Eunbin Yoo

Data Science Project

Car Insurance Customer Segmentation

Customer Analysis Using PCA and K-means Clustering Algorithm

Problem Statement

An insurance company seeks to expand its offerings by introducing vehicle insurance to its existing health insurance customers. The project aims to identify and analyze distinct segments of health insurance customers who show interest in car insurance, in order to tailor marketing strategies and optimize customer engagement.

Project Task:

The goal of this project is to conduct K-means clustering on different subsets of dataset to determine the optimal number of clusters among health insurance customers interested in car insurance. The project also aims to derive insights from the data that can be used to inform business decisions.

Dataset Used

- The dataset used in this project is “Health Insurance Cross Sell dataset” uploaded on Kaggle under a public domain.
- This dataset involves information about demographics, vehicles and insurance policy. Each row represents a health insurance customer.

Data Dictionary

Variable	Definition	Key
id		
Gender	Gender of a customer	'Male', 'Female'
Age	Age of a customer	
Driving_License	Does a customer have a driving license?	0=No, 1=Yes
Region_Code		
Previously_Insured	Is a customer insured previously?	0=No, 1=Yes
Vehicle_Age	Age of a customer's vehicle	'1-2 Year', 'Less than 1 Year', 'More than 2 Years'
Annual_Premium	Annual premium ammount	
Policy_Sales_Channel	Sales channel	
Vintage	Days customer associated for with company	
Response		

Data Head

	Gender	Age	Driving_License	Region_Code	Previously_Insured	Vehicle_Damage	Annual_Premium	Policy_Sales_Channel	Vintage	Response	Age_Group	Vehicle_Age_num
0	1.0	44	1.0	28.0	0.0	1.0	40454	24.0	217	1	2.0	2.0
1	1.0	76	1.0	3.0	0.0	0.0	33536	24.0	183	0	5.0	1.0
2	1.0	47	1.0	28.0	0.0	1.0	38294	24.0	27	1	2.0	2.0
3	1.0	21	1.0	11.0	1.0	0.0	28619	145.0	203	0	0.0	0.0
4	0.0	29	1.0	41.0	1.0	0.0	27496	145.0	39	0	0.0	0.0

Approach

The project involves these steps:

1. **Data Pre-processing**

   Data cleaning, handling missing values, feature scaling, encoding categorical variables, and subset dataset by variables

2. **Principal Component Analysis (PCA)**

   Dimension reduction using Principal Component Analysis

3. **Cluster Analysis**

   K-means clustering on two different subsets of dataset (business and demographics), Within Clusters Sum of Square (WCSS), Silhouette score

4. **Segmentation Analysis**

   Define and visualize segments, and derive actionable insights

Cluster Analysis

The cluster analysis will be conducted on the two different subsets of dataset:

Business Cluster

'Business cluster' covers the trends from four business related variables:

- *Annual Premium*
- *Policy Sales Channel*
- *Vintage*
- *Region Code*

Demographic Cluster

'Demographic cluster' groups customers based on common demographic characteristics:

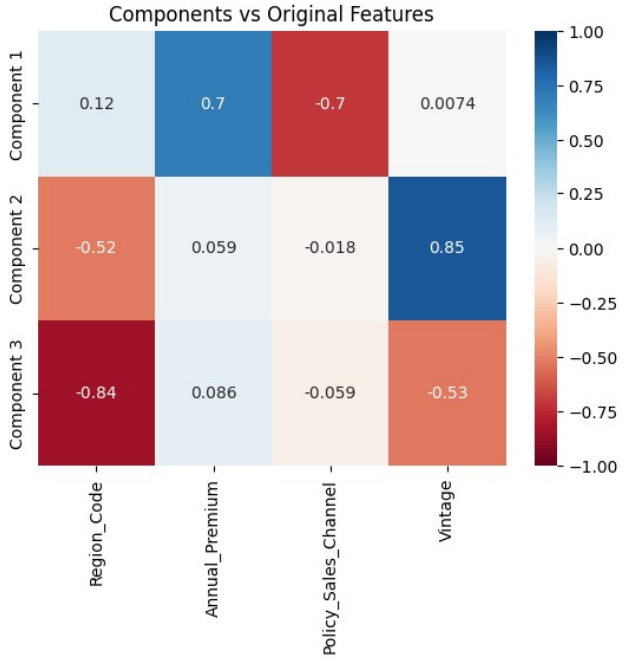
- *Age*
- *Gender*
- *Region Code*
- *Driving License*
- *Vehicle Age*

Business Cluster Analysis

K-means clustering using
business factors

Principal Component Analysis

3 components explained 80% of the variance of the data. Our features have been reduced to three components.



1. **The first component:**

   A high positive correlation to Annual Premium (0.7) and negative to Policy Sales Channel (-0.7) shows higher-paying customers prefer specific channels.

2. **The second component:**

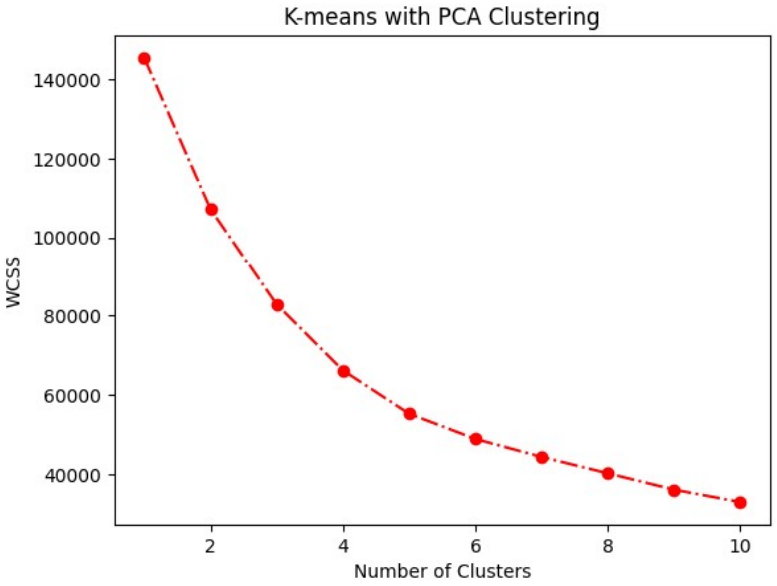
   A high positive correlation to Vintage (0.85) and negative to Region Code (-0.52), indicates long-standing customers cluster in certain regions.

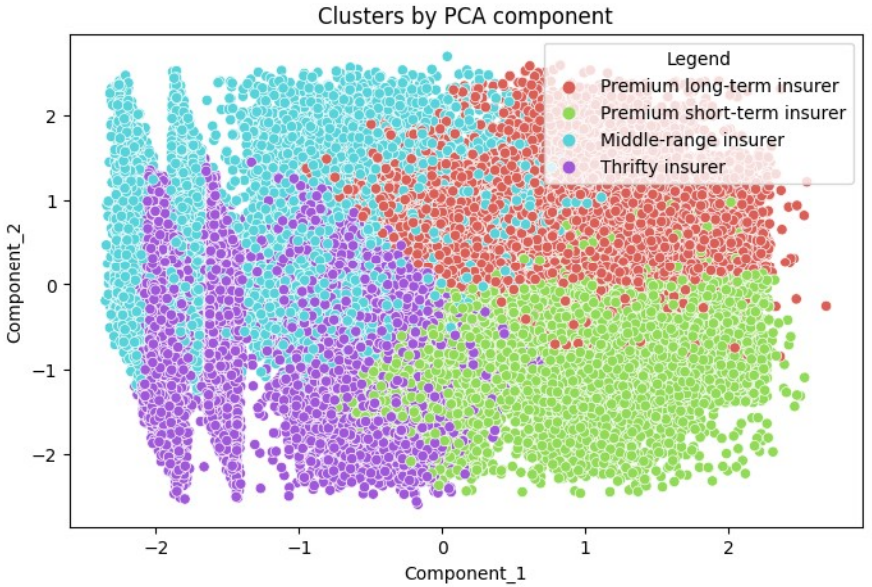
3. **The final component:**

   Strong negative correlation to Region Code (-0.84) represents newer customers from specific geographic areas.

K-means Clustering

Evaluation using WCSS and clustering using K-means clustering algorithm





Using WCSS and elbow method, four clusters were chosen as the optimal number of cluster for ‘Business Cluster’. The four segments are labeled based on the values of variables and components.

Segment Profiling

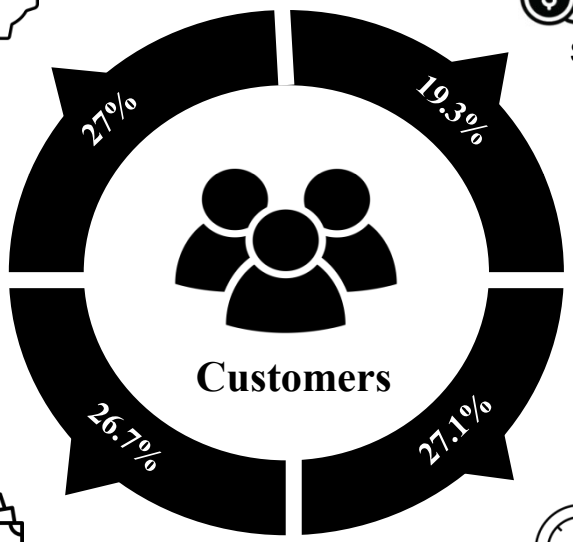
Segment profiling for Business Cluster

Thrifty Insurer

Segment that pays the lowest average annual premium ($18587 on average) and prefers the highest Policy Sales Channel (code 125 on average)

Mid-Range Insurer

Segment that pays the middle range price of average annual premium ($22335 on average) compared to other clusters



Segment that pays the highest annual premium ($38522 on average) and have been associated with the company for the shortest period (77 days on average)

Short-Term Premium Insurer

Segment that pays the high average annual premium ($38395 on average) and have been associated with the company for the longest period (234 days on average)

Long-Term Premium Insurer

Segment Analysis

Customer segmentation metrics

Total Revenue

Total Revenue Potential

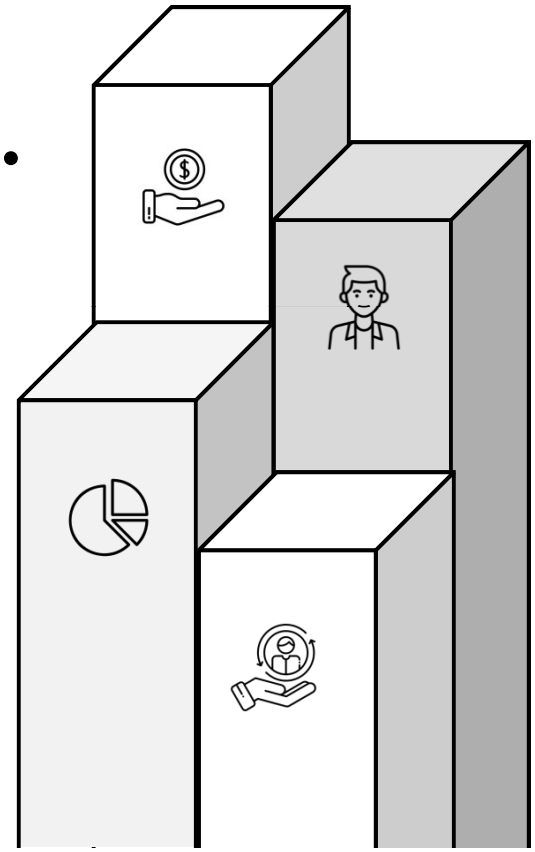
1.Long-Term Premium Insurer: $470,191,165

2.Short-Term Premium Insurer: $463,969,003

Market Size

Market Size in Percentage

1. Long-Term Premium Insurer: 27.1%
2. Thrifty Insurer: 26.9%



Customer Value

Average annual premium spend per customer

1. Long-Term Premium Insurer: $38,395
2. Short-Term Premium Insurer: $ 38,522

Customer retention

Average days of association

1. Longest retention:

Premium Long-Term Insurer (234 days)

2. Shortest retention:

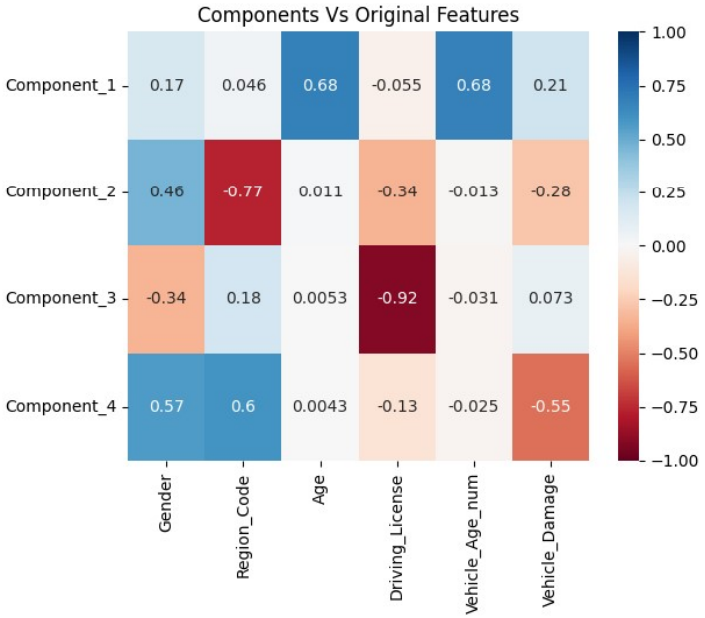
Premium Short-Term Insurer (76 days)

Demographic Cluster Analysis

K-means clustering using
demographic factors

Principal Component Analysis

4 components explained 77% of the variance of the data, so I reduced the features to four components.



1. **The first component:**

   A strong positive correlation to Age (0.68) and Vehicle Age (0.68) shows a pattern where older age groups tend to own older vehicles.

2. **The third component:**

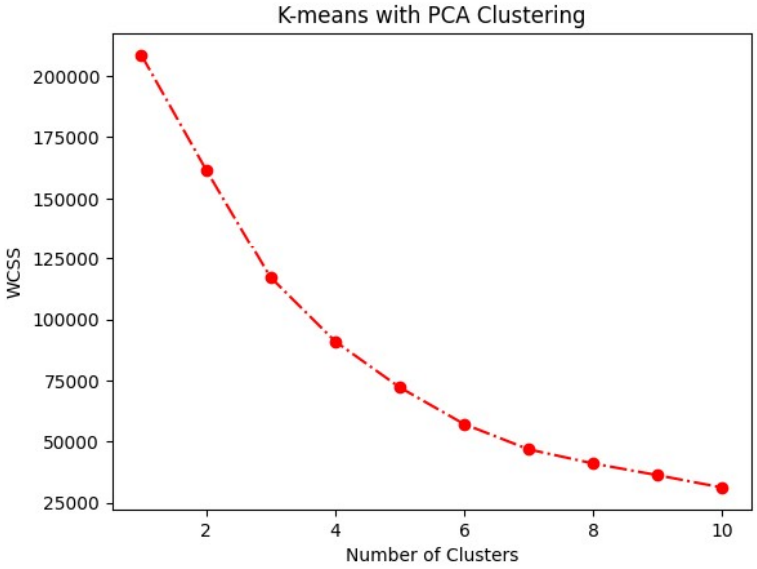
   A strong negative correlation to Driving License (-0.92) indicates customers who do not have driving license.

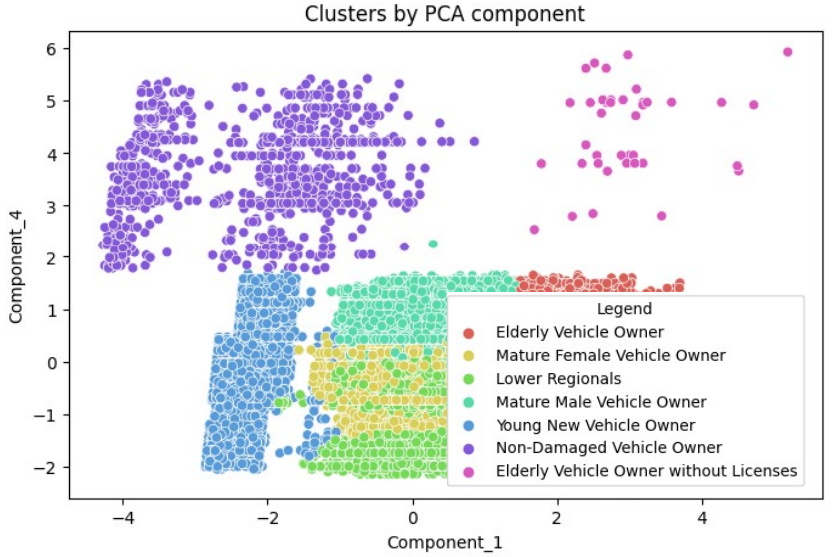
3. **The final component:**

   Positive correlation to Gender (0.57) and Region Code (0.6) and negative correlation to Vehicle Damage (-0.55) represent a pattern where customers in particular gender group and region are less likely to have damage on their vehicles.

K-means Clustering

Evaluation using WCSS and clustering using K-means clustering algorithm

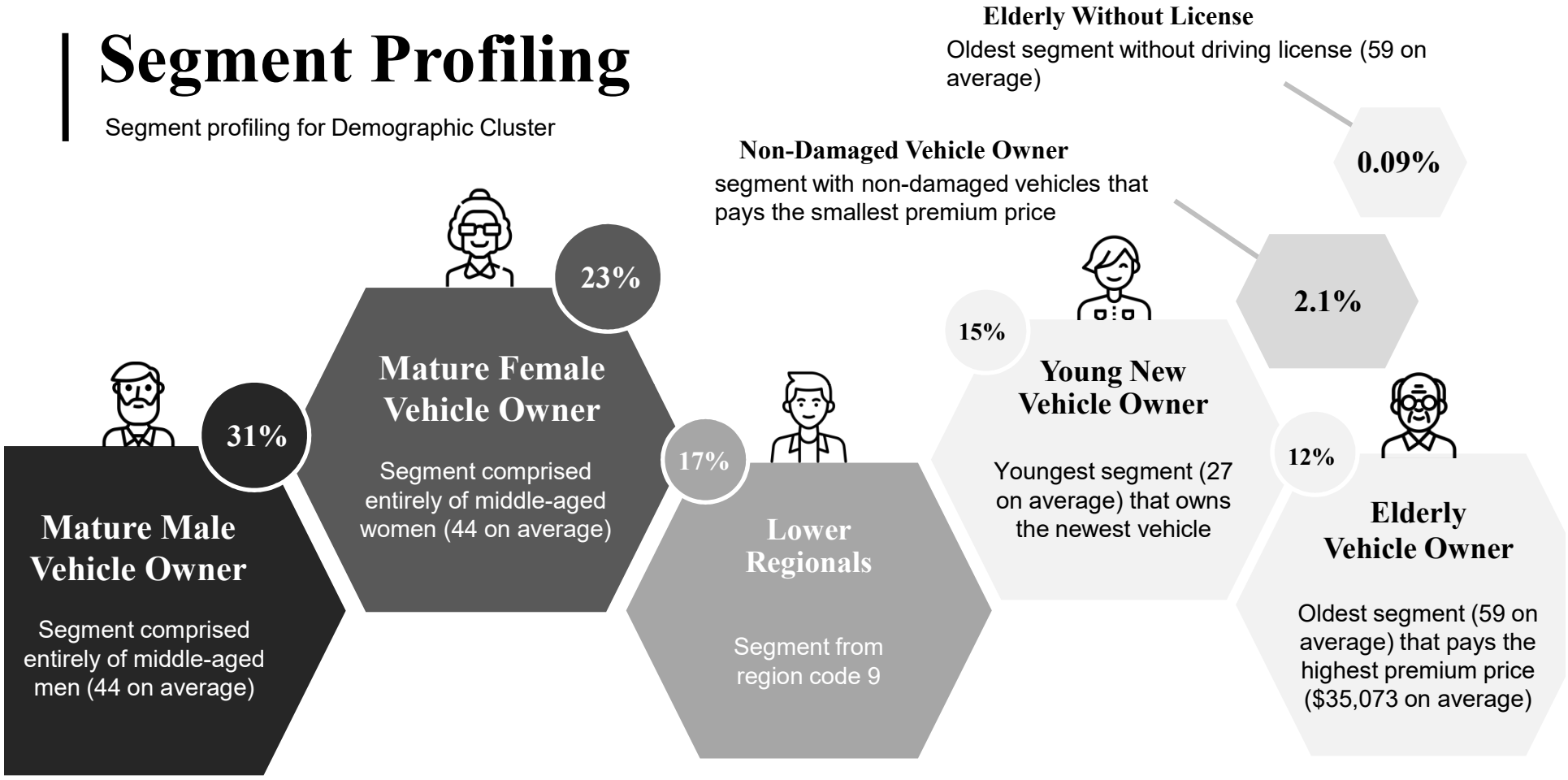




Using WCSS and elbow method, seven clusters were chosen as the optimal number of cluster for ‘Demographic Cluster’. The seven segments are labeled based on the values of variables and components.

Segment Profiling

Segment profiling for Demographic Cluster



Segment Analysis

Customer segmentation metrics

Total Revenue

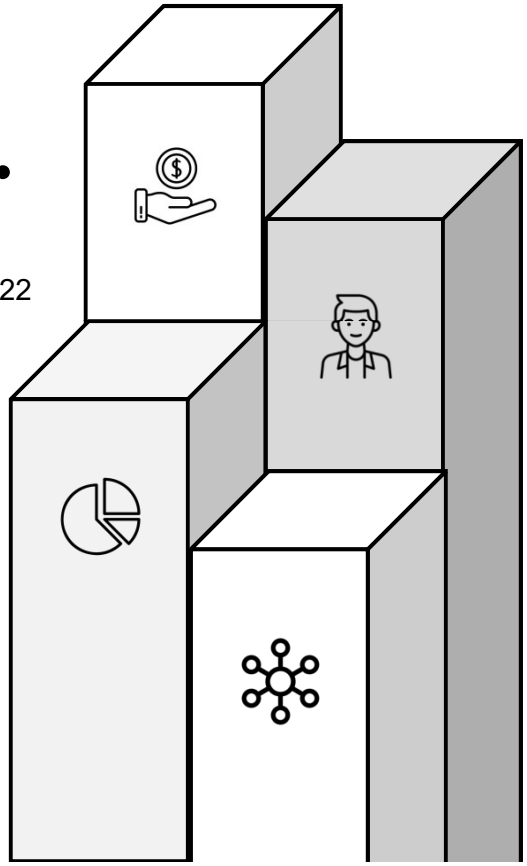
Total Revenue Potential

1. Mature Male Vehicle Owner: $425,896,572
2. Mature Female Vehicle Owner : $319,937,722

Market Size

Market Size in Percentage

1. Mature Male Vehicle Owner: 30.6%
2. Mature Female Vehicle Owner : 23.0%



Customer Value

Average annual premium spend per customer

1. Elderly Vehicle Owner : $35,073
2. Elderly Vehicle Owner without Licenses: $32062

Channel Preference

Average Sales Policy Channel code

1. 60~70: Elderly Vehicle Owner, Elderly Vehicle Owner without Licenses
2. 80~90: Mature Female Vehicle Owner, Mature Male Vehicle Owner
3. 110~120: Young New Vehicle Owner, Non-Damaged Vehicle Owner

Findings

❶ Which segments are the most valuable to the company in terms of revenue potential, loyalty, or growth opportunities?

'Premium long-term insurer' and 'Mature Male Vehicle Owner' seem to be the most valuable segment to the company. 'Premium long-term insurer' is a segment with the largest total revenue ($470,191,165) and the longest average duration (234 days on average), suggesting a higher level of loyalty or retention among these customers. 'Mature Male Vehicle Owner' emerges as the segment with the highest revenue potential ($425,896,572) and has the largest market size which present greater growth opportunities.

❷ Do different segments prefer different communication channels?

There is a tendency that insurers who pay higher annual premium and are in older age groups prefer to be outreached by lower Policy Sales Channel code.

'Premium long-term insurer' and 'Premium short-term insurer' from business cluster and 'Elderly Vehicle Owner' and 'Elderly Vehicle Owner without License' from demographic cluster prefer the channel code around between 60 and 70. 'Thrifty insurer' from business cluster and 'Young New Vehicle Owner', the youngest segment from the demographic cluster, prefer the code greater than 120.

Findings

❸ Are there differences in duration of a customer's association with the company across segments?

‘Short-Term Premium Insurer' has the shortest duration of association with the company (77 days on average). Despite paying the highest price of average annual premium, customers in this segment are relatively new to the company, indicating a recent acquisition of customers who opt for higher-priced insurance plans.

In contrast, ‘Long-Term Premium Insurer' has the longest duration of association with the company (234 days on average). Despite also paying a high price of average annual premium, customers in this segment have been with the company for a significantly longer period, suggesting a higher level of loyalty or retention among these customers.

❹ Are there the age-related trends and preferences?

Customers in older age groups pay a high average annual premium price, suggesting a potential preference for comprehensive coverage and higher levels of insurance protection.

The largest age group is 40-50 age range (70% in total), contributing the most to total revenue potential (949,627,836 dollars in total). Younger age groups in their 20s and 30s tend to have relatively newer vehicles and pay a lower average annual premium price than older age groups. Elderly segments, both with and without licenses, spend the highest average annual premium per customer.

Conclusion

In conclusion, the customer segmentation analysis has provided valuable insights into the characteristics, behaviors, and preferences of health insurance customers interested in vehicle insurance. By leveraging PCA and k-means clustering algorithms, we identified distinct customer segments based on both business and demographic variables.

From the analysis, we have identified key customer segments such as ‘Long-Term Premium Insurer' from the business cluster and 'Mature Male Vehicle Owner' from the demographic cluster, which demonstrate high revenue potential, market size, and growth opportunities. Understanding the Policy Sales Channel preferences of each segment can optimize marketing strategies and improve campaign effectiveness. The age-related trends and preferences identified underscore the importance of understanding demographic dynamics in shaping customer behavior and preferences.

By leveraging these insights, businesses can tailor their products, services, and marketing strategies to better meet the needs of different customer segments, ultimately driving customer satisfaction, loyalty, and revenue growth.

Thank you

- Link to Original Dataset: https:
  www.kaggle.com/datasets/anmolkumar/health-insurance-cross-sell-prediction/data?select=train.csv
- Github Link for code, resources and more information:
  https://github.com/bebe5004/Eunbin-Yoo-s-Portfolio/tree/main/Car%20Insurance%20Customer%20Segmentation%20Analysis
- Github page for more projects:
  https://github.com/bebe5004/Eunbin-Yoo-s-Portfolio/tree/main?tab=readme-ov-file#readme