Project Part 1 – Preliminary Data Analysis

Prajakta Pohare

1. Data Wrangling

1.1 Loading the Data

The dataset was loaded into an R data frame named myData using the read.xlsx function. It contains 2240 observations and 28 variables representing customer information, purchasing behavior, and engagement metrics.

1.2 Checking Data Structure

The str(myData) function was used to inspect the structure, confirming that each variable's data type was appropriate. Key variables included numerical fields like Income and Recency, categorical fields like Education and Marital_Status, and date values in Dt_Customer.

1.3 Identifying Missing Values

The is.na() function identified missing values across various columns:

- **Income**: Several missing values were found.
- **Recency**: A few missing values detected.
- **Country** and **Marital_Status**: Some observations lacked entries.

1.4 Data Imputation

- **Numerical Imputation**: Missing values in Income, Recency, and MntWines were replaced with their respective mean values using mean() to ensure minimal data loss.
- **Categorical Imputation**: For categorical fields like Marital_Status and Country, missing values were imputed with the mode. Specifically, Marital_Status was imputed as "Married" and Country as "Spain" (most frequent values).

1.5 Date Conversion

- The Dt_Customer column, initially in numeric format, was converted to Date format with as.Date.numeric(), setting a reference origin date. This allows for easier date-based analysis and transformation.

1.6 Categorical Conversion

- The Education, Marital_Status, and Country columns were converted to factor types using as.factor(), which enables efficient handling of categorical data in R.

2. Data Transformation

2.1 Binning

- The Income variable was binned into three categories: Low, Medium, and High using cut(). The defined bins help categorize income levels for better analysis, producing:
  - Low: Income < 20,000
  - Medium: 20,000 ≤ Income < 70,000
  - High: Income ≥ 70,000
- This segmentation helps in targeting marketing strategies for different income levels.

2.2 Date Transformation

- Extracted the month from Dt_Customer, creating a new column Mnt_Customer representing the month of customer registration.
- Monthly trends show spikes in specific months, suggesting seasonal or campaign-driven engagement patterns.

2.3 Categorical to Numeric Conversion

- Converted Education and Marital_Status into numerical codes to simplify statistical analysis.
- Created dummy variables for each level of Education (e.g., Education_Graduation, Education_PhD) using ifelse(). This facilitates analysis of each education level individually.

3. Summary Measures

3.1 Numerical Variables

- **Mean**: Calculated for Income (52,247.25) and MntWines (303.74) to understand the central tendency of spending and income.
- **Median**: MntFruits median = 8, meaning half the customers spent less than this amount on fruits. A low median for MntFruits suggests limited engagement in fruit purchases.
- **Variance and Standard Deviation**: Variance and standard deviation were computed for NumStorePurchases and NumWebPurchases, revealing their spread and variability in customer purchases. NumStorePurchases variance = 10.57, high variance in store purchases reflects diverse shopping behaviors.
- **Coefficient of Variation**: This metric, found for NumStorePurchases and NumWebPurchases, highlighted relative variability, with NumWebPurchases showing more variation compared to its mean than NumStorePurchases.
- **Range and Interquartile Range (IQR)**: Range and IQR for NumStorePurchases and NumWebPurchases were calculated to measure data spread. NumStorePurchases range = 13, NumWebPurchases range = 27, with similar IQRs. Broad range suggests some customers make far more purchases than others, indicating possible high-value segments.

3.2 Categorical Variables

- **Mode**: Marital_Status and Education modes were determined to be "Married" and "Graduation", respectively.
- **Group Means**: Grouped means were calculated for Income by Kidhome and Teenhome categories, revealing households with children had a lower average income.

3.3 Outlier Detection

- Outliers were identified using box plots for NumStorePurchases and NumWebPurchases. Observations beyond 1.5 * IQR were marked as outliers, and extreme values in spending categories (e.g., MntWines) were excluded to ensure robustness in analysis.

4. Data Visualization

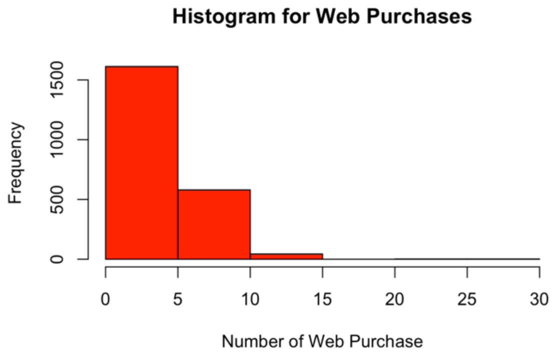
1. Frequency Distributions for Numerical Variables

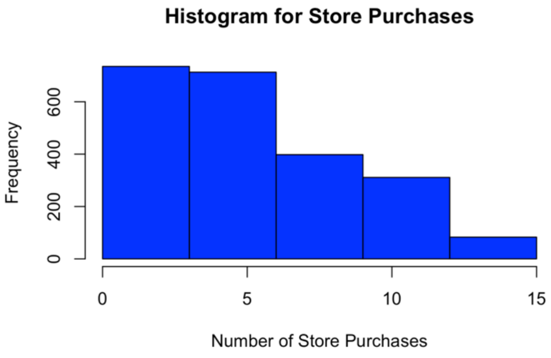
- We created frequency distributions for the Income variable by setting intervals and categorizing income levels into bins. We then generated frequency and proportion tables for these intervals.
- The majority of customers fall within the lower income ranges, with nearly 47.32% earning between 0 and 50,000, and 52.10% earning between 50,000 and 100,000.
- Very few customers are in the higher income ranges (200,000 and above), with many income bins having a frequency of zero.
- This distribution indicates a concentration of customers with lower income levels, suggesting a target demographic within the 0 to 100,000 income range.

2. Frequency Distributions for Categorical Variables

- **Country and Education Level Frequency**: Frequency tables were created for Country and Education to show the distribution across different categories.
- **Spain** has the highest number of customers, with 1,096, indicating a significant concentration of the customer base in this region.
- **Saudi Arabia** and **Canada** also have a substantial number of customers, with 337 and 268 respectively.
- **Mexico** has the lowest representation, with only 3 customers.
- The distribution shows a diverse customer base, with a notable concentration in certain countries, which can guide region-specific marketing and engagement strategies.

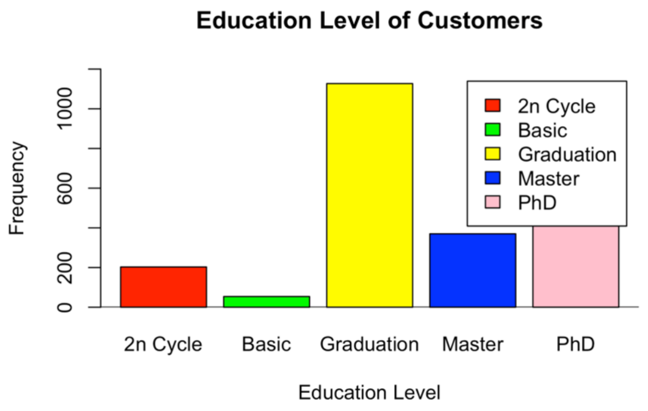
3. Histogram for Numerical Variables





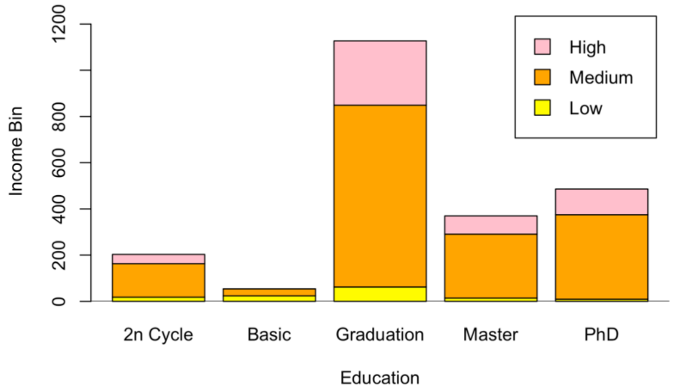
- Histograms were generated for NumWebPurchases and NumStorePurchases.
- The histograms show the frequency of purchases within defined intervals for online and in-store purchases.
- Most customers have made a low number of web purchases, with the highest frequency in the 0–5 range. There is a sharp decline in frequency as the number of purchases increases, suggesting that web purchases are generally infrequent among the customer base.
- Unlike web purchases, store purchases have a more even distribution across the lower purchase ranges. The frequency still declines as the number of purchases increases, but the pattern suggests that in-store purchases are slightly more common than web purchases.

4. Bar Chart for Categorical Variables



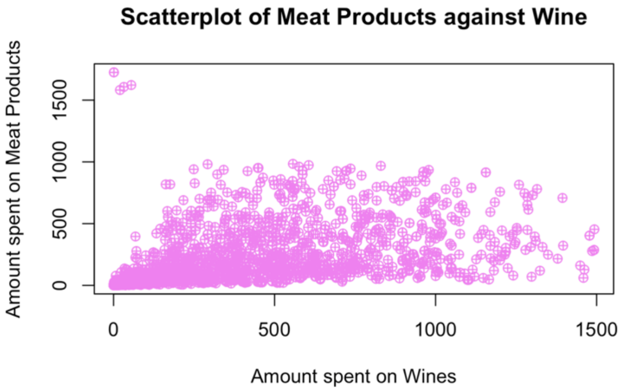
- This bar chart illustrates the distribution of education levels among the customers.
- Customers with a “Graduation” level of education are the most common, followed by those with “Master” and “PhD” degrees. There is a low representation of customers with “Basic” education, indicating a highly educated customer base.

5. Stacked Column Chart for Income and Education



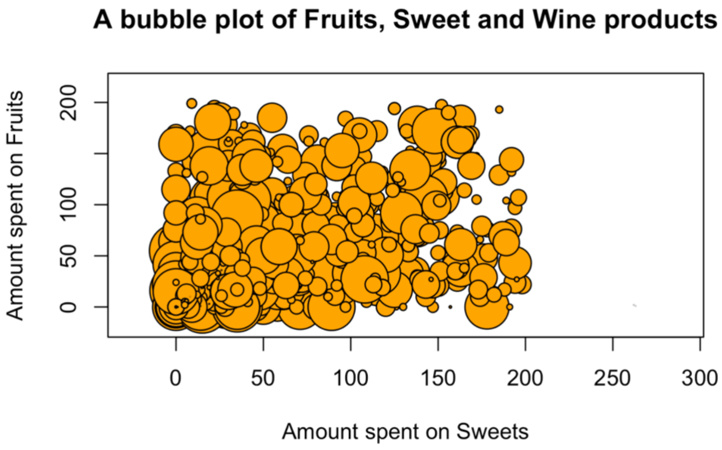
- This chart further illustrates the relationship between income bins and education, with each education level shown as a segment within income bins.
- The variation in colors and sizes across income levels reveals demographic insights, such as whether higher education correlates with higher incomes. The chart indicates that the “Graduation” group has the widest range of income levels, followed by “Master” and “PhD” groups.

6. Scatter Plot for Amount Spent on Meat and Wine Products



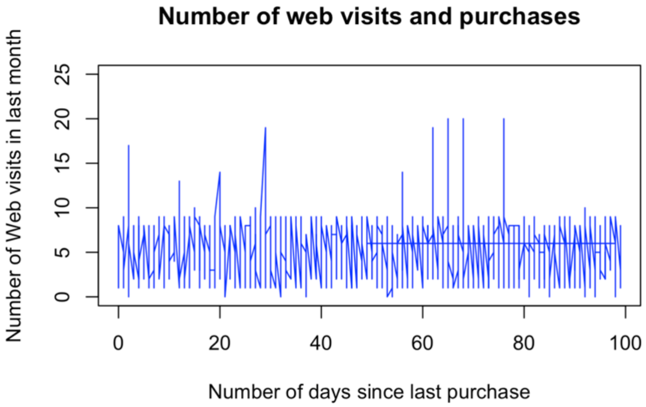
- A scatter plot was used to explore the correlation between spending on meat and wine products.
- There appears to be a weak positive relationship between spending on meat products and wine, suggesting that customers who spend more on wine also tend to spend slightly more on meat products.
- This insight can be valuable for cross-promotional strategies targeting these product categories.

7. Bubble Chart for Amount Spent on Fruits and Sweet Products



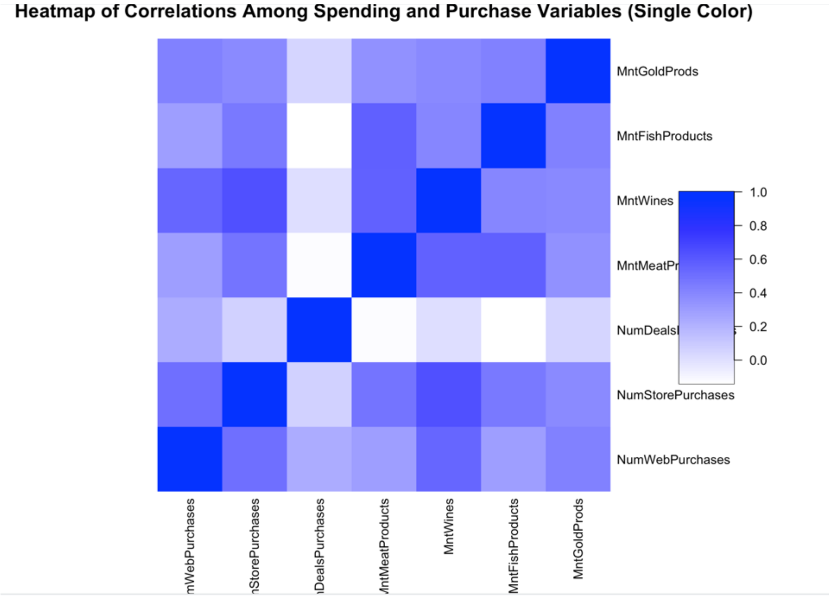
- This chart visualizes spending on fruits and sweet products, with bubble sizes indicating the amount spent on wine.
- Larger bubbles (higher wine spending) are concentrated in the moderate spending range for sweets and fruits, indicating that customers who spend on wine also tend to purchase these items.
- This could be useful for bundling these products in marketing efforts.

8. Line Chart for Web Visits and Recency



- **Line Chart Explanation**: The relationship between recency (days since last purchase) and web visits per month is displayed in a line chart.
- There is a high variance in the number of web visits regardless of recency, though the trend suggests that recent purchasers have higher web visits.
- This suggests that maintaining engagement shortly after a purchase could be effective in driving repeat visits.

9. Heatmap for Spending and Purchase Behaviors



- We created a heatmap to visualize the relationship among various spending and purchasing behaviors, using a single-color gradient to indicate the strength of the correlations. Darker blue shades represent stronger correlations between pairs of variables, while lighter shades indicate weaker or no correlations.
- **Spending Patterns**: The heatmap shows a moderate correlation between spending on meat products and wine (MntMeatProducts and MntWines), suggesting that customers who buy one of these items may be more likely to buy the other. This insight could guide cross-promotional efforts targeting customers interested in both products.
- **Purchase Behavior**: There is also a moderate correlation between NumStorePurchases and NumDealsPurchases, suggesting that customers who frequently shop in-store are more likely to take advantage of deals. This behavior can help identify customer segments that respond well to in-store promotions.
- **Independent Spending**: Most categories, such as MntGoldProds, MntFishProducts, and NumWebPurchases, show little to no correlation with each other or with income, indicating that these categories are largely independent of each other and are not influenced by a common purchasing pattern.

Dataset Citation

Saw, D. (2022). *Marketing dataset* [Data set]. Kaggle. https://www.kaggle.com/datasets/deepaksaw/marketing-dataset