INTRODUCTION TO THE DATASET

For my project, I selected a dataset that is not overly complex but allows for the completion of all project steps effectively. The dataset comprises the following columns: Customer ID, Education, Age, Marital Status, Job, Product Category, Quantity, and Price per unit. It provides information about customers and their purchases.

SUMMARY OF THE DATA CLEANING/PREPROCESSING STEPS

I began by performing data cleaning and preprocessing. Initially, I addressed missing values, removing 20 rows from the dataset due to its large size (over 10 thousand rows). Subsequently, I identified and removed four outlier rows in the age column, as these entries appeared unrealistic. Next, I standardized the capitalization in the Education, Job, and Marital Status columns to lowercase to ensure consistency. Additionally, I renamed the value "unknown" to "other" in the Job and Education columns. Before proceeding with analysis, I checked for duplicates and added a new column, "Total Cost," calculated as the product of Price per Unit and Quantity.

KEY FINDINGS FROM THE EDA

1. **Distribution of Key Numerical Variables Using Histograms:** The age distribution appears roughly symmetric overall, with the majority of customers falling within the 25–40 age range. Total cost exhibits a right-skewed distribution, with a predominant frequency of small purchases, and over 75% of purchases totaling under 500 units of currency.

2. **Relationships Between Variables Using Scatter Plots and Correlation Matrices:** No strong association is observed between age and total cost, as evidenced by the scattered data points across the plot. Consequently, there is no significant correlation between age, quantity, and price per unit. Therefore, the relationship between total cost and price or quantity is disregarded, as total cost is a derived value.

3. **Group Comparisons Using Box Plots and Bar Charts:** Students exhibit the highest median total cost, while retired individuals demonstrate the lowest. Notably, entrepreneurs, technicians, services, and self-employed individuals display above-average total costs, making them potential focal points for further analysis.

RESULTS OF THE STATISTICAL TESTS

1. **Hypothesis 1 (Age Distribution):**

Null Hypothesis (H0): The age distribution of customers follows a normal distribution, with a mean age within the range of 25 to 40 years.

Alternative Hypothesis (H1): The age distribution of customers does not follow a normal distribution within the specified age range.

Results: The Shapiro-Wilk test yielded a p-value of 0.0, leading to the rejection of the null hypothesis. Although the age distribution does not adhere to normality, the mean and median ages falling within the 25–40 range provide opportunities for targeting this customer segment.

2. **Hypothesis 2 (Total Cost by Occupation):**

Null Hypothesis (H0): There is no significant difference in the median total cost among different occupations.

Alternative Hypothesis (H1): There is a significant difference in the median total cost among different occupations.

Results: The Kruskal-Wallis test resulted in a test statistic of 477.47 and a p-value of 2.03e-95, leading to the rejection of the null hypothesis. These findings corroborate the initial analysis, indicating significant differences in median costs across various occupations. Consequently, further investigation into the spending behaviors of different occupational groups is warranted, along with strategies to optimize spending among retired individuals.

CONCLUSIONS FROM THE ANALYSIS

Age Distribution and Spending Patterns:

The analysis revealed that the age distribution of customers is not normally distributed, with a significant deviation from normality according to the Shapiro-Wilk test. Despite this, the majority of customers fall within the age range of 25–40 years, indicating a prime demographic for targeted marketing and product offerings.

Occupational Influence on Spending Behavior:

Significant differences in median total costs among various occupations were observed, as evidenced by the rejection of the null hypothesis in the Kruskal-Wallis test. Notably, students exhibited the highest median total costs, while retired individuals showed the lowest.

Entrepreneurs, technicians, services, and self-employed individuals displayed above-average spending tendencies, highlighting them as potential key demographics for tailored marketing strategies.

Potential Areas for Further Investigation:

The analysis identified several avenues for further exploration. Understanding the underlying factors contributing to the spending behaviors of different occupational groups, particularly retirees, could provide valuable insights for developing targeted promotional campaigns or product offerings. Additionally, investigating the spending patterns of outliers, such as students with exceptionally high spending, may uncover unique preferences or consumption trends that can be leveraged for strategic advantage.

Implications for Marketing and Business Strategy:

The insights derived from the analysis have significant implications for marketing and business strategy. By targeting specific age demographics and occupational groups with tailored marketing campaigns and product offerings, businesses can optimize their marketing efforts and maximize revenue potential. Furthermore, understanding the factors driving spending behaviors can inform the development of personalized customer experiences and loyalty programs, ultimately fostering stronger customer relationships and brand loyalty.

Continuous Monitoring and Iterative Analysis:

It is essential for businesses to adopt a data-driven approach to marketing and decision-making, continually monitoring customer trends and preferences to adapt strategies in real-time. Iterative analysis and refinement of marketing strategies based on evolving customer insights will be critical for maintaining competitiveness and driving sustainable growth in the dynamic business landscape.

In conclusion, the comprehensive analysis of customer demographics, spending patterns, and occupational influences provides valuable insights for businesses seeking to optimize their marketing strategies, enhance customer engagement, and drive revenue growth in an increasingly competitive marketplace. By leveraging these insights and adopting a proactive and data-driven approach to marketing and business strategy, organizations can position themselves for long-term success and sustained profitability.

INDIVIDUAL REFLECTION

The first challenge I faced was choosing an appropriate dataset that would allow me to complete all the tasks outlined in the project. During this process, I had to switch datasets and start the steps again.

Data cleaning and preprocessing were relatively straightforward tasks for me, and I completed them with ease. However, I found the subsequent steps, including exploratory data analysis (EDA) and statistical analysis, to be quite challenging. Statistics has never been my strong suit, and I struggled with conducting the analyses effectively. Additionally, I felt that my dataset was relatively small, which may have limited the depth of analysis I could perform. Nonetheless, I was able to leverage some knowledge gained from previous programs to navigate through these steps.

Further questions for exploration and additional analyses could include:

Segmentation Analysis:

Explore the possibility of segmenting customers based on demographic characteristics, such as age, education, marital status, and occupation, using advanced clustering techniques like K-means clustering or hierarchical clustering. This segmentation can provide deeper insights into distinct customer groups with unique preferences and behaviors, enabling targeted marketing strategies.

Predictive Modeling:

Develop predictive models to forecast future customer behavior, such as purchase propensity, churn likelihood, or lifetime value, using advanced machine learning algorithms like logistic regression, decision trees, or neural networks. By predicting customer outcomes, businesses can proactively identify and address potential opportunities or risks, optimizing resource allocation and decision-making.

Market Basket Analysis:

Conduct a market basket analysis to uncover associations and patterns in customers' purchase behaviors, identifying frequently co-occurring products or product categories. Advanced techniques such as association rule mining or sequential pattern mining can reveal cross-selling or upselling opportunities, informing product bundling strategies and personalized recommendations.

Sentiment Analysis:

Perform sentiment analysis on customer reviews, feedback, or social media conversations using natural language processing (NLP) techniques. By analyzing customer sentiments and opinions, businesses can gain valuable insights into customer satisfaction levels, identify pain points or areas for improvement, and tailor their communication and service offerings accordingly.

Customer Lifetime Value (CLV) Analysis:

Calculate and analyze customer lifetime value metrics to quantify the long-term profitability of different customer segments. Advanced techniques such as cohort analysis or Markov chain modeling can provide insights into customer retention, acquisition costs, and revenue potential, guiding strategic decisions related to customer acquisition, retention, and loyalty programs.

Geospatial Analysis:

Explore geospatial data to analyze the geographic distribution of customers, identify regional trends or preferences, and optimize store locations or distribution channels. Geospatial techniques such as spatial autocorrelation analysis or hotspot identification can uncover spatial patterns and clusters, informing targeted marketing campaigns and expansion strategies.

Cross-Channel Attribution Modeling:

Implement cross-channel attribution models to assess the impact of different marketing channels and touchpoints on customer conversion and engagement. Advanced attribution models like Shapley value regression or time-series analysis can provide insights into the effectiveness of marketing campaigns across various channels, guiding budget allocation and optimization strategies.

By exploring these additional questions and conducting more advanced analyses, businesses can gain deeper insights into customer behavior, optimize marketing strategies, and drive sustainable growth and competitive advantage in today's dynamic marketplace.