**ISOM 837**

**Storytelling Report**

**Dataset – Customer Segmentation**

**Dharmin Sanghvi**

**Business Understanding:**

Understanding a dataset containing customer information, demographic data, behavioral patterns, and insurance-related metrics offers businesses a wealth of opportunities to optimize their operations and enhance customer experiences. By segmenting customers based on demographic factors such as age, gender, marital status, and education level, organizations can tailor their marketing strategies and product offerings to better meet the diverse needs of different customer groups. Furthermore, analysing behavioural data, purchase history, and interactions with customer service enables businesses to gain insights into customer preferences, buying patterns, and satisfaction levels, empowering them to optimize their products, services, and customer support processes accordingly.

In the insurance industry specifically, analysing insurance-related metrics such as insurance products owned, coverage amounts, premium amounts, and policy types can provide valuable insights into customer insurance needs and preferences. By understanding these factors, insurance companies can develop personalized insurance solutions, set competitive premiums, and target the right customers with relevant offerings. Additionally, leveraging customer data for risk assessment and underwriting allows insurers to accurately assess the likelihood of claims, determine appropriate coverage levels, and mitigate potential risks, ultimately improving underwriting accuracy and profitability.

Moreover, understanding customer data enables businesses to implement effective customer retention and acquisition strategies. By identifying factors that contribute to customer churn or loyalty, organizations can develop targeted retention initiatives and acquisition campaigns to improve customer satisfaction, loyalty, and lifetime value. Furthermore, ensuring compliance with regulatory requirements and industry standards is essential for businesses operating in the insurance sector. By monitoring and ensuring compliance with data privacy laws, anti-discrimination regulations, and other relevant regulations, organizations can build trust with customers and mitigate legal and reputational risks. Overall, leveraging customer data effectively enables businesses to make informed decisions, drive operational efficiency, and deliver personalized experiences that meet the evolving needs of their customers.

**Data Understanding :**

The Customer Segmentation dataset has the following :

- 53503 Rows of Data

- 20 Columns

**Customer ID:** A unique identifier for each customer.

**1)Age:** The age of the customer.

**2)Gender:** The gender of the customer (e.g., Male, Female).

**3) Marital Status:** The marital status of the customer (e.g., Married, Single, Divorced).

**4)Education Level:** The highest level of education attained by the customer (e.g., Associate Degree, Bachelor's Degree).

**5)Geographic Information:** The geographic location of the customer (e.g., states or regions).

**6)Occupation:** The occupation of the customer (e.g., Entrepreneur, Manager).

**7)Income Level:** The annual income level of the customer.

**8)Behavioral Data:** Information related to the customer's behavior, indicated by policies (e.g., policy5).

**9)Purchase History:** The date of a purchase made by the customer.

**10)Interactions with Customer Service:** The mode of interaction with customer service (e.g., Phone, Chat, Email).

**11)Insurance Products Owned:** The type of insurance policy owned by the customer (e.g., policy1, policy2).

**12)Coverage Amount:** The amount of coverage provided by the insurance.

**13)Premium Amount:** The premium amount paid for the insurance.

**14)Policy Type:** The type of insurance policy (e.g., Group, Family).

**15)Customer Preferences:** Customer's preferences (e.g., Email, Mail, Text).

**16)Preferred Communication Channel:** The preferred channel for communication with the customer (e.g., In-Person Meeting, Mail).

**17)Preferred Contact Time:** The preferred time for contacting the customer (e.g., Morning, Afternoon, Evening).

**18)Preferred Language:** The language preferred by the customer (e.g., English, French).

**19)Segmentation Group:** The segmentation group assigned to the customer (e.g., Segment2, Segment3).

**Research Questions :**

Question: How do demographic factors (such as age, gender, marital status, and education level) influence the segmentation group of a customer?

Question: What is the relationship between a customer's interactions with customer service (mode and frequency) and their segmentation group?

Question: How do policy type preferences (e.g., group vs. family policies) and preferred communication channels (e.g., email, chat, in-person meeting) correlate with customer segmentation groups?

A screenshot of a computer

Description automatically generated

**Class Variable Summary Statistics**

Behavioral\_Data: This variable has 5 categories and no missing values. The most common category is 'policy3', occurring 29.46% of the time. The second mode is 'policy1' at 20.09%.

Customer\_Preferences: Comprised of 5 levels with no missing data. 'Mail' is the most common preference at 22.22%, and 'Email' follows closely at 21.15%.

Education\_Level: This variable has 5 different educational levels with no missing entries. The 'Associate Degree' is most frequent with a percentage of 22.83, and 'Doctorate' is next with 22.62%.

Gender: There are 2 categories and no missing values. 'Male' is the mode with 51.40%, and 'Female' is close at 48.60%.

Geographic\_Information: It consists of 35 categories with no missing entries. 'Lakshadweep' has a very low percentage of 0.67, while 'Himachal Pradesh' has 3.67%.

Insurance\_Products\_Owned: With 5 levels and no missing values, 'policy1' is the most common product at 24.98%. 'policy2' is the second most common at 22.36%.

Interactions\_with\_Customer\_Service: Also, with 5 levels without missing data, the 'Chat' mode is at 26.45%, followed by 'Mobile App' at 23.21%.

Marital\_Status: There are 5 categories, no missing values, 'Married' is the most common status at 24.71%, with 'Divorced' not far behind at 24.58%.

Occupation: Comprised of 5 categories with no missing data, 'Salesperson' is the most frequently occurring at 14.80%, and 'Entrepreneur' follows with 12.40%.

Policy\_Type: It has 4 categories with no missing entries. The 'Group' category is most common, with a 34.12% occurrence, and 'Business' has a 26.14% share.

Preferred\_Communication\_Channel: This variable shows 4 levels and no missing data. 'In-Person Meeting' is the most preferred at 26.75%, and 'Phone' follows at 25.19%.

Preferred\_Contact\_Time: With 5 categories and no missing data, 'Weekends' is the preferred time at 26.34%, and 'Morning' is next with 25.73%.

Preferred\_Language: Consisting of 5 levels with no missing values, 'German' is the mode with 27.01%, and 'French' follows with 24.97%.

Segmentation\_Group: As a target variable with 5 categories and no missing data, 'Segment5' is most common at 26.12%, and 'Segment2' is the second mode with 21.39%.

A screenshot of a computer

Description automatically generated

**Interval Class Variable Summary Statistics**

Age: The average age in the dataset is approximately 44.14 years, with a standard deviation of about 15.07, indicating moderate age variation among individuals. The age distribution is slightly right-skewed (skewness of 0.116304), suggesting a longer tail towards the older ages. The kurtosis is -1.1771, indicating a flatter distribution with fewer outliers than a normal distribution (platykurtic).

Coverage Amount: On average, the coverage amount is 492580.8 with a high standard deviation of 268405.5, showing significant variability in the coverage amounts. The distribution has a slight right skew (skewness of 0.165374), and the kurtosis is -1.2161, which implies a flatter peak compared to a normal distribution.

Income Level: The mean income level is 82768.32 with a large standard deviation of 36651.08, pointing to substantial differences in income levels among individuals. The distribution has a slight right skew (skewness of 0.109402), and the kurtosis is -1.42459, indicating a flatter distribution.

Premium Amount: The average premium amount is 3023.702 with a standard deviation of 1285.834, reflecting moderate variability in premium amounts paid. The skewness of -0.274318 indicates a slight left skew, and the kurtosis of -1.20639 indicates a flatter distribution.

Purchase History: The dataset shows an average purchase history of 22281.62 with a standard deviation of 631.4466, suggesting that most values cluster around the mean. The skewness is -0.02053, which is very slight and indicates a distribution that is almost symmetrical. The kurtosis of -1.18969 suggests that the distribution is flatter than a normal distribution.

**EDA: Variable Worth**

A screenshot of a graph

Description automatically generated

The bar for "Geographic Information" is the tallest, suggesting it has the highest worth in this model, while the bar for "Gender" is the shortest, suggesting it has the lowest worth. This visualization helps in understanding which features might be most predictive for the customer ‘Segmentation’ target variable.

**Chi Square Plot**

A screenshot of a computer

Description automatically generated

Geographic Information: This variable has the highest chi-square value, indicating it is likely the most significant predictor among those listed in distinguishing between different customer segments.

Occupation: The second highest value, suggesting it's also a strong predictor.

Insurance\_Products\_Owned: Also seems to have a strong association with the target.

Wealth\_Status: A bit less associated compared to the first three, but still significant.

Bankcard\_Interest: This has a lower chi-square value, indicating a weaker, but still present, relationship.

Interactions\_With\_Customer\_Service: Even lower in predictive power compared to the variables listed so far.

Customer Preferences: Quite low, suggesting it's not a strong predictor.

Gender: Has a very low chi-square value, suggesting it has little to no predictive power with respect to customer segmentation in this context.

Education\_Level: Similar to gender, indicating a weak association.

Preferred\_Communication\_Channel: Also a weak predictor based on the low chi-square value.

Preferred\_Contact\_Time: Very low, suggesting it might not be useful for predicting customer segment.

Preferred\_Language: The lowest chi-square value, implying it has the least association with customer segmentation among the variables listed.

**Missing Values:**

All variables have 53503 observations, indicating no missing values within the dataset.

**Data Error:**

No specific data errors are apparent from the provided summary statistics. However, thorough data validation and quality checks should be conducted to ensure accuracy and reliability.

**Skewness and Kurtosis :**

Skewness provides a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive, zero, negative, or undefined.

Age has a skewness of 0.116046, which suggests a very slight bias towards the right tail; the distribution is nearly symmetrical.

Coverage\_Amount has a skewness of 0.165374, also indicating a slight right tail, but it's nearly symmetrical.

Customer\_ID has a slight negative skewness, -0.1864, indicating a slight bias towards the left tail.

Income\_Level has a skewness of 0.109042, suggesting a slight right-tail bias.

Premium\_Amount has a negative skewness of -0.27418, which suggests the distribution leans slightly towards the left.

Purchase\_History has a very slight negative skewness of -0.02053, indicating a distribution that's almost symmetrical.

**Kurtosis**

Greater than 3 indicates a distribution with tails heavier than a normal distribution (Leptokurtic).

Equal to 3 indicates a normal distribution (Mesokurtic).

Less than 3 indicates a distribution with tails lighter than a normal distribution (Platykurtic).

All the variables have kurtosis values less than 3, indicating that they have lighter tails than the normal distribution. This can imply that there are fewer extreme values (outliers) in the tails .

**Variables:**

A screenshot of a computer

Description automatically generated

**Appendix :**

https://www.kaggle.com/datasets/ravalsmit/customer-segmentation-data