Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance – Insurance Company

Group EU

Diogo Morgado, number: 20221392

Tomás Gama, number: 20221354

December, 2022

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

INDEX

1. Introduction

Acquiring new customers is essential for the success of any business. One way to do this is by gathering information from current customers and using it to understand the needs and preferences of different market segments. This includes analyzing factors such as geographic location, age, personality, and purchasing habits. By dividing the market into smaller groups based on these characteristics, companies can tailor their marketing strategies and make informed decisions about product development, pricing, and targeted advertising. By understanding and effectively targeting specific customer segments, businesses can improve their ability to meet the needs of their current customers and attract new ones.

A2Z Insurance is a reputable Portuguese insurance company that offers a range of services, including motor, household, health, life, and work compensation insurance. A2Z primarily serves customers in Portugal, but a significant number of new customers also come from the company's website. Customers have the option to sign up for A2Z services through branches, by phone, or online.

To effectively understand and target specific customer segments, we used various approaches and perspectives to segment the customers and analyzed the results. A2Z can benefit from gaining insight into the value and demographics of each segment, as well as determining which types of insurance are most appealing to them. This can help the company better serve its customers and make informed decisions about marketing and product development.

2. Data Exploration

We start by exploring the dataset and performing some initial exploration tasks. First, we rename the columns to make them easier to understand. Then, we retrieve information from all columns and check for any null values. We also verify that the data types are correct, and we notice that the "BirthYear" and "First_Policy" columns have float data types, which we need to change to integer because they represent years. During this process, we also observe that the dataset may contain outliers and that several columns have missing values (Table 1). To address the missing values, we consider various imputation methods and decide to use Mode imputation and Median Imputation, as the percentage of missing data is not too high to justify the use of other methods.

2.1. Handling Missing Values

We decided to approach the fact of having some missing values, in diverse ways for each case. In the variables “First_Policy”, “Birth_Year”, “Salary” and “Area” we decided to fill the missing values with the median, between the data we have for each variable. For this we used the function median(), which splits the higher half of the data or probability distribution from the lower half. For the variable “Children”, we assumed that the Nan means that they don´t have kids, so we filled with the zero value. Finally, for the variable “Education” we filled the missing values using the mode, with the method mode(), that provides with the values that appear most often. Concluding, we do a last check to see if we missed some value, and we transform the variables “First_Policy”, “Birth_Year”, “Salary” and “Area” to Integer type.

2.2. Handling Outliers

First, we split the variables in metric and non-metric features, now we take a visualization of the non-numeric and numeric variables before the outlier removal (Figure 1,2,3).

Next, we opted for using two different methods of removing the outliers, manually and using the IQR method. Using the manual method, we defined by ourselves, with graphic assistance and interpretation, to remove the values which we thought would be right to remove, this method can vary from session to session, because it depends on the interpretation and the view, may be different from person to person, using this method we kept 77% of the data.

Using the IQR method we managed to keep 85% of the data. This method consists of defining an upper and lower limit of the quantile removing the values that are out of the range (Figure 4,5,6).

3. Data Preprocessing

After exploring the dataset, we have a better understanding of the variables and how we can use them for data clustering. To enhance the performance of the clustering, we perform feature engineering to create new variables that might give us an advantage. We also conduct a coherence check to ensure that the data in our dataset is consistent and make sense, and we remove any outliers that we find during this process.

3.1. Feature Engineering

During the feature engineering process, process of creating and selecting features that can improve the performance of machine learning models**,** we created several new variables: "Age", "Customer_Years", "Total_Premium", and "Salary_Rate". To obtain the "Age" variable, we subtracted the year of the database from the "Birth_Year". We calculated the "Customer_Years" variable by subtracting the current year of the database from the year of the "First_Policy". We transformed Education into Ordinal Encoding, meaning 4 would be corresponding to PhD and 1 to Basic Education, and all negative values of premiums into 0, because it meant the customer had already left the insurance or did already pay, and created the "Total_Premium" variable as the sum of all the customer's premiums. Finally, we created the "Salary_Rate" variable by dividing the "Total_Premium" by the customer's annual salary (calculated by multiplying their salary by 14) and multiplying the result by 100, which gives us the rate of salary they invest in the company's insurance.

3.2. Coherence checking

When performing coherence checking, we looked for values in the dataset that did not make sense. We began by analyzing the difference between the "Age" and "Customer_Years" variables to ensure that a customer cannot be older than they have been a customer of the insurance company. We also checked for underage individuals with children, individuals with a PhD but not the minimum age to obtain one, and individuals with ages below the minimum required for a BSc/MSc. As these values could potentially influence the results, we removed them from the dataset.

4. Data Partition

In this step we do a Data Partition, consisting of splitting in Demographic and Insurance Data. The Demographic Data contains “Birth_Year”, “Age”, “Education”, “Salary”, “Area”, “Children”, basically the information more focused on the client’s information and Insurance Data the information related about the insurance company such as “First_Policy”, “CMV”, “Claims”, “Motor”, “Household”, “Life”, “Health”, “Work”, “Customer_Years”, “Total_Premium”, and “Salary_Rate”.

4.1. Feature Selection

For feature Selection we plotted two correlation matrixes, each one for Demographic (Figure 7) and Insurance Data (Figure 8). After interpreting both the graphics and the correlation between each variable we decided to drop “Birth_Year”, “Age” from Demographic Data and “Total_Premium”, “Claims” from the Insurance Data, due to high correlation.

4.2. Data Standardization/Normalization

In this procedure we first try out three different Scalers before choosing the one we used: the Standard Scaler (works well with outlier detection, the features need to be the same type and normalizing your data will scale most of your data to a small interval if you have outliers in your feature), the MinMax Scaler (preserves the shape of the original distribution, the importance of outlier values doesn´t affect, so those can be used for outlier detection algorithms) and the Robust Scaler (doesn´t work well for outlier detection and reduces the effect of the outliers).

After evaluating all the three methods we opted to go with Standard Scaler, because using it would bring more advantages.

5. Dimensionality reduction

Despite previously performing feature selection, we decided to use dimensionality reduction to further simplify the dataset while preserving as much information as possible. The high dimensionality of the data can make it difficult to process and visualize, and it can also increase the risk of overfitting. Therefore, we applied principal component analysis (PCA) to our insurance dataframe to improve the output and performance of machine learning algorithms. From the results of PCA (Figure 9), we selected 4 components and applied PCA again with those components. We then interpreted the values of each principal component and decided to drop "PC2" because it did not add additional value to the dataframe.

In conclusion, using principal component analysis (PCA) for data clustering can be an effective way to reduce the dimensionality of the data and improve the performance of clustering algorithms. By transforming the data into a new set of linearly uncorrelated variables called principal components, PCA allows us to capture the most important information in the data and use it to cluster the observations into groups.

6. Data Clustering

Data clustering is a technique used to group similar observations together into clusters, based on their characteristics and patterns. In the context of an insurance company, data clustering can be used to understand and analyse the characteristics of different groups of policyholders, identify trends and patterns in their behaviours and characteristics, and make informed decisions about how to best serve their needs.

By applying data clustering to insurance data, companies can better develop targeted marketing and underwriting strategies. Clustering can also be used to identify subgroups within the policyholder population that may have different needs or preferences, and tailor insurance products and services to meet those needs.

6.1. K-means

First, we plotted the inertia over the number of clusters to determine the optimal number of clusters for the dataset (Figure 10). Inertia is a measure of the sum of the squared distances between the data points and the centroid of their respective clusters. As the number of clusters increases, the inertia decreases, indicating that the clusters are becoming more compact. However, at some point, the decrease in inertia may become diminishing, and increasing the number of clusters beyond this point may not provide any additional benefit. By examining the plot of inertia over the number of clusters, we can identify the "elbow" in the curve, which represents the point at which the decrease in inertia becomes diminishing and may be a good choice for the number of clusters.

Next, we checked the silhouette score and average silhouette score to further evaluate the quality of the clusters (Figure 11). The silhouette score is a measure of how well-defined each cluster is and how similar the data points within the cluster are to each other. It is calculated for each data point by comparing the distance between the point and the centroid of its cluster to the distance between the point and the centroid of the nearest cluster. A higher silhouette score indicates a more distinct and well-defined cluster, while a lower score indicates a less distinct or more overlapping cluster.

Finally, we selected the number of clusters that provided the best combination of low inertia and high silhouette scores. In our final solution, we selected 3 as the number of clusters based on these evaluation metrics.

6.2. SOM

Self-organizing maps (SOMs) are a type of unsupervised machine learning algorithm that can be used for data visualization, clustering, and feature extraction. In this analysis, we used SOMs to analyze a dataset and obtained a final quantization error of 1,70 after training the model for 25,36 seconds.

The quantization error is a measure of the difference between the original data and the representation of the data on the SOM and the training time is the amount of time it took for the SOM to learn the structure of the data and adjust the weights of the neurons.

One potential advantage of using SOMs is their ability to visualize high-dimensional data in a low-dimensional space and reveal patterns and relationships within the data that may not be immediately apparent. They can also be used for clustering and feature extraction by identifying the neurons that are most activated by data points or groups of points. We also plotted a U-matrix which is a visualization of the distances between the neurons on the SOM and can be used to identify clusters or patterns within the data and a hit-map that is a visualization of the data points on the SOM, with each data point represented by a coloured dot. (Figure 13,14,15).

6.3. K-means on top of SOM units

In this approach, the SOM is used to pre-process the data and reduce the dimensionality, and then the k-means algorithm is applied to the reduced dataset to perform the actual clustering.

The performance of the k-means algorithm may be affected by the quality of the reduced dataset produced by the SOM, and the choice of the number of clusters may also have an impact on the results. It may be necessary to try different parameters and configurations to find the best combination of SOM and k-means for a particular dataset and task. We plotted a figure which we can see how many clusters we should retain (Figure 16).

6.4. Hierarchical Clustering on top of SOM units

In our analysis, we applied Hierarchical Clustering on top of Self-Organizing Map (SOM) units to the data. We first visualized the R^2 scores for each cluster solution on demographic variables to assess the quality of the clusters. We then performed Hierarchical clustering on top of the SOM units, resulting in 4 clusters (Figure 17,18).

6.5. *Mean shift clustering*

Based on the results of mean shift clustering using a bandwidth of, approximately, 2.31, the algorithm identified a total of 12 clusters in the data. The R^2 value of 0.2556 for the cluster solution indicates that the clusters can explain approximately 26% of the variance in the data.

Overall, the results of mean shift clustering using a bandwidth of, approximately, 2.31 and identifying 12 clusters suggest that the data exhibits a moderate degree of structure, but additional factors may also be influencing the data.

6.6. *DBSCAN*

DBSCAN identified 2 clusters in the data and classified 218 rows as noise. The R^2 value for the clusters was 0.1124, meaning that the clusters only explain about 11% of the variance in the data. This suggests that the data has a low degree of structure, with few distinct clusters and many observations classified as noise (218 rows). The low R^2 value also indicates that the clusters do not capture a significant amount of variance in the data.

6.7. *GMM*

Using the criteria of the Akaike information criterion (AIC) and the Bayesian information criterion (BIC), we selected a Gaussian mixture model (GMM) with 6 components to cluster the data (Figure 19). The R^2 value of 0.3258 for the cluster solution indicates that the clusters can explain approximately 33% of the variance in the data.

The results of the GMM clustering suggest that the data has a moderate degree of structure, with distinct clusters and some overlap between them. The relatively high R^2 value indicates that the clusters can capture a significant amount of variance in the data.

6.8. *K-means and Hierarchical clustering*

We applied K-means on top of hierarchical clustering to the data and visualized the R² scores for each cluster solution on demographic and insurance variables (Figure 20, 21). We selected the right clustering algorithm and number of clusters for each perspective, ultimately deciding on 4 clusters for each perspective and merging them using hierarchical clustering (Figure 22). This led to the definition of 6 clusters for the hierarchical clustering solution, which were identified using centroids for the previously defined clusters and a threshold. The use of both K-means and hierarchical clustering allowed us to identify clusters based on both demographic and insurance variables, and the R² scores indicate that the clusters can capture a significant amount of variance in the data.

7. Cluster Analysis

Cluster analysis is a common technique in data mining and machine learning, and it can be used to discover hidden patterns and trends in the data, identify groups of similar observations, and make predictions about future behavior.

There are many different approaches to performing cluster analysis, including K-means clustering, hierarchical clustering, and density-based clustering algorithms, that we previously used. The choice of which algorithm to use will depend on the characteristics of the data and the specific goals of the analysis.

After analyzing the result that came from the Cluster Profiling (Figure 23). We have divided the demographic and insurance clustering into 4 clusters. After merging the labels, we got the result of 6 clusters, where we can see that the characteristics of each cluster suggest that they may represent a group of individuals with a particular insurance need or risk profile.

In conclusion, the clustering analysis identified 6 distinct groups of clients for the company to consider when developing marketing strategies. These clusters may have different insurance needs or risk profiles, and careful evaluation of the clusters is necessary to determine their usefulness for the specific application. Some characteristics or variables that stand out across the clusters include "Salary," "Motor," and "Household" insurance, as well as the remaining types of insurance. It may be logical to create different strategies for these groups to optimize marketing efforts. It is important to carefully consider the unique needs and characteristics of each cluster to effectively target marketing efforts and better serve the diverse needs of the company's customer base.

7.1. Cluster Visualization using t-SNE

Cluster visualization is an important step in the cluster analysis process, as it allows us to understand and interpret the structure and patterns in the data. One common technique for visualizing clusters is t-SNE (t-distributed stochastic neighbor embedding), which is a non-linear dimensionality reduction method that can effectively visualize high-dimensional data in two or three dimensions.

Based on the results of using t-SNE to visualize 6 clusters in the data (Figure 24), we can see the relative positions and relationships between the different clusters. This can help us understand how the clusters differ from each other and identify any patterns or trends that may be present.

Overall, cluster visualization using t-SNE can be a valuable tool for understanding and interpreting the results of cluster analysis, and for identifying patterns and trends in the data.

8. Conclusion

Following on our clustering analysis, we can observe some notable differences in the insurance profile of the different clusters. For example, Cluster 0 has a high concentration of motor insurance, while the other clusters have a lower concentration of insurance policies. Cluster 1 has a high concentration of CMV and motor insurance, while the other clusters have a lower number of insurance policies. Cluster 2 has a small sample size, but we can see that they have a high salary and a high concentration of CMV insurance, as well as a low concentration of motor insurance and a high concentration of household insurance. The remaining insurance policies have a lower concentration in this cluster. Cluster 3 also has a small sample size, and we can see that they have a low salary compared to the other clusters, a high concentration of CMV and household insurance, and a low concentration of motor insurance and customer years. This cluster also has a high salary rate, indicating that they spend a significant proportion of their annual salary on insurance. Cluster 4 has a high salary and a low concentration of CMV and motor insurance, while the remaining insurance policies have a medium concentration in this cluster. Finally, Cluster 5 has a low salary and a low concentration of CMV and motor insurance, with a medium concentration of the remaining insurance policies. This cluster also has a small value for customer years and a high salary rate, indicating that they spend a large proportion of their annual salary on insurance.

Based on the results of our analysis, we can create a marketing strategy that targets the specific needs and risks of each segmentation group. For example, we might consider targeting Cluster 0 and 1 with high-end car insurance options, while offering a good deal for the remaining clusters. We can also consider targeting Cluster 2 and 3, which have a high concentration of household insurance, with better deals for the remaining insurance types. These clusters may have a large house or multiple houses, and therefore may have higher insurance needs. We can also consider applying this strategy to Cluster 0 and 1, which have a lower concentration of household insurance.

For the remaining variables such as health, life, and work insurance, we can observe that medium salary clusters tend to spend less on these insurance types. Therefore, we could create a campaign or package that offers these insurance types as a bundle for this group. For the other clusters, which have a medium concentration of these insurance types, we could offer better deals with the insurance company. Finally, we can target clients in Cluster 3 and 5, who are more likely to pay for insurance due to their high salary rates, with marketing efforts focused on these clusters. Overall, our marketing segmentation analysis can help us to better understand the unique needs and characteristics of our customer base and develop targeted marketing strategies that effectively meet these needs.

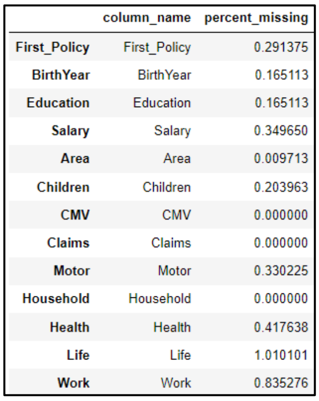
9. References

Provost, F., & Fawcett, T. (2013). *Data Science For Business.* O'Reilly Media, Inc.

S. Linoff, G., & J. A. Berry, M. (2011). *Data Mining Techniques, For Marketing, Sales, and Customer Relationship Management.* Wiley Publishing, Inc.

VanderPlas, J. (2016). *Pyhton Data Science Handbook.* O'Reilly Media, Inc.

10. Appendix



	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 1 – Percentage of Missing Data

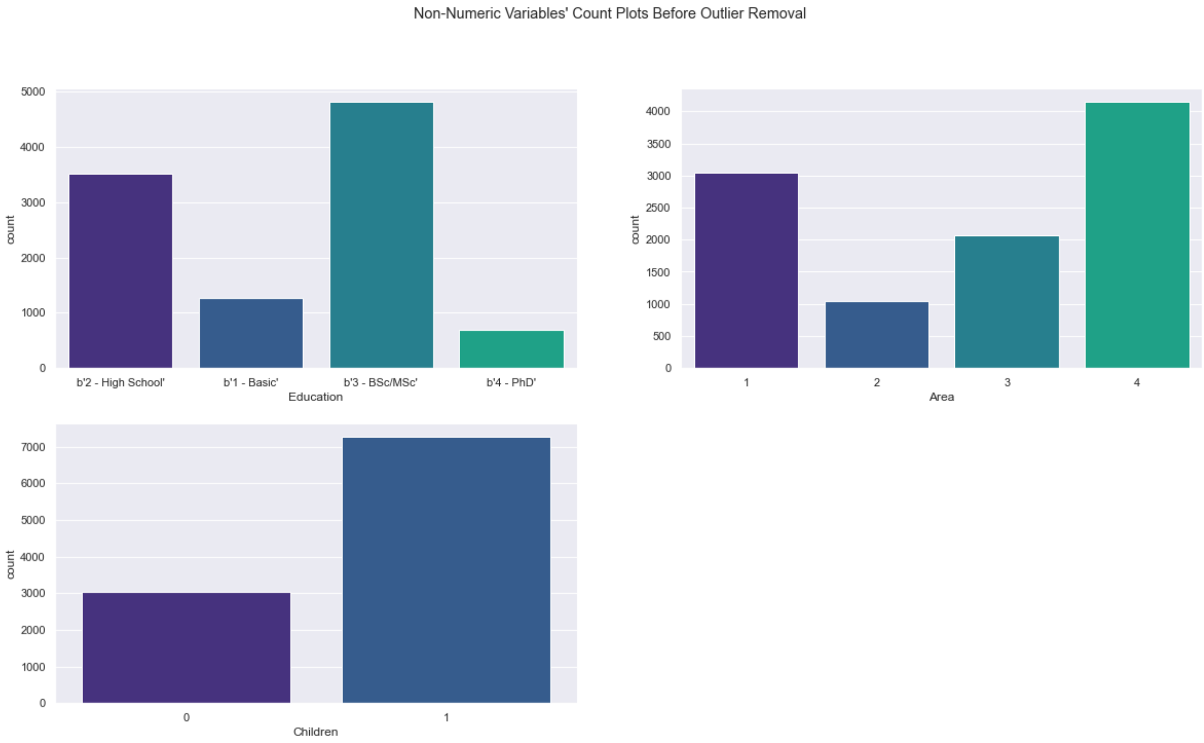


Figure 1 – Non-Numeric Variables’ Count Plots Before Outlier Removal

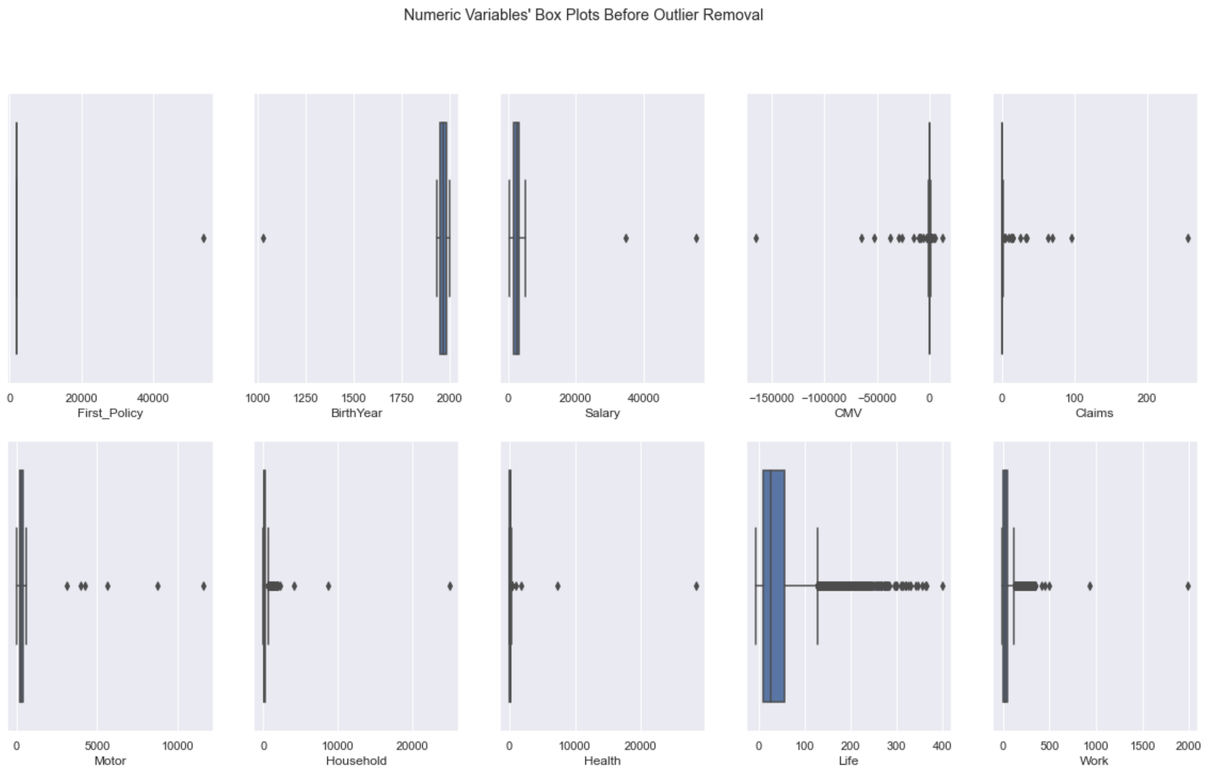


Figure 2 – Numeric Variables’ Box Plots Before Outlier Removal

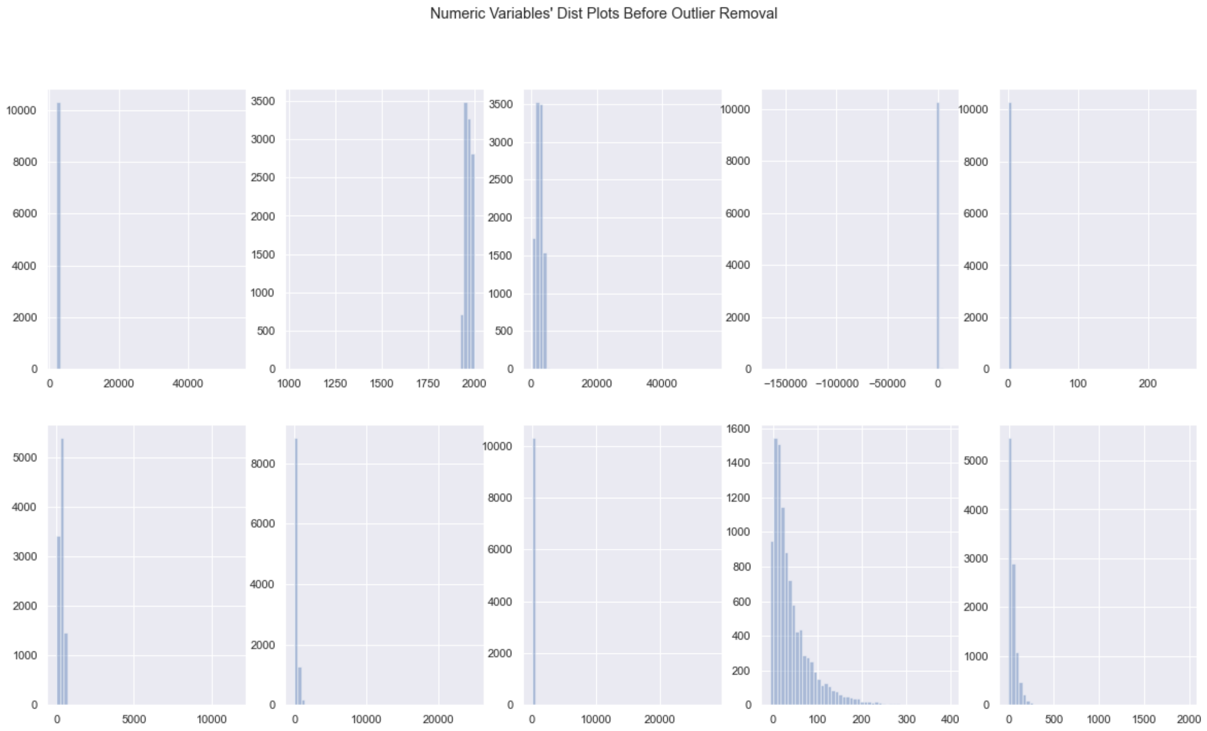


Figure 3 – Numeric Variables’ Dist Plots Before Outlier Removal

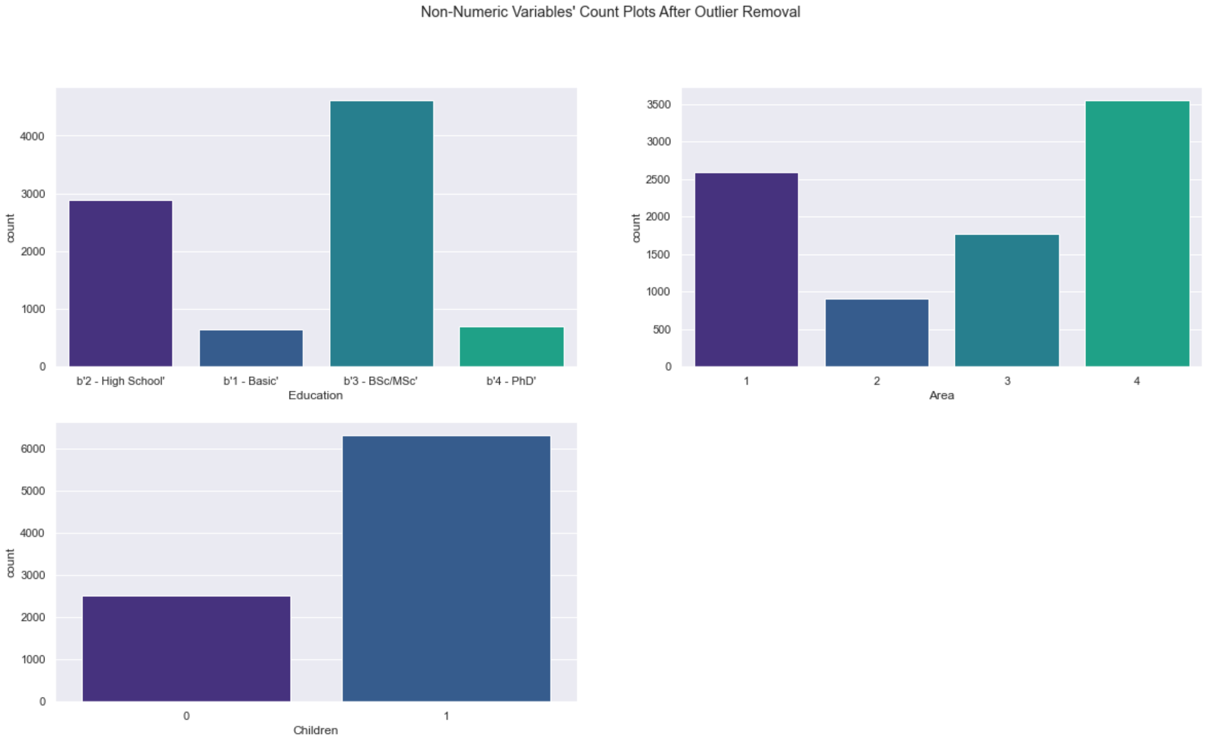


Figure 4 – Non-Numeric Variables' Count Plots After Outlier Removal

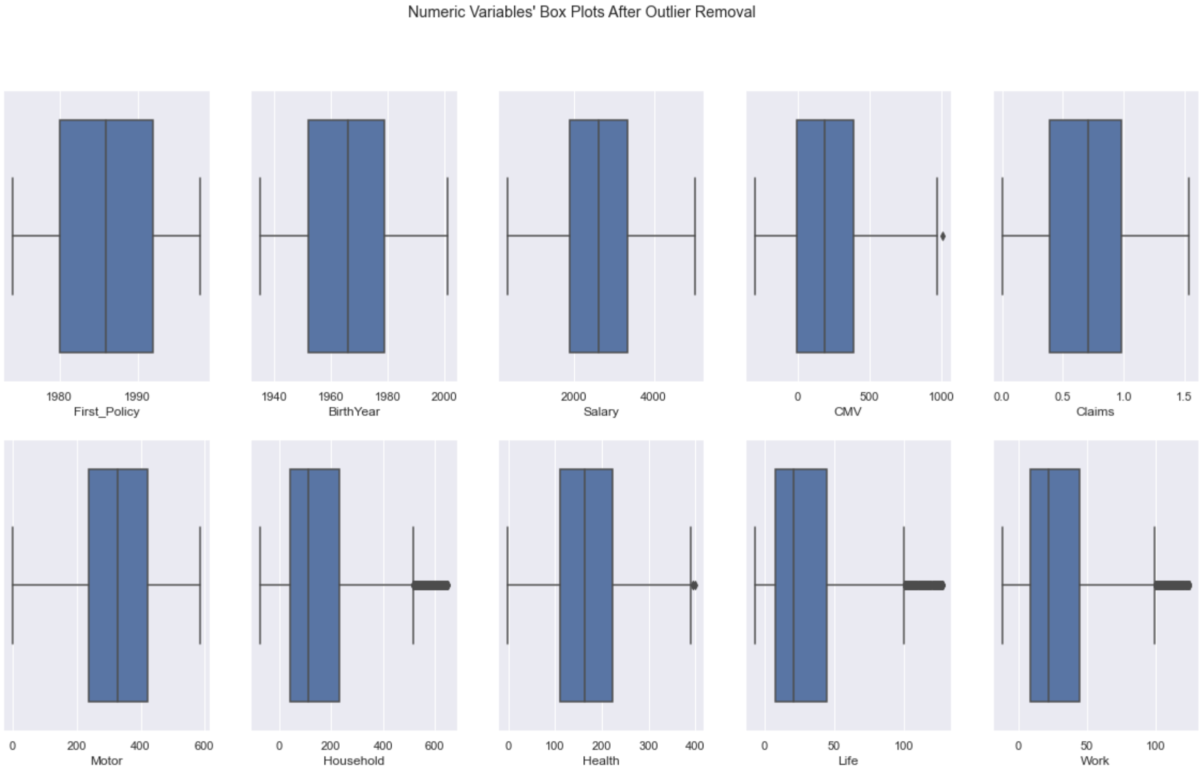


Figure 5 – Numeric Variables' Box Plots After Outlier Removal

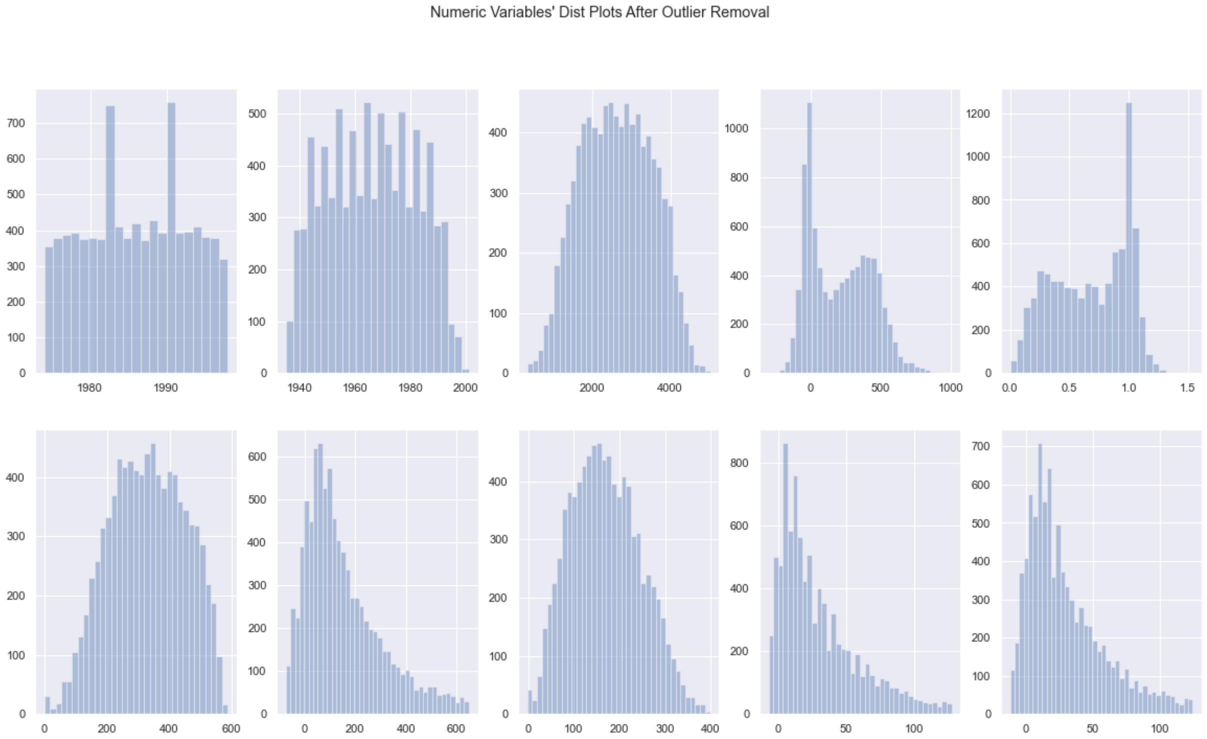


Figure 6 – Numeric Variables’ Dist Plots After Outlier Removal

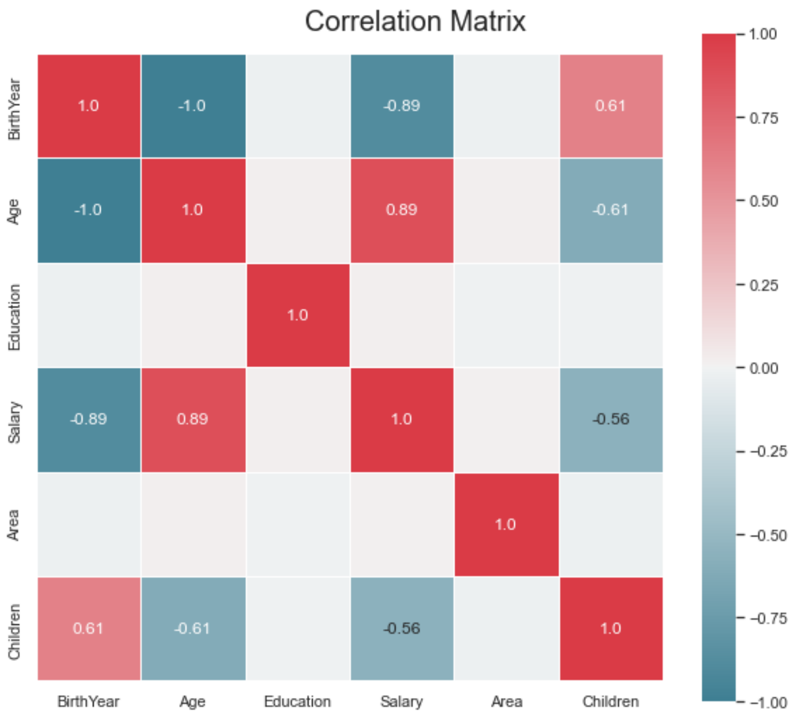


Figure 7 – Correlation Matrix for Demographic DataFrame

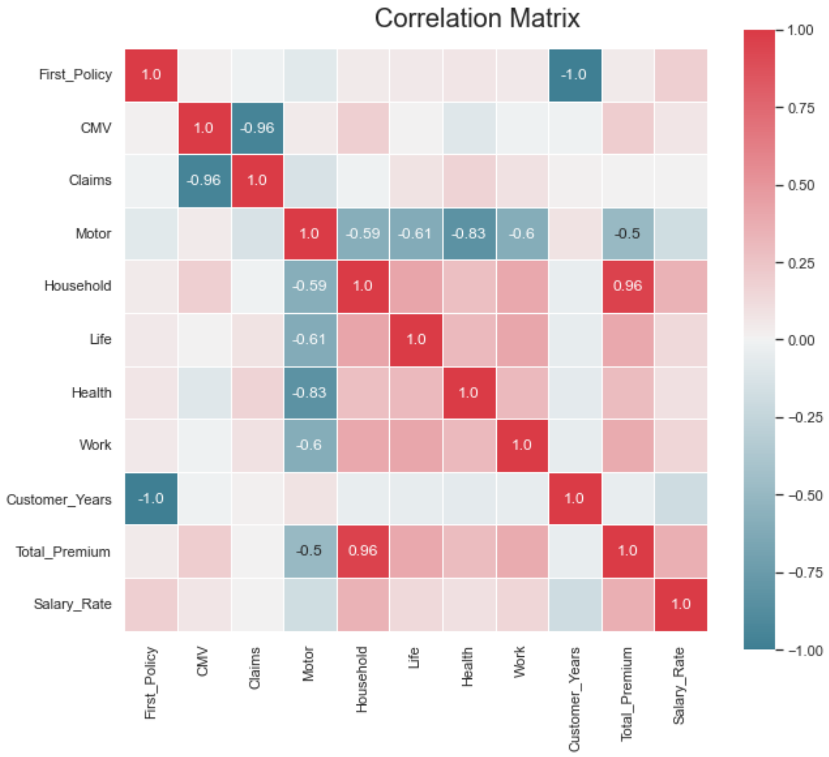


Figure 8 – Correlation Matrix for Insurance DataFrame

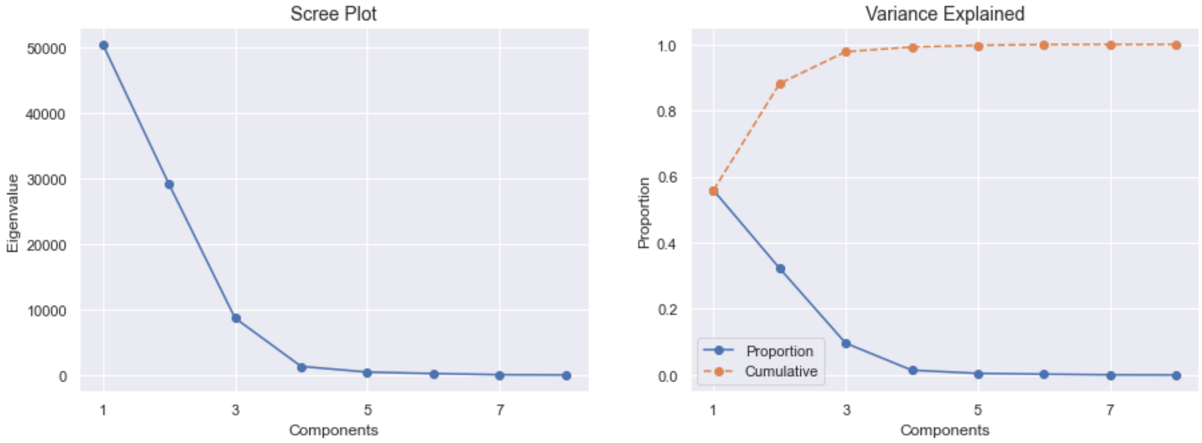


Figure 9 – Scree Plot and Variance Plot for Number of Components (PCA)

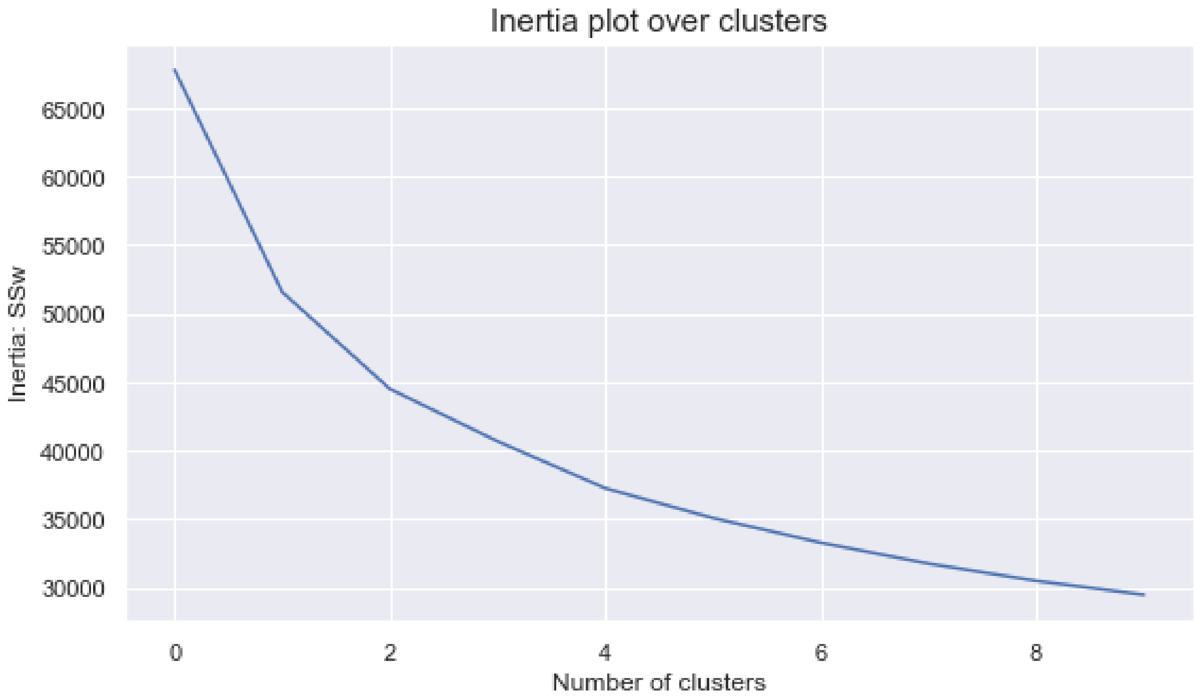


Figure 10 – Inertia Plot (K-Means)

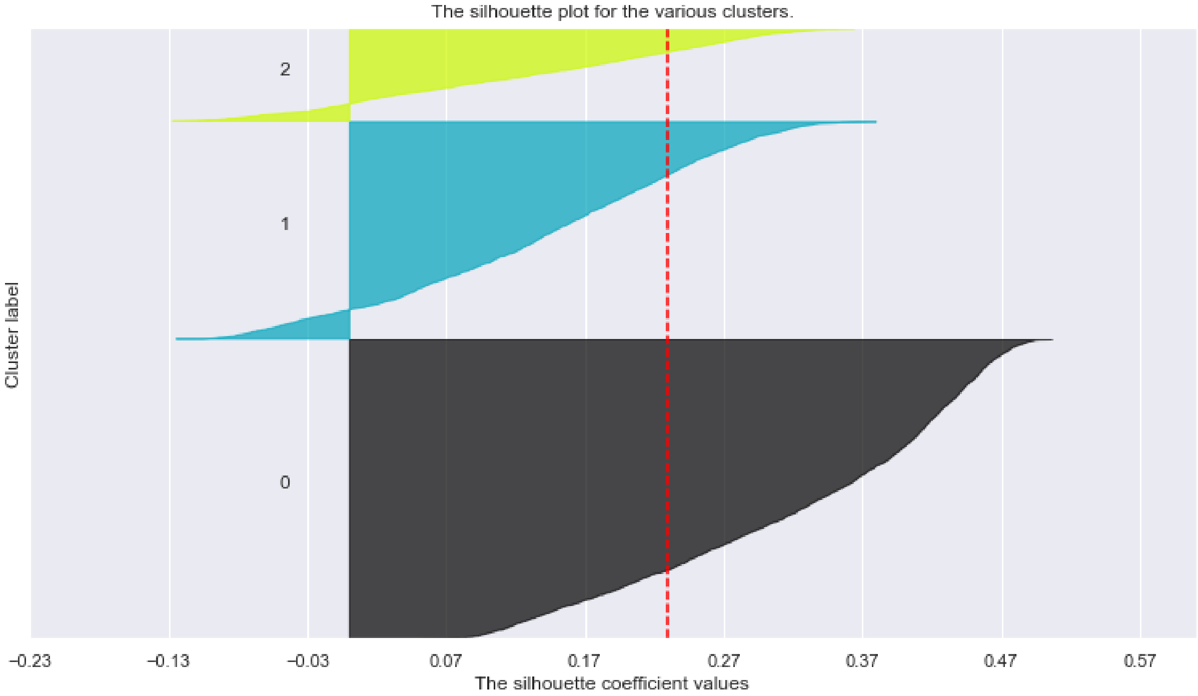


Figure 11 – Silhouette Plot for 3 Clusters (K-Means)

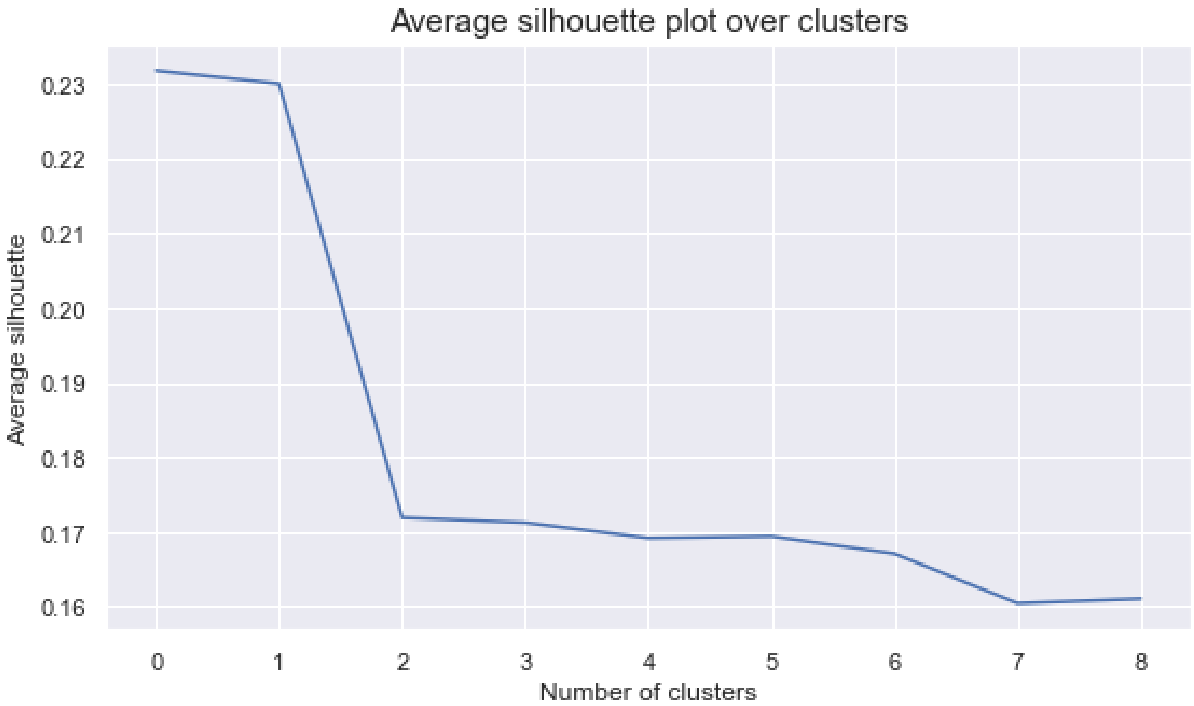


Figure 12 – Average Silhouette Plot Over Clusters (K-Means)

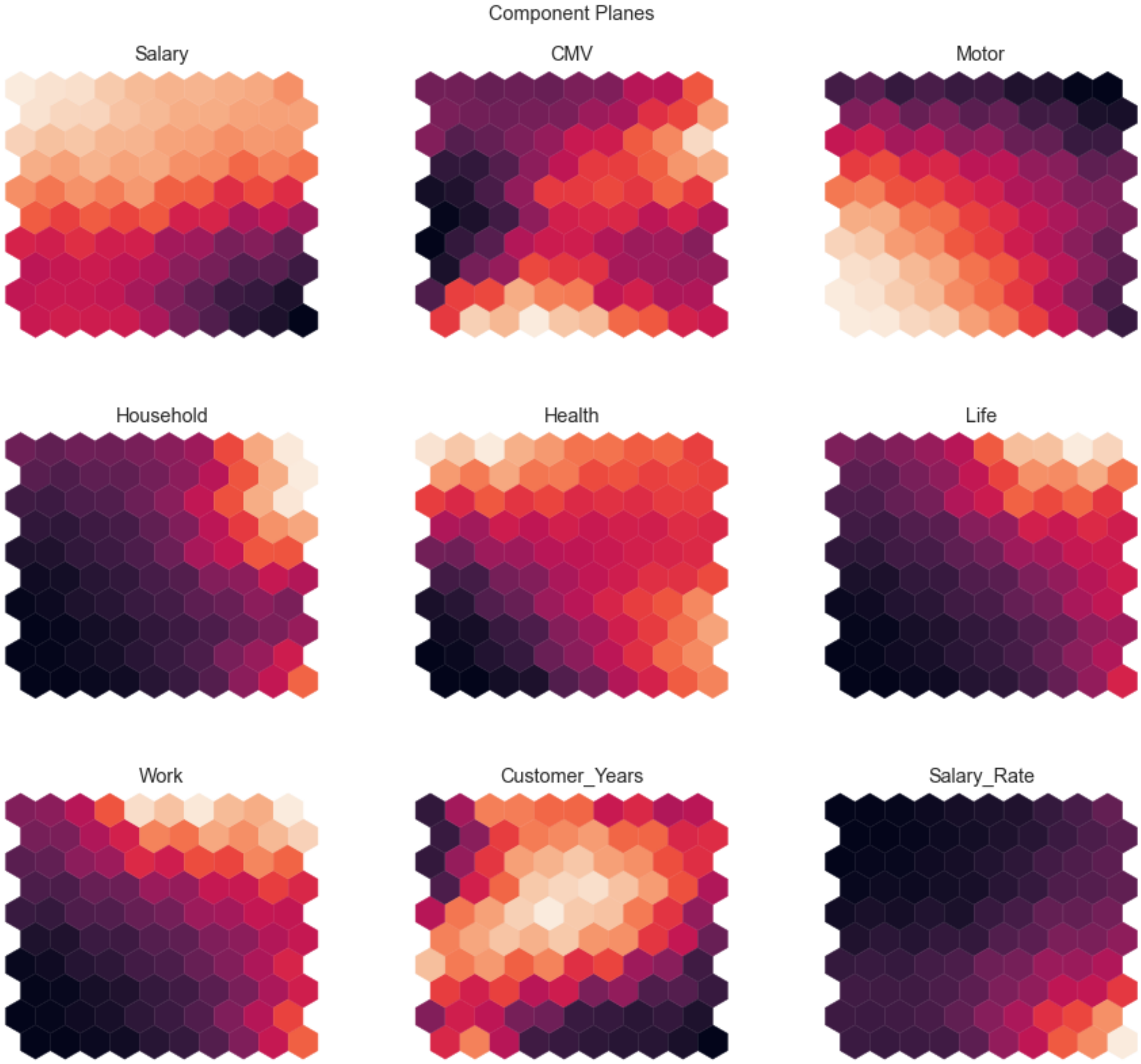


Figure 13 – Component Planes (SOM)

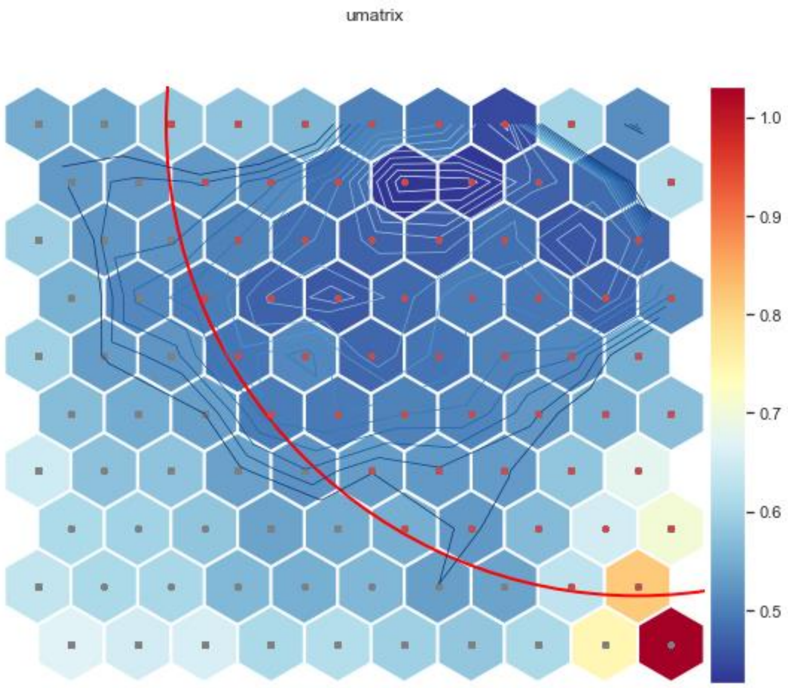


Figure 14 – U-Matrix (SOM)

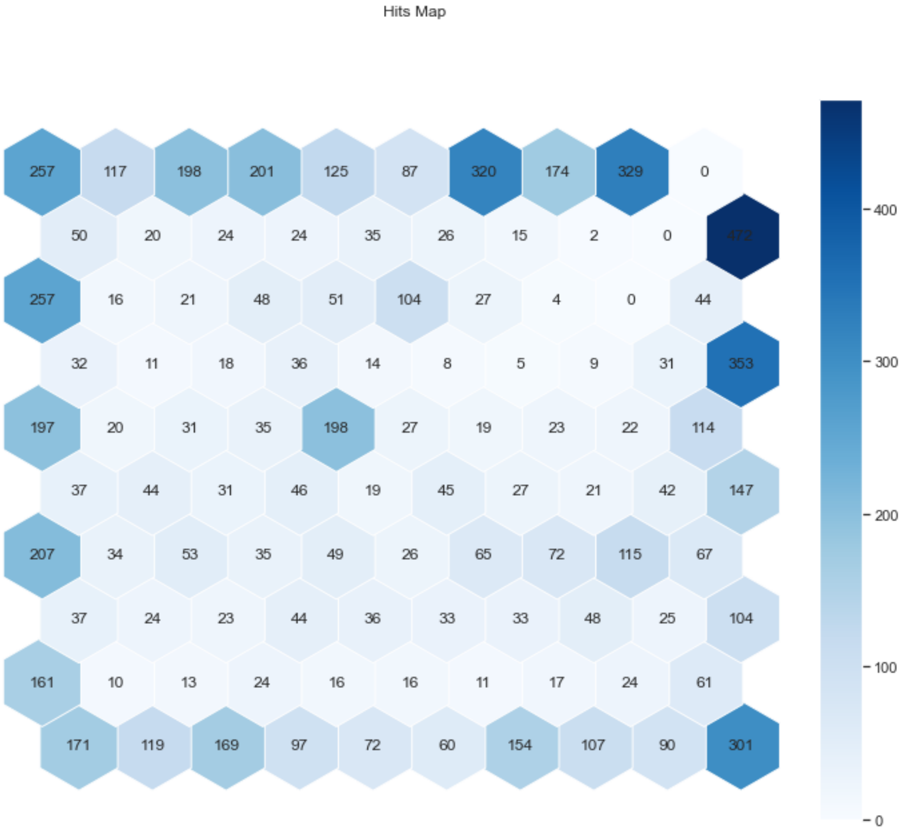


Figure 15 – Hit-Map (SOM)

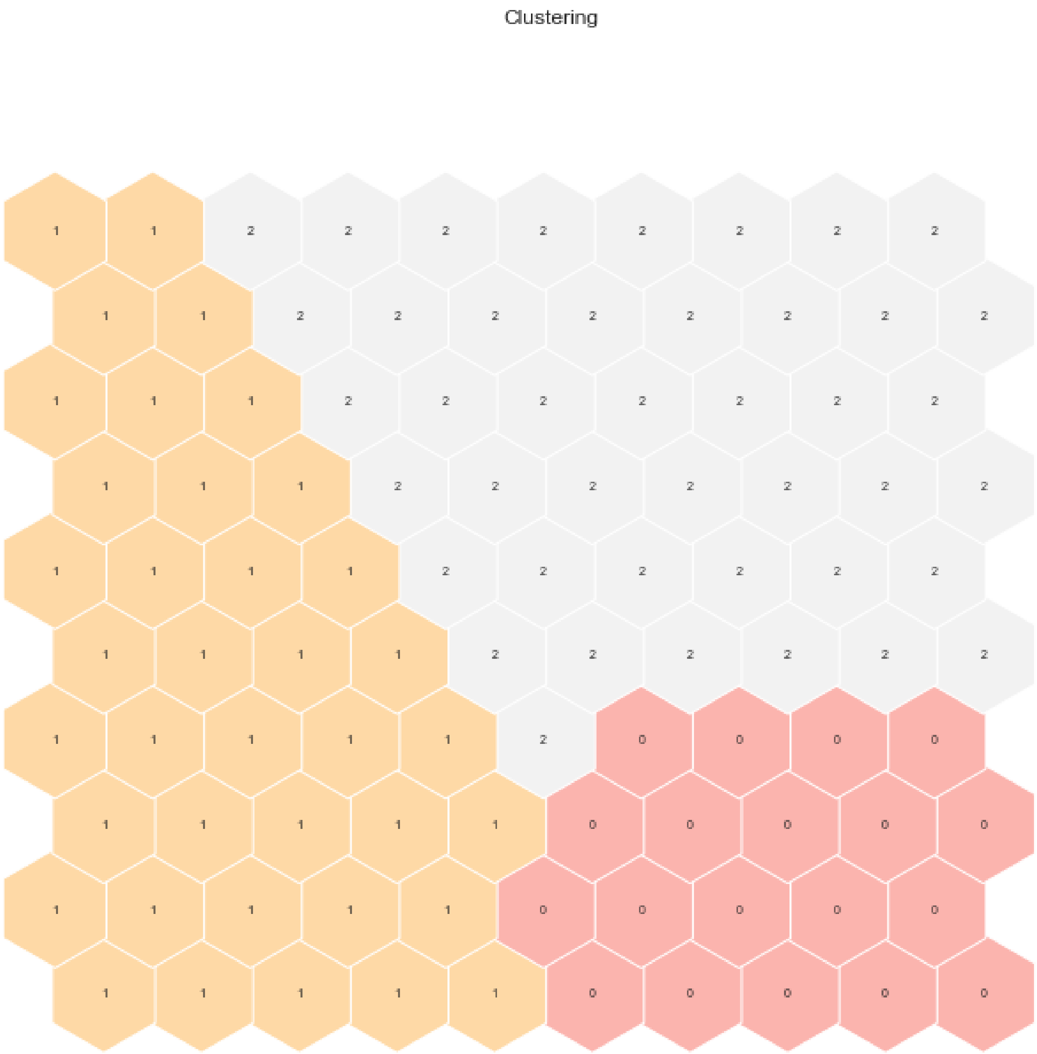


Figure 16 – Clustering (K-Means on top of SOM units)

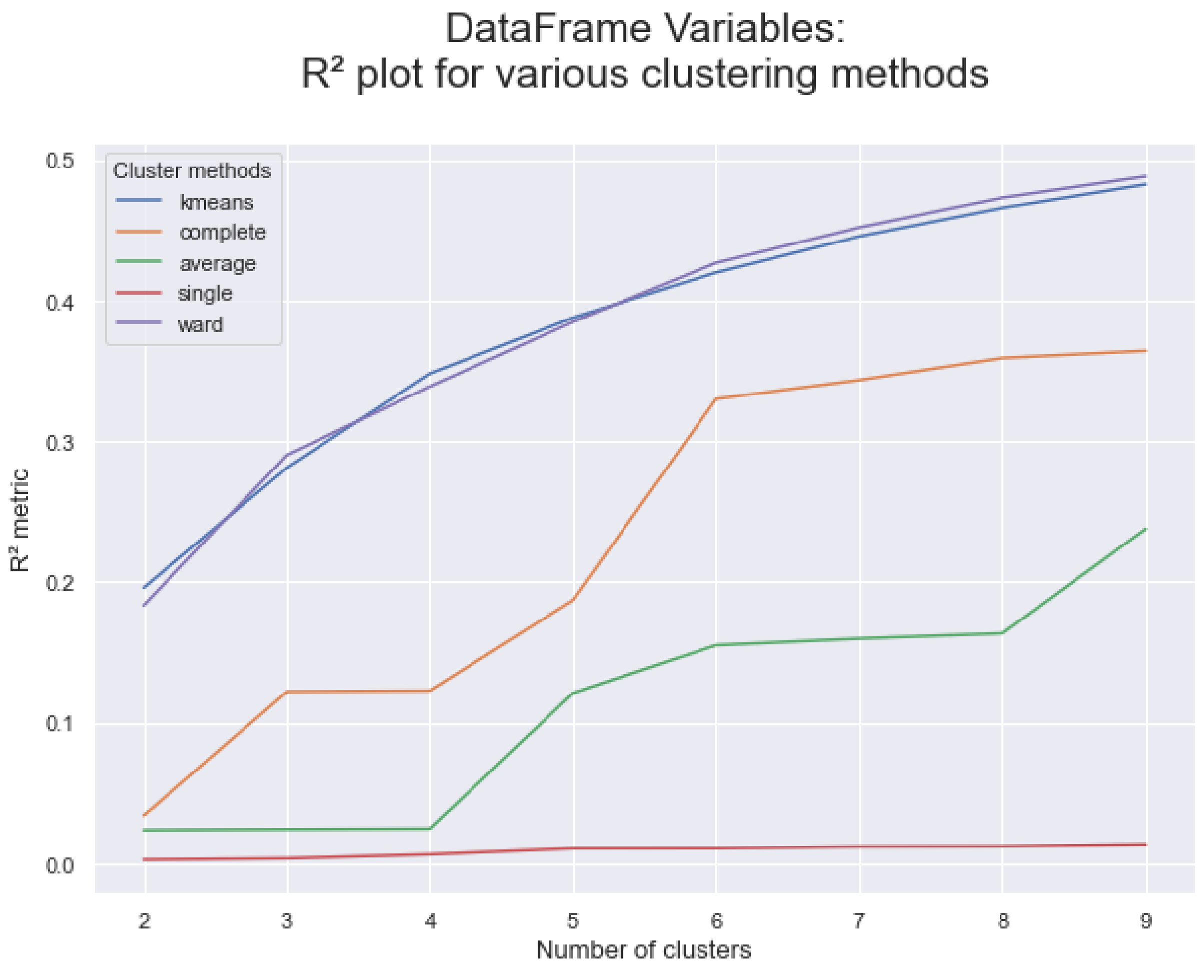
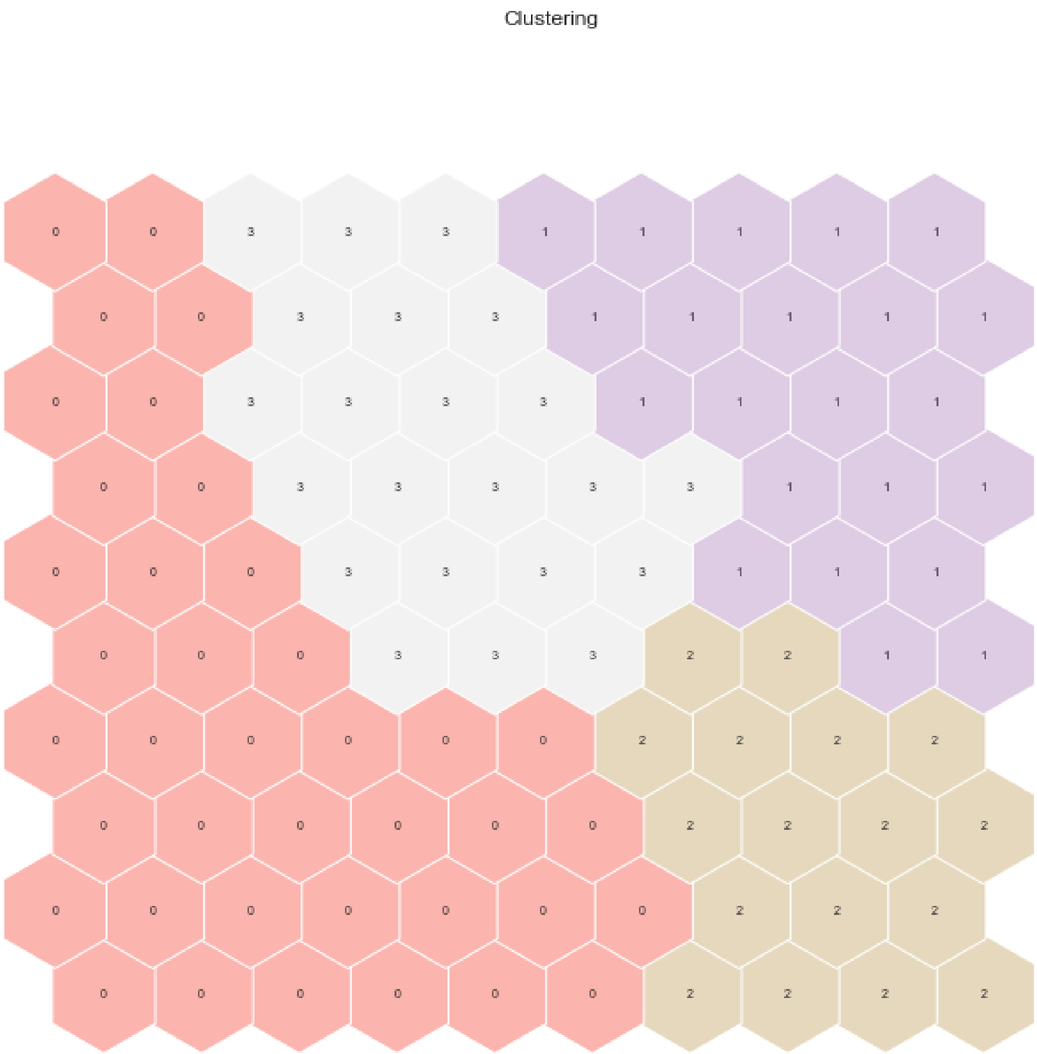


Figure 17 – R² Plot for various clustering methods applied to the DataFrame



0 0 3 3 3 1 1 1 1 1
0 0 3 3 3 1 1 1 1 1
0 0 3 3 3 3 1 1 1 1
0 0 3 3 3 3 3 1 1 1
0 0 0 3 3 3 3 1 1 1
0 0 0 3 3 3 2 2 1 1
0 0 0 0 0 0 2 2 2 2
0 0 0 0 0 0 2 2 2 2
0 0 0 0 0 0 0 2 2 2
0 0 0 0 0 0 2 2 2 2

Figure 18 – Clustering (Hierarchical Clustering on top of SOM units)

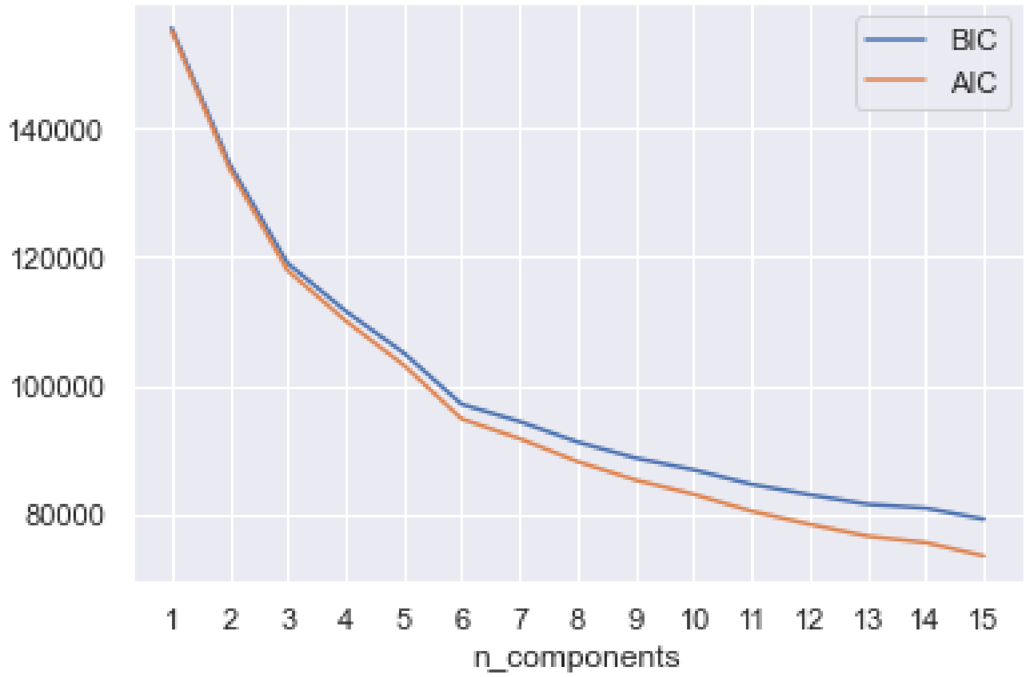


Figure 19 – Number of components using the BIC and AIC criteria (GMM)

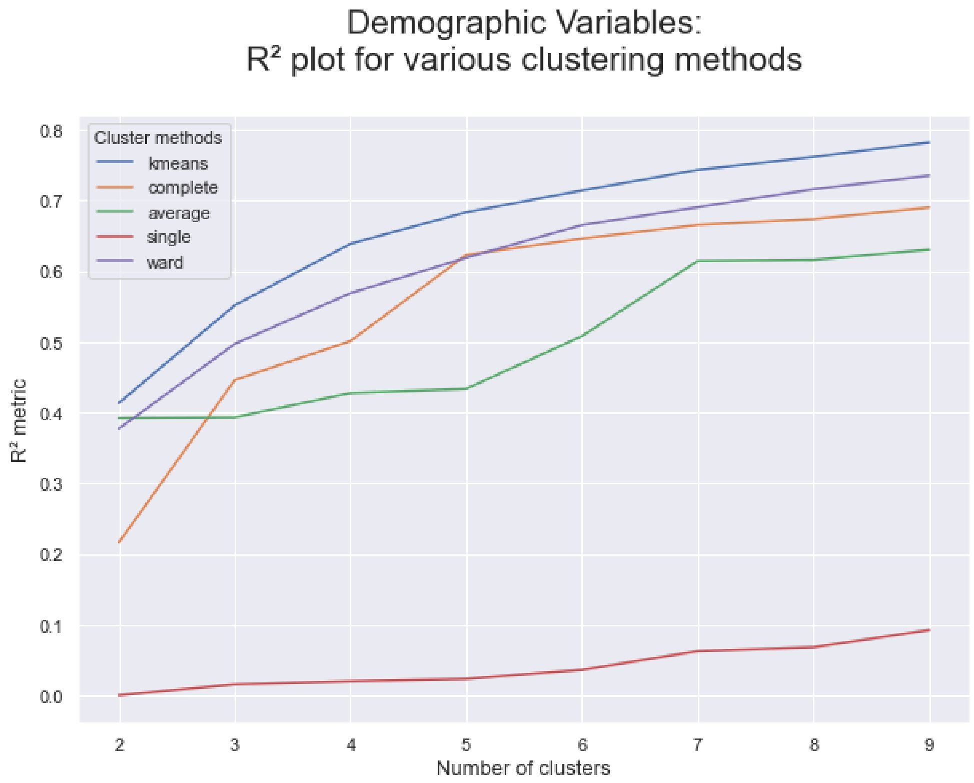


Figure 20 – R² Plot for various clustering methods applied to the Demographic DataFrame

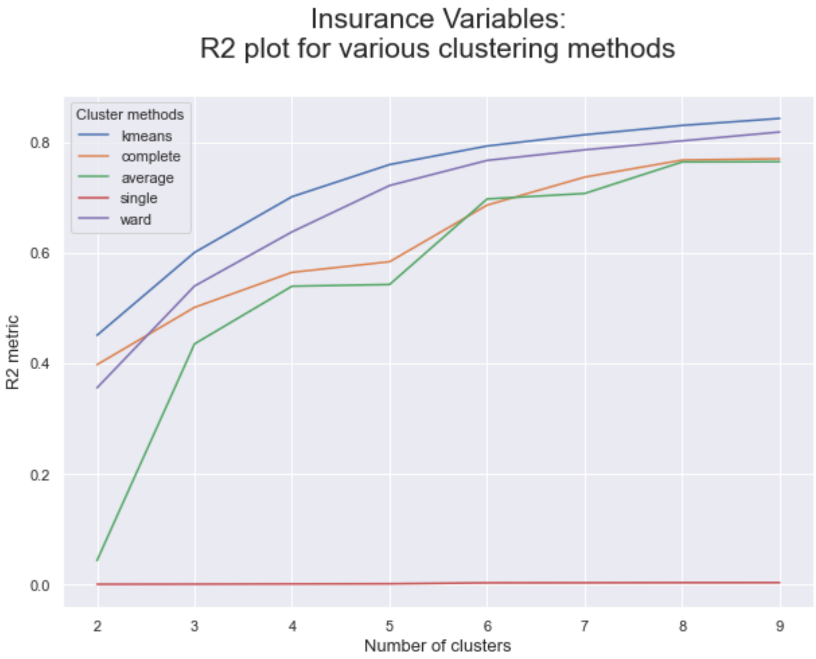


Figure 21 – R² Plot for various clustering methods applied to the Insurance DataFrame

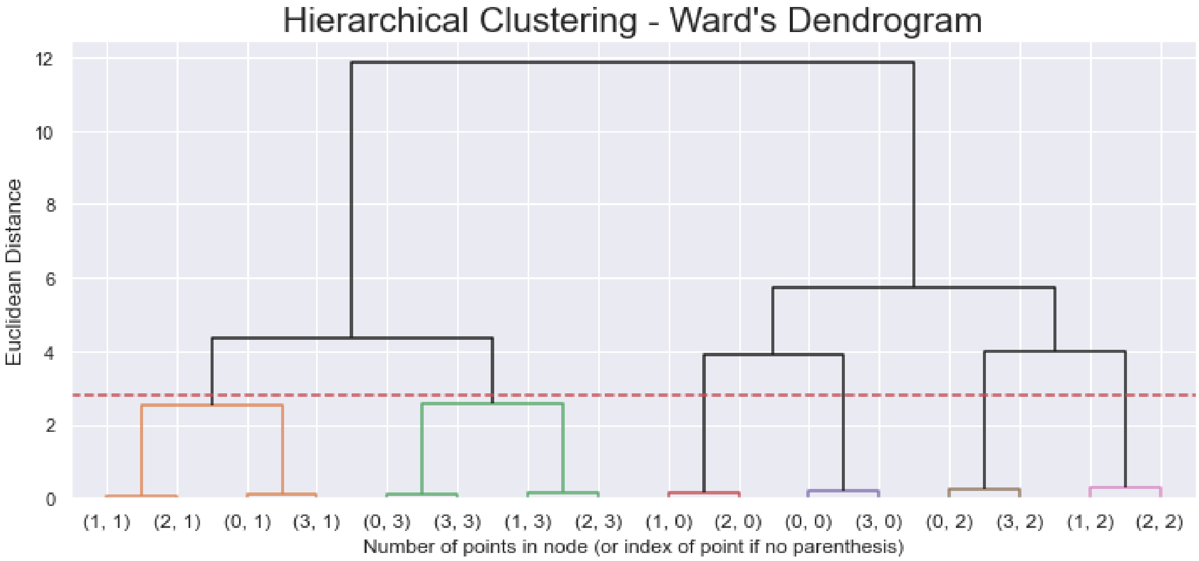


Figure 22 – Hierarchical Clustering - Ward's Dendrogram

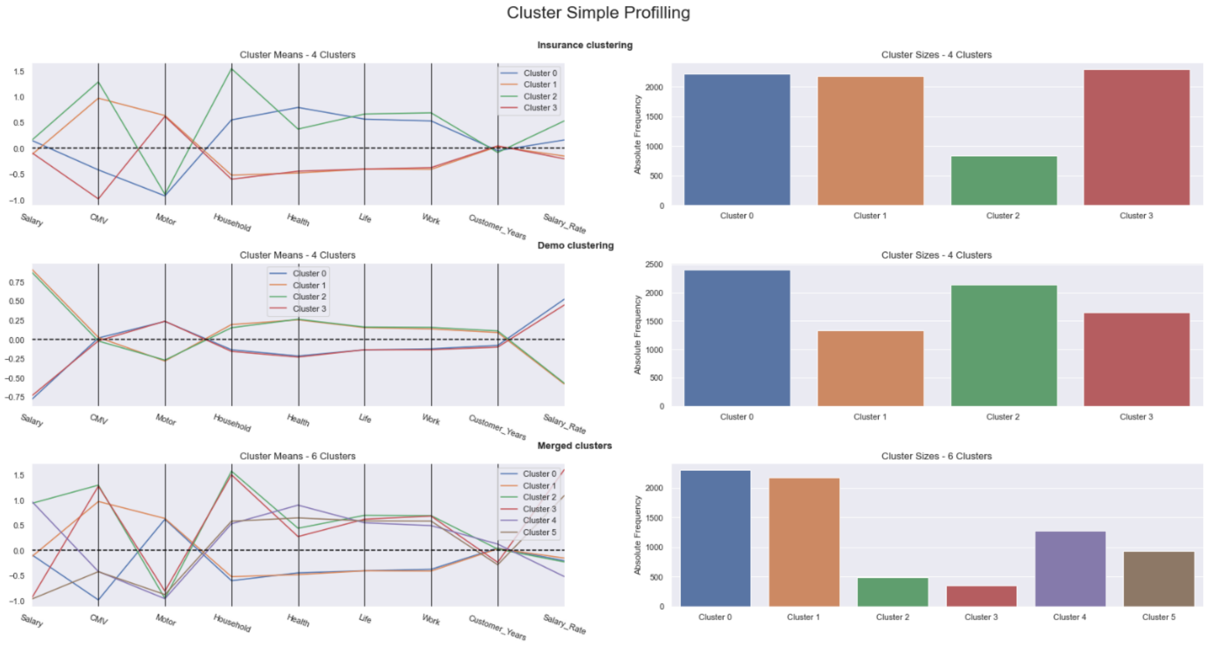


Figure 23 – Cluster Profiling for Demographic, Insurance and Merged DataFrames

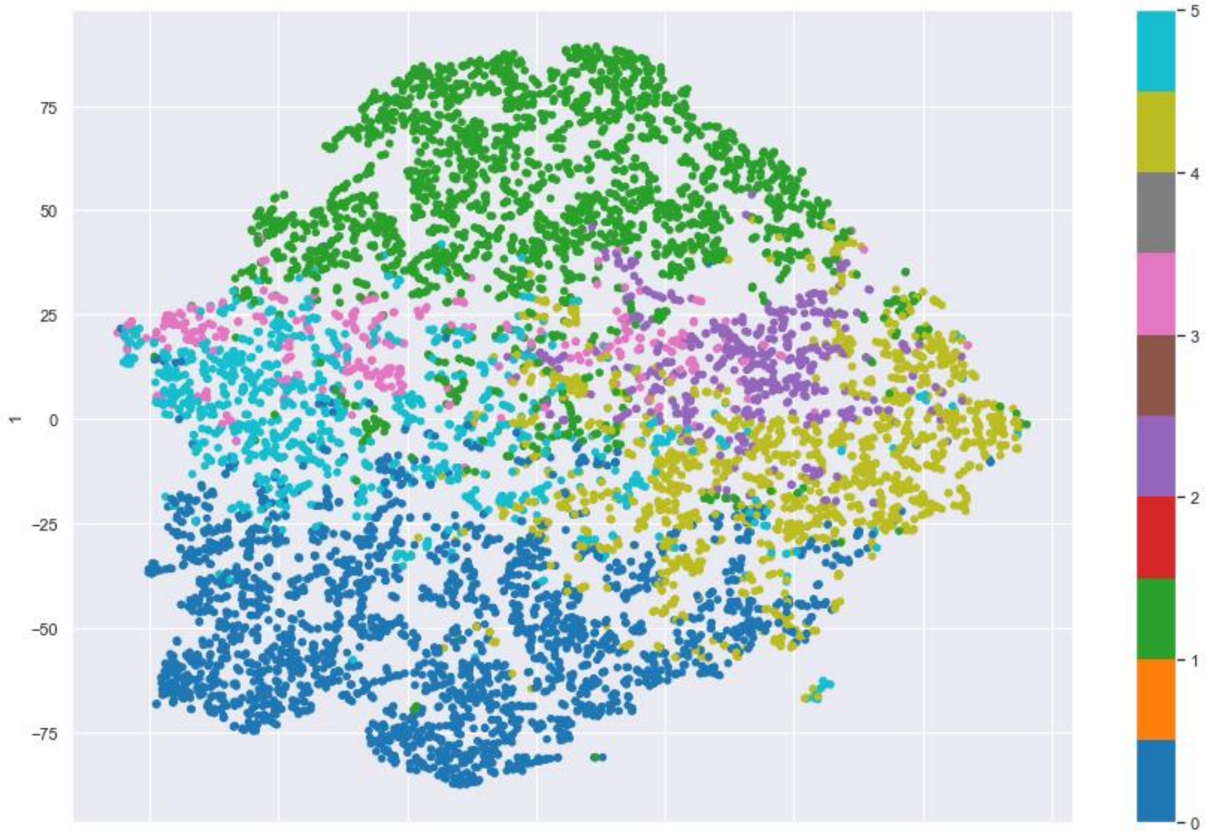


Figure 24 – Cluster Visualization (t-SNE)