Trevor Maxwell

DSC 680

2024-03-31

Customer Segmentation Using K-Means Clustering

Companies have loads of data on their customer's attributes and behavior, but it is not always used efficiently to drive the company forward. This vast amount of data getting created, processed, and stored by organizations is derived from the development of information technology (Verdenhofs, Atis, and Tatjana Tambovceva, 2019). Customer segmentation is a way to uncover insights into the customer base and achieve actionable items from what could seem to be a random pool of customers. This technique uses the customer data by clustering them into groups and the customers in each group behave similarly when compared to the customers in other groups (Dhandayudam and Krishnamurthi, 2014). The problem with large volumes of customer data is the difficulty of understanding the customer base in its entirety, and this is solvable with customer segmentation. Shirole, Rahul, et al used these techniques on E-commerce data and discovered the customer population was categorized into four groups and which groups generated the highest revenue versus the lowest revenue (2021). The analysis in this paper used the K-Means clustering algorithm similarly to group the customers into clusters to discover actionable insights behind the data.

The dataset used for this analysis included personal, transactional, and geographical attributes of customers of an insurance company. The personal attributes included age, education level, income, marital status, and occupation among a few others. The transactional attributes included the customer's premium, coverage amount, policy type, and purchase history. The geographical information included the market of the customer.

The main attributes used were the transactional data, age, and market of the customer. Since insurance company policies vary in price and coverage from market to market, the location with the highest volume of customers was the base population for this analysis. This was customers with the location of Lakshadweep with a count of 2,140. Additionally, policy types vary similarly. The coverage of a business insurance policy differs from that of an individual insurance policy. The policy with the largest count in the dataset was the group policy type, therefore the group policy was also part of the base population for this analysis. The number of customers in a group policy in the Lakshadweep market was 746.

Additional preprocessing steps included checking for duplicate customers based on the Customer ID. While there were many duplicates, it appeared as though the Customer ID did not reflect the same customer even though the ID was the same. For example, the Customer ID "1" had 20 different observations that varied in nearly all features. Therefore, it was concluded that there were no duplicate customers, and the column was dropped. There were also no completely duplicated observations or missing values in any observations. However, there were two different date formats for the purchase history date, and this was cleaned to have a unified date for that feature and used during the exploratory data analysis portion of the analysis. Ultimately the features that remained for the modelling were the customer's age and transactional attributes (income, premium amount, and coverage amount). These remaining features were scaled using a min-max scaler since K-Means clustering is a distance-based algorithm.

To obtain the optimal number of clusters, the elbow method was performed. This method finds the optimal number by calculating the sum of the square distance between points in a cluster and the cluster centroid and plots them in a graph (Saji, 2024). Then, the optimal number

of clusters is the “elbow point” on the graph. For this analysis, the optimal number of clusters is 3.

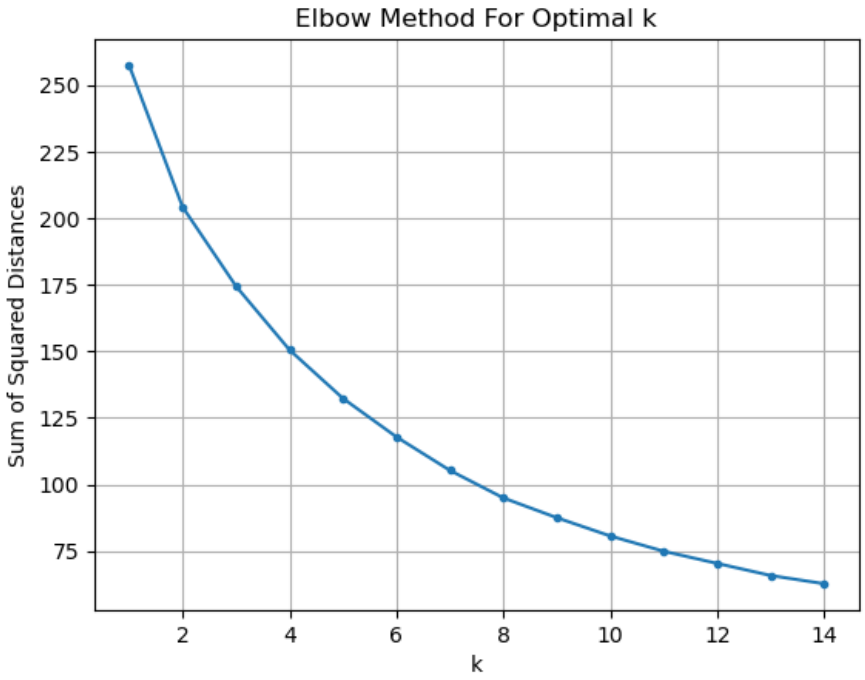
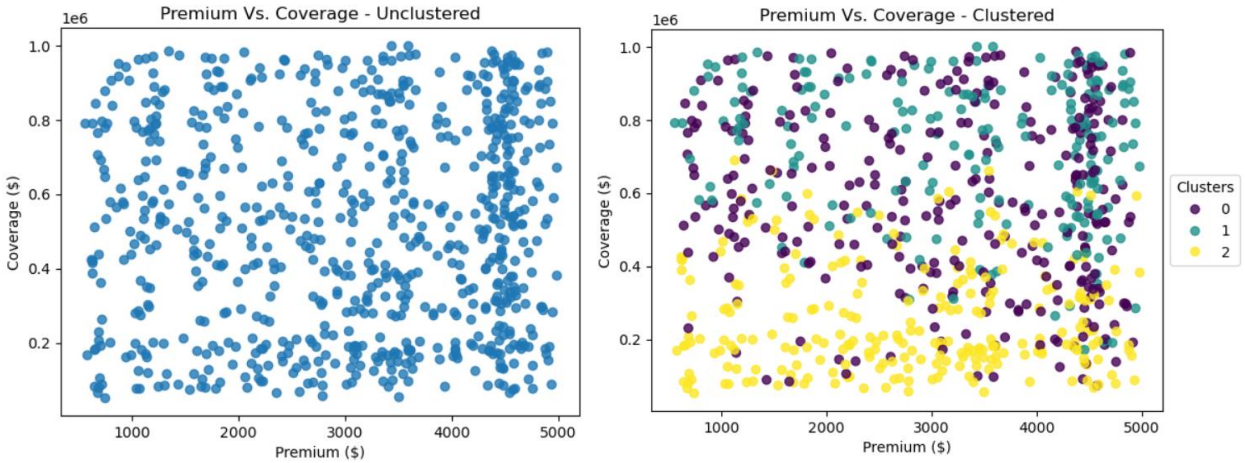


Fig 1: Elbow plot to determine number of optimal clusters.

Once the customers were added to a cluster, additional analysis was performed which uncovered greater insights into the data. As seen below, the scatter plots for premium amount, coverage amount, and income seemed random before clustering and once clustered a clear grouping can be observed.



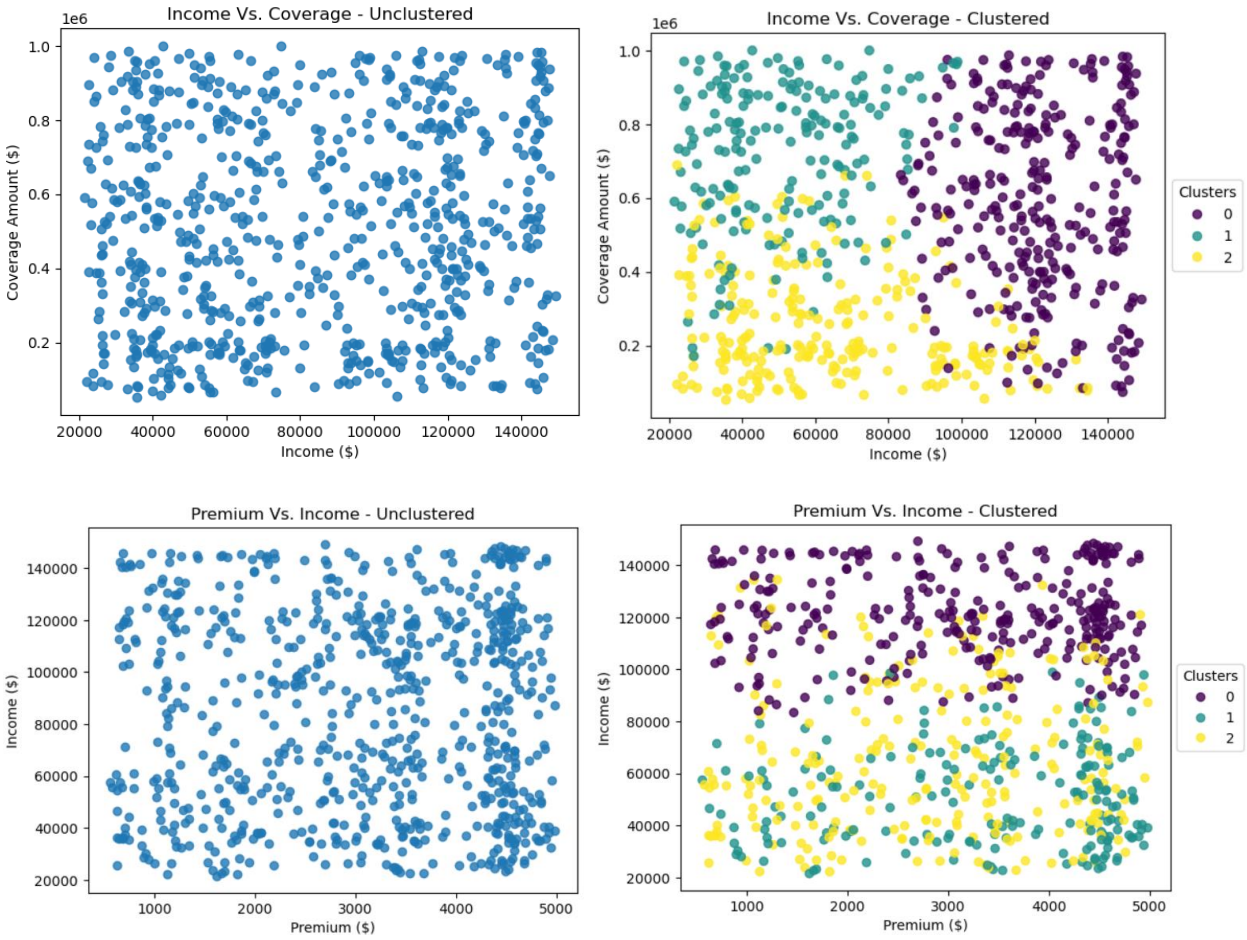


Fig. 2: Scatter Plots of income, premium, and coverage amounts before and after clustering.

The distinct groups were observed in the scatter plots after the K-Means clustering algorithm was deployed. Additionally, the average premium and coverage amount were calculated for each cluster to determine which cluster was most cost-effective for the business. This was cluster 2 with an average premium amount close to the other two clusters, however, the average coverage amount was significantly less than the other two clusters.

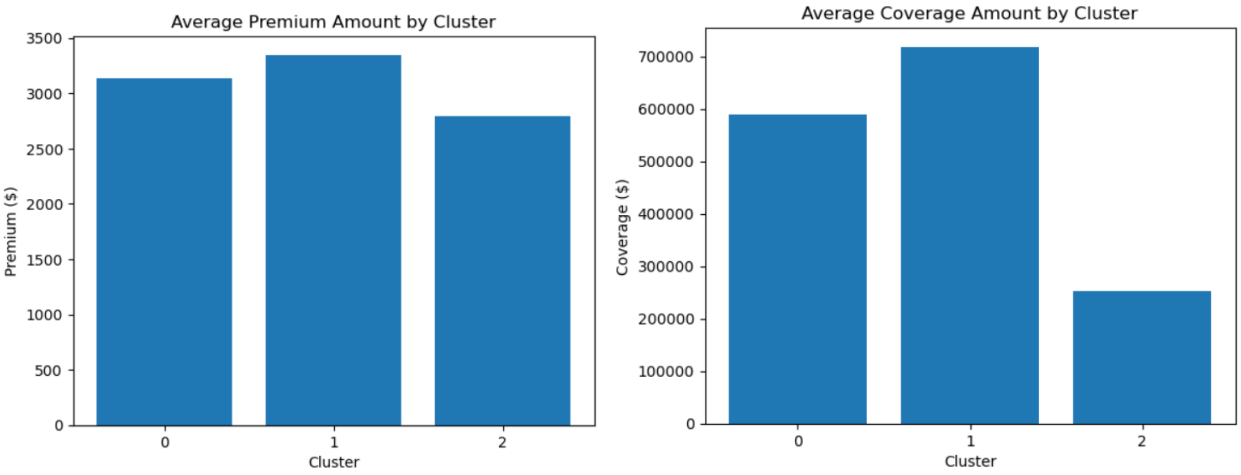


Fig. 3: Average coverage and premiums amounts by cluster.

There are several actions the insurance company could deploy to become more generate more revenue based on the results of this analysis. The company could deploy marketing campaigns to move customers from the two less cost-effective clusters into similar policies that the customers in the most cost-effective cluster are in. Another option could be adopting principles from the policies in the most cost-effective cluster to the policies in the other two clusters. This could also be an opportunity to develop new policies similar to the most cost-effective cluster's policies. These new policies could be marketed to new and existing customers. This analysis provides strategies to generate more revenue, improve current policies, and generate new policies.

Limitations/Challenges

A few challenges arose throughout the exploratory data analysis and modeling portions of this analysis. The purchase history was a transactional feature in the dataset and contained the date the policy was purchased, but it was noisy when plotted and did not add value to the model. To input this data into the model, two different techniques were tried. It was broken down into the month (1 through 12) as well as the quarter (1 through 4) of the purchase. Neither of these two

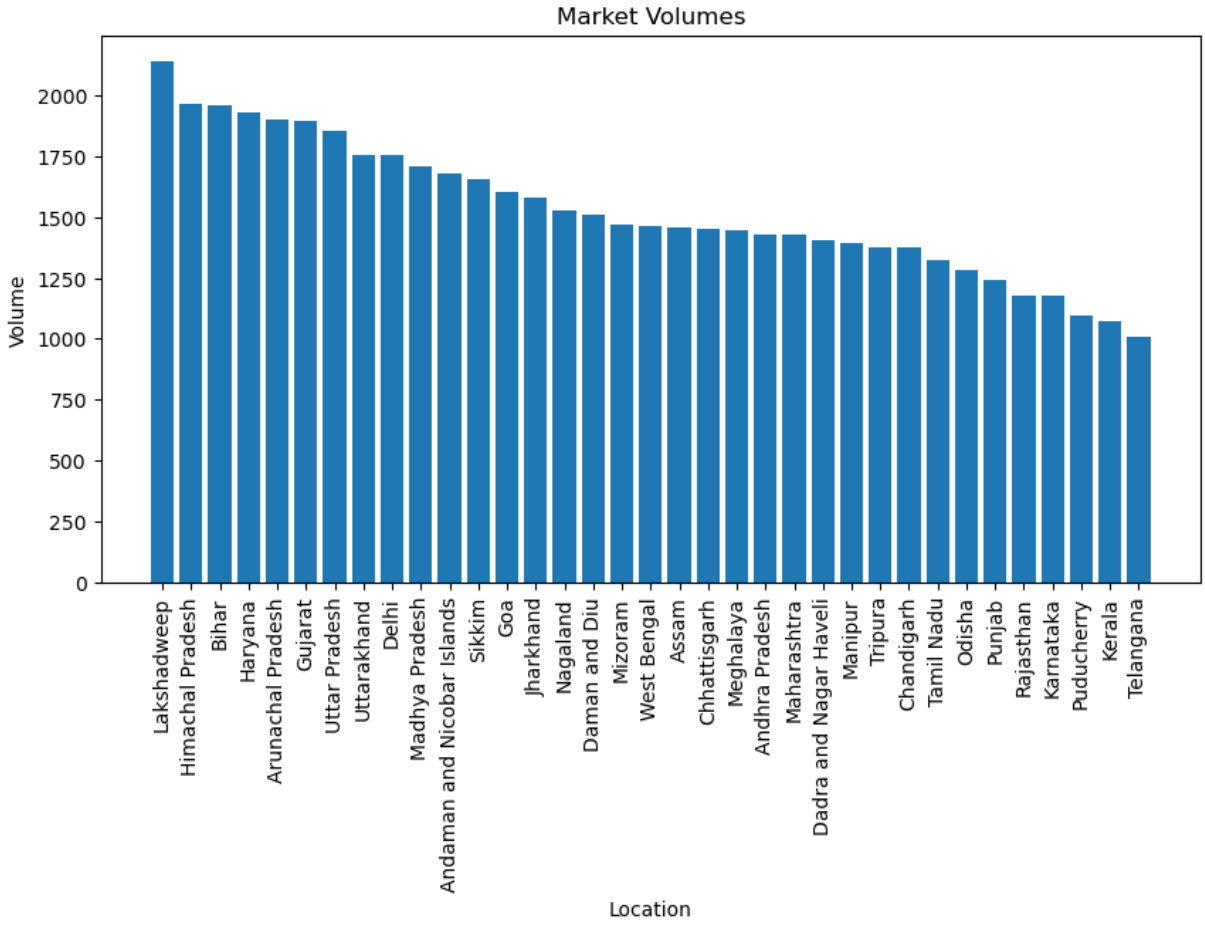
techniques added value to the model. Another attempted technique was one-hot encoding the categorical features so they could be inputted into the model. When the data was clustered with these in the model, the scatter plots were just as random as the scatter plots before clustering. Therefore, all categorical columns were dropped. Another limitation is not knowing exactly what products/offerings were included with each policy.

Ethical Assessment

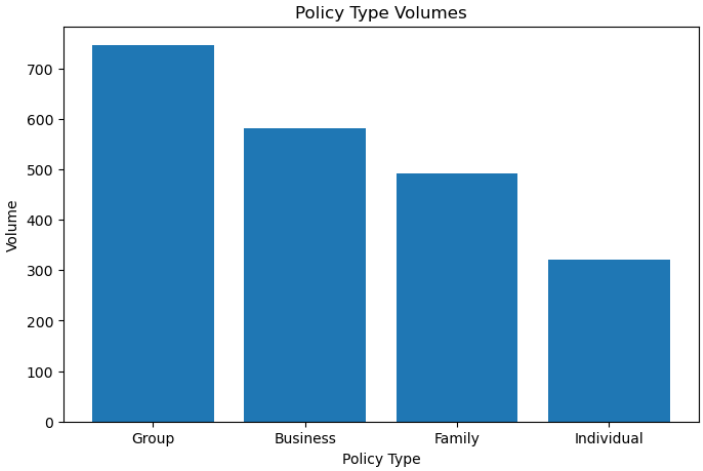
Ethical considerations arise during the recommendations suggested to the company based on the clusters. Removing perks from policies in the less cost-effective clusters to be more like the policies in the most cost-effective cluster could result in customers losing their needs for the policy. Also increasing premiums to become more cost-effective would affect the customers with lower income more than the customers with higher income. Another consideration would be decreasing the coverage amount overall which could lead to more out-of-pocket costs for the customers should they submit claims and reach the coverage amount quicker due to the decrease.

10 Questions an Audience Would Ask You

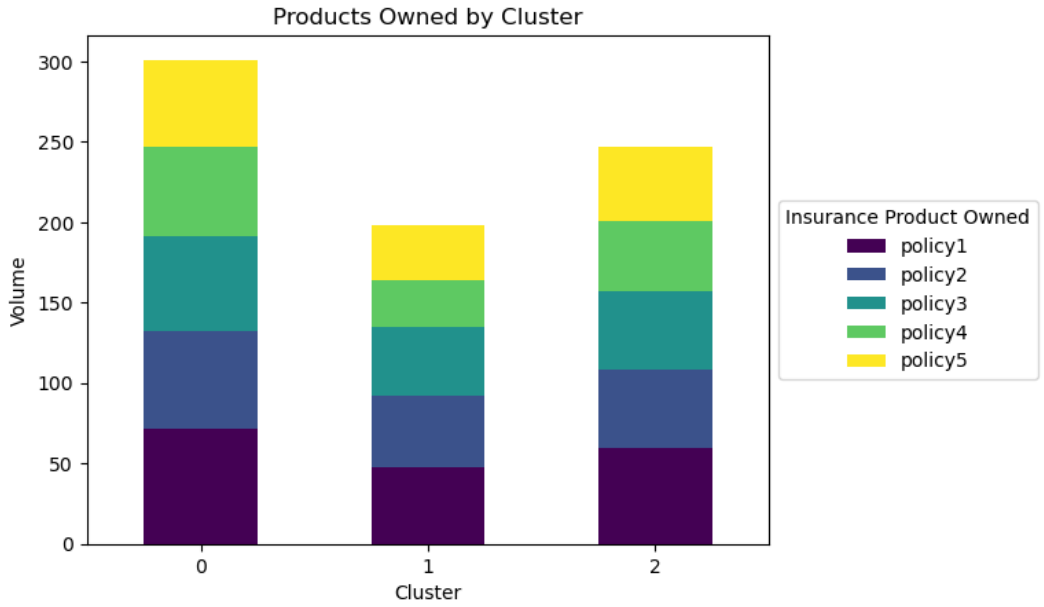
1. What is the breakdown of volumes for the other markets?



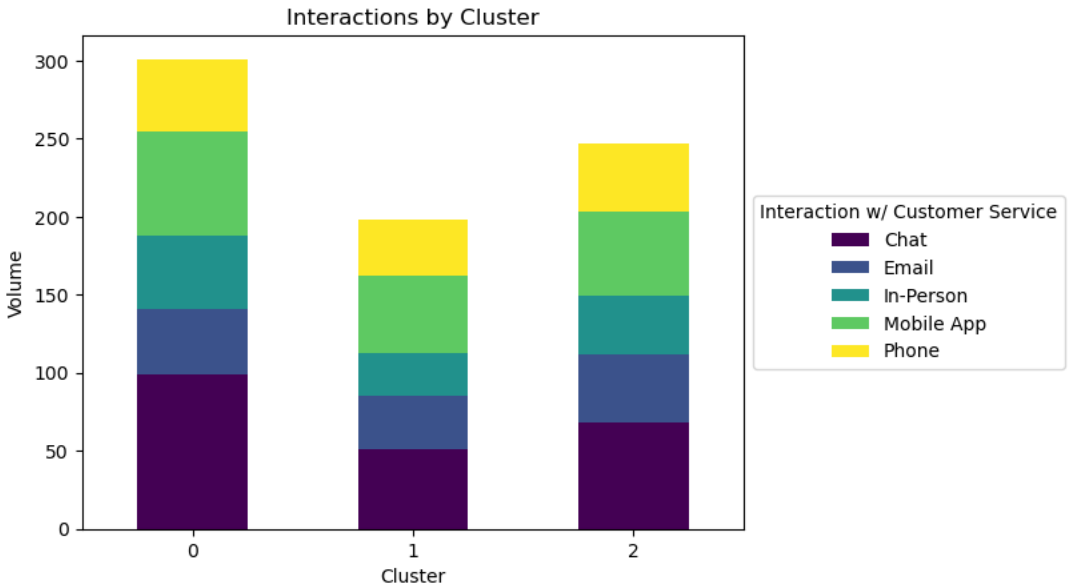
2. What is the breakdown of volumes for the other policy types?



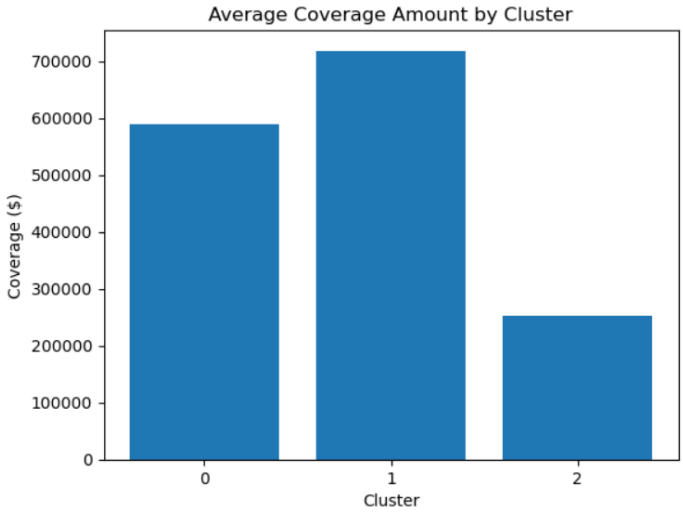
3. Would this analysis work for other markets?
    a. Yes, the analysis would work for other markets.
4. Would this analysis work for other policy types?
    a. Yes, the analysis would work for other policy types.
5. What is the breakdown of insurance products owned by cluster?

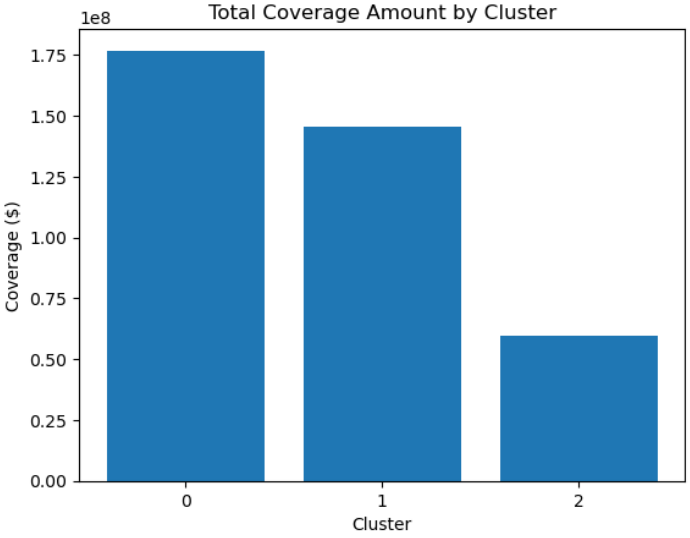


6. Did the clusters interact with customer service differently?

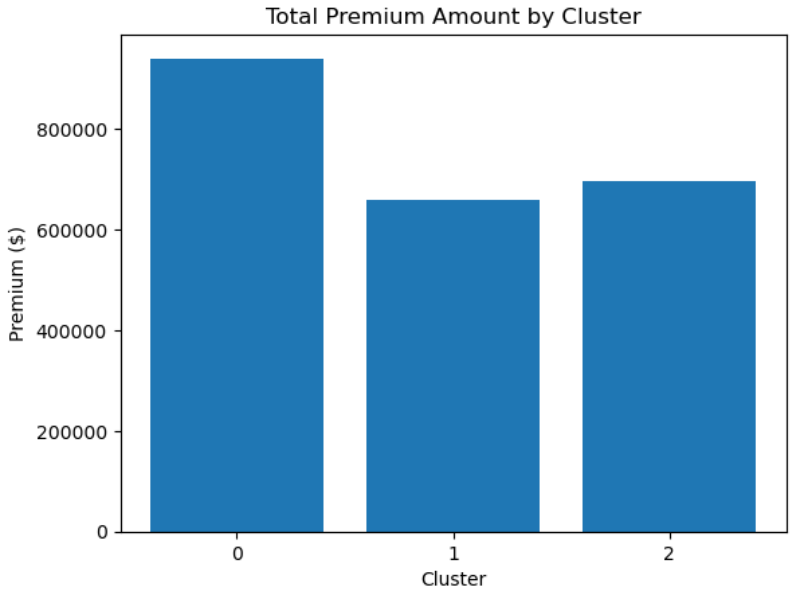


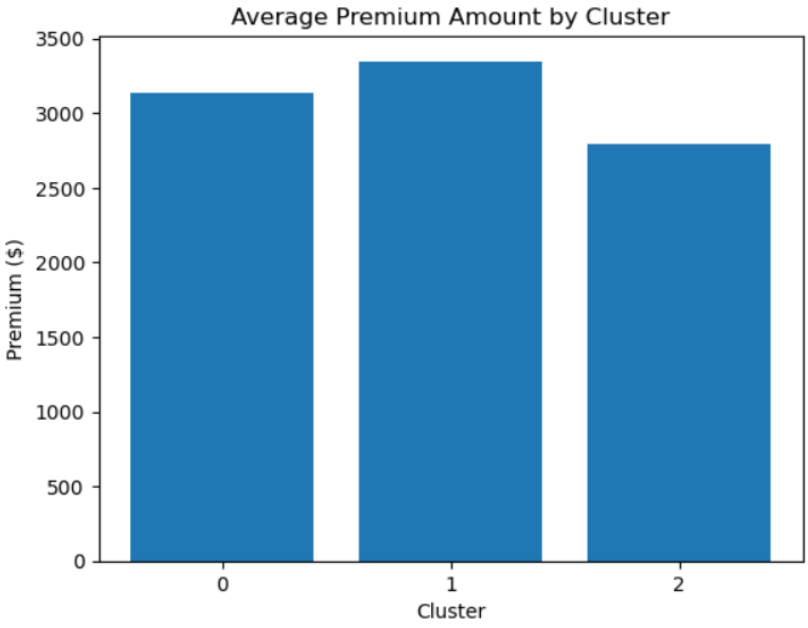
7. What feature(s) impacted the clustering the most?
    a. Difficult to determine, but it seems as if income had the most impact based off the scatter plots.
8. How many customers are in each cluster?
    a. Cluster 0: 301
    b. Cluster 1: 198
    c. Cluster 2: 247
9. What are the total and average coverage amounts for the dataset?





10. What are the total and average premium amounts for the dataset?





References

Dhandayudam, Prabha, and Ilango Krishnamurthi. “A rough set approach for customer segmentation.” *Data Science Journal*, vol. 13, no. 0, 9 Apr. 2014, pp. 1–1, https://doi.org/10.2481/dsj.13-019.

Saji, Basil. “Elbow Method for Finding the Optimal Number of Clusters in K-Means.” *Analytics Vidhya*, 7 Jan. 2024, www.analyticsvidhya.com/blog/2021/01/in-depth-intuition-of-k-means-clustering-algorithm-in-machine-learning/.

Shirole, Rahul, et al. “Customer segmentation using RFM model and K-means clustering.” *International Journal of Scientific Research in Science and Technology*, 1 June 2021, pp. 591–597, https://doi.org/10.32628/ijsrst2183118.

Verdenhofs, Atis, and Tatjana Tambovceva. “Evolution of customer segmentation in the era of Big Data.” *Marketing and Management of Innovations*, 2019, pp. 238–243, https://doi.org/10.21272/mmi.2019.1-20.