Data Mining Project

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

A2Z INSURANCE – A CUSTOMER SEGMENTATION

Group AB

David Santos: R20181082

Foazul Islam: M20200750

Tomás Jordão: M20210664

January - 2021

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

INDEX

1. Introduction

In this report we will show a customer segmentation model that has the objective to find clusters and understand the customer´s behavior of an insurance company A2Z Insurance, a fictional Portuguese company. In order to perform this study, we analyzed a total of 13 variables and 10296 observations.

During this project various data preparation techniques were applied such as treatment of data incoherence, missing values, filtering of outliers, data normalization together with clustering techniques like k-means, hierarchical clustering, t-SNE.

Using these techniques and with the insights that they provided us about the data we were able to construct various groups of customers that represent the patterns present in the data and the different customers' profiles.

In the end considering all the previous work on the development of the clusters and with the knowledge we gained about the data we were able to develop a marketing approach that we think best suits each of the different customers' profiles.

GitHub: https://github.com/TJordao/Project_Data_Mining_2021

2. Exploration of the Dataset

First, before performing any clustering technique we analyzed and explored the dataset to better understand the variables and any data nuances that may exist. We checked the number of variables, their data types and any incoherence they may have.

We began by loading the a2z_insurance dataset and set the variable "CustID" as index. Following this action, we looked at the first five rows of the dataset and the variables data types where we observed that the dataset is composed of 10296 records and 13 variables.

After this first observation we noticed that there were various problems with two of the variables, "BirthYear" and "FirstPolYear". The main problem was that there were 19.4% customers with birthyear that was more recent than the first policy year or the year they became customers of the insurance company. To deal with this problem we decided to drop these two variables because we were unable to determine from which variable the problem originated and the number of records with the problem was great.

In the next step we began by defining the metric and non-metric features, checked for missing values, checked the distribution of the variables and created new variables.

We decided that the variables, "EducDeg", "Children" and "GeoLivArea" should be considered non-metric features and the rest of the variables are metric.

Regarding the missing values we observed that there were several variables with missing values as it can be seen in the **annex 1**. These missing values were then filled using measures of central tendency as is explained in the next section of this report.

In relation to the distribution of the variables we used the method "describe(include = 'all')" from pandas that allowed us to observe a number of important characteristics of each variable, such as the frequency, unique values, mean, standard deviation and quantiles as it can be seen in **annex 2**. These metrics showed several important aspects that permitted us to better understand the data.

The first being that the most common academic degree is BSc/MSc, the second is that there are more customers with children than without and finally we were able to observe that most of the variables present in this data set presented outliers. This assumption is easily seen by observing the **annex 2** where most variables show that the variable max is very distanced from the variable mean implying that it can be outliers.

3. Transforming the Dataset

3.1. Missing Values

In this step we began by identifying and selecting techniques to deal with the values that were missing in our dataset. We identified the following number of outliers as it is shown in **annex 1**:

As is possible to observe, there are missing values in both numerical variables and categorical variables.

Considering for that fact we decided to use two techniques:

- For categorical variables or non-metric features we decide to fill the missing values using the mode.
- For numerical variables or metric features we decided to fill the missing values using the median mainly because it is not affected by outliers

3.2. Creating New Variables

In this step we only created a new variable that we thought would give good knowledge about each customer average spending in terms of types of insurances. This new variable “Avg_Premiuns” give us the mean spent by each customer in insurances.

3.3. Outliers

In this step we used different techniques for identifying and filtering outliers. Those techniques were thresholding, IQR, Z-Score, Local Outlier Factor (LOF) and One Class SVM.

3.3.1. Thresholding

Firstly, we began by doing a filtering where we removed some outliers. These was necessary since when constructing the histograms and boxplots we were unable to get any relevant information of them.

After this initial filtration we constructed boxplots and histograms, which enabled us to define thresholds that removed approximately 1.1% of observations we considered outliers.

3.3.2. IQR

The next technique we tried was intra quartile range. When using the standard values for the classification of observations as outliers we observed that we would remove approximately 14.6% of the observations.

Due to this fact and although we tried different values for defining the upper and lower limit, we could never comply with the rule of thumb of not removing more than approximately 3% of the observations we decided not to use this technique for filtering the outliers from our dataset.

3.3.3. Z-Score

This technique defines as outlier observations that are at a predefined value of standard deviations of the mean value of what is being observed.

In our case we defined as outlier an observation we a z score greater than 3.5 and considering this value we removed about 2.55% of the observations in our dataset, a value that complies with the rule of thumb stated previously.

3.3.4. Local Outlier Factor – LOF

LOF is and unsupervised anomaly detection method that measures the local density deviation of a data point in relation with its neighbours. This technique considers outliers observations with substantially lower density than their neighbours.

The main parameter of LOF is “n_neighbors” and based on the documentation found in scikit-learn a value of 20 usually works well.

In our case we needed to define a value of 50 for the parameter “n_neighbors” that contradicts the documentation but that could be linked with the characteristics of our dataset.

3.3.5. One Class SVM

One Class SVM is an unsupervised outlier detection technique that uses support vector machines. It has two main parameters the “kernel” and “nu”. The first parameter cobined with the second define a frontier where observations that lie outside are considered as outliers. The parameter “nu” corresponds to the probability of finding a observations outside of that frontier.

In our problem we used the default value for the parameter “kernel” and decided to use the value 0.025 for the parameter “nu” complying with the rule of thumb stated above for the removal of observations from the dataset.

3.3.6. Technique used for removing outliers - Z-Score and LOF

For the final filtration of outliers, we decided to combine two of techniques mentioned previously, Z-Score and LOF. First, we removed observations that presented a z-score value lower than 3.5. After this, we used LOF with a value of 50 for the parameter “n_neighbors” to remove any outliers that were not considered by the z-score method.

In the end, we removed 3.02% of the observations a value that complies with the rule of thumb stated previously.

3.4. Feature Selection

To select the relevant metric features we utilized a correlation matrix that can be seen in **annex 3**. We began by constructing scatterplots that enabled us to identify relationships between numerical variables. After observing these graphs, we realized that some of the relationships may not be linear. Considering this fact, we decided to use the method “spearman” to construct the correlation matrix.

In the end we decided to eliminate the variable “PremMotor” because presented a considerable correlation with various variables.

3.5. Standardization

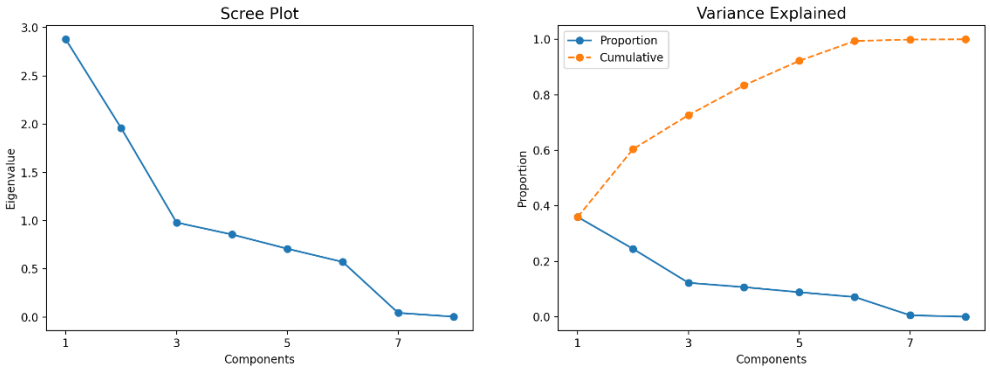
The metric features in our dataset present different scale and in order to not bias the clustering phase we decided to standardize them. In that sense, we used Standard Scaler from python so that all metric features had mean of 0 and standard deviation of 1.

3.6. PCA – Principal Components Analysis

After all the previous steps, we decided to use a dimensional reduction technique called Principal Components Analysis. This technique transforms possibly correlated variables into linearly uncorrelated variables called principal components.

In our project we used the scree plot and the plot of variance explained to decided on how much principal components to use, these two graphs can be found in bellow. After this observation, we concluded that the optimal number of principal components was 4.

We will use the principal components for clustering, but they will not be the final solution because they are more difficult to interpret, and the results were very similar to other techniques used during this project.



4. Clustering

4.1. K-Means – All Features

K-means is algorithm that uses centroids and assigns each sample to the closest centroid and therefor creating a cluster for each centroid. The objective of the k-means is to minimize the inertia or intra-cluster variance. This algorithm is very sensitive to the initial seeds or centroids and only works for metric variables.

Regardless, to tackle these limitations we used the default value for the parameter "init = kmean++" and used our metric features. We began by calculating the inertia for different values of k, number of clusters. After this calculation, we applied the elbow method to the inertia plot (**annex 4**) and decided that the optimal number should be 3 or 4 clusters.

In addition, we created a silhouette plot for each different value of k, where we observed that for a value of 3 for k there was a cluster with not well defined. With the information from the inertia and silhouette plots we decided that our best value for k should be 4.

With the value of k defined and using a plot to see the means of each metric variable for each cluster (**annex 5**) we can see four different behaviors.

Cluster 0, with 2453 samples, presents a salary (MonthSal) slightly above the mean, a customer monetary value (CustMonVal) well below the mean, a claims rate well above the mean (ClaimsRate) and then the amount paid in premiums bellow the mean.

Cluster 1, with 1482 samples, presents a salary (MonthSal) well below the mean, a customer monetary value (CustMonVal) well above the mean, a claims rate slightly below the mean (ClaimsRate) and then the amount paid in premiums various spending well above the mean for the premiums in household, life and work and slightly below for premium heath.

Cluster 2, with 2735 samples, presents a salary (MonthSal) slightly above the mean, a customer monetary value (CustMonVal) well above the mean, a claims rate well below the mean (ClaimsRate) and then the amount paid in premiums bellow the mean.

Cluster 3, with 3315 samples, presents a salary (MonthSal) slightly above the mean, a customer monetary value (CustMonVal) above the mean, a claims rate slightly above the mean (ClaimsRate) and then the amount paid in premiums various spending below the mean for the premiums in household and slightly above for premium life, work and heath. Being the latest the furthest away from the mean.

We still tried with a value of 3 for the number of clusters but considered that the clusters where to large and not enough heterogeneous to retrieve any relevant information.

4.2. K-Means – PCA

After performing the k-means clustering with the metric features we decided to use the principal components instead of the metric features.

We applied the same techniques as the ones mentioned above and found that the optimal number of clusters based on the inertia plot was 4.

As it can be seen in **annex 6**, the final solution was very similar to the previous one, so we decided to continue with the previous solution found.

4.3. K-Means -Different Perspectives

Before profiling the final solution, we still tried another technique. This technique used k-means on two different perspectives.

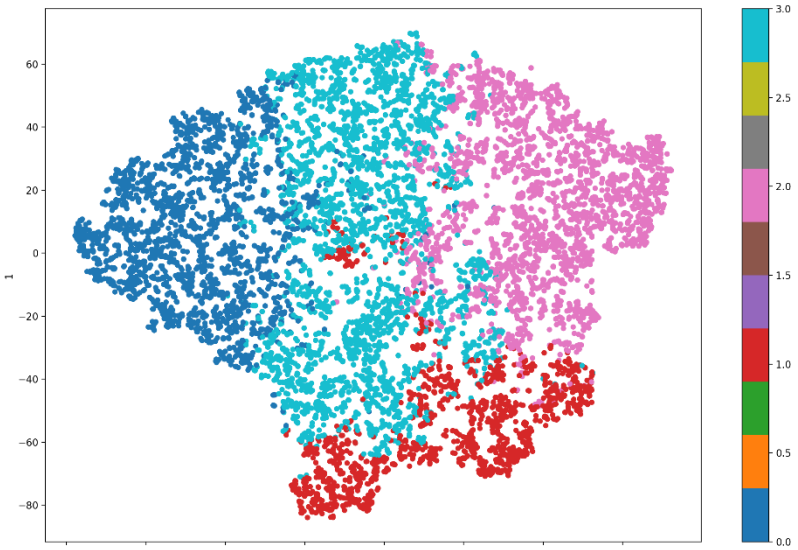
The first perspective was the “Costumer Value” and contained the metric variables “MonthSal”, “CustMonVal” and “ClaimsRate”. The second perspective was “Customer Spent” and contained the variables “PremHousehol”, “PremHealth”, “PremLife”, “PremWork” and “Avg_Premiuns”.

Using the inertia plot we decided that for both perspectives the optimal number of clusters should be 3. After forming the clusters and observed the contingency table we merged the different perspectives using Hierarchical clustering. Ending with 4 clusters for the final solution for this technique.

In the end and with the help of the graphics present in **annex 7** the decision was that this solution found was not better than the solution found previously using all metric features together. This decision was based on, as seen in the graphics, the clusters formed were very similar to each other and some of the clusters and very few observations.

4.4. Cluster Visualization with t-SNE

T-SNE is technique for the visualization of high-dimensional data. We decide to use this technique to confirm the results found previously. For this dataset, we used the default parameters of 2 for the number of components and 30.0 for the perplexity. The result can be seen in the following figure.



4.5. Cluster Profiling and Marketing Campaign

Using the solution found previously for the k-means using all metric features together we can provide a profile for each of the clusters.

The table below provides a summary of each variable for each cluster can be found in the next table.

	Custer 0	Custer 1	Custer 2	Custer 3
MonthSal		Individuals with income bellow average		
CustMonVal			Individuals with value above average	
ClaimsRate	Individuals with value above average		Individuals with value below average	
PremHousehold		Individuals who spend the most		
PremHealth				Individuals who spend the most
PremLife		Individuals who spend the most		
PremWork		Individuals who spend the most		
Avg_Premiuns				
b'2 - High School	More Individuals			
b'3 - BSc/MSc				More Individuals
b'4 - PhD		More Individuals	No individuals	
Children			Least Individuals with children	More Individuals with children
GeoLivArea	Most Individuals live in area 1			

By observing this table and the graphics that were produced during the clustering phase we can profile the clusters and make decisions on what is the best marketing strategy for each one.

Cluster 0

This cluster is composed of individuals with a low level of education, with a value of salary slightly above average and with a great number of individuals having children. They are the individuals with the greatest value above average for claims rate and with the lowest value for customer monetary value. This fact can be tied with the fact that they are the ones who least spend in premiums and therefor on insurance.

The marketing strategy should pass by trying to increase their spendings in premiums to contradict their above average value of claims rate.

Cluster 1

This cluster is composed of individuals with a value of salary below average but with a great range of values. They have the greatest number of individuals with the higher education (PhD) and are the second clusters that has more individuals with children. These individuals focus most of their spendings in premiums on household, life and work.

The marketing strategy should pass by trying to increase their spendings in premiums linked to health with a promotion on those insurances leading them to be an even more valuable client.

Cluster 2

This cluster is composed of individuals with a value of salary slightly above average. They have zero individuals with the higher education (PhD) and are the cluster that has least individuals with children. These individuals have spent below average on premiums but are the most valuable for the insurance company with the lowest value of claims rate.

The marketing strategy should pass by trying to increase their spendings in premiums to be an even more valuable client.

Cluster 3

This cluster is composed of individuals with a value of salary above average but with a great range of values. They are the cluster that has most individuals with children. These individuals have spent below average on premiums for household but spend the most on premiums for health.

The marketing strategy should pass by trying to increase their spendings in premiums for health, household or work to contradict theirs above average value for claims rate. A good strategy could pass by raising billboards near schools with information related to those insurances giving the fact that they are the group where there more individuals with children.

4.6. Classification of Outliers

In the end, we decided to classify the outliers that have been removed from the dataset. We classified these observations with the results that we get from the k-means clustering and a Decision Tree. Splitting the dataset into 80% training and 20% testing the model could estimate on average that it can predict 82.97% of the individuals correctly.

As a result, we concluded that, in the outlier's dataset, we have 269 observations belonging to cluster 0 and 42 belonging to cluster 1.

5. Conclusion

On this project we were able to explore and modify the a2z_insurance dataset and finally create a solution based on a clustering technique. The data preparation was full of challenges from incoherent data to missing values, but we implemented various techniques to deal with those problems that provide confidence on the results displayed on this report. We removed variables, filled missing values and removed outliers using multidimensional techniques such as Z-Score and LOF.

Although various clustering techniques were applied and tested during this project, we fill that the one we presented is the best solution for the problem in question as it shown in the results displayed previously. K-Means applied to all metric features gave the most different values for the features that were present in our dataset, allowing us to find and characterize 4 clusters to classify the different individuals that compose them.

In the end we present four clusters of different individuals who possess different characteristics, high and low income, and different backgrounds, education. These allow us to come up with a marketing strategy thar we think will best apply to each one of the clusters.

References

Philip Wenig, (2018), Local Outlier Factor For Anomaly Detection

https://towardsdatascience.com/local-outlier-factor-for-anomaly-detection-cc0c770d2ebe

scikit-learn documentation (2021) OneClassSVM

https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html

scikit-learn documentation (2021) LocalOutlierFactor

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html

scikit-learn documentation (2021) Outlier detection with Local Outlier Factor

https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

6. Appendix (Doesn't count for the 10page limit)

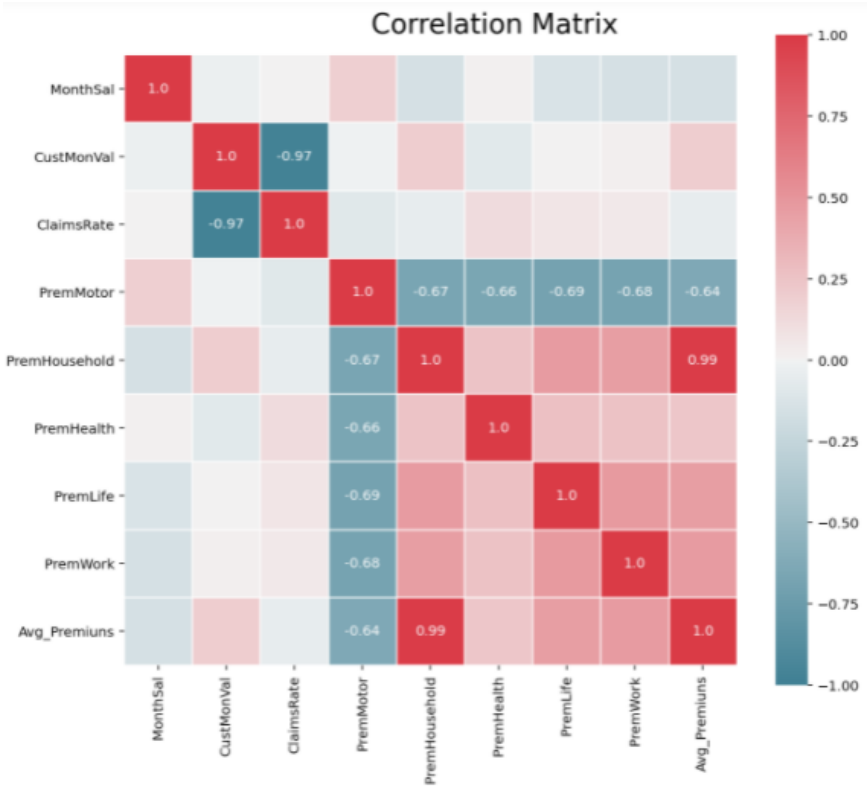
6.1. Annex 1- Missing Values

EducDeg	17
MonthSal	36
GeoLivArea	1
Children	21
CustMonVal	0
ClaimsRate	0
PremMotor	34
PremHousehold	0
PremHealth	43
PremLife	104
PremWork	86

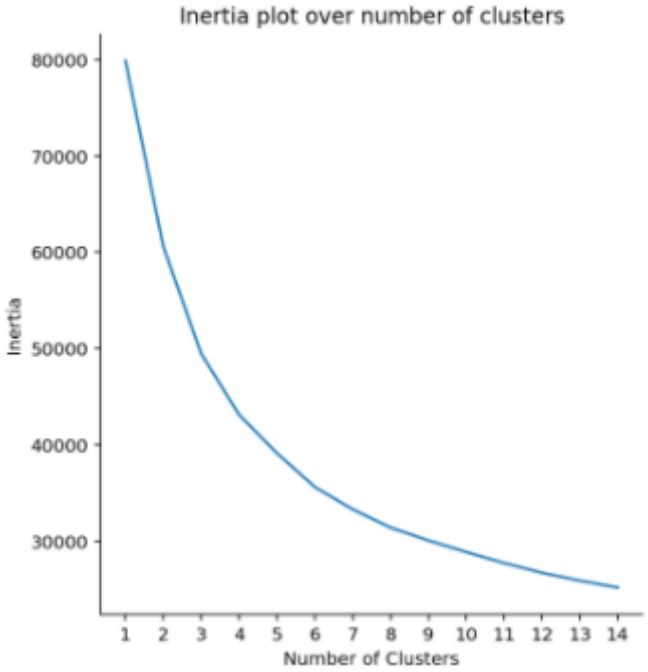
6.2. Annex 2- Distribution of the Variables

Variable	Mean		Min Value	Max Value
EducDeg	NaN		NaN	NaN
MonthSal	2506.64899		333	55215
GeoLivArea	2.709887		1	4
Children	NaN		NaN	NaN
CustMonVal	177.892605		-165680.42	11875.89
ClaimsRate	0.742772		0	256.2
PremMotor	300.464109		-4.11	11604.42
PremHousehold	210.431192	-75	25048.8	
PremHealth	171.544203	-2.11	28272	
PremLife	41.691178	-7	398.3	
PremWork	41.147148	-12	1988.7	

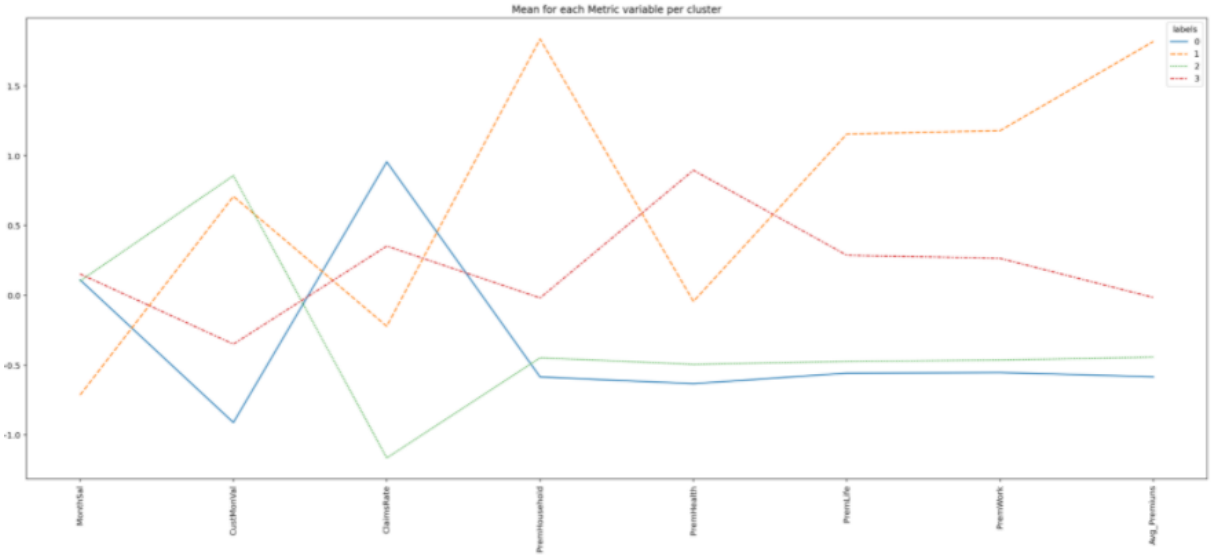
6.3. Annex 3- Correlation Matrix



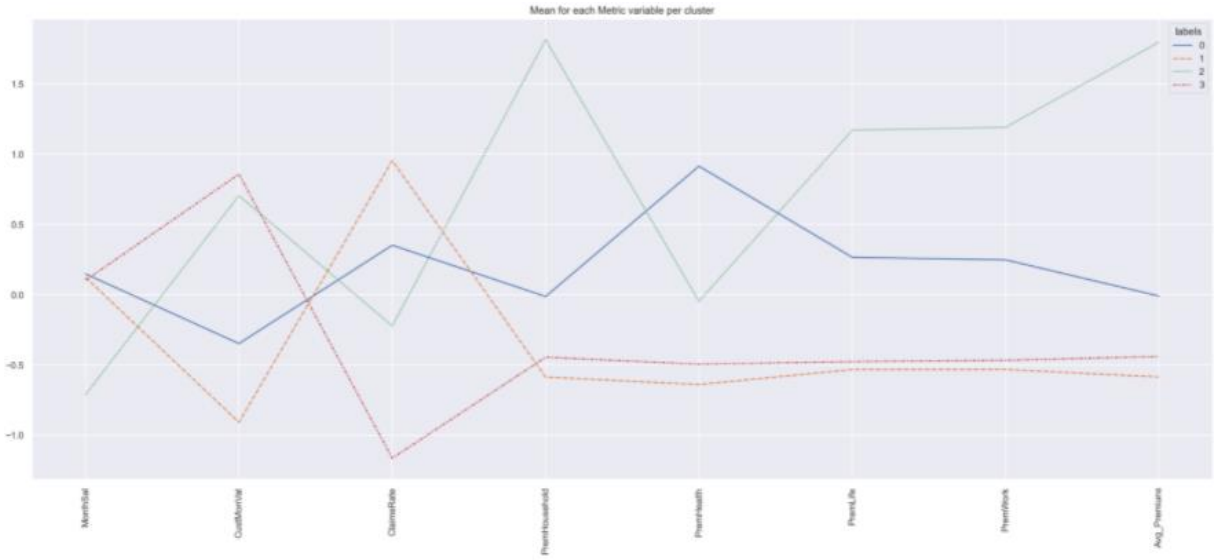
6.4. Annex 4- Inertia Plot



6.5. Annex 5- Cluster Means Plot



6.6. Annex 6- Cluster Means Plot for PCA



6.7. Annex 7- Clustering with Different Perspectives