Data Mining Project

MASTER'S DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance Customer Segmentation and Clustering: Enhancing Business Strategy.

Mahmoud Ahmed, number: 20221397

Dec, 2022

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

INDEX

1. Introduction

In today's competitive business environment, it is essential for companies to effectively target and serve their customer base in order to remain competitive and achieve growth. One important aspect of this is market segmentation, which involves identifying and analyzing different groups of customers based on their characteristics and behaviors. By understanding the differences between these segments, companies can tailor their marketing, sales, and product offerings to better meet the needs and preferences of their target customers.

A2Z Insurance is a long-standing Portuguese company that offers a variety of insurance services, including motor, household, health, life, and work compensation coverage. While A2Z primarily serves Portuguese customers, it also acquires a significant portion of its customer base through its website. To improve its targeting and marketing efforts, A2Z has provided us with an Analytic Based Table (ABT) containing data on a sample of 10,290 active customers. Our task is to use this data to segment the customer base and identify relevant clusters of customers based on their characteristics and insurance interests.

By applying data mining techniques to the A2Z customer database, we hope to provide the company with valuable insights that can inform its future marketing and sales strategies. We aim to understand the demographics and value of each customer segment, as well as the types of insurance that they are most interested in purchasing. Through this process, we hope to help A2Z better serve the needs of its customers and identify new opportunities for growth and expansion.

2. Data Understanding

The proposed dataset ABT (Analytic Based Table) contains data regarding **10.290** Customers from its active database. These are customers that had at least one insurance service with the company at the time the dataset was extracted.

Each customer is described by a set of **14 attributes** (originally), which provide information on both customer demographics and insurance service. All the statistics and considerations were made before the data cleaning phase.

2.1. Data Semantics

This section contains a brief description of each attribute in the original dataset Table 2.1 .

2.2. Exploratory Data Analysis (EDA)

- Exploratory Data Analysis (EDA) is a crucial step in the data mining process that involves exploring and analyzing the data to understand its characteristics and patterns.

The data provided consists of 13 columns and 10,296 rows. After initially analyzing the data, we have noticed the following:

- There are missing values in some of the columns. We will need to decide how to handle these missing values before performing any statistical analysis.
- The data types of the columns consist of 11 float64 columns (numerical data) and 1 object column (categorical data). We will need to ensure that we are using appropriate techniques and tools for each data type.
- The Children column is a binary variable that can be treated as a categorical variable as it represents two distinct categories ("has children" and "does not have children").
- The GeoLivArea column can also be treated as a categorical variable as it represents distinct categories (1, 2, 3, 4).
- To understand the distribution of values in each column, we will visualize the data using histograms or other appropriate plots. This will help us to understand the shape of the distribution and identify any outliers or anomalies in the data.

2.2.1. Data Distributions and Statistics

- Data distributions for the columns in the provided dataset were analyzed using histograms and other appropriate plots. The resulting visualizations allowed us to better understand the shape of the distributions and identify any potential outliers or anomalies in the data.

Define Metric and Non-Metric Features

In the data exploration phase, it is important to identify the features of the dataset that will be used for analysis. These features can be divided into two categories: metric and non-metric features.

- **Metric Features:** FirstPolYear, BirthYear, MonthSal, CustMonVal, ClaimsRate, PremMotor, PremHousehold, PremHealth, PremLife and PremWork.
- **Non-Metric Features:** GeoLivArea, Children and EducDeg.

Uni-variate distribution

To understand the distribution of values in each column, we visualized the metric and Non-metric data distributions using box plots Figure 2.1 and histograms Figure 2.2 , Figure 2.3 . These plots gave us a better understanding of the shape of the distribution and helped to identify any outliers or anomalies in the data.

Statistics Summary

We also calculated several descriptive statistics for each column, including the mean, standard deviation, minimum, maximum, and percentiles. These statistics provided us with information about the central tendency and spread of the data, which can be useful for identifying patterns and trends in the data.

Some observations and insights that can be drawn:

- **Demographics Related:**
  - The average age of customers in the dataset is approximately 50 years.
  - The majority of customers in the dataset have an education degree of b'3 - BSc/MSc.
  - The average monthly salary of customers in the dataset is approximately 2623 euros.
  - Most customers in the dataset live in an area with a geographical classification of 2 or 3.
  - About 70% of customers have children.
- **Insurance Related:**
  - The average loyalty years, or the number of years a customer has been with the company, is approximately 24 years.
  - The average customer monetary value is approximately 204 euros.
  - The average claims rate, or the proportion of insurance claims made by a customer, is approximately 0.68.
  - The average motor insurance premium paid by customers is approximately 329 euros.
  - The average household insurance premium paid by customers is approximately 154 euros.
  - The average health insurance premium paid by customers is approximately 168 euros.
  - The average life insurance premium paid by customers is approximately 30 euros.
  - The average work compensation insurance premium paid by customers is approximately 30 euros.

Overall, these visualizations and descriptive statistics helped us to understand the characteristics of each column in the dataset and provided us with insights that will be useful for the next steps in our analysis.

2.2.2. Outliers Detection

The last step for data quality assessment was to identify and analyze any outliers in the data.

- This is an important step in the data exploration process because outliers can significantly impact the results of any statistical analysis. To detect outliers, we used various methods such as the Interquartile Range (IQR) method and Z-score.

Inter-Quartile Range (IQR)

We used IQR method to detect outliers in our dataset. This method involves calculating the difference between the 75th percentile and the 25th percentile of the data and using it to identify values that fall outside of a certain range.

IQR limitations

There were some limitations with IQR using 25th and 75th percentile due to a large amount of data ended detecting a higher percentage of our data points as outliers around 15% of records.

Introduced Solution

To overcome this limitation, we revised the range of IQR and used a wider range 10th and 90th percentile to detect outliers. This allowed us to identify a smaller percentage of data points as outliers, which was more in line with the recommended range. We also validated our results by comparing them to other methods, such as the z-score method, to ensure that our outliers' detection was accurate and reliable. As a result, approximately only 4% of records were defined as outliers.

Z-Score

we utilized the z-score method to detect outliers in our dataset. The z-score is a statistical measure that describes how many standard deviations a data point is from the mean. This method helps us identify values that fall outside of the normal range of values in our dataset, as they may potentially impact the reliability and robustness of our analysis. As a result of the Z-Score, approximately 0.5% of records were defined as outliers.

2.2.3. Correlations Analysis

- The purpose of the correlations analysis is to identify relationships between variables in the dataset. In other words, it helps us understand how different variables are related to each other and how much they change together.

To identify correlations in our dataset, we used various statistical techniques such as Pearson's correlation coefficient Figure 2.4, and Spearman's rank correlation coefficient Figure 2.5. Correlation

coefficients could provide us with a quantitative measure of the strength and direction of the relationship.

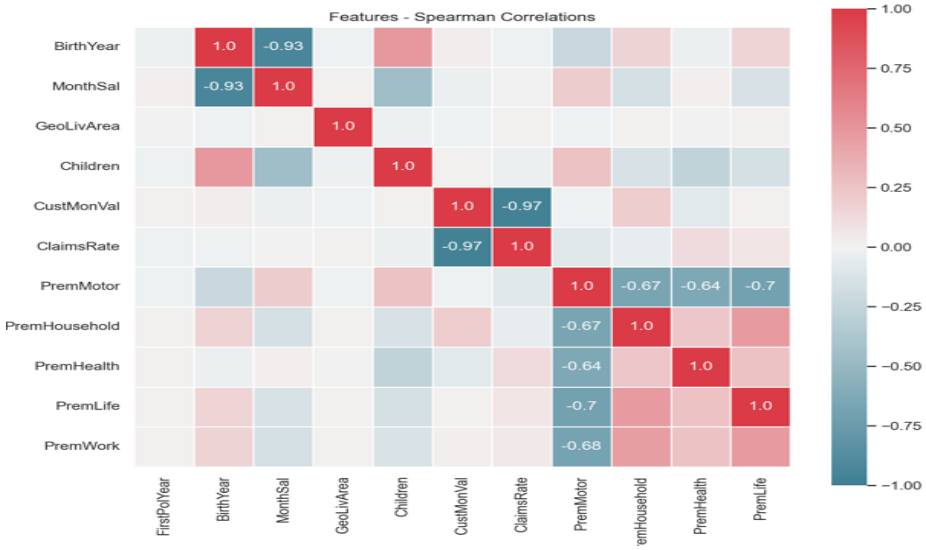

Figure 2.5 - Spearman Correlations

Some key observations and insights:

- In the **Spearman** correlation matrix, some **strong negative correlations** can be seen between certain variables, such as between **BirthYear and MonthSal** (-0.87), **PremHousehold and PremMotor** (-0.79), and **PremLife and PremWork** (-0.75). This suggests that as one of these variables increases, the other tends to decrease.
- In the **Pearson** correlation matrix, similar **strong negative correlations** can be seen between certain variables, such as between **BirthYear and MonthSal** (-0.70), **PremHousehold and PremMotor** (-0.28), and **PremLife and PremWork** (-0.35).
- There is also **a strong positive correlation** between **CustMonVal and ClaimsRate** (0.99 for Spearman and -0.99 for Pearson), which may indicate that as **CustMonVal increases, ClaimsRate also tends to increase**.
- There is a **moderate positive correlation** between **Children and BirthYear** (0.44 for Spearman and 0.44 for Pearson), **which suggests that as the number of Children increases, the BirthYear tends to increase as well**.

In general, it appears that there are several strong correlations between different variables in the dataset, which may be useful to consider when conducting further analysis.

3. Data Pre-Processing

- Data preprocessing is a crucial step that can help to improve the quality and reliability of the analysis by ensuring that the data is accurate, consistent, and relevant. Data preprocessing typically involves a series of steps, including data cleaning, missing value imputation, outlier detection and treatment, and feature scaling. In this report, we will provide an overview of the data preprocessing steps that were taken to prepare the dataset for analysis.

3.1. Data Cleansing

- In this step, we identified and resolved any corrupt, inaccurate, or irrelevant data points in our dataset through data processing.

3.1.1. Data Imputation

In the Data Imputation step, we applied various techniques to handle missing data points in our dataset. First, we identified the columns and rows that contained missing or invalid data Table 3.1.

	column_name	percent_missing
First_Policy	First_Policy	0.291375
BirthYear	BirthYear	0.165113
Education	Education	0.165113
Salary	Salary	0.349650
Area	Area	0.009713
Children	Children	0.203963
CMV	CMV	0.000000
Claims	Claims	0.000000
Motor	Motor	0.330225
Household	Household	0.000000
Health	Health	0.417638
Life	Life	1.010101
Work	Work	0.835276

Table 3.1 – Percentage of Missing Data

Next, we used statistical methods such as median, and mode imputation to estimate the values for these missing data points.

- **Mean Imputation**
  - We decided to handle missing values in **PremMotor, PremHealth, PremLife** and **PremWork** features by replacing with the mean of similar non-missing values in datapoints group by **ClaimsRate**.
- **Median Imputation**
  - We decided to handle missing values in **FirstPolYear**, **BirthYear, GeoLivArea** and **MonthSal** features by replacing with the median of the non-missing values for those features.
- **Mode Imputation**
  - We decided to handle missing values in **EducDeg** feature by replacing with the mode of the non-missing values for those features.
- **Zero Imputation**
  - We decided to handle missing values in **Children** features by replacing with Zero (0) assuming Nan means that they don´t have kids.

We carefully considered the potential impacts of each imputation method on our analysis and chose the method that would have the least impact. Finally, we validated the imputed data points to ensure that they did not have any significant impact on our results.

3.1.2. Coherence checking

- Coherence checking is an important step in the data preprocessing process as it helps to ensure that the data is consistent and logical. In this step, we examined the data for any inconsistencies or errors that may have occurred during data entry or data collection. To do this, we checked for any inconsistencies or discrepancies between different columns within the same row of the dataset.

To ensure the reliability and robustness of our analysis, we performed coherence checks on the data. We identified several inconsistencies that needed to be addressed, including:

1. **Approximately 50%** of the customers had a **First Policy Year that occurred before they turned 18**, which is not allowed according to Portuguese insurance laws.
2. **Approximately 20%** of the customers had a **First Policy Year that occurred before their Birth Year**.
3. **Approximately 20%** of customers **aged 23 or older had an education level of basic or high school**, and **their income more than average income of people their age** which may be an indication of an error or inconsistency in the data.
4. A small percentage of customers had an **incorrect Birth Year**, indicating that they were either **over 110 years old or under 18** years old, which may make them ineligible for certain insurance policies.
5. A small percentage of customers who were **under 18 years old also had children**.

These inconsistencies could potentially impact the reliability and robustness of our analysis if not addressed. We will need to carefully consider how to handle these inconsistencies in the data before proceeding with further analysis.

We decided to take the following steps:

1. For the issue of customers with an incorrect First Policy Year that occurred before they turned 18, we decided to impute these values by **setting the First Policy Year to the year the customer turned 18**.
2. For the issue of customers with an incorrect First Policy Year that occurred before their Birth Year, we decided to impute these values by **swapping the Birth Year and First Policy Year values**.
3. For the issue of customers aged 23 or older with a basic level or high school education, we decided to impute these values **by setting the education level to to align with the customer's age and income**.
4. For the issue of customers with an incorrect Birth Year that indicated they were either over 110 years old or under 18 years old, we decided to **drop these rows** from the dataset as we cannot accurately impute the correct values.
5. We also decided to **drop the rows** for customers who were under 18 years old and had children, as we cannot accurately impute the correct values for these records.

3.1.3. Handling Outliers

- Determining whether a value is an outlier that should be removed or not is very subjective. And while there are certainly valid reasons for throwing away outliers if they are the result of a computer glitch or a human error, eliminating every extreme value is not always a good idea.

As we discussed in section 2.2.2. Outliers Detection we used Interquartile Range (IQR) and Z-score methods to detect the outliers in our dataset. We unionid results from both methods. After reviewing the summary statistics of the results, we identified several outliers that may impact the reliability and robustness of our analysis. **Some possible insights from the data could include the following**:

1. There are a few extremely high or low values for each type of premium. These values may represent unusual or exceptional cases and could potentially be considered outliers.
2. The standard deviations for most of the premiums are relatively large compared to the means, which could indicate that there is a significant amount of variation in the data.
3. The minimum and maximum values for each premium are quite different from the other values, which could also suggest the presence of outliers.
4. The maximum value for the column "MonthSal" is significantly higher than the 75th percentile,
5. The minimum value for the column "ClaimsRate" is significantly lower than the 25th percentile.

These extreme values may be errors or inconsistencies in the data and could skew our results if not addressed properly. Therefore, we decided to drop these outliers from the dataset as a precautionary measure to ensure the accuracy of our analysis.

3.2. Feature Engineering

In this step, we transformed and manipulated the existing features in our dataset to create new ones that may better represent the data and improve the performance of our model.

- This process is a crucial step in the data preprocessing phase and can have a significant impact on model performance. We carefully selected and created new features that we believed would be most relevant and informative for our analysis.

3.2.1. Feature Extraction

In this step, we aimed to identify and extract important features from the raw data that will be useful in building our predictive models. This involves selecting relevant features and engineering new features.

Engineering new features

Creating additional features from the existing data, such as aggregating or combining existing features. Overall, the goal of feature extraction is to identify and extract a subset of features that will provide the most predictive power for our models.

To describe in a better way the customer demographics, insurance policy and to improve data quality, we extracted some features for each customer Table 3.2 .

Attribute	Type	Notes
Age	Numerical	More relevant than birth year in this context because it reflects the current age of the customer. ***= Current Year (2016) – BirthYear***
LoyaltyYears	Numerical	More relevant than first policy year because it reflects the length of time the customer has been with the company. ***= Current Year (2016) - FirstPolYear***
TotalPremium	Numerical	It represents the total amount of premiums paid by a customer. ***= PremLife + PremWork + PremMotor + PremHealth + PremHousehold***
ClaimsAmount	Numerical	The actual claims amount paid by the insurance company. ***= ClaimsRate X TotalPremium***
AnnualProfit	Numerical	The annual profit for the insurance company from each customer. ***= TotalPremium - ClaimsAmount***
AcquisitionCost	Numerical	It refers to the cost of acquiring new customer (One Time) ***= AnnualProfit - CustMonVal***

Table 3.2 – New Features

3.2.2. Categorical Encoding

In this step, we di ordinal encoding for Education Degree we have encoded degrees in a way that reflects the fact that a high school degree is a higher level than basic degree, bachelor's degree is a higher level of education than a high school degree, master's degree is a higher level of education than a bachelor's degree and PhD is the highest between.

- b'1 – Basic (1)
- b'2 - High School (2)
- b'3 - BSc/MSc (3)
- b'4 – PhD (4)

3.2.3. Feature Scaling

As part of our feature engineering process, we applied both normalization and standardization techniques to our dataset. In this step, we applied both normalization and standardization techniques to our data. We used different Scalers: StandardScaler and MinMaxScaler from scikit-learn.

Original Data Distributions as show in Figure 3.1.

Normalization scaled the features of our dataset to a common range, typically between 0 and 1 Figure 3.2. This is often useful when the magnitude of the features in our dataset varies significantly, as it allows us to weigh all features equally in our analysis.

Standardization transformed our features to have a mean of 0 and a standard deviation of 1 Figure 3.3. This is often useful as the features in our dataset have a Gaussian distribution, as it allows us to easily compare the features to one another.

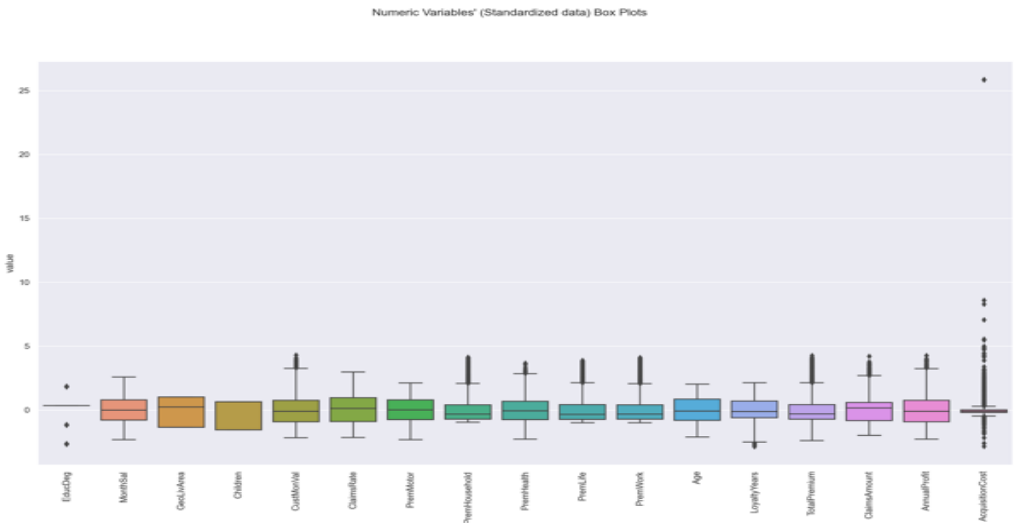

Figure 3.3 - Standardized Data

And after evaluating the results we chose to use Standard Scaler as it offered the most advantages for our data and obtained results were more interesting Figure 3.4. Overall, both normalization and standardization proved to be useful techniques in preprocessing our dataset and allowing us to better analyze and understand the relationships between our features.

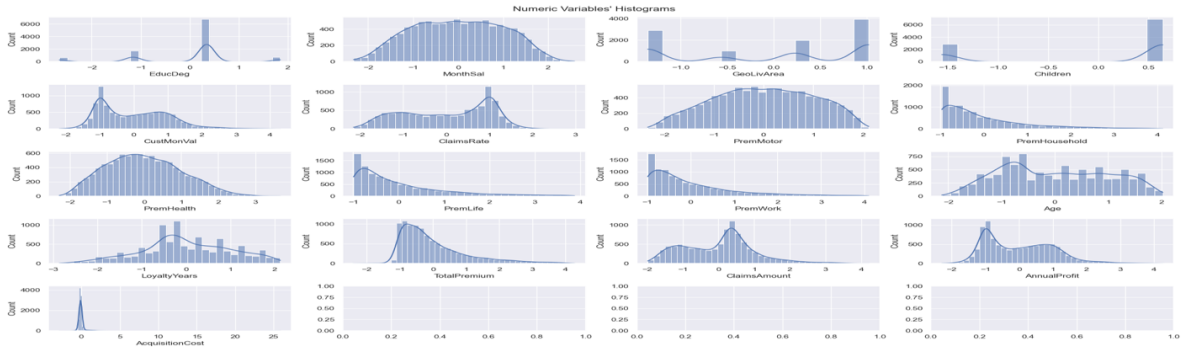

Figure 3.4 - Scaled Data Distributions

4. Dimensionality Reduction

- In the Dimensionality Reduction step, we sought to reduce the number of features in the dataset while retaining as much information as possible. We applied both feature selection and manifold learning techniques to identify the most relevant features and reduce the number of dimensions in our dataset.

4.1. Feature Selection

In this step of the project, we used correlation methods to identify the most important features in our dataset. Specifically, we used the Pearson and Spearman methods to calculate the correlations between different features. Based on these correlations, we were able to select the most relevant features for our model.

4.1.1. Demographics Related

After applying correlation methods to measure the strength and direction of the relationship between Demographics variables Figure 4.1. We could define lowest correlations between variables in EducDeg and GeoLivArea.

We decided to drop these variables with the lowest correlations because they were not contributing significantly and could potentially even be adding noise to the model. By removing these variables, we can improve the interpretability and simplicity of the models.

4.1.2. Insurance Related

After applying correlation methods above to measure the strength and direction of the relationship between insurance related variables Figure 4.2. We noticed some features are strongly correlated with each other that can cause redundancy and may not provide additional information to the model, in such cases, it may be beneficial to remove one of the highly correlated features to avoid overfitting and to improve the interpretability of the model.

We decided finally to drop theses variable LoyaltyYears, PremWork, PremLife, PremHealth, TotalPremium and AcquisitionCost to improve the interpretability and simplicity of our models.

Feature selection process allowed us to build a more accurate and efficient model by focusing on the features that had the greatest impact on the target variable Figure 4.3.

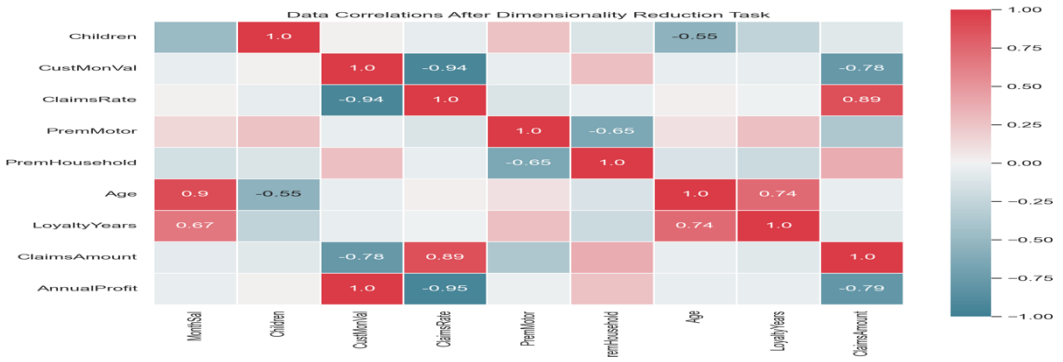

Figure 4.3 – Merged Correlations

4.2. Principal Component Analysis (PCA)

In our project, we used Principal Component Analysis (PCA) as a dimensionality reduction technique to reduce the number of features in our dataset.

- PCA is a statistical method that rotates the data in such a way that the rotated features are statistically uncorrelated. This is useful because it reduces the complexity of the data and makes it easier to analyze and interpret.

To use PCA, we first calculated the covariance matrix of the features in our dataset. The covariance matrix is a measure of how the features vary with respect to each other.

4.2.1. Eigenvectors and Eigenvalues

Next, we calculated the eigenvectors and eigenvalues of the covariance matrix, which are used to determine the principal components of the data.

- The eigenvectors of the covariance matrix are the principal components, and the corresponding eigenvalues represent the magnitude of the variation Figure 4.4.
- **Scree plot**: The "elbow" of the plot above, where the rate of change in the variances decreases sharply give us a guideline for selecting the number of components.
- **Cumulative Explained Variance**: It's the sum of the variances explained by each component up to a certain number of components.

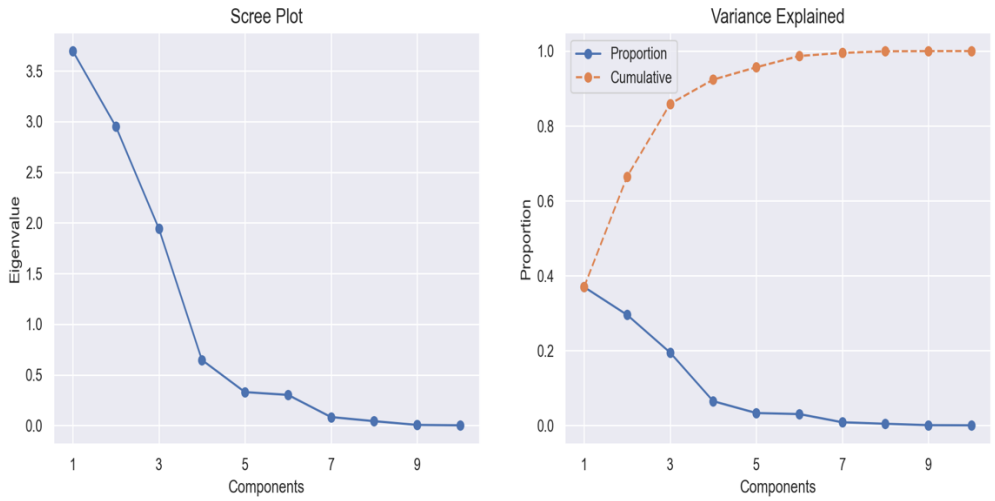

Figure 4.4 - PCA and Variances

Number Of Components

Based on the results of both the scree plot and the cumulative explained variance, we conclude that selecting 5 components would provide the best performance. The scree plot and cumulative explained variance show that these 5 components explain at least 95% of the variance in the data.

4.2.2. Loadings

- The best loadings in principal component analysis (PCA) are the weights or coefficients that represent the importance of each feature in the dataset in constructing the principal components.

In the next step of our analysis, we utilized Principal Component Analysis (PCA) to identify the most important features in our dataset. By training the PCA on a selected number of components (5) from previous step, we were able to determine the best loadings of our dataset which could help us to identify which features contribute the most to the variance in our dataset.

By understanding the best loading of our dataset, we could make informed decisions about which features to focus on or potentially eliminate from our analysis Table 4.1.

4.2.3. Best Components

There are several factors we considered when deciding which components Table 4.2 to keep or discard in PCA.

- Explained variance: Percentage of the total variance in the data that is explained by each component. it is recommended to keep the components that explain a significant portion of the variance in the data, as they capture the most important patterns in the data.
- Loadings: Correlations between the original features and the components. Components with high loadings on a particular feature are likely to capture important patterns in that feature. We used the loadings in previous step to identify which features are most important for each component.
- Interpretability: It is easier to interpret and communicate the meaning of components that are composed of a small number of features with high loadings, rather than components that are composed of many features with low loadings.

	PC0	PC1	PC2	PC3	PC4
MonthSal	-0.056373	-0.931434	-0.015154	-0.065561	0.313639
Children	0.104191	0.661837	-0.523435	-0.497898	0.168806
CustMonVal	0.972795	-0.041928	0.155194	-0.058827	-0.010378
ClaimsRate	-0.987392	0.045878	-0.049943	0.003859	0.002658
PremMotor	0.180801	-0.040441	-0.909017	0.357820	0.027315
Age	-0.060761	-0.963717	-0.019326	-0.063960	0.121369
LoyaltyYears	-0.008847	-0.818002	-0.347296	-0.271226	-0.365294
ClaimsAmount	-0.938540	0.056266	0.141122	-0.099312	-0.023417
AnnualProfit	0.974941	-0.042106	0.149702	-0.056383	-0.010218

Table 4.2 - Principal Components

Attribute	Variance Ratio
ClaimsRate	0.369588
Age	0.295146
PremHousehold	0.194320
Children	0.064642
PremMotor	0.032925

Table 4.1 - Best Columns

Trade-Off

It is worth noting that selecting the number of components is often a trade-off between complexity and performance.

- A larger number of components may capture more variance in the data but may also increase the complexity of the data and may not necessarily improve the performance of a machine learning model.
- On the other hand, a smaller number of components may reduce the complexity of the data but may also limit the ability to capture important patterns and trends in the data.

It is generally a good idea to drop components that have low variance, as they may not contribute much to the analysis and may even cause problems with certain algorithms.

After considering these factors and checking the profile report, we decided to drop components with the lowest variances (PC3, PC4).

5. Clustering

- The process of performing clustering in our project involved selecting an appropriate algorithm, training the model, evaluating its performance, and interpreting the results to gain insights about the data. This allowed us to discover patterns and group similar data points into clusters, which can be useful for various business and marketing purposes.

5.1. Hierarchical clustering

5.1.1. Parameters and distance function

To ensure that we were using the best hyperparameters for this algorithm, we conducted a grid search. This involved testing different combinations of hyperparameters and evaluating their performance using various metrics and techniques.

Out of all the possibilities, we've chosen Single, Complete, Average, and Ward methods and we applied them to different metrics (Euclidean Distance, Cosine Similarity, and Manhattan Distance).

After analyzing the results of the grid search, we were able to determine the optimal combination of hyperparameters: **{'affinity': 'euclidean', 'linkage': 'ward', 'n_clusters': 3}**

5.1.2. Dendrograms analysis

We've run the hierarchical clustering algorithm with different combinations of methods and metrics will list the most relevant one for conciseness:

- The Ward method returned very interesting results with the Euclidean Distance metric: this confirms that it's a good metric with numerical values. The resulting dendrogram is shown in Figure 5.1.

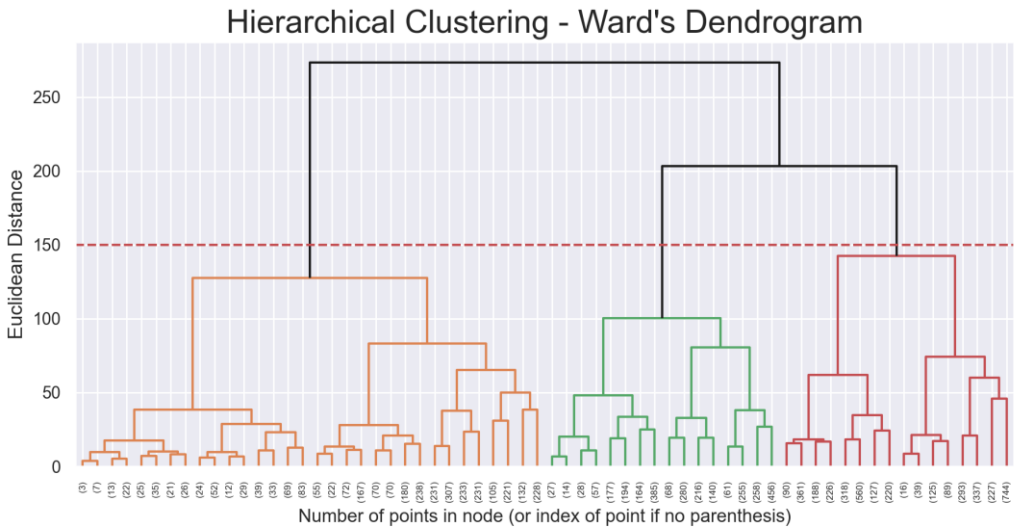

Figure 5.1 - Ward's Dendrogram

5.1.3. Cluster Analysis

The scatter plot Figure 5.2 of the clusters shows three distinct groups of data points. All three groups appear slightly dispersed, but still shows a significant degree of similarity. Overall, the plot suggests that the data can be divided into three distinct clusters based on the similarity of the data points.

5.2. K-Means Clustering

5.2.1. Parameters and distance function

To determine the Optimal Number of Clusters, we used different methods:

- **Elbow method** fitting and plotting the model for a range of values for the number of clusters and selecting the number of clusters where the plot (WCSS) starts to flatten out Figure 5.3.
- **Silhouette method** involves calculating the silhouette score for a range of values for the number of clusters and selecting the number of clusters with the highest average silhouette score Figure 5.4.

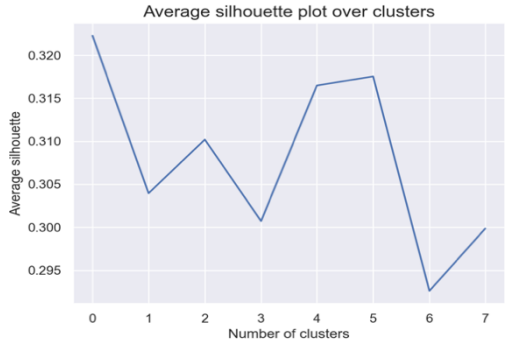

Figure 5.3 - Silhouette Method

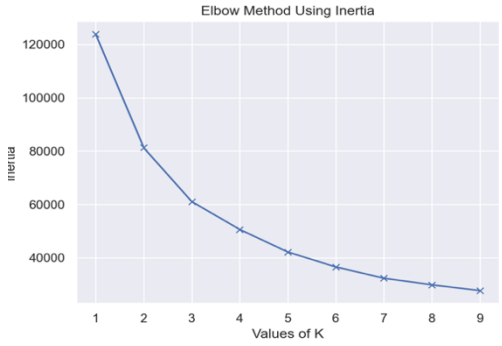

Figure 5.4 - Elbow Method

5.2.2. Clusters Centroids

Centroids can be useful for understanding the characteristics of a cluster, we have calculated clusters centroids of K-means model give results as:

We can notice Squared distance within each cluster Table 5.1.

Parallel coordinates

We've also used Parallel coordinates in order to find patterns: the plot is shown in Figure 5.5.

In our dataset, which has 6 numerical attributes and 3 different classes representing the clusters, we can see that there is a significant variance within the clusters and the data points within the clusters, indicating a less shared behavior across clusters: for this reason, we can safely say that there's a different pattern for each cluster among all of them.

- **Cluster (0)** appears to have the highest claim rate and the lowest Customer Monthly Value (CMV). Additionally, this cluster has the oldest customers among all the clusters.

- These characteristics may be important to consider when determining the risk level and potential profitability of this cluster.

- **Cluster (1)** appears to have the highest customer monthly value (CMV) and the lowest claim rate, with an average age and average premiums.
  - This suggests that this cluster may consist of customers who have a high value to the insurance company, as they are generating a high level of revenue through their premiums and have a low number of claims.
  - This could be a valuable group for the company to focus on retaining and potentially targeting for additional products or services. On the other hand, if the company has a high number of claims from this group, it may be worth further investigation to understand the reasons for the high claims rate and address any potential issues.
- **Cluster (2)** appears to have a higher claims rate. This may be due to a variety of factors such as the age of the customers in this cluster or the type of insurance policies they have.
  - It is also worth noting that the customers in this cluster have the lowest motor premiums and are the youngest in age, indicating that they may not have as much experience with driving or may not own as many vehicles.
  - Finally, it seems that this cluster has the highest percentage of customers with children, which may also contribute to their higher claims rate. Further analysis and investigation may be necessary to determine the specific reasons for the patterns observed in this cluster.

5.2.3. Cluster Analysis

The scatter plot Figure 5.6 of the clusters shows three distinct groups of data points. The first group appears to be tightly clustered together, indicating a high degree of similarity among these data points. The second group is slightly more dispersed, but still shows a significant degree of similarity. The third group is the most dispersed, indicating a lower degree of similarity among these data points. Overall, the plot suggests that the data can be divided into three distinct clusters based on the similarity of the data points.

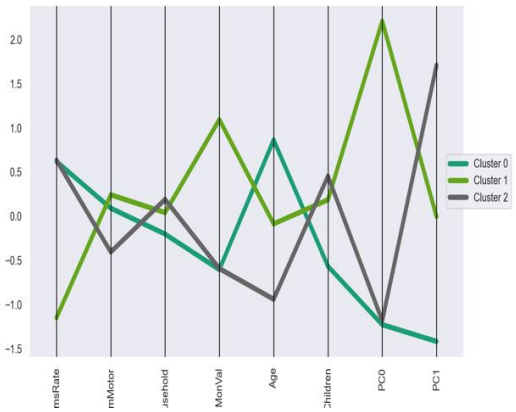

Figure 5.5 - Parallel coordinates

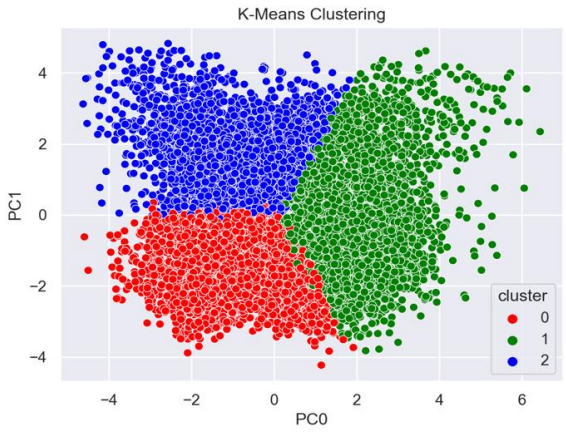

Figure 5.6 – K-means Scatter Plot

5.3. DBScan (Density-Based Clustering)

5.3.1. Parameters and distance function

DBSCAN uses these two parameters:

- Eps ε (epsilon): the maximum distance between two samples for them to be considered as in the same neighborhood.
- MinPts: the number of samples in a neighborhood for a point to be considered as a core point. This includes the point itself.

Since there's no automatic way to decide the value of these parameters, we did some research and found that a standard setup is to choose min samples as twice the dimensionality of the dataset. After analyzing the results of the grid search, we were able to determine the optimal combination of hyperparameters: {'eps': 1.0, 'min_samples': 16}

5.3.2. Cluster Analysis

The scatter plot Figure 5.7 of the clusters shows 5 distinct groups of data points. The first group (-1) is detected as Outliers (Noise), while the other groups are more dispersed, indicating a lower degree of similarity among their data points.

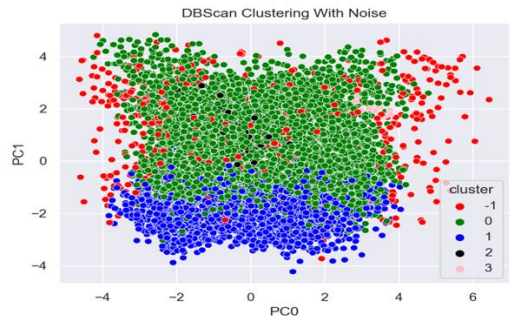

Figure 5.7 – DBSCAN-Noise Scatter Plot

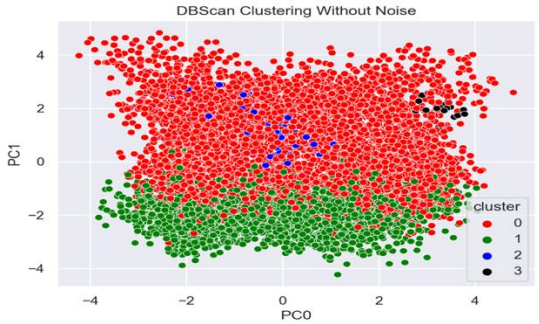

Figure 5.8 - DBSCAN Scatter Plot

5.4. Gaussian Mixture Model (Distribution-based clustering)

5.4.1. Parameters and distance function

The following are some of the parameters that can be specified when using GMM:

- n_components: The optimal number of clusters can be determined using techniques such as the elbow method or silhouette score.
- covariance_type: The type of covariance matrix used to model the data. The options include "spherical", "diagonal", "tied", and "full".
- init_params: The method used to initialize the cluster means. The options include "kmeans" and "random".

After analyzing the results of the grid search, we were able to determine the optimal combination of hyperparameters: {'covariance_type': 'full','init_params': 'kmeans','n_components': 2}

5.4.2. Cluster Analysis

The scatter plot Figure 5.9 of the clusters shows 2 distinct groups of data points. Both has slightly dispersed, but still shows a significant degree of similarity. Overall, the plot suggests that the data can be divided into 2 distinct clusters based on the similarity of the data points.

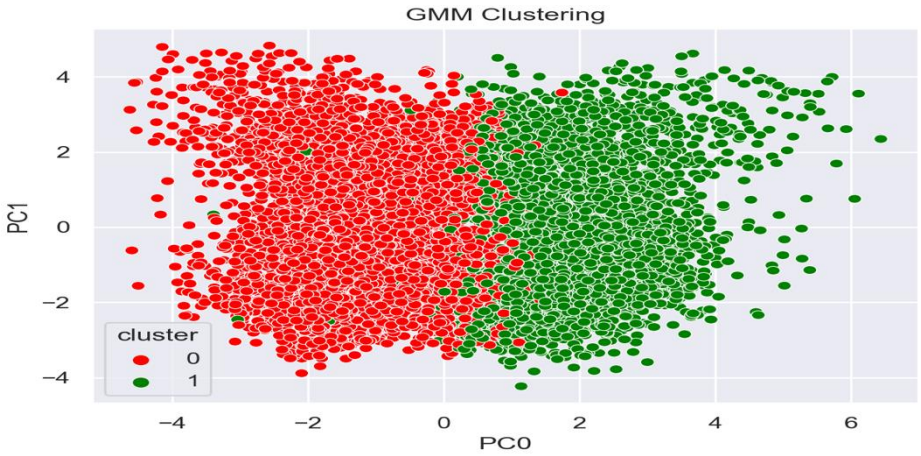

Figure 5.9 - GMM Scatter Plot

5.5. Evaluations Of Clustering Approaches

To conclude the clustering section, we'll now compare the different clustering algorithms and the obtained clusters.

There are several ways to evaluate the results of clustering models:

- **Silhouette Score** Measure of how similar an object is to its own cluster compared to other clusters. It ranges from -1 to 1, with a higher score indicating a better fit.
- **R2 Score** Measure of proportion of the variance in the target cluster that is explained by the model. The score ranges from 0 to 1, with a higher score indicating better model performance.

▪ Algorithm	R^2 Score	Silhouette Score
Hierarchical	0.4684	0.2785
K-Means	0.5068	0.3039
DBSCAN	0.2529	0.1525
GMM	0.3307	0.1393

Table 6.1 – Model Evaluations

K-Means

produced a good distribution for our data. As we said in Section 2.1.2, we've obtained 3 clusters that can identify 3 different real-life customer profiles.

DBSCAN,

no matter what the values of the initial parameters, we couldn't obtain a good clustering because it always produced a huge cluster with most of the points and another negligible one.

Hierarchical approach

produced very unbalanced clustering with some linkage methods (e.g., Single, Ward). Instead, when using other combinations (like Complete Method with Euclidean Distance), we've obtained something like the results of the K-Means approach that can reflect the previously discussed customer profiles.

So, in summary, K-Means and Hierarchical approaches returned good clustering's, but we consider more interesting the K-Means one.

6. Cluster Interpreting and Profiling

- Interpreting the results of a clustering analysis involves examining the characteristics of the clusters that were identified and determining what they mean in the context of the problem being solved. This can involve looking at the size, shape, and distribution of the clusters, as well as the characteristics of the data points within each cluster.

6.1. Clustering Using Different Perspectives

As A2Z would like to Understand the value and demographics of each customer segment, as well as Understand which types of insurance they will be more interested in buying. We've splitted data set into customer demographics data and insurance related data.

- Demographic Features: Age, EducDeg, MonthSal, GeoLivArea and Children.
- Coverage Features: LoyaltyYears, CustMonVal, ClaimsRate, PremMotor, PremHousehold, PremHealth, PremLife, PremWork, TotalPremium, ClaimsAmount, 'AnnualProfit and AcquisitionCost.

6.1.1. Evaluate Perspectives

In this step we evaluated each perspective using R^2 and Silhouette Metrics.

Demographics evaluations score Figure 6.1 and Coverage evaluations scores Figure 6.2

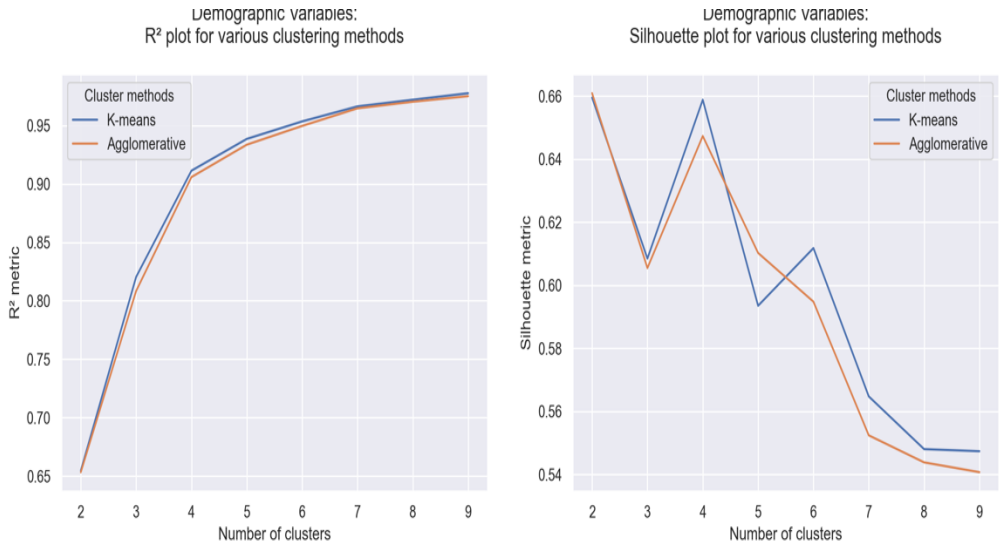

Figure 6.2 Coverage

6.1.2. Clusters Distributions and Statistics

To Understand each cluster characteristics and Distributions, we plotted Demographics data distributions between its clusters Figure 6.3 and same for Coverage clusters Figure 6.4.

6.1.3. Insights Into Customers Demographics!

- **The first cluster (demographic cluster 0)** has a relatively young average age, lower education level, and lower monthly salary compared to the other clusters. This suggests that the

customers in this cluster may be less financially secure or may have fewer resources available to them.

- **The second cluster (demographic cluster 1)** has a much older average age and higher education level compared to the other clusters. This suggests that the customers in this cluster may be more financially secure and may have more resources available to them.
- **The third cluster (demographic cluster 2)** has an average age and education level that is intermediate between the first and second clusters. This suggests that the customers in this cluster may be more financially secure than those in the first cluster, but less financially secure than those in the second cluster.

6.1.4. Insights Into Customers Coverage!

- **The first cluster (coverage cluster 0)** This cluster has relatively high premiums for motor, household, health, and life coverage, and relatively high claims amount and acquisition cost. This suggests that the customers in this cluster may be more likely to make claims and may be more expensive for the insurance company to cover.
- **The second cluster (coverage cluster 1)** This cluster has relatively high premiums for motor and work coverage, and relatively low premiums for household, health, and life coverage, as well as relatively low claims amount and acquisition cost. This suggests that the customers in this cluster may be less likely to make claims and may be less expensive for the insurance company to cover.
- **The third cluster (coverage cluster 2)** This cluster also has relatively high premiums for motor and work coverage, and relatively low premiums for household, health, and life coverage, as well as relatively high claims amount and acquisition cost. This suggests that the customers in this cluster may be more likely to make claims and may be more expensive for the insurance company to cover, like the customers in the first cluster.

It would be useful to further investigate these characteristics and behaviors to understand the differences between the clusters and how they may impact the business. For example,

1. understanding why customers in cluster 1 have higher customer monetary values and lower claims rates may help the business target similar customers or identify strategies to increase customer monetary values and reduce claims rates for other customers. Similarly,
2. Understanding why customers in cluster 2 have higher premiums for health and life insurance, as well as higher total premiums and claims amounts, may help the business tailor its insurance offerings or identify opportunities to upsell these customers on additional insurance products.

6.2. Merge Perspectives using Hierarchical Clustering

In our clustering analysis step, upon further investigation, we determined that these two distinct perspectives that emerged from our data were closely related and could be merged into a single cluster. This decision was based on the similarity of the characteristics and behaviors within the two perspectives, as well as the potential benefits of combining them for further analysis and decision-making.

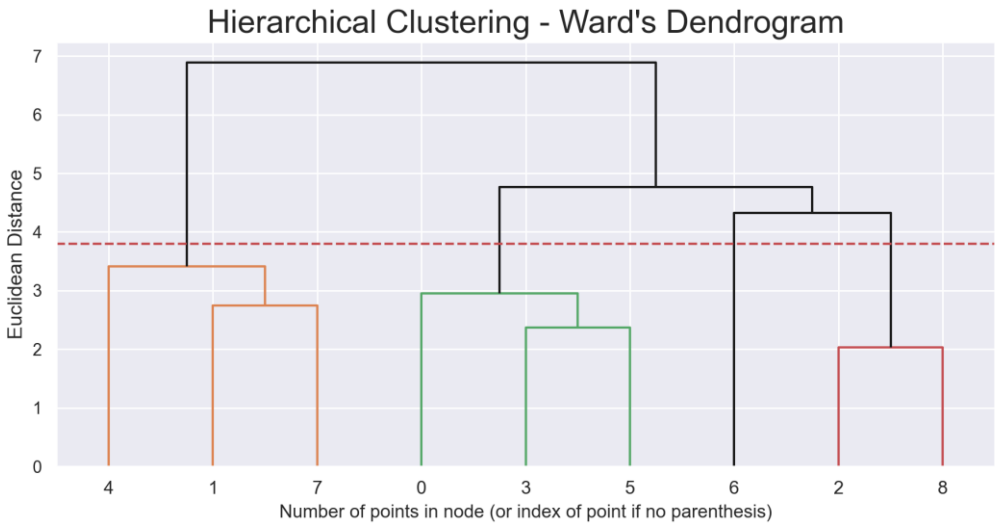

Figure 6.5 - Perspectives Dendrogram

6.3. Cluster Visualization Using t-SNE

Figure 6.5 The t-SNE plot shows 4 distinct clusters of data points. Cluster 0 has the highest Annual Profit and lowest Claims Amount, while cluster 3 has the highest Claims Amount and lowest Annual Profit. Cluster 1 has a higher Acquisition Cost and lower Total Premium compared to the other clusters, while cluster 2 has a lower Acquisition Cost and higher Total Premium. The Age, Education Level, and Monthly Salary of the data points in each cluster also vary. There is some overlap between the clusters, indicating a lower degree of similarity among their data points.

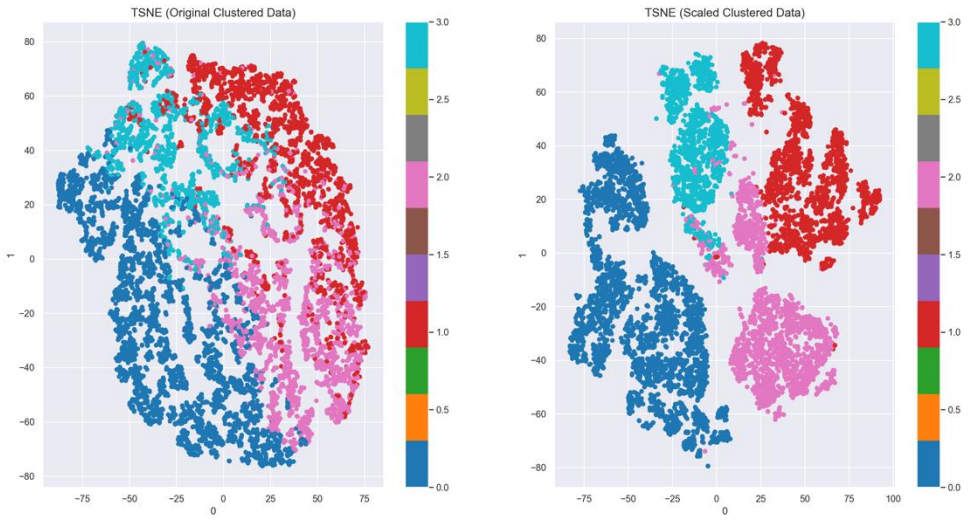

Figure 6.5 - T-SNE

7. Predictive Analysis

7.1. Decision Tree

In this step of the analysis, we used decision tree classification to evaluate and predict the clusters for each group. This allowed us to use the insights gained from the clustering step to better understand the characteristics of each group and make more accurate predictions about their behavior. Overall, this helped us to further refine our understanding of the data and gain additional insights about the relationships between different variables in the dataset.

- It is estimated that in average, we could predict 97.81% of the customers correctly.

Model Evaluation

Checking classification report and confusion matrix Figure 7.1 our prediction had f1 score 98%.

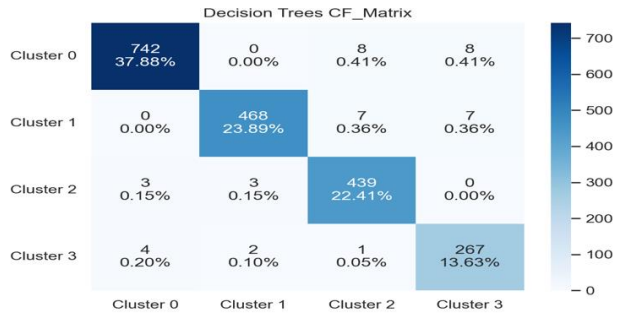

Figure 7.1 – Confusion Matrix

7.2. Assessing Features Importance

Figure 7.2 We can see that; AnnualProfit (Engineered Features) is an important factor on deciding customer cluster. Also, PremMotor, Children, Age, CMV and ClaimsRate seem to have influence on customers clusters. On the other side, rest of factors seem to have less importance for the customers clustering.

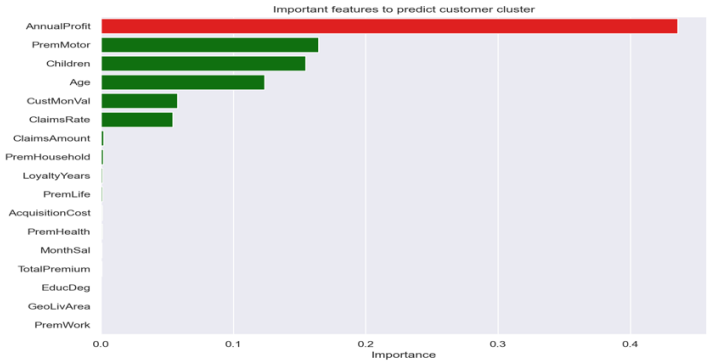

Figure 7.2 – Features Importance

8. Conclusion

Business

Based on the customer segmentation and clustering analysis, we have identified several key patterns and trends in the data.

The first cluster has a relatively young average age, lower education level, and lower monthly salary, and has relatively high premiums for motor, household, health, and life coverage, as well as relatively high claims amount and acquisition cost. This suggests that the customers in this cluster may be more financially insecure and may be more likely to make claims, which could be an area of concern for the insurance company.

On the other hand, the second cluster has a much older average age and higher education level, as well as relatively high premiums for motor and work coverage, and relatively low premiums for household, health, and life coverage, as well as relatively low claims amount and acquisition cost. This suggests that the customers in this cluster may be more financially secure and less likely to make claims, which could be a valuable group for the company to focus on retaining and potentially targeting for additional products or services.

The third cluster has an average age and education level that is intermediate between the first and second clusters and has relatively high premiums for motor and work coverage, as well as relatively low premiums for household, health, and life coverage, and relatively high claims amount and acquisition cost. This suggests that the customers in this cluster may be more likely to make claims and may be more expensive for the company to cover, like the customers in the first cluster.

Marketing

Based on the customer segmentation and clustering analysis, we have identified several key patterns and trends in the data that could be useful for targeted marketing efforts. The first cluster has a relatively young average age and lower education level, and may be more financially insecure, which could inform marketing strategies aimed at attracting and retaining this group of customers.

On the other hand, the second cluster has a much older average age and higher education level, and may be more financially secure, which could inform marketing strategies aimed at upselling or cross-selling to this group.

The third cluster has an average age and education level that is intermediate between the first and second clusters and may be more likely to make claims and more expensive for the company to cover, like the customers in the first cluster. Therefore, marketing efforts for this group may need to focus on mitigating risk and ensuring customer satisfaction to reduce the likelihood of claims.

Overall, understanding the characteristics and behavior of each customer cluster can help the marketing team to tailor their efforts and optimize their efforts to reach the right customers with the right message.

9. References

Arumawadu, H. I., Rathnayaka, R. M., & Illangarathne, S. K. (2015). Mining Profitability of Telecommunication Customers Using K-Means Clustering. Journal of Data Analysis and Information Processing, 3, 63-71.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding Groups in Data: An Introduction to Cluster Analysis. 116-118.

Sukup, J. (2018). When K-Means Clustering Fails: Alternatives for Segmenting Noisy Data. Retrieved from DataScience.com

Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. Prentice Hall, New Jersey (USA).

Cross, G. & Thompson, W. (2008). Understanding your customer: Segmentation Techniques for Gaining Customer Insight and Predicting Risk in the Telecom Industry.

Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering: Algorithms and applications. CRC Press.

Campello, R. J. G. B., Moulavi, D., Zimek, A., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In Proceedings of the ACM SIGMOD International Conference on Management of Data (pp. 593-604). ACM.

van der Maaten, L., & Hinton, G. (2012). Visualizing high-dimensional data using t-SNE. Journal of Machine Learning Research, 9, 2579-2605.

10. Appendix

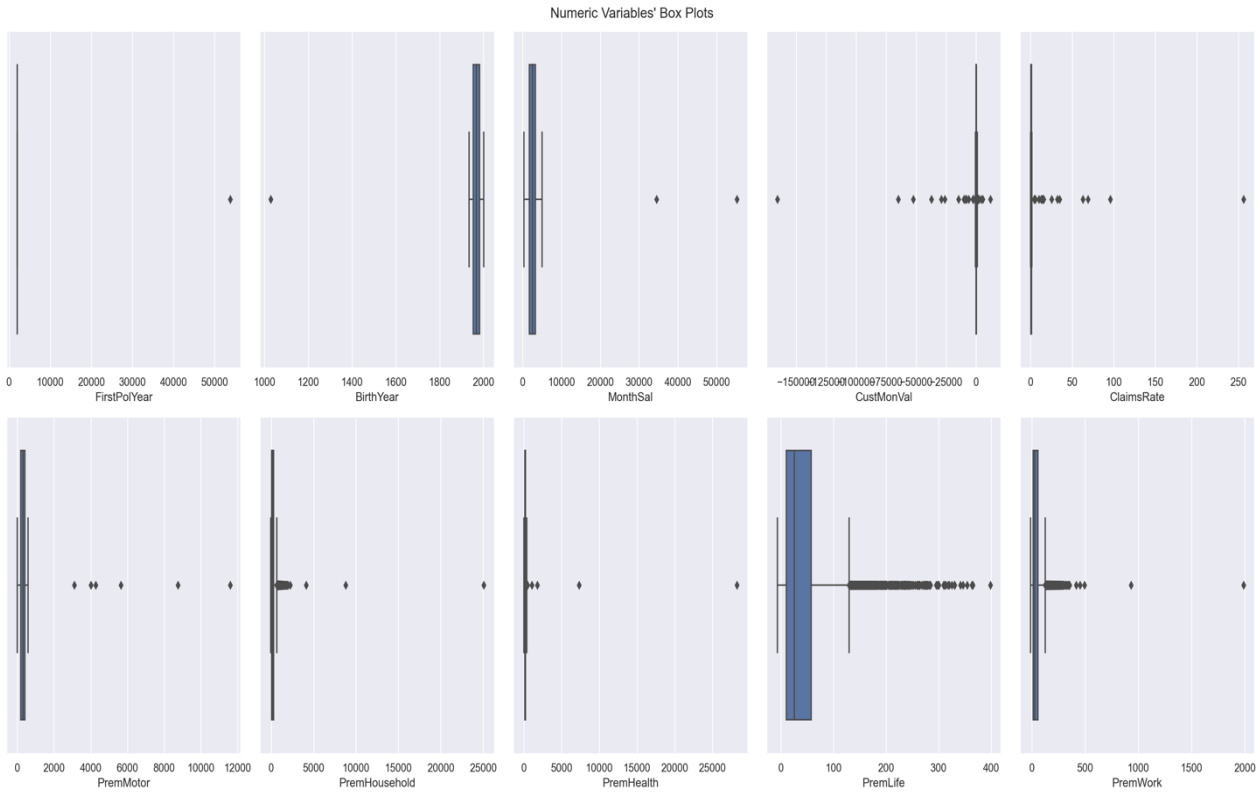

Figure 2.1 – Metric Features (Boxplots)

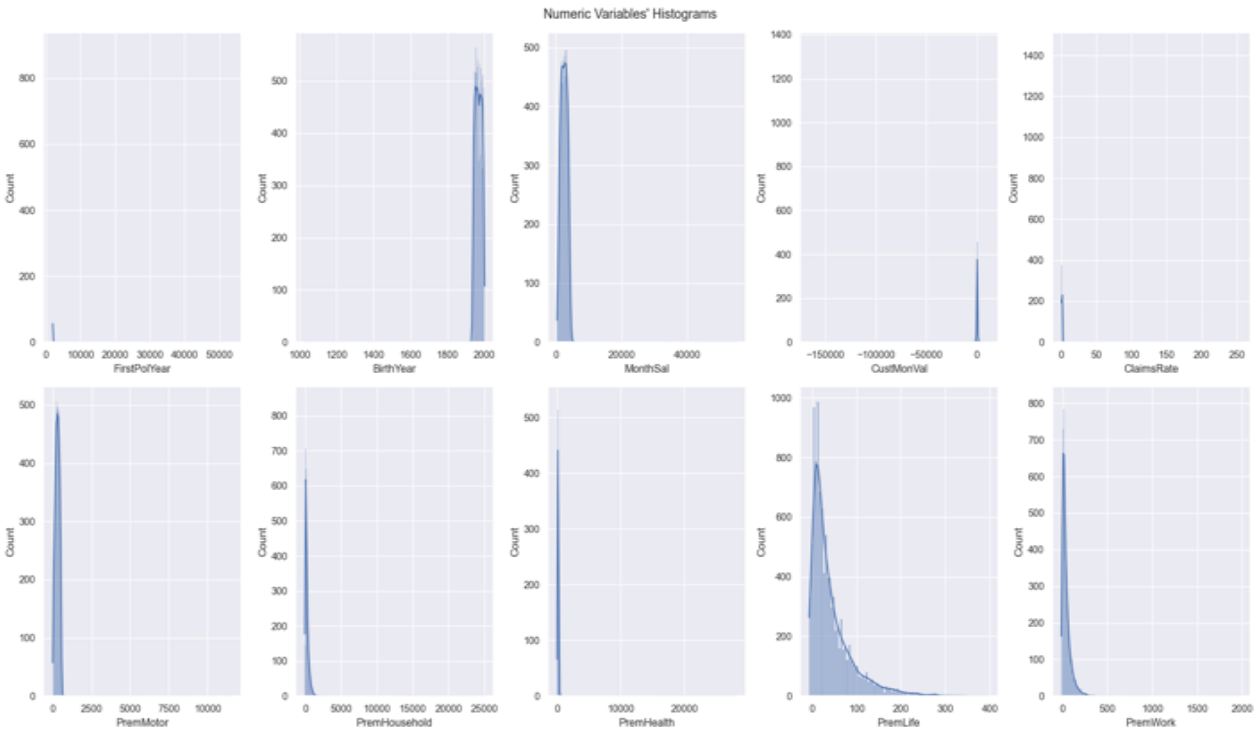

Figure 2.2 – Metric Features (Histograms)

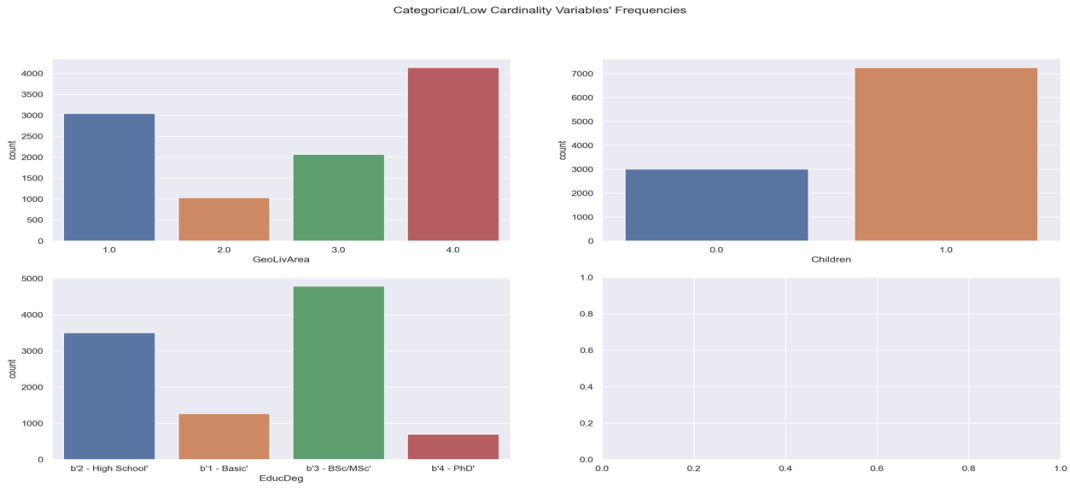

Figure 2.3 – Non-Metric Features (Histograms).

Attribute	Type	Description	Additional Information
CustID	Numerical	Unique numeric ID assigned to the customer.	
FirstPolYear	Numerical	Year of the customer's first policy	May be considered as the first year as a customer
BirthYear	Numerical	Customer's Birthday Year	The current year of the database is 2016
EducDeg	Categorical	Academic Degree	
MonSal	Numerical	Gross monthly salary (€)	
GeoLivArea	Categorical	Living area	No further information provided about the meaning of the area codes
Children	Numerical	Binary variable (Y=1)	
CustMonVal	Numerical	Customer Monetary Value	Lifetime value = (annual profit from the customer) X (number of years that they are a customer) - (acquisition cost)
ClaimsRate	Numerical	Claims Rate	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremMotor	Numerical	Premiums (€) in LOB: Motor	Annual Premiums (2016). Negative premiums may manifest reversals occurred in the current year, paid in previous one(s)
PremHousehold	Numerical	Premiums (€) in LOB: Household	
PremHealth	Numerical	Premiums (€) in LOB: Health	
PremLife	Numerical	Premiums (€) in LOB: Life	
PremWork	Numerical	Premiums (€) in LOB: Work Compensations	

Table 2.1 Semantic Data

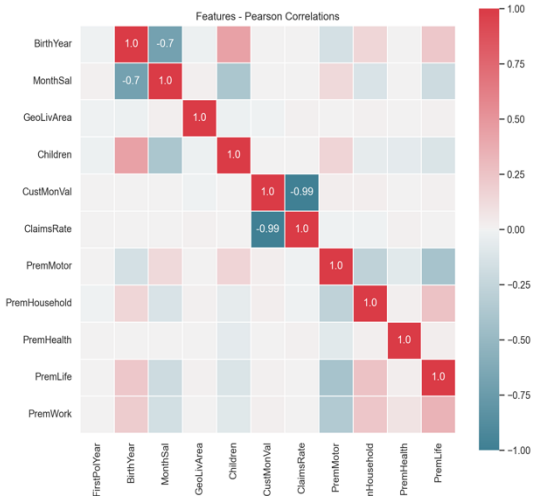

Figure 2.4 - Pearson Correlations

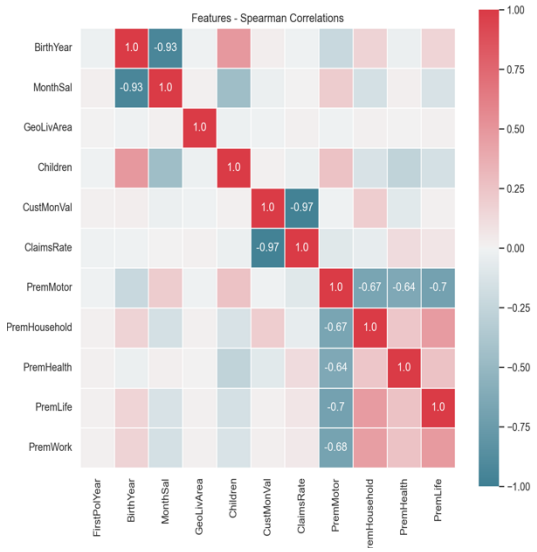

Figure 2.5 - Spearman Correlations

Attribute	Type	Notes
Age	Numerical	More relevant than birth year in this context because it reflects the current age of the customer. ***= Current Year (2016) – BirthYear***
LoyaltyYears	Numerical	More relevant than first policy year because it reflects the length of time the customer has been with the company. ***= Current Year (2016) - FirstPolYear***
TotalPremium	Numerical	It represents the total amount of premiums paid by a customer. ***= PremLife + PremWork + PremMotor + PremHealth + PremHousehold***
ClaimsAmount	Numerical	The actual claims amount paid by the insurance company. ***= ClaimsRate X TotalPremium***
AnnualProfit	Numerical	The annual profit for the insurance company from each customer. ***= TotalPremium - ClaimsAmount***
AcquisitionCost	Numerical	It refers to the cost of acquiring new customer (One Time) ***= AnnualProfit - CustMonVal***

Table 3.2 – New Features

	column_name	percent_missing
First_Policy	First_Policy	0.291375
BirthYear	BirthYear	0.165113
Education	Education	0.165113
Salary	Salary	0.349650
Area	Area	0.009713
Children	Children	0.203963
CMV	CMV	0.000000
Claims	Claims	0.000000
Motor	Motor	0.330225
Household	Household	0.000000
Health	Health	0.417638
Life	Life	1.010101
Work	Work	0.835276

Table 3.1 – Percentage of Missing Data

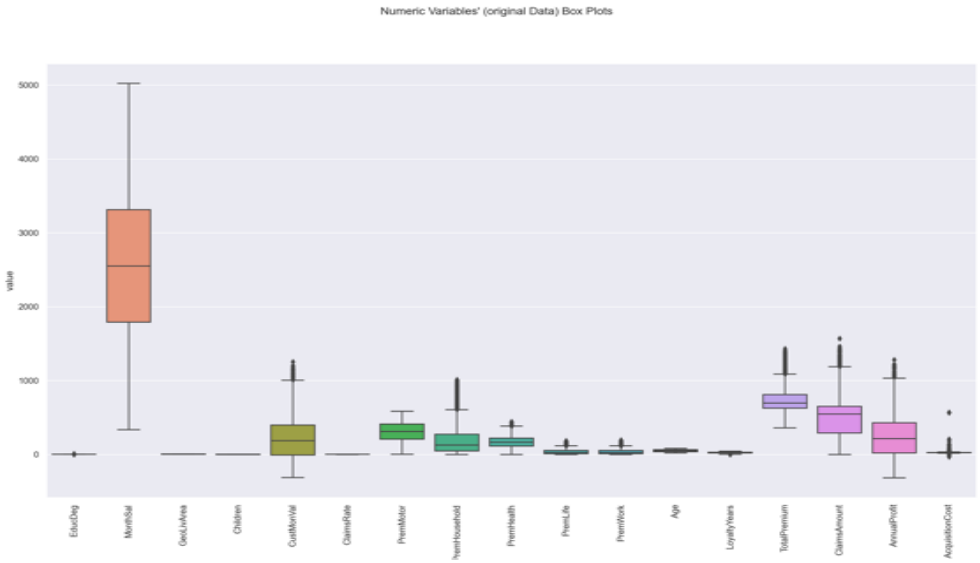

Figure 3.1 Original Data

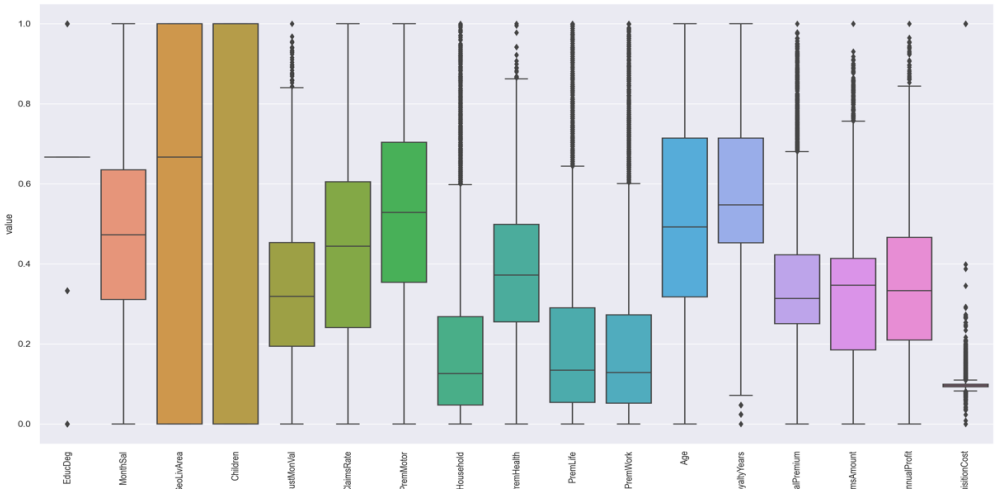

Figure 1.2 - Normalized Data

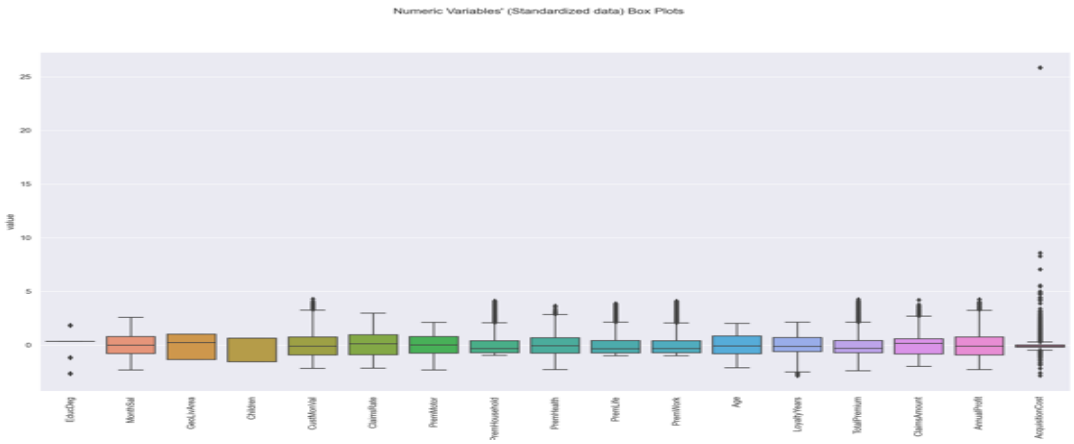

Figure 3.3 - Standardized Data

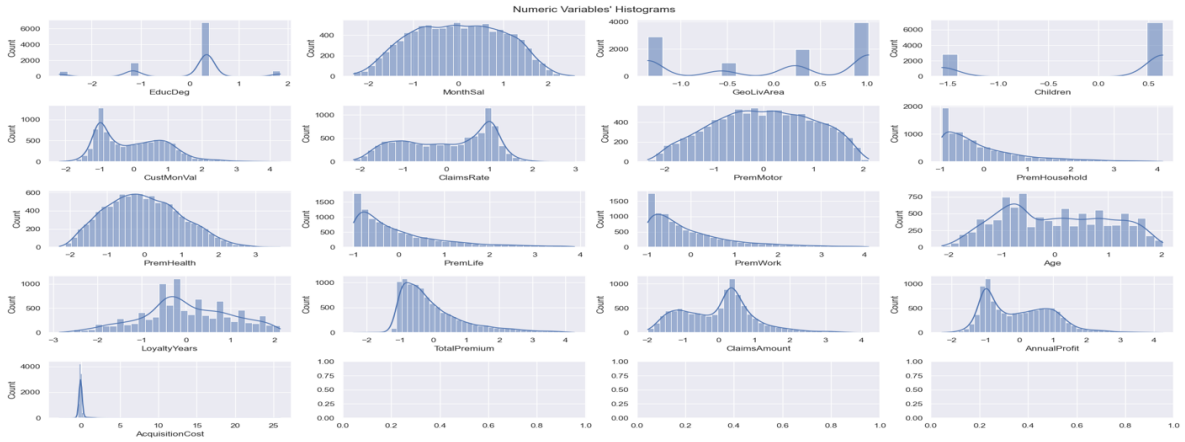

Figure 3.4 - Scaled Data Distributions

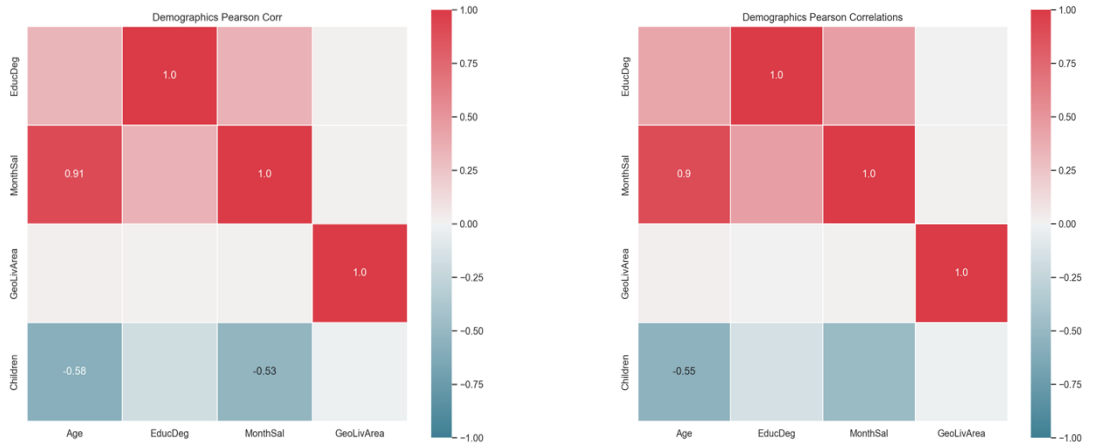

Figure 4.1 - Demographics Correlations

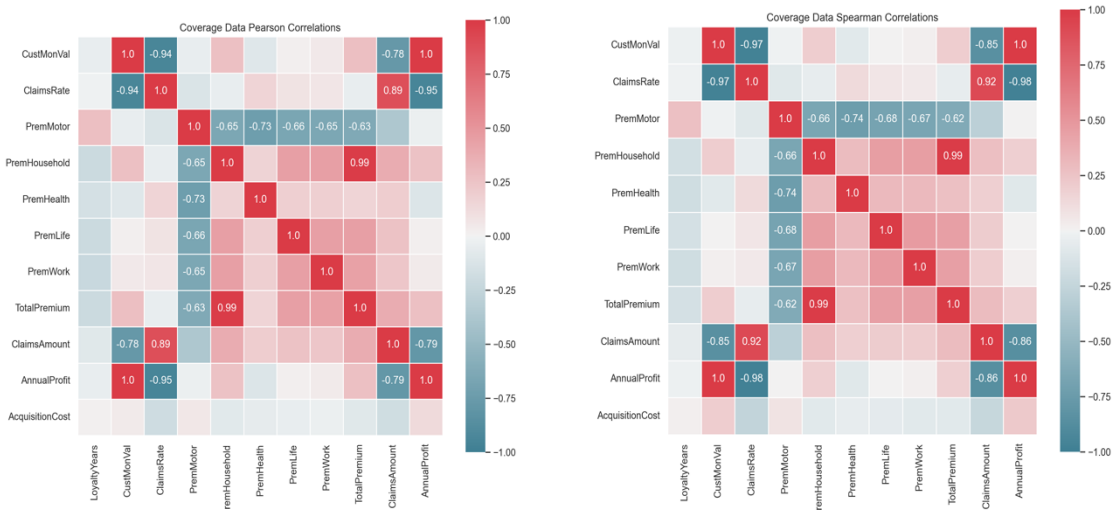

Figure 4.2 – Insurance Correlations

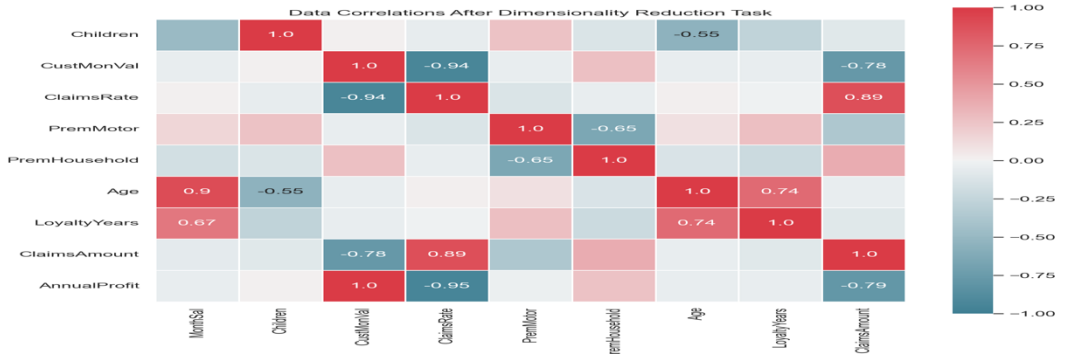

Figure 4.3 – Merged Correlations

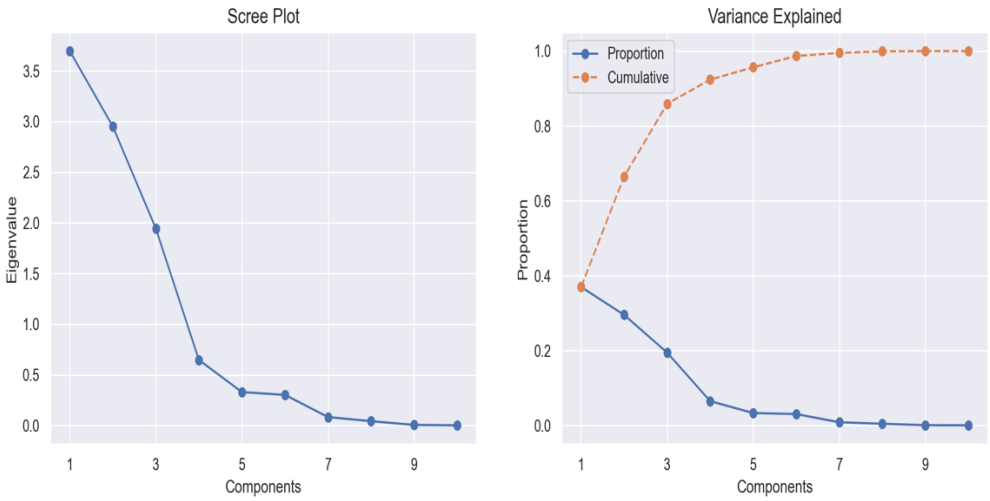

Figure 4.4 - PCA and Variances

	PC0	PC1	PC2	PC3	PC4
MonthSal	-0.056373	-0.931434	-0.015154	-0.065561	0.313639
Children	0.104191	0.661837	-0.523435	-0.497898	0.168806
CustMonVal	0.972795	-0.041928	0.155194	-0.058827	-0.010378
ClaimsRate	-0.987392	0.045878	-0.049943	0.003859	0.002658
PremMotor	0.180801	-0.040441	-0.909017	0.357820	0.027315
Age	-0.060761	-0.963717	-0.019326	-0.063960	0.121369
LoyaltyYears	-0.008847	-0.818002	-0.347296	-0.271226	-0.365294
ClaimsAmount	-0.938540	0.056266	0.141122	-0.099312	-0.023417
AnnualProfit	0.974941	-0.042106	0.149702	-0.056383	-0.010218

Table 4.2 - Principal Components

Attribute	Variance Ratio
ClaimsRate	0.369588
Age	0.295146
PremHousehold	0.194320
Children	0.064642
PremMotor	0.032925

Table 4.1 - Best Columns

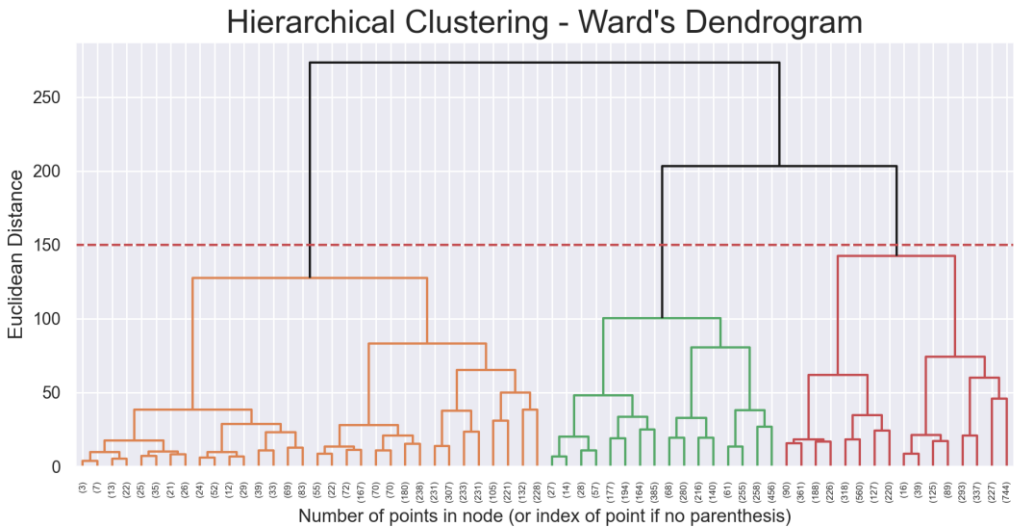

Figure 5.1 - Ward's Dendrogram

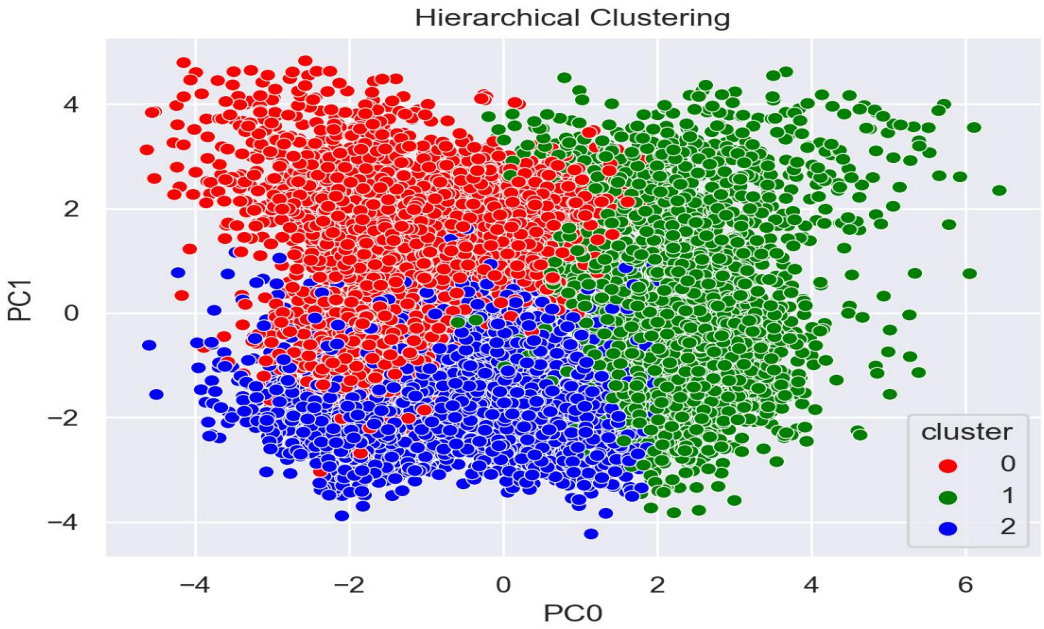

Figure 5.2 - Hierarchical Scatter Plot

Figure 5.3 - Silhouette Method

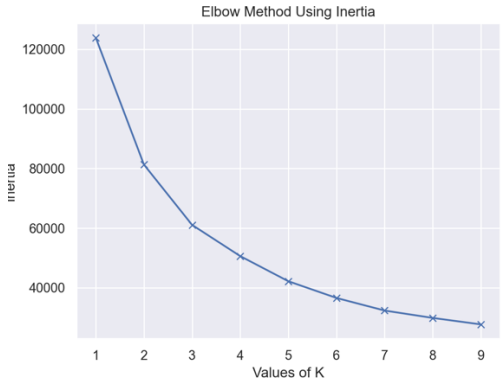

Figure 5.4 - Elbow Method

Cluster	# Members	Squared Distance
Cluster 0	3457	19341.48
Cluster 1	3461	25905.55
Cluster 2	2877	15839.88

Table 5.1 - K-Means Centroids

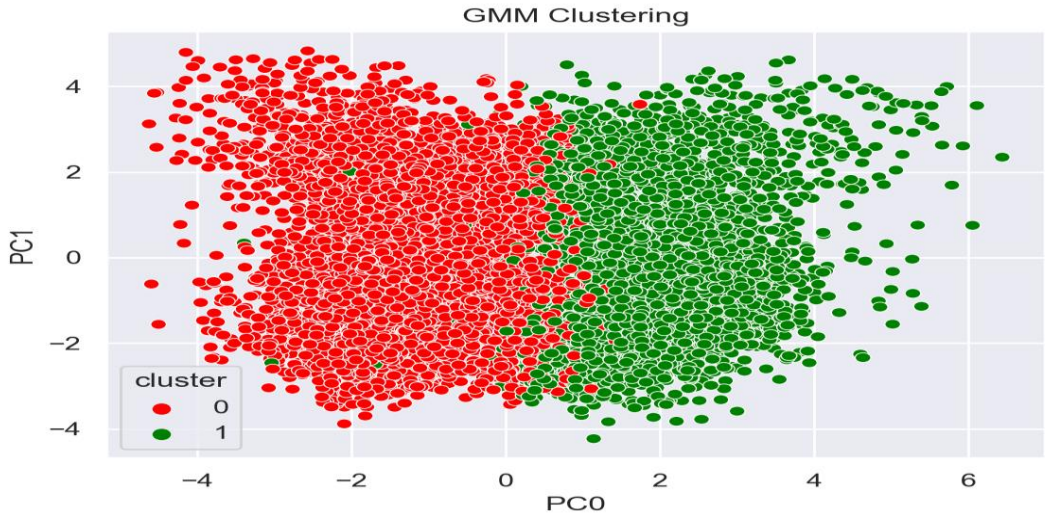

Figure 5.9 - GMM Scatter Plot

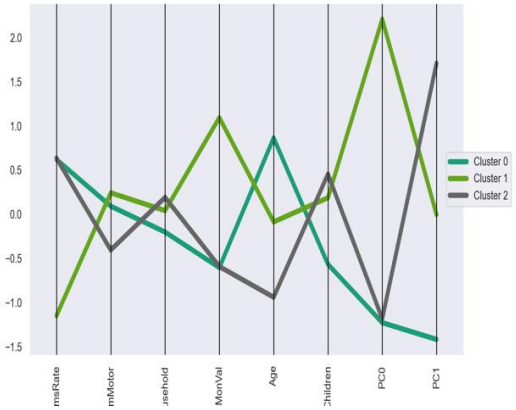

Figure 5.5 - Parallel coordinates

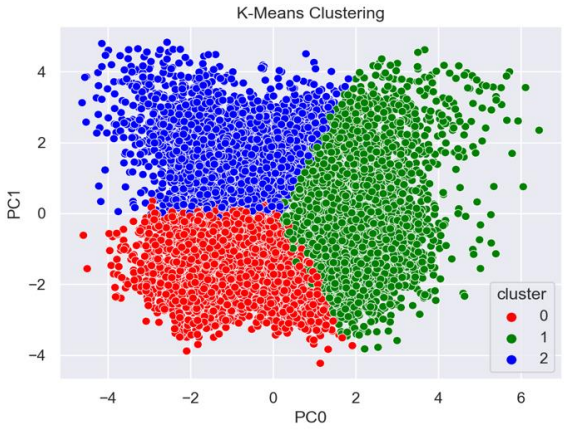

Figure 5.6 - Kmeans Scatter Plot

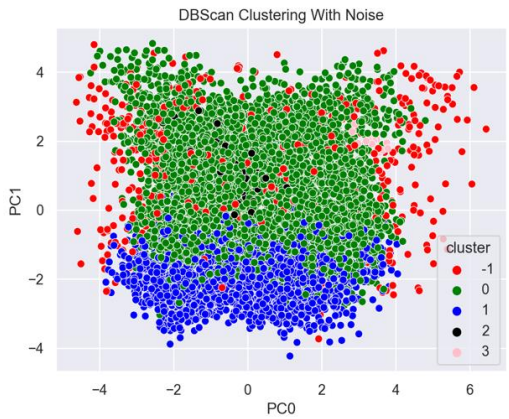

Figure 5.7 - DBSCAN Scatter Plot

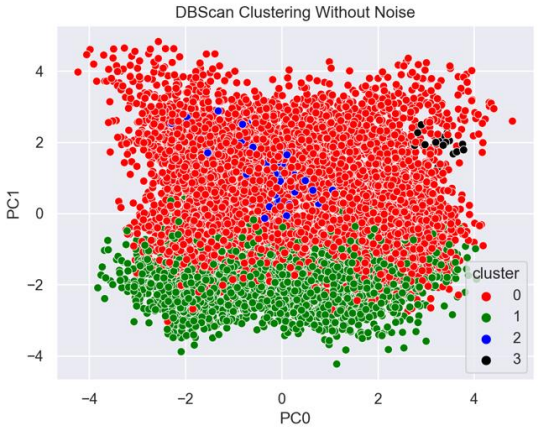

Figure 5.8 - DBSCAN Scatter Plot

Algorithm	R^2 Score	Silhouette Score
Hierarchical	0.4684	0.2785
K-Means	0.5068	0.3039
DBSCAN	0.2529	0.1525
GMM	0.3307	0.1393

Table 6.1 – Model Evaluations

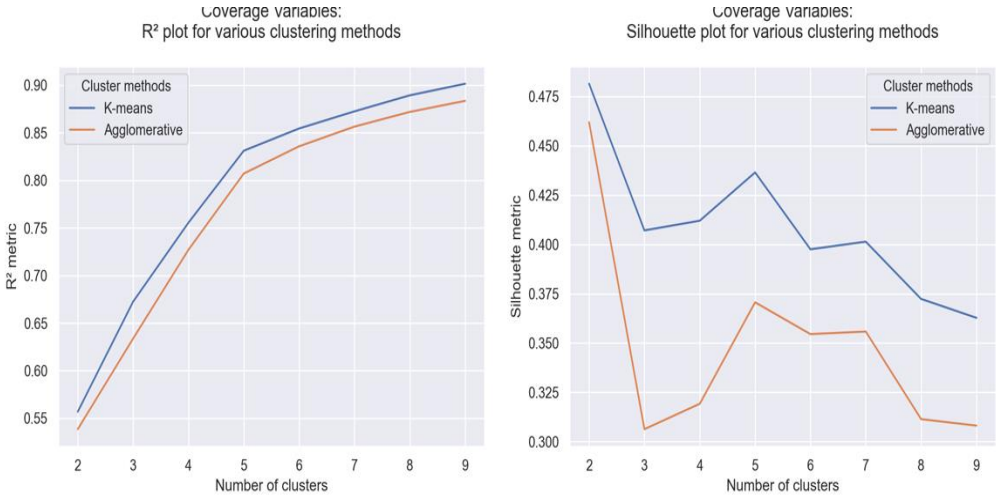

Figure 6.1 - Demographics

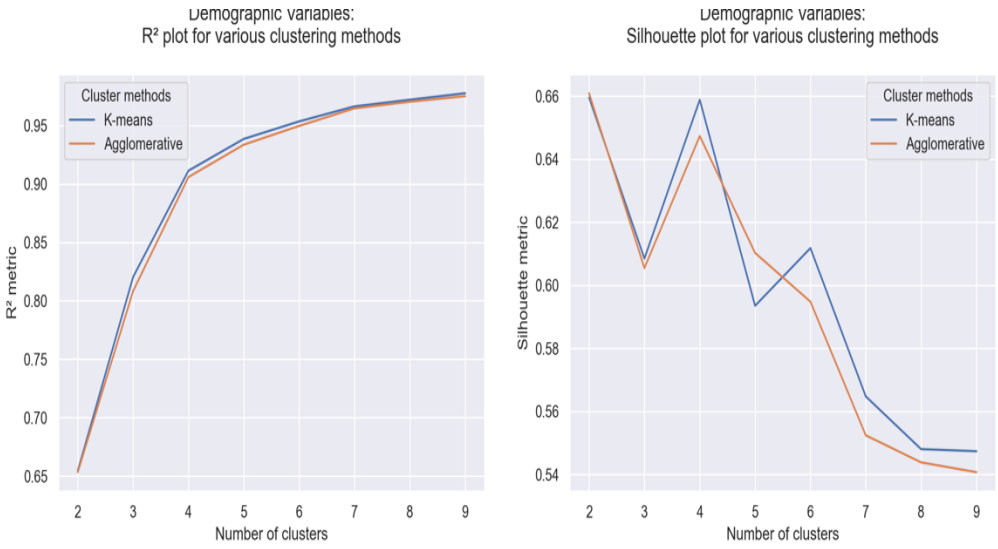

Figure 6.2 Coverage

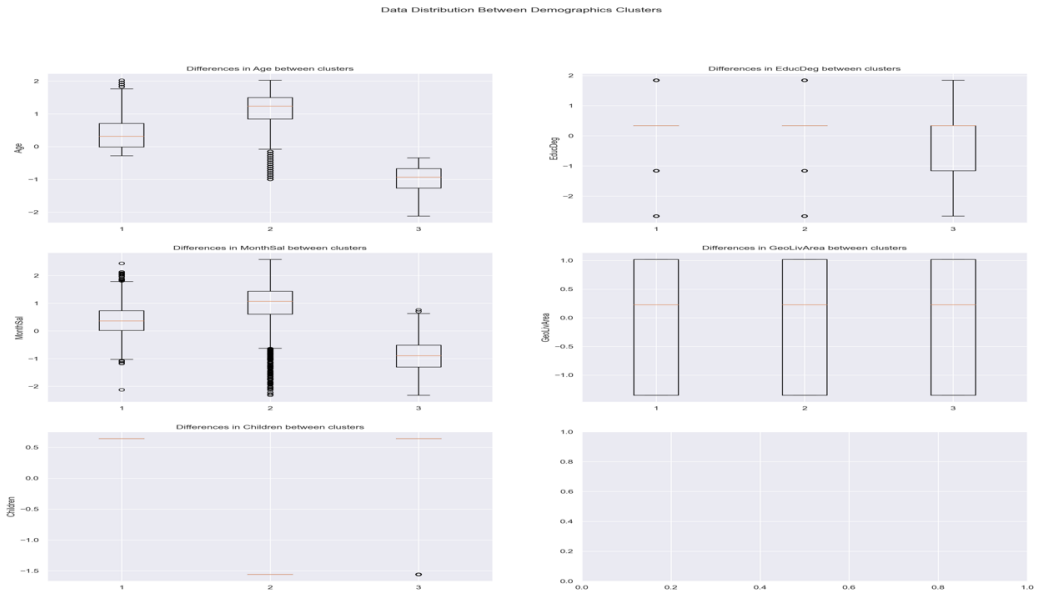

Figure 6.3 Demographics Clusters

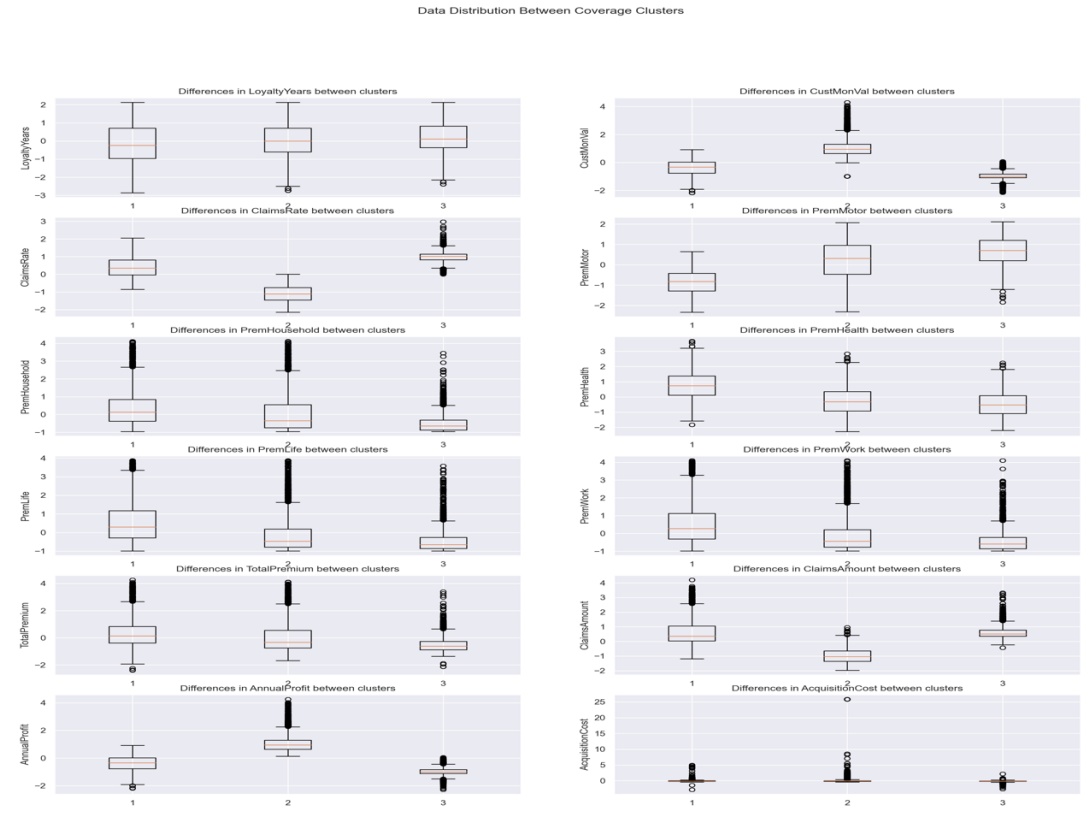

Figure 6.4 Coverage Clusters

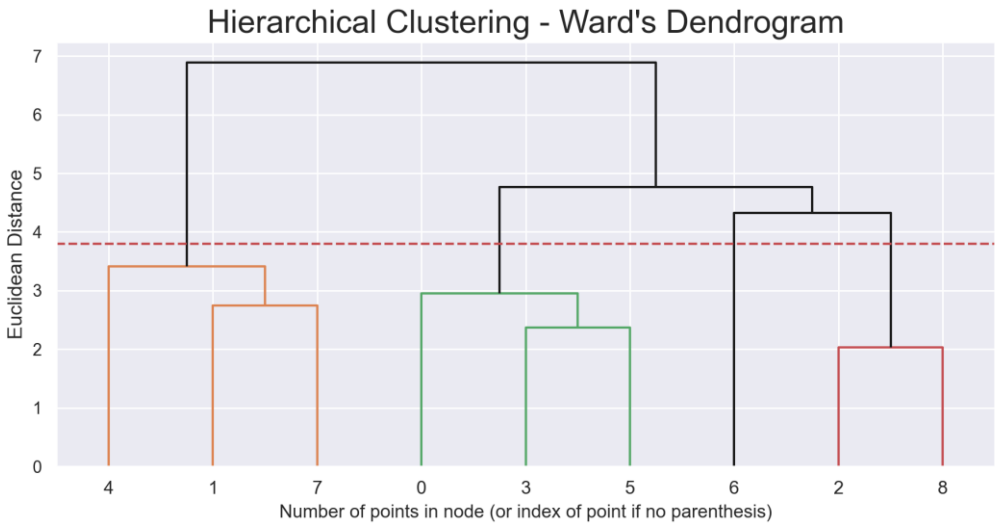

Figure 6.5 - Perspectives Dendrogram

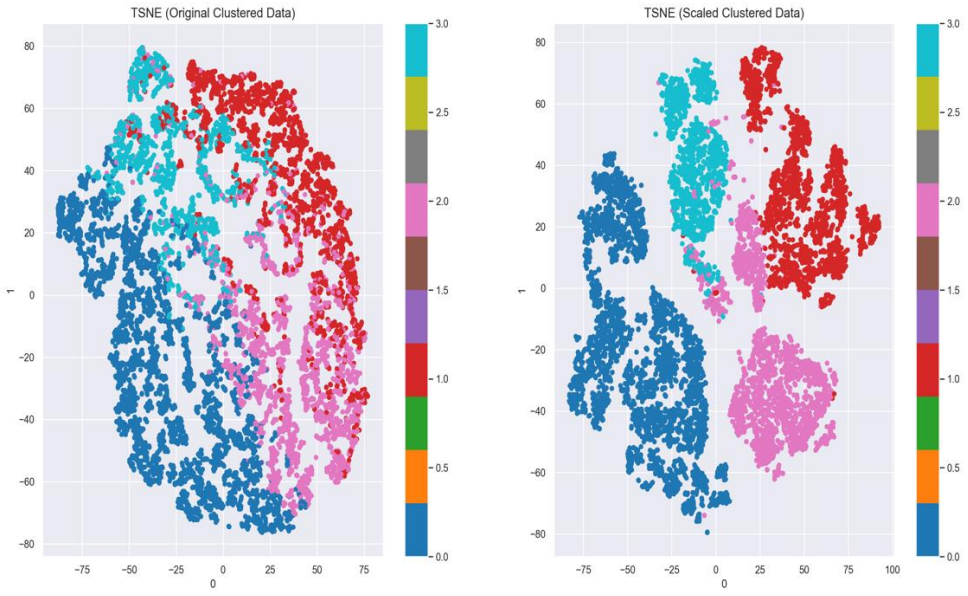

Figure 6.6 - T-SNE

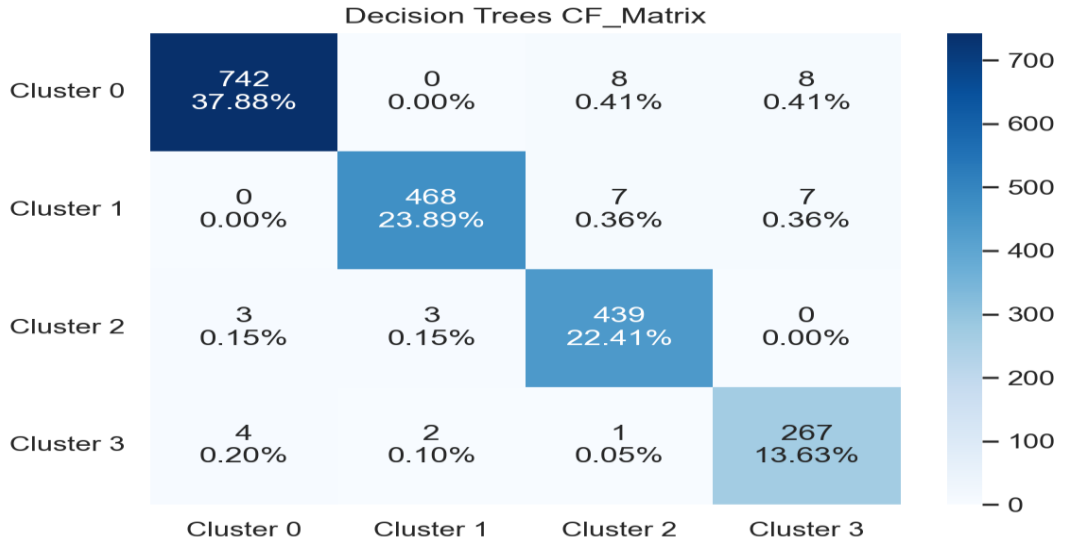

Figure 7.1 – Confusion Matrix

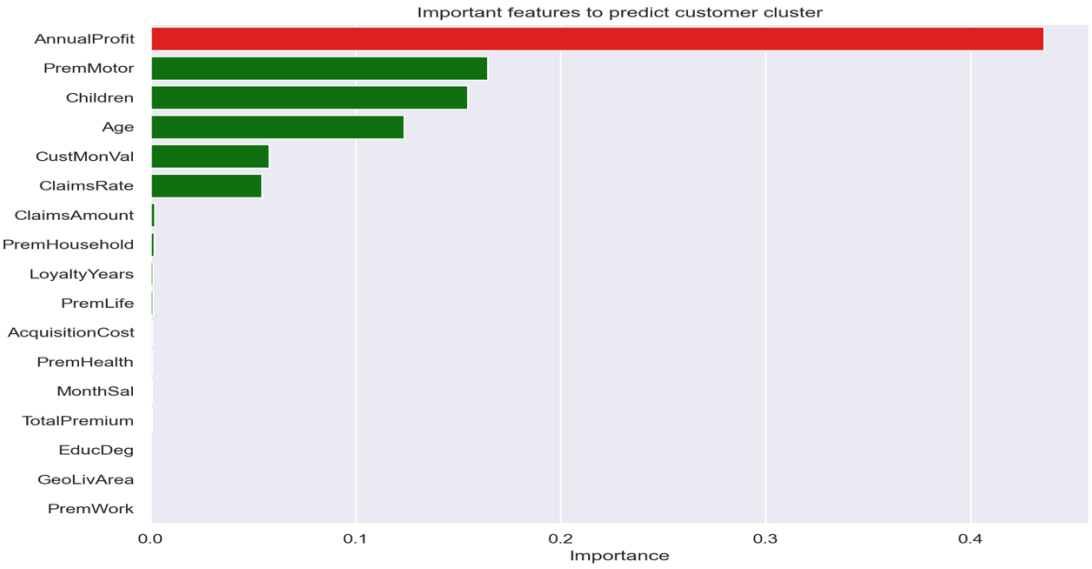

Figure 7.2 – Features Importance