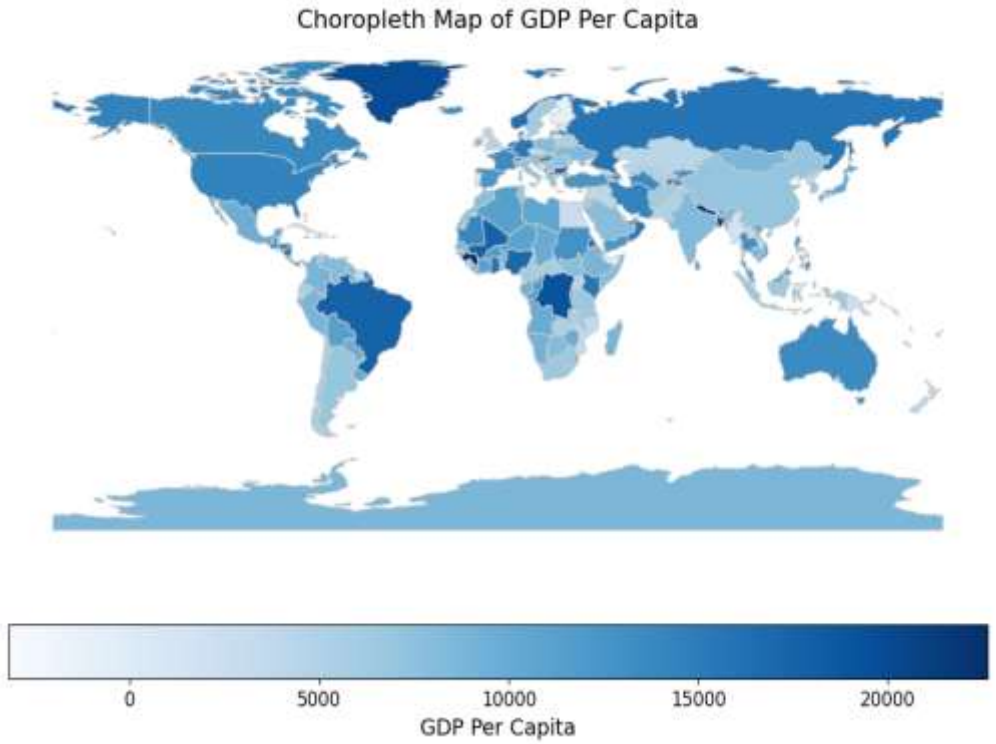


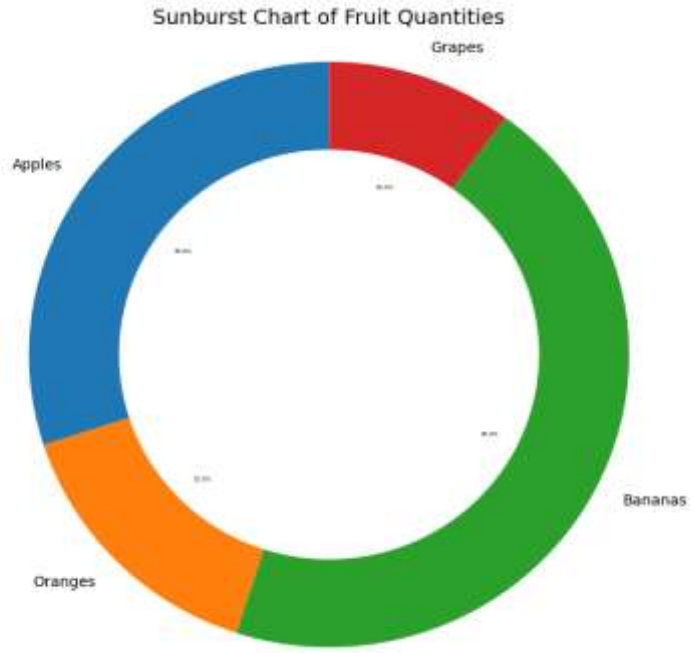
# TYPE –DRIVEN DATA VISUALIZATION CHOICES

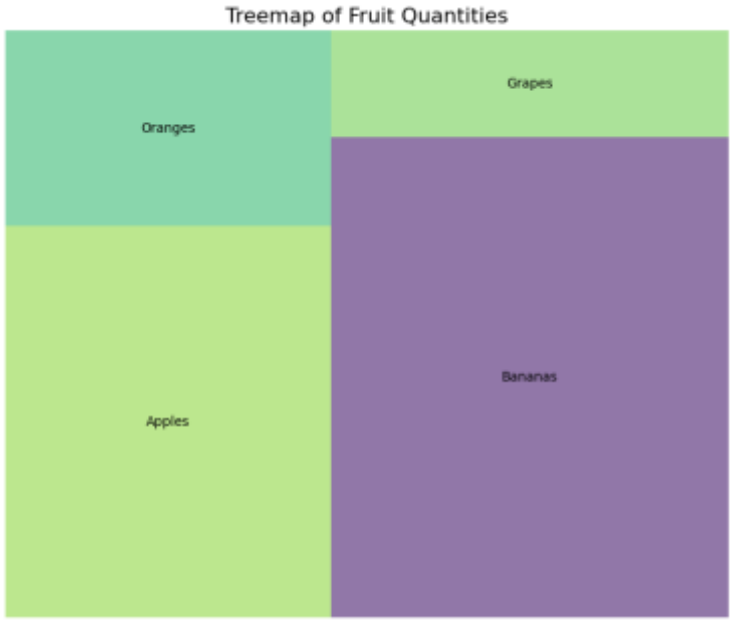
- **Geospatial Data (e.g., customer distribution by region):**
  - Requires showing data across geographical locations.
  - Best Visualizations: Maps (such as choropleth maps for intensity and bubble maps for point-specific data).
  - These enable spatial understanding, showing variations in location-based data.



# TYPE –DRIVEN DATA VISUALIZATION CHOICES

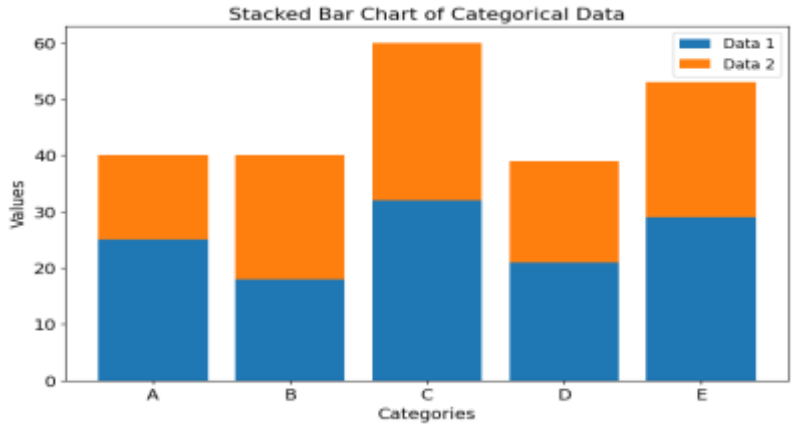
- **Hierarchical Data** (e.g., organizational structures, category breakdowns): Data with nested levels or groupings.
  - **Best Visualizations**: Tree maps, sunburst charts, and dendrograms. Hierarchical visuals are effective in showcasing part-to-whole relationships within nested categories.

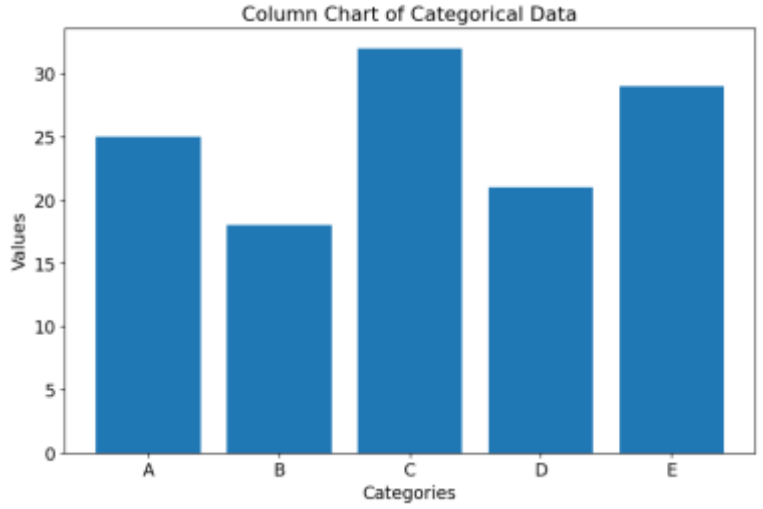


# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

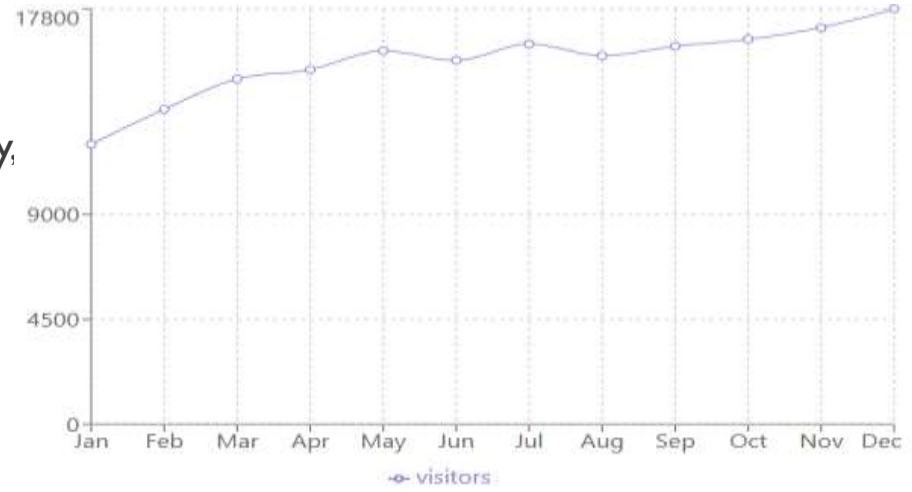
- **1. Comparing Values Across Categories**
  - **Purpose**: To compare quantities or values across different categories or groups.
  - **Recommended Visualizations**:
    - **Bar Charts**: Ideal for comparing discrete categories, such as sales by product category, revenue by region, or survey responses.
      - Example: A bar chart showing quarterly revenue across different product lines helps quickly compare which products perform best.
    - **Column Charts**: Similar to bar charts, but vertical; useful when categories represent time periods.
      - Example: Monthly sales figures for the past year, with each month represented as a column, make it easy to see seasonal trends.
    - **Stacked Bar/Column Charts**: Show parts of a whole within categories while comparing category totals.
      - Example: Sales contribution by different departments within each region.
- **When to Use**: When comparing distinct, categorical values, especially where showing relative or absolute differences is essential.



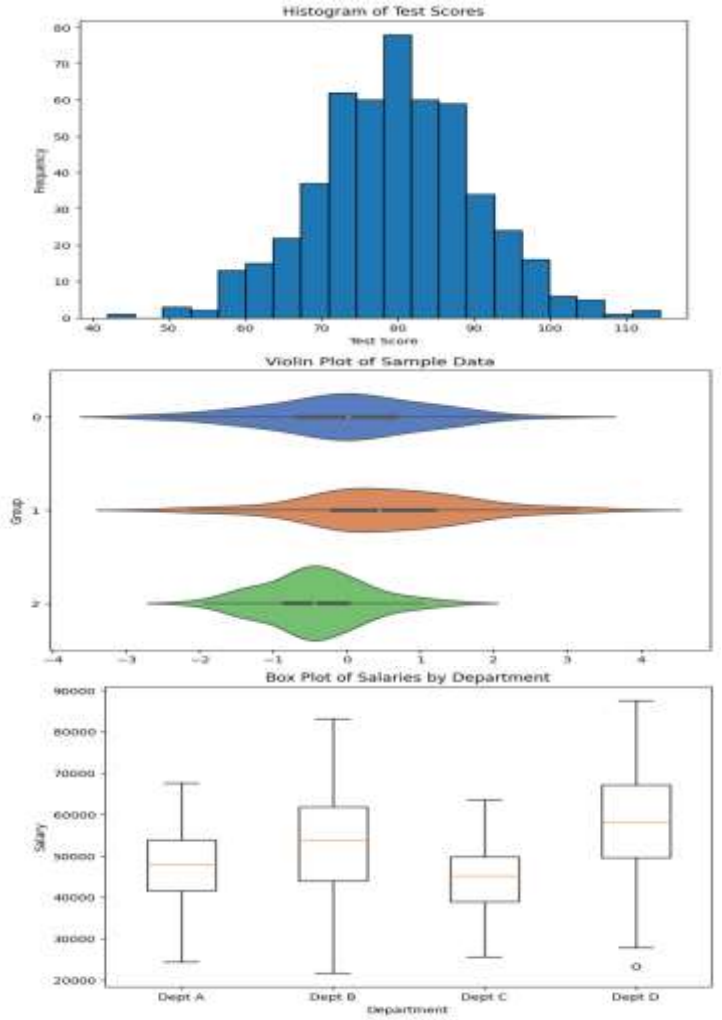
# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **2. Showing Trends Over Time**
  - **Purpose**: To show how values change over time, capturing trends, seasonality, or patterns.
  - **Recommended Visualizations**:
    - **Line Charts**: The most common choice for time-series data, as they clearly display changes over intervals.
      - Example: A line chart showing website traffic over the last year helps visualize trends, spikes, and dips.
    - **Area Charts**: Good for showing cumulative values over time, with the area under the line indicating volume.
      - Example: Monthly sales growth, where the area chart reveals both the growth rate and total cumulative sales.
    - **Dual-Axis Charts**: Useful when comparing two related variables over time.
      - Example: Revenue and advertising expenditure over time to identify how spending impacts income.
- **When to Use**: When time is a key variable, as these charts highlight trends, making it easy to analyze growth, decline, or cycles.

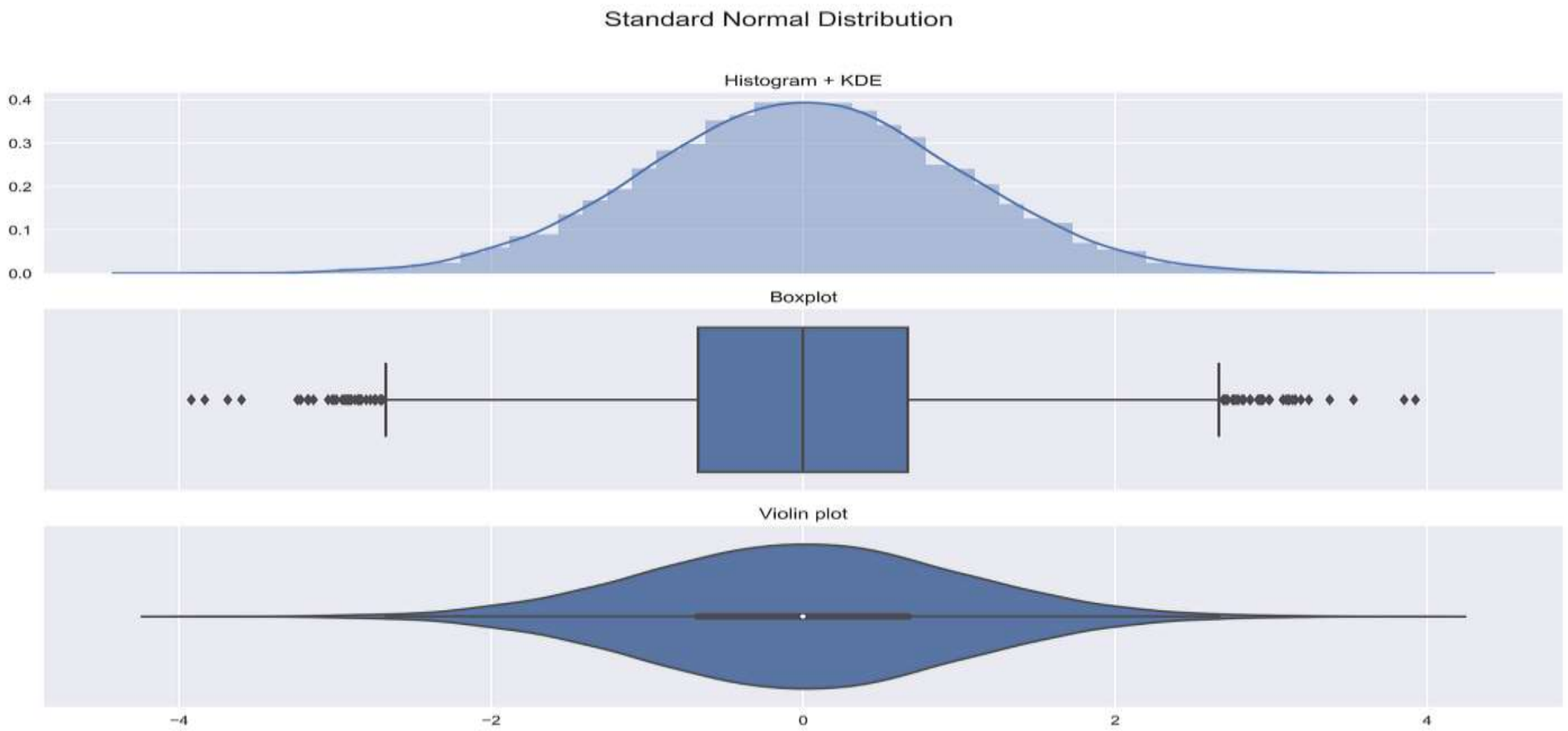



# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **3. Displaying Distributions of a Variable**
  - **Purpose**: To show the spread or concentration of data points, helping understand data variability, central tendency, and outliers.
  - **Recommended Visualizations**:
    - **Histograms**: Ideal for continuous data; bins group values to show the frequency of observations within each range.
      - Example: A histogram showing test scores helps reveal whether scores cluster around the mean or if there's a wide spread.
    - **Violin Plots**: Combine box plots and density plots to show the data's distribution shape and density.
      - Example: Visualizing customer purchase frequency to understand which segments make more frequent purchases.
    - **Box Plots**: Display data distributions by quartiles, useful for spotting outliers and comparing distributions.
      - Example: Box plots of salaries in different departments can show disparities in pay and identify outliers.
- **When to Use**: When you need to understand the range, central tendencies, and spread of the data or to identify patterns such as skewness or outliers.

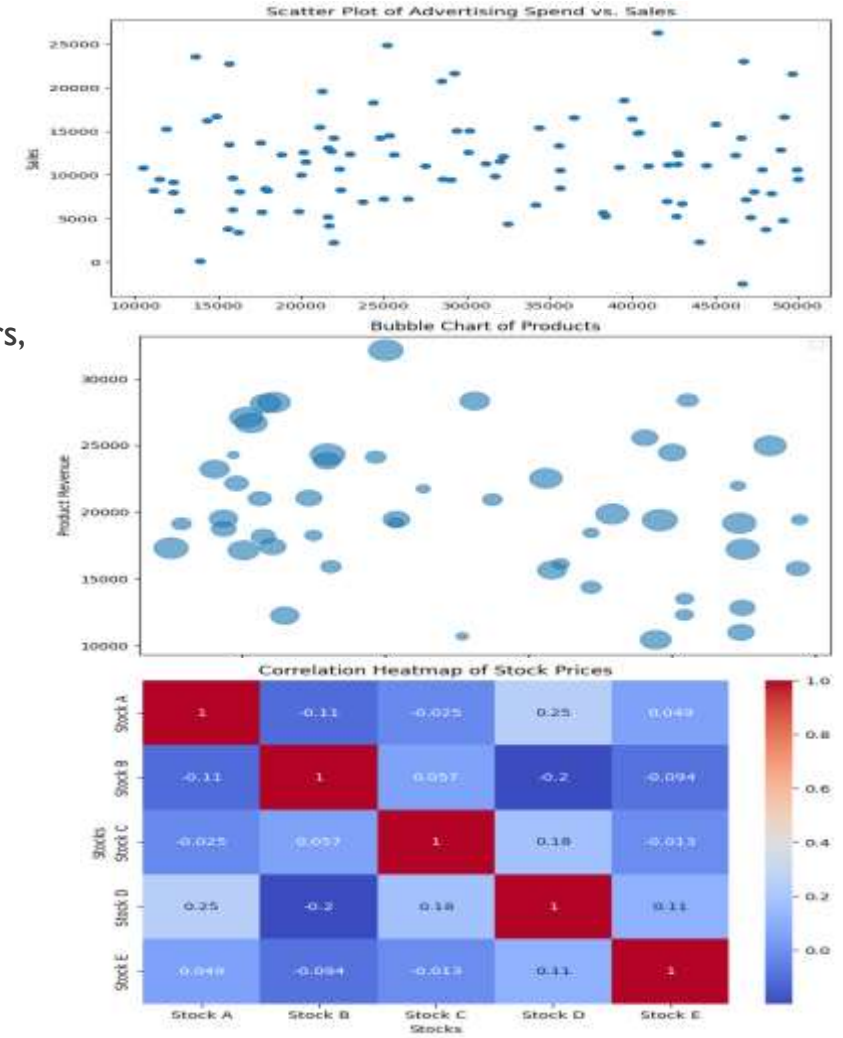


# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES



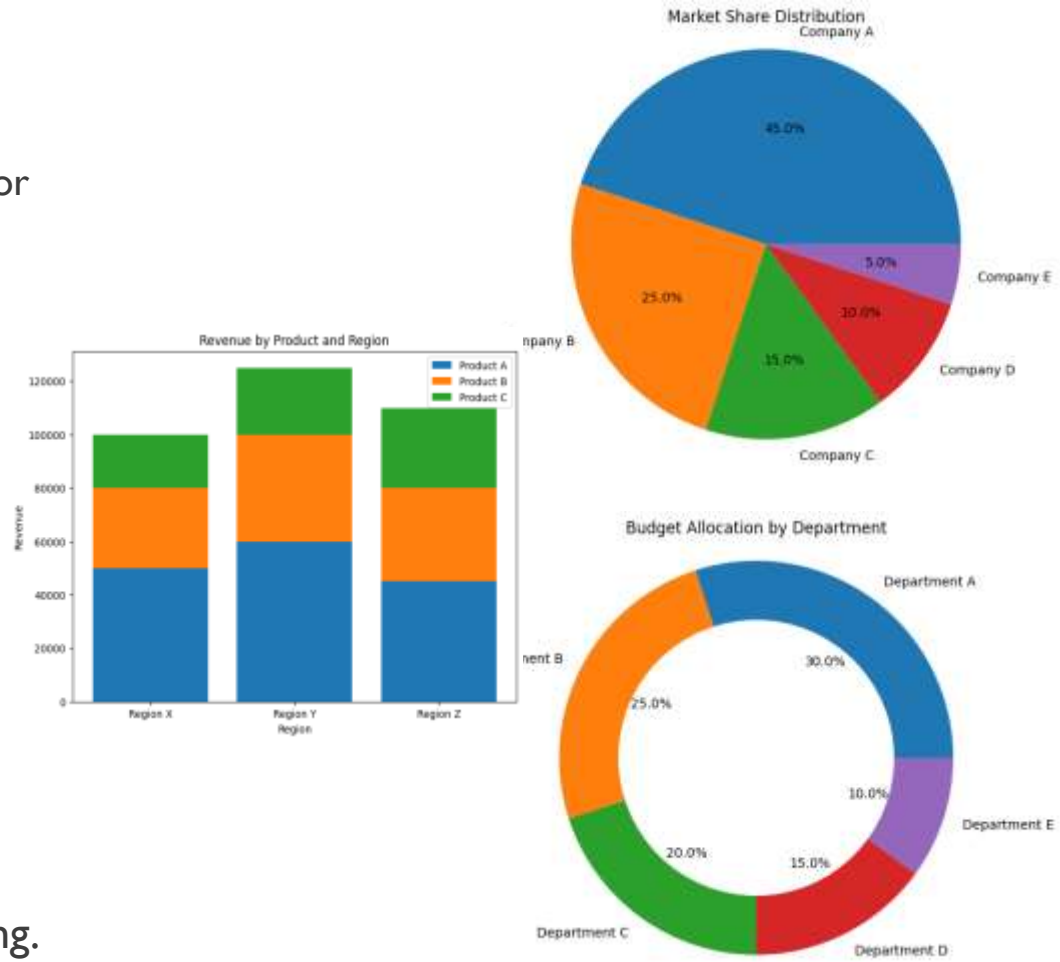
# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **4. Examining Relationships Between Variables**
  - **Purpose**: To analyze how two or more variables interact, identifying correlations or patterns.
  - **Recommended Visualizations**:
    - **Scatter Plots**: Show relationships between two numerical variables, helping spot trends, clusters, or correlations.
      - Example: A scatter plot of advertising spend vs. sales to explore if more spending correlates with increased sales.
    - **Bubble Charts**: An extension of scatter plots where a third variable is represented by the size of bubbles.
      - Example: Each bubble represents a product, with axes showing cost and revenue, and bubble size representing the number of units sold.
    - **Heatmaps**: Represent relationships in tabular data, with color intensity showing the value magnitude.
      - Example: A correlation heatmap for various stock prices highlights which stocks move similarly.
- **When to Use**: When exploring potential interactions or relationships between two or more variables, especially if looking for trends or patterns.



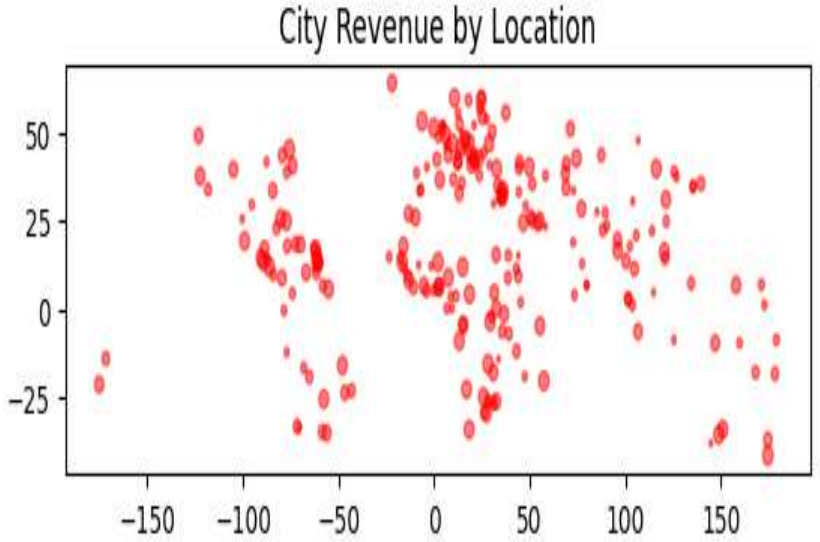
# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **5. Displaying Parts of a Whole**
  - **Purpose**: To show how parts contribute to a total, illustrating proportions or shares.
  - **Recommended Visualizations**:
    - **Pie Charts**: Show percentage breakdowns within a whole. Best for limited categories.
      - Example: Market share distribution among competitors.
    - **Donut Charts**: Similar to pie charts but with a hollow center, adding focus on the total while showing parts.
      - Example: Proportion of budget allocation across departments.
    - **Stacked Bar/Column Charts**: Useful when the breakdown is across multiple categories.
      - Example: Revenue contributions from various products across several regions.
- **When to Use**: For simple, proportional comparisons, but avoid when there are too many categories, as it can become cluttered and confusing.



# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **6. Geospatial Data Visualization**
  - **Purpose**: To represent data that has geographical or spatial components, useful for showing distribution or density across locations.
  - **Recommended Visualizations**:
    - **Maps**: Heat maps, choropleth maps, or dot maps are commonly used for showing regional differences.
      - Example: Population density map by region or customer distribution map across cities.
    - **Bubble Maps**: Use bubbles to indicate magnitude or volume of a variable at specific locations.
      - Example: Revenue generated by different store locations.
- **When to Use**: When location is an important dimension and you need to show spatial patterns or distributions.



# PURPOSE-DRIVEN DATA VISUALIZATION CHOICES

- **7. Displaying Hierarchical Data**
  - **Purpose**: To present data with a hierarchical structure, often in nested categories.
  - **Recommended Visualizations**:
    - **Tree Maps**: Represent data in nested rectangles, with size and color representing metrics.
      - Example: Revenue breakdown by product categories and subcategories.
    - **Sunburst Charts**: Show hierarchical levels in concentric circles.
      - Example: Organizational structure from company to department level.
- **When to Use**: When there's a need to show both the hierarchy and quantitative size of categories and subcategories.

# ADVANCED DATA VISUALIZATION TECHNIQUES

- Advanced visualization techniques go beyond basic charts and graphs to provide richer, more detailed views of data.
- By exploring multiple dimensions, analyzing trends over time, or enabling interactive elements, these methods help users gain a nuanced and comprehensive understanding of their data.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **MULTIVARIATE VISUALIZATIONS**

- **Multivariate Visualizations**
  - **Definition**: Multivariate visualizations display multiple variables or dimensions within a single visualization. These are useful for exploring complex datasets and for understanding relationships between several variables simultaneously.
  - **Purpose**: Multivariate techniques help uncover patterns and correlations in datasets with multiple variables, enabling deeper insights. They are commonly used when two-dimensional plots (e.g., simple bar or line charts) are insufficient for showing the full scope of the data.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **MULTIVARIATE VISUALIZATIONS**

- **Examples of Multivariate Visualizations**:
  - **Scatter Plot Matrix**: Displays multiple scatter plots for pairs of variables, allowing quick identification of correlations and trends between variables.
  - **Parallel Coordinates Plot**: Each variable is shown as a vertical axis, with lines connecting data points across these axes. It's ideal for comparing individual data points across several dimensions.
  - **Heatmaps with Annotations**: Shows data intensity or magnitude across two variables, with color variations indicating value differences and annotations providing context.
  - **Bubble Chart**: Extends a scatter plot by adding a third dimension through bubble size or color, making it useful for exploring relationships in three-variable datasets.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **TEMPORAL VISUALIZATION TECHNIQUES**

- **Temporal Visualization Techniques**
  - **Definition**: Temporal visualizations focus on changes and trends over time. These techniques are essential for analyzing time series data, where understanding the temporal progression of values is key.
  - **Purpose**: Temporal visualizations are used to track trends, cycles, and seasonality, making it easier to see how data points evolve over time. They are commonly applied in fields like finance, sales, and climatology.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **TEMPORAL VISUALIZATION TECHNIQUES**

- **Examples of Temporal Visualizations**:
  - **Line Chart**: The most straightforward way to display changes over time; it shows a continuous trend and highlights increases or decreases in data values.
  - **Time-Series Heatmaps**: Represent time on one axis and another variable on the other axis, with color variations showing intensity. They're useful for visualizing seasonal or periodic trends.
  - **Gantt Chart**: Primarily used in project management, Gantt charts show tasks over time, with each task represented by a horizontal bar whose length represents duration.
  - **Stacked Area Chart**: Used to display the cumulative contribution of multiple categories over time, highlighting both the total trend and each component's contribution.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **INTERACTIVE VISUALIZATIONS**

- **Interactive Visualizations**
  - **Definition**: Interactive visualizations enable users to engage with data, allowing them to filter, zoom, or explore different aspects of the visualization in real-time.
  - **Purpose**: Interactivity makes complex data more accessible and engaging by letting users explore data at their own pace, focus on specific insights, and understand data from multiple perspectives. Interactive elements are commonly used in dashboards, data reports, and presentations.

# ADVANCED DATA VISUALIZATION TECHNIQUES: **INTERACTIVE VISUALIZATIONS**

- **Examples of Interactive Visualizations**:
  - **Interactive Dashboards**: Dashboards combine several interactive charts, such as bar charts, line charts, and maps, allowing users to filter data by category, time, or region.
  - **Hover and Tooltip Effects**: Information appears when hovering over data points, allowing users to see details like precise values or additional data attributes without cluttering the main visualization.
  - **Zooming and Panning**: Especially useful for geographical or temporal data, zooming and panning enable users to focus on specific regions or time periods.
  - **Linked Visualizations**: Changes in one chart (e.g., filtering) update related charts, making it easier to see relationships between multiple datasets in real-time.

# CASE STUDY

1. **Introduction: Context and Purpose**
   - **Objective**: Start by describing the purpose of the presentation and giving the context for the data being visualized.
   - **Key Points**:
     - **Background**: What is the dataset about? What questions are being addressed?
     - **Purpose of Visualizations**: Explain why specific visualizations are used (e.g., to analyze trends, compare categories, or show relationships).
2. **Description of Visuals**
   - **Objective**: Describe each visualization in terms of its structure, type, and elements.
   - **Key Points**:
     - **Chart Type**: Name the type of chart (bar, line, scatter plot, heatmap, etc.) and explain why it's appropriate for the data.
     - **Data Breakdown**: Briefly explain the axes, labels, colors, legend, and any other components to ensure clarity.

# CASE STUDY

3. **Highlighting Key Insights**
   - **Objective**: Guide the audience through the main takeaways of each visualization.
   - **Key Points:**
     - Trends and Patterns: Describe notable trends (e.g., "sales increased steadily over the year") or anomalies (e.g., "a sharp drop in sales in Q3").
     - Comparisons: If relevant, compare different categories or groups (e.g., "Product A outperformed Product B by 20%").
     - Statistical Insights: Point out any significant statistics, like averages, maximums, or minimums, if applicable.
4. **Interpreting the Findings**
   - **Objective**: Move beyond the visuals to interpret what the data might imply or suggest.
   - **Key Points:**
     - Potential Causes: Explain factors that may have influenced the trends (e.g., seasonality, market shifts).
     - Implications: Discuss what these findings mean for decision-making or future analysis (e.g., "This suggests a need to increase marketing in Q3").

# CASE STUDY

5. **Concluding Summary**
   - **Objective**: Summarize the presentation's key points, reinforcing main insights and interpretations.
   - **Key Points**:
     - **Main Insights Recap**: Briefly list the top insights.
     - **Recommendations**: If applicable, propose next steps or strategies based on the findings.
     - **Future Analysis Suggestions**: Suggest areas for further investigation or data collection if relevant.
6. **Q&A and Audience Engagement**
   - **Objective**: Open the floor for questions and encourage the audience to explore and question the findings.
   - **Key Points**:
     - **Invite Questions**: Engage the audience in discussion, encouraging critical thinking.
     - **Ask Reflective Questions**: Pose questions like, "What additional data could support these insights?" to prompt deeper engagement.

# DATA VISUALIZATION TOOLS

# DATA VISUALIZATION TOOLS: **MATPLOTLIB**

- **Overview**:
  - **Matplotlib** is one of the oldest and most widely-used Python libraries for data visualization. It offers comprehensive control over charts and supports a wide range of basic and advanced visualizations.
- **Key Features**:
  - **High Customizability**: Every aspect of a plot can be modified, including colors, labels, legends, and axis scales.
  - **2D Plotting Capabilities**: Supports line plots, bar charts, scatter plots, histograms, and more.
  - **Integration**: Works well with other Python libraries, including NumPy and Pandas, and supports exporting plots in various formats (PNG, PDF, etc.).
  - **Basic Interactive Features**: Limited interactivity but offers basic functions like zooming and panning within plots.
- **Ideal For**: Users who need detailed control over simple visualizations, such as static plots in reports or presentations.
- **Limitations**:
  - Steeper learning curve for complex customizations.
  - Limited interactivity compared to other tools.

# VISUALIZATION TOOLS: SEABORN

- **Overview**: Built on top of Matplotlib, **Seaborn** provides a more user-friendly API and focuses on statistical data visualization, especially for data exploration. It integrates well with Pandas and adds color themes and options for data visualization.
- **Key Features**:
  - **Easy-to-Use Syntax**: Simplifies the process of creating aesthetically pleasing and informative statistical visualizations.
  - **Built-In Statistical Plots**: Supports heatmaps, pair plots, violin plots, and other statistical plots out-of-the-box.
  - **Color Palettes**: Offers pre-defined color themes and palettes, allowing users to create visually appealing charts with minimal customization.
  - **High-Level Abstractions**: Automatically handles a lot of formatting based on data structure, making it easy to create complex plots quickly.
- **Ideal For**: Quick data exploration, statistical analysis, and visualizations involving relationships and distributions.
- **Limitations**:
  - Limited customization compared to Matplotlib.
  - Some visualizations may be challenging to customize beyond built-in options.

# VISUALIZATION TOOLS: IA BASED TOOLS

- Advantages of LLM-Based visualization tools:
  - It makes visualization tools accessible for non-technical users and improve efficiency for data professionals by generating insights and visuals quickly.
  - It helps users focus on interpreting data rather than creating charts manually.
  - It allows users to generate complex charts and graphs using natural language prompts or guidance, making them highly suitable for non-experts and enhancing productivity for experienced users.

Please generate a set of data visualizations to explore the following dataset:

```
{
  "age": [35, 42, 28, 51, 39, 27, 44, 33, 36, 30],
  "salary": [62000, 55000, 48000, 75000, 59000, 52000, 68000, 60000, 57000, 50000],
  "experience": [8, 12, 5, 15, 10, 4, 13, 7, 9, 6],
  "department": ["Sales", "Tech", "HR", "Tech", "Sales", "HR", "Tech", "Sales", "HR", "Tech"]
}
```

The visualizations should include:
- A histogram to show the distribution of ages
- A box plot to compare salaries across departments
- A scatter plot to explore the relationship between experience and salary, colored by department
- A correlation heatmap to understand the relationships between the numerical variables

Please provide the code to generate these visualizations, as well as a brief interpretation of the insights you can gather from the data.

# VISUALIZATION TOOLS: IA BASED TOOLS

- **Plotly's AI-Powered Chart Creation:** Plotly has introduced AI-based chart creation tools that integrate natural language processing into its dashboard creation platform. With Plotly's tools, users can describe the type of visualization they want to see, and the system will generate the code or chart accordingly.
- **Key Features**:
  - **Text-to-Visualization**: Users can type commands or descriptions (e.g., "scatter plot of sales over time with categories by region") and receive a Plotly graph.
  - **Code Suggestions**: For those using Python code, Plotly can generate code snippets to create custom visualizations based on natural language input.
  - **Interactive Dashboards**: Users can quickly create interactive dashboards with prompts, which is ideal for prototyping and sharing insights in real-time.
- **Ideal For**: Users familiar with Plotly who want a quicker, intuitive way to create complex or custom visualizations without writing detailed code.

# VISUALIZATION TOOLS: IA BASED TOOLS

- **DataRobot's AI Visualizations:** DataRobot offers automated visualization suggestions and insights as part of its broader automated machine learning (AutoML) platform. The AI assistant provides users with recommended charts and data insights based on the dataset being analyzed.
- **Key Features**:
  - **Automated Insights**: Analyzes data and suggests appropriate visualizations, highlighting key patterns, trends, or correlations.
  - **Natural Language Explanations**: Descriptions accompany visualizations, helping users interpret what they see in simple language.
  - **Multiple Chart Types**: Based on the type of data and analysis, DataRobot suggests scatter plots, heatmaps, bar charts, and more.
- **Ideal For**: Businesses or data scientists using DataRobot for predictive modeling and exploratory data analysis, who want quick, data-driven visualization recommendations.

# ANNEX DATA VISUALIZATION: BAR PLOT

```
import matplotlib.pyplot as plt

# Data for bar chart
courses = ['Python', 'Machine Learning', 'SQL', 'Data Visualization']
students = [120, 90, 70, 60]

# Plotting the bar chart
plt.figure(figsize=(8, 5))
plt.bar(courses, students, color='skyblue')

# Adding titles and labels
plt.title('Student Enrollment in Courses')
plt.xlabel('Courses')
plt.ylabel('Number of Students')
plt.show()
```

# ANNEX DATA VISUALIZATION: STACKED BAR PLOT

```
import matplotlib.pyplot as plt

# Data for stacked bar chart
categories = ['Q1', 'Q2', 'Q3', 'Q4']
sales_a = [200, 250, 300, 400]  # Sales by Department A
sales_b = [150, 200, 250, 300]  # Sales by Department B

# Plotting the stacked bar chart
plt.figure(figsize=(8, 5))
plt.bar(categories, sales_a, color='skyblue', label='Department A')
plt.bar(categories, sales_b, bottom=sales_a, color='salmon', label='Department B')

# Adding titles, labels, and legend
plt.title('Quarterly Sales by Department')
plt.xlabel('Quarter')
plt.ylabel('Sales')
plt.legend()
plt.show()
```

# ANNEX DATA VISUALIZATION: BUBBLE PLOT

```
import matplotlib.pyplot as plt


# Data for bubble chart
study_hours = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
test_scores = [52, 55, 60, 62, 65, 68, 72, 75, 78, 80]
effort_level = [20, 30, 40, 50, 60, 70, 80, 90, 100, 110]  # Bubble size representing effo

# Plotting the bubble chart
plt.figure(figsize=(8, 5))
plt.scatter(study_hours, test_scores, s=effort_level, color='teal', alpha=0.5)

# Adding titles and labels
plt.title('Study Hours vs Test Scores (Bubble Size = Effort Level)')
plt.xlabel('Study Hours')
plt.ylabel('Test Scores')
plt.show()
```

# ANNEX DATA VISUALIZATION: LINE CHART PLOT

```python
import matplotlib.pyplot as plt

# Data for line chart
months = ['Jan', 'Feb', 'Mar', 'Apr', 'May', 'Jun', 'Jul', 'Aug', 'Sep', 'Oct', 'Nov',
revenue = [2000, 2200, 2100, 2500, 2700, 3000, 2800, 3200, 3300, 3100, 3400, 3600]

# Plotting the line chart
plt.figure(figsize=(10, 6))
plt.plot(months, revenue, marker='o', color='purple')

# Adding titles and labels
plt.title('Monthly Revenue')
plt.xlabel('Month')
plt.ylabel('Revenue ($)')
plt.grid(True)  # Optional: Adds grid for readability
plt.show()
```

# ANNEX DATA VISUALIZATION: PIE PLOT

```python
import matplotlib.pyplot as plt

# Data for pie chart
activities = ['Work', 'Sleep', 'Exercise', 'Leisure']
time_spent = [8, 7, 2, 7]  # hours in a day

# Plotting the pie chart
plt.figure(figsize=(7, 7))
plt.pie(time_spent, labels=activities, autopct='%1.1f%%', startangle=140)

# Adding title
plt.title('Time Spent on Daily Activities')
plt.show()
```

# ANNEX DATA VISUALIZATION: SCATTER PLOT

```python
import matplotlib.pyplot as plt

# Data for scatter plot
study_hours = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
test_scores = [52, 55, 60, 62, 65, 68, 72, 75, 78, 80]

# Plotting the scatter plot
plt.figure(figsize=(8, 5))
plt.scatter(study_hours, test_scores, color='orange')

# Adding titles and labels
plt.title('Study Hours vs Test Scores')
plt.xlabel('Study Hours')
plt.ylabel('Test Scores')
plt.show()
```

# ANNEX DATA VISUALIZATION: HIST PLOT

```python
import matplotlib.pyplot as plt

# Data for histogram
scores = [55, 58, 60, 62, 65, 65, 70, 72, 75, 78, 80, 82, 85, 88, 90]

# Plotting the histogram
plt.figure(figsize=(8, 5))
plt.hist(scores, bins=5, color='teal', edgecolor='black')

# Adding titles and labels
plt.title('Distribution of Exam Scores')
plt.xlabel('Score Ranges')
plt.ylabel('Number of Students')
plt.show()
```

# ANNEX DATA VISUALIZATION: BOX PLOT

```python
import seaborn as sns
import matplotlib.pyplot as plt

# Data for box plot
salaries = {
    'Department A': [45, 50, 55, 60, 65, 70, 75],
    'Department B': [40, 45, 50, 52, 56, 60, 64]
}

# Converting data to list of lists for Seaborn
data = [salaries['Department A'], salaries['Department B']]

# Plotting the box plot
plt.figure(figsize=(8, 5))
sns.boxplot(data=data, palette='Set2')
plt.xticks([0, 1], ['Department A', 'Department B'])  # Setting custom x-axis labels

# Adding title
plt.title('Salary Distribution by Department')
plt.xlabel('Department')
plt.ylabel('Salary (Thousands)')
plt.show()
```

# ANNEX DATA VISUALIZATION: HEATMAP PLOT

```python
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

# Data for heatmap (example correlation matrix)
data = np.array([[1, 0.5, 0.2], [0.5, 1, 0.3], [0.2, 0.3, 1]])
labels = ['Feature A', 'Feature B', 'Feature C']

# Plotting the heatmap
plt.figure(figsize=(6, 5))
sns.heatmap(data, annot=True, xticklabels=labels, yticklabels=labels, cmap='coolwarm')

# Adding title
plt.title('Feature Correlation Matrix')
plt.show()
```

# ANNEX DATA VISUALIZATION: GEODATA PLOT

```python
import geopandas as gpd
import matplotlib.pyplot as plt

# Load a GeoDataFrame containing US states
gdf = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))
usa = gdf[gdf['continent'] == 'North America']

# Example data: adding a random population density column for demonstration
import numpy as np
usa['pop_density'] = np.random.randint(10, 100, size=len(usa))

# Plotting the choropleth map
plt.figure(figsize=(10, 6))
usa.plot(column='pop_density', cmap='Blues', legend=True)

# Adding title
plt.title('Population Density by State (Example Data)')
plt.show()
```

# ANNEX DATA VISUALIZATION: VIOLIN PLOT

```
# Generate a sample dataset
np.random.seed(42)
data = {
    "Category": np.repeat(["A", "B", "C"], 100),
    "Value": np.concatenate([
        np.random.normal(50, 10, 100),
        np.random.normal(60, 15, 100),
        np.random.normal(55, 5, 100),
    ])
}

df = pd.DataFrame(data)

# Create a violin plot
plt.figure(figsize=(8, 6))
sns.violinplot(x="Category", y="Value", data=df, palette="muted")

# Add labels and title
plt.title("Violin Plot of Values by Category", fontsize=16)
plt.xlabel("Category", fontsize=14)
plt.ylabel("Value", fontsize=14)

# Show the plot
plt.tight_layout()
plt.show()
```

# CONCLUSION

- Data visualization transforms raw data into visuals, making insights easier to understand.
- Key concepts: visual encodings, chart types, interactivity, and storytelling.
- It simplifies complexity, aids decision-making, and enhances communication and engagement.

# REFERENCES


- Beautiful Visualization: Looking at Data Through the Eyes of Experts" by **Julie Steele and Noah Iliinsky**
  - This compilation features insights from 24 experts on the design process behind successful data visualizations, focusing on storytelling and communication.
- The Visual Display of Quantitative Information" by **Edward R. Tufte**
  - A classic text that discusses the theory and design of data graphics, illustrated with numerous examples that highlight effective and ineffective practices.
- Storytelling with Data: A Data Visualization Guide for Business Professionals" by **Cole Nussbaumer Knaflic**
  - This guide focuses on how to tell compelling stories through data visualizations, emphasizing the importance of context and audience understanding.
- Data Visualization: A Handbook for Data Driven Design" by **Andy Kirk**
  - This comprehensive handbook reviews various methods and techniques for creating effective data visualizations, making it a practical resource for practitioners.