**Telecommunications Customer Segmentation: Data Due Diligence and Exploratory Data Analysis**

**Adeline Casali**

**Data Preprocessing**

The data was processed to ensure it was clean, consistent, and ready for in-depth analysis and feature engineering. By addressing issues such as missing values, inconsistencies, and data type conversions, the dataset’s suitability for exploration and further analysis was enhanced. The steps taken to process the data are described below.

1. *Removing Empty Column*:

The column “x” was removed as it was empty and didn’t contribute to the analysis.

1. *Separating into Qualitative and Quantitative Data*:

The dataset was divided into two subsets based on whether the variables represent categorical attributes or numerical measurements. This segmentation allowed for targeted preprocessing and analysis.

1. *Checking the Structure of and Converting Data Types in Qualitative Data*:

The structure of the qualitative data subset was checked to ensure accuracy. The ‘TownSize’ variable was converted from a character to an integer type, as it contains integers.

1. *Checking the Structure of and Cleaning Quantitative Data*:

The structure of the quantitative data subset was also checked to ensure accuracy. Monetary variables (‘HHIncome’, ‘CarValue’, ‘CardSpendMonth’, ‘VoiceLastMonth’, ‘VoiceOverTenure’, ‘EquipmentLastMonth’, ‘EquipmentOverTenure’, ‘DataLastMonth’, and ‘DataOverTenure’) were cleaned to remove the special characters “$”, “,”, and “-” and were converted to numeric format for meaningful analysis. Values originally depicted as “$ - ” were replaced with 0 to maintain data integrity. Additionally, ‘CommuteTime’ was converted to an integer type to facilitate numerical operations.

1. *Searching for Unique Values and Cleaning Qualitative Data*:

The unique values of each qualitative variable were reviewed to ensure validity. Both ‘Region’ and ‘TownSize’ were converted to include descriptions in place of numerical values for better utilization in further analysis (i.e. ‘Very Large’ in place of ‘1’ for ‘TownSize’ and ‘Northeast’ in place of ‘1’ for ‘Region’, etc.). Missing data in both ‘CarOwnership’ and ‘CarBrand’, originally depicted as ‘-1’, were replaced with NA for consistency and to signify missing data. Additionally, the missing values in ‘Internet,’ originally depicted as ‘2’, ‘3’, and ‘4’, were all replaced with NA.

1. *Checking for Outliers and Ensuring Data Validity*:

Summary statistics were computed to identify potential outliers or inconsistencies in numerical variables, ensuring data validity and reliability.

1. *Merging Cleaned Data*:

The cleaned qualitative and quantitative datasets were merged based on the common identifier ‘CustomerID’, consolidating the cleaned data for further analysis.

1. *Handling Missing Values in Cleaned Data*:

The dataset was searched for NA values by column. In columns where the total number of NA values equaled less than 5% of the dataset, rows with NA values were removed. Otherwise, NA values were kept to avoid removing excess data.

1. *Confirming the Cleanliness and Validity of the Cleaned Dataset*:

Finally, the structure, summary values, unique values, and NA values were checked in the cleaned dataset.

**Feature Engineering**

Below is a detailed description of each feature engineering step taken to create 10 new columns and their justifications. Each feature engineering step adds valuable insights into customers’ demographics, financial behaviors, and consumption patterns.

1. *Location*:

The ‘Location’ column categorizes customers into rural or urban areas based on the ‘TownSize’ column. Towns categorized as “Small” or “Very Small” were categorized as “Rural”, while the remaining options, “Medium”, “Large”, and “Very Large”, were categorized as “Urban”. Understanding the geographic distribution of customers helps in tailoring marketing strategies and services to specific regions, as well as resource allocation based on urban or rural demographics.

1. *Generation*:

The ‘Generation’ column categorizes customers into different generational cohorts based on their age. Customers ages 12-27 were categorized as “Gen Z”, ages 28-43 as “Millennials”, ages 44-59 as “Gen X”, and ages 60-79 as “Boomer” (*Beresford Research*, 2024). Generational cohorts often have distinct preferences, behaviors, and communication styles. By segmenting customers into generations, marketing efforts can be customized to resonate with each group’s unique characteristics.

1. *Education Level*:

The ‘EducationLevel’ column categorizes customers based on their years of education. Customers with fewer than 12 years of education are categorized as “Less than high school”, 12-13 years as “High School”, 14-15 years as “Some college or Associate’s”, 16-17 years as “Bachelor’s degree”, and 18 or greater years as “Graduate degree”. These are estimated categories, as years of education for degree requirements vary. However, education level influences consumer behavior, preferences, and decision-making processes. Segmenting customers by education level allows for targeted marketing to each segment.

1. *Total Debt*:

The ‘AllDebt’ column calculates customers' total debt by summing the ‘CreditDebt’ and ‘OtherDebt’ columns. Total debt is a critical financial indicator that impacts customers’ purchasing power and financial stability. By aggregating all types of debt, we can gain insights into customers’ overall financial health.

1. *Pets per Person*:

The column ‘PetsPerPerson’ was calculated as the average number of pets per person in each household. Pet ownership reflects lifestyle preferences and household dynamics, influencing purchasing decisions and product preferences.

1. *Total Over Tenure*:

The ‘TotalOverTenure’ column calculates the total usage of data, equipment, and voice services over tenure. Tracking total usage over customer tenure consolidates information and provides insights into service utilization patterns.

1. *Average Card Spending per Item*:

The ‘AvgCardSpendPerItem’ column was created to calculate the average amount of money spent per item on credit cards. The average amount spent per item provides insights into customers’ spending habits and purchasing preferences. It helps to identify high-value customers who make frequent and substantial transactions, enabling targeted marketing efforts.

1. *Age to Employment Ratio*:

The ‘AgeToEmploymentRatio’ column computes the ratio of age to employment length for each customer. This ratio reflects the stage of a customer’s career relative to their age and gives a standardized measure to indicate factors such as career stability and financial independence.

1. *Total Spent Last Month*:

The ‘TotalSpentLastMonth’ column calculates the total amount spent on data, equipment, and voice services in the last month. Total spending in the last month provides insights into short-term consumption patterns and immediate purchase behavior. It helps consolidate the spending data into one column and works to identify customers with high spending potential.

1. *Card Spending to Income Ratio*:

The ‘CardSpendingToIncomeRatio’ column computes the proportion of income spent on credit card transactions. Card spending relative to income indicates customers’ financial management practices and credit utilization behavior. It helps to identify customers with high debt burdens, guiding risk assessment.

**Categorization**

Below are the categories and a brief description of each variable. These categories provide a framework for organizing and understanding the variables in the dataset, facilitating analysis and interpretation.

1. *Demographic*

* Region: Geographic region where the customer resides.
* TownSize: Size of the town where the customer resides.
* Gender: Gender of customer.
* MaritalStatus: Martital status of the customer.
* Age: Age of the customer.
* EducationYears: Number of years of education completed by the customer.
* EducationLevel: Level of education, based on education years.
* EmploymentLength: Number of years the customer has been employed.
* AgeToEmploymentRatio: Ratio of age to number of years the customer has been employed.
* HouseholdSize: Number of people living in the customer’s household.
* JobCategory: Category of the customer’s job.
* HomeOwner: Whether the customer owns their home.
* Location: Whether the customer lives in an urban or rural location.
* Generation: Generation category of the customer.

1. *Financial*

* DebtToIncomeRatio: Ratio of the customer’s debt to their income.
* CreditDebt: Amount of credit debt a customer has.
* OtherDebt: Amount of other debt not related to credit accounts.
* AllDebt: Total amount of debt a customer has.
* CarsOwned: Number of cars owned by the customer.
* CarOwnership: Whether the customer owns or leases their car.
* CarValue: Value of the customer’s car(s).
* HHIncome: Household income of the customer.
* CardSpendingToIncomeRatio: Ratio of the customer’s card spending to their income.
* AvgCardSpendPerItem: Average amount spent per item on a credit card.
* CardItemsMonthly: Number of items charged to the primary credit card is monthly.
* CardSpendMonthly: Total primary credit card spending in a month.
* LoanDefault: Whether the customer has defaulted on a loan.
* CreditCard: The primary credit card type used by the customer.
* CardTenure: Length of time the customer has had their primary credit card.

1. *Behavioral*

* CommuteTime: Time spent commuting to work.
* UnionMember: Whether the customer is a member of a labor union.
* Retired: Whether the customer is retired.
* ActiveLifestyle: Whether the customer leads an active lifestyle.
* Pager: Whether the customer uses a pager.
* NewsSubscriber: Whether the customer subscribes to news services.
* OwnsPC: Whether the customer owns a personal computer.
* OwnsMobileDevice: Whether the customer owns a mobile device.
* OwnsGameSystem: Whether the customer owns a gaming system.
* OwnsFax: Whether the customer owns a fax machine.
* NumberPets: Total number of pets owned by the customer.
* NumberDogs: Number of dogs owned by the customer.
* NumberCats: Number of cats owned by the customer.
* NumberBirds: Number of birds owned by the customer.
* PetsPerPerson: Number of pets per person in the customer’s household.
* TVWatchingHours: Number of hours spent watching TV.
* PoliticalPartyMem: Whether the customer is a member of a political party.
* CarBrand: Origin of the brand of car the customer owns.
* Votes: Whether or not the customer votes.

1. *Telecom usage*

* CustomerID: Unique identifier for each customer
* WirelessData: Whether the customer uses wireless data services.
* Multiline: Whether the customer has multiple phone lines.
* Internet: Whether the customer has internet services.
* PhoneCoTenure: Tenure with the phone company, in months.
* VoiceLastMonth: Amount spent on voice usage in the last month.
* VoiceOverTenure: Amount spent on voice usage over tenure.
* EquipmentLastMonth: Amount spent on equipment in the last month.
* EquipmentOverTenure: Amount spent on equipment over tenure.
* DataLastMonth: Amount spent on data in the last month.
* DataOverTenure: Amount spent on data over tenure.
* TotalSpentLastMonth: Total spent on data, equipment, and voice in the last month.
* TotalOverTenure: Total spent on data, voice, and equipment over tenure.
* EquipmentRental: Whether the customer rents any equipment.
* CallingCard: Whether the customer uses a calling card.
* CallerID: Whether the customer subscribes to Caller ID services.
* CallWait: Whether the customer has Call Waiting.
* CallForwarding: Whether the customer uses Call Forwarding services.
* ThreeWayCalling: Whether the customer has Three Way Calling capability.
* EBilling: Whether the customer uses electronic billing.
* VM: Whether the customer uses Voicemail services.

**Variables of Interest**

These selected variables encompass demographic, socioeconomic, and geographic dimensions, offering a comprehensive overview of customers’ characteristics, behaviors, and preferences.

* HHIncome: Household income provides insights into customers’ purchasing power and socioeconomic status, which directly influence their spending behavior and affordability of services.
* Gender: Gender is a demographic variable that can significantly impact consumer behavior, preferences, and purchasing decisions.
* EducationLevel: Education level indicates customers’ educational attainment, which correlates with income levels and lifestyle preferences (Wolla and Sullivan, 2017). Marketing tailored by education level can enhance relevance and resonate with target audiences.
* Region and Location: Both region and location delineate customers based on geographical location, allowing for targeting service offerings and marketing tailored to regional preferences.
* MaritalStatus: Marital status influences household dynamics and purchasing decisions (Srinivasan et al., 2015). Tailoring marketing messages and product offerings to marital status segments can enhance relevance and engagement.
* JobCategory and UnionMember: The customer’s job category and whether they are a union member provides valuable insights into their lifestyle, and understanding this distribution and the behavioral patterns associated with each segment can enable targeted marketing efforts.
* DebtToIncomeRatio: Debt to income ratio reflects customers’ financial health and debt management capabilities, influencing their ability to spend and engage with services. Monitoring this ratio can also aid in risk assessment.
* LoanDefault: Whether a customer has defaulted on a loan highlights their credit risk and financial behavior. It is a critical variable in informing risk management strategies.
* Generation: Different generations have unique characteristics, preferences, and consumption patterns. Understanding generational differences allows for tailored marketing strategies and product offerings.
* AllDebt: Total debt provides a holistic view of customers’ financial obligations. Monitoring this level can inform risk assessment.
* AvgCardSpendingtoIncomeRatio: This ratio indicates customers’ reliance on credit and can inform risk assessment decision-making.
* AvgCardSpendPerItem: This calculation offers insights into customers’ purchasing habits and transactional behavior. Understanding this metric can guide promotional offerings and other marketing services.
* TotalSpentLastMonth: The amount spent on telecommunications services in the past month reflects the customer’s overall consumption patterns and financial activity. Monitoring total spending levels can inform revenue forecasting and customer engagement strategies.
* WirelessData: Wireless data usage is a key indicator of service utilization and can inform the development of marketing and pricing strategies.
* PhoneCoTenure: Phone company tenure indicates the length of time the customer has been with the telecommunications company, reflecting loyalty and long-term relationships. Understanding tenure can inform customer retention strategies and predict churn likelihood.
* TotalOverTenure: The total amount spent over tenure with the phone company provides insights into customer loyalty and engagement. The company can identify high-value customers and their characteristics.

**Graphical Analysis**

*Single Variable Visualizations*

A map of the united states

Description automatically generated

Figure 1. Customers by Region: This graph visualizes the distribution of customers across the United States by region. The customers are relatively equally distributed, with the Southeast having the most and the Southwest having the fewest.

A graph of a number of orange colored bars

Description automatically generated

Figure 2. Distribution of Phone Company Tenure: There is an even distribution of customers across phone tenure, with the highest concentration at tenures of 60 months and higher, indicating good customer retention.

A blue and pink circle with white text

Description automatically generated

Figure 3. Distribution of Gender: This dataset has a very equal gender ratio, which makes for excellent comparisons between groups without bias.

A graph of a number of people

Description automatically generated with medium confidence

Figure 4. Distribution of Generation: Most customers are older, with the largest segment being Boomers and the smallest being Gen Z. This indicates an older customer base.

A graph showing the amount of income

Description automatically generated

Figure 5. Household Income Distribution: The vast majority of households have incomes around $20,000, with the few larger outliers causing a strong right skew.

A graph of a number of purple bars

Description automatically generated with medium confidence

Figure 6. Debt to Income Ratio Distribution: The debt-to-income ratio is normally distributed with a slight right skew due to a few customers with a larger ratio. This indicates a financially solid customer base.

A diagram of a pie chart

Description automatically generated

Figure 7. Distribution of Education Level: Most customers have at least a high school degree, indicating a well-educated customer base.

*Multivariable Visualizations*

A graph of a number of people

Description automatically generated

Figure 1. Job Category and Union Member Distribution: Most customers are employed in white-collar professions, with few union members.

A screenshot of a graph

Description automatically generated

Figure 2. Phone Company Tenure by Loan Default Status: Customers who have not defaulted on a loan have a much higher tenure. This indicates that customers with a clean credit history have a higher retention rate.

A chart of different shades of red

Description automatically generated

Figure 3. Phone Company Tenure by Location and Region: Urban customers have a higher retention rate than rural customers, with urban Western and urban Southeastern customers having the longest tenure.

A graph of a number of people

Description automatically generated

Figure 4. Phone Company Tenure by Generation: Phone company tenure increases as generations age, which is intuitive as older customers would have had more time to purchase services.

A graph showing a green line

Description automatically generated

Figure 5. Total Spent Last Month vs. Total Over Tenure: Intuitively, there is a strong positive correlation between customers' spending in the previous month and their total spending over their tenure with the phone company.   
A graph of a credit card

Description automatically generated

Figure 6. Avg. Credit Card to Income Ratio by Gender and Marital Status: Male customers tend to have a higher ratio of card spending to income. Married females have a higher spending ratio, while the trend is reversed in males. This is informative financial and demographic information for future marketing and decision-making.

A graph showing a number of blue dots

Description automatically generated

Figure 7. Total Debt vs. Total Amount Spent Over Tenure: A slight positive correlation exists between customers' total debt and the amount they have spent over their tenure with the phone company. This indicates that customers with more debt may be customers with higher value.

A graph of a credit card

Description automatically generated

Figure 8. Avg. Amount Spent Per Item on Credit Card by Wireless Data Use: Customers who use wireless data services have a higher average card spending per item. These customers may be good marketing targets for wireless data services.

*References*:

*Age range by generation*. Beresford Research. (2024, February 29). https://www.beresfordresearch.com/age-range-by-generation/

Srinivasan, R., Srivastava, R. K., & Bhanot, S. (2015). Impact of education on purchase behaviour of luxury brands. *IOSR Journal of Business and Management*, *16*(11), 63–74. https://doi.org/10.9790/487x-161116374

Wolla, S. A., & Sullivan, J. (2017, January). *Education, income, and wealth*. Economic Research - Federal Reserve Bank of St. Louis. https://research.stlouisfed.org/publications/page1-econ/2017/01/03/education-income-and-wealth