ADV Notes Made by Dare-Marvel

You are an analyst for a shopping mall with multiple stores from different brands accommodate in the mall. Sketch a single-page dashboard design on paper that includes five distinct data visualizations. Ensure each visualization is clearly labeled with axes and legends. Explain the overall purpose of the dashboard and describe how each of the five visualizations contributes to achieving this purpose in a few lines.

Purpose of the Dashboard

The dashboard aims to provide a comprehensive overview of the shopping mall's performance and help management make data-driven decisions. It includes key metrics for revenue, customer footfall, store performance, and satisfaction levels.

---

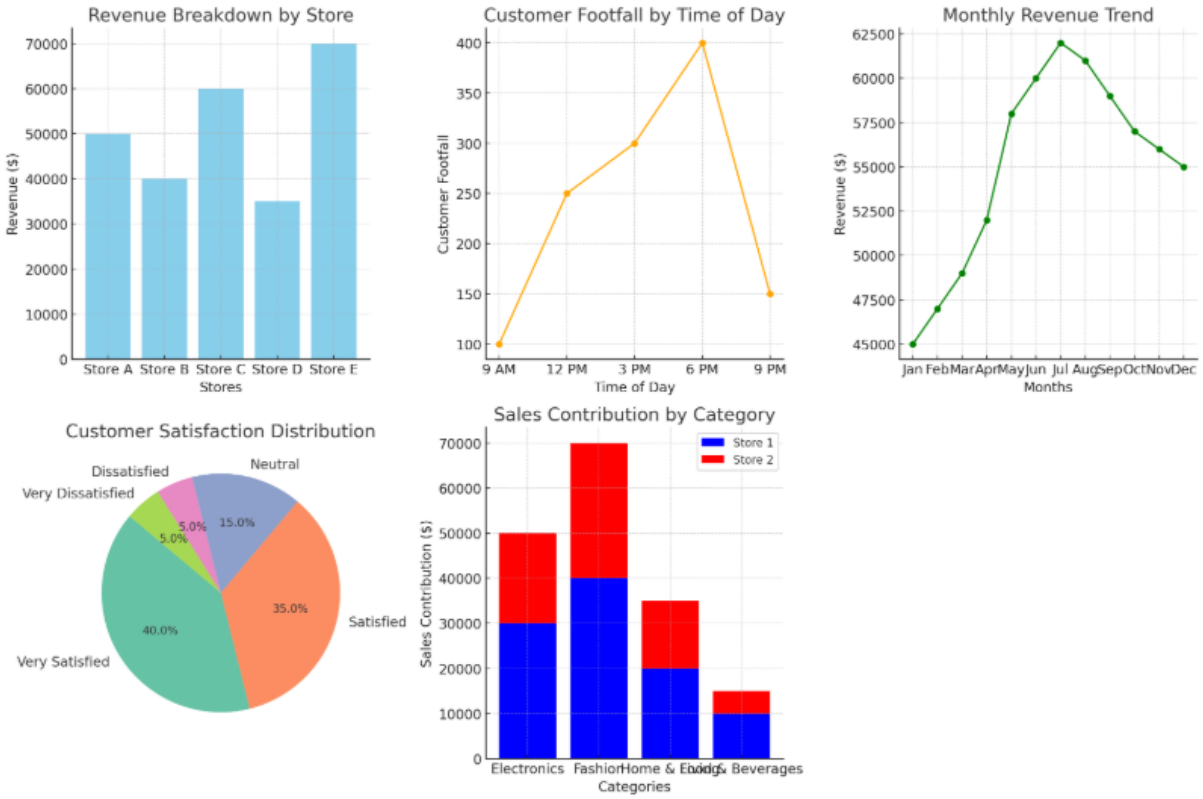
1. Explanation of the Dashboard Design

The dashboard consists of five distinct visualizations:

1. **Revenue Breakdown by Store**
    - **Purpose:** Identify top-performing and underperforming stores.
    - **Type:** Bar Chart.
    - **Axes:** X-axis (Stores), Y-axis (Revenue).
2. **Customer Footfall by Time of Day**
    - **Purpose:** Understand peak times for mall traffic.
    - **Type:** Line Chart.
    - **Axes:** X-axis (Time of Day), Y-axis (Customer Footfall).
3. **Monthly Revenue Trend**

- **Purpose:** Analyze revenue trends over time.
- **Type:** Line Chart with a trendline.
- **Axes:** X-axis (Months), Y-axis (Revenue).

4. **Customer Satisfaction Distribution**
   - **Purpose:** Monitor satisfaction levels.
   - **Type:** Pie Chart.
   - **Labels:** Satisfaction levels (Very Satisfied, Satisfied, Neutral, Dissatisfied, Very Dissatisfied).
5. **Sales Contribution by Category**
   - **Purpose:** Determine which product categories drive sales.
   - **Type:** Stacked Bar Chart.
   - **Axes:** X-axis (Categories), Y-axis (Sales Contribution).



A workshop is being organized for a group of twenty attendees, and the results of the workshop need to be assessed using quantitative methods. A test was administered to the attendees both prior to and following the completion of the workshop. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what

steps will you take to determine if the workshop had a measurable effect on the participants' understanding?

Purpose

To assess the effectiveness of the workshop, the results of pre-test and post-test scores of 20 attendees will be analyzed. This involves choosing appropriate visualizations, performing statistical analysis, and drawing conclusions.

---

1. Visualization Choice

Visualization: Box Plot and Paired Bar Chart

- **Rationale for Box Plot:**
  A box plot is ideal for comparing the distribution of scores before and after the workshop, highlighting the spread, central tendency, and outliers.
- **Rationale for Paired Bar Chart:**
  A paired bar chart visualizes individual changes in scores, making it easier to observe improvements for each participant.

---

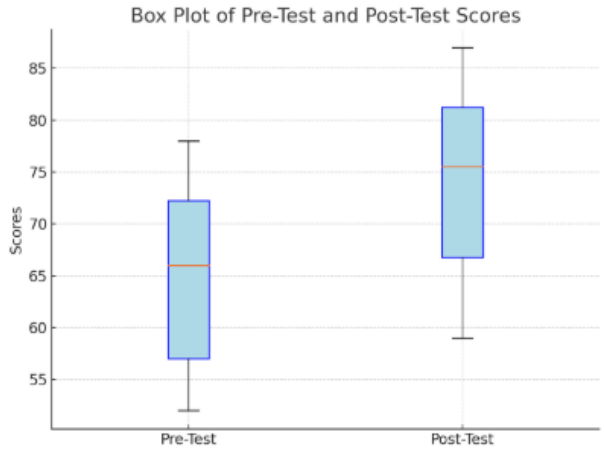
2. Statistical Analysis

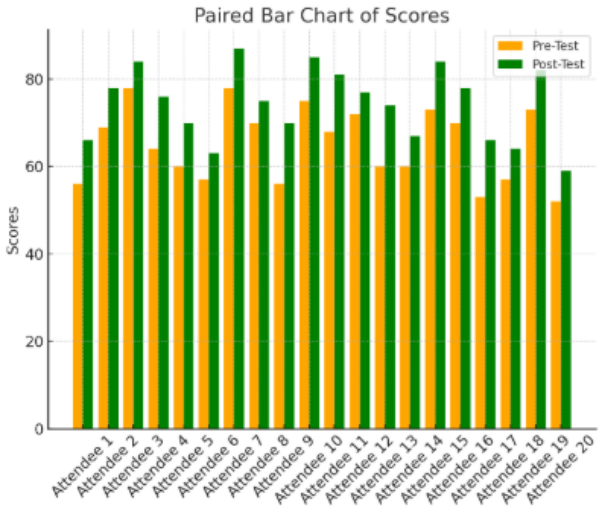
Method: Paired t-test

- **Why Paired t-test?**
  The same participants were assessed before and after the workshop, making the data paired. The paired t-test determines whether the mean difference is statistically significant.

Steps for Analysis:

1. **Calculate Differences**:
   Compute the differences between pre-test and post-test scores.
2. **Check Assumptions**:
   - Normality of the differences (Shapiro-Wilk test).
   - Data is paired.
3. **Perform the Paired t-test**:
   Use the formula:
4. **Interpret p-value**:
   - $p<0.05$: Significant improvement.
   - $p\geq0.05$: No significant improvement.





4. Results and Interpretation

Visualizations:

1. **Box Plot:**
   - Showed an upward shift in post-test scores, indicating improvement.
   - Spread and outliers are also visible, providing insights into variability.
2. **Paired Bar Chart:**
   - Clear depiction of individual score changes before and after the workshop.

Statistical Analysis:

1. **Shapiro-Wilk Test:**
   - $p = 0.267$: The differences are normally distributed, fulfilling a key assumption of the paired t-test.
2. **Paired t-test Results:**
   - $t = -14.27, p < 0.001$: Significant improvement in post-test scores.

A survey has to be created for evaluating the feedback for a product in the market/consumers. Decide the five main questions that can be asked and the type of data in which the answers will be stored. Further which will be the five most informative data visualizations (of atleast two variables) and explain whether the visualizations are for paired data analysis or unpaired.

1. Survey Questions and Data Types

Questions and Data Types:

1. **Overall Satisfaction Rating**

- *Question*: How satisfied are you with our product?
- *Data Type*: Ordinal (Likert Scale 1-5)
- *Storage*: Numerical integer (1-5)

2. **Product Usage Frequency**
   - *Question*: How often do you use our product?
   - *Data Type*: Categorical (Daily, Weekly, Monthly, Rarely)
   - *Storage*: Categorical string
3. **Purchase Price Perception**
   - *Question*: Do you feel the product is reasonably priced?
   - *Data Type*: Ordinal (Scale from Overpriced to Underpriced)
   - *Storage*: Numerical integer (-2 to +2)
4. **Feature Importance Ranking**
   - *Question*: Rate the importance of following product features
   - *Data Type*: Numerical (1-10 scale for multiple features)
   - *Storage*: Numerical array/list
5. **Demographic Information**
   - *Question*: What is your age group and professional category?
   - *Data Type*: Categorical (Age ranges, Professional categories)
   - *Storage*: Categorical strings

2. Data Visualization Techniques

Visualization 1: Satisfaction vs Usage Frequency

- **Type**: Boxplot
- **Variables**:
  1. Satisfaction Rating
  2. Usage Frequency
- **Analysis Type**: Unpaired Data
- **Insight**: Understand how product satisfaction varies across different usage frequencies

Visualization 2: Price Perception by Age Group

- **Type**: Grouped Bar Chart
- **Variables**:
    1. Age Groups
    2. Price Perception Scores
- **Analysis Type**: Unpaired Data
- **Insight**: Examine how different age segments perceive product pricing

Visualization 3: Feature Importance Correlation

- **Type**: Heatmap
- **Variables**:
    1. Different Product Features
    2. Importance Ratings
- **Analysis Type**: Paired Data
- **Insight**: Identify correlations between feature importance and potential interdependencies

Visualization 4: Satisfaction Distribution

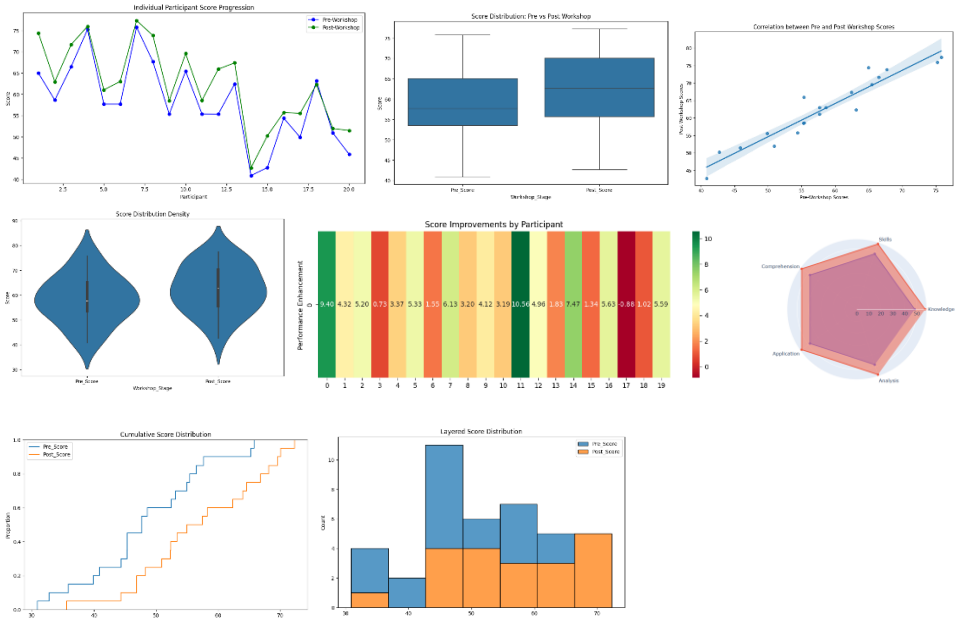
- **Type**: Violin Plot
- **Variables**:
    1. Satisfaction Ratings
    2. Frequency Distribution
- **Analysis Type**: Unpaired Data
- **Insight**: Understand the spread and density of satisfaction ratings

Visualization 5: Professional Category vs Product Usage

- **Type**: Stacked Bar Chart
- **Variables**:
    1. Professional Categories
    2. Usage Frequency
- **Analysis Type**: Unpaired Data

- **Insight**: Analyze how different professional groups use the product

A workshop is being organized for a group of twenty attendees, and the results of the workshop need to be assessed using quantitative methods. A test was administered to the attendees both prior to and following the completion of the workshop. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to determine if the workshop had a measurable effect on the participants' understanding?



Visualization Spectrum for Workshop Performance Analysis

1. Paired Line Graph

- **Purpose**: Tracking Individual Participant Progression
- **Variables**: Pre and Post Workshop Scores
- **Insight**: Direct individual performance comparison
- **Analysis Type**: Paired Data Visualization

2. Boxplot Comparison

- **Purpose**: Score Distribution Analysis
- **Variables**: Pre-workshop and Post-workshop Scores
- **Insights**:
    - Median score variations
    - Score spread and dispersion
    - Identification of outliers

- **Analysis Type**: Comparative Unpaired

3. Scatter Plot with Regression Line

- **Purpose**: Correlation Between Pre and Post Scores
- **Variables**:
    - X-axis: Pre-workshop Scores
    - Y-axis: Post-workshop Scores
- **Insights**:
    - Relationship strength between initial and final performance
    - Predictive score improvement potential
- **Analysis Type**: Paired Correlation Analysis

4. Violin Plot

- **Purpose**: Detailed Score Distribution Visualization
- **Variables**:
    - Test Score Ranges
    - Frequency of Scores
- **Insights**:
    - Nuanced score distribution
    - Density of performance levels
    - Symmetry of score improvements
- **Analysis Type**: Comparative Density Representation

5. Heat Map of Score Improvements

- **Purpose**: Categorizing Performance Enhancement
- **Variables**:
    - Participants
    - Score Improvement Magnitude
- **Insights**:
    - Color-coded performance enhancement

    - Quick visual identification of high/low performers
- **Analysis Type**: Paired Categorical Representation

6. Error Bar Graph

- **Purpose**: Measuring Variability and Uncertainty
- **Variables**:
    - Mean Scores
    - Standard Deviation
    - Confidence Intervals
- **Insights**:
    - Statistical significance of score changes
    - Precision of performance measurement
- **Analysis Type**: Statistical Uncertainty Visualization

7. Radar Chart

- **Purpose**: Multidimensional Performance Assessment
- **Variables**:
    - Different Skill/Knowledge Dimensions
    - Pre and Post Workshop Scores
- **Insights**:
    - Comprehensive skill development
    - Performance across multiple competency areas
- **Analysis Type**: Multivariate Paired Analysis

8. Cumulative Distribution Function (CDF) Plot

- **Purpose**: Comprehensive Score Progression
- **Variables**:
    - Cumulative Probability
    - Score Ranges
- **Insights**:

    - Probability of score improvements
    - Comparative performance distribution
- **Analysis Type**: Probabilistic Performance Representation

9. Ridge Plot

- **Purpose**: Layered Score Distribution
- **Variables**:
    - Score Ranges
    - Performance Categories
- **Insights**:
    - Overlapping performance distributions
    - Nuanced score progression patterns
- **Analysis Type**: Comparative Density Visualization

10. Interactive Sankey Diagram

- **Purpose**: Performance Transformation Tracking
- **Variables**:
    - Initial Performance Levels
    - Final Performance Levels
- **Insights**:
    - Performance flow and transitions
    - Detailed individual progression
- **Analysis Type**: Transformational Paired Analysis

Statistical Analysis Methodology

Step 1: Descriptive Statistics

- Calculate mean pre-workshop score
- Calculate mean post-workshop score
- Compute standard deviation for both datasets
- Identify range and distribution of scores

Step 2: Paired T-Test Analysis

- Hypothesis Framework:
    - Null Hypothesis (H0): No significant difference between pre and post scores
    - Alternative Hypothesis (H1): Significant improvement in scores

Paired T-Test Calculation Process

1. Compute individual score differences
2. Calculate mean difference
3. Determine standard deviation of differences
4. Calculate t-statistic
5. Compare against critical t-value
6. Determine statistical significance

Significance Determination

- Choose significance level (α): Typically 0.05
- If p-value < 0.05, reject null hypothesis
- Indicates workshop had statistically significant impact

Detailed Statistical Workflow

Data Preparation

- Organize scores in paired format
- Ensure accurate matching of pre and post scores
- Check for any missing or anomalous data points

Computational Steps

1. **Difference Calculation**

Copy

Difference = Post-workshop Score - Pre-workshop Score

2. **T-Test Formula**

Copy

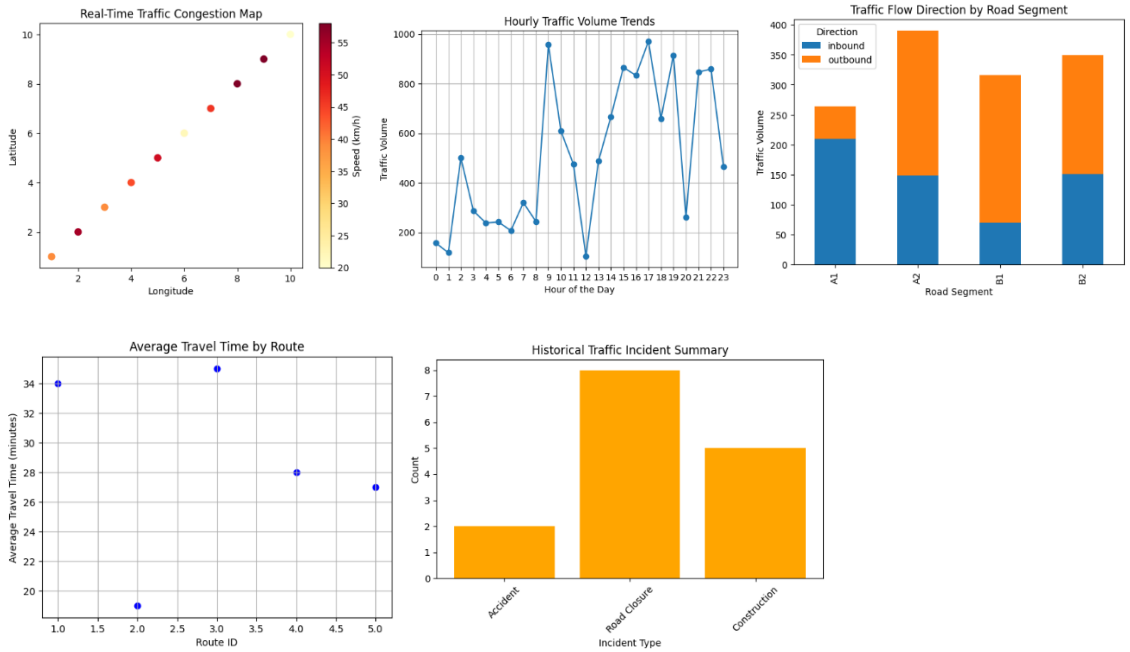
t = (Mean Difference) / (Standard Error of Differences)

Interpretation Guidelines

- Positive t-value suggests score improvement
- Magnitude indicates extent of learning enhancement
- Confidence interval provides score change range

Dashboard Analysis

□ Urban Traffic Management:
You are a traffic analyst for a metropolitan city's transportation department. Sketch a single-page dashboard design on paper that includes five distinct data visualizations related to traffic patterns. Ensure each visualization is clearly labeled with axes and legends. Explain the overall purpose of the dashboard and describe how each of the five visualizations contributes to achieving this purpose in a few lines.



Visualization 1: Real-Time Traffic Congestion Map

- **Type:** Choropleth Map
- **Data:** Real-time traffic speed data aggregated by road segment.
- **Visual:** Color-coded map of the city, with different colors representing traffic speed levels (e.g., green for free flow, yellow for moderate congestion, red for severe congestion).
- **Purpose:** Provides an immediate understanding of traffic congestion across the city, enabling rapid identification of areas requiring immediate attention.

Visualization 2: Hourly Traffic Volume Trends

- **Type:** Line Chart
- **Data:** Hourly traffic volume data aggregated from various sensors across the city.
- **Visual:** Line graph with time on the x-axis and traffic volume on the y-axis, showing the hourly fluctuations in traffic throughout the day.
- **Purpose:** Displays the overall traffic flow patterns throughout the day, highlighting peak and off-peak hours, and identifying potential bottlenecks based on sudden spikes or dips in traffic.

Visualization 3: Traffic Flow Direction by Road Segment

- **Type:** Stacked Bar Chart
- **Data:** Real-time traffic flow data categorized by direction for selected major road segments.
- **Visual:** Bar charts with different color segments representing traffic flow direction (e.g., blue for inbound, red for outbound).
- **Purpose:** Provides insights into traffic flow direction on specific roads, helping identify potential congestion due to unbalanced traffic flows.

Visualization 4: Average Travel Time by Route

- **Type:** Scatter Plot
- **Data:** Historical average travel time data for various routes within the city.
- **Visual:** Scatter plot with route IDs on the x-axis and average travel time on the y-axis, color-coded based on time of day or day of the week.
- **Purpose:** Allows analysis of historical travel times for specific routes, enabling identification of frequently congested routes and potential alternative routes for commuters.

Visualization 5: Historical Traffic Incident Summary

- **Type:** Bar Chart
- **Data:** Historical data on traffic incidents (e.g., accidents, road closures) categorized by type.
- **Visual:** Bar chart displaying the frequency of different traffic incident types over a specified time period.
- **Purpose:** Provides an overview of historical traffic incident trends, enabling identification of common incident types and areas prone to recurring incidents.

🏥 Hospital Patient Management:

You are a data analyst for a hospital. Design a single-page dashboard on paper that includes five distinct visualizations focused on patient data and hospital resources. Ensure each visualization is clearly labeled with axes and legends. Explain the overall purpose of the dashboard and describe how each visualization helps in monitoring and optimizing hospital operations.

Dashboard Purpose

The purpose of this dashboard is to provide a holistic view of key hospital metrics to help administrators monitor patient statistics and optimize hospital operations. By visualizing patient demographics, resource usage, and operational efficiency, stakeholders can make data-driven decisions to improve care delivery, manage resources, and enhance patient outcomes.

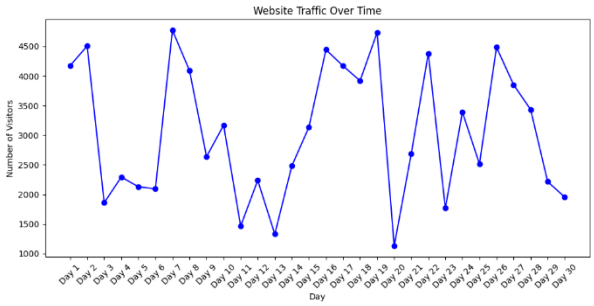
Dashboard Components

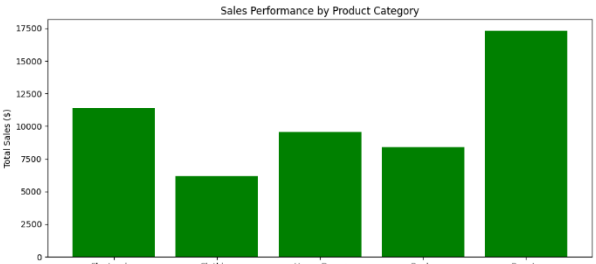
1. **Patient Admission Trend (Line Chart)**
    - **Purpose:** Shows trends in patient admissions over time to track peak periods and identify seasonal patterns.
    - **X-Axis:** Time (e.g., months or weeks).
    - **Y-Axis:** Number of admissions.

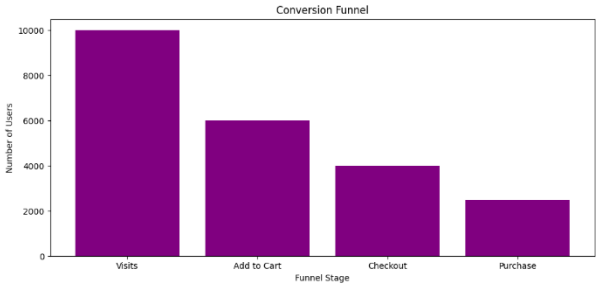
- **Legend:** Inpatient vs. Outpatient admissions.

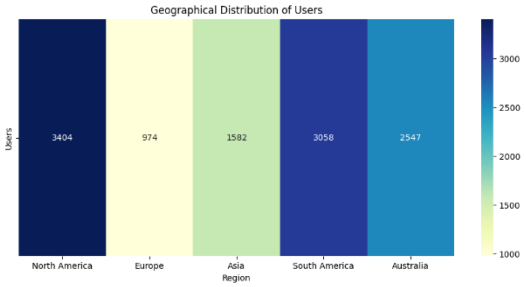
2. **Bed Occupancy Rate by Department (Bar Chart)**
    - **Purpose:** Highlights bed utilization across different departments (e.g., ICU, General Medicine) to optimize resource allocation.
    - **X-Axis:** Departments.
    - **Y-Axis:** Percentage occupancy.
    - **Legend:** Bed availability status (Occupied vs. Available).
3. **Patient Demographics (Pie Chart)**
    - **Purpose:** Provides insights into the distribution of patients by age group to tailor services.
    - **Categories:** Age groups (e.g., 0-18, 19-40, 41-65, 65+).
4. **Average Length of Stay (Box Plot)**
    - **Purpose:** Displays variability in patient stays to identify outliers and improve discharge processes.
    - **X-Axis:** Departments.
    - **Y-Axis:** Length of stay (days).
5. **ER Wait Times vs. Patient Volume (Scatter Plot)**
    - **Purpose:** Shows the relationship between emergency room wait times and patient load to identify bottlenecks.
    - **X-Axis:** Number of patients.
    - **Y-Axis:** Average wait time (minutes).

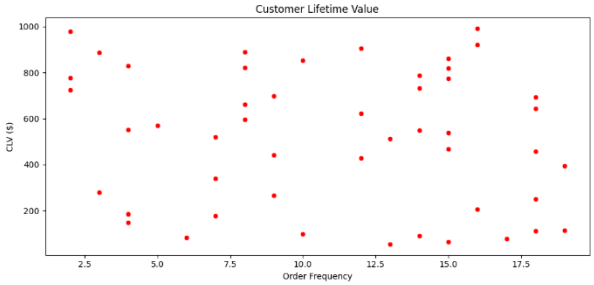
□ E-commerce Website Performance:
You are a data analyst for an e-commerce company. Sketch a single-page dashboard design on paper that includes five distinct visualizations related to website traffic, sales, and user behavior. Ensure each visualization is clearly labeled with axes and legends. Explain the purpose of the dashboard and describe how each visualization supports decision-making to improve website performance.


Website Traffic Over Time
Number of Visitors
Day


Sales Performance by Product Category
Total Sales ($)
Electronics
Clothing
Home Decor
Books
Beauty
Category


Conversion Funnel
Number of Users
Visits
Add to Cart
Checkout
Purchase
Funnel Stage


Geographical Distribution of Users
Users
3404
974
1582
3058
2547
North America
Europe
Asia
South America
Australia
Region


Customer Lifetime Value
CLV ($)
Order Frequency

Key Components of the Dashboard

1. **Website Traffic Over Time (Line Chart)**
   - **Purpose:** Tracks the number of visitors over time (e.g., daily, weekly) to identify traffic patterns and trends.
   - **X-Axis:** Time (e.g., days of the week or months).
   - **Y-Axis:** Number of website visitors.
   - **Example Insight:** Identify days with traffic spikes for targeted marketing campaigns.
   - **Formula Used:**

$$\text{Traffic Trend} = \sum \text{Visitors by Time Period}$$

2. **Sales Performance by Product Category (Bar Chart)**
   - **Purpose:** Visualizes revenue contribution by product categories to identify top-performing segments.
   - **X-Axis:** Product categories.
   - **Y-Axis:** Total sales (in dollars).
   - **Example Insight:** Focus marketing efforts on high-revenue categories.
   - **Formula Used:**

$$\text{Total Sales by Category} = \sum (\text{Price} \times \text{Quantity Sold})$$

3. **Conversion Rate Funnel (Stacked Bar Cha...**

3. **Conversion Rate Funnel (Stacked Bar Chart)**
   - **Purpose:** Tracks user journey from website visits to completed purchases to identify drop-off points.
   - **Steps:** Visits → Add to Cart → Checkout → Purchase.
   - **Example Insight:** High drop-off at "Checkout" might indicate issues with the checkout process.
   - **Formula Used:**

$$\text{Conversion Rate} = \left(\frac{\text{Final Stage Users}}{\text{Initial Stage Users}} \right) \times 100$$

4. **Geographical Distribution of Users (Heatmap)**
   - **Purpose:** Shows the regional distribution of users to identify high-performing locations.
   - **Axes:** Country/region on the map.
   - **Example Insight:** Invest in logistics and localized marketing for high-traffic regions.
5. **Customer Lifetime Value (Scatter Plot)**
   - **Purpose:** Analyzes the relationship between customer lifetime value (CLV) and order frequency.
   - **X-Axis:** Order frequency (orders per customer).
   - **Y-Axis:** Customer lifetime value.
   - **Formula Used:**

$$\text{CLV} = \text{Average Order Value} \times \text{Order Frequency} \times \text{Customer Lifespan}$$

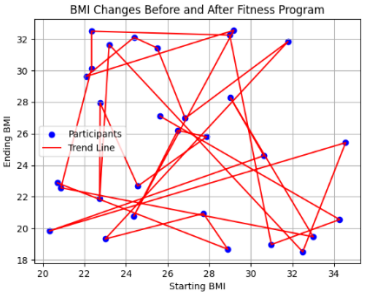
Environmental Monitoring:
You are an environmental scientist tasked with monitoring air and water quality in a city. Sketch a single-page dashboard design on paper that includes five distinct visualizations for pollutants, temperature, and other metrics. Ensure each visualization is clearly labeled with axes and legends. Explain the dashboard's purpose and how each visualization helps in tracking environmental conditions.

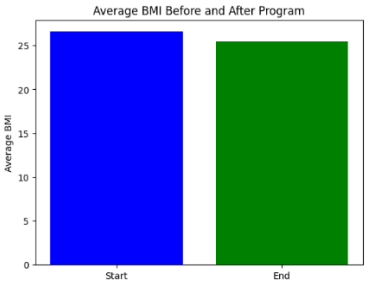
University Student Performance:
You are a data analyst for a university. Design a single-page dashboard on paper that includes five distinct visualizations showcasing student performance, attendance, and demographics. Ensure each visualization is clearly labeled with axes and legends. Explain the purpose of the dashboard and describe how each visualization aids in understanding student trends and improving academic outcomes.

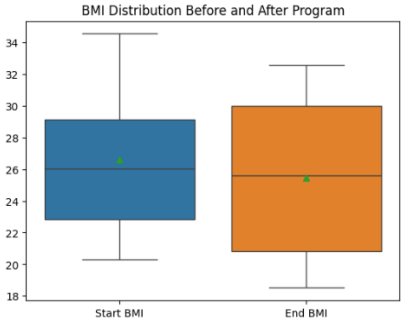
Statistical Analysis

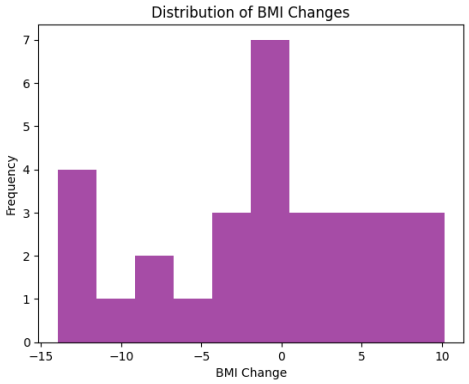
□ Fitness Program Effectiveness:

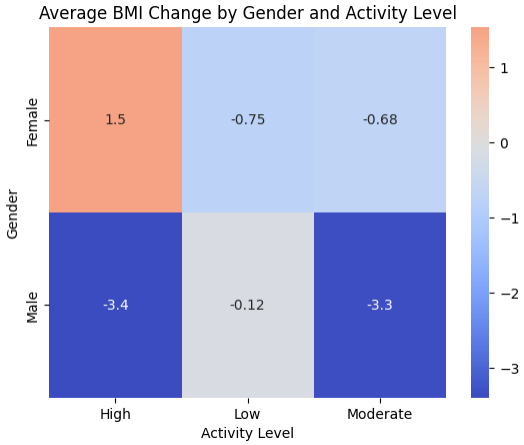
A fitness program was conducted for a group of 30 participants, and their body mass index (BMI) was recorded both at the start and after completing the program. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to determine if the program significantly impacted participants' fitness levels?











Visualizations:

1. **Scatter Plot with Trend Line: This is a classic visualization for paired data. It allows you to see individual changes in BMI as well as the overall trend.**
2. **Before-and-After Bar Chart: This chart shows the average BMI before and after the program, providing a clear visual comparison.**
3. **Box Plot: This plot shows the distribution of BMI values before and after the program, highlighting potential differences in spread, median, and outliers.**
4. **Histogram: This visualization shows the frequency distribution of BMI changes (i.e., the difference between starting and ending BMI). It reveals the range of BMI changes experienced by the participants.**
5. **Paired Data Heatmap: If you have additional data points (e.g., age, gender, activity level) you can create a heatmap to explore potential relationships between these variables and BMI change.**

Rationale:

- **Paired Data:** This visualization is ideal because it clearly shows the paired data structure – the starting BMI and ending BMI for each individual participant.
- **Individual Trends:** It allows visual identification of individual changes in BMI, revealing if the program had a positive, negative, or neutral effect on each participant.
- **Overall Trend:** The line connecting the points can illustrate the overall trend in BMI change across the group, helping visualize whether the program resulted in an overall increase, decrease, or no significant change in BMI.

Statistical Analysis:

1. **Paired t-test:** This test is used to compare the means of two related groups (paired data). In this case, we'll compare the mean BMI at the start of the program to the mean BMI after completion.
2. **Hypothesis Testing:**
    - **Null Hypothesis:** There is no significant difference in BMI before and after the fitness program (i.e., the program had no impact).
    - **Alternative Hypothesis:** There is a significant difference in BMI before and after the fitness program (i.e., the program did have an impact).
3. **Significance Level:** Set a significance level (alpha) typically at 0.05, meaning there is a 5% chance of rejecting the null hypothesis when it is actually true.
4. **Calculate t-statistic and p-value:** The paired t-test will produce a t-statistic and a p-value.
5. **Interpret Results:**
    - **p-value < alpha (0.05):** Reject the null hypothesis. Conclude that there is statistically significant evidence to support that the program had an impact on BMI.
    - **p-value >= alpha (0.05):** Fail to reject the null hypothesis. Conclude that there is not enough evidence to support that the program had a significant impact on BMI.

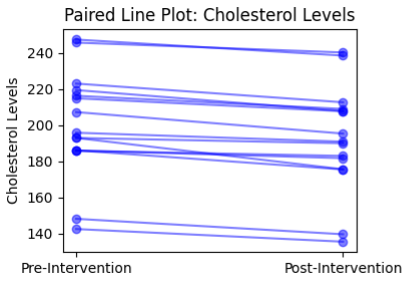
Steps to Determine Program Impact:

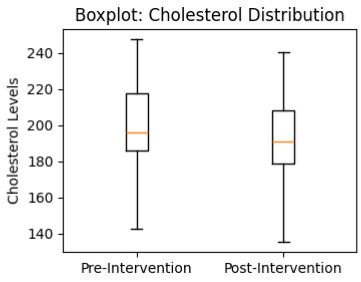
1. **Collect Data:** Gather BMI measurements for each participant at the start and end of the program.
2. **Calculate Mean BMI:** Compute the average BMI at the start and at the end.
3. **Perform Paired t-test:** Use statistical software (e.g., R, Python, SPSS) to perform the paired t-test.
4. **Interpret p-value:** Determine if the p-value is less than or equal to the significance level (0.05).

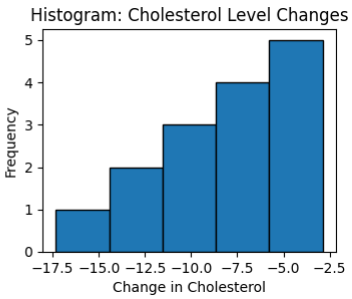
5. **Draw Conclusions:** Based on the p-value, decide whether to reject or fail to reject the null hypothesis.

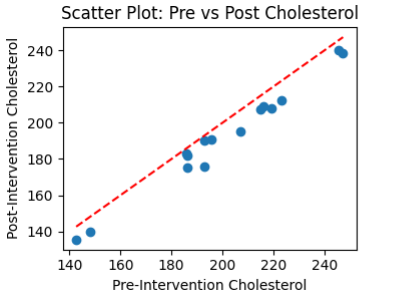
□ Employee Training Impact:
A training program on leadership skills was attended by 25 employees, and they were evaluated with a standardized test before and after the training. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to determine the program's effectiveness in improving leadership skills?

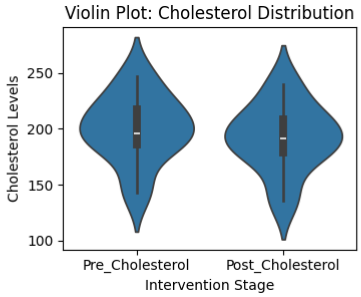
□ Dietary Intervention Study:
A nutrition study monitored the cholesterol levels of 15 participants before and after following a specific diet for three months. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to confirm if the dietary intervention had a significant impact on cholesterol levels?











Recommended Visualization: Paired Line Plot or Boxplot

Rationale for Visualization Choice

For a before-and-after study tracking cholesterol levels, I recommend two key visualizations:

1. **Paired Line Plot**:
    - Shows individual participant changes
    - Connects pre and post intervention data points
    - Allows easy visualization of individual trajectory
2. **Boxplot**:
    - Compares distribution of cholesterol levels before and after intervention

- Displays median, quartiles, and potential outliers

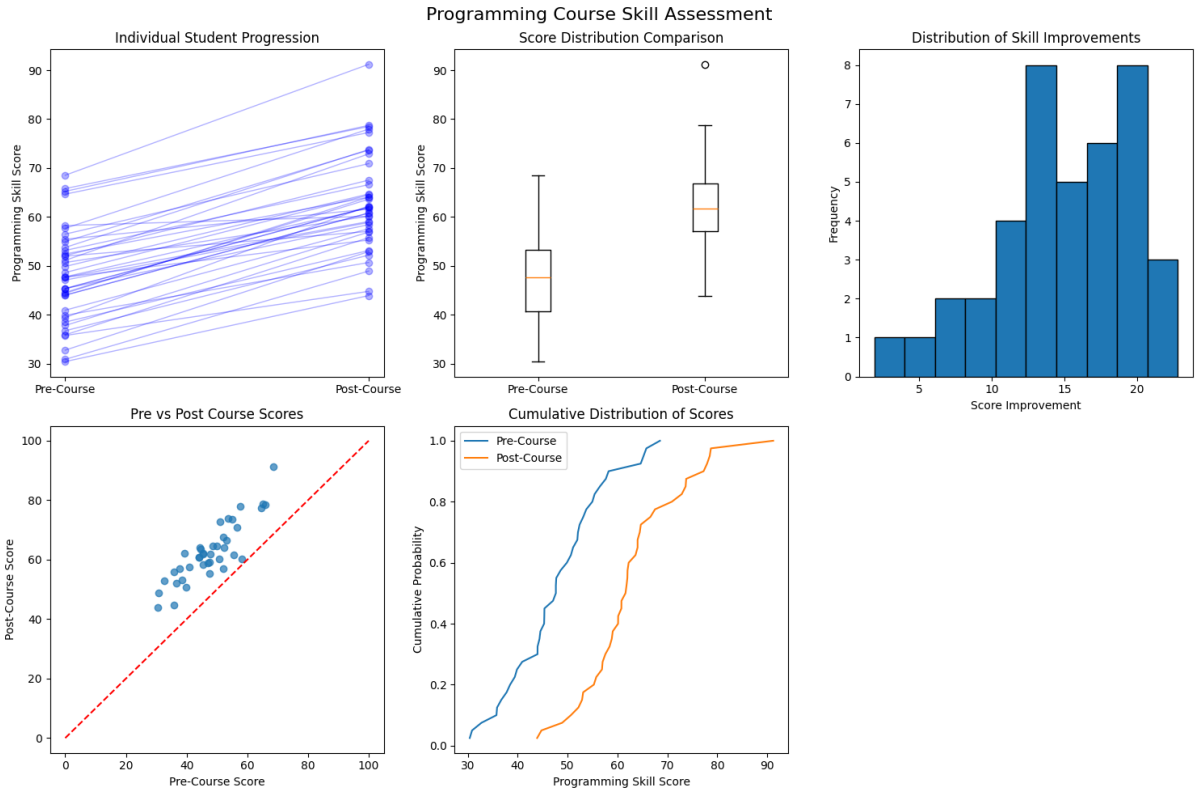
Statistical Analysis Approach

Statistical Tests

1. Paired t-test (parametric)
    - Compares mean cholesterol levels before and after intervention
    - Determines statistical significance
    - Assumes normal distribution of differences
2. Wilcoxon Signed-Rank Test (non-parametric)
    - Alternative if data is not normally distributed
    - Tests for significant changes without assuming normal distribution

□ Online Course Assessment:

A group of 40 students participated in an online programming course, and their coding proficiency was tested both before and after the course. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to evaluate whether the course improved their programming skills?



□ Paired Line Plot

- Tracks individual student progression
- Reveals individual learning trajectories
- Highlights variability in skill improvement

 Boxplot Comparison

- Compares pre and post-course skill distributions
- Shows median, quartiles, and potential outliers
- Provides quick overview of overall group performance

 Histogram of Skill Changes

- Displays distribution of individual skill improvements
- Helps understand the range and frequency of skill gains

 Scatter Plot

- Correlates pre and post-course scores
- Identifies relationship between initial skill and improvement potential

 Cumulative Distribution Function (CDF)

- Shows how skill levels are distributed before and after the course
- Provides insights into collective skill progression

Statistical Testing

1. **Paired t-test**
    - Compares mean scores before and after the course
    - Assumes normally distributed differences
2. **Wilcoxon Signed-Rank Test**
    - Non-parametric alternative
    - Robust to non-normal distributions
3. **Effect Size Calculation**
    - Cohen's d measures practical significance
    - Interprets magnitude of improvement

□ Medical Treatment Study:
A clinical study was conducted to measure the effect of a new medication on blood pressure levels for 50 patients. Blood pressure readings were taken before and after a 6-week treatment period. Select a suitable visualization to represent this data and provide a rationale for your choice. How will you statistically analyze the data, and what steps will you take to determine the medication's effectiveness?

1. Suitable Visualization

Proposed Visualizations:

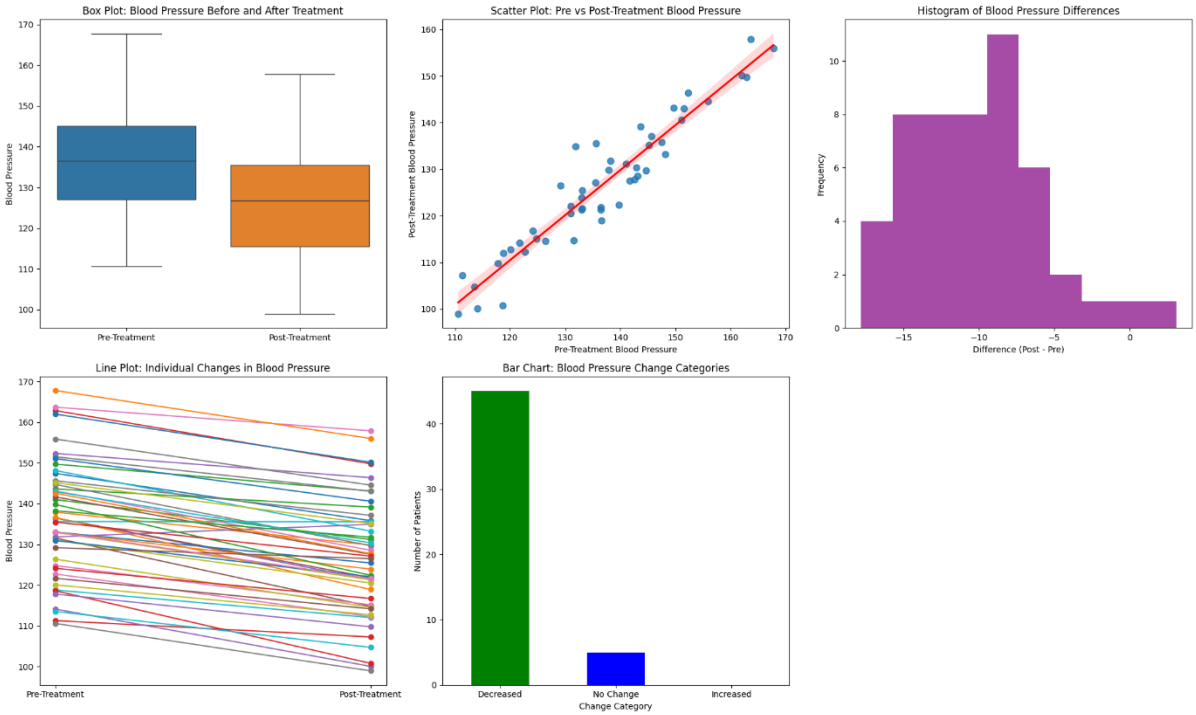
1. **Box Plot**: To compare the distribution of blood pressure readings before and after treatment.
2. **Scatter Plot with Trend Line**: To observe individual patient changes in blood pressure levels (paired data).
3. **Histogram of Differences**: To analyze the frequency distribution of the differences (post-treatment minus pre-treatment).
4. **Line Plot**: To show paired blood pressure levels for each patient before and after treatment.
5. **Bar Chart**: To aggregate changes in blood pressure by group (e.g., patients with positive, no, or negative changes).

---

2. Statistical Analysis

Steps:

1. **Calculate the Differences**: Difference=Post-Treatment BP−Pre-Treatment BP\text{Difference} = \text{Post-Treatment BP} - \text{Pre-Treatment BP}Difference=Post-Treatment BP−Pre-Treatment BP
2. **Test for Normality**: Use the Shapiro-Wilk or Kolmogorov-Smirnov test to assess if the differences are normally distributed.
3. **Paired t-Test or Wilcoxon Signed-Rank Test**:
    - If differences are normally distributed: **Paired t-Test**.
      Formula: t=fracbardsd/sqrtnt = \\frac{\\bar{d}}{s_d / \\sqrt{n}}t=fracbardsd/sqrtn
      Where bard\\bar{d}bard is the mean of the differences, sds_dsd is the standard deviation of differences, and nnn is the number of patients.
    - If differences are not normally distributed: **Wilcoxon Signed-Rank Test**.
4. **Effect Size**: Calculate Cohen's ddd for effect size: d=fracbardsdd = \\frac{\\bar{d}}{s_d}d=fracbardsd



Questionnaire Type

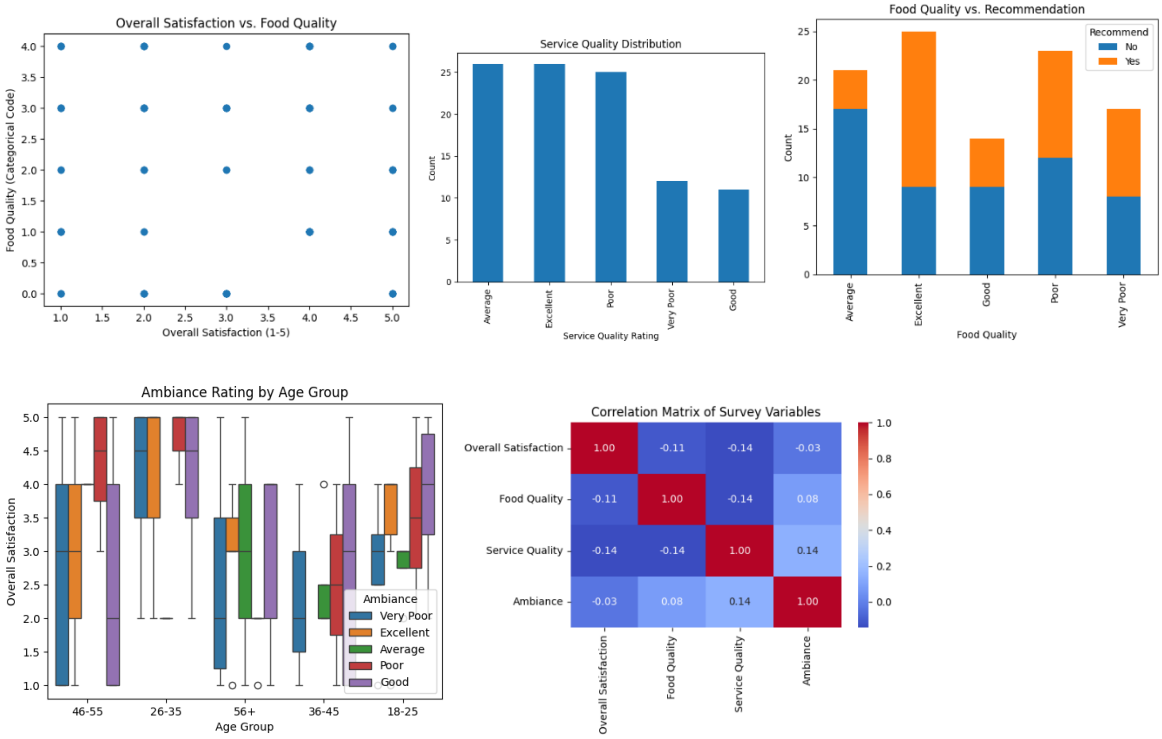
🔍 Restaurant Customer Feedback:
A survey has to be created to evaluate customer satisfaction at a newly opened restaurant. Decide the five main questions that can be asked and the type of data in which the answers will be stored. Further, determine the five most informative data visualizations (of at least two variables) and explain whether the visualizations are for paired data analysis or unpaired.

Five Main Questions:

1. **Overall Satisfaction:** "On a scale of 1 to 5, with 1 being "Very Dissatisfied" and 5 being "Very Satisfied," how satisfied were you with your dining experience at [Restaurant Name]?"
    - **Data Type:** Numerical (Ordinal)
2. **Food Quality:** "How would you rate the quality of the food?"
    - **Data Type:** Categorical (Multiple Choice) - Options: Excellent, Good, Average, Poor, Very Poor
3. **Service Quality:** "How would you rate the service you received from our staff?"
    - **Data Type:** Categorical (Multiple Choice) - Options: Excellent, Good, Average, Poor, Very Poor
4. **Ambiance:** "How would you rate the ambiance of the restaurant?"
    - **Data Type:** Categorical (Multiple Choice) - Options: Excellent, Good, Average, Poor, Very Poor

5. **Would you recommend this restaurant to friends and family?**

    - **Data Type:** Categorical (Binary) - Options: Yes, No



Five Most Informative Data Visualizations:

1. **Scatter Plot: Overall Satisfaction vs. Food Quality**

    - **Paired Data:** Yes

    - **Explanation:** This visualization shows the relationship between overall satisfaction and perceived food quality. It reveals if higher food quality leads to greater satisfaction.

2. **Bar Chart: Service Quality Distribution**

    - **Unpaired Data:** Yes

    - **Explanation:** This chart displays the frequency of different responses regarding service quality, providing insights into the overall service experience.

3. **Stacked Bar Chart: Food Quality vs. Recommendation**

    - **Paired Data:** Yes

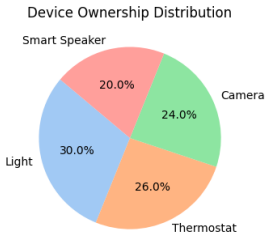
    - **Explanation:** This chart combines food quality ratings with the customer's recommendation. It shows which food quality levels are associated with higher chances of recommending the restaurant.

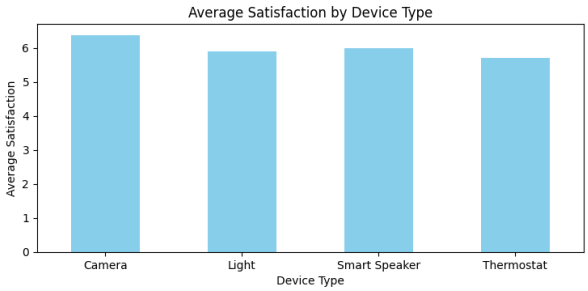
4. **Box Plot: Ambiance Rating by Age Group**

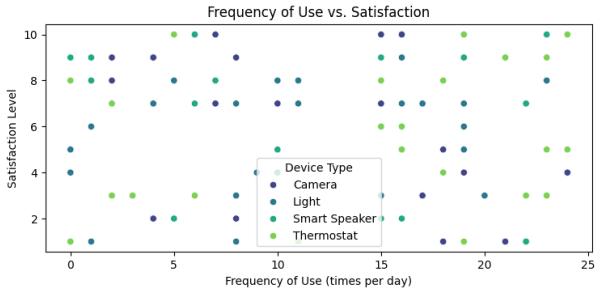
- **Unpaired Data:** Yes
- **Explanation:** This plot analyzes the distribution of ambiance ratings across different age groups, highlighting potential variations in preferences.

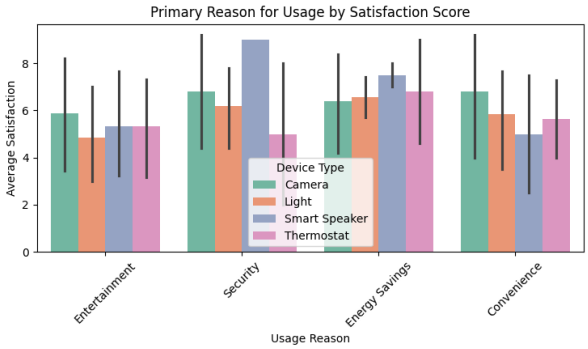
5. **Heatmap: Correlation Matrix of All Survey Variables**
    - **Paired Data:** No
    - **Explanation:** This heatmap displays the correlation coefficients between all survey variables. It reveals potential relationships and patterns between different aspects of the dining experience.

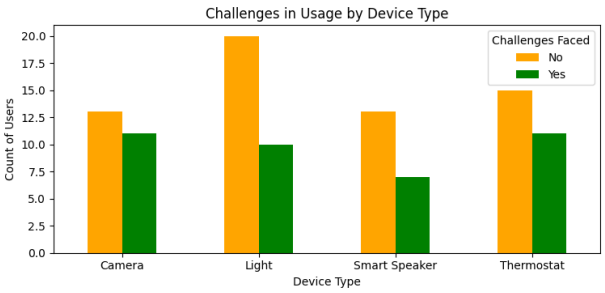
A survey is being developed to explore consumer feedback about the adoption and usability of smart home devices. Decide on five critical survey questions and specify the type of data expected. Suggest five informative visualizations (using at least two variables) and explain whether the visualizations are based on paired or unpaired data analysis.











Survey Questions and Data Types

1. **Question 1:** What type of smart home devices do you currently own? (e.g., Smart Speakers, Thermostats, Cameras, Lights)
    - **Data Type:** Categorical (Nominal).

2. **Question 2:** On a scale of 1 to 10, how satisfied are you with the usability of your smart home devices?
    - **Data Type:** Ordinal (Likert Scale).
3. **Question 3:** How frequently do you use your smart home devices per day? (e.g., 0-24 times)
    - **Data Type:** Quantitative (Discrete).
4. **Question 4:** What is the primary reason for using smart home devices? (e.g., Convenience, Security, Energy Savings, Entertainment)
    - **Data Type:** Categorical (Nominal).
5. **Question 5:** Have you faced any challenges in setting up or using your smart home devices? (Yes/No)
    - **Data Type:** Categorical (Binary).

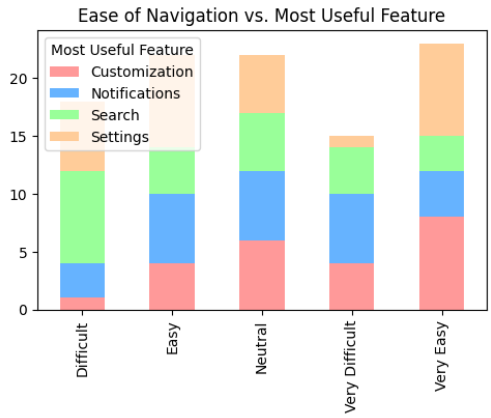
Suggested Visualizations and Data Types

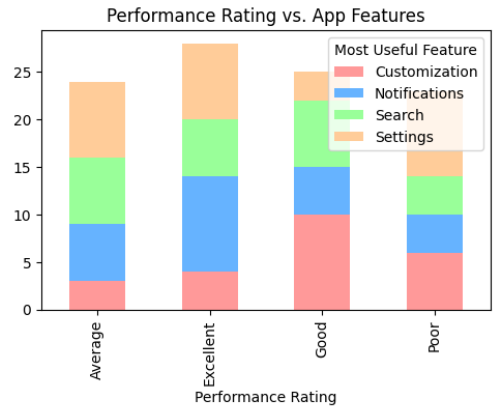
1. **Device Ownership Distribution (Pie Chart)**
    - **Variables:** Device type (Categorical) and percentage distribution.
    - **Type:** Unpaired Data.
2. **Satisfaction Score by Device Type (Bar Chart)**
    - **Variables:** Device type (Categorical) and average satisfaction score (Quantitative).
    - **Type:** Paired Data.
3. **Frequency of Use vs. Satisfaction (Scatter Plot)**
    - **Variables:** Frequency of use (Quantitative) and satisfaction score (Quantitative).
    - **Type:** Paired Data.
4. **Primary Reason for Usage by Satisfaction Score (Stacked Bar Chart)**
    - **Variables:** Usage reason (Categorical) and satisfaction score (Quantitative).
    - **Type:** Paired Data.
5. **Challenges in Usage by Device Type (Grouped Bar Chart)**
    - **Variables:** Device type (Categorical) and proportion of users reporting challenges (Categorical).
    - **Type:** Paired Data.

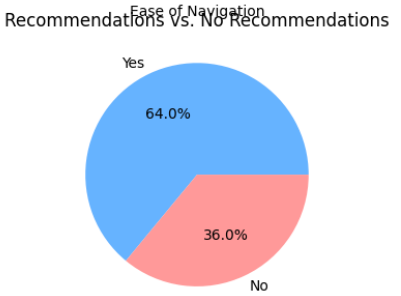
Explanation of Visualizations

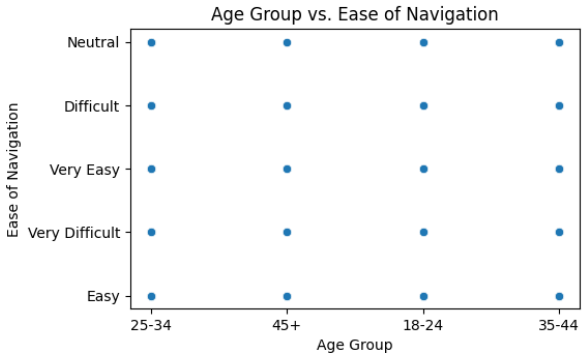
1. **Device Ownership Distribution (Pie Chart)**
    - **Insight:** Highlights the most commonly owned smart home devices, helping companies identify popular product categories.
    - **Data Type:** Unpaired.
2. **Satisfaction Score by Device Type (Bar Chart)**
    - **Insight:** Reveals which device types have higher satisfaction levels, informing product design improvements.
    - **Data Type:** Paired.
3. **Frequency of Use vs. Satisfaction (Scatter Plot)**
    - **Insight:** Examines the relationship between usage frequency and satisfaction, helping assess usability.
    - **Data Type:** Paired.
4. **Primary Reason for Usage by Satisfaction Score (Stacked Bar Chart)**
    - **Insight:** Shows how different usage motivations relate to satisfaction across device types.
    - **Data Type:** Paired.
5. **Challenges in Usage by Device Type (Grouped Bar Chart)**
    - **Insight:** Analyzes the prevalence of challenges, segmented by device type, to identify areas needing improvement.
    - **Data Type:** Paired.

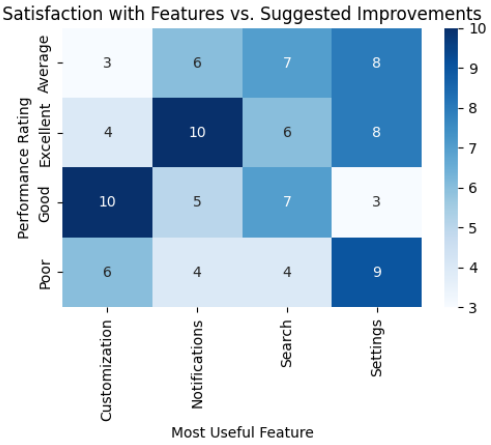
□ Mobile App Usability Study:
A survey needs to be designed to collect user feedback about a mobile app's usability. Identify five key questions to include and specify the type of data in which responses will be stored. Also, propose five informative data visualizations (of at least two variables) and clarify whether they involve paired data analysis or unpaired.











1. **How easy was it to navigate through the app?**
    - **Type of Data:** Ordinal (Scale from "Very Easy" to "Very Difficult" with intermediate ratings like "Easy", "Neutral", "Difficult").
2. **What feature did you find most useful in the app?**
    - **Type of Data:** Categorical (e.g., "Search", "Notifications", "Settings", "Customization").
3. **How would you rate the overall performance of the app (e.g., speed, responsiveness)?**
    - **Type of Data:** Ordinal (Scale from "Excellent" to "Poor").
4. **What improvements would you suggest for the app?**

- **Type of Data:** Text (Open-ended question for feedback).

5. **Would you recommend this app to others?**
    - **Type of Data:** Nominal (Binary: "Yes", "No").

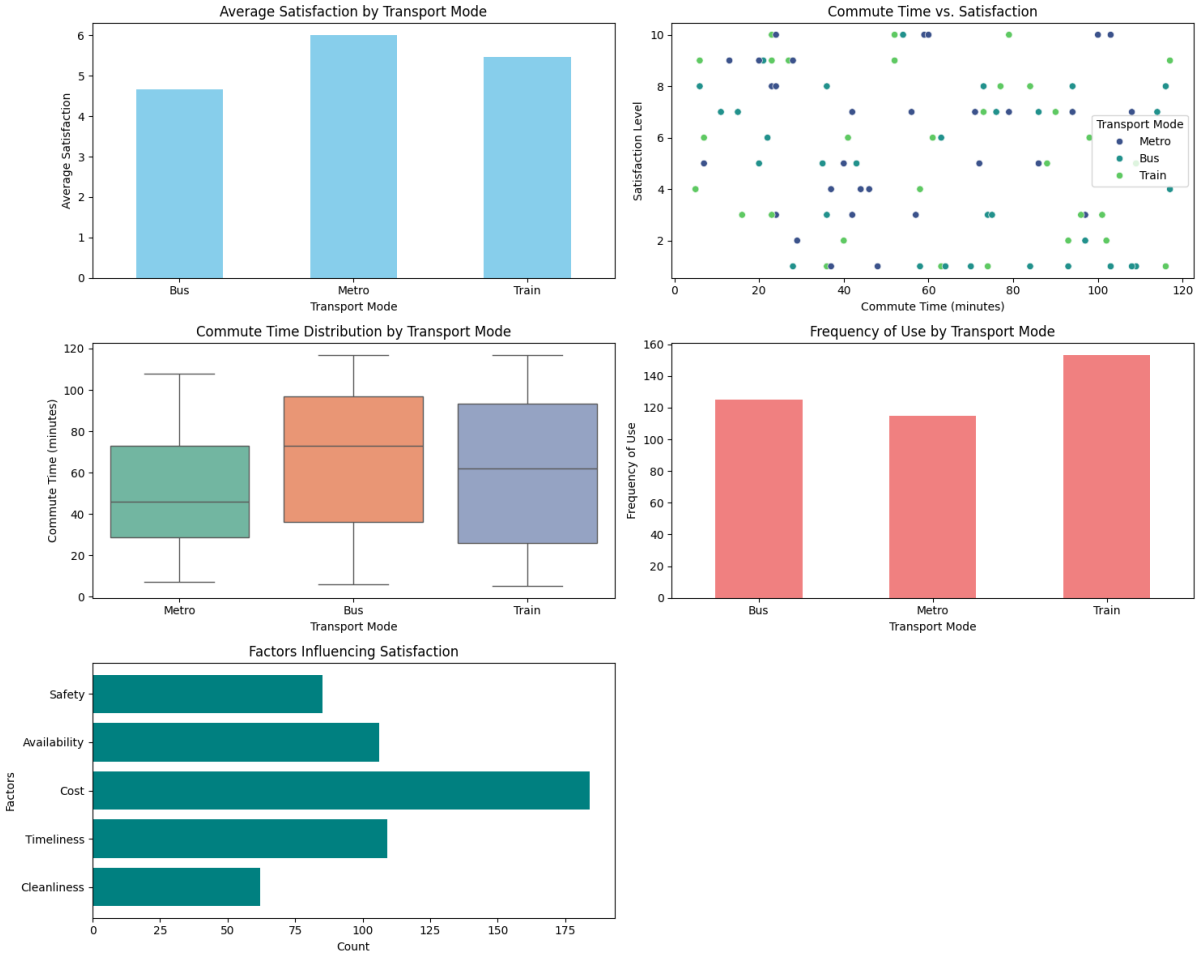
Informative Data Visualizations

Here are five proposed visualizations, including whether they pertain to paired or unpaired data:

1. **Bar Chart: Ease of Navigation vs. Most Useful Feature**
    - **Data Analysis Type:** Unpaired (Comparing the distribution of two categorical variables: "Ease of Navigation" and "Most Useful Feature").
2. **Stacked Bar Chart: Performance Rating vs. App Features**
    - **Data Analysis Type:** Paired (Comparing performance ratings for different app features to assess which feature has the highest/lowest performance rating).
3. **Pie Chart: Recommendations vs. No Recommendations**
    - **Data Analysis Type:** Unpaired (Comparing binary responses "Yes" and "No" for whether users would recommend the app).
4. **Scatter Plot: Age Group vs. Ease of Navigation**
    - **Data Analysis Type:** Paired (Assessing how ease of navigation correlates with different age groups).
5. **Heatmap: Satisfaction with Features vs. Suggested Improvements**
    - **Data Analysis Type:** Paired (Assessing the correlation between satisfaction levels for different features and suggestions for improvements).

 Public Transport

A survey is being created to assess the satisfaction of commuters with a city's public transport services. Define five key questions for the survey and specify the type of data each will collect. Suggest five visualizations (involving at least two variables) and clarify whether these visualizations are for paired or unpaired data analysis.



1. **Question 1:** How often do you use public transport in a week?
    - **Data Type:** Quantitative (Continuous or Discrete depending on format).
    - Example Responses: 0-7 times.
2. **Question 2:** On a scale of 1 to 10, how satisfied are you with the cleanliness of public transport?
    - **Data Type:** Ordinal (Likert Scale).
3. **Question 3:** What is your primary mode of public transport? (e.g., Bus, Train, Metro)
    - **Data Type:** Categorical (Nominal).
4. **Question 4:** How much time does your average commute take using public transport? (in minutes)
    - **Data Type:** Quantitative (Continuous).
5. **Question 5:** What factors influence your satisfaction with public transport the most? (Select all that apply: Cleanliness, Timeliness, Cost, Availability, Safety)
    - **Data Type:** Categorical (Multiple Choice).

Suggested Visualizations and Data Types

1. **Average Satisfaction Score by Transport Mode (Bar Chart)**
    - **Variables:** Mode of transport (Categorical) and Average satisfaction (Quantitative).
    - **Type:** Paired Data.
2. **Commute Time vs. Satisfaction Level (Scatter Plot)**
    - **Variables:** Commute time (Quantitative) and Satisfaction (Quantitative).
    - **Type:** Paired Data.
3. **Distribution of Commute Times (Box Plot)**
    - **Variables:** Commute time grouped by transport mode (Quantitative, Categorical).
    - **Type:** Unpaired Data.
4. **Frequency of Use by Transport Mode (Stacked Bar Chart)**
    - **Variables:** Transport mode (Categorical) and Frequency (Quantitative).
    - **Type:** Paired Data.
5. **Top Factors Influencing Satisfaction (Horizontal Bar Chart)**
    - **Variables:** Count of responses per factor (Categorical).
    - **Type:** Unpaired Data.

Explanation of Visualizations

1. **Average Satisfaction Score by Transport Mode**
    - **Insight:** Helps identify which transport mode receives the highest satisfaction, aiding in mode-specific improvements.
2. **Commute Time vs. Satisfaction Level**
    - **Insight:** Analyzes how commute duration impacts satisfaction, highlighting potential pain points for long commutes.
3. **Distribution of Commute Times**
    - **Insight:** Provides a comparative view of commute time variability across modes, which can inform scheduling or frequency optimization.
4. **Frequency of Use by Transport Mode**

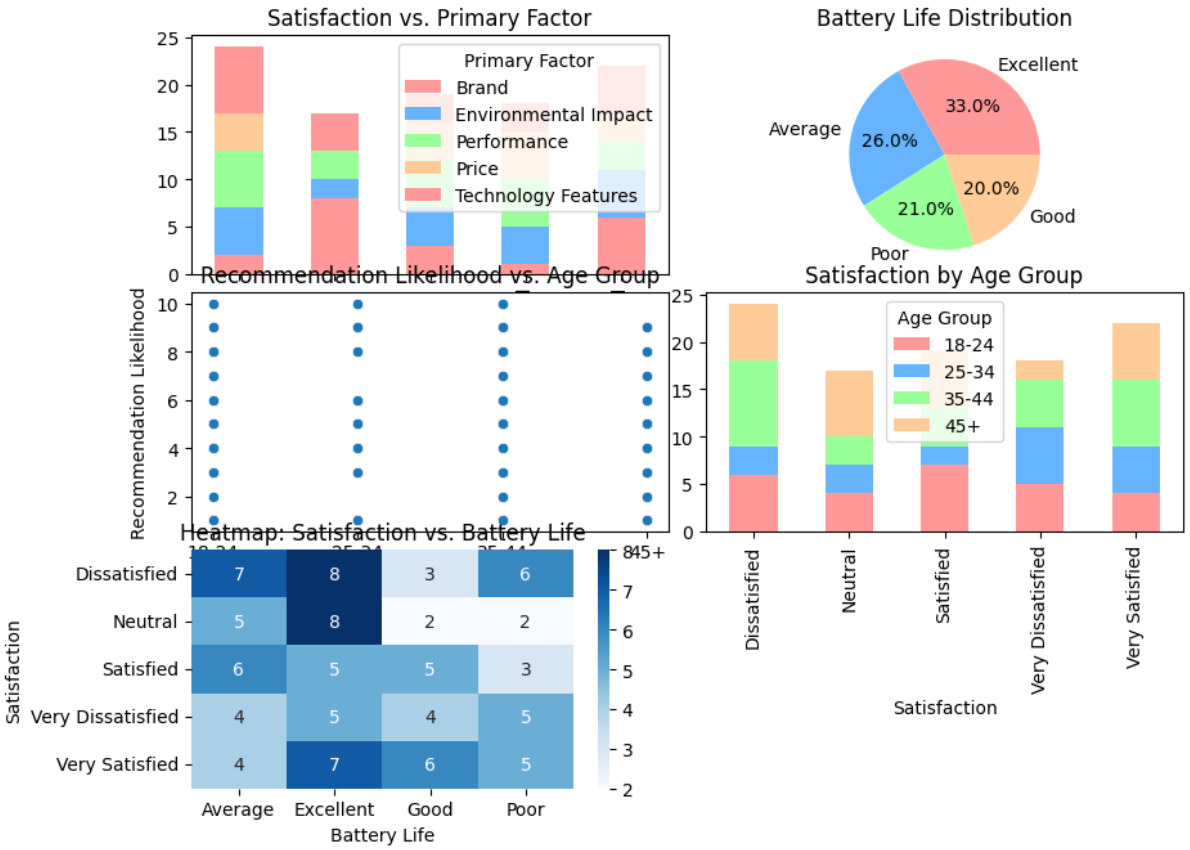
- **Insight:** Identifies the most frequently used transport mode, helping prioritize resource allocation.

5. **Top Factors Influencing Satisfaction**
    - **Insight:** Highlights key satisfaction drivers, enabling targeted improvements in service quality.

□ Electric Vehicle (EV) User Feedback:
A survey is being developed to evaluate consumer feedback about a new electric vehicle model. Define the five main questions to include and the type of data expected from their answers. Propose five effective visualizations (of at least two variables) and indicate whether these visualizations pertain to paired or unpaired data analysis.

1. Survey Questions and Expected Data Types

1. **How satisfied are you with the performance of the electric vehicle (EV)?**
    - **Type of Data:** Ordinal (e.g., "Very Satisfied", "Satisfied", "Neutral", "Dissatisfied", "Very Dissatisfied")
    - This is categorical data with an inherent order.
2. **What is the primary factor influencing your decision to purchase the EV?**
    - **Type of Data:** Nominal (e.g., "Price", "Performance", "Brand", "Environmental Impact", "Technology Features")
    - This is categorical data without any inherent order.
3. **How do you rate the battery life of the vehicle?**
    - **Type of Data:** Ordinal (e.g., "Excellent", "Good", "Average", "Poor")
    - This is categorical data with a ranking.
4. **How likely are you to recommend this EV to others?**
    - **Type of Data:** Interval/Ratio (e.g., scale of 1 to 10, where 1 = "Not Likely" and 10 = "Very Likely")
    - This is numerical data representing a continuous scale.
5. **What is your age group?**
    - **Type of Data:** Ordinal/Nominal (e.g., "18-24", "25-34", "35-44", "45+")
    - This is categorical data based on age ranges.



---

2. Proposed Visualizations

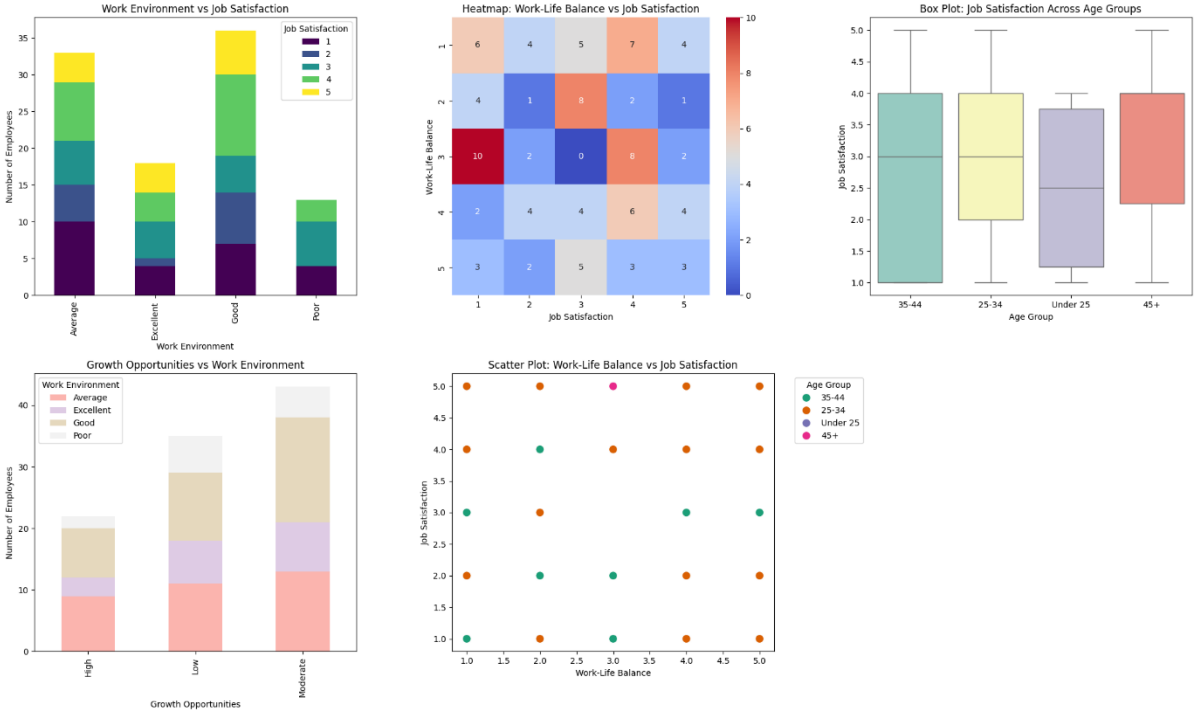
1. **Bar chart of satisfaction level vs. primary factor influencing purchase**
    - **Purpose:** To analyze how different satisfaction levels correlate with the reasons people buy the EV.
    - **Data Type:** Paired data analysis (Satisfaction level and the factors influencing purchase).
2. **Pie chart for battery life ratings**
    - **Purpose:** To show the distribution of battery life ratings given by respondents.
    - **Data Type:** Unpaired data analysis (Battery life ratings as a single variable).
3. **Scatter plot of recommendation likelihood vs. age group**
    - **Purpose:** To show the relationship between how likely people are to recommend the EV and their age group.
    - **Data Type:** Paired data analysis (Likelihood to recommend and Age group).
4. **Stacked bar chart of satisfaction by age group**

- **Purpose:** To compare satisfaction levels across different age groups.
- **Data Type:** Paired data analysis (Satisfaction and Age group).

5. **Heatmap of satisfaction and battery life ratings**
   - **Purpose:** To explore the correlation between satisfaction with the vehicle and battery life.
   - **Data Type:** Paired data analysis (Satisfaction level and Battery life rating).

□ Employee Satisfaction Survey:

A survey is to be designed to measure employee satisfaction within an organization. Decide on five key questions and specify the type of data for each answer. Then, select five visualizations (with at least two variables) that will best represent the results and explain if the visualizations are for paired data analysis or unpaired.



Key Questions:

1. **How would you rate the work environment?** *(Categorical: 'Poor', 'Average', 'Good', 'Excellent')*
2. **How satisfied are you with your job?** *(Ordinal: Scale of 1 to 5)*
3. **How would you rate your work-life balance?** *(Ordinal: Scale of 1 to 5)*
4. **How do you perceive growth opportunities in the organization?** *(Categorical: 'Low', 'Moderate', 'High')*

5. **What is your age group?** *(Categorical: 'Under 25', '25-34', '35-44', '45+')*

Visualizations:

1. **Bar Chart**: Work Environment vs Job Satisfaction (Paired Analysis)
2. **Heatmap**: Work-Life Balance vs Job Satisfaction (Paired Analysis)
3. **Box Plot**: Job Satisfaction Across Age Groups (Unpaired Analysis)
4. **Stacked Bar Chart**: Growth Opportunities vs Work Environment (Paired Analysis)
5. **Scatter Plot**: Work-Life Balance vs Job Satisfaction (Paired Analysis)