Key Drivers of Consumer Buying Patterns: Location, Promotions, Demographics, and Loyalty

Figure 1: Distribution of Consumer Preferences by Product Category

- **Description of the figure**: This figure is a bar chart showing the most preferred categories by shoppers. It presents the distribution of commonly purchased categories across the entire dataset. Clothing is the most popular category, while outerwear is the least preferred.
- **Mark and Channels**:
    - **Mark**: Bar chart
    - **Channels**:
        - **X-axis**: Product Category (N), showing each category (Clothing, Accessories, Footwear, Outerwear).
        - **Y-axis**: Count of Purchases (Q), representing the number of purchases in each category.
        - **Color**: Not used in this visualization, as the focus is on clear comparison between categories.
- **Discussion**: This bar chart was chosen to easily visualize and compare the frequency of purchases across different product categories. The clear and straightforward nature of the bar chart makes it effective for understanding which categories dominate the shopping preferences of the dataset. The simplicity of this visualization allows viewers to quickly grasp which categories are most popular and which are less favored, aiding in targeted marketing and product planning.

Figure 2: Distribution of Age

- **Description of the figure**: This figure is a histogram showing the count of purchasers by age. Each bin represents five years, with eleven buckets encompassing the entire dataset. Naturally, there are fewer younger buyers, but the count remains remarkably stable between ages 25 and 70.
- **Mark and Channels**:
    - **Mark**: Histogram
    - **Channels**:
        - **X-axis**: Age (N), where each bin represents a five-year age range (e.g., 18-22, 23-27, 28-32, etc.). This breaks down the distribution of age groups within the dataset.
        - **Y-axis**: Count of Purchasers (Q), showing the number of individuals who fall within each age range.
- **Discussion**: The histogram was chosen to effectively display the distribution of ages among purchasers in the dataset. This visualization method is particularly useful for identifying the frequency of purchases across different age ranges and understanding where the majority of consumers fall. The use of five-year bins helps clarify trends without overwhelming viewers with too much data in each bar. The stable distribution between ages 25 and 70 suggests a consistent

pattern of purchasing across middle to later adulthood, indicating that brand loyalty or preference for certain product categories may extend throughout this life stage.

Figure 3: Distribution of Purchase Amount

- **Description of the figure**: This figure is a histogram showing the count of basket size across the dataset. Ranging from $20 to $100, the figure indicates a small parabolic trend, where more purchases occur at lower basket sizes ($20-$40) and higher basket sizes ($85+), while baskets in the middle occur with less frequency. This informs us that customers are often making large or small purchases, putting one or two things in their cart, or many.
- **Mark and Channels**:
    - **Mark**: Histogram
    - **Channels**:
        - **X-axis**: Purchase Amount ($), divided into bins of $5 (e.g., 20-25, 25-30, etc.), covering the range from $20 to $100.
        - **Y-axis**: Count of Purchases (Q), representing the number of purchases made within each basket size range.
- **Discussion**: The histogram was chosen to display the distribution of basket sizes, providing insights into shopping behavior patterns. The small parabolic shape reveals that there is a preference for either very large or very small basket sizes, with fewer purchases in the middle. This trend suggests that shoppers are likely making impulse buys or planned large purchases, indicating different shopping strategies and habits. The histogram is effective in conveying this behavior, making it easier for store owners to understand how to encourage different purchasing behaviors among their customers.

Figure 4: Items Purchased By Season

Figure 4.1: Bar plot

- **Description of the figure**: This chart shows the items purchased by season, where each item is colored by category. The viewer is able to switch between seasons or view them in the aggregate to drill down onto more specific data. This stacked bar chart shows that generally, clothing are the highest purchased items, making up four out of the five most purchased items across all seasons. As one would expect, the outerwear category sees higher sales in the Fall and Winter compared to the summer and fall. That being said, the dataset also comes with some key surprises. For instance, Pants are the leading item in the summer, and skirt is one of the lowest selling items of the summer. This is potentially due to people planning their shopping ahead (buying pants for the fall because they already bought skirts in the spring, which is corroborated in the data).
- **Mark and Channels**:
    - **Mark**: Stacked Bar Chart
    - **Channels**:
        - **X-axis**: Season (N), showing the different seasons (Fall, Winter, Spring, Summer).
        - **Y-axis**: Item Purchased (N), displaying items like Pants, Skirt, etc., listed in descending order of purchase frequency.

        - **Color**: Category (N), using different colors to distinguish between Clothing, Accessories, Footwear, and Outerwear.
        - **Tooltip**: Showing Item Purchased, Count, Category, and Season for a detailed view of data.
- **Discussion**: This visualization was chosen to illustrate the seasonal variation in purchases, providing a clear understanding of how preferences change throughout the year. The stacked bar chart is particularly effective in showing the breakdown of item purchases by season, highlighting patterns and trends that may not be immediately obvious. The use of color by category allows viewers to see which product types are more or less popular during specific seasons. The ability to interact with the chart—switching between seasons—adds another layer of depth to the analysis, enabling a detailed view of seasonal purchasing behavior.

Figure 4.2: Boxplot

- **Description of the figure**: This figure is a boxplot that displays the distribution of purchases across different seasons for each product category (Clothing, Accessories, Footwear, Outerwear). The viewer can easily see the variation in purchasing patterns, such as median values and spread across seasons. The whiskers represent the range of purchases, while the interquartile range (IQR) provides insights into the central tendency of data. By grouping the data by season, this boxplot allows for an understanding of how purchasing behavior changes throughout the year, identifying peaks in sales for different product categories. It complements the bar chart by providing a deeper look into the spread and variability of purchases by season, beyond just the count.
- **Mark and Channels for Figure 4.2**:
    - **Mark**: Boxplot
    - **Channels**:
        - **X-axis**: Season (N), showing the different seasons (Fall, Winter, Spring, Summer).
        - **Y-axis**: Item Purchased (N), listing items like Pants, Skirt, etc., for each category.
        - **Color**: Category (N), using different colors to distinguish between Clothing, Accessories, Footwear, and Outerwear.
        - **Tooltip**: Displaying Item Purchased, Season, Count, and Category for detailed insights.
- **Discussion**: The boxplot was chosen to provide a detailed view of the spread and central tendency of purchases across different seasons. This visualization technique helps in understanding the variability within categories, such as how some items may have a larger range of purchases in certain seasons. By grouping data by season, the boxplot highlights seasonal trends and allows viewers to discern the frequency and variation of purchases, aiding in inventory management and marketing strategies based on seasonal demands.

Figure 5: Relationship between Demographics and Purchases

- **Description of the figure**: For store owners, knowing the demographics for shopping behaviors and trends is important in order to target consumers for online and physical advertising. The chart

shows the sales total purchase amounts for the primary genders, divided into the purchase categories. It can be seen that males spend nearly double overall than females, outspending them in every product category. However, for both males and females, the relative spending per category is the same, i.e., clothing is the largest category, followed by accessories, footwear, then outerwear. This suggests that males spend more overall when shopping, with the largest product category being clothing.

- **Mark and Channels**:
    - **Mark**: Side-by-Side Bar Charts
    - **Channels**:
        - **X-axis**: Product Category (N), with items like Clothing, Accessories, Footwear, Outerwear, plotted for both genders.
        - **Y-axis**: Total Purchase Amount ($), showing the total spending by gender for each category.
        - **Color**: Gender (N), using different colors (e.g., blue for male, pink for female) to differentiate between the groups.
        - **Tooltip**: Displaying Product Category, Total Purchase Amount, Gender, and relative spending proportions.
- **Discussion**: The side-by-side bar charts were chosen to effectively compare the spending patterns between males and females across different product categories. This visualization clearly shows the differences in total spending while also emphasizing that the relative preference per category (e.g., Clothing > Accessories > Footwear > Outerwear) remains consistent across genders. The use of color not only enhances visibility but also makes the comparison intuitive. This allows store owners to tailor marketing strategies based on gender-specific preferences and spending habits, providing a strategic advantage in targeting consumers more effectively.

Explanation of Work Split: In the pre-processing phase, Darin is responsible for preparing the dataset, which includes: cleaning and handling missing values or outliers, ensuring consistency in data types across columns, encoding categorical variables such as gender or shopping preferences using label or one-hot encoding, and normalizing or scaling continuous variables when needed.

During the exploratory data analysis (EDA) stage, Bea takes charge of summarizing the dataset's key characteristics (e.g., mean, median, standard deviation), analyzing the distribution of continuous variables like purchase amount and age, and creating a correlation matrix to investigate relationships among factors like promotions, loyalty, demographics, and purchase behavior. She will also generate static visualizations such as bar charts, histograms, and heatmaps to illustrate data patterns. Michael complements this by creating interactive visualizations (using tools like Plotly, Altair Tooltip, or D3) that enable deeper exploration of demographic factors such as age, gender, and location.

In the analysis phase, all team members collaborate to investigate the influence of location on purchase frequency and basket size, analyze the impact of promotions on purchase behavior through time-series analysis, assess demographic correlations with shopping habits, and use regression models to evaluate how loyalty programs affect purchase frequency and basket size. Additionally, the entire team will work together to design and develop the project's website, ensuring that it effectively presents our findings and visualizations.

We all worked together to build out website, final deliverables, and presentation materials.

Data Description: The dataset is from Kaggle originally composed by Sir Sourav Banerjee at CogniTensor and last updated a year ago. It contains two tables covering demographic information, purchase history, product preference, and preferred shopping channels with the first table focusing on customer behavior while the second on customer trends. Both contain about 3900 observations across 18 columns, including both categorical (such as gender, size, season) and continuous variables (purchase amount, age). Data is licensed under Creative Commons and thus is free to use.

Notes: As per feedback, we tried out scatter and hexplot for the first 2nd and 3rd plots, but there did not appear to be any clearly defined relationship. We understand there isn't an obvious trend in the plots 2 and 3, but chose to include the descriptive information to build more context around our dataset.