Francesca Marchese, Ian Smith, Emmanuella Acheampong, Andrew Mukurazita

BUS-212A- Report 1

BUS 212A – Advanced Data Analytics

Report 1: Data Wrangling

January 26, 2024

Table of Contents

Introduction

Data Consultancy, Source and Credibility

Customer segmentation is essential to ensuring one's product or service appeals and successfully reaches the "right" consumer. How does one know, though, who, in fact, the "right" consumer is? Before a marketing team can develop an ad campaign to communicate its unique product or services' value proposition, the team must begin by conducting market research to bring their ideal persona to life. A persona is a user-centered, fictionalized character created by a brand or company to represent a user type; the consumer behavior gathered through market research allows a persona to come to life, as buying habits, age demographics and relationships surface, successfully identifying areas of growth and opportunity within the market. Value can then be successfully communicated to the ideal consumer through marketing campaigns and channels of distribution, maximizing efforts and invested resources, while additionally increasing the company's bottom line.

As analytic tools increase in value and popularity, marketing research tools and techniques are evolving; where focus groups and surveys once dominated the field, customer personality analysis is now implored to better understand consumer behavior. Businesses find inherent value in the data's ability to provide detailed information about its customers to ensure marketing efforts are focused, as well as to modify products and services according to specific consumer needs, concerns and behaviors in particular segments. Ultimately, analytical tools are incredibly powerful and impactive at providing CMOs, marketers, and entrepreneurs with data that aids in better understanding target audiences, consumer behavior, opportunities, and of course, overall results. However, tension between growth and efficiency exists, as the knowledge gap between traditional marketers and data scientists widens.[1] Traditional Marketers struggle to fully trust and appreciate how data science can enable and improve decision making; surveys and focus groups certainly have their time and place, growth and efficiency is what the C-Suite expects and traditional methods of analyzing consumer preferences and behavior can be costly and time consuming. Data scientists, however, get caught up in the numbers and often forget the purpose of marketing: to foster personalized and meaningful connections with consumers to build brand awareness all while providing unmatched value. In an evolving and expensive economic climate, marketing budgets are the first to go. In 2022, there was an average 8% reduction in marketing

[1] Boudet, J., Brodherson, M., Robinson, K., & Stein, E. (2023, June 26). *Beyond belt-tightening: How marketing can drive resiliency during uncertain times.* McKinsey & Company.

expenditures across 36 companies[2]. Ultimately, CMOs are expected to do more with less, and, therefore, must simultaneously relieve the tension between growth and efficiency and appeal to the "right" consumer to grow their revenues. Customer personality analysis is exactly what marketers need to increase their company's bottom line, while maximizing their resources and efforts.

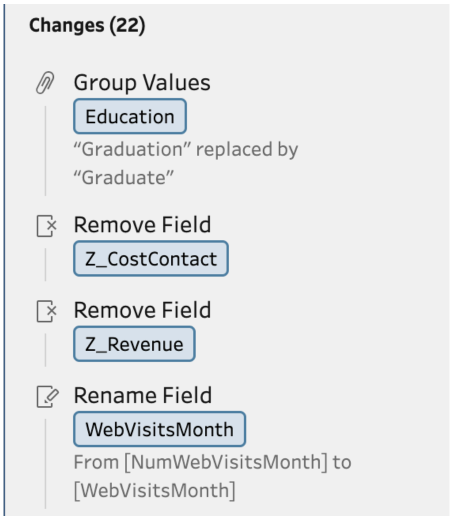
For our consultancy, we have chosen to address the tension between growth and efficiency that exists in part because of underutilization of marketing analytical tools within companies. In this project, our consultancy, DataWise Consultants, is working with MarketSphere Inc., a leading retail company, to unlock insights into their customer base. We've developed several innovative features to enable a more detailed analysis of MarketSphere's customers. Our aim is to employ data analytics techniques on customer personality analysis to determine growth opportunities for companies' in determining the specific needs, concerns and behaviors of their "right" consumer to increase their product value and marketing efforts. We will be using publicly available customer personality data to understand consumer preferences, behaviors and trends across different markets in the United States based on factors such as demographic, income and purchase history. We want to aid businesses in better understanding the competitive advantage marketing analytics can provide.

[2] Checa, A., Heller, C., Stein, E., & Wilkie, J. (2023, April 5). *Modern marketing: Six capabilities for multidisciplinary teams.* McKinsey & Company.

Data Preparation

Quality, Cleaning and Summary

2.1 INITIAL DATA QUALITY AND PREPARATION



Changes (22)

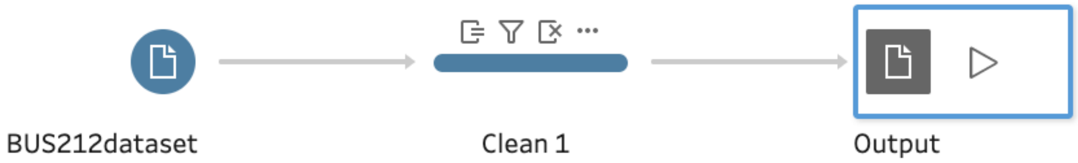
Group Values
Education
“Graduation” replaced by “Graduate”

Remove Field
Z_CostContact

Remove Field
Z_Revenue

Rename Field
WebVisitsMonth
From [NumWebVisitsMonth] to [WebVisitsMonth]

The data used by our consultancy is sourced from a Customer Personality Analysis dataset published to Kaggle, a credible datasource for data related projects, and last updated in 2022.[3] The dataset contains roughly 2,240 observations of 28 variables on customer demographics and household information, product purchase patterns, purchase platform preferences, and customer responses to company offers and campaigns. It had a usability rate of 9.7, indicating its appropriateness for our project. From a first glance, the initial quality of the data from a traditional analysis is high. None of the included columns showed a significant presence of either outliers (less than 5%) or missing values (only 24 in the case of household income). On two occasions, initially, during the cleaning process, two variables were trimmed of immediately apparent outliers. These were the cases in which the three individuals born in the 19th century and the one individual earning above $700,000 were removed, both because age and household income are possible explanatory variables, but also because the behaviors of these customers are not likely to be representative of the overall sample. Outside of this, the largest lapse in initial data usability was a lack of proper column and variable value labeling. This is corrected in the following section.

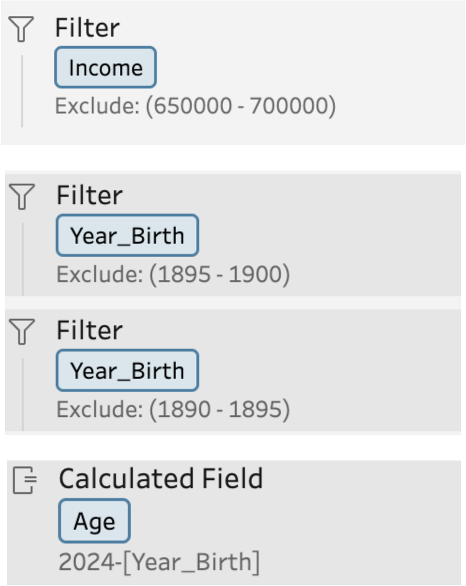


The above diagram shows our steps taken to manipulate the dataset into a more user friendly format. This included a combination of field renaming, field removal in two cases, value regrouping, and the creation of a new field for customer age, calculated as 2024 - *Year_Birth*. The first step in this process was to reconcile the labels of some of our categorical variables. In particular, with the variable *Education*, which initially took the values of “Basic,” “2n Cycle,” “Graduation,” “Master,” and “Phd,”

[3] Data sourced from Kaggle repository: “Customer Personality Analysis: Analysis of Company’s Ideal Customers.” Provided by Dr. Omar Romero-Hernadez, Uploaded by Akash Patel. Last updated 2022. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

the label for “Graduation” was converted to simply “Graduate,” denoting college graduates. Since none of the values for *Education* were initially defined, we have also taken “Basic” to mean an elementary level education, and “2n Cycle” to mean a maximum of a highschool level education. This process was also conducted for *Marital_Status*, in which the self-reported values of “YOLO,” “absurd,” and “alone” were converted to the existing label of “Single.” This speaks, in part, to the inevitable vulnerability of data to the reliability of survey respondents.

The next important step was to rename a number of the fields, which were often convoluted and unintuitive. This included both dramatic changes, like converting *Recency* to *Last_Purchase* (or the number of days since the customer’s last purchase) and *NumDealsPurchases* to *DiscountPurchases* by customer, as well as simple changes for consistency and efficiency. In many cases, we removed the preface, *Num*, from variable names like *NumStorePurchases* to maintain more concise labeling. For product preference variables, indicating the number of a given product each customer has purchased (including Fish, Meats, Fruits, Gold, Sweets, and Wines), we converted each of the long, bulky names, like *MntMeatProds*, to simply *Meats*, for example.



Filter
Income
Exclude: (650000 - 700000)

Filter
Year_Birth
Exclude: (1895 - 1900)

Filter
Year_Birth
Exclude: (1890 - 1895)

Calculated Field
Age
2024-[Year_Birth]

In the case of the variables *Z_CostContact* and *Z_Revenue*, these were removed altogether due to a lack of any variation across customers. *Dt_Customer* (now *JoinDate*) was also edited so that all the dates were in the same format of dd/mm/yyyy. In addition, the following section shows the final distribution of a number of our variables and other adjustments made prior to theoretical modeling and analysis, indicating if any additional extreme values need to be removed.

2.2. DATA PROCESSING – STAGE 2

Following the initial data cleaning performed in Tableau Prep, the dataset was subsequently imported into R to facilitate further data preprocessing tasks. During this phase, a strategic decision was made to exclude the 'Year_Birth' variable from the dataset. This choice was made due to the presence of the 'Age' variable, which serves a similar purpose. This decision was primarily aimed at optimizing runtime efficiency and conserving memory resources. A number of additional variable modifications were made.

2.2.1. TRANSFORMATIONS

1. *Age Data Type Transformation:* To enhance the robustness of the 'Age' variable, its data type was transformed from 'numeric' to 'integer'.

2. *Handling Missing Values:* The next step we took involved identifying rows within the dataset that contained missing values. Initially a mean imputation was used during the first cleaning step in Tableau Prep. However, upon visualizing the Income distribution, we realized the presence of outliers (from chart below). Consequently, utilizing mean imputation was deemed impractical due to the sensitivity of the mean to outliers, which can lead to a skewed result. Instead, a prudent choice was made to implement median imputation. This entailed filling the missing values in the 'Income' variable with the calculated median value (please refer to Appendix 1). Consequently, the median value of 51,371 was employed to replace the missing entries in the 'Income' variable. This transformation shifted the mean slightly from 51,951 to 51,953.

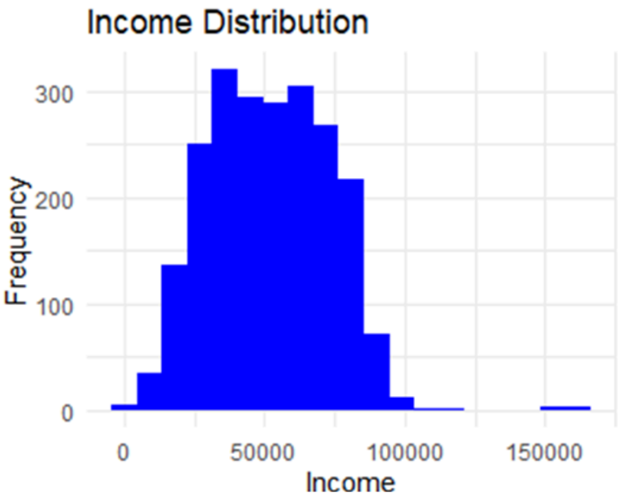


Figure 1 - Income Distribution

2.2.2. FEATURE ENGINEERING

1. Income Segmentation

Here, we segmented income levels into 'Low', 'Medium', 'High', and 'Very High' categories for MarketSphere's customers, based on income brackets with $30,000 increments to understand the financial diversity within the customer base.

2. Age Segmentation

Recognizing the varied age groups MarketSphere serves, we categorized customers into 'Young', 'Adult', 'Middle-Aged', and 'Senior', based on 20-year age intervals to understand the distribution and interactions within the different life stages and across other variables.

3. Recency of Purchase

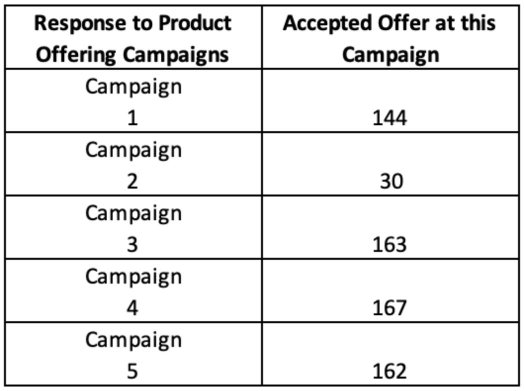
To analyze customer engagement levels, we classified the days since their last purchase into '0-30 days', '31-60 days', '61-90 days', and 'More than 90 days'. This we believe will provide insights into how recent interactions with MarketSphere influence future purchasing behaviors.

4. Total Purchases

The 'total_purchases' feature sums spending across all product categories for each customer, offering a holistic view of customer spending at MarketSphere. This data is key to identifying high-value customers and understanding spending habits.

5. Total Campaign Count

We summed up responses to the last five campaigns to create the 'total_campaign_count' feature. With this, we aim to understand the impact of MarketSphere's marketing efforts and help them strategize future campaigns for maximum engagement. Response rates for each campaign is outlined in the table below.



Response to Product Offering Campaigns	Accepted Offer at this Campaign
Campaign 1	144
Campaign 2	30
Campaign 3	163
Campaign 4	167
Campaign 5	162

2.2.3. VISUALIZING VARIABLES FOR OUTLIER DETECTION AND TREATMENT

For the numerical variables like income, we decided to plot each variable first in order to examine their distribution. Once outliers were detected, we then treated and visualized the variables again to see the effect.

2.2.3.1 Financial Feature

a. Income: Numerical

Income distribution was visualized using a box plot for outlier detection. From the plot below, it is evident that Income has outliers even after the median imputation done earlier. It's clear there are some customers who are real standouts in terms of their income, sitting far above the rest. These few, just 8 in number, appear to be extraordinary, not just in their earnings but in their academic achievements, with half of them holding PhDs (please see Table 1 below) and accounting for 2.6% of the combined income of all customers. In light of this discovery, we decided to investigate the relationship between income and education in the Exploratory Data Analytics Section

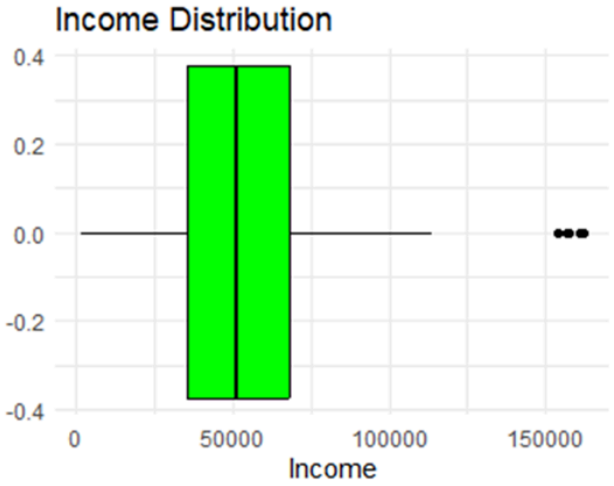
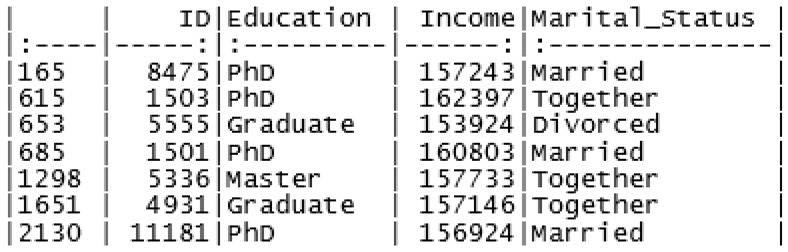


Figure 2 - Income distribution with outliers



```
|     |     ID|Education  | Income|Marital_Status |
|:----|-----:|:----------|------:|:--------------|
|165   |   8475|PhD        | 157243|Married        |
|615   |   1503|PhD        | 162397|Together       |
|653   |   5555|Graduate   | 153924|Divorced       |
|685   |   1501|PhD        | 160803|Married        |
|1298  |   5336|Master     | 157733|Together       |
|1651  |   4931|Graduate   | 157146|Together       |
|2130  |  11181|PhD        | 156924|Married        |
```

Table 1 – Extract of top income earners

In order to handle the outliers in Income, we calculated the quantiles for the distribution and substituted the values beyond the upper and lower bounds with the maximum and minimum values of the upper and lower bounds respectively. This ensured that all the values fitted within the quantiles without any outliers that would skew or bias our understanding of the typical customer. The image below shows the transformation after handling the outliers (image on the left) and the new income distribution (image on the right). This reflects where most of the customers stand financially.

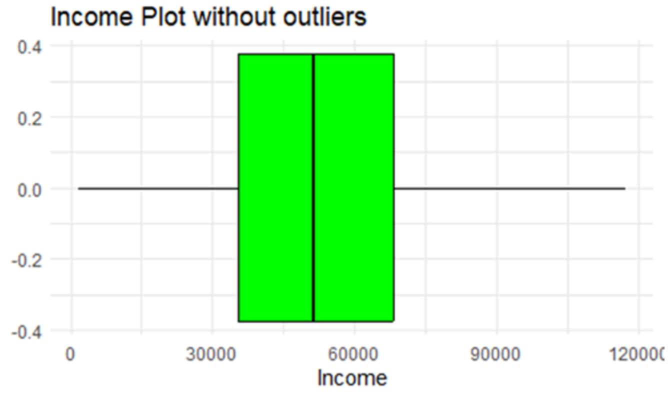


Figure 3 - Income boxplot with outliers removed.

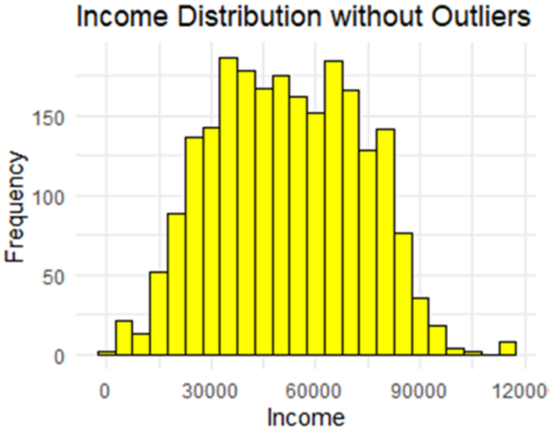


Figure 4 - New Income Distribution

2.2.3.2. Spending Pattern

a. Total Wines Bought: Numerical

The box plot below reveals the distribution and spending patterns of wines bought by customers of MarketSphere Inc over 2 years. Though most customers have their spending within a reasonable range, the plot also shows several points above the main distribution, indicating that there are customers

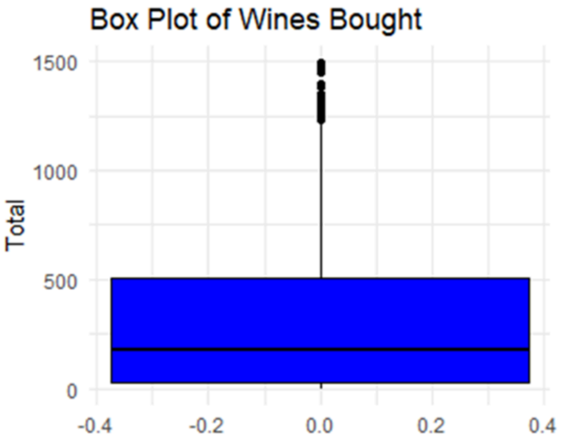


Figure 5 - Total Wines bought in 2 years with outliers.

who spend significantly more on wines than the average buyer.

Upon closer inspection we found that this group consists of 35 individuals, although a few could represent a segment with a particularly strong preference for wines. This behavior is worth noting, as it may influence how MarketSphere Inc. tailors its wine offerings and marketing strategies to appeal to such high-value customers and also entice the average buyers to buy more.

Before resolving outliers, we decided to visualize the distribution of the top 10 Wine products buyers. The "Top 10 Wine Buyers" scatter plot showcases the individual purchase volumes of wines by the top-spending customers, as identified by their unique Customer IDs. The high spending could be indicative of a dedicated segment that has a strong affinity for MarketSphere's wine selection, possibly reflecting a combination of high engagement and a taste for premium products.

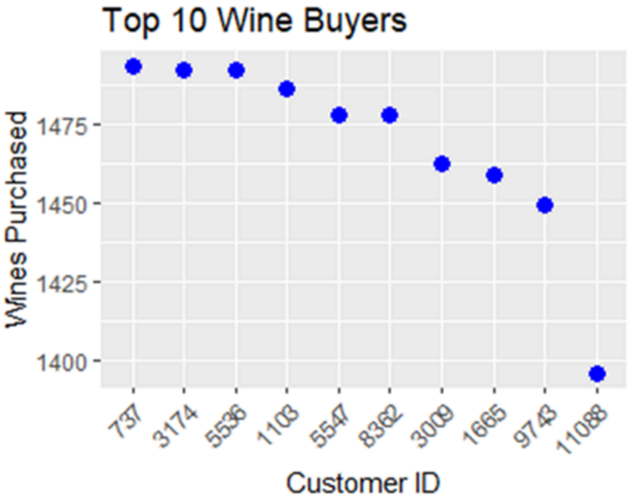


Figure 6 - Top 10 Wine Products buyers

When comparing the box plot of wines bought with outliers (the previous plot) to the new plot without outliers (figure7 on right), those extreme high-spending customers have been removed from the dataset. The removal of outliers has effectively compressed the upper tail of the distribution, resulting in a more compact and centered distribution of wine purchases among the majority of customers.

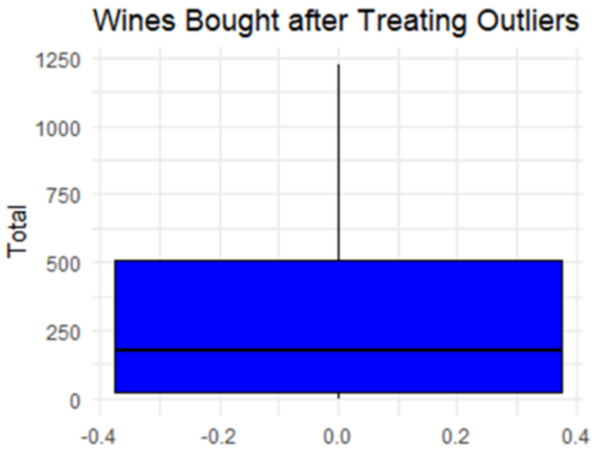


Figure 7 - Total Wines bought without outliers.

b. Fruits Bought: Numerical

The distribution of fruit purchases shows a pattern where most customers bought a smaller quantity of fruits, with fewer customers making larger purchases. The majority of data points cluster towards the lower end of the scale, indicating that a significant portion of customers prefers smaller fruit quantities. During our investigation, we identified 227 outliers, signifying a substantial number of fruit purchases that deviate from the typical buying pattern. These outliers could be attributed to various factors, such as customers buying fruits in bulk for business needs or for personal stockpiling.

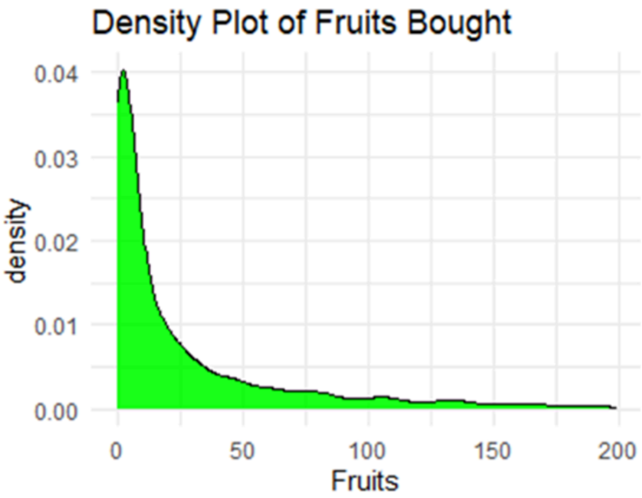


Figure 8 - Total Fruits bought in 2 years with outliers.

Compared to the original distribution which showed an 'L' shaped distribution (figure 8), the distribution after treating outliers depicts a 'U' shaped curve (figure 9). This is because the outliers in the original distribution were pulled towards the upper quartile.

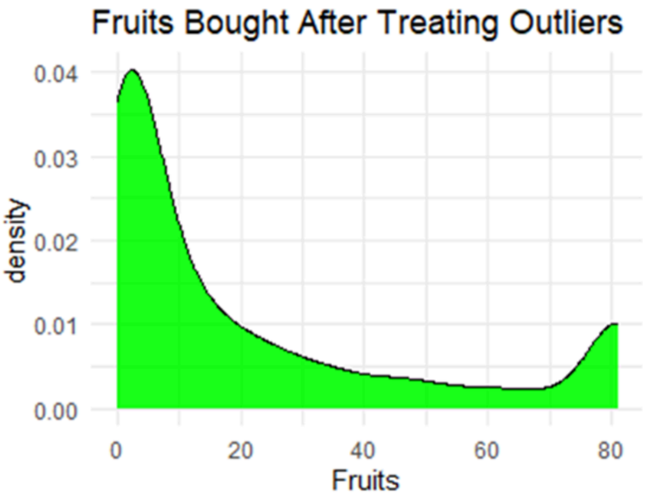


Figure 9 - Total Fruits Bought without Outliers

c. Fishes Bought: Numerical

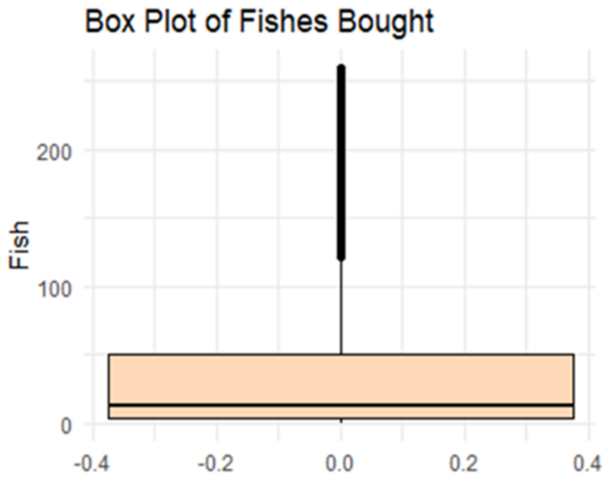


Figure 10 - Total Fishes Bought with outliers.

Similar to the distributions of fruits and wines, the distribution of fish purchases also revealed that a larger number of customers tend to buy smaller quantities of fish compared to those who make larger purchases. We identified 223 outliers in the fish purchase data. The presence of these outliers may be attributed to some customers buying fish in bulk, either for restaurant use, special events, or personal consumption.

The bar chart in figure 11 depicts a uniform spending pattern among the highest-spending customers, each purchasing just over 200 units of fish. This consistency suggests a similar demand or preference for fish among these top customers, possibly due to standard pricing or package deals from MarketSphere.

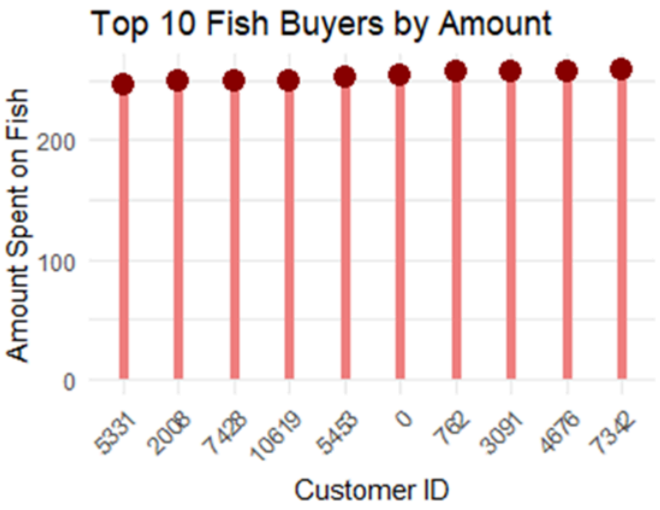


Figure 11 - Top 10 Fish Buyers

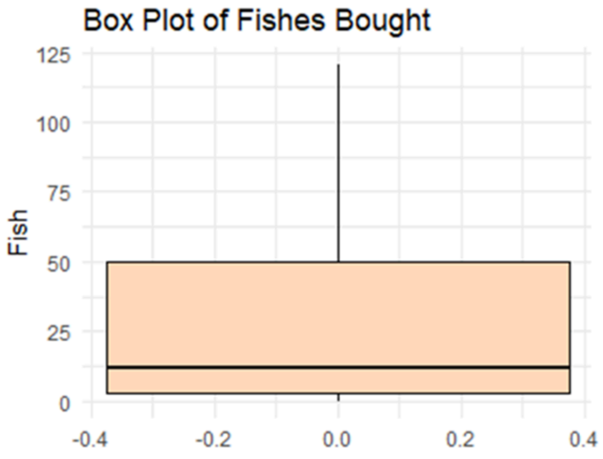


Figure 12 - Total Fishes bought without outliers.

The adjusted data on meat purchases shown by figure 12, following the treatment for outliers, presents a more standardized purchasing behavior across the customer base. The box plot shows the everyday consumption patterns after removing typical bulk or high buying events.

d. Meats Bought: Numerical

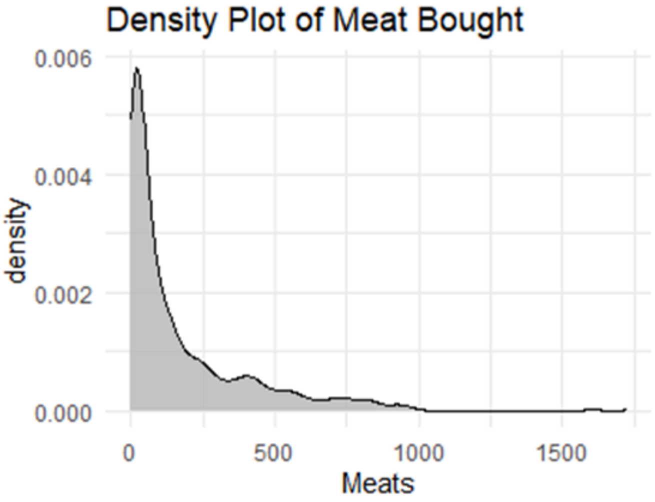


Figure 13 - Total Meat bought with outliers.

The distribution of meat purchases reveals that most customers buy moderate quantities of meat, while a smaller group makes notably larger purchases, as evident from the tail of the distribution. This purchasing pattern mirrors what we've observed in other product categories like fruits, wines, and fish, hinting at a consistent trend within the customer base. Within the meat product purchases, we identified 174 outliers. These outliers represent purchases that deviate from the typical range of the distribution. It could be due to special occasions or events, such as hosting a large gathering or stocking up for a holiday season barbecue.

2.2.3.3. Promotion Preference

a. Discount Purchase

The box plot for discount purchases (figure 14), even with 86 outliers present, suggests a normally distributed pattern of discount product buying among consumers. This indicates that most customers are likely to purchase similar quantities of discounted products, with a few engaging in significantly more or less purchasing behavior. This could imply that discount strategies are effectively reaching a broad customer base, promoting regular purchasing patterns. The outliers may represent a segment of customers who are either particularly discount-driven, possibly waiting for promotional periods to make bulk purchases, or those less responsive to discount offerings. Understanding the causes behind these

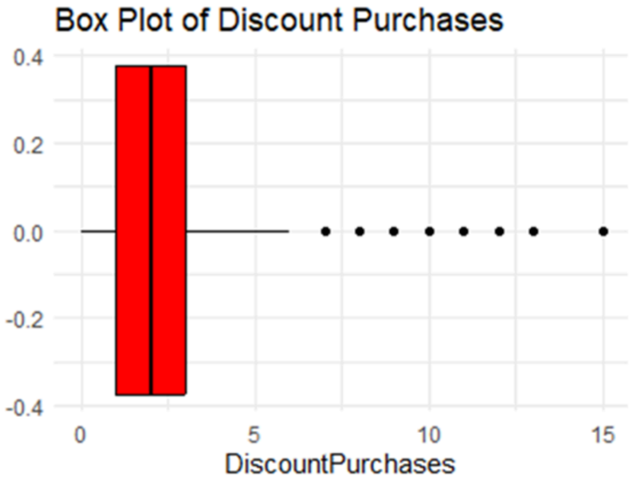


Figure 14 - Number of discount purchases

outlier purchases can help refine discount strategies, tailoring them to both the average consumer and the more extreme buyer segments. The distribution after treating outliers looked normally distributed.

2.2.3.4. Place of Purchase Preference

a. Web Purchases

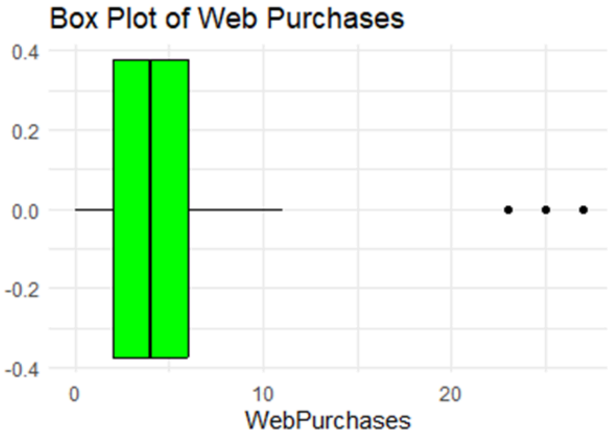


Figure 15 - Web Purchases

The "Box Plot of Web Purchases" (figure 15) shows a standard distribution for online transactions, with most customers making a uniform number of purchases. The few outliers indicate that some customers make significantly more purchases online, potentially due to targeted online promotions or bulk online buying habits. This information is valuable for tailoring online marketing strategies to encourage more frequent purchases across the customer base. The distribution after resolving outliers looked normalized.

b. Catalog Purchases

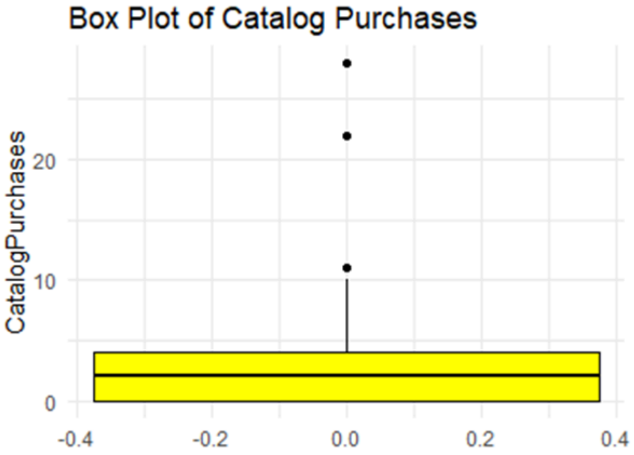


Figure 16 - Catalog Purchases

The Catalog Purchase distribution before outlier treatment shows several purchases that fall well outside the typical range, indicating occasional significantly higher spending through catalog orders.

After treating outliers, the distribution appears to be normal, suggesting that aside from these few cases, catalog purchase volumes are relatively consistent among the majority of customers.

c. Store PurchasesStore

Purchases distribution indicates a relatively broad distribution of in-store purchase volumes with no outliers, suggesting a varied purchasing behavior among customers at physical store locations. The absence of outliers implies that there are no extreme purchasing behaviors, which could suggest stable and predictable sales from the storefront for inventory management.

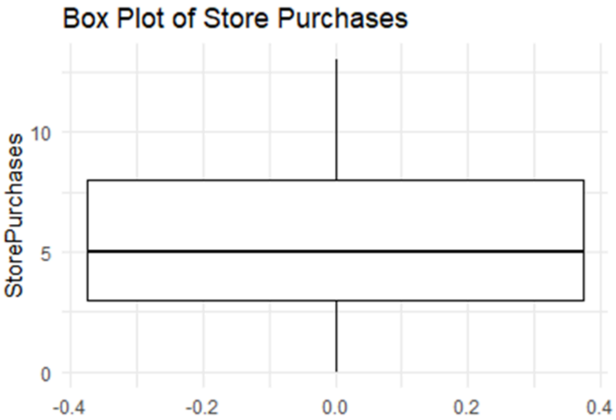


Figure 17 - Store purchases

Comment on Final Data Quality

We have ensured high data quality for our project which had 2241 unique records by meticulously refining them down to 2237, eliminating extreme outliers. We enhanced data completeness and accuracy through careful preprocessing, which included changing data types for consistency, renaming columns for clarity, and removing unnecessary columns. Missing values and outliers were addressed through imputation and Winsorization respectively, and feature engineering was employed to create relevant variables for analysis. These efforts resulted in a clean, reliable dataset, ready for effective exploratory analysis, clustering, and modeling

3.0 EXPLORATORY DATA ANALYSIS

In the process of preparing for clustering and regression analysis, we deemed it crucial to examine the interrelations and correlations among various variables

3.1 Demographic Features Analysis

For the demographic features within our dataset, which predominantly consist of categorical data, we have employed bar plots to visually represent the distribution of these categories. For the numerical demographic variables, we have utilized both histograms and box plots to effectively illustrate the data distribution. These visualizations are selected to facilitate a geared understanding of demographic trends and patterns, which will be pivotal in informing our subsequent modeling efforts.

3.1.1 Analysis of Education by Marital Status

Graduates have the highest count compared to all educational levels, indicating that the majority of customers have graduate degrees. Most of them are married as well. This suggests that targeting marketing campaigns towards married individuals with graduate-level education could be particularly effective.

Possible clusters: The largest cluster appears among married individuals with graduate degrees, signaling a significant market segment, likely suitable for premium product targeting. The second noticeable cluster is PhD. A third cluster can be seen among individuals with masters to basic education.

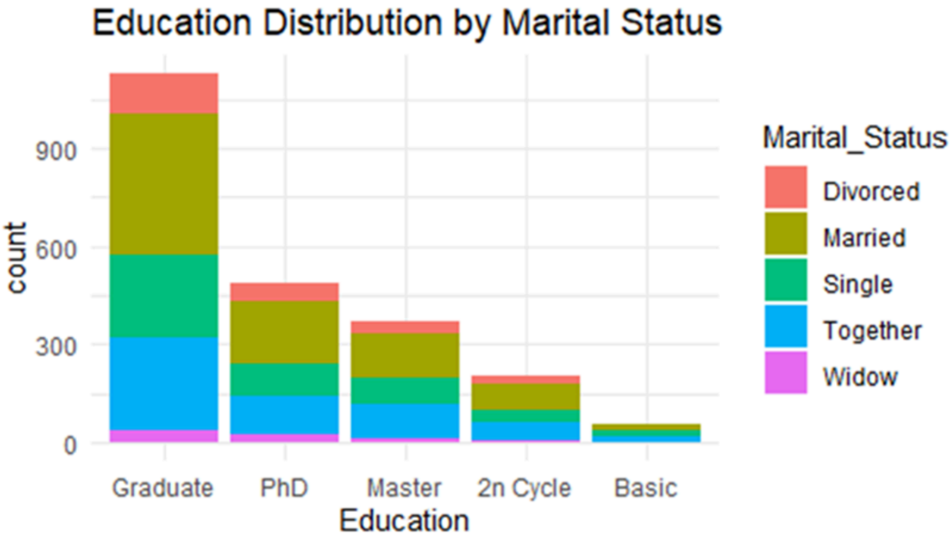


Figure 18 - Education vs. Marital Status

3.1.2 Analysis of Education by Marital Status

Figure 19 indicates married individuals as the predominant group across income levels, especially in the middle to high range. Single and Together statuses are more evenly spread across all income ranges. Widowed individuals are less common in the low and very high income interval.

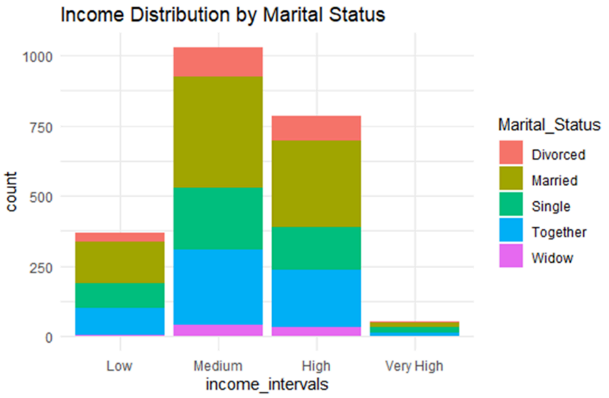


Figure 19 - Income vs. Marital status

Potential Clusters and Product Suggestions

1. Married High Earners: This cluster is ideal for MarketSphere’s premium offerings like fine wines and gold, which align with a higher disposable income.

2. Single and Together Middle Earners: MarketSphere could target this cluster with a mix of practical items like meats, fruits, and sweets, catering to the lifestyle of single professionals.

3. Diverse Middle Earners (Divorced and Widows): A wide range of selection including essential items like fruits and meats could appeal to this varied group.

For regression analysis, "Income" could be used as a target variable to predict spending patterns on these product categories, helping MarketSphere tailor its marketing strategies to these customer segments.

3.1.3 Analysis of Income by Age Group

The income distribution by age group suggests varying earning potentials across life stages. Young individuals, with incomes skewed towards the lower end, are potential targets for MarketSphere's more affordable products. Adults, with a wider range of incomes, might be interested in a broader product selection, including both staples and occasional luxuries. Seniors and Middle-aged

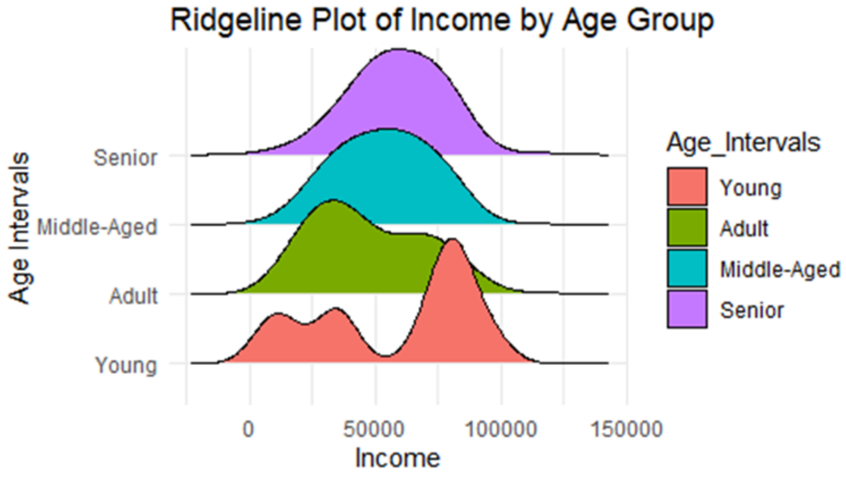


Figure 20 - Income vs. Age group

consumers, often at a higher income peak, represent a market for premium items. "Income" could serve as a key metric in a regression model to predict spending habits, informing MarketSphere's age-targeted marketing strategies.

3.1.4 Analysis of Income by Education

The "Ridgeline Plot Income by Education" indicates a clear upward trend in income with higher educational attainment. The data points observed at the right tailed ends of each educational level slopes more downward compared to Phd holders. This indicates that PhD holders earn higher income.

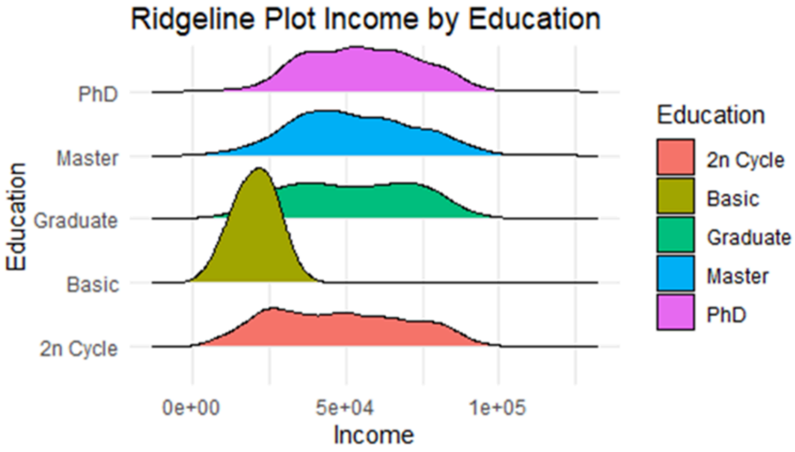


Figure 21 - Income vs. Education

We can say that there is a positive correlation between the levels of education and income. MarketSphere’s Graduates (1127) outweigh the number of Phd holders (486), hence though Phd holders earn higher income than Graduates, the proportion of income controlled by Graduates (58 million) is approximately 2 times that of Phd holders (27 million). The highest paid Phd holder earns 162,397 units (unit of measurement was missing in the dataset), while the highest paid Graduate earns 157,146 units. In order to test the hypothesis as to whether total income earned by these two groups are similar or not, we averaged the income of Graduates (52,168 units) and multiplied it by the number of PhD holders (486) to get a balanced assessment. The result showed that on average 486 graduates (the same number as Phd holders) earn 25 million compared to the 27 million earned by the same number of Phd holders (486). Those with 2n Cycle or Basic education levels tend to have lower incomes, suggesting they may prioritize essential and affordable MarketSphere products. In contrast, individuals with Graduate and Master’s degrees, indicate a potential interest in a diverse range of MarketSphere offerings, including both basic items and occasional luxury goods. Phd holders and Graduate holders, with the broadest and highest income distribution, represent a prime market for MarketSphere’s premium products such as gold and fine wines.

3.1.5 Analysis of Age Distribution

Age initially didn't show outliers so we decided to delve deep into age groups. To our surprise, there were some outliers in the senior group. We may have to resolve this outlier in the future should we decide to create interaction terms between different levels of age and a different variable, the age outlier in the senior level will bias our estimate. For the moment, this outlier concern can be set aside since age in the broader sense has no outliers, but it warrants attention before proceeding with more complex modeling involving age interactions.

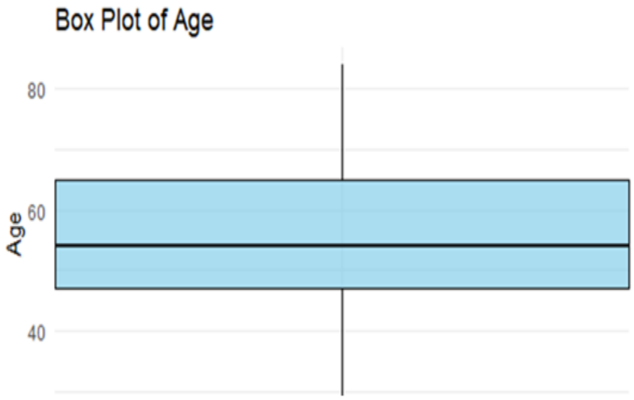


Figure 22 - Age Distribution without outliers

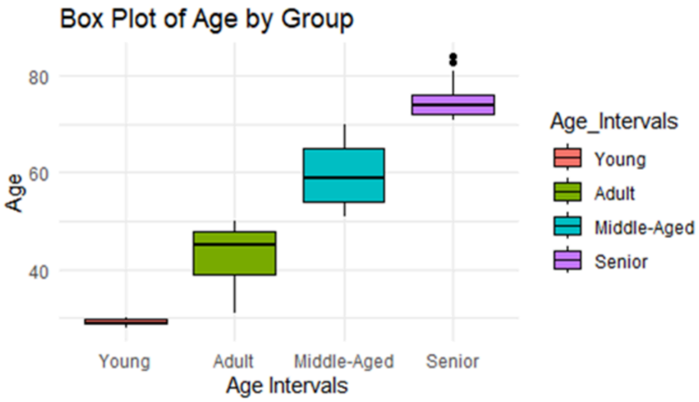


Figure 23 - Age by Age groups

3.1.6 Analysis of Marital Status by Groups

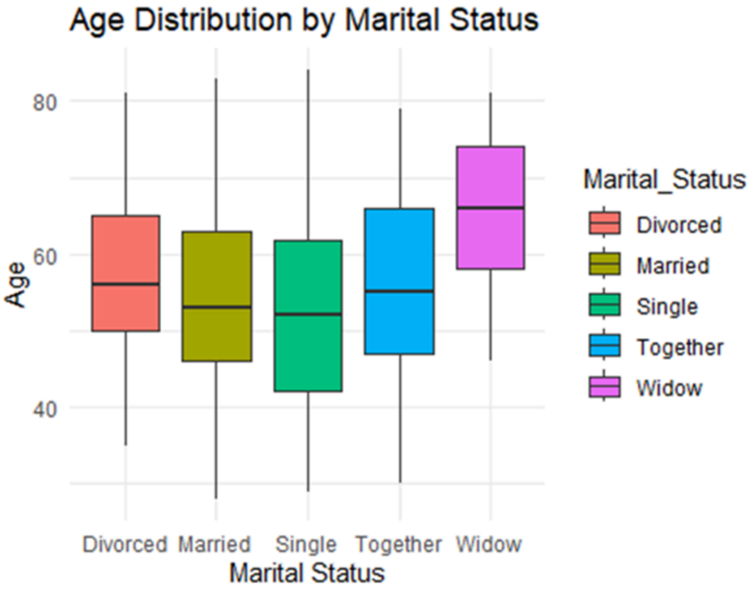


Figure 24 - Age vs. Marital status

The box plot presents the age distribution of individuals across different marital statuses: Divorced, Married, Single, Together, and Widow. The median age of divorced individuals appears slightly lower than that of married ones, with a slightly similar range of ages. Singles, Widows and Together show a uniform age distribution. Overall, the age range of MarketSphere's customers is between 28 to 84, suggesting a diverse age demographic. Only 2 people are aged 28 and both married without children and earn above 7500 units. On the other hand, it is interesting to note that the oldest customer is a single Phd holder who earns 51141 units. This age and marital status data can inform MarketSphere's approach to product targeting, as the varying age distributions likely correlate with different product preferences and spending habits. For instance, younger singles might lean towards trendier MarketSphere offerings, while older widowed customers may prefer more traditional goods. In a regression model, age could be predicted by marital status, aiding MarketSphere in targeting products to customer segments based on their life stage.

3.1.7 Analysis of Purchase Recency

The "Days since last purchase" box plot details the distribution of days elapsed since customers last made a purchase.

The "Cumulative Density of Purchase Recency" chart shows a straightforward cumulative frequency distribution of days since last purchase. The boxplot in figure 25 suggests that 50% of the customers (1118) have made a purchase within 90 days while 50% have not. Further, we observed from the cumulative distribution in figure 26 that as the number of days increases, there are more people who have not made a purchase over 90 days. It would have been insightful if the dataset had information concerning profitability. We would have assessed the impact on profitability given this discovery.

Together, these charts provide a snapshot of purchase frequency and recency. Strategies could include targeted promotions to reduce the number of days between purchases, especially for those customers approaching the higher end of the distribution.

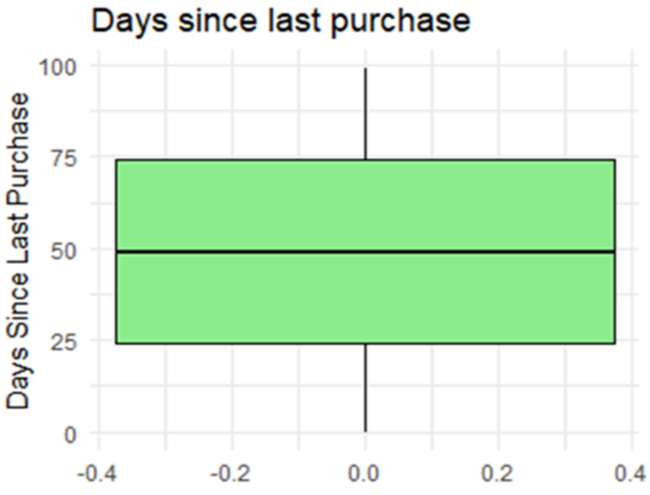


Figure 25 - Days since last purchase

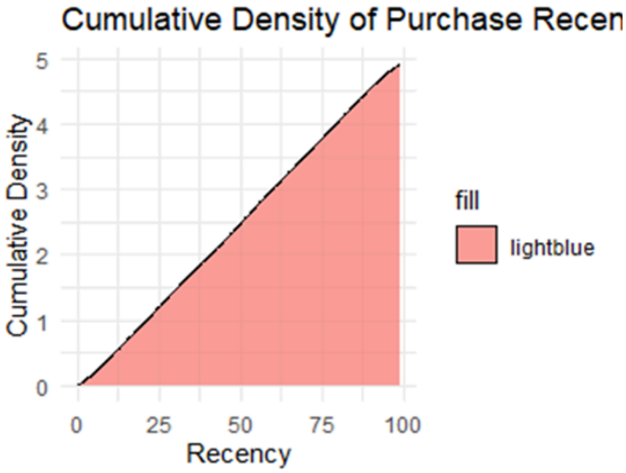
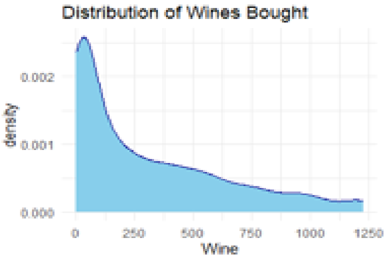
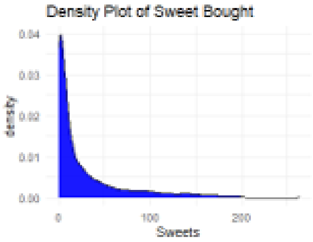
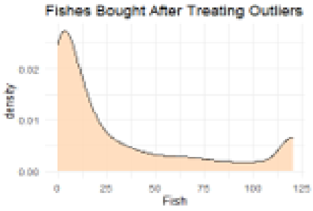
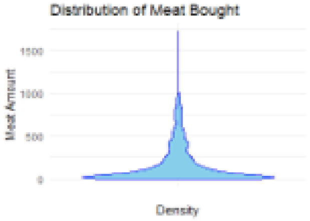


Figure 26 - CDF of days since last purchase









Overall, the product distributions for MarketSphere show that customers generally buy small quantities of wines, fish, and sweets, with purchases dropping off as quantity increases. Potential customer clusters might include those who buy in low, moderate, or high volumes. For regression models, the quantity bought could serve as the dependent variable, with demographics, purchase channels, and promotional responses as predictors, helping to tailor MarketSphere's marketing and inventory strategies.

Reflection

The Role of Tableau Prep & R in our Data Analytics Consultancy

To effectively analyze the Customer Personality Analysis data set, we utilized Tableau Prep Builder. An efficient and user-friendly software, Tableau Prep allowed us to connect our data source as either an Excel file or as a CSV (command-separated values) file. This feature was essential to ensuring that the few outliers we had as well as the handful of missing values in the household income column were appropriately handled. Data quality is another key component of successful teams and business and Tableau Prep Builder has many interface options to ensure that throughout the cleaning and manipulation process, all of our data is properly accounted for. Additionally, consistency is made possible with Tableau Prep, as grouping mechanisms allow us to ensure that data is clean to prevent misspellings and duplicates that can affect outcomes and conclusions. While Tableau Prep Builder is a familiar data preparation tool for nearly every member of our team, we were still pleased with the interface and the options available for wrangling and cleaning our data; this aided in our ability to better understand the behaviors of different consumers according to different variables.

After using Tableau Prep to manipulate our data appropriately, as well as rename fields more intuitively, we imported our dataset into R to further visualize the consumer personality information. R allowed us to properly visualize the different fields to bracket and segment based on income, age, and recency of purchase. The Income distribution also allowed us to properly impute the missing values, using the median values rather than the mean income. R was used to assess summary statistics and undertake feature engineering and visualization. Ultimately, visualizing the distribution of numerical values, including products, income and age further allowed better understanding of purchasing behavior and correctly accounted for outliers.

Tableau Prep allows team members to save and publish data preparation workflows. Tableau prep shares these workflows with others, making it easy to reuse and reproduce the same data preparation steps.

This sharing capability enhances consistency and reduces duplication of effort. Also Tableau Prep supports the integration of data quality checks into the workflow. Team members can define data validation rules and tests to ensure data quality throughout the project. This collaborative approach helps maintain data integrity.

While Tableau Prep Builder and R have many benefits for teams and business organizations, these softwares share a key drawback: data analysts/users are unable to collaborate within Tableau Prep Builder and R without additional purchases. Therefore, businesses must acquire an additional cost monthly to ensure analysts can clean, shape and visualize data together, which is a growing need today, as many businesses have adopted hybrid work environments.

General Reflections on Report 1

The analysis of various charts revealed a rich dataset with diverse variables, providing a robust foundation for understanding customer behavior and preferences. Notably, purchasing patterns across different products were similar, showcasing its strength in predicting customer behavior and enabling precise segmentation.

However, we observed complexity of some relationships in the data which can complicate model development. Fewer data points in certain categories also poses a challenge of overfitting and biased estimates in predictive models, potentially affecting the reliability of segment analysis. Furthermore, the monetary unit of measurement for income and amount of products bought was not specified in the dataset. Cost and revenue variables were static figures (3 and 11, respectively) for all customers, which limited the depth of financial analysis. Moreover, the project presented total purchases over the two years, but trends in purchases over time were not provided, which could have offered valuable insights. Nonetheless, the project's variables provide a strong basis for detailed customer analysis and predictive analytics. However, achieving success will depend on striking a balance between harnessing the depth of insight from the data and remaining cognizant of its limitations, alongside adapting to the dynamic market landscape.

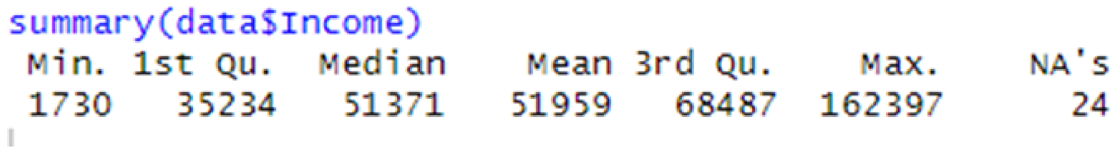
Citations

Boudet, J., Brodherson, M., Robinson, K., & Stein, E. (2023, June 26). *Beyond belt-tightening: How marketing can drive resiliency during uncertain times.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/beyond-belt-tightening-how-marketing-can-drive-resiliency-during-uncertain-times#/

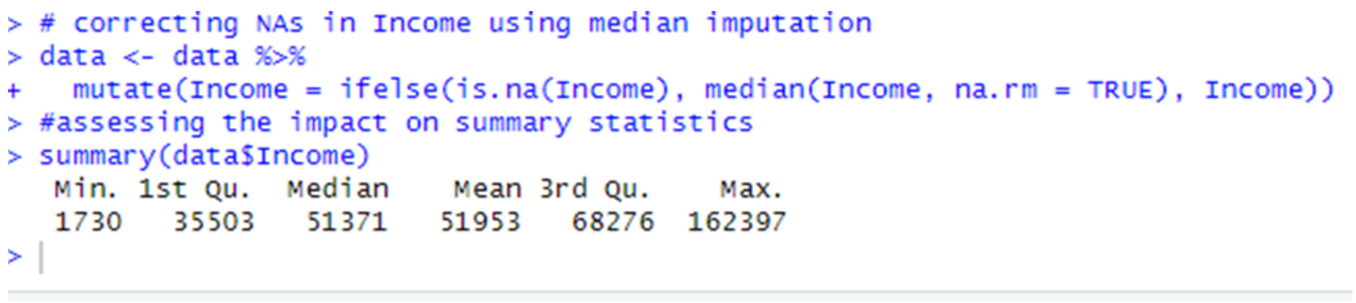
Checa, A., Heller, C., Stein, E., & Wilkie, J. (2023, April 5). *Modern marketing: Six capabilities for multidisciplinary teams.* McKinsey & Company. https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/modern-marketing-six-capabilities-for-multidisciplinary-teams

Customer Personality Analysis. (n.d.). Www.kaggle.com. https://www.kaggle.com/datasets/imakash3011/customer-personality-analysis/data

Appendix Back to Top

1. 

```
summary(data$Income)
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 1730   35234   51371   51959   68487  162397      24 
```

2. 

```
> # correcting NAs in Income using median imputation
> data <- data %>%
+   mutate(Income = ifelse(is.na(Income), median(Income, na.rm = TRUE), Income))
> #assessing the impact on summary statistics
> summary(data$Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1730   35503   51371   51953   68276  162397 
> 
```