Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z Insurance

Group CZ

Miguel Ramos Nº 20210581

José Matos Nº 20220607

Eduardo Palma Nº 20221022

January, 2023

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

1. Introduction

The purpose of this project is to perform market segmentation for "A2Z Insurance" in order to effectively target specific customer groups with customized marketing strategies. In order to achieve this, we will follow a multi-step process which includes exploring, cleaning, and processing the data, defining segments, clustering the segments, cross-tabbing the segments, and profiling each final cluster to identify unique groups of customers. By targeting these customer groups with personalized offers that align with their interests and insurance consumption patterns, we aim to not only reduce costs for the company but also increase revenues through more effective marketing efforts. The final outcome of the project will be a set of well-defined customer segments and a proposal for their corresponding marketing campaigns, ultimately resulting in enhanced success for "A2Z Insurance."

2. Data Preparation and Pre-Processing

2.1. Basic Data Exploration

To start our data exploration, we imported the dataset and analysed the number of observations and variables it contained. The dataset had 10296 observations and 13 variables. We set the "CustID" variable as the index and examined the descriptive statistics (*Figure 1*) to become familiar with the data and understand it better. While examining the descriptive statistics, we noticed some patterns. We will not go into detail here as we will discuss them further in the report, but it is worth mentioning that we identified the presence of outliers in nearly all variables based on the minimum and maximum values, as well as the significant gap between the minimum and first quartile, and the third quartile and maximum. We also noticed that some variables have fewer than 10296 assigned values, indicating the presence of missing values.

We then noticed that 3 pairs of rows had different "CustID" but had equal values for the other variables, and so we decided to remove one row from each pair. We also discovered that 17 observations had empty values for "EducDeg", which were not identified in the descriptive statistics. We temporarily replaced these values with null values so that we could keep track of them. In order to extract knowledge from all the variables, we decided to transform the values of "EducDeg" into ordinal values. This will allow us to use them in the correlation matrix and during clustering, for example.

2.2. Consistency Check

In the second step of our analysis, we checked to see if the relationships between the variables made sense, as some combinations may not be logical.

- ***"BirthYear" and FistPolYear":*** We found that in 1997 observations, the "FirstPolYear" was earlier than the "BirthYear", which is not possible since an individual must be alive to purchase insurance. One of the variables is clearly incorrect, but at this point we do not have enough information to determine which one is incorrect. We will address this issue later. We then checked "BirthYear" for individuals older than 110 years and found one such observation. We also checked "FirstPolYear" and found an observation with a value of 53784, which is not possible. Both of these outliers will be addressed during the outlier removal process.

- ***"BirthYear" and "EducDeg":*** Upon examining both variables together, everything appears to be in order. The minimum ages for completing Basic School, High School, a BSc/MSc, and a PhD are 15, 18, 20, and 23, respectively, and these values align with our expectations. There is nothing to note in this regard.

- ***"BirthYear" and "MonthSal":*** We chose to investigate this relationship because it is unlikely for someone born after 2000 to have a monthly salary. In fact, we found 12 observations of individuals born in 2001 (at 15 years old) who had a monthly salary. Since relations is not something we are thinking about addressing during our outlier removal, we decided to delete these observations at this stage.

2.3. Identifying and Removing Outliers

Before starting the outlier removal, we created histograms (*Figure 2*), box plots (*Figure 3*) and bar charts (*Figure 4*) for the variables to gain a better understanding of the data and assess the level of outlier presence. We created a correlation matrix (*Figure 5*) to better understand the relationships between certain variables. To account for potentially non-linear relationships, we chose to use the "Spearman" method for the correlation matrix. Upon analysing the box plots, we observed two types of outliers: ones that were clearly mistakes in data entry, and others that were correct but extremely rare and not representative of any group of people due to their uniqueness. These latter outliers will also be removed as they do not aid in generalization for clustering.

We decided to try two approaches for removing the outliers. One approach involves setting an arbitrary threshold for each variable, and the other approach involves defining a threshold for each variable based on the mean and standard deviation of that variable. Ultimately, we chose the second approach (*Figure 6*) for removing the outliers. For certain variables, such as "FirstPolYear", "BirthYear", "MonthSal", "CustMonVal", "ClaimsRate", "PremMotor", "PremHousehold", and "PremHealth", we eliminated any observation that distanced more than three standard deviations away from the mean. For other variables, such as "PremLife" and "PremWork", we eliminated any observation that distanced more than four standard deviations away from the mean. This decision was made because these latter variables often had more extreme values, and we didn't want to penalize them too severely. This process resulted in the elimination of a certain number of observations for each variable: 1 for "FirstPolYear", 1 for "BirthYear", 2 for "MonthSal", 13 for "CustMonVal", 12 for "ClaimsRate", 6 for "PremMotor", 35 for "PremHousehold", 3 for "PremHealth", 78 for "PremLife", and 61 for "PremWork". In addition, any outliers identified during the consistency check were also eliminated using this technique.

2.4. Filling Missing Values

Before we began filling in the missing values (*Figure 7*), we examined the correlation matrix (*Figure 8*) after outliers had been removed. This was done because it may be helpful to predict some of the missing values using a regression or KNN Imputer, and we wanted to use variables that had a high correlation with the variables containing missing values.

- ***"Premiums":*** For the missing values in the variables "PremMotor," "PremHealth," "PremLife," and "PremWork," we assumed that these customers chose not to purchase these particular types of insurance and instead opted for the other available options. Therefore, we set the

missing values to 0 to reflect this. After making these changes, we checked for customers who had all five premiums set to 0, indicating that they did not have any insurance with our company. We found 12 such customers and removed them from the dataset.

- ***"BirthYear":*** We used a linear regression model to estimate their values based on the "MonthSal" and "Children" columns, which have high correlations with "BirthYear" (-0.93 and 0.51, respectively). After filling in the missing values with our predictions, we converted the "BirthYear" column to integer values, as it should only contain whole numbers.

- ***"MonthSal":*** We used a similar approach as before, utilizing "BirthYear" and "Children" as estimators based on their correlations with "MonthSal" (-0.93 and -0.48, respectively). After filling in the missing values with our predictions, we converted the "MonthSal" column to integer values, as it previously only contained whole numbers.

- ***"Children":*** "To handle the missing values we used KNN Imputer with the most correlated variables: "BirthYear", "MonthSal", "PremMotor". and "PremHealth" (with correlation coefficients of 0.51, -0.48, 0.27, and -0.27, respectively). We had to scale the data for the KNN Imputer because it relies on distances. To fill in the missing values, we used the average of the 5 closest neighbours, which may result in decimal values. Therefore, we converted the "Children" column to integer values because it is a binary variable that can only take on the values 0 or 1."

- ***"FirstPolYear" and "EducDeg":*** For the missing values in "FirstPolYear", we chose to delete the rows because it is not correlated with any other variables, which we found strange and plan to address later. By deleting the 30 rows with missing values for "FirstPolYear", we were left with only 2 rows that had missing values for "EducDeg." We filled these missing values using the mode, as there were only 2 observations.

Now that we have addressed outliers and missing values, we can revisit the descriptive statistics (*Figure 9*) to get a better understanding of the data. Additionally, we can examine the updated histograms (*Figure 10*) and box plots (*Figure 11*) of each variable and the pairwise relationships between them (*Figure 12*) to identify potential bivariate outliers and more clearly understand the relationships between the variables.

2.5. Feature Engineering

Since we have a large number of features, we decided to perform feature engineering to transform some raw features into combinations that can help reduce the dimensionality of our clusters, which is an important consideration in data mining.

- ***"Age":*** Calculated by subtracting their "BirthYear" from 2016 and found that this variable is more intuitive than "BirthYear." As a result, we removed "BirthYear" from the analysis.

- **"Fidelity":** Computed using the same logic as before. It is important to note that this number does not represent the total number of years the customer has had an insurance with the company, but rather the number of years since their first insurance policy with the company. Deleted "FirstPolYear" from the analysis.

- **“Annual_Salary”:** Derived by multiplying the monthly salary by total number of months within a year. Erased “MonthSal” from the analysis.

- **“Total_Premium”:** Estimated by summing the values for “PremMotor”, “PremHousehold”, “PremHealth”, “PremLife”, “PremWork”, and represent the total amount a customer spends per year on insurances with the company.

- **“Insurance_Burden”:** Determined by dividing “Total_Premium” by the “Annual_Salary”, which represents the customer’s willingness to pay for insurance based on their salary.

- **“Ratios”:** Inferred the ratios by dividing each of the individual premiums ("PremHousehold", "PremHealth", "PremLife", "PremWork", "PremMotor") by "Total_Premium." This resulted in the creation of 5 new ratio variables: "Household_Ratio", "Health_Ratio", "Life_Ratio", "Work_Ratio", and "Motor_Ratio." It is important to note that in cases where a premium was reversed, we considered the ratio for that premium to be 0 because it would not make sense to have a negative ratio. These ratios will be helpful in understanding which types of insurance the customer spends the most on in proportion to the others.

- **“Cessation”:** We created a dummy variable that takes a value of 1 if any of the insurance premiums was cancelled, and 0 if none of the premiums were cancelled. This will be useful in understanding the company's ability to retain customers.

After creating these new features, we felt it was important to again examine the updated histograms (*Figure 13*), box plots (*Figure 14*), and correlation matrix (*Figure 15*) to ensure a thorough understanding of the data.

Upon examining the histograms and box plots, we identified some outliers in the newly created feature engineering variables. To address these outliers, we established maximum thresholds of 0.80 for "Household_Ratio," 0.70 for "Health_Ratio," and 1 for "Motor_Ratio. It resulted on the elimination of 84 observations.

2.6. Further Considerations

We will not include "Fidelity" in our analysis going forward, as previously mentioned in the report. There are several indications that this variable was not constructed correctly. First, it is only correlated between -0.02 and 0.02 with the other variables, which is unusual as one would expect it to be correlated with "Age" and "CustMonValue" at least. Additionally, during the consistency check, we discovered that 1997 observations had a higher "Fidelity" value than their "Age", which is impossible since a customer cannot have insurance before they are born. While it could be argued that "Age" was the incorrectly computed variable, it does show correlations that make sense with "MonthSal" and "Children," and the relationship between "Age" and "EducDeg" appears logical during the consistency check. Based on all of these factors, we have decided to exclude "Fidelity" from our analysis at this time.

We will also not be using "GeoLivArea" in our analysis because it appears to have limited discriminatory power. During simple comparisons with other variables, we found that it does not vary in the way we would expect. For example, when we compare it with "Children" (*Figure 16*), we would expect certain regions to have a higher proportion of “Children” compared to other regions, but this

is not the case. The same is true when comparing it with "EducDeg" (*Figure 17*) , “Annual_Salary” (*Figure 18)*, and “PremHousehold” (*Figure 19*) as there are no significant variations across the different regions. After examining the correlation matrix, we decided to continue using "Children" and "Education" as they are correlated with several other variables.

3. Segmentation

We separated our variables into three segments based on the initial variables and those generated during the feature engineering process. The goal is to use various clustering techniques on each segment and determine which method works best in each, providing us with more options for analysis. The segments created are presented below:

“Demographic”: Includes "Age", "Children", "EducDeg", and "Annual_Salary". These variables are characteristics that describe a population, with "Age" and "Children" reflecting the age structure of a population and "EducDeg" and "Annual_Salary" indicating socio-economic status. "Age" and "Annual_Salary" being highly correlated could potentially harm the clustering algorithm, so we decided to cluster removing one of them at the time. Regardless of which one we remove, we will always include it again when profiling the clusters for this segment. As we are using distance-based clustering methods, we created two datasets for each configuration, one using min max scaling and the other using standard scaling. This resulted in a total of four datasets to apply the clustering algorithms.

“Insurance Package”: Comprises "PremMotor", "PremHousehold", "PremHealth", "PremLife" and "PremWork." These variables describe the insurance package that each customer subscribes to and are important in understanding the insurance preferences of each customer. We created a variation of this segment using the ratios of these variables: "Motor_Ratio", "Household_Ratio", "Health_Ratio", "Life_Ratio", and "Work_Ratio". For both variations, we did not cluster with "PremMotor" (or "Motor_Ratio") as it is highly correlated with the other premiums (or other ratios). However, we will include it again when profiling the clusters for this segment. As with the other segments, we created two datasets for each configuration, one using min max scaling and the other using standard scaling. We ended up with a total of four datasets to use for the clustering algorithms.

“Customer Importance”: Contains "CustMonVal", "ClaimsRate", "Total_Premium", "Cessation", and "Insurance_Burden". These variables describe the type of customer in terms of profitability for the company and their insurance behaviour. "CustMonVal" and "ClaimsRate" are highly correlated, so we decided to cluster using only one of them at a time, but always including both when profiling the cluster for this segment. As with the other segments, we created two datasets for each configuration, one using min max scaling and the other using standard scaling, resulting in a total of four datasets for the clustering algorithms.

4. Clustering by Segment

For this phase, we will be using a range of clustering algorithms on the datasets created for each segment. We have chosen Hierarchical Clustering, K-Means, and DB-Scan as we think they will be the most effective with our datasets. We are testing multiple algorithms because each has its own pros and cons, and the ultimate goal is to choose the algorithm and dataset that offers the most comprehensive profiling of each segment.

4.1. Demographic Segment Clustering

The best results for demographic segmentation were achieved by using the K-Means Clustering method on the "EducDeg", "Annual_Salary", and "Children" data with standard scaling. Although "Age" was not directly included in the clustering analysis, it was considered when evaluating the resulting cluster profiles. One of the aspects of K-Means is that the number of clusters must be determined beforehand. To determine the optimal number of clusters, we applied the elbow method to the inertia plot (*Figure 20*) and found that three clusters was the best number of clusters. The silhouette plot (*Figure 21*) for three clusters showed that the clusters had similar sizes, high silhouette scores, and almost no tail, with an average silhouette score of 0.41 (*Figure 22*).

By looking at the histograms for each variable within each cluster (*Figure 23*), we can determine the demographic characteristics of the three clusters and classify the customers into three groups based on these characteristics:

- ***Older Professional*:** Cluster 0, which is depicted in *Figure 23*, consists primarily of older customers who do not have children, have moderate levels of education, and high salaries. It is noteworthy that the data indicates that it is more probable for an older person without children to have insurance. Composed by 3706 observations.

- ***Mid-Career Parent*:** Cluster 1, also depicted in *Figure 23*, consists primarily of mid-age individuals with children, strong levels of education, and average salaries. Composed by 3198 observations.

- ***Lower-Educated Parent:*** Cluster 2, also depicted in *Figure 23,* consists primarily of mid-age individuals with children, low levels of education, and average salaries. Composed by 3051 observations.

DB-Scan was also effective in grouping the clusters for the demographic segment, as it took into account the categorical variables and grouped the clusters based on density. Initially, there were many clusters, but when the most similar ones were combined, the results were satisfactory. However, the clusters from the K-Means method were found to have better distributions for each cluster.

4.2. Insurance Package Segment Clustering

The best results for insurance package segmentation were achieved by using the Hierarchical Clustering method on the “PremHousehold”, “PremHealth”, “PremLife”, and “PremWork” data with standard scaling. Although “PremMotor” was not directly included in the clustering analysis, it was considered when evaluating the resulting cluster profiles. One of the features of Hierarchical Clustering is that the linkage method and the number of clusters must be specified beforehand as parameters. A R^2 plot (*Figure 24*) was constructed with a range of different numbers of clusters and a line for each linkage method to determine the best linkage method. It can be seen that for all numbers of clusters, the "ward" method was the best, as the higher R^2 shows. The "ward" method is known for clustering that aims to minimize distances within each cluster as its foundation. To determine the number of clusters, a dendrogram was plotted, and the optimal number of clusters was determined to be 3 by examining the defined threshold.

By looking at the histograms for each variable within each cluster (*Figure 26*), we can determine the insurance package characteristics of the three clusters and classify the customers purchase pattern into three groups based on these characteristics:

- **Complete Package:** Cluster 0, shown in *Figure 26*, is a well-rounded package that has higher premiums in the categories of "PremWork", "PremLife", and "PremHousehold". However, the premiums in the other categories are not particularly low. Out of the three packages, it is the most expensive and comprehensive. Composed of 2535 observations.

- **Vehicle Focused Package:** Cluster 1, shown in *Figure 26*, is a all-rounded package that has "PremMotor" as the highest premium, and presents relatively low premiums for the other coverage types. Among the three packages, this one is particularly specialized in vehicle insurance. Composed of 4024 observations.

- **Medical Coverage Package:** Cluster 2, shown in *Figure 26*, is an all-rounded package that has "PremHealth" as highest premium among the three packages, but offers relatively low premiums for other coverage types. This package is particularly focused on health insurance. Composed of 3707 observations.

During the clustering for this segment, we attempted to utilize the ratios previously described to cluster, but none of the methods produced as accurate profiles as using the standard premiums.

4.3. Customer Importance Segment Clustering

The best results for customer importance segmentation were achieved by using the Hierarchical Clustering method on "ClaimsRate", "Total_Premium", and "Insurance_Burden" data with standard scaling. Although "CustMonVal" was not directly included in the clustering analysis, it was considered when evaluating the resulting cluster profiles. We used the same approach as before and looked at the R2 plot (*Figure 27*). This showed that the "ward" linkage method was the best choice. Additionally, examining the dendrogram (*Figure 28*) indicated that 3 clusters were optimal based on our threshold.

By looking at the histograms for each variable within each cluster (*Figure 29*), we can determine the customer characteristics of the three clusters and classify the customers importance into three groups based on these characteristics:

- **Premium Client:** Cluster 0, depicted in *Figure 29*, consists of the highest spending clients on the insurance company. They have the highest "CustMonVal" among the three types of clients. While these clients have a high "ClaimsRate", the "Total_Premium" they pay compensates for it. These clients save a significant portion of their yearly salary for insurance purchases, as indicated by the high values in "Insurance_Burden." Additionally, their low "Cessation" rate suggests that they are unlikely to cancel any of their policies. Composed of 2268 observations.

- **Non-Profitable Client:** Cluster 1, depicted in *Figure 29*, consists of clients who do not generate profit for the company, as indicated by the sometimes negative and low "CustMonVal". This is likely due to the small "Total_Premium" they pay and their high "ClaimsRate". Additionally,

their higher “Cessation” rate suggests that they are more likely to cancel their policies. Composed of 4854 observations.

- **Stable Client:** Cluster 2, depicted in *Figure 29*, consists of highly profitable clients for the company. They pay an acceptable “Total_Premium” and have a very low “ClaimsRate”. Although these clients are not more profitable than those in Cluster 0, they still generate a good profit for the company. However, their lower “ClaimsRate” indicates that they use their insurance less frequently, which is reflected in their higher “Cessation rate” - they are more likely to cancel one of their policies. Composed of 4833 observations.

5. Final Clusters

By considering the three different clusters for each segment, we obtained 27 clusters in total when we cross-referenced them in a table (*Figure 30*). Given the impracticality of creating a marketing strategy for 27 distinct customer groups, we will need to generalize and combine some of the clusters into larger groups. To merge clusters that are similar, we will utilize Hierarchical Clustering again. Examining the dendrogram, we can see that six clusters is the optimal number based on our defined threshold (*Figure 31)*.

In *Figures 32 and 33*, we present the centroids and distributions of the six final clusters. U-Map Dimensionality Reduction available on *Figure 34*. We will then describe each cluster and propose a marketing strategy for each one of them:

- **Older Professional / Health Insurance / Mildly Profitable:** Cluster 0 composed of older individuals who do not have children. They have a median level of education and a high annual salary. Their motor insurance premiums are considerable, possibly due to owning a high-value vehicle. However, their household and work insurance premiums are relatively low, as many of them may be retired. They have a high health insurance premium, as is common for older individuals, and a relatively low life insurance premium. Overall, this segment is mildly profitable, with a moderate total premium and claims rate. They rarely cancel their insurance and only dedicate a small portion of their salary to insurance premiums. Composed by 2307 observations.

- **Educated Parent / Health Insurance / Mildly Profitable:** Cluster 1 consists mostly of younger individuals, many of whom have children. They have a median level of education and a moderate to low annual salary. Their motor insurance premiums are considerable, possibly due to a lack of driving experience. Their household and work insurance premiums are relatively low, while their health insurance premiums are high. This segment has a moderate total premium and claims rate, making them mildly profitable. They do not frequently cancel their insurance, but they only dedicate a small portion of their salary to insurance premiums, likely due to their lower income requiring them to allocate funds elsewhere. Composed by 1851 observations.

- **Youngest Parent / Household Insurance / Highly Profitable:** Cluster 2 is the youngest of the groups. Many of them have children, while others may be planning to have children in the future. They have a low level of education and a low annual salary, possibly due to being in university. They have a low motor insurance premiums, possibly because they do not own a

car or have a low-value vehicle. Their household insurance premiums are high, while their health insurance premiums are relatively low. They also have low to well-rounded work and life insurance premiums, which, may mean they have riskier jobs, justifying their high claims rate. Despite this, they are profitable, with a high total premium that is largely derived from their high household insurance premiums. They never cancel their insurance. Composed by 1184 observations.

- **Middle Age Parent / Motor Insurance / Highly Profitable:** Cluster 3 is composed of middle-aged individuals, many of whom have children. They have a median to high level of education and a median annual salary. Their motor insurance premiums are very high, possibly due to owning a high-value vehicle. Their household insurance premiums are low, as are their health, life, and work insurance premiums. This segment is highly profitable, with a moderate total premium and low claims rate. They only dedicate a small portion of their salary to insurance premiums and frequently cancel their insurance, possibly because they feel they do not need the insurance. Composed by 589 observations.

- **Young Parent / Well-Rounded / Low Profit:** Cluster 4 is composed of young individuals, many of whom have children or are planning to have children in the future. They have a low to median level of education, possibly due to being in university. They have a low annual salary and low to well-rounded motor insurance premiums. Their household insurance premiums are low, while their health insurance premiums are moderate. They have well-rounded life and work insurance premiums. However, they are not very profitable, with a high claims rate and moderate total premium. They rarely cancel their insurance and only dedicate a small portion of their salary to insurance premiums. Composed by 1874 observations.

- **Middle Age Parent / Motor Insurance / Unprofitable:** Cluster 5 is composed of middle-aged individuals, many of whom have children. They have a median to high level of education and a median annual salary. Their motor insurance premiums are very high, possibly due to owning a high-value vehicle as a result of a mid-life crisis. Their household insurance premiums are low, as are their health, life, and work insurance premiums. However, this segment is unprofitable or possibly break-even, with a very high claims rate and moderate total premiums. They frequently cancel their insurance, possibly due to having their coverage cancelled by the insurance company, and only dedicate a small portion of their salary to insurance premiums. Composed by 2150 observations.

6. Marketing Strategies

Older Professional / Health Insurance / Mildly Profitable (Cluster 0): The target audience for this insurance product is older people who have a lot of money and are loyal to their insurance company, but prefer more traditional methods. They may be dropped by other insurance companies as they age and are more likely to become sick. A good strategy for reaching this audience would be to advertise a full coverage health insurance product with no caveats at a high price through their insurance broker or through conventional mail. This group values personal contact and takes their mail seriously.

Educated Parent / Health Insurance / Mildly Profitable (Cluster 1): We should target young people with families who have some money to spend on insurance but may not have a high annual salary. A good option for them could be a low-cost life insurance policy. This would not be too expensive for

them, so they may still be interested. We could advertise this through online and social media ads, focusing on the idea of protecting their families.

Youngest Parent / Household Insurance / Highly Profitable (Cluster 2): These people have already committed a high chunk of their annual salary to insurance premiums, it's unlikely they will be willing to spend more money on insurances. The strategy for this group of people would be not to target, to keep the company from spending money unnecessarily (without any return).

Middle Age Parent / Motor Insurance / Highly Profitable (Cluster 3): These people rarely activate their insurance, a good practice could be to lower their premiums. They have income to spare, and many probably don't even have a household insurance, even though most probably have a house of their own having kids in mind, a median salary and are middle-aged. It would be a good practice to target these to them (probability making a low-cost variant available for them).

Young Parent / Well-Rounded / Low Profit (Cluster 4): These people are not very profitable but they still have some Money to spend, a good practice would be to raise their premiums. Maybe targeting them with household insurance ads on social media could be a good idea. They're probably thinking about buying their first home soon and the advertising for their insurance could trigger that decision, making them remember that trigger (most likely would come to insure their house on our company).

Middle Age Parent / Motor Insurance / Unprofitable (Cluster 5): Given their low profit profile, a good business practice would be to raise the premiums on most of these customers to make them profitable. An alternative could be to end our relationship with them, but our company would always rather find them a solution to keep in business with us, even if it is a more expensive one. They still have some Money to spend. Billboard and tv ads are still relatively popular with this age gap, so one idea could be to target them with a more complete (a little more expensive) health insurance. This ad campaign should always be developed having in mind the goal of transmitting the idea the audience should be starting to plan for their future (and our product will fill that need).

6. Conclusion

To conclude, our group has successfully met the objectives outlined in the beginning of the project. We analysed and interpreted data through multiple phases to identify 3 main types of customer segments, which were then clustered into 27 unique groups. By reducing these clusters based on similarity criteria, we were able to identify 6 distinct customer segments and create personalized marketing strategies for each segment. These targeted approaches will not only reduce costs for the company, but also increase revenue by providing more effective marketing efforts. Implementing these strategies will significantly improve the success of A2Z Insurance.

7. References

[1] https://pandas.pydata.org/docs/user_guide/10min.html (accessed january 1, 2023).

[2] https://seaborn.pydata.org/tutorial/introduction (accessed january 3, 2023).

[3] https://matplotlib.org/2.0.2/users/pyplot_tutorial.html (accessed january 2, 2023).

[4] https://docs.python.org/3/library/math.html (accessed january 2, 2023).

[5] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.preprocessing (accessed january 2, 2023).

[6] https://arrow.apache.org/docs/python/index.html (accessed january 4, 2023).

[7] https://docs.python.org/3/library/datetime.html (accessed january 6, 2023).

[8] https://pypi.org/project/sas7bdat/ (accessed january 4, 2023).

[9] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.decomposition (accessed january 5, 2023).

[10] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.impute (accessed january 5, 2023).

[11] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.mixture (accessed january 5, 2023).

[12] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.neighbors (accessed january 6, 2023).

[13] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.cluster (accessed january 2, 2023).

[14] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.manifold (accessed january 3, 2023).

[15] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.metrics (accessed january 3, 2023).

[16] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.linear_model (accessed january 3, 2023).

[17] https://scikit-learn.org/stable/modules/classes.html#module-sklearn.tree (accessed january 6, 2023).

[18] https://matplotlib.org/stable/tutorials/colors/colormaps.html (accessed january 6, 2023).

[19] https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html (accessed january 6, 2023).

[20] https://umap-learn.readthedocs.io/en/latest/basic_usage.html (accessed january 6, 2023).

7. Appendix

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10266.0	1991.062634	511.267913	1974.00	1980.00	1986.00	1992.0000	53784.00
BirthYear	10279.0	1968.007783	19.709476	1028.00	1953.00	1968.00	1983.0000	2001.00
MonthSal	10260.0	2506.667057	1157.449634	333.00	1706.00	2501.50	3290.2500	55215.00
GeoLivArea	10295.0	2.709859	1.266291	1.00	1.00	3.00	4.0000	4.00
Children	10275.0	0.706764	0.455268	0.00	0.00	1.00	1.0000	1.00
CustMonVal	10296.0	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	0.742772	2.916964	0.00	0.39	0.72	0.9800	256.20
PremMotor	10262.0	300.470252	211.914997	-4.11	190.59	298.61	408.3000	11604.42
PremHousehold	10296.0	210.431192	352.595984	-75.00	49.45	132.80	290.0500	25048.80
PremHealth	10253.0	171.580833	296.405976	-2.11	111.80	162.81	219.8200	28272.00
PremLife	10192.0	41.855782	47.480632	-7.00	9.89	25.56	57.7900	398.30
PremWork	10210.0	41.277514	51.513572	-12.00	10.67	25.67	56.7900	1988.70

Figure 1 – Initial Descriptive Statistics

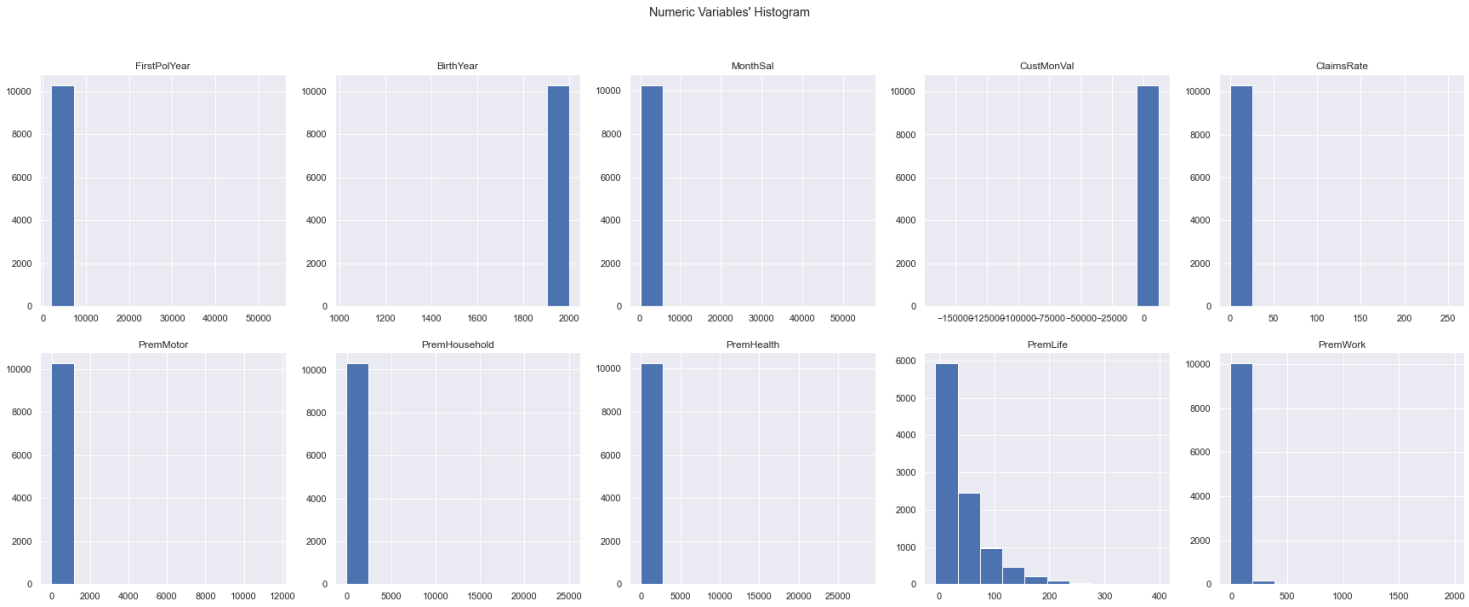


Figure 2 – Initial Histograms

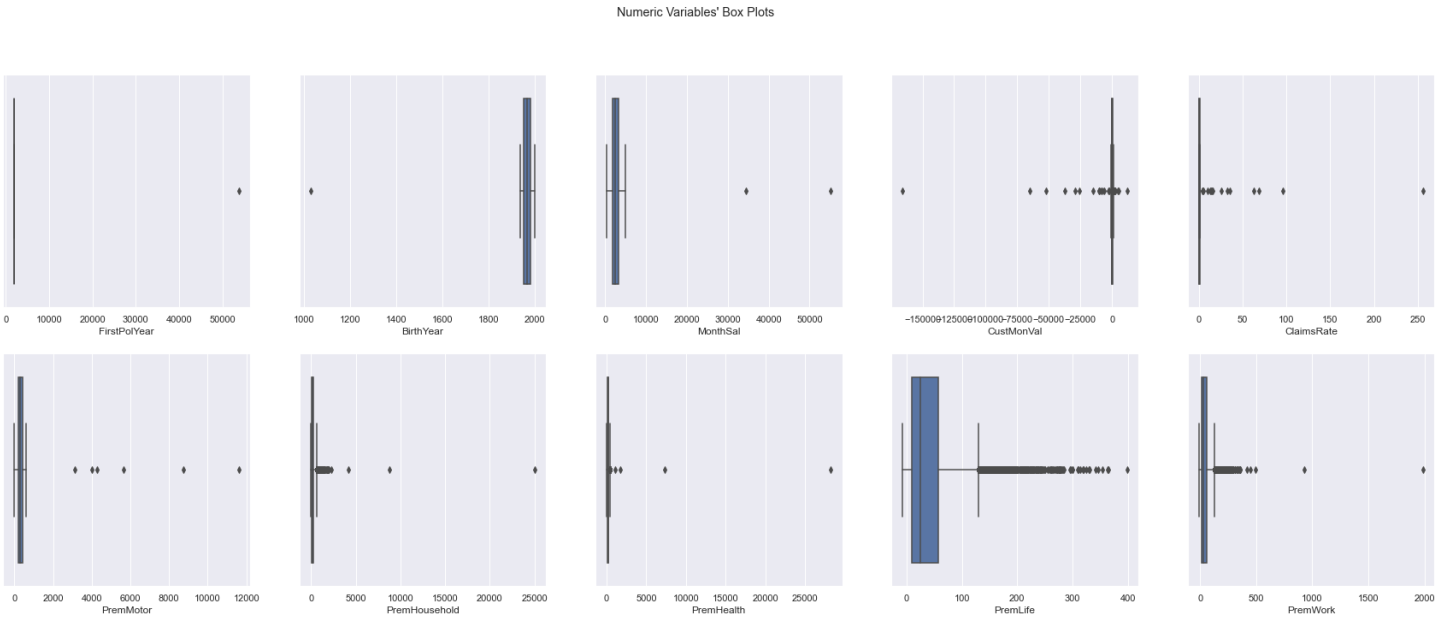


Figure 3 – Initial Box Plots

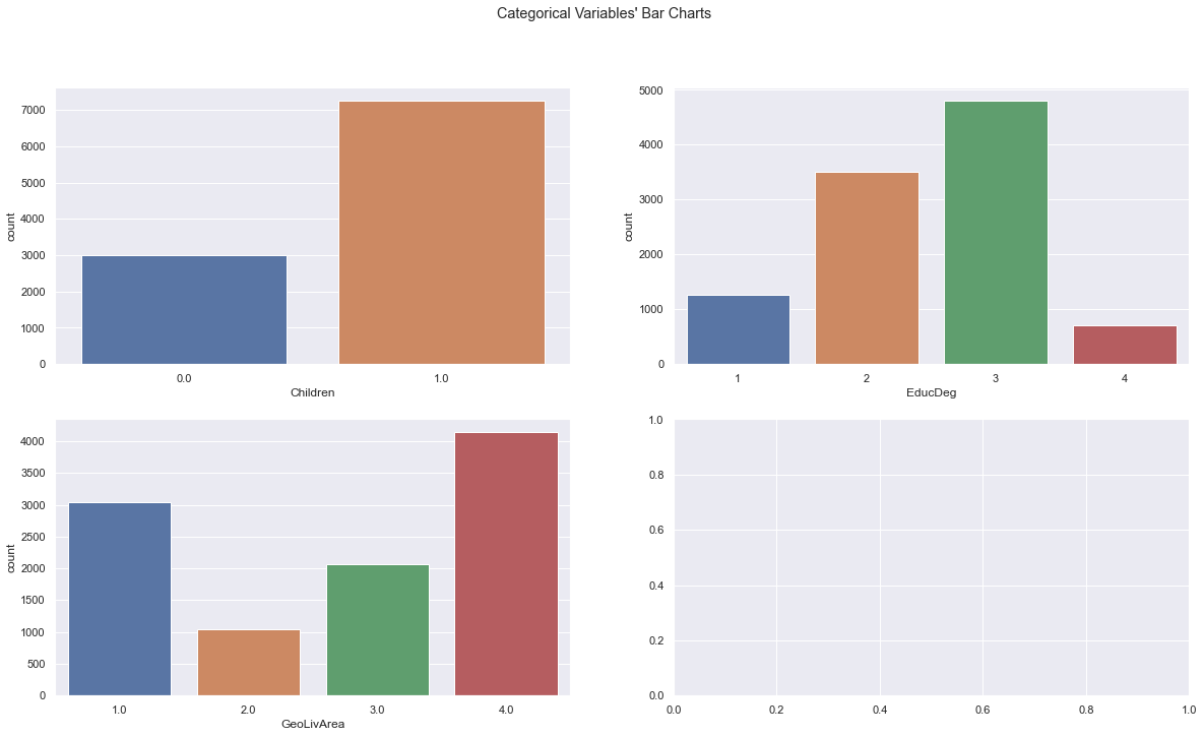


Figure 4 – Initial Bar Charts

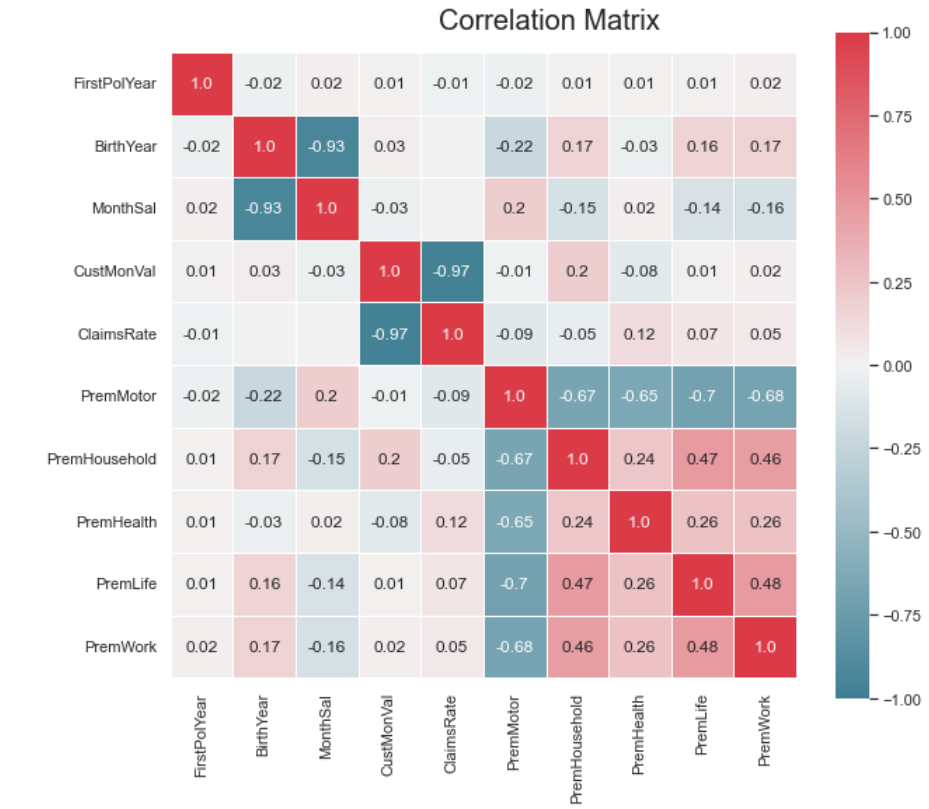
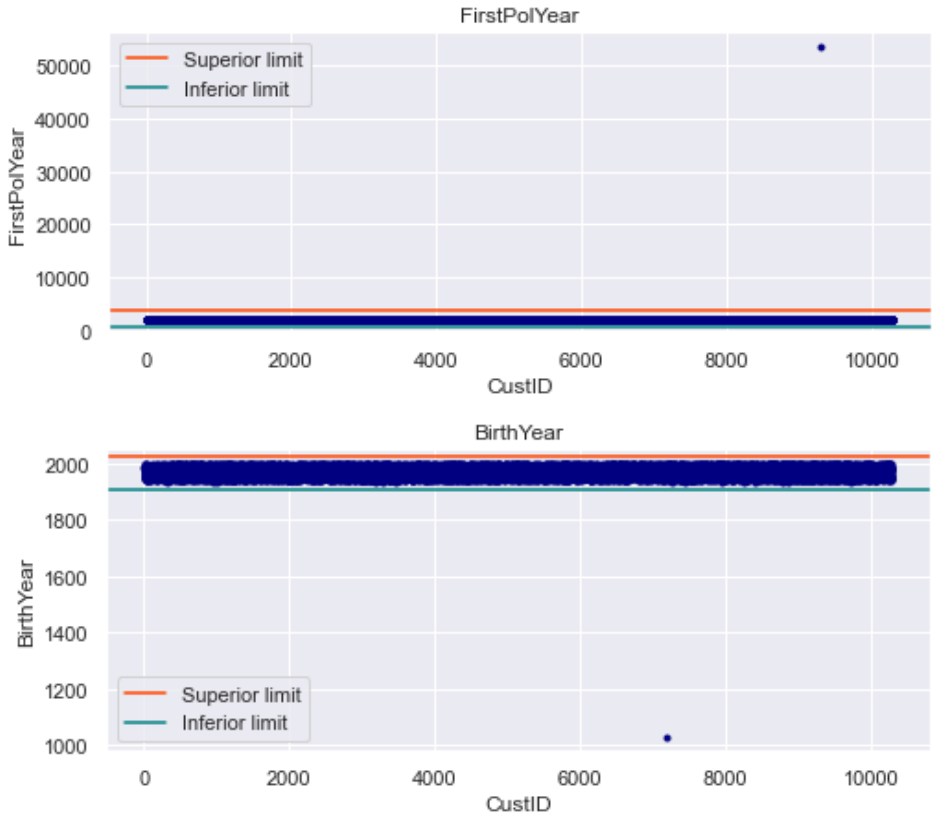
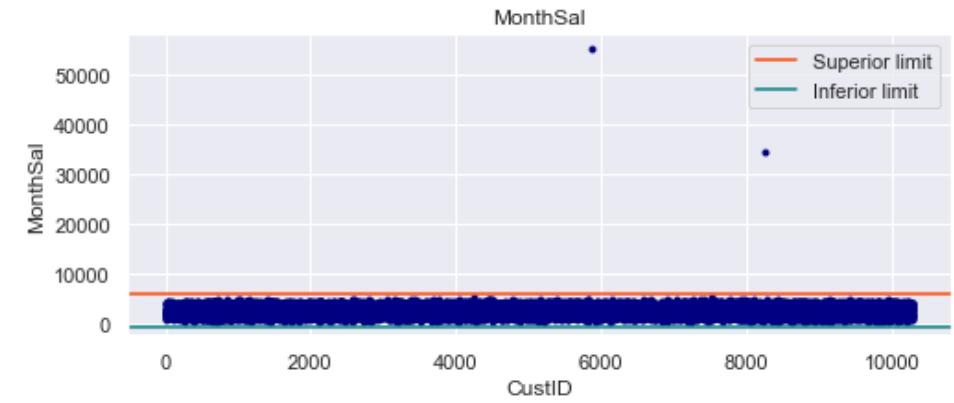
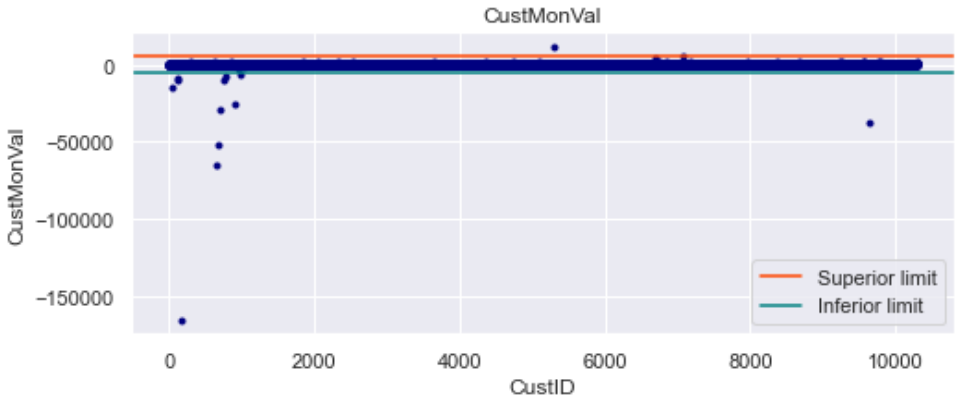
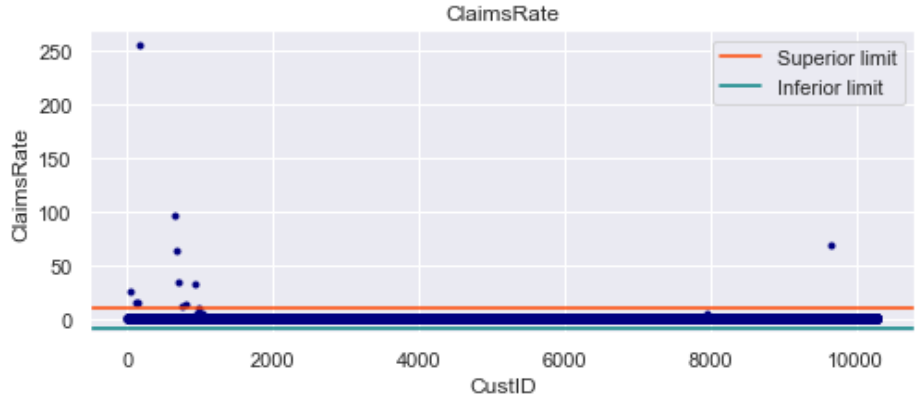


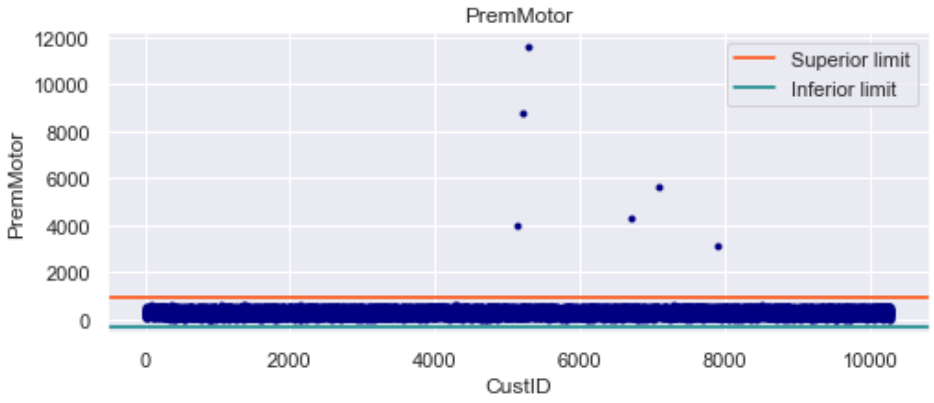
Figure 5 – Initial Correlation Matrix




MonthSal
Superior limit
Inferior limit
MonthSal
50000
40000
30000
20000
10000
0
0
2000
4000
6000
8000
10000
CustID


CustMonVal
Superior limit
Inferior limit
CustMonVal
0
−50000
−100000
−150000
0
2000
4000
6000
8000
10000
CustID


ClaimsRate
Superior limit
Inferior limit
ClaimsRate
250
200
150
100
50
0
0
2000
4000
6000
8000
10000
CustID


PremMotor
Superior limit
Inferior limit
PremMotor
12000
10000
8000
6000
4000
2000
0
0
2000
4000
6000
8000
10000
CustID

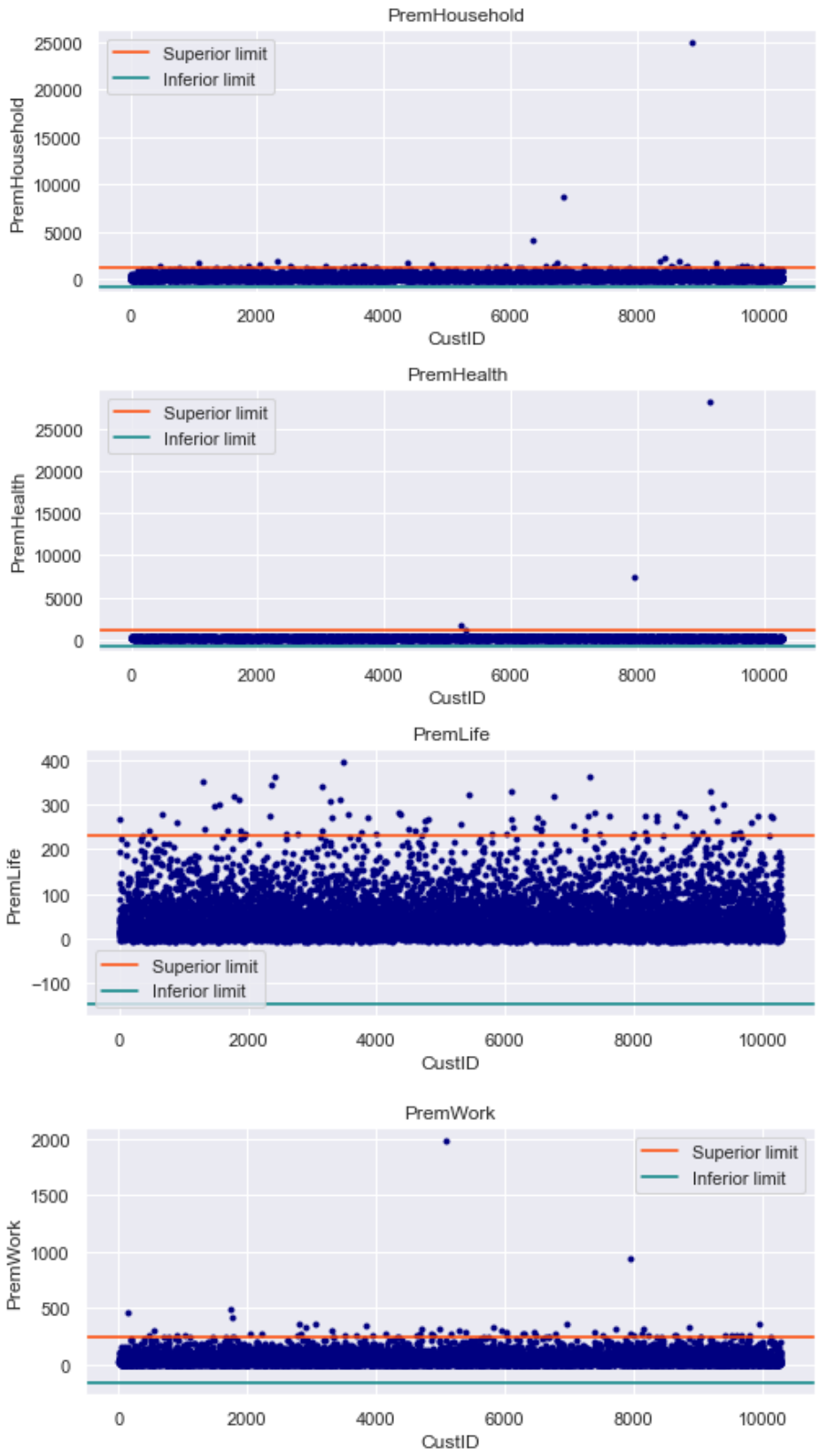


Figure 6 – Thresholds defining the Outliers

FirstPolYear	30
BirthYear	17
EducDeg	17
MonthSal	36
GeoLivArea	1
Children	21
CustMonVal	0
ClaimsRate	0
PremMotor	33
PremHousehold	0
PremHealth	42
PremLife	103
PremWork	86

Figure 7 – Missing Values per Feature

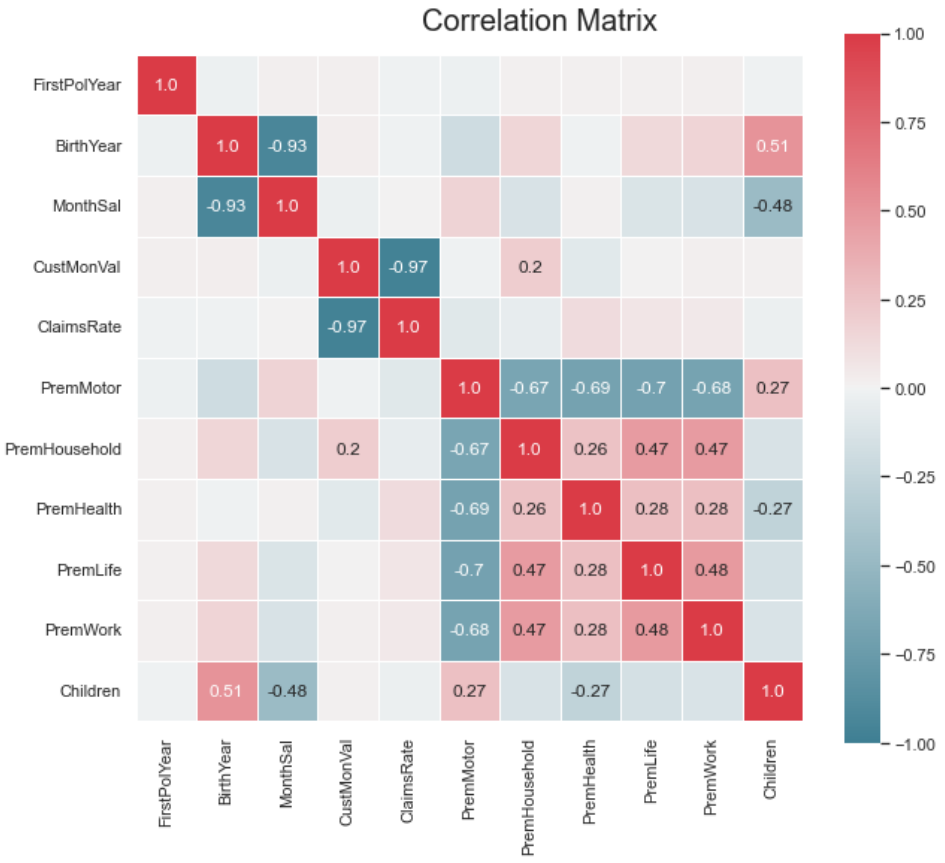


Figure 8 – Correlation Matrix after Outlier Removal

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10039.0	1986.020520	6.600185	1974.00	1980.00	1986.00	1992.000	1998.00
BirthYear	10039.0	1967.683136	17.147529	1935.00	1953.00	1967.00	1982.000	2000.00
EducDeg	10039.0	2.499153	0.784370	1.00	2.00	3.00	3.000	4.00
MonthSal	10039.0	2519.403626	973.963808	333.00	1743.00	2526.00	3299.000	5021.00
GeoLivArea	10039.0	2.709832	1.266837	1.00	1.00	3.00	4.000	4.00
Children	10039.0	0.707242	0.455051	0.00	0.00	1.00	1.000	1.00
CustMonVal	10039.0	216.371797	250.939089	-402.96	-8.44	187.03	397.915	1448.28
ClaimsRate	10039.0	0.680620	0.316793	0.00	0.39	0.72	0.980	1.62
PremMotor	10039.0	300.555821	135.931895	-4.11	197.26	302.39	409.300	585.22
PremHousehold	10039.0	200.230152	221.909347	-75.00	48.90	131.70	283.950	1255.25
PremHealth	10039.0	168.355873	74.649226	-2.11	112.02	163.81	219.930	442.86
PremLife	10039.0	39.326430	42.585699	-7.00	9.78	24.56	55.010	230.60
PremWork	10039.0	38.971053	42.621191	-12.00	9.89	25.34	54.790	242.60

Figure 9 – Descriptive Statistics after Outlier Removal and Filling Missing Values

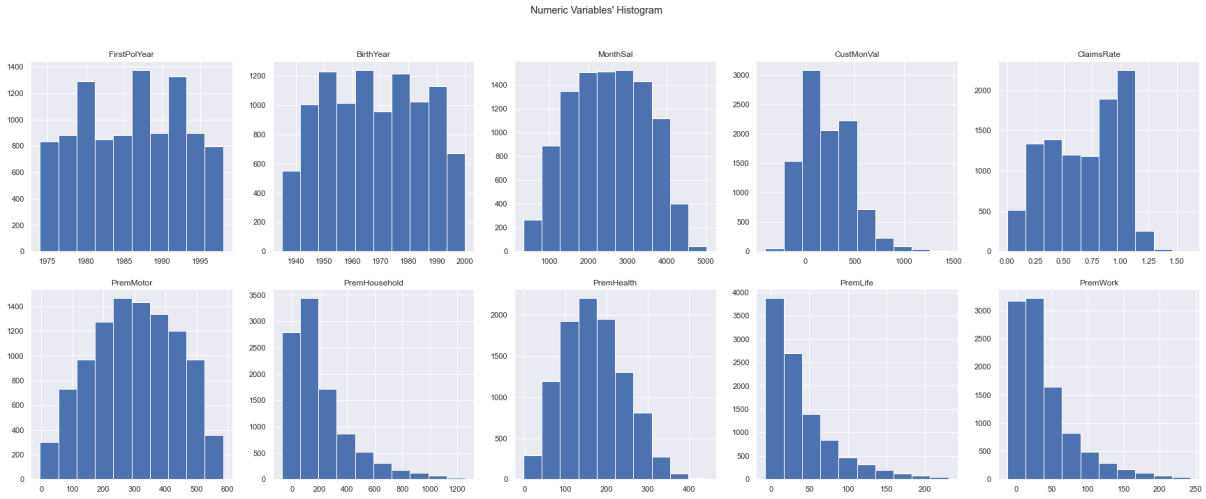


Figure 10 – Updated Histograms after Outlier Removal and Filling Missing Values

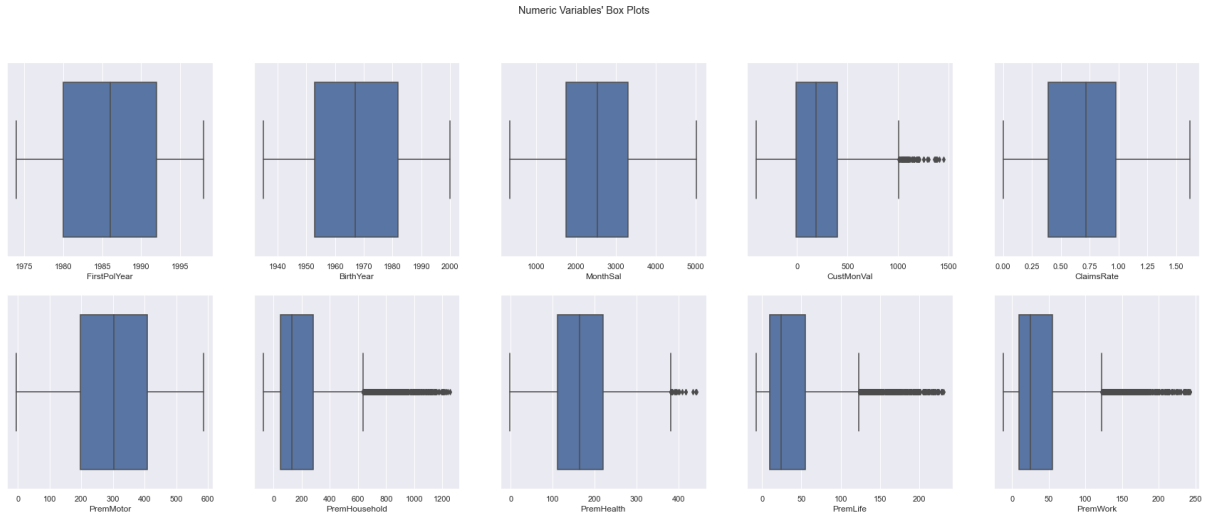


Figure 11 – Updated Box Plot after Outlier Removal and Filling Missing Vales

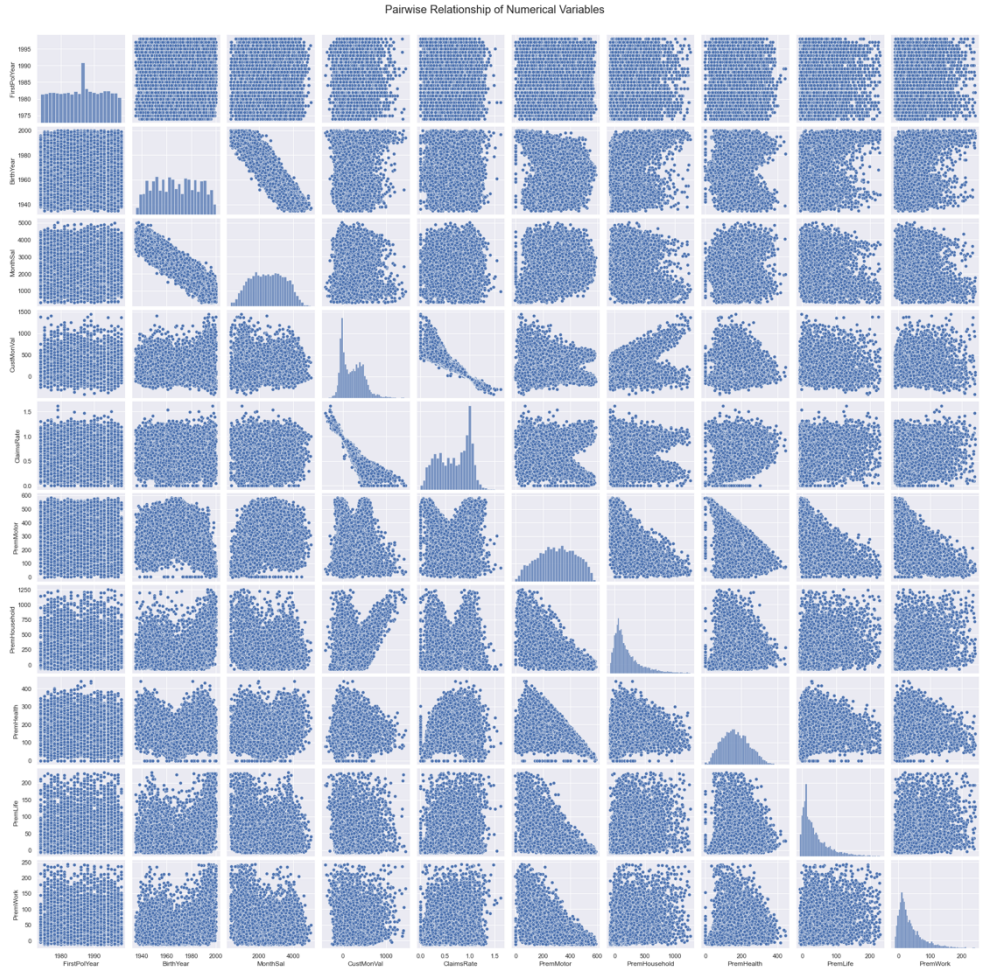


Figure 12 – Pairwise Relationship between all Variables

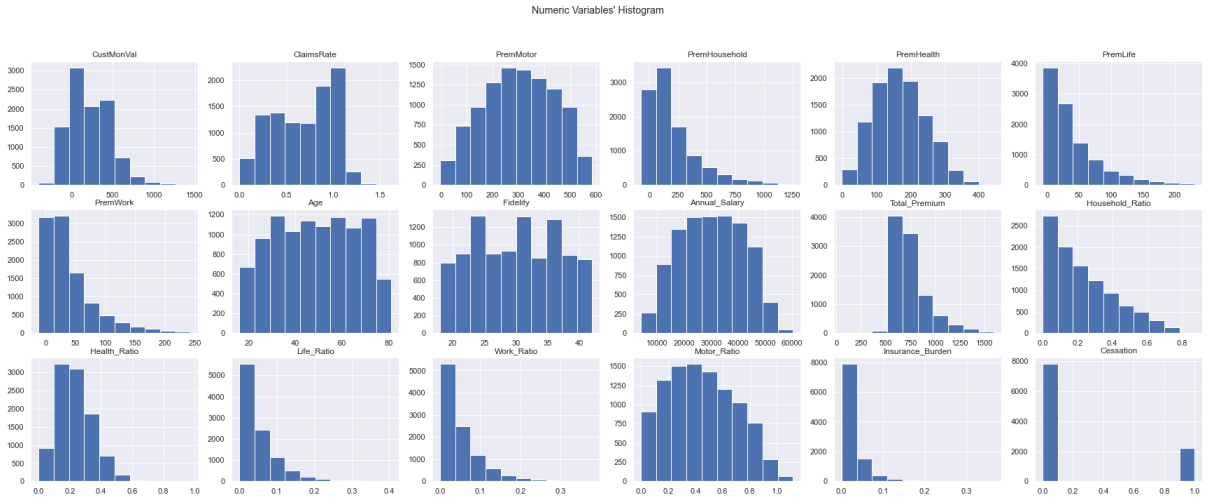


Figure 13 – Updated Histograms with Feature Engineering Features

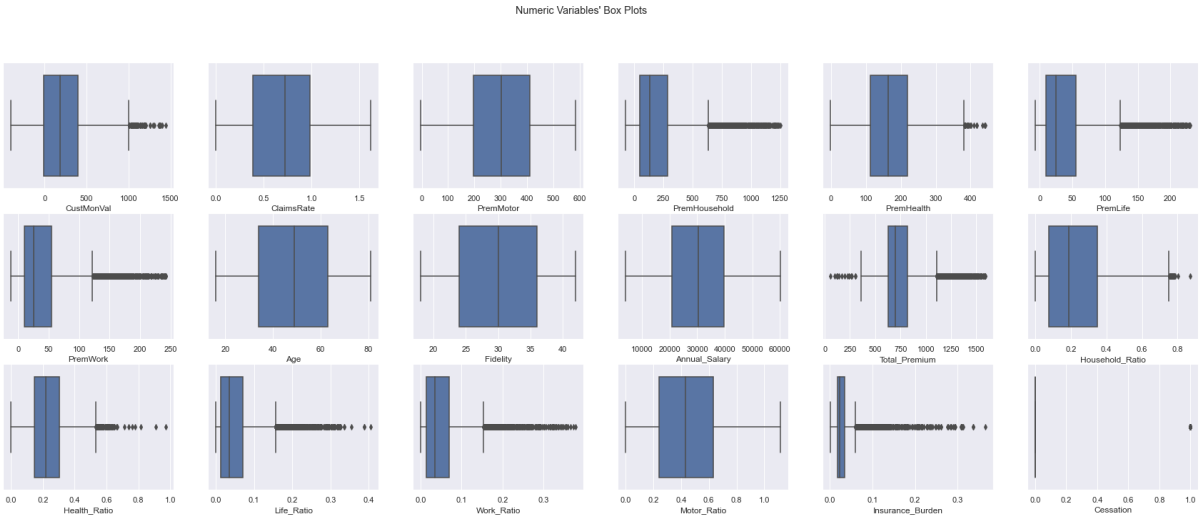


Figure 14 – Updated Box Plots with Feature Engineering Features

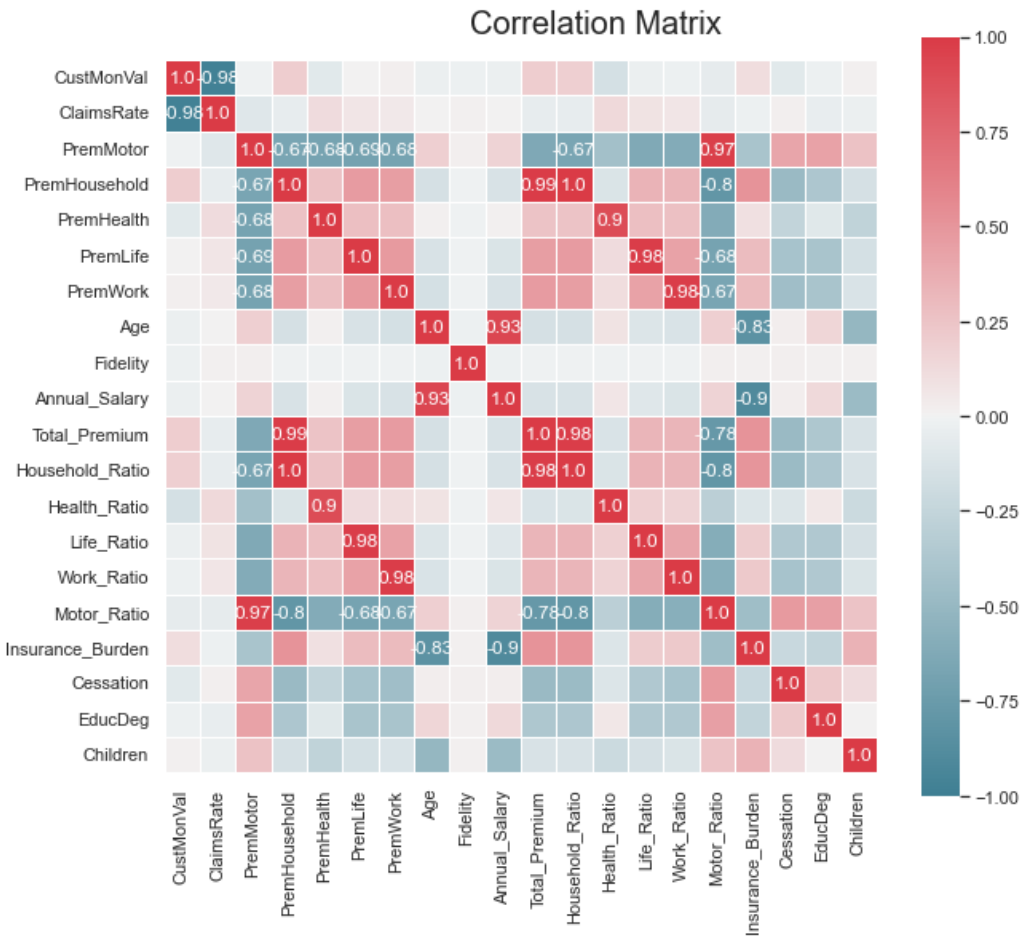


Figure 15 – Correlation Matrix with Feature Engineering Features

Children GeoLivArea	0	1
1	28.1%	71.9%
2	30.6%	69.4%
3	28.4%	71.6%
4	30.6%	69.4%

Figure 16 – Relative Proportion between "Children" and "GeoLivArea"

EducDeg GeoLivArea	1	2	3	4
1	11.5%	33.7%	47.7%	7.1%
2	11.7%	34.8%	46.8%	6.8%
3	11.2%	33.5%	48.7%	6.7%
4	11.5%	35.5%	46.6%	6.4%

Figure 17 – Relative Proportion between "EducDeg" and "GeoLivArea"

GeoLivArea	Annual_Salary
1	29910.610169
2	30908.103586
3	30056.794355

Figure 18 – "Annual_Salary" Mean by each "GeoLivArea"

GeoLivArea	PremHousehold
1	201.783726
2	195.667490
3	197.648554
4	201.509068

Figure 19 – "PremHousehold" Mean by each "GeoLivArea"

Figure 20 – Inertia Plot for demo_df1_standard using K-Means

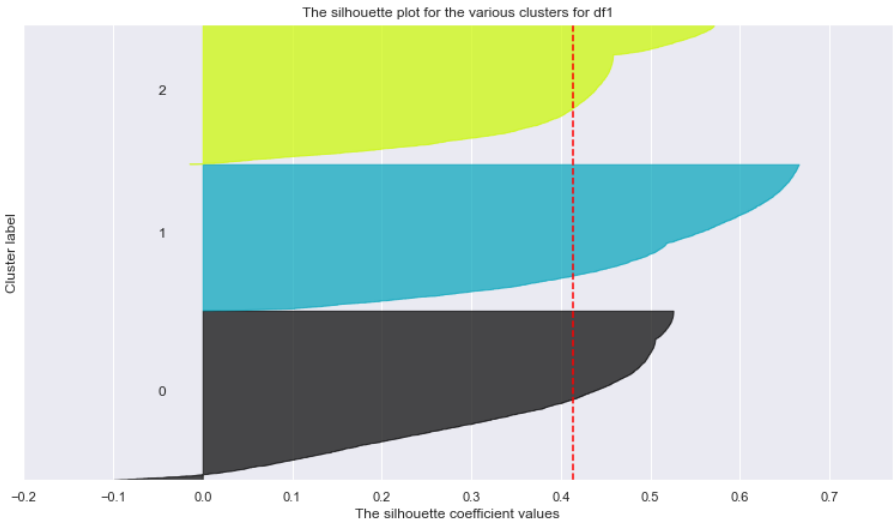


Figure 21 –Silhouette Plot for demo_df1_standard using 3 Clusters on K-Means

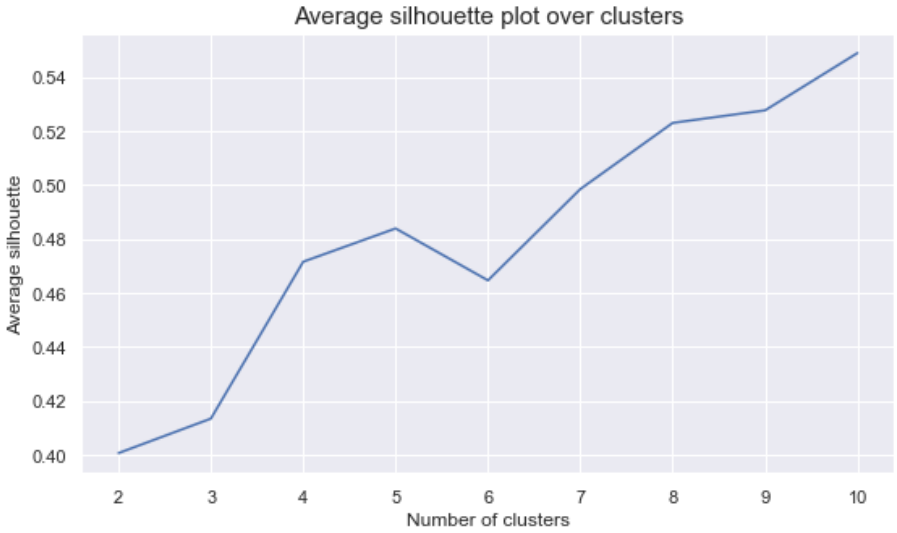


Figure 22 – Average Silhouette Plot for demo_df1_standard using K-Means

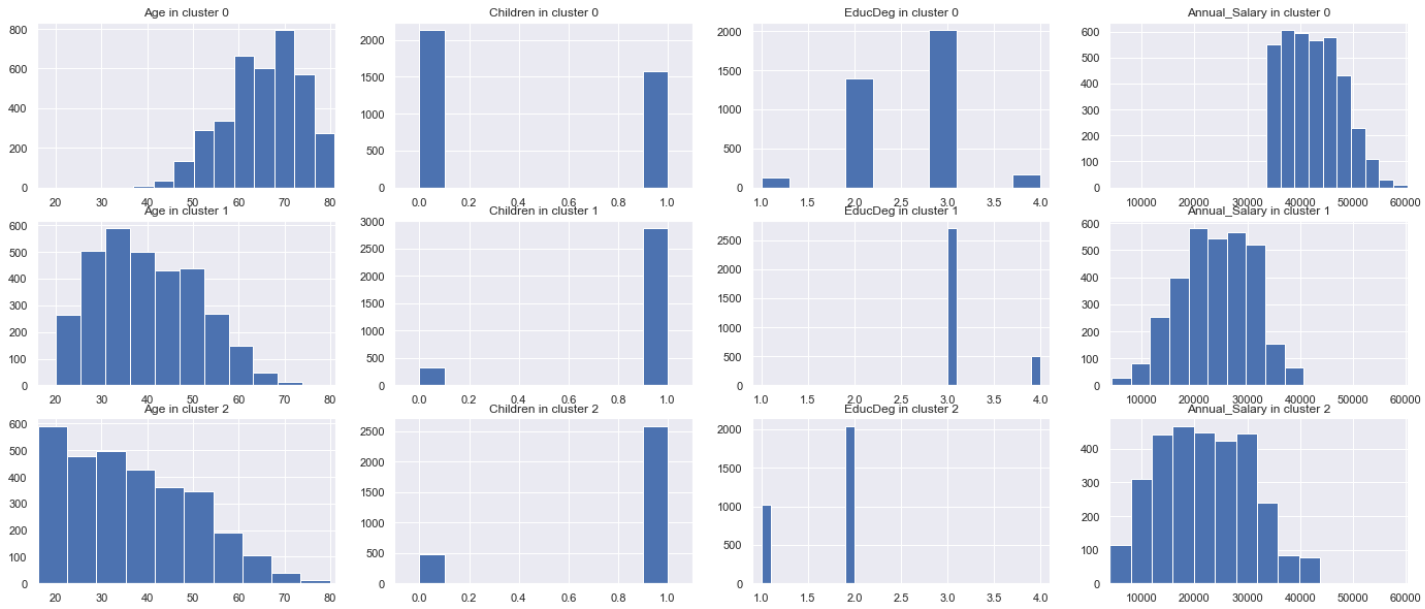


Figure 23 – Demographic Segment Clustering Variables' Histograms

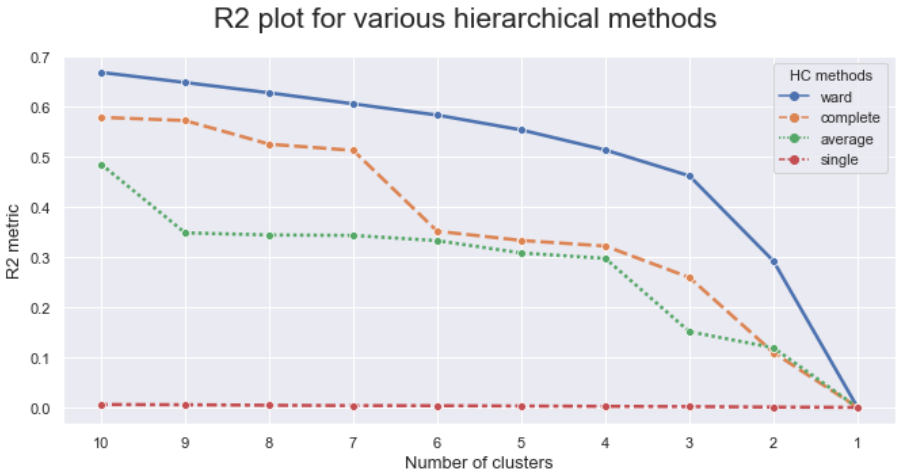


Figure 24 - R^2 Plot for pack_df1_standard changing the Linkage Method

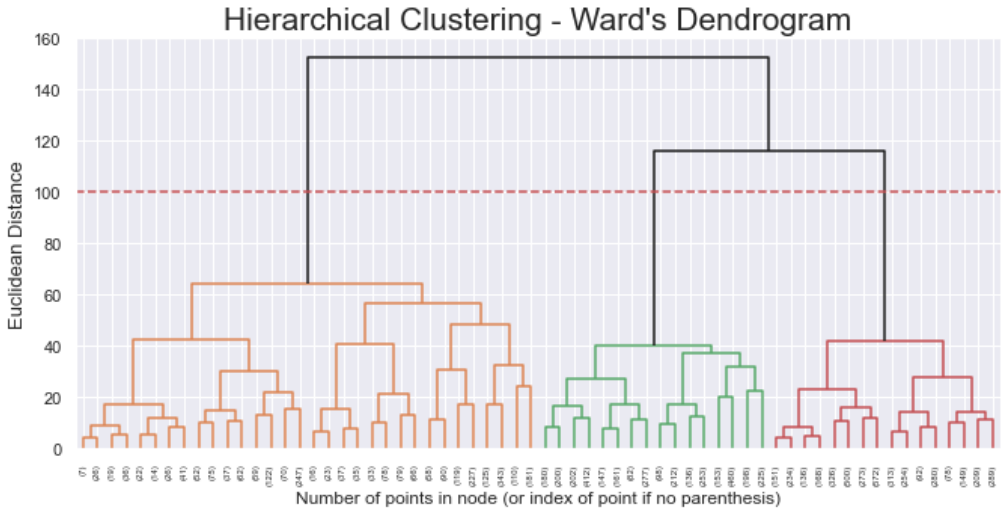


Figure 25 – Dendrogam for pack_df1_standard

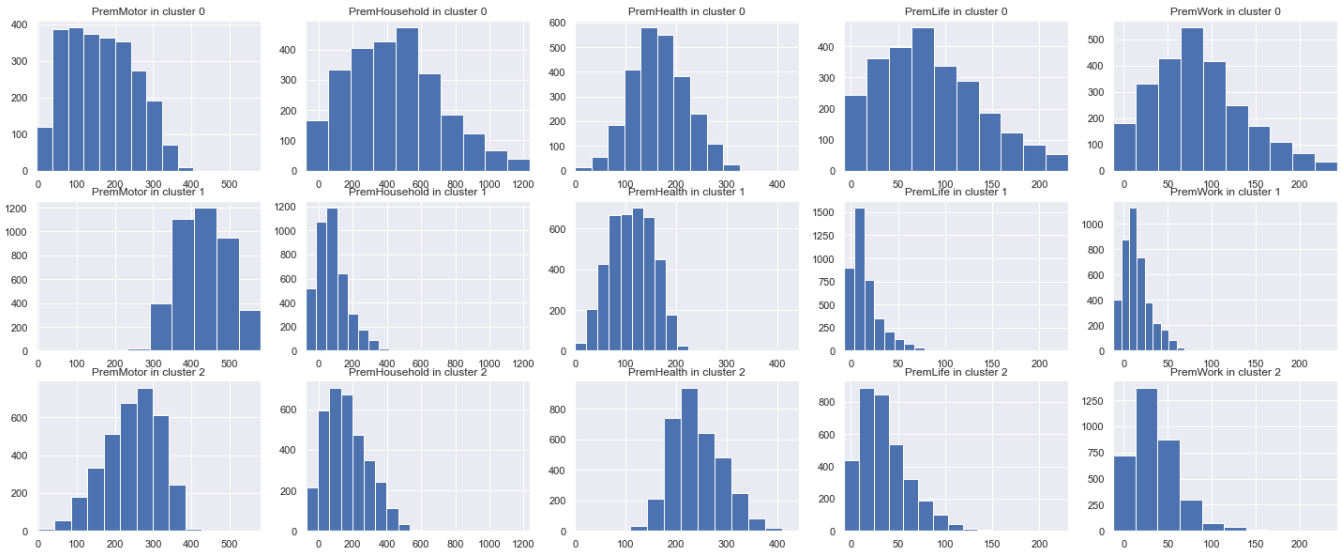


Figure 26 – Insurance Package Segment Clustering Variables' Histograms

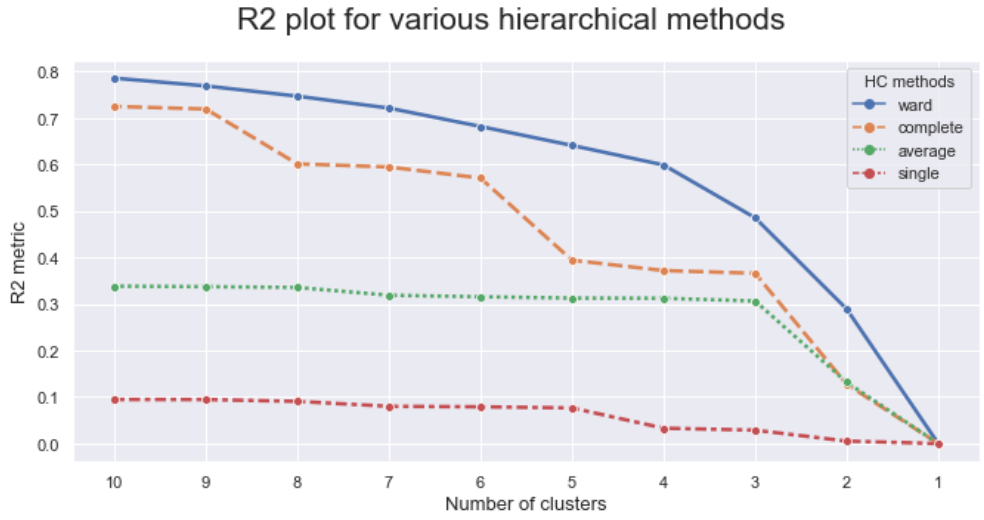


Figure 27 - R^2 Plot for import_df1_standard changing the Linkage Method

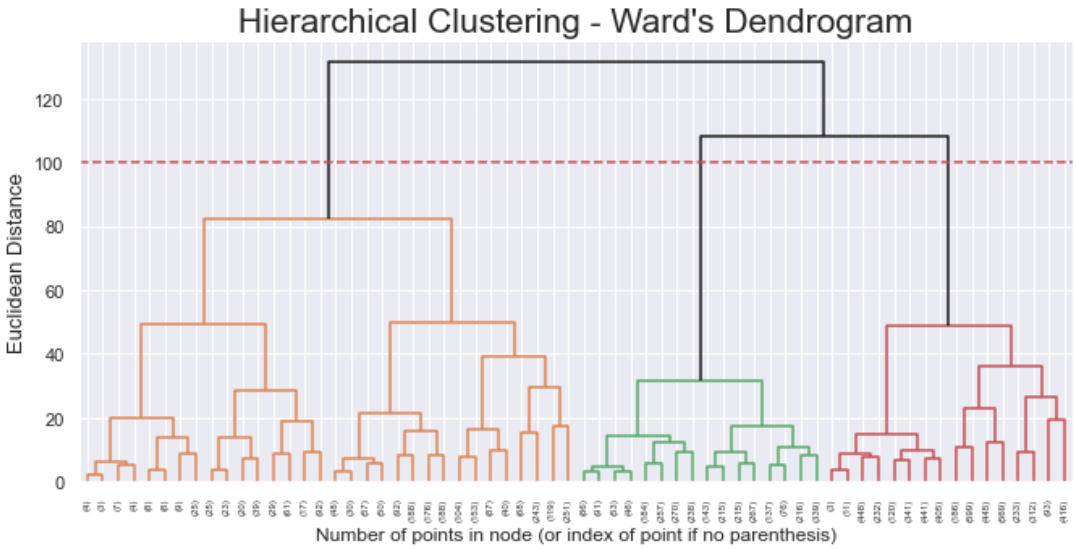


Figure 28 – Dendrogram for import_df1_standard

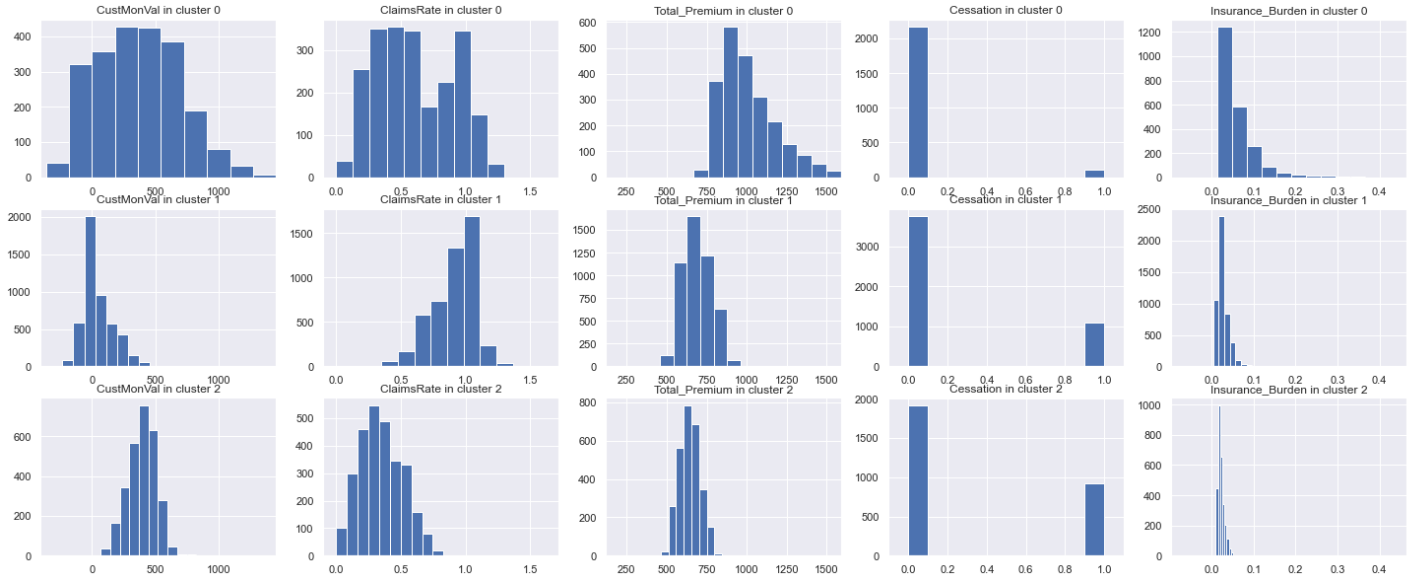


Figure 29 – Customer Importance Segment Clustering Variables' Histograms

									0
Package_Labels			0			1			2
Import_Labels	0	1	2	0	1	2	0	1	2
Demo_Labels									
0	444	221	97	62	747	590	205	869	471
1	205	116	37	47	965	821	128	609	270
2	979	360	76	43	438	311	155	529	160

Figure 30 – Cross Table for the Clusters from the Different Segments

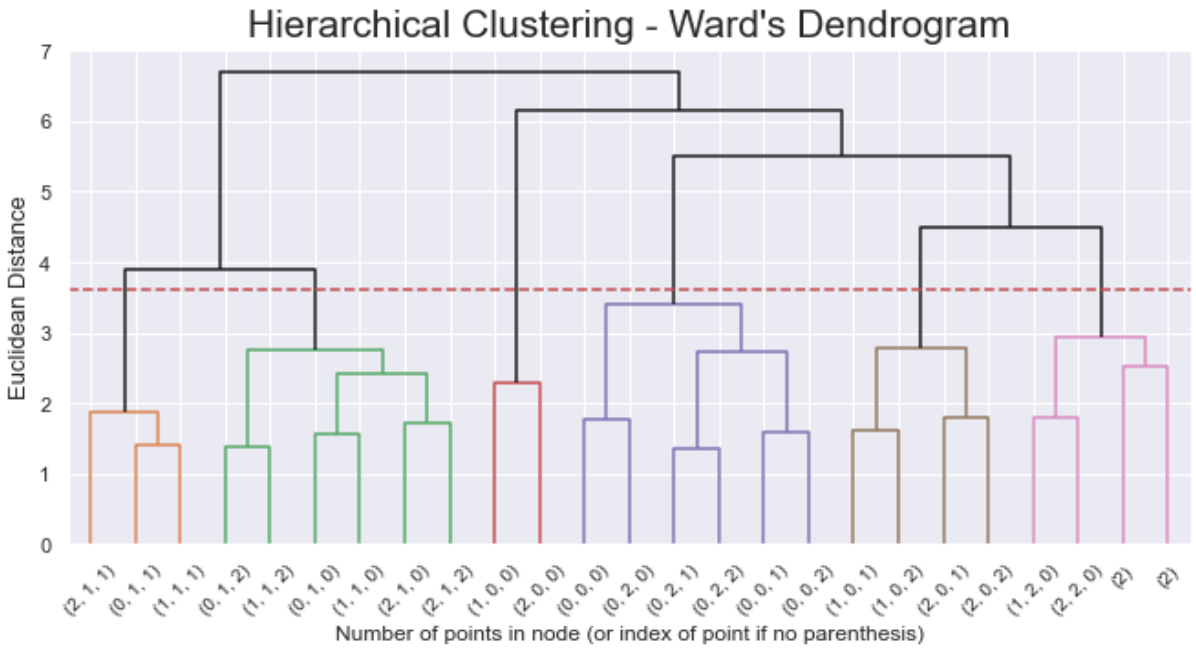


Figure 31 – Dendrogram to Merge the Segment's Clusters

Labels	0	1	2	3	4	5
Age	67.739489	35.632631	29.765203	50.331910	34.981324	50.447442
Children	0.268314	0.905997	0.742399	0.842583	0.769100	0.845581
EducDeg	2.496749	2.508374	1.754223	2.789755	1.998302	2.761395
Annual_Salary	43502.902471	21608.408428	17930.280405	31527.240128	21350.587436	31425.661395
PremMotor	230.461847	243.352204	129.160160	432.365747	187.089847	428.219991
PremHousehold	235.810750	166.231199	616.774958	79.373613	178.065110	81.869140
PremHealth	218.173927	240.146337	161.109772	109.456772	170.244160	112.497526
PremLife	46.288036	36.058936	87.811199	15.282732	95.101681	14.996460
PremWork	45.872007	35.670573	88.962601	14.523116	94.748608	14.388572
CustMonVal	205.057165	178.410573	365.075583	439.411547	117.134958	14.475991
ClaimsRate	0.712501	0.722129	0.640811	0.279733	0.803243	0.948372
Total_Premium	776.606567	721.459249	1083.818691	651.001980	725.249406	651.971688
Cessation	0.114868	0.156672	0.031250	0.375133	0.084890	0.364186
Insurance_Burden	0.018116	0.038301	0.079516	0.022462	0.040159	0.022730

Figure 32 – Centroids of each Final Cluster

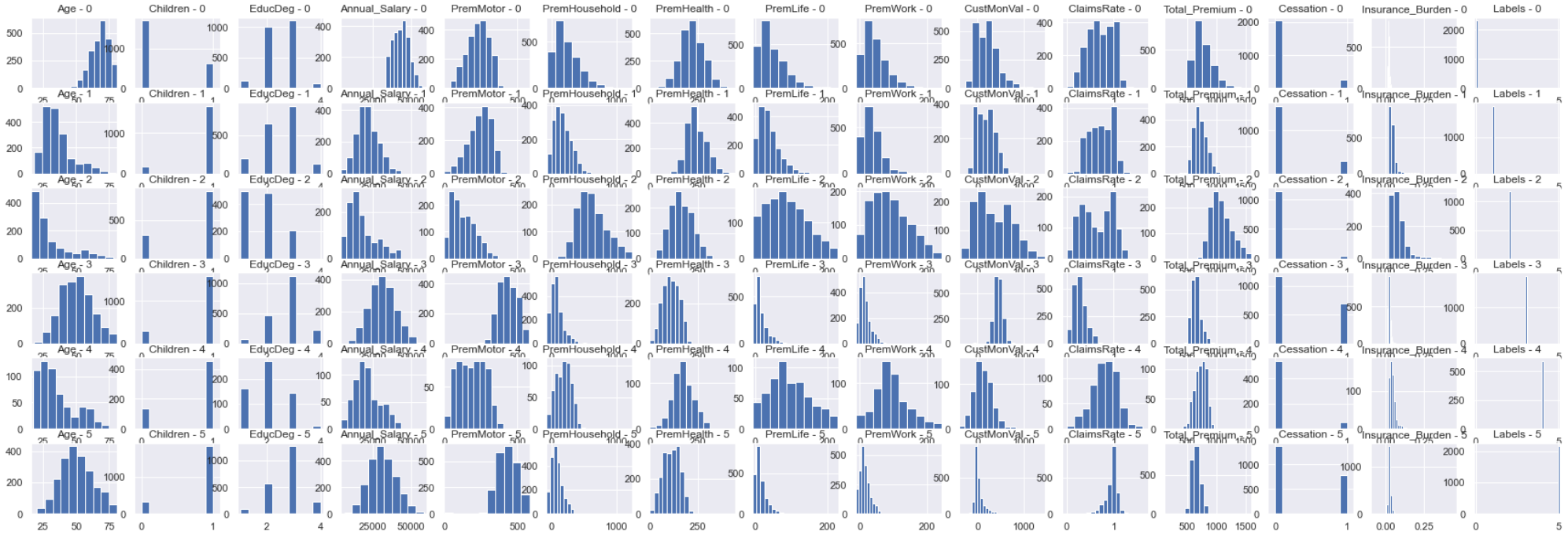


Figure 33 – Final Clustering Histograms

Figure 34 - U-Map of the final Clustering Solution