NOVA IMS Information Management School

MDS

MASTER DEGREE PROGRAM IN DATA SCIENCE AND ADVANCED ANALYTICS

Data Mining Project

Clustering on Insurance Data

Group AN

Inês Rocha, number: 20220052

Isabel Dias, number: 20191215

Joana Sousa, number: 20191205

January, 2022

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

INDEX

1. Introduction

In today's competitive business environment, it is essential for organizations to understand their customers in order to effectively meet their needs and drive growth. One way to achieve this understanding is through market segmentation, the process of identifying groups of customers with similar characteristics and behaviours. By understanding the differences between these segments, organizations can make more informed strategic decisions about opportunities, product definition, positioning, promotions, pricing, and target marketing.

Clustering is a powerful data mining technique that can be used to identify and understand customer segments. It involves grouping similar data points together into clusters, based on their characteristics and behaviours. In this report, we applied clustering techniques to the customer database of A2Z Insurance, a Portuguese insurance company, in order to identify and understand the different segments within the customer base. Our goal was to not only understand the demographics and value of each segment, but also to identify opportunities for targeted marketing and cross-selling of insurance products.

A2Z Insurance provided us with a sample of 10,290 customers from its active database, along with data on their characteristics and behaviours. Our analysis involved cleaning and pre-processing the data, selecting appropriate clustering algorithms, and evaluating the resulting clusters to determine their relevance and value to the organization. By using data-driven approaches to market segmentation, A2Z can better serve its existing customers and improve its targeting of prospective customers.

2. Data

The process started by importing a variety of libraries for data manipulation, visualization, and clustering. Loaded the A2Z Insurance dataset, which includes a total of 13 variables (Table 1), into a pandas DataFrame using the pyreadstat library and set the *CustID* column as the index. The nunique method was used to check for duplicate rows based on the *CustID* column. With these initial steps, we have successfully loaded and prepared the dataset for further analysis.

3. Data Exploration

The data exploration process for the A2Z Insurance dataset began by checking for missing values in the data using the isnull and sum methods (Figure 1). The data types of the variables *BirthYear*, *FirstPolYear*, *GeoLivArea*, and *Children* were changed to integers due to being discrete variables, using the astype() method. The describe method was used to generate summary statistics for the numerical variables where we were able to see that the variable *BirthYear* had an impossible minimum value and the variable *FirstPolYear* had an impossible maximum value. The describe(include="O") method was used to generate summary statistics for the categorical variables, including count and unique values.

Next, the inconsistencies and duplicates in the data were checked. Using the value_counts method on the variable *BirthYear*, it was able to be seen that there was only one value that was inconsistent: 1024. Next it was checked if there were any observations where the first year of policy came before the birth

year. There were 1997 records where this happened, which is a very significant amount of the dataset. The dataset has 117 minors, which cannot have insurance contracts in their name in Portugal. It was verified that in each education category, there were no records that seemed too young.

Value counts and distribution graphs for various variables were generated to gain a better understanding of their characteristics and patterns. With this, the *EducDeg* variable does have missing values that were masked as blank spaces instead of the Nan. They were changed to Nan values, to have all missing values marked the same way.

The distribution graphs help checking for extreme or outlying values that may require further investigation or treatment.

Several variables were identified as having with extreme values which were *MonthSal, CustMonVal*, ClaimsRate, *PremMotor*, *PremHousehold*, *PremHealth*, and *PremWork*. In the following outlier analysis, these values were explored in more depth and a decision was made on how to handle them.

In this analysis, also several variables were found to have skewed distributions, namely *CustMonVal*, *ClaimsRate*, *PremHousehold*, *PremLife*, and *PremWork*. Due to the outliers, it is hard to sure that all these variables were skewed, so histograms filtering the outliers were made. Firstly, the variable *MonthSal* (Figure 2) that had a distribution almost Normal without the outliers. After, the *CustMonVal* variable histogram (Figure 3) that showed that there was a high number of customers just below zero. The *ClaimsRate* histogram (Figure 4) showed most customer between 0 and 2. The *PremMotor* histogram (Figure 5) also seemed to show that this variable had a bell-shaped distribution. We are going to zoom in some of these by filtering the outliers. These variables may require further investigation or transformation to better inform the clustering process.

Then, the missing values were checked using the msno.matrix function that provides information about the correlation between variables' missing values (Figure 6). We concluded that the *FirstPolYear* and *EducDeg* were Missing Not at Random (MNAR), since the observations in which they had missing values were the same. With this same matrix, we noticed that there was only one variable related to the premium's value that did not have any missing values, *PremHousehold*.

Zeros were checked for in the other variables related to the premium's values and none were found. It was concluded that this was because the missing values represented people who did not have the premium as indicated in the *PremHousehold* variable.

The next section was to explore the relationships between variables. Firstly, a pairplot (Figure 7) was used to detect possible relationships, then those variables were plotted against each other with scatterplots and histograms.

There is a strong correlation between *CustMonVal* and *ClaimsRate* in the dataset (Figure 8). As identified through the metadata and observed in the distribution graphs, this correlation is logical given the nature of the data. However, there were also some outliers present, including an extreme outlier in the CustMonVal variable representing a person who received significantly more money from the company than they paid in premiums. It was noticed that *BirthYear* and *Salary* are very positively linearly correlated (Figure 9), and that people that were born until around the 1950s seem less prone to have *Children* (=1) (Figure 10). Lastly, the variance and mean of *PremMotor* decreases with the increase of *PremLife* (Figure 11).

Another tool used to get information about the data was the ProfileReport function, which provided all kinds of descriptive statistics and visualizations that characterize the variables and their relationships. It confirmed the conclusion taken about the skewed variables and the correlated variables.

4. Data Cleaning

To start the cleaning process, a copy of the dataset was created to preserve its' initial qualities. Then, the following outliers were removed according to the exploratory analysis:

1. *MonthSal* above 8000,
2. *CustMonVal* below -15000,
3. *ClaimsRate* above 50,
4. *PremMotor* above 6000,
5. *PremHousehold* above 5000,
6. *PremHealth* above 5000,
7. *PremWork* above 750.

These instances accounted for less than 1% of the dataset, so it wasn't a substantial loss of information.

After that, new histplots were computed to verify whether there was skewness in the data or if was caused by the outliers.

In variables *CustMonVal* (Figure 12), *ClaimsRate* (Figure 13) and *PremMotor* (Figure 14) a few more outliers were removed, as those observations had a significant impact on the tails of the distributions. The variables *PreHousehold* (Figure 15), *PremLife* (Figure 16) and *PremWork* (Figure 17), had very heavy tails. To combat this, a log transformation was made to the variables: $\log(X - \min(X) + 1)$, so that the negative values would be able to be computed. After these new variables were created, their distribution became closer to Normal, and thus, the algorithms could better identify the differences between values in the denser areas of the distribution.

Some inconsistencies were found in the data, including a person who had their first year of policy after 2016, and someone whose *BirthYear* was below 1900. These observations were removed. All the records younger than 18 were also dropped as minors cannot have contracts in their name. After this, we kept 98% of the data. For the people whose *FirstPolYear* was below their *BirthYear*, it was opted to make both variables null in order to impute these values as it was not feasible to just remove these observations because they represented around 20% of the dataset and would be a significant loss of information. A possible reason found for having this inconsistency was that the first policy was made by parents or family members, and the record is of a beneficiary of that policy. However, since this is just a hypothesis, it is best not to assume it as true.

The next step was to convert the categorical variable into an ordinal variable, in this case, *EducDeg,* using the first digit in the category. It was also an option to dummify the variable, but it was decided against it.

Then, the missing values were imputed. For the variables *PremMotor, PremHealth, PremLife, PremWork* the missing values were imputed with "0", as it was assumed that the missing values were due to the customer not having premium in these areas. The variables *log_PremLife* and

log_PremWork were imputed with the new value for zero in those variables. For *FirstPolYear* and *BirthYear,* the KNNImputer was used, as without an underlying assumption for the reason of missingness, it was better to use a more sophisticated method of imputation. The results were compared using a barplot (Figure 18, Figure 19), and we concluded that the distribution of the variables did not change significantly with the imputation.

Then, new variables were created. *TotalPrem* was created by the sum of the premium variables, to give an idea of the monetary value of the customer. And *Age* was created by subtracting the *BirthYear* to 2016, as it is more interpretable. After that, the dataset was divided into two: *df_cat* with the categorical variables, and *df_num* with the numerical variables.

Finally, the data was scaled both with MinMaxScaler and RobustScaler so that results of the algorithm could be compared using each of them. It was noted that the latter performed better, since it considered the skewed nature of some distributions.

5. Feature Selection

For the Feature selection, PCA (Principal Components Analysis) was tried. From the Scree Plot and Variance Explained Plot the biggest jump was at two components (Figure 20), where the cumulative variance is already about 90%.

When looking at the loading for the 1st Principal Component, only one variable had more than 0.5, *MonthSal*. This might be due to the importance of this variable in the dataset. For the 2nd Principal Component, the variables *CustMonVal*, *PremHousehold*, *log_PremHousehold* and *TotalPrem* had high positive loadings while *PremMotor* had a high negative value.

A correlation matrix (Figure 21) was created to identify relationships between the numerical variables in the dataset. The heatmap plot showed that *CustMonVal* had a very high negative correlation with *ClaimsRate*, as it was seen before, leading to the decision to not use *ClaimsRate* in further analysis. Additionally, *TotalPrem* was found to have a high correlation with *PremHousehold*, so *TotalPrem* was also not included in further analysis.The *Age* variable was also found to have a high positive correlation with *MonthSal*. It is generally advisable to remove variables that are strongly correlated from your dataset before running a clustering algorithm, as they can distort the results of the algorithm. However, it is not necessary to remove one of the variables in this case, since due to the specific goals of the analysis. As this variable is important for the analysis, it will be kept in the analysis.

The dataset was divided into two perspectives, one for the data that was about the clients as persons, which is called customer perspective on this report, and the other about everything that was related to the relationship between the customer and the insurance company, that is called insurance perspective in the report. For this perspective, the variables used were *CustMonVal*, *PremMotor*, *log_PremHousehold*, *PremHealth*, *log_PremLife*, and *FirstPolYear*, *log_PremWork*, which were later taken off as it was not helping our clusters. From the perspective about the customer, the variables used were *MonthSal*, *GeoLivArea*, *Children*, *EducDeg* and *Age*.

When we were already trying to cluster our data, we noticed that there were some multidimensional outliers still and that Mean-shift was identifying them particularly well (Figure 22, Figure 23). So, we used this algorithm to take these last outliers off our dataset. After the removal of these last outliers,

there was still had less than 3% of the data removed and the data didn't seem to have any big outliers (Figure 24, Figure 25).

6. Clustering

The cluster models were applied in both perspectives and then tuned according to what the perspective needed. Two functions were created get_ss that calculates the sum of squares for all the variables; and r2_score that uses get_ss to calculate the R^2 score of a given solution

6.1 Hierarchical Clustering

To be able to apply this method, the linkage method must be chosen. A function was created that calculates the R^2 for each cluster solution, from 1 to 5 clusters, per linkage type. The R^2 score for each linkage type values per number of clusters was plotted and the appropriate method was chosen. After choosing the linkage method that was deemed the best one, a dendrogram was done to determine the best number of clusters to use. Here the distances between the clusters are evaluated in order to find the highest one, where the threshold is placed, determining the number of clusters to use. Finally, the AgglomerativeClustering algorithm is applied with the chosen linkage method and the chosen number of clusters. The mean of each cluster for each variable is done to see if the clusters are different between each other and the silhouette score is calculated. The final solution is plotted on a UMap for visualization purposes.

For the insurance perspective, the linkage method plot (Figure 26) was not very clear, and many combinations of number of clusters and linkage method were tried. In the end, the method that was used was the ward method, as it seems to be more consistent in giving best silhouette scores and the difference to the other method in the R^2 score was negligible. Then, on the dendrogram (Figure 27), the threshold was placed at 4, resulting in 2 clusters. After, the clustering was applied to our data, it was concluded that the two clusters had a difference between each other and a reasonable silhouette score. The UMap also seemed well divided (Figure 30, Figure 31).

For the client perspective, the linkage method used was complete as that seemed to be the best after trying many combinations as well, since the plot (Figure 28) was also not very clear. In the dendrogram (Figure 29), the threshold was placed at4 , which meant 2 clusters. After applying clustering to the data, it was determined that the resulting clusters were distinct from each other and had a good silhouette score. The UMap visualization also showed clear separation between the clusters (Figure 30, Figure 32).

6.2 K-Means Cluster

As the number of clusters is needed in order to do the KMeans clustering, the inertia was plotted. This plot was then compared to the silhouette scores, that were also plotted, and based on both a number of clusters was chosen. Then, the final cluster solution was calculated and the means of the clusters for each perspective variable were done and a UMap was plotted for visualization purposes.

In the insurance perspective, the inertia elbow (Figure 34) was at 2 clusters, and this choice was supported by the silhouette score (Figure 35). So, the model was done with two clusters and it was

determined that there were differences between the two clusters through the means and the UMap (Figure 37, Figure 38).

For the customer perspective the inertia plot (Figure 36) seemed best between and 2 and 3 clusters and after looking at the silhouette score (Figure 37), 3 clusters were chosen. The model was completed using 3 clusters and the differences between the clusters were apparent based on their means and the UMap visualization(Figure 39, Figure 40).

6.3 Self-Organizing Map (SOM)

To start SOM, a random seed was set so the results would be reproducible. Next, the SOM algorithm was built, and combination of parameters were tried for it. After that a function for visualizing the component planes was created and using it the component plane were plotted. A U-Matrix and a Hit-Map were plotted as well for better visualization of the possible clusters and how many.

The variables of the insurance perspective seemed to have been good for clustering on the component planes (Figure 42) and both the U-Matrix (Figure 43) and the Hit-map (Figure 44) showed some possible clusters.

The variables related to the customer perspective appeared to be effective for creating clusters on the component planes (Figure 44), as indicated by the U-Matrix(Figure 46) and Hit-map(Figure 46), which showed potential separation between the clusters.

Emergent SOM Hierarchical Clustering as well as K-Means Emergent SOM, to analyze the data. This method involves using a very large number of units and is effective at detecting the underlying structure of the data through clear U-Matrices. This technique can be used in conjunction with other clustering algorithms to provide a more comprehensive analysis. Emergent SOM was employed to thoroughly examine the data and identify any patterns or structures.

Emergent SOM with K-Means

A similar process to what was done with K-Means was also applied here, however instead of being applied in the data directly, it was applied on the SOM nodes. So, inertia was used to choose the number of clusters and then performed the K-Means algorithm. Then the HitMap was done for this solution. Once again, the mean of each variable per cluster and the UMap were done to see how defined and different between each other the clusters are.

For the insurance perspective, the inertia plot (Figure 48) showed an elbow at 2 clusters, it was the one chosen in the end to perform the algorithm. The Hit-Map (Figure 49) seemed to be very well defined, which was confirmed by the means of the variables by cluster and the UMap().

Based on the inertia plot(Figure 47) for the customer perspective, it was determined that 2 clusters provided the most clear separation. This decision was supported by the well-defined Hit-Map and the distinct means of the variables for each cluster, as well as the UMap(Figure 50, Figure 51) visualization.

Emergent SOM with Hierarchical Clustering

The Hierarchical algorithm was used in a similar way, but it was applied to the nodes of the SOM instead of the data directly. First, the most appropriate linkage method was chosen and then a dendrogram was used to choose the ideal number of clusters to use with the final algorithm. A HitMap was created based on the solution that was obtained by applying the Hierarchical algorithm to the nodes of the SOM. The mean of each variable was calculated for each cluster and a UMap was also generated to assess the distinctness and separation of the clusters.

For the insurance perspective, he ward linkage method was the chosen from the linkage method plot and 2 clusters were chosen as the ideal number of clusters. Based on the HitMap and the analysis of the mean of the variables for each cluster and the UMap, the clusters appeared to be cleary defined and distinct from one another.

For the customer perspective, the ward linkage method was selected and two clusters were identified as the optimal number based on the linkage method plot. The distinctness of the clusters was supported by the HitMap, the mean of the variables for each cluster, and the UMap visualization.

6.4 Mean Shift Clustering

For the Mean Shift algorithm, a silhouette score plot was done in order to choose the best quantile to have in the bandwidth parameter. After that, the bandwidth was chosen and the algorithm was calculated, giving us a number of clusters. Following this, the means of the variables were calculated by cluster as was as the R^2score. Once again, the UMap for the solution was done to get a better understanding of the clusters.

The best quantile value was 0.3 for the insurance perspective Figure , with a bandwidth of 1,59. This gave us 2 clusters when calculating the algorithm. The R^2 of this solution was 0,03. This solution had very distinct but unbalanced clusters as it can be seen on the UMap.

It was determined that the best quantile value for the customer perspective was 0,03, however this resulted in 29 cluster, which was not feasible, so the value 0,15 was chosen as it also had a relatively high silhouette score but gave us 2 clusters instead, resulting in a bandwidth of 1,09. Training the algorithm with this value resulted in a R^2 of -1,72.

6.5 DB Scan

The DBSCAN algorithm requires two parameters in order to cluster the dataset, minpts, and eps. Usually, the minpts value is chosen based on this rule (minPts = 2 x dim), in our case, would be minPts = 10 in both perspectives. However, because we have a large dataset, we opted to increase this value in both solutions.

The value for ε(eps) can then be chosen by using a k-distance graph. We plotted this curve and choose an eps in the "elbow" of the plot.

In the insurance perspective, we created a solution with 2 clusters, based on the K-distance graph (), 2177 points as noise/outliers and an R^2 of 0.9496.

In the customer perspective, we ended up with a solution with 8 clusters, based on the K-distance graph (), 108 points as noise/outliers, and an R^2 of 0.6053.

7. Final Models

7.1 Insurance Perspective

After tuning and fitting all the models, the R^2 of each one was calculated through the function r2_scores and they were kept in a DataFrame. Through seeing the values in the DataFrame and doing a bar plot, the SOM with K-Means was decided as the best model for this perspective. The labels were found through the best matching unit and added to a new DataFrame called df_final.

7.2 Customer Perspective

A DataFrame with all the R^2 scores for each model was also calculated for this perspective and compared in a bar plot. The DBScan was the model with the highest R^2, so it was the chosen one. It was noticed that the clusters seemed defined in the UMap. However, 9 clusters were to many clusters, as when joining the perspective, there would be 36 clusters. So, with the help of the UMAP visualization, the clusters were joined into 4 clusters, as shown in Figure 78.

8. Cluster Profiling and Advertising

In order to get the end solution, we saw the number of observations in each cluster, and concluded than neither was a very small subset of data, so there weren't outliers. Then, we used an Hierarchical Clustering algorithm with the Ward's linkage to join clusters. The best deemed solution was 4 clusters with the aid of a dendrogram, from which was created the final solution.

To better visualize and interpret those clusters, we used line plots and barplots for each of the clusters, so that they could be compared side by side. A radar chart was also created, to have a joint visual representation of the different cluster mean values. Finally, we also used groupby() to see the mean differences between clusters, and evaluate their characteristics in relation to one another. The interpretations were as follows (Figure 83 and 84).

Cluster 0:

- This group has the highest values for the household, health, work and Life premiums;
- Lowest value in the motor premiums;
- This group has the second highest Customer Monetary Value (average of 262), in the long term these are also some of the insurance's valuable customers, since it has a high value close to cluster 1;
- High School as the mode of this group;
- Youngest group of the dataset;

These customers are the youngest ones, so it is natural that on average they earn less. Still this cluster have the second highest monetary value and second lowest claims rate. This is due to the fact that all the premiums have high values except for the PremMotor. This cluster does not use motor premiums,

as they might not need it. It can be beneficial to do an analysis od the needs of these customers, as they are costumers that spend money on a very big variety of products.

Cluster 1:

- This group has the highest Customer Monetary Value (average of 271), in the long term these are the insurance's valuable customers;
- BSc/MSc as the mode for this cluster;
- Highest average of month salaries in this dataset, clients in this cluster tend to earn more than its counterparts;
- This group has the lowest proportion of children than the others clusters, and similarly to the next two clusters, it has older people;
- This is the cluster that has the lowest Claims Rate;
- From the Premiums, this cluster is among the lowest in all types of premiums except for the health. In this area, it is the most profitable cluster.

This cluster is the one with the highest Monetary Value and have the highest capacity of spending money (since they have a higher salary), so it is important that these customers are retained. As they have the highest consumption in health premiums, this area is the one that should be invested in the most, as the other premiums do not seem to be popular among this group. An exception could be the Life premiums, as these are fairly associated with the health premiums, and this cluster is also composed of older people.

Cluster 2:

- This group has the highest proportion of children, compared to the other clusters;
- Similarly to the 1st cluster, they are older clients (in age) who have also been clients the longest.
- This cluster is also composed of less education
- This is the group that earns the least, and not surprisingly, has the lowest Customer Monetary value.
- They also have the highest Claims Rate.
- In the Premiums, they are never in the extremes, but are more likely to spend less than other clusters. Although in PremMotor and PremWork, they are the group with the second highest.
- Has the highest value for the claims rate variable, meaning that the insurance company paid more to these clients. This combined with the fact that they are the lowest in terms of Customer Monetary Value means that this most likely is a group of clients that present lower profit margins for the company

Given these characteristics, we recommend advertising that is geared towards families, and bringing appealing to that component, as these customers have a median spending on the premiums but then to activate its premiums more often.

Cluster 3:

- Has the Highest premiums in Premmotor and the lowest in the others;
- BSc/MSc as the mode for this cluster;
- Has the second highest value for the claims rate variable, meaning that the insurance company paid more to these clients. This combined with the fact that they are the second lowest cluster in terms of Customer Monetary Value means that this most likely is a group of clients that present lower profit margins for the company;
- This cluster might be composed of clients that are more prone to have accidents, since they have higher values I the premium motor and are activating its insurance more often, given the high value of Claims Rates;

The insurance company should try to implement a stategy of premium package deals in this group, since its one of the group with the highest salary, mostly spends its money on only one type of insurance and its one of the most expensive cluster for the insurance company. We also we recommend advertising that is geared towards families since they are the second highest group with children in the dataset.

9. References

Fonseca, J. & Pontejos, F. (2022). Data-Mining-22-23. https://github.com/joaopfonseca/Data-Mining-22-23.(2023)

Moosavi, V & Packmann, S & Vall'es, I. (2014). SOMPY: A Python Library for Self Organizing Map(SOM). https://github.com/sevamoo/SOMPY. (2023)

Holtz, Y.(2017). The-Python-Graph-Gallery.https://github.com/holtzy/The-Python-Graph-Gallery.(2023). notebooks/391-radar-chart-with-several-individuals.ipynb

Pedregosa *et al.,* Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

Buitinck *et al.*, 2013. API design for machine learning software: experiences from the scikit-learn project

J. D. Hunter.(2007). Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering. vol. 9, no. 3, pp. 90-95.

McKinney, W., & others. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51–56).

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., … Oliphant, T. E. (2020). Array programming with NumPy. Nature, 585, 357–362.

McInnes, L, Healy, J.(2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426

Waskom, M. et al. (2017). *mwaskom/seaborn: v0.8.1*. Zenodo.

Appendix

Table 1- Dataset Variables

Variable	Description	Data type	Observations
CustID	ID	Numerical	
FirstPolYear	Year of the customer's first policy	Numerical	May be considered as the first year as a customer
BirthYear	Customer's Birthday Year	Numerical	The current year of the database is 2016
EducDeg	Academic Degree	Categorical	
MonthSal	Gross monthly salary (€)	Numerical	
GeoLivArea	Living area	Numerical	No further information provided about the meaning of the area codes
Children	If customer has children (Y=1)	Binary	
CustMonVal	Customer Monetary Value	Numerical	Lifetime value = (annual profit from the customer) X (number of years that they are a customer)- (acquisition cost)
ClaimsRate	Claims Rate	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremMotor	Premiums (€) in LOB: Motor	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremHousehold	Premiums (€) in LOB: Household	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremHealth	Premiums (€) in LOB: Health	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremLife	Premiums (€) in LOB: Life	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
PremWork	Premiums (€) in LOB: Work	Numerical	Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years

FirstPolYear	30
BirthYear	17
EducDeg	0
MonthSal	36
GeoLivArea	1
Children	21
CustMonVal	0
ClaimsRate	0
PremMotor	34
PremHousehold	0
PremHealth	43
PremLife	104
PremWork	86

Figure 1-Missing values

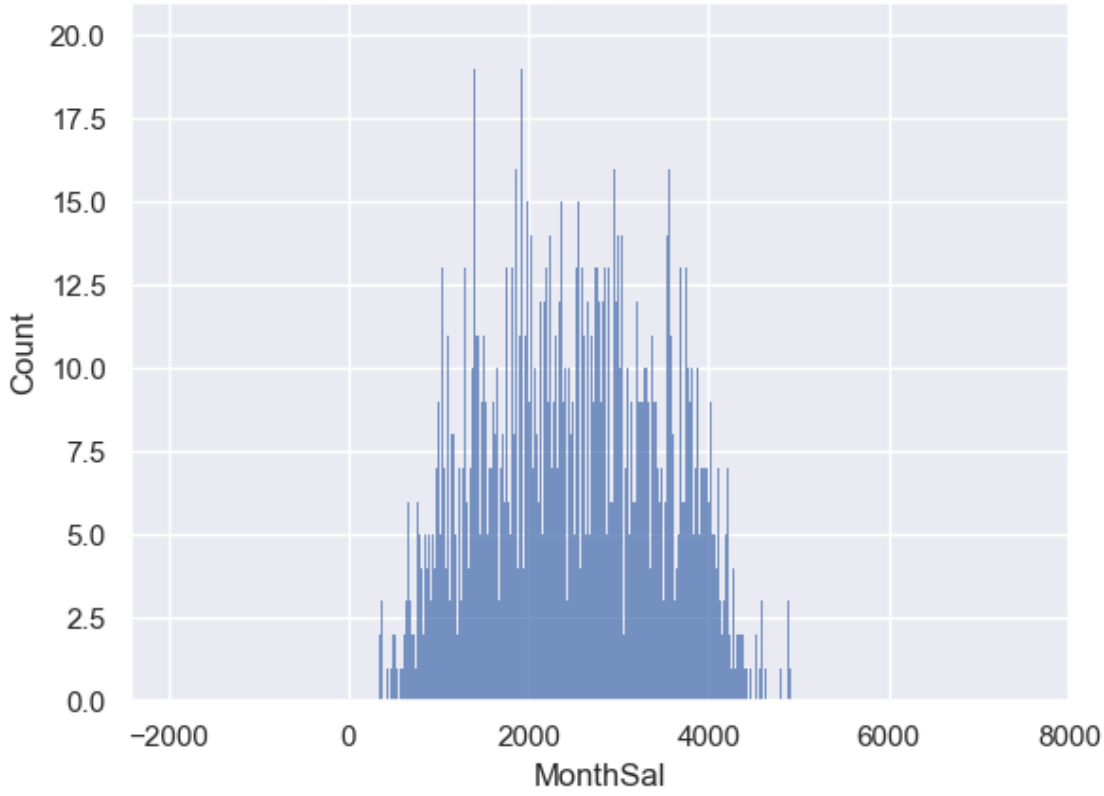

Figure 2-MonthSal without outliers

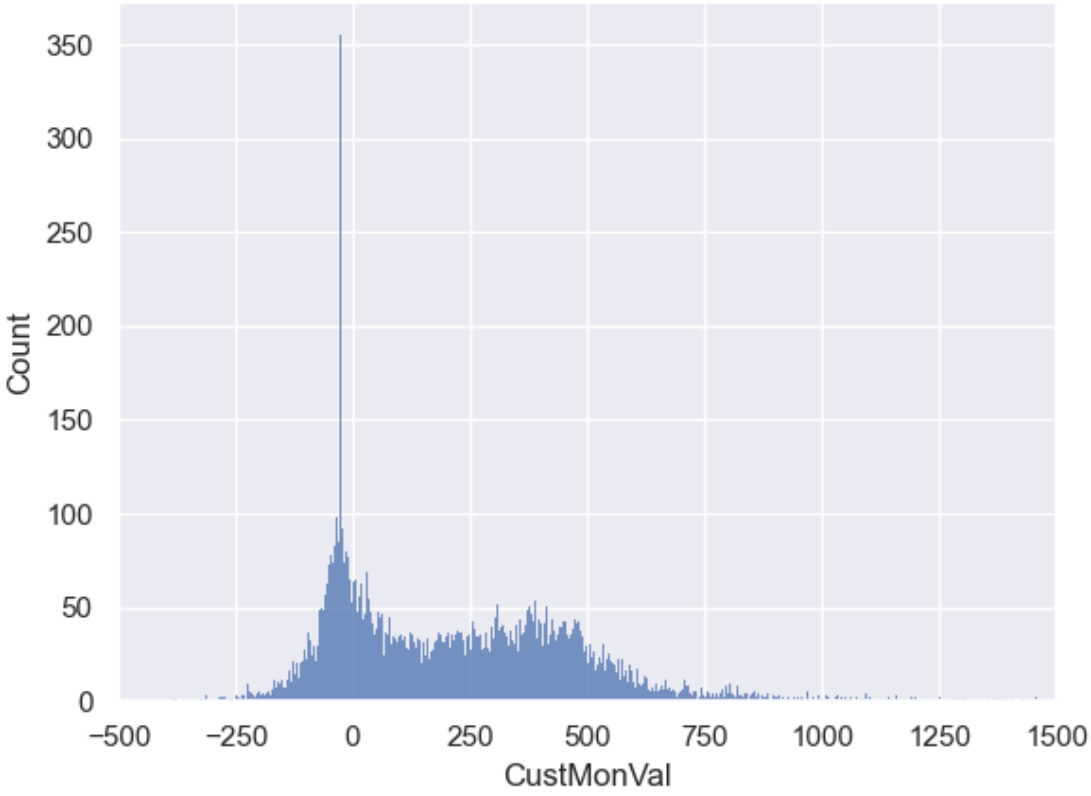

Figure 3 - CustMonVal without oultiers

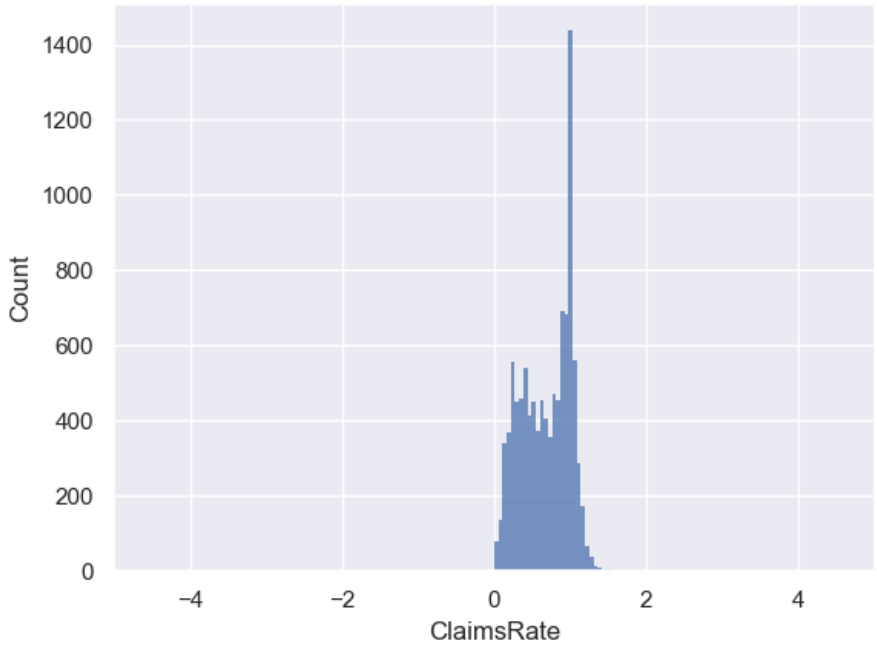

Figure 4 - Claims Rate without outliers.

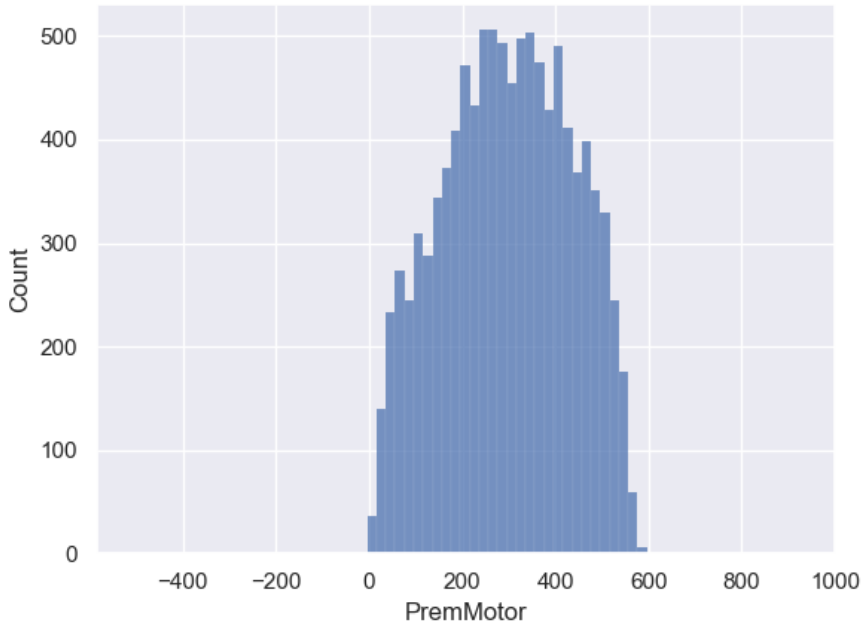

Figure 5 - PremMotor without outliers

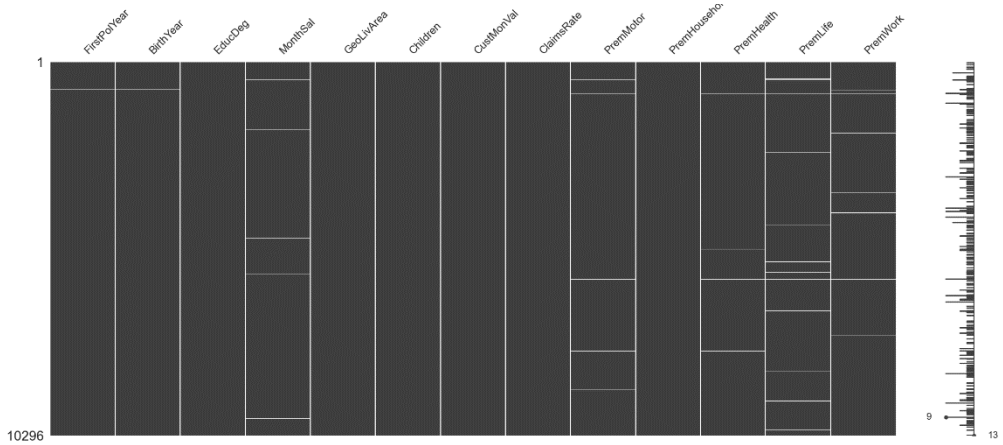

Figure 6 - Missing values matrix

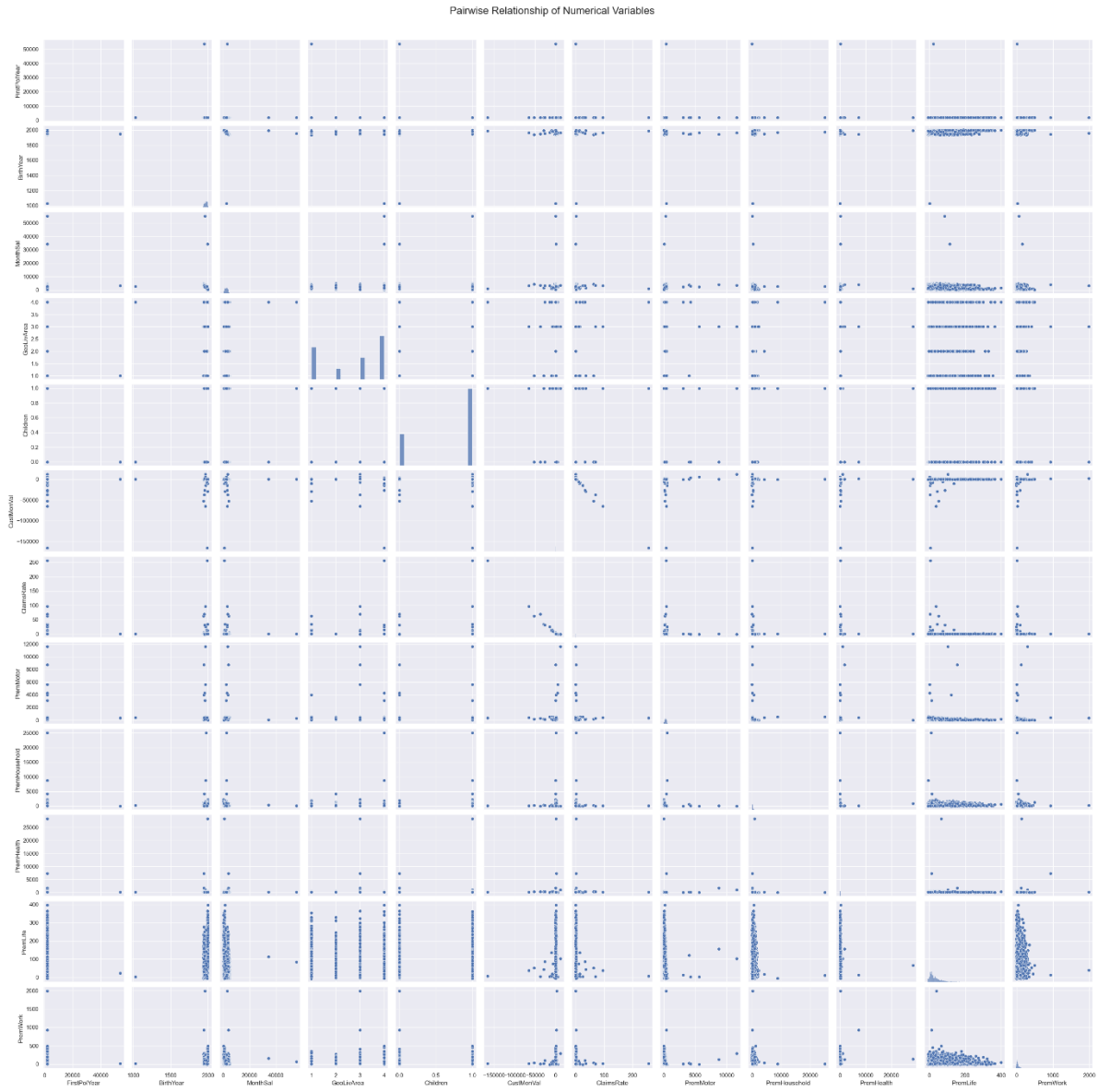

Figure 7 - Pairplot with all variables

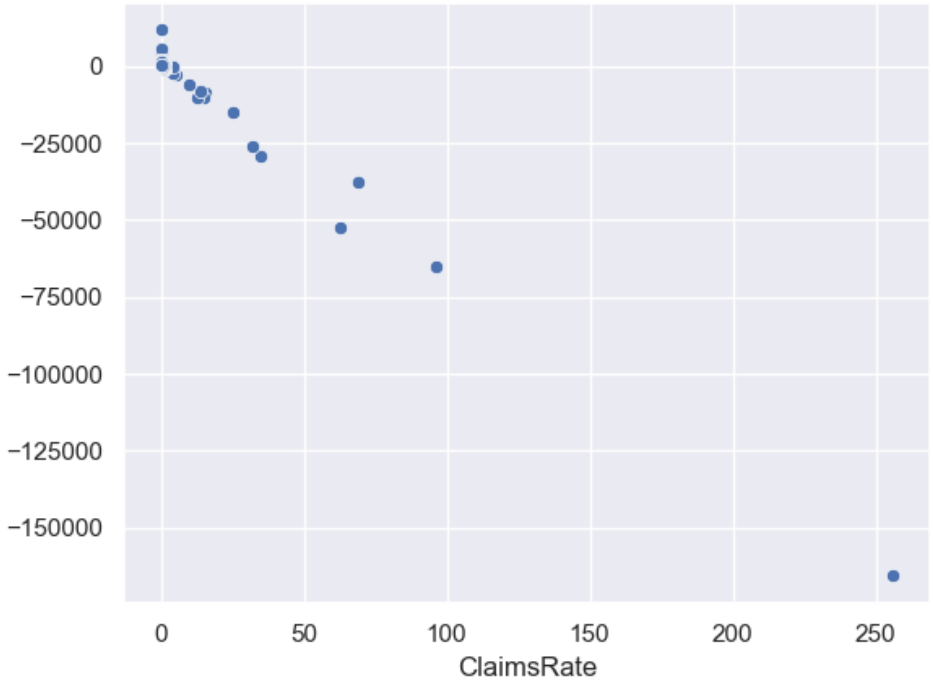

Figure 8 -CustMonVal vs. ClaimsRate

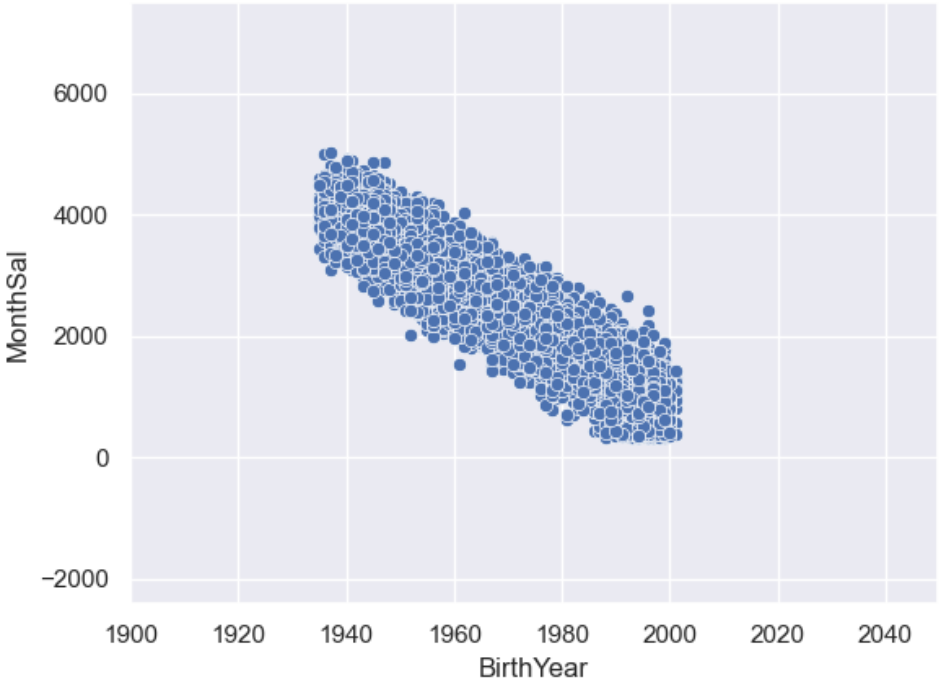

Figure 9 - BirthYear vs. MonthSal

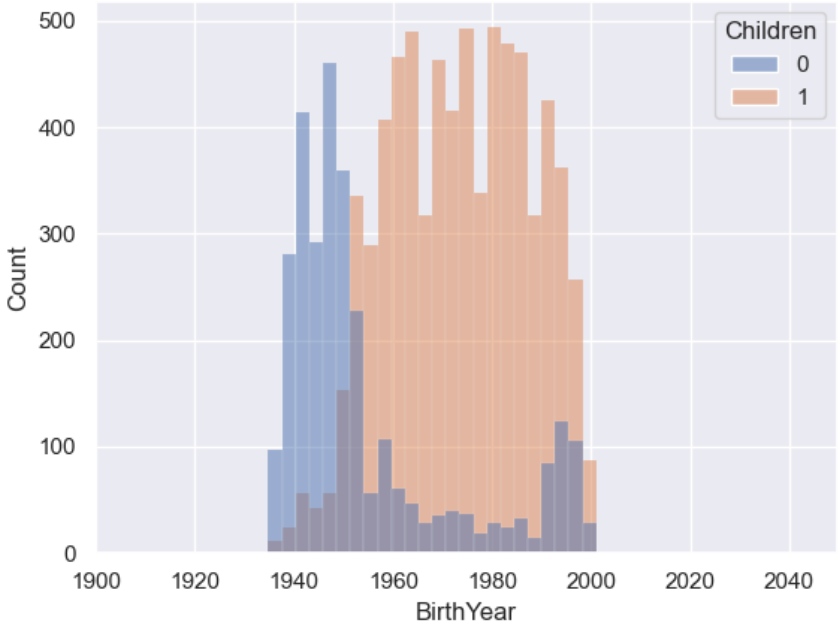

Figure 10 – Histogram for BirthYear and Children

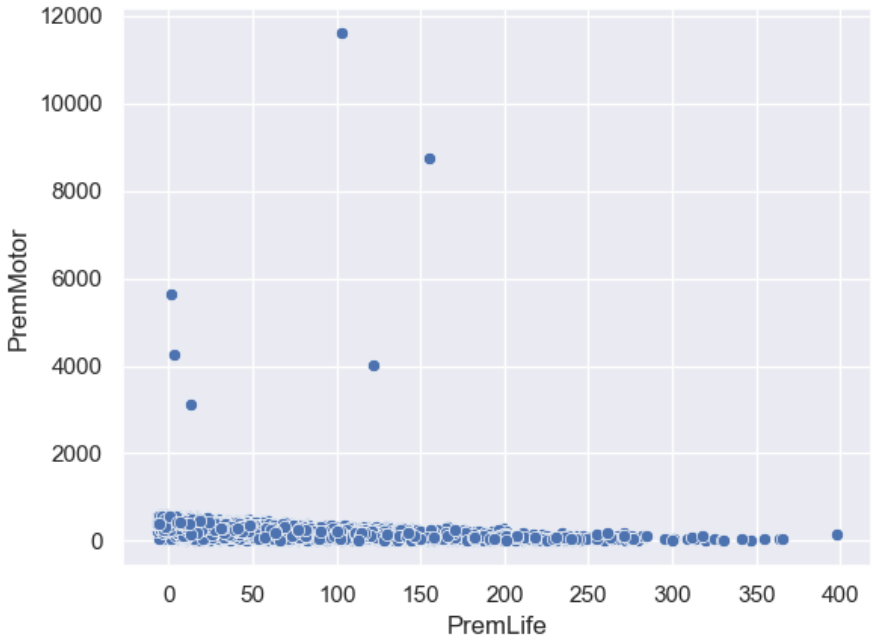

Figure 11 - PremLife vs. PremMotor

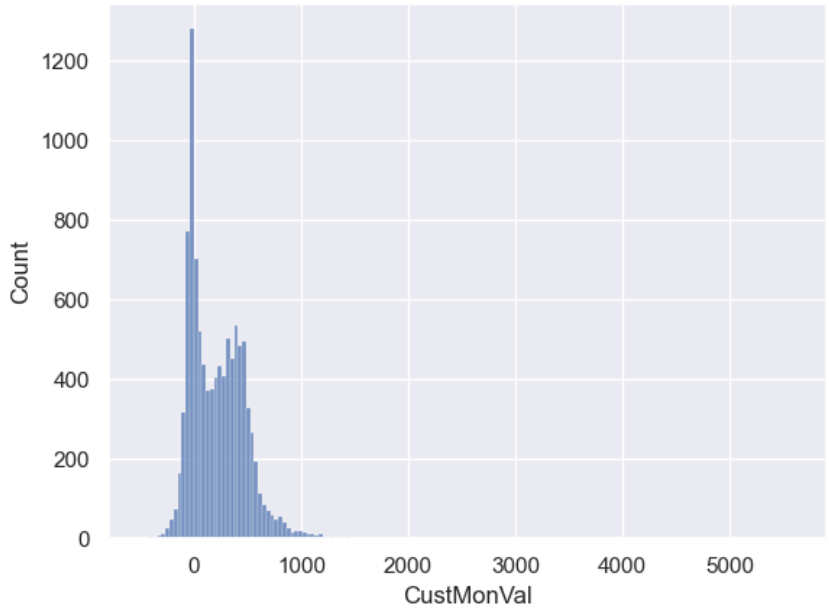

Figure 12 - CustMonVal without Extra Outliers

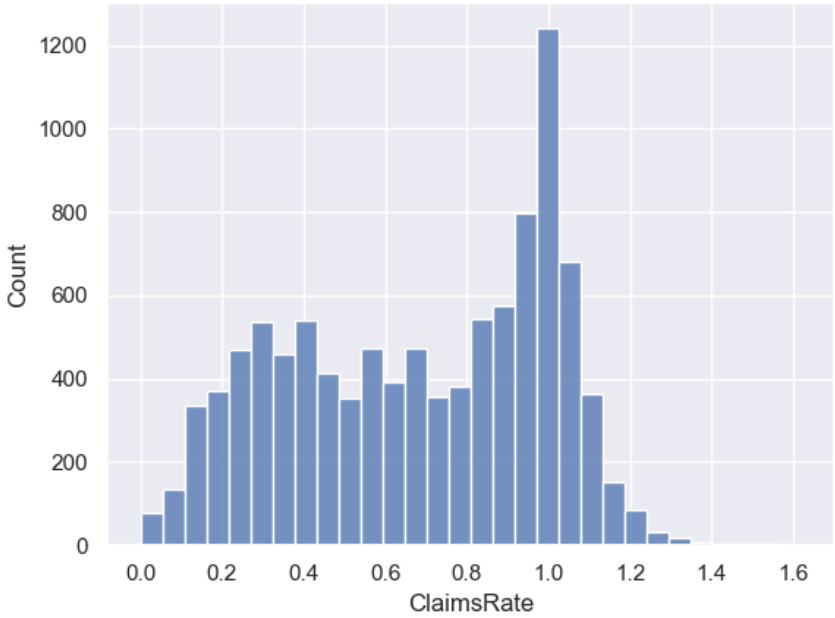

Figure 13 -ClaimsRate without extra Ouliers

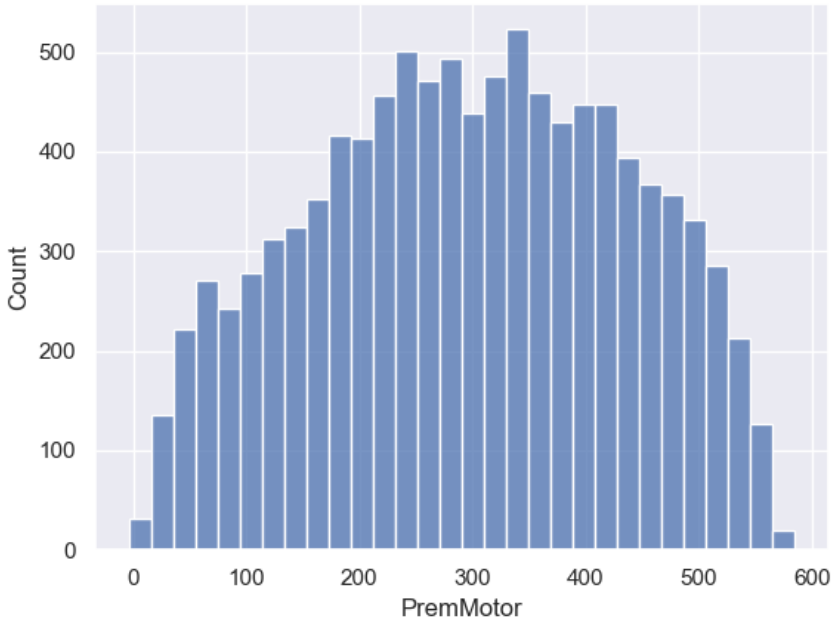

Figure 14 - PremMotor without extra outliers

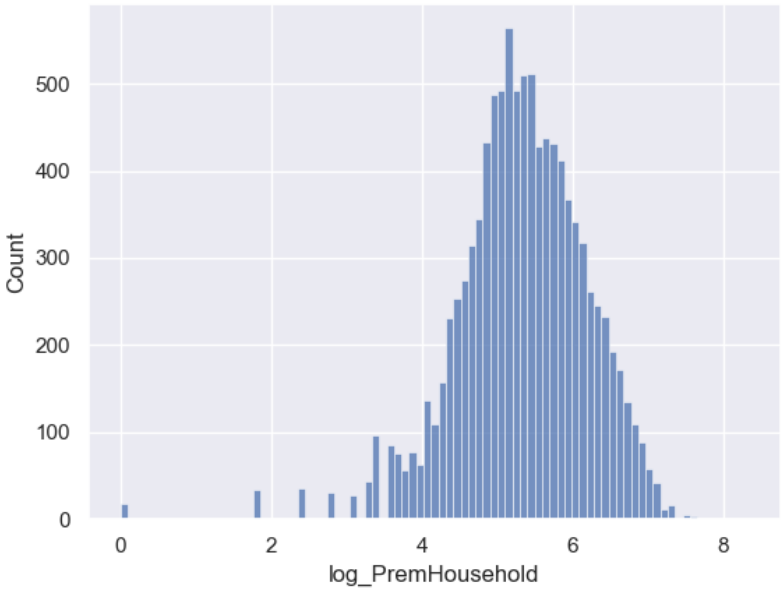

Figure 15 - Log transformation of PremHousehold

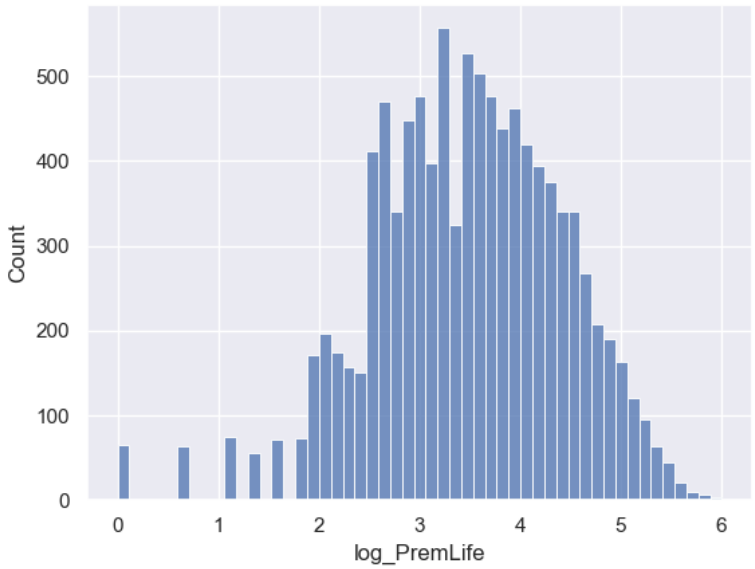

Figure 16 - Log transformation of PremLife

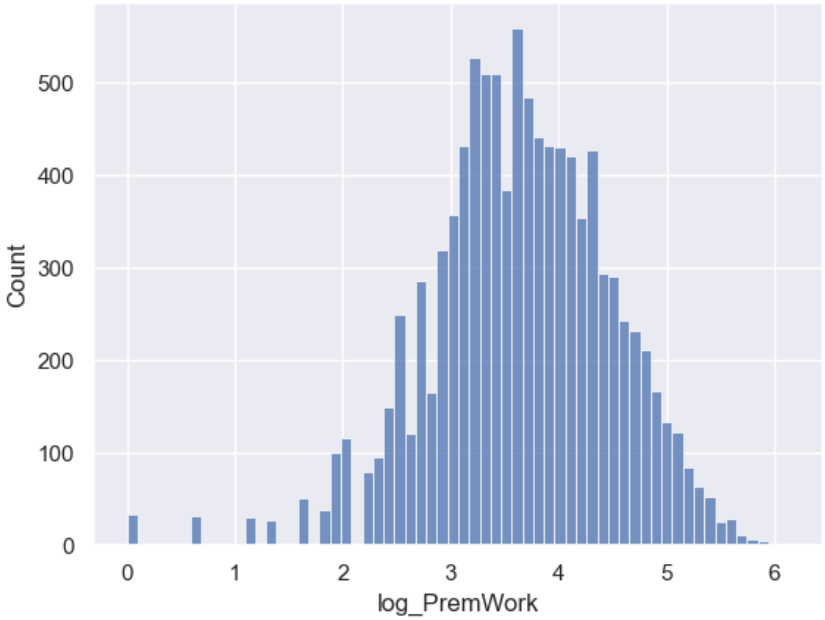

Figure 17 - Log transformation of PremWork

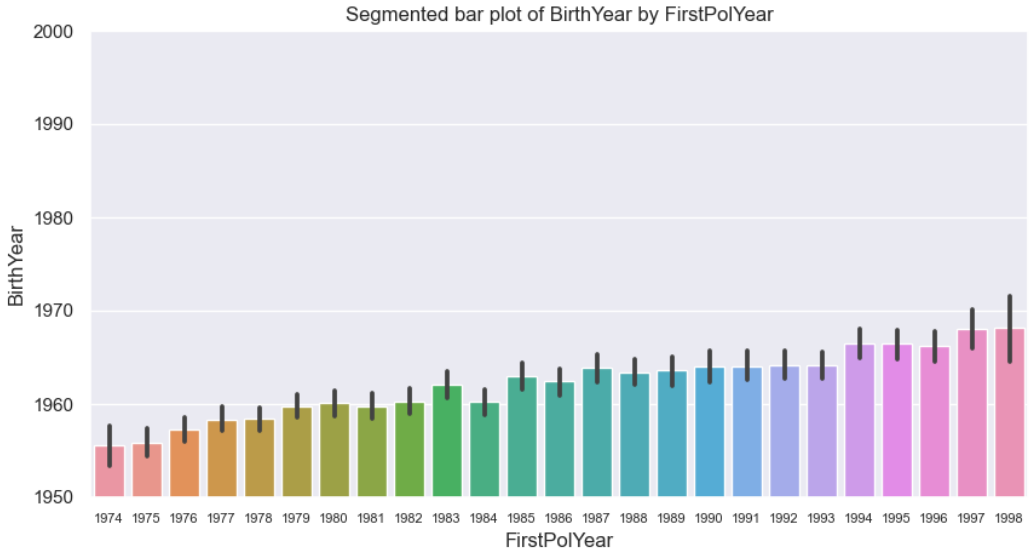

Figure 18 - Distribution before KNNImputer

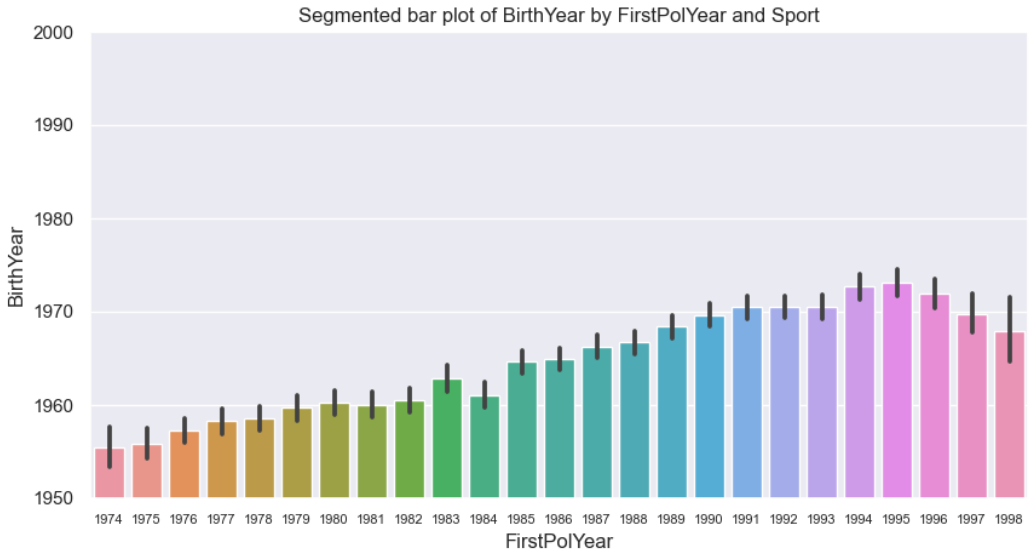

Figure 19 - Distribution after KNNImputer

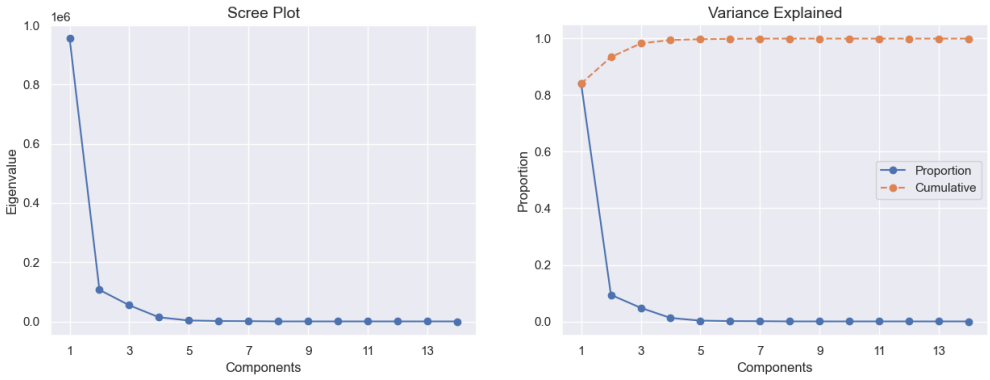

Figure 20 - Scree and Variance plots for PCA

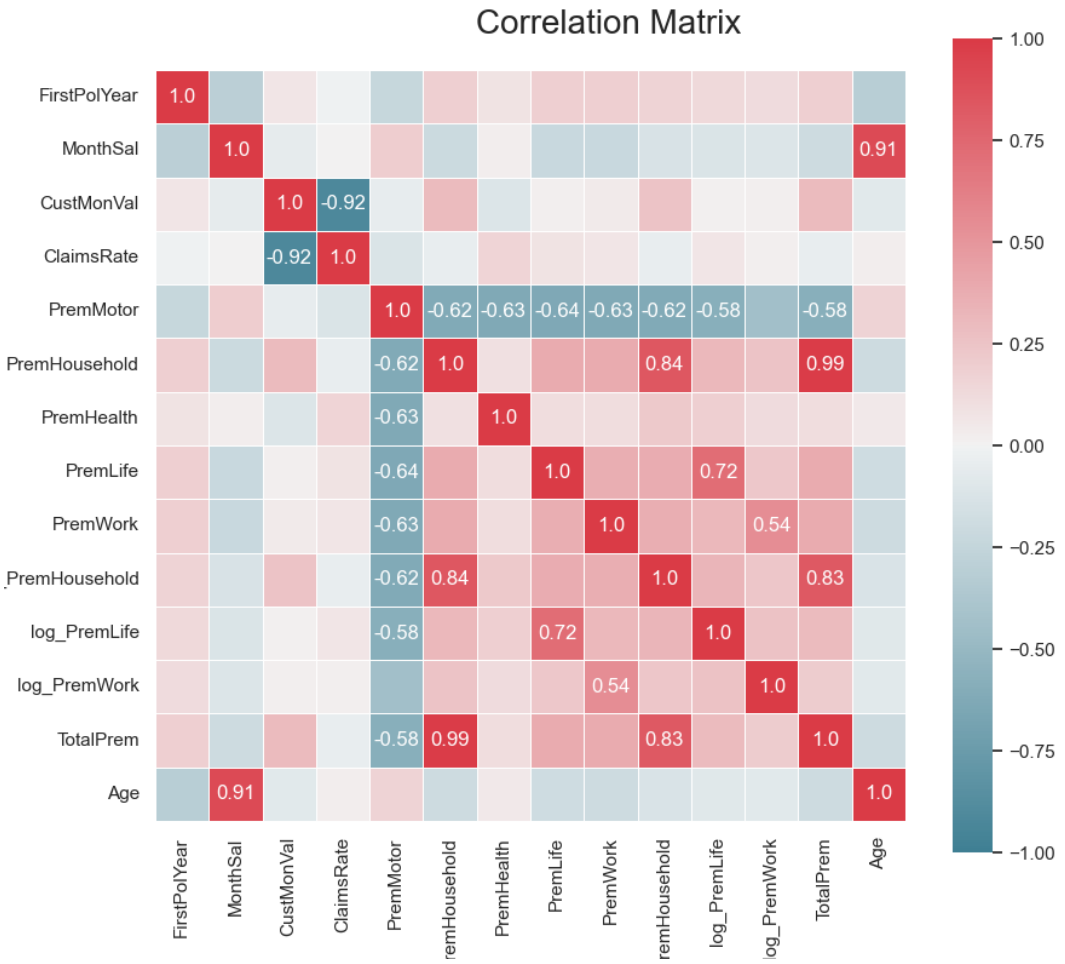

Figure 21 - Correlation Matrix

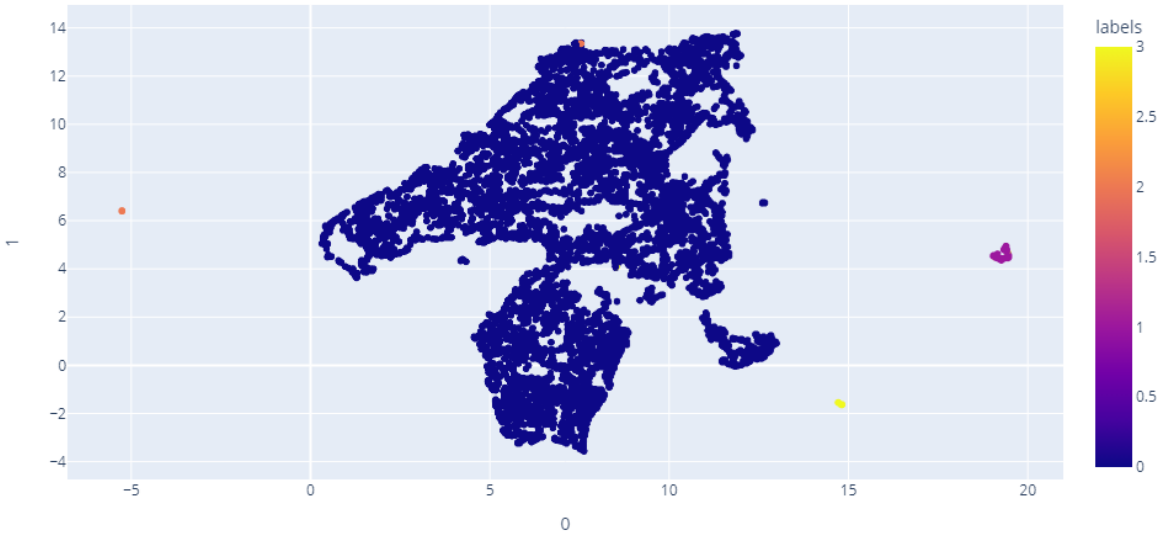

Figure 22 - 2D view of Mean Shift to remove outliers

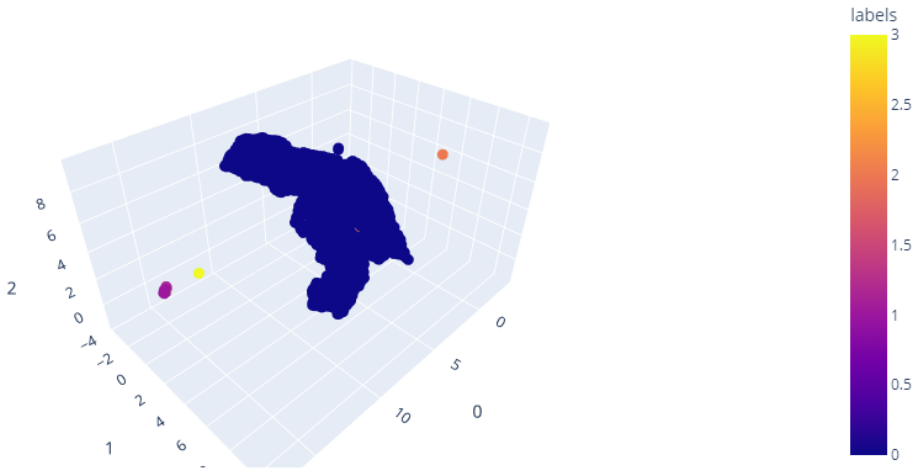

Figure 23 - 3D view of the Mean Shift to remove outliers

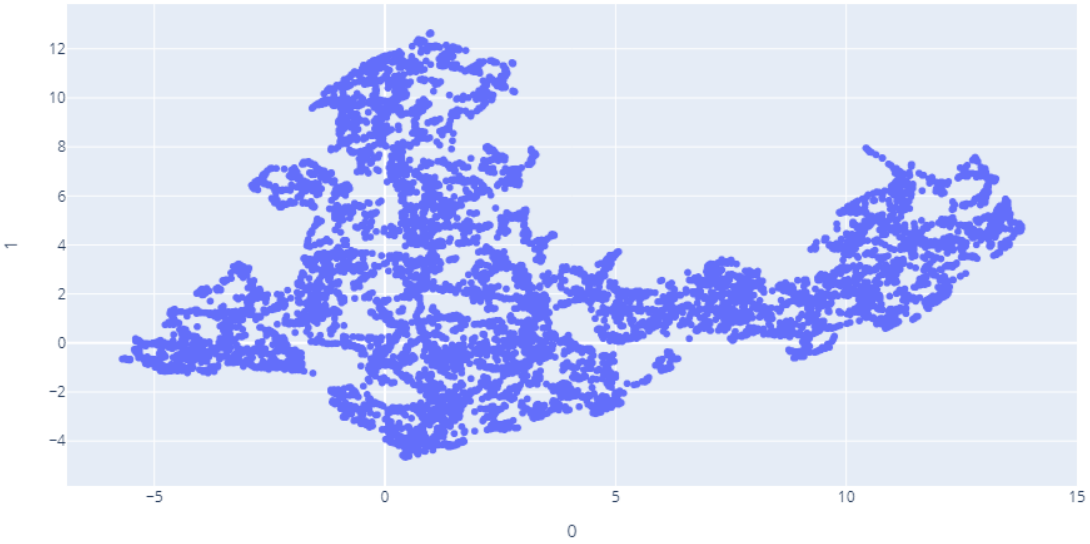

Figure 24 - 2D view after outlier removal

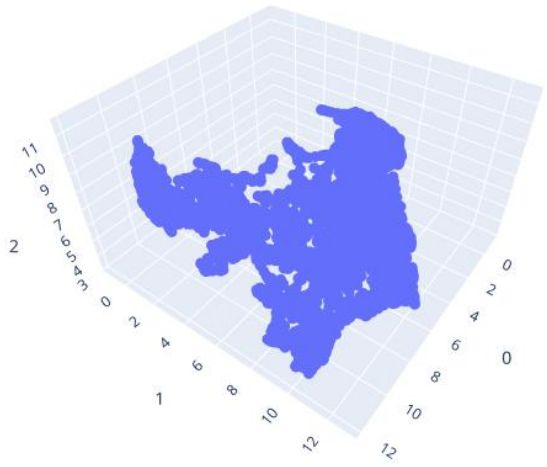

Figure 25 - 3D view after outlier removal

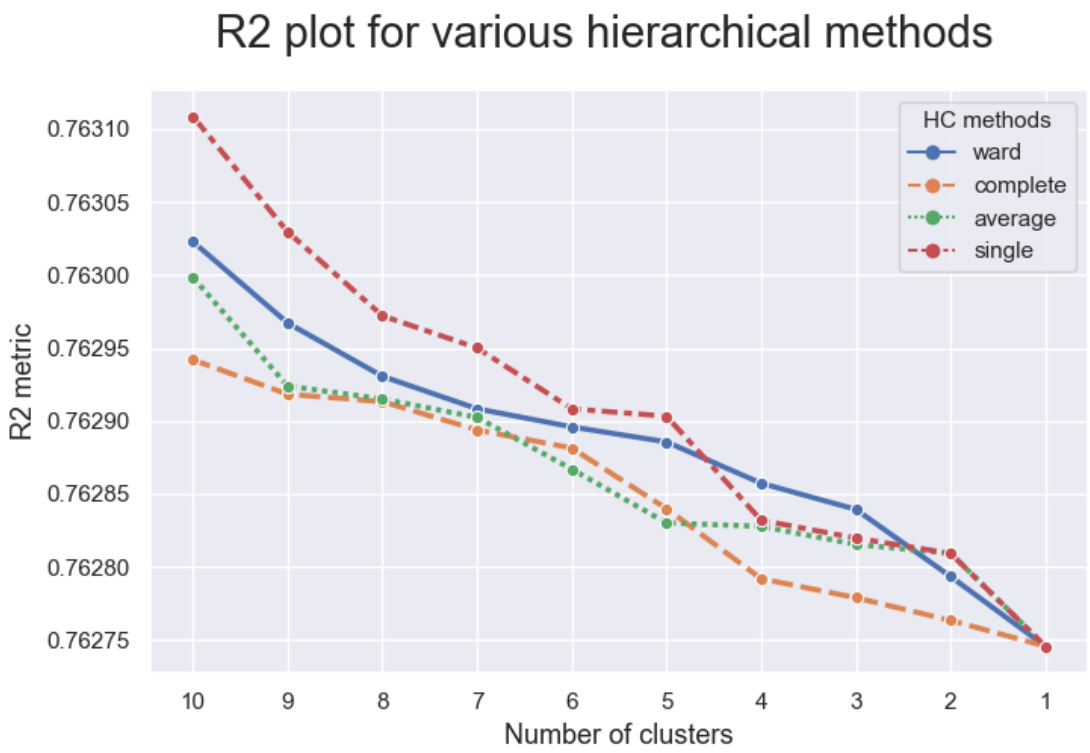

Figure 26 - R2 plot for Hierarchical Clustering of the Insurance perspective

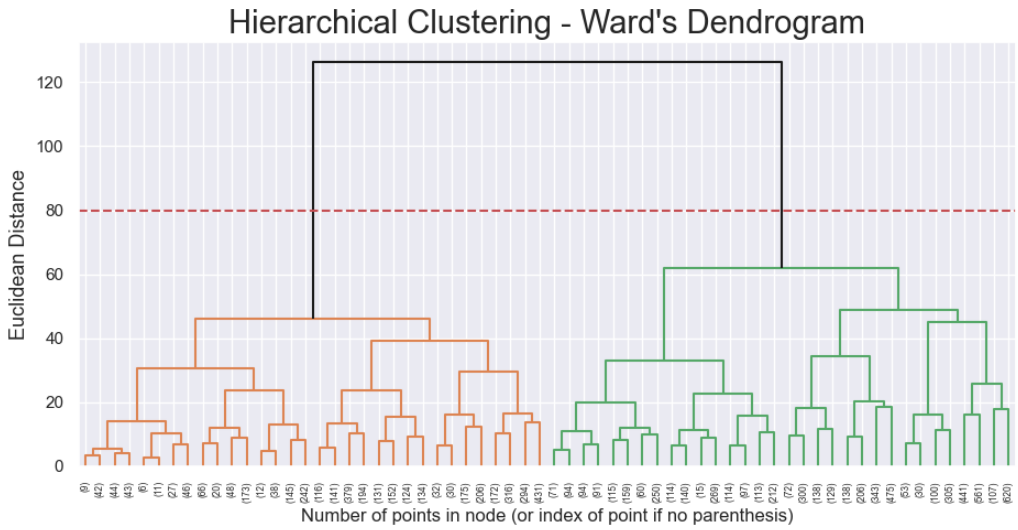

Figure 27 - Dendogram for Insurance perspective

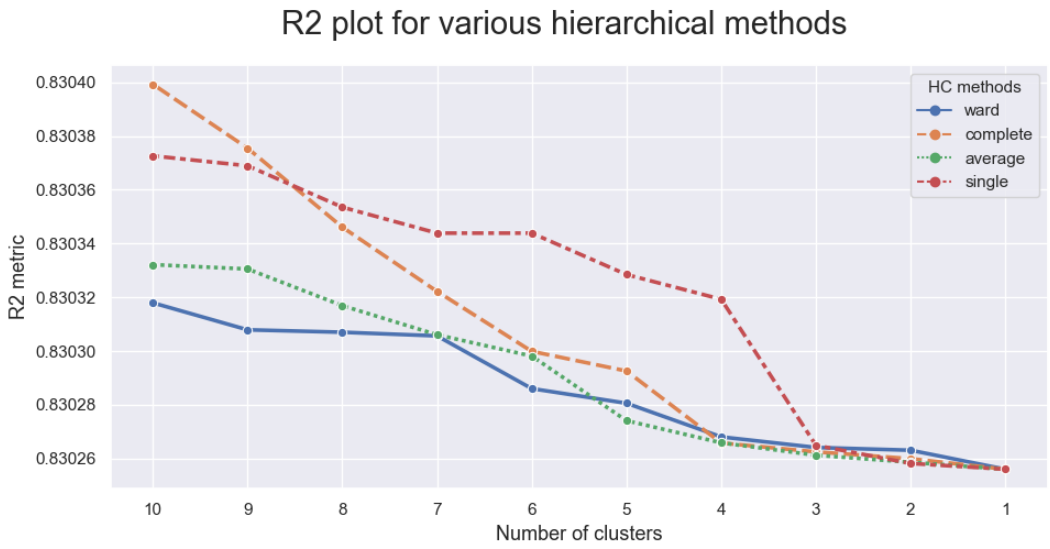

Figure 28 - R2 plot for Hierarchical Clustering of the Customer perspective

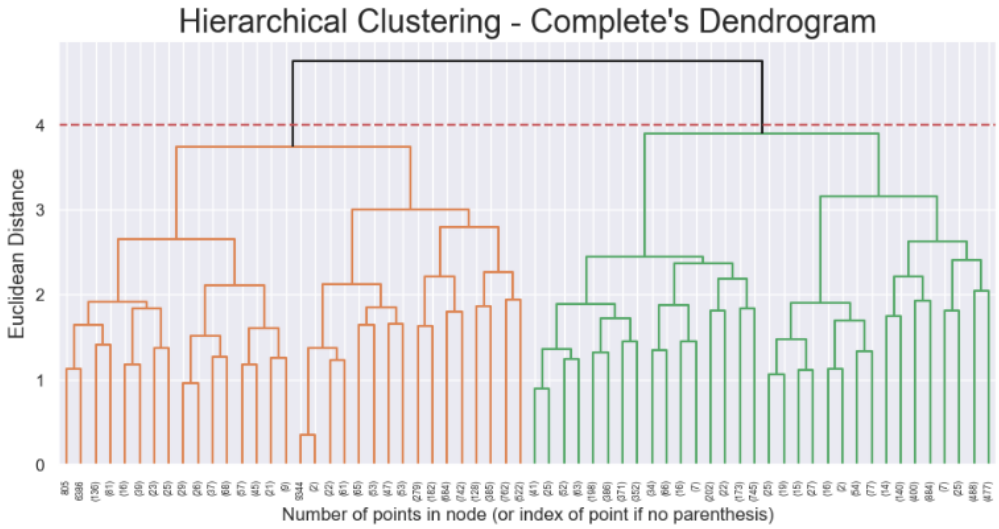

Figure 29 - Dendogram for Customer perspective

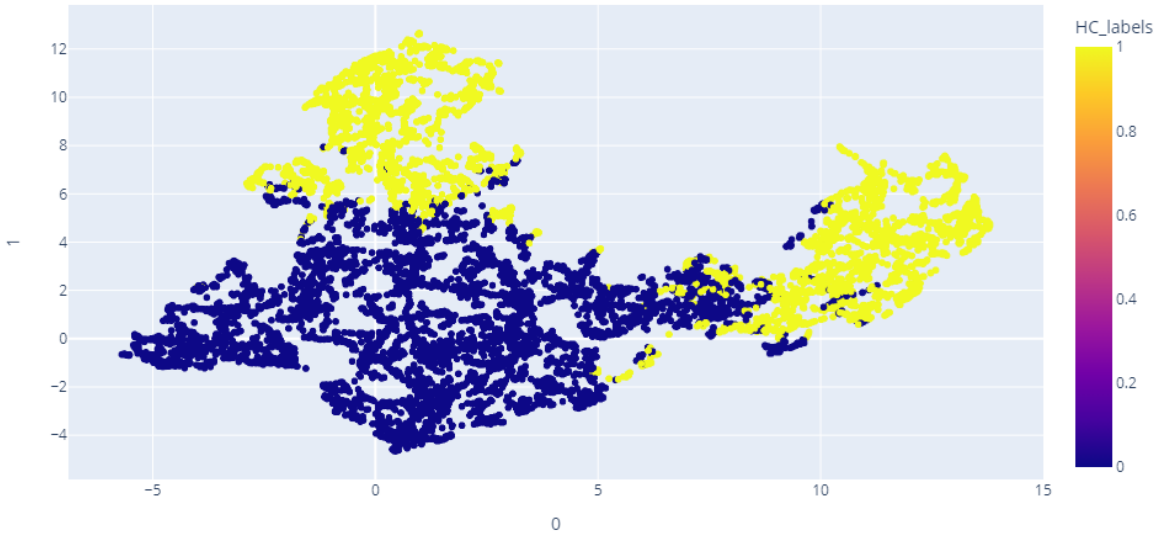

Figure 30 - 2D for Hierarchical Clustering of the Insurance perspective

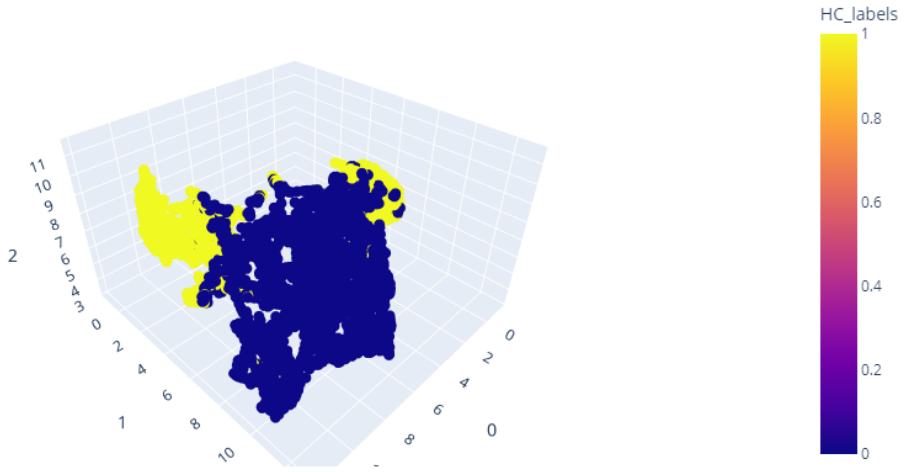

Figure 31 - 3D for Hierarchical Clustering of the Insurance perspective

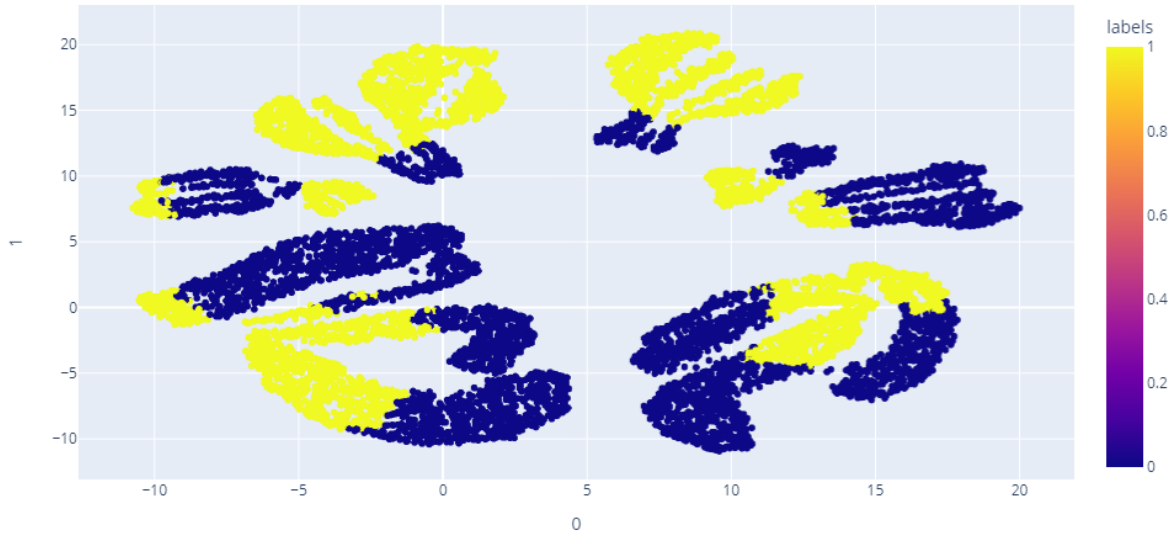

Figure 32 -2D for Hierarchical Clustering of the Customer perspective

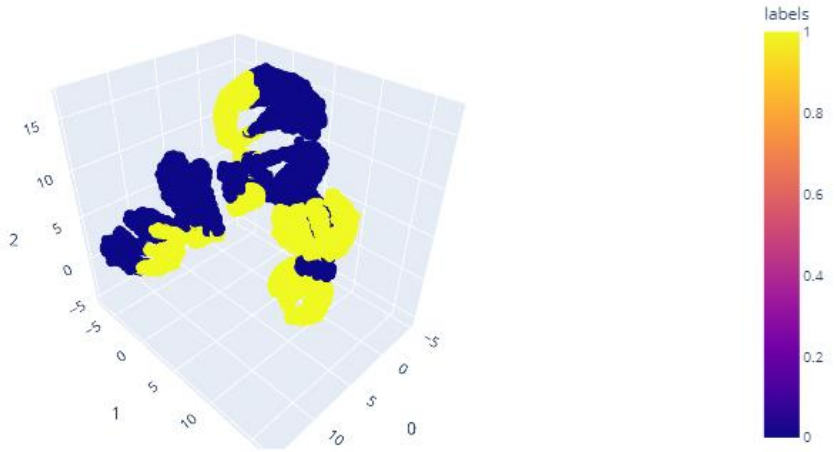

Figure 33 - 3D for Hierarchical Clustering of the Customer perspective

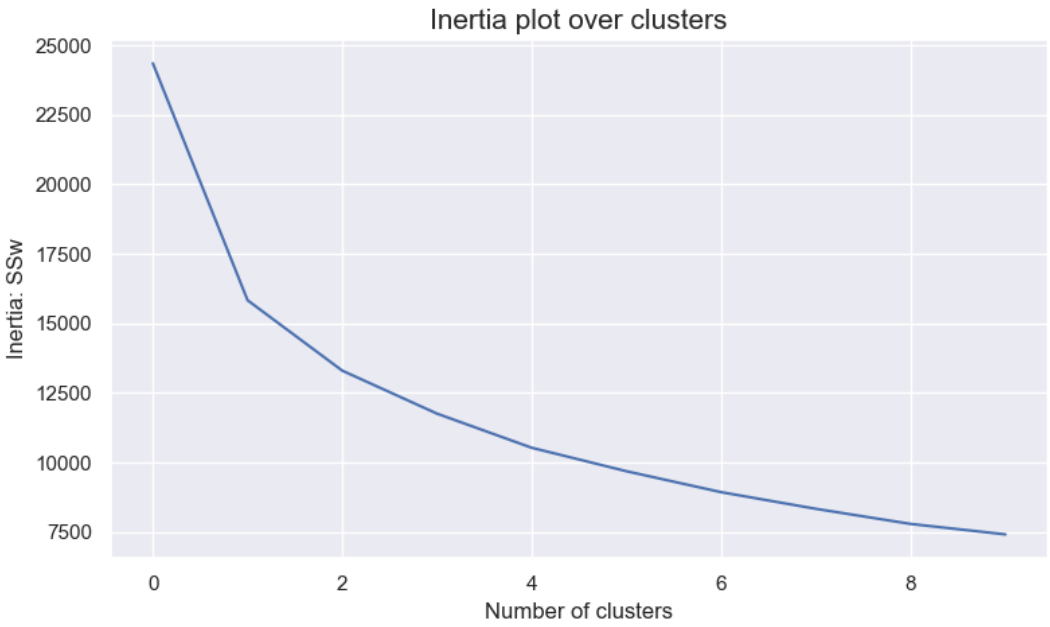

Figure 34 - Inertia for K-Means of the Insurance perspective

Figure 35 - Silhouette plot for K-Means of the Insurance perspective

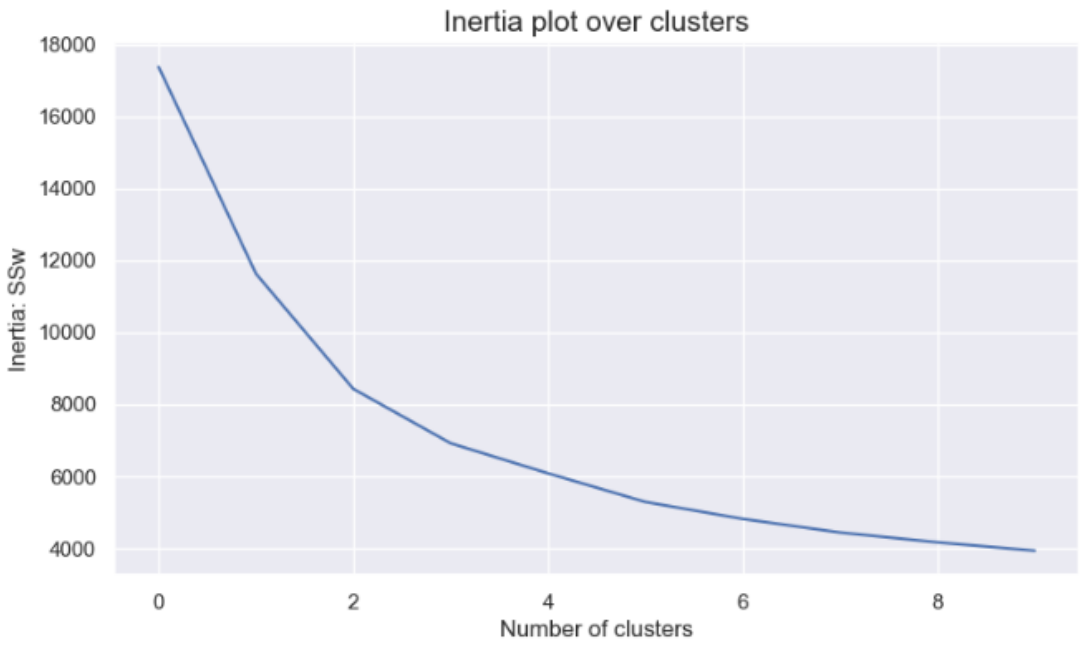

Figure 36 - Inertia for K-Means of the Customer perspective

Figure 37 - Silhouette plot for K-Means of the Customer perspective

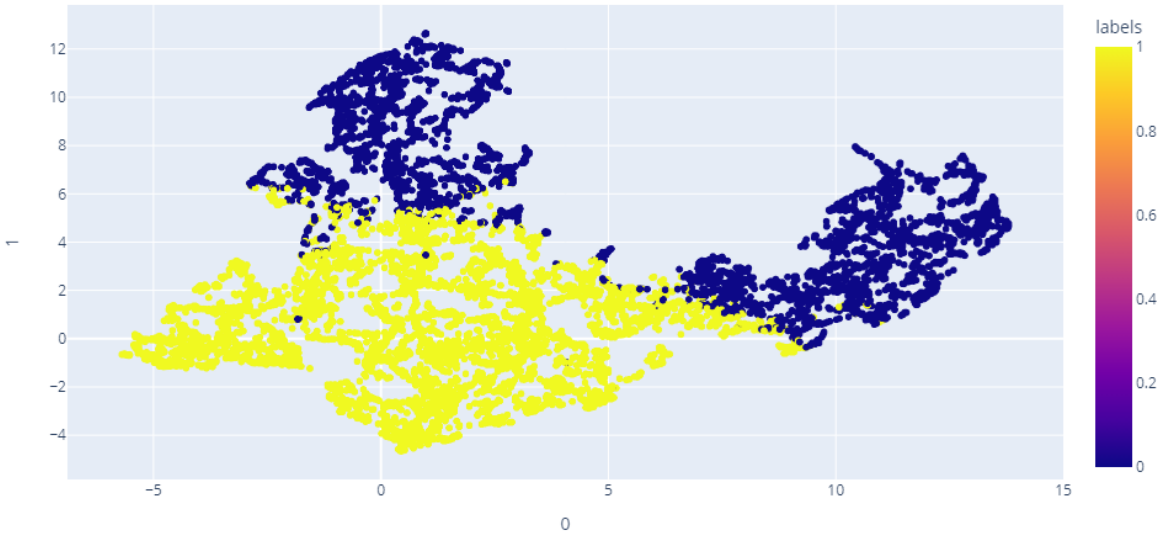

Figure 38 - 2D for K-Means of the Insurance perspective

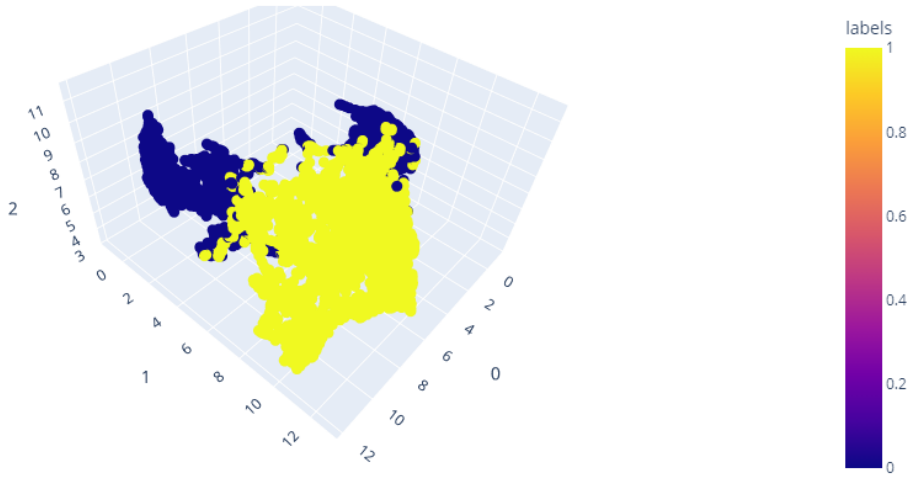

Figure 39 - 3D for K-Means of the Insurance perspective

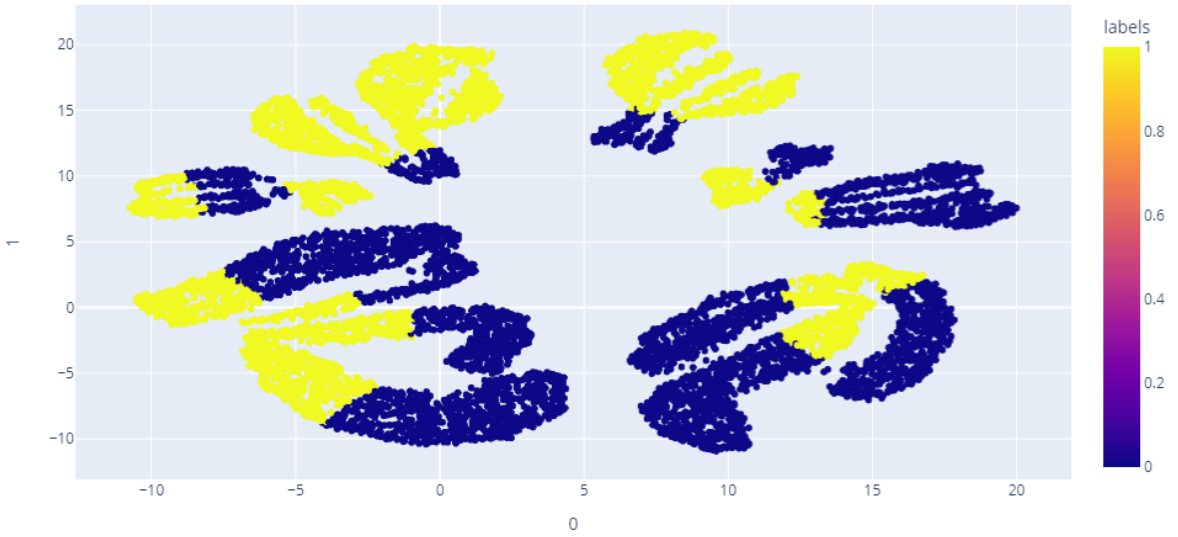

Figure 40 - 2D for K-Means of the Customer perspective

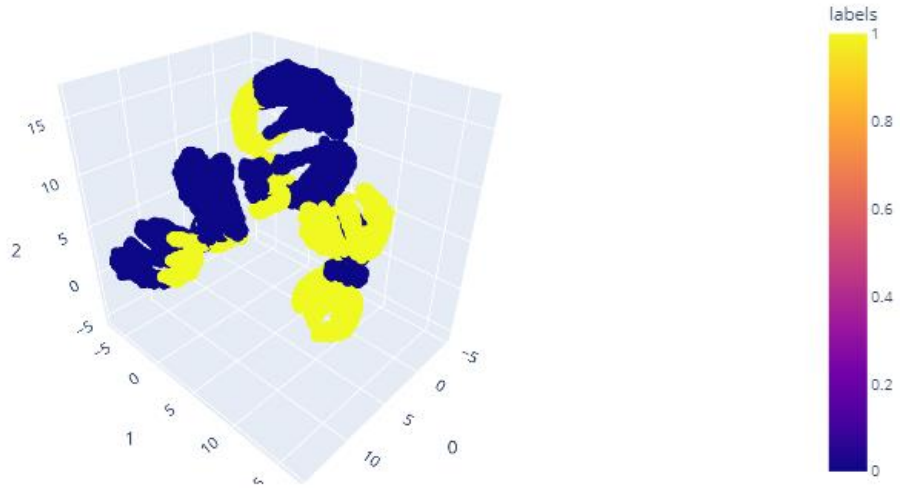

Figure 41 - 3D for K-Means of the Customer perspective

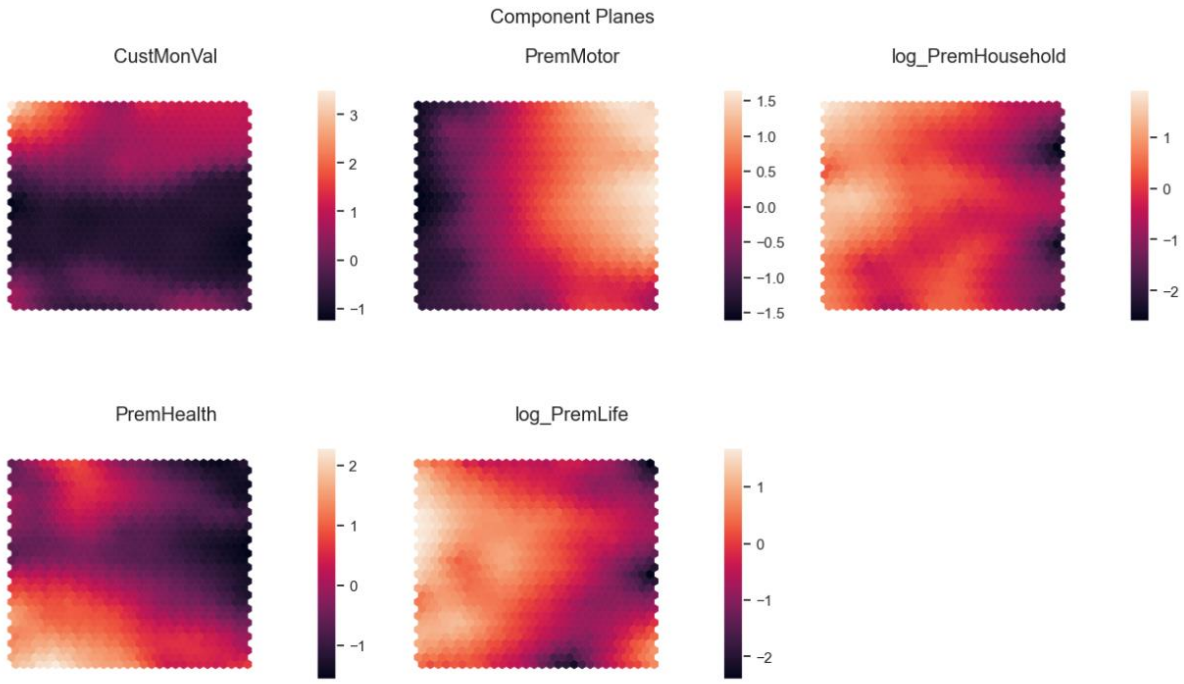

Figure 42 - Component Planes for Insurance Perspective

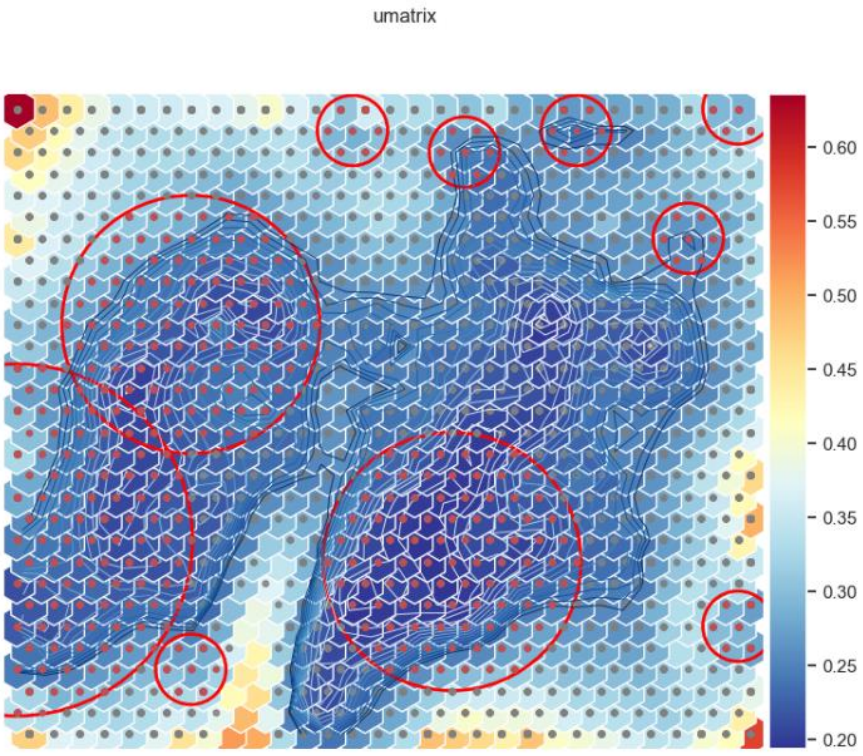

Figure 43 - U-Matrix for the Insurance Perspective

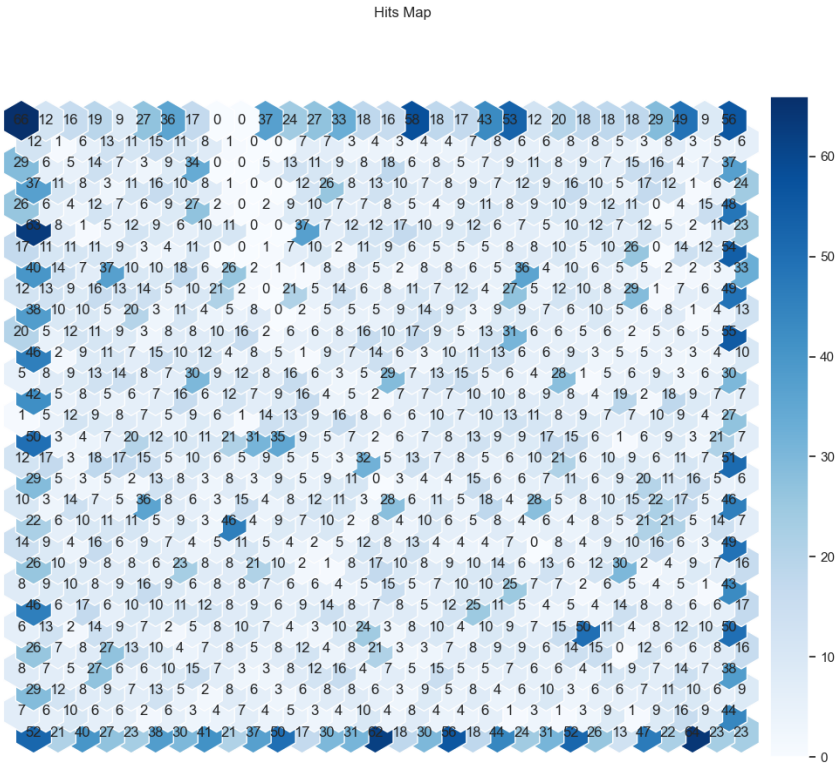

Figure 44 - Hit Map for the Insurance Perspective

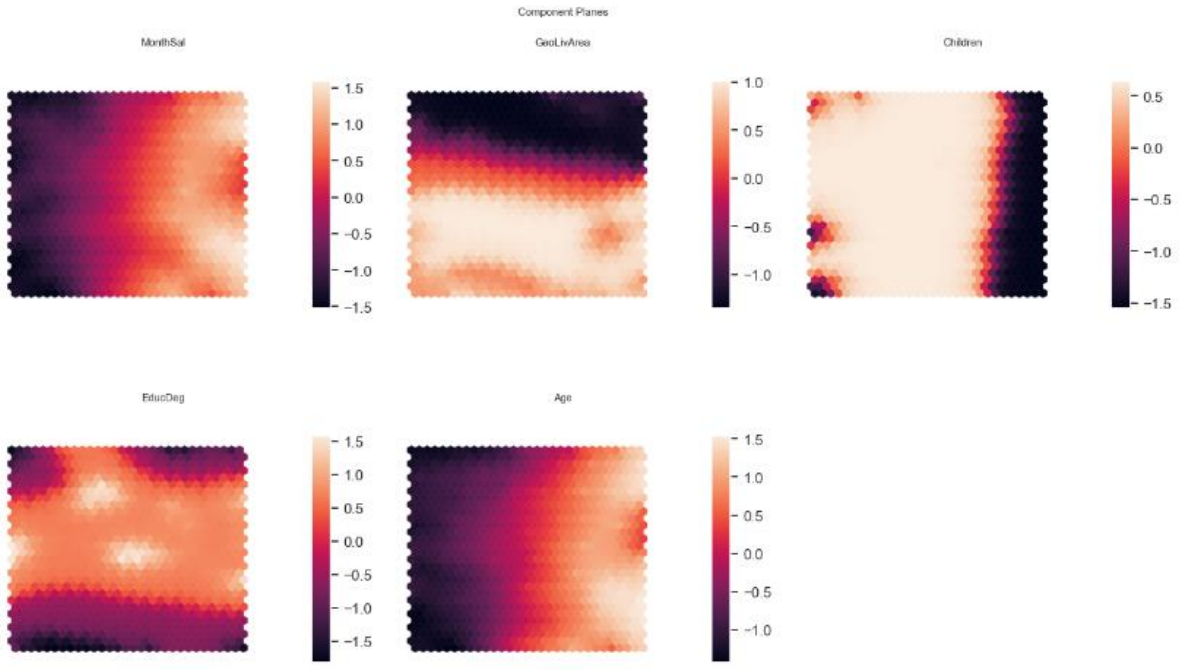

Figure 45 - Component Planes for Customer Perspective

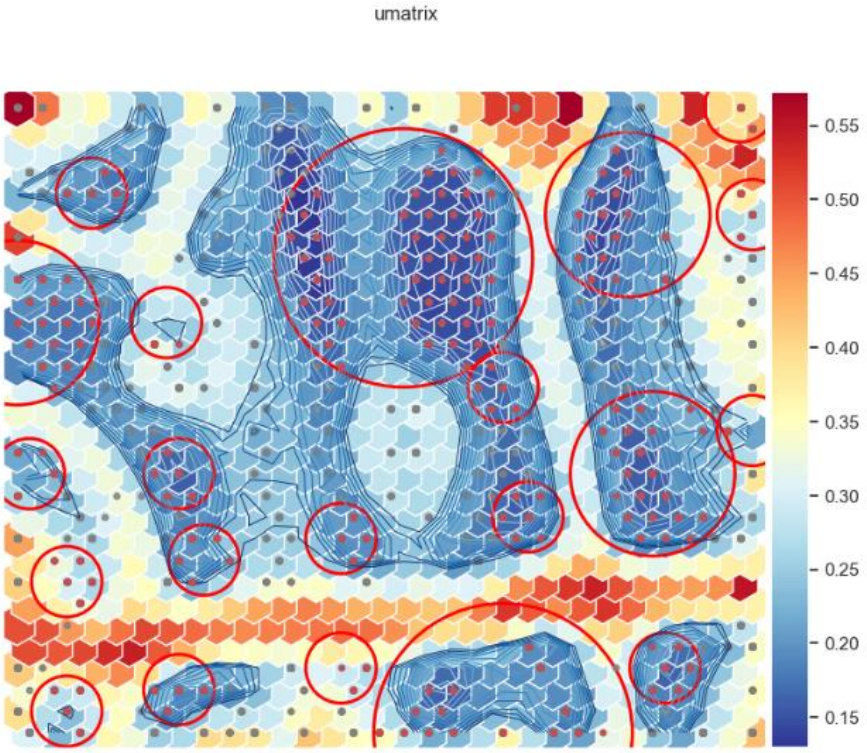

Figure 46 - U-Matrix for the Customer Perspective

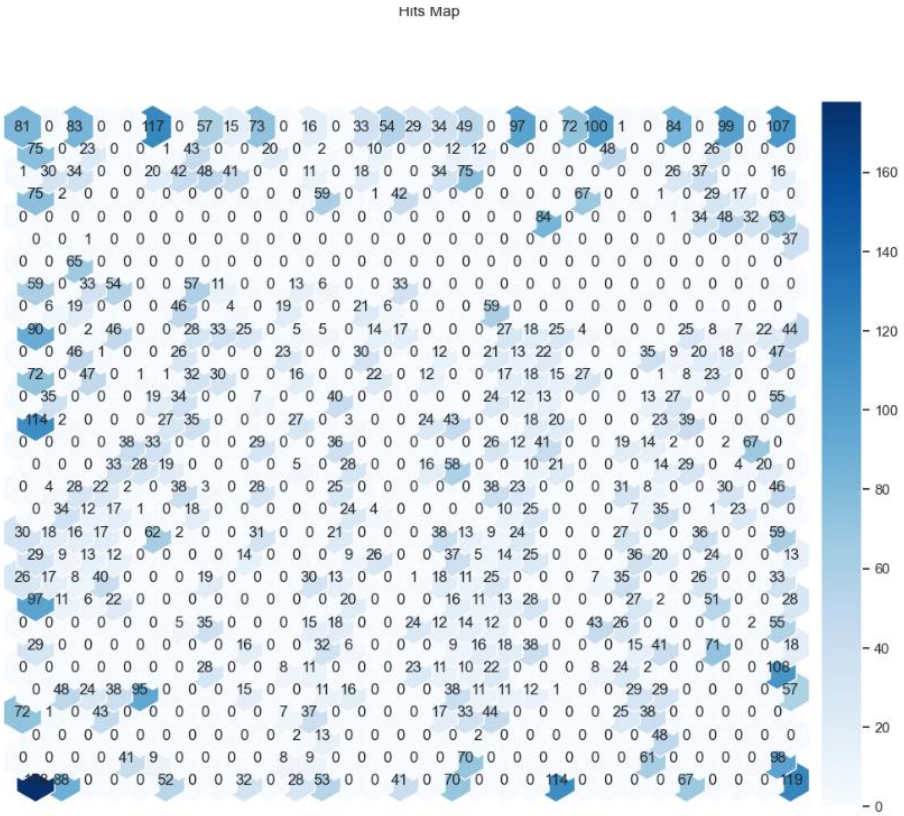

Figure 47 - Hit Map for the Customer Perspective

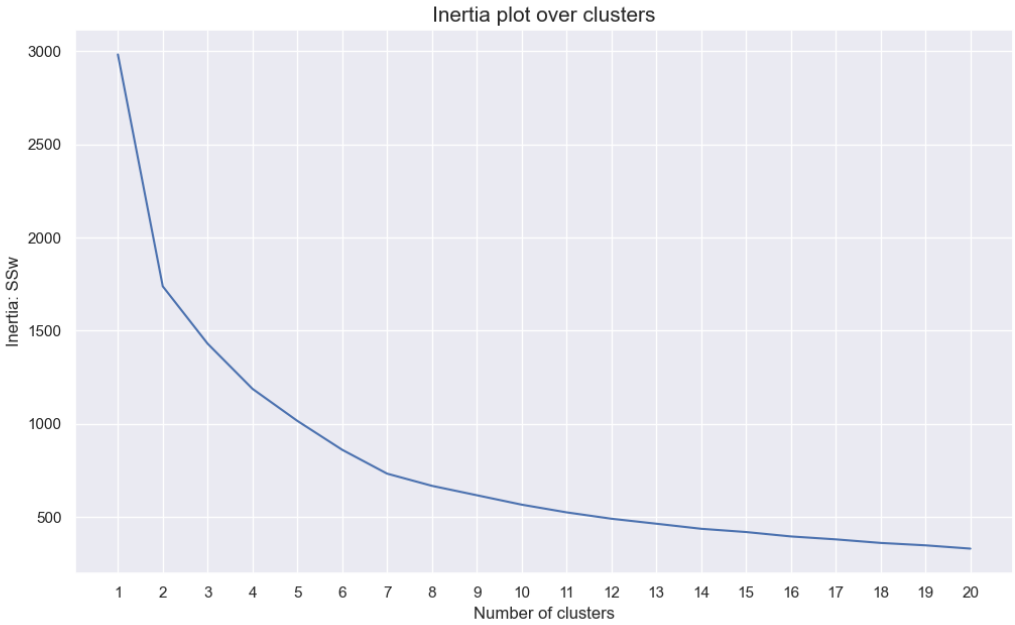

Figure 48 - Inertia for K-Means with SOM for Insurance perspective

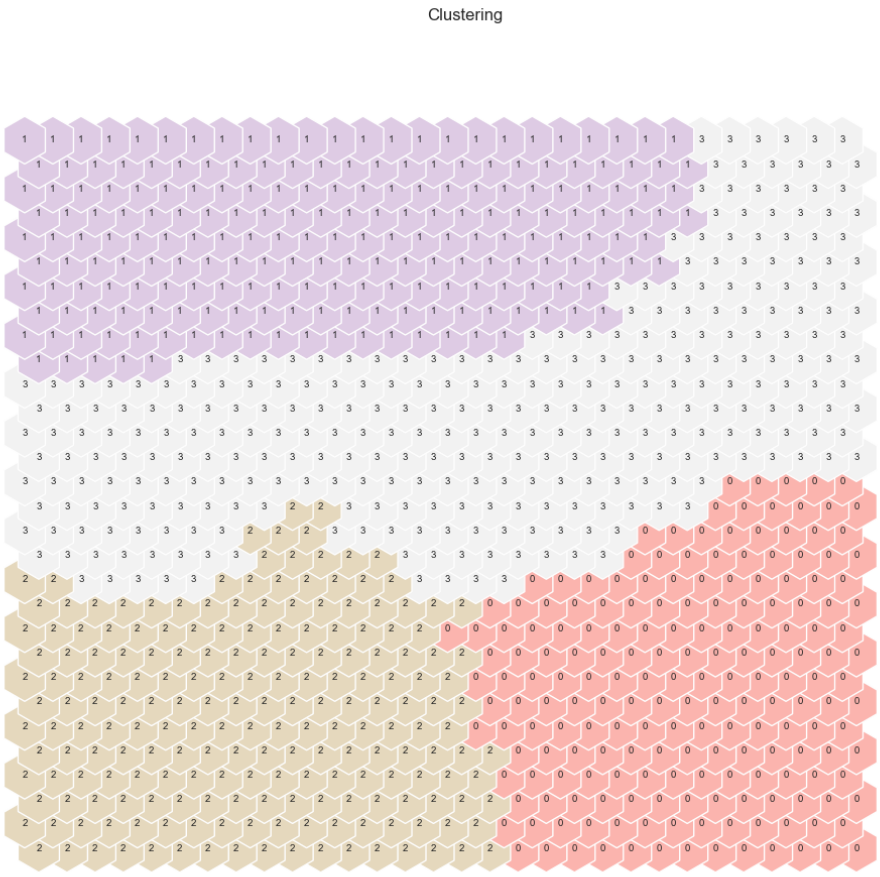

Figure 49 - Hit Map for K-Means with SOM for Insurance Perspective

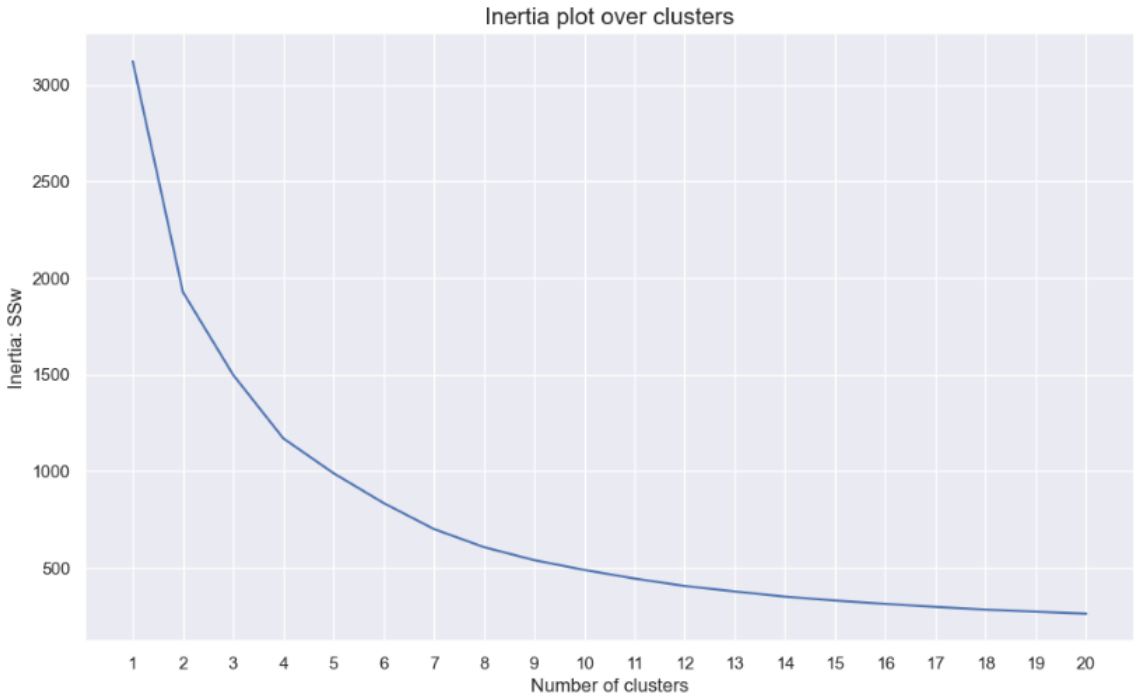

Figure 50 - Inertia for K-Means with SOM for Customer perspective

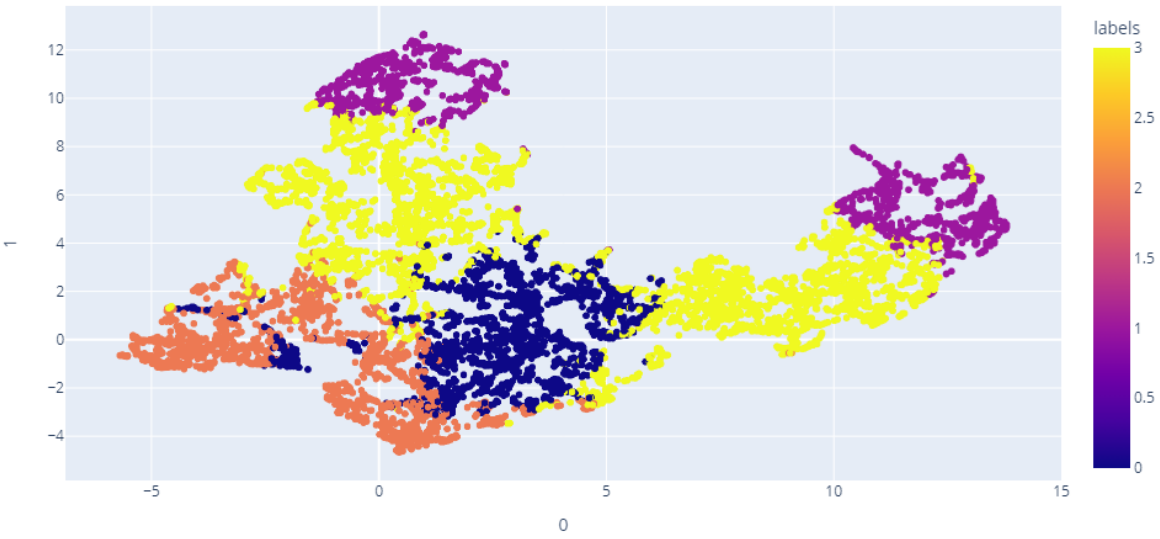

Figure 51- 2D for K-Means with SOM of the Insurance perspective

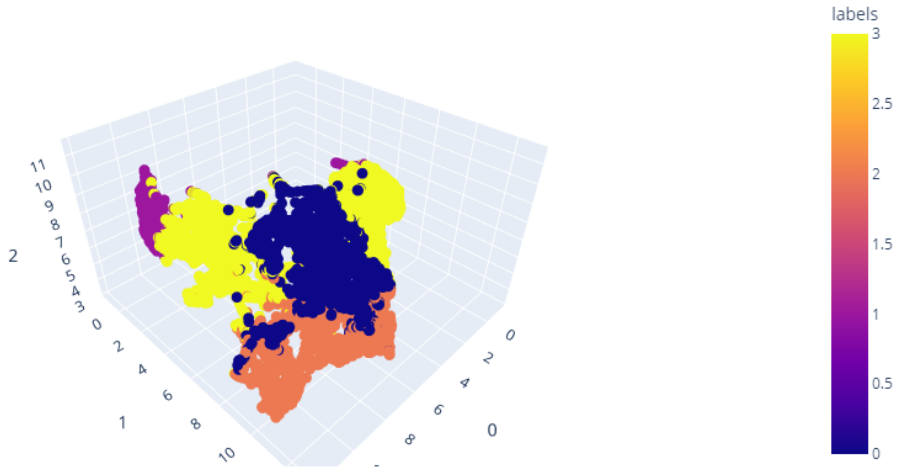

Figure 52 - 3D for K-Means with SOM of the Insurance perspective

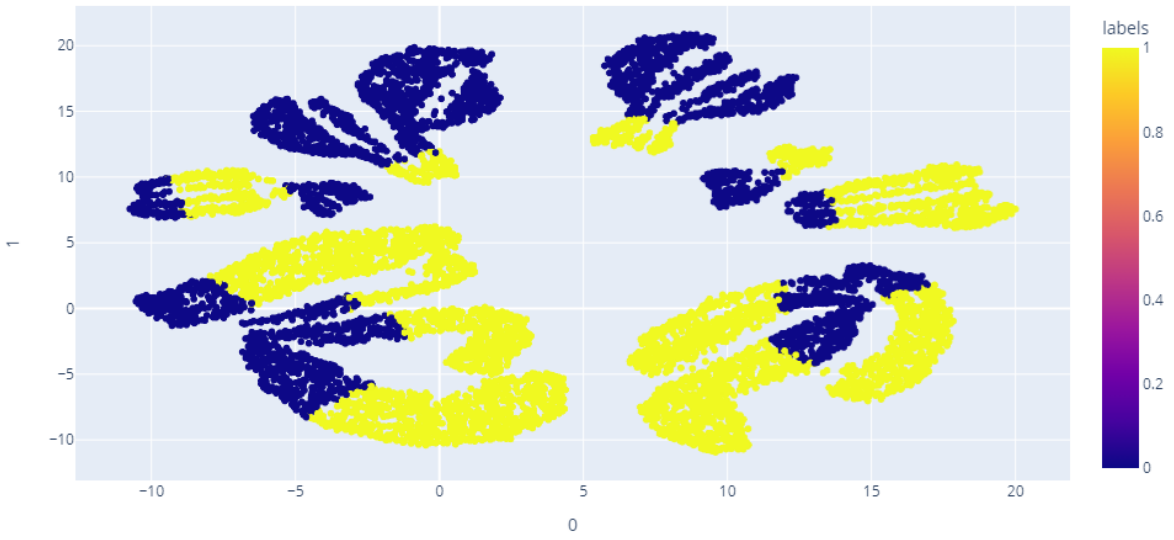

Figure 53- 2D for K-Means with SOM of the Customer perspective

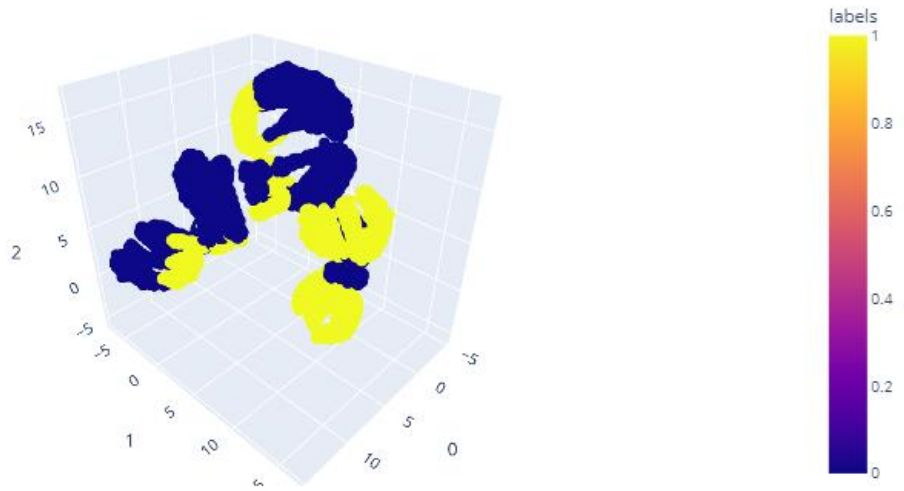

Figure 54 - 3D for K-Means with SOM of the Customer perspective

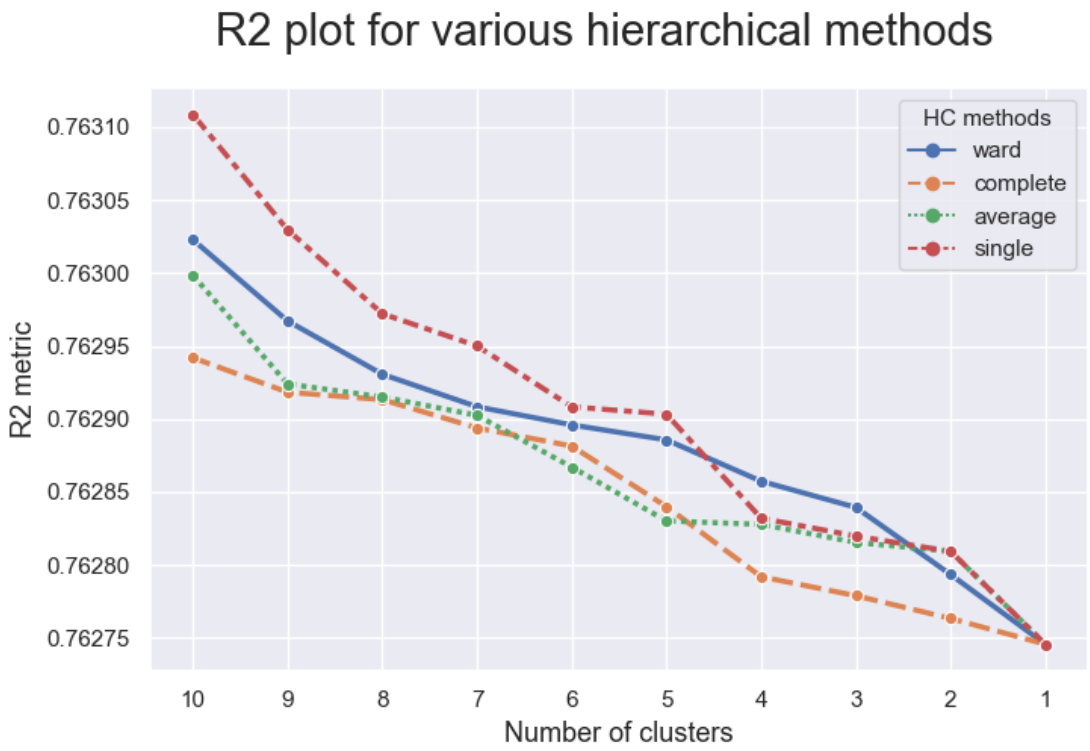

Figure 55 - R2 plot for Hierarchical Clustering with SOM of the Insurance perspective

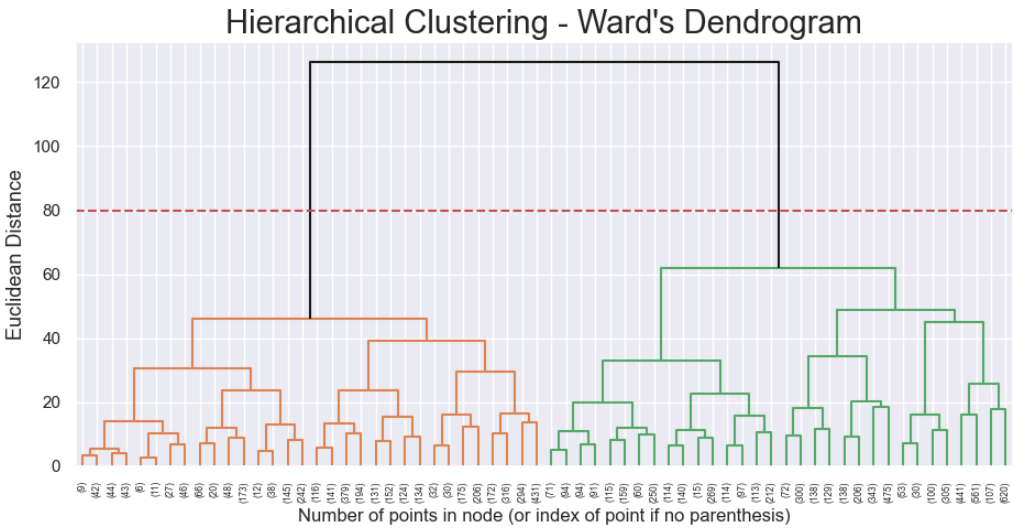

Figure 56 - Dendogram for Insurance perspective (HC with SOM)

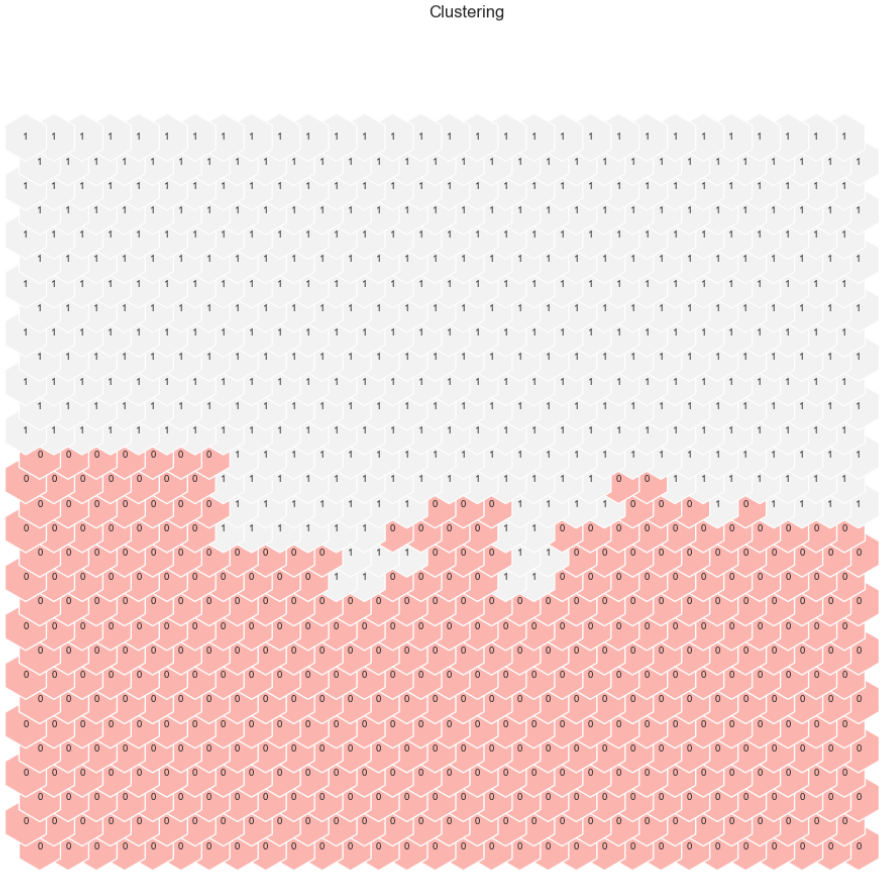

Figure 57 - Hit Map for Hierarchical with SOM for Insurance Perspective

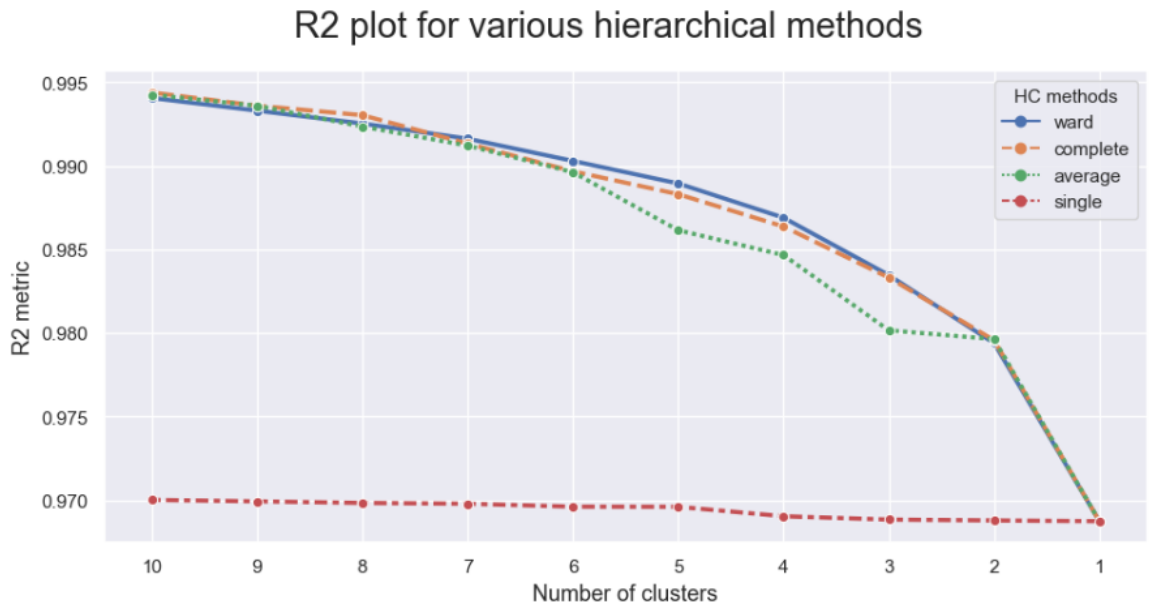

Figure 58 - R2 plot for Hierarchical Clustering with SOM of the Customer perspective

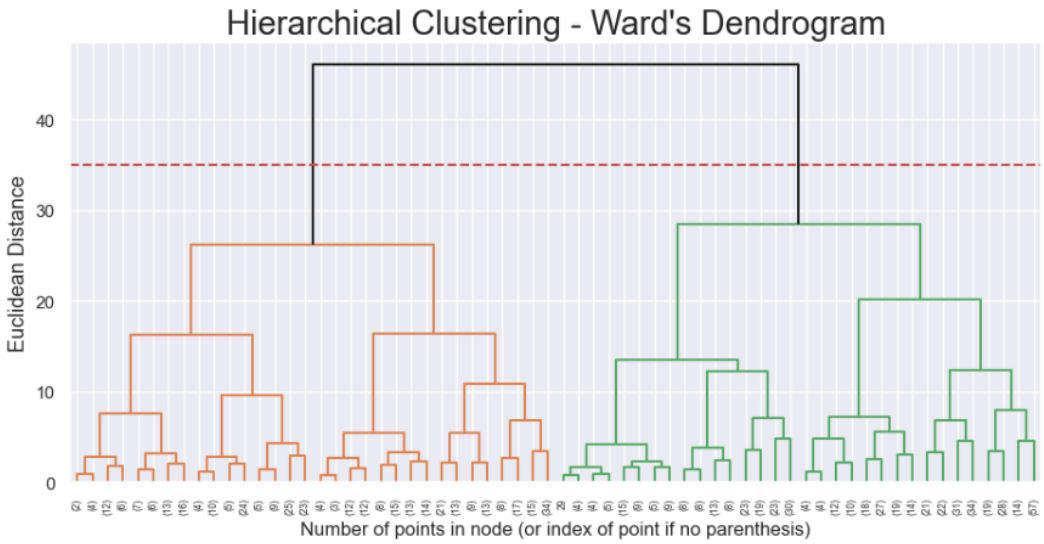

Figure 59 - Dendogram for Customer perspective (HC witb SOM)

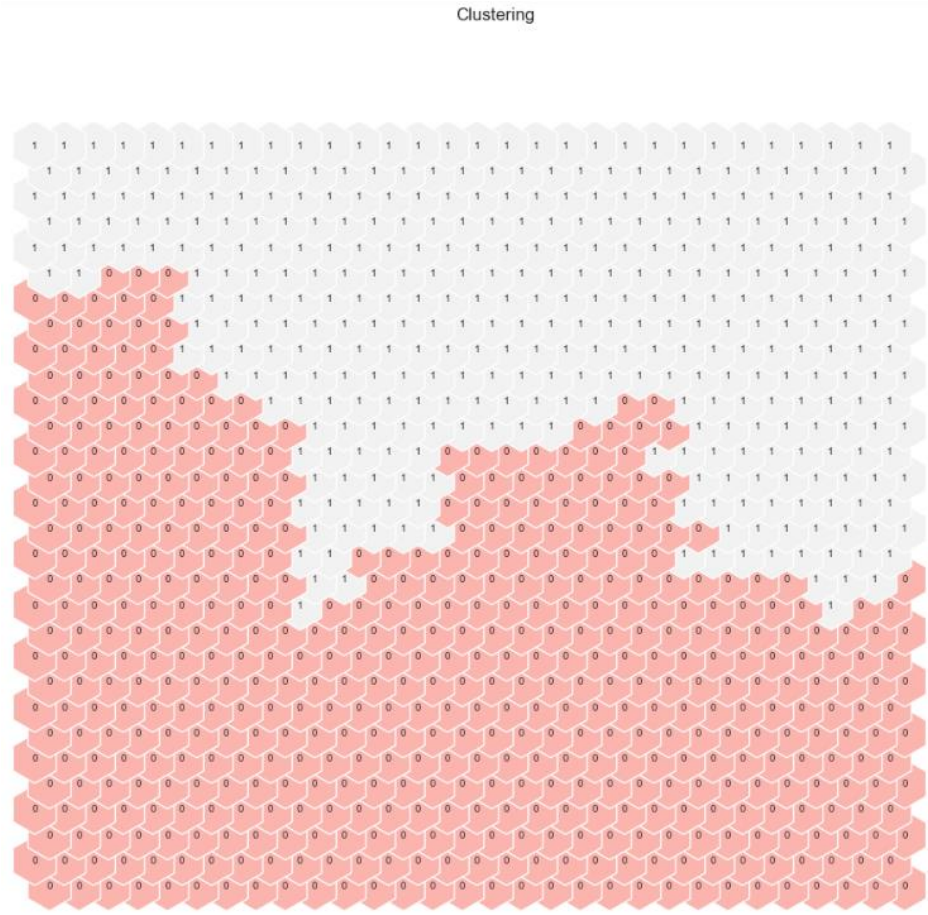

Figure 61 - Hit Map for Hierarchical with SOM for Customer Perspective

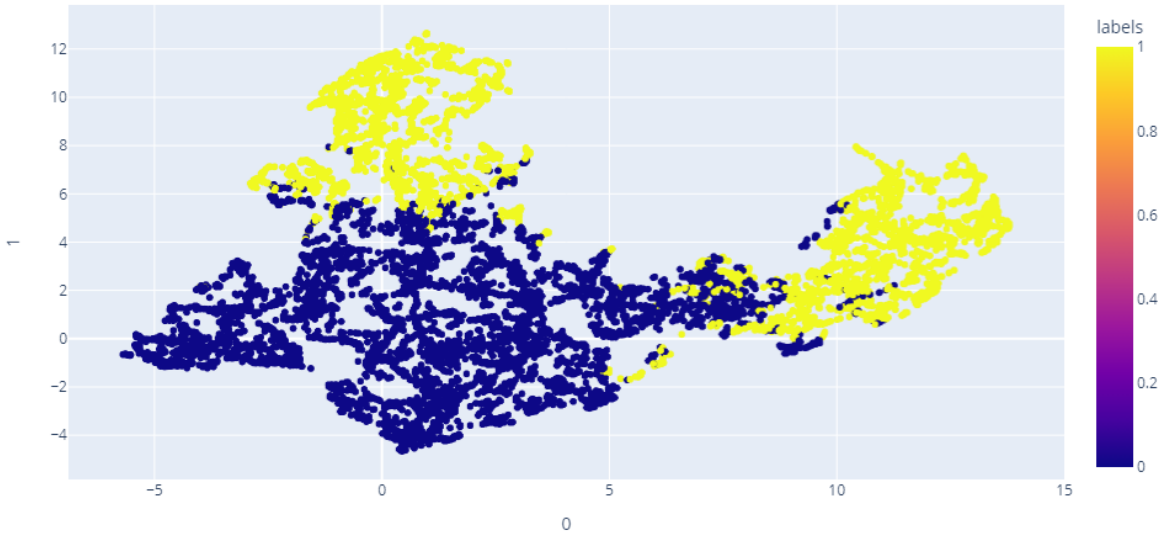

Figure 62- 2D for K-Means with SOM of the Customer perspective

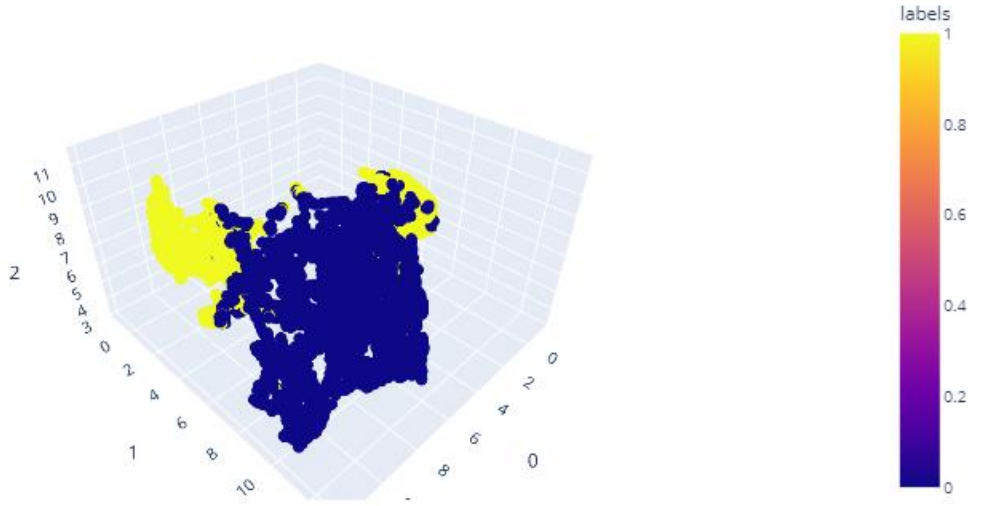

Figure 63 - 3D for K-Means with SOM of the Customer perspective

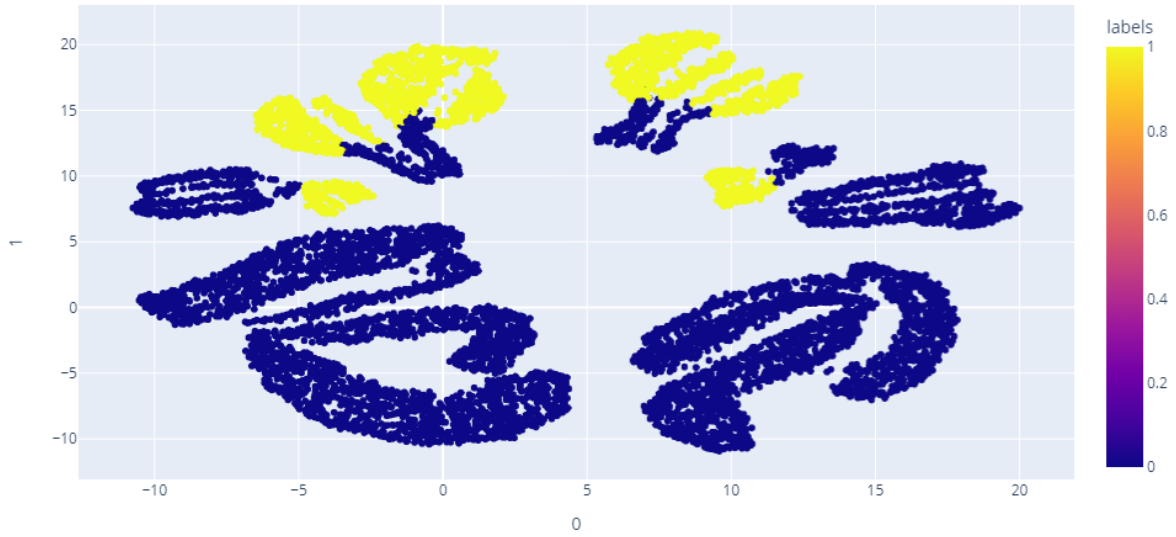

Figure 64- 2D for K-Means with SOM of the Customer perspective

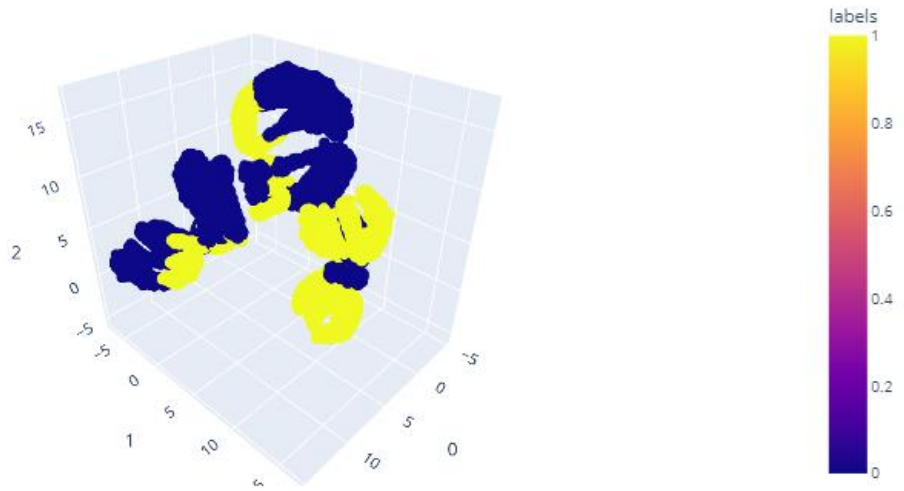

Figure 65 - 3D for K-Means with SOM of the Customer perspective

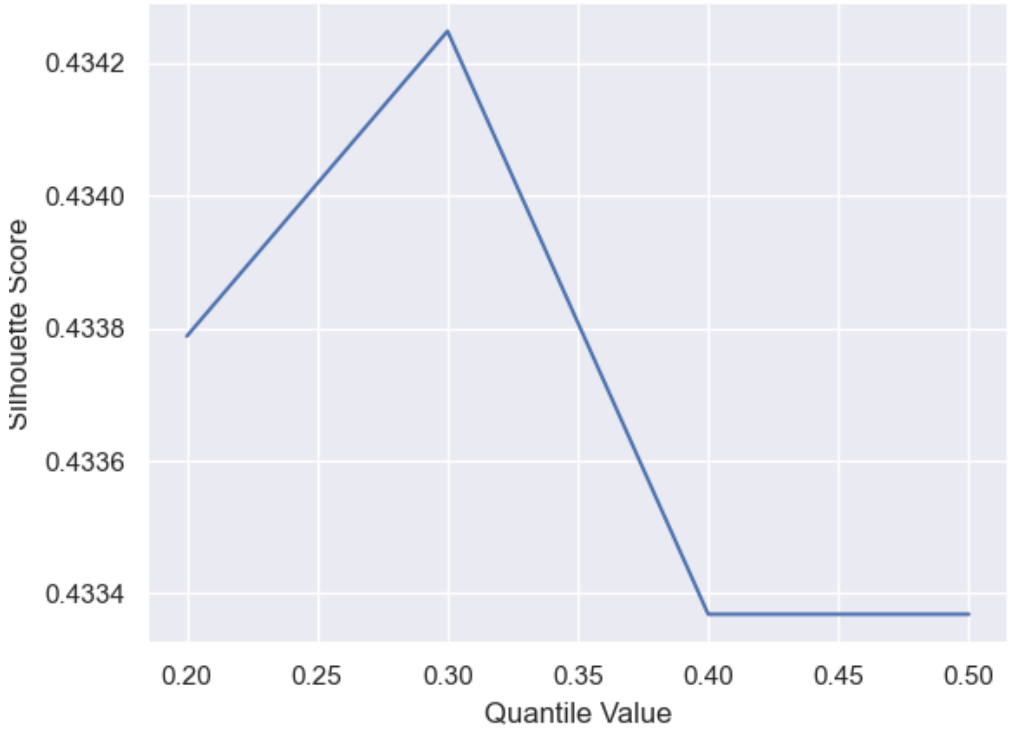

Figure 66 - Silhouette Score for the Quantile of the Mean-Shift Algorithm

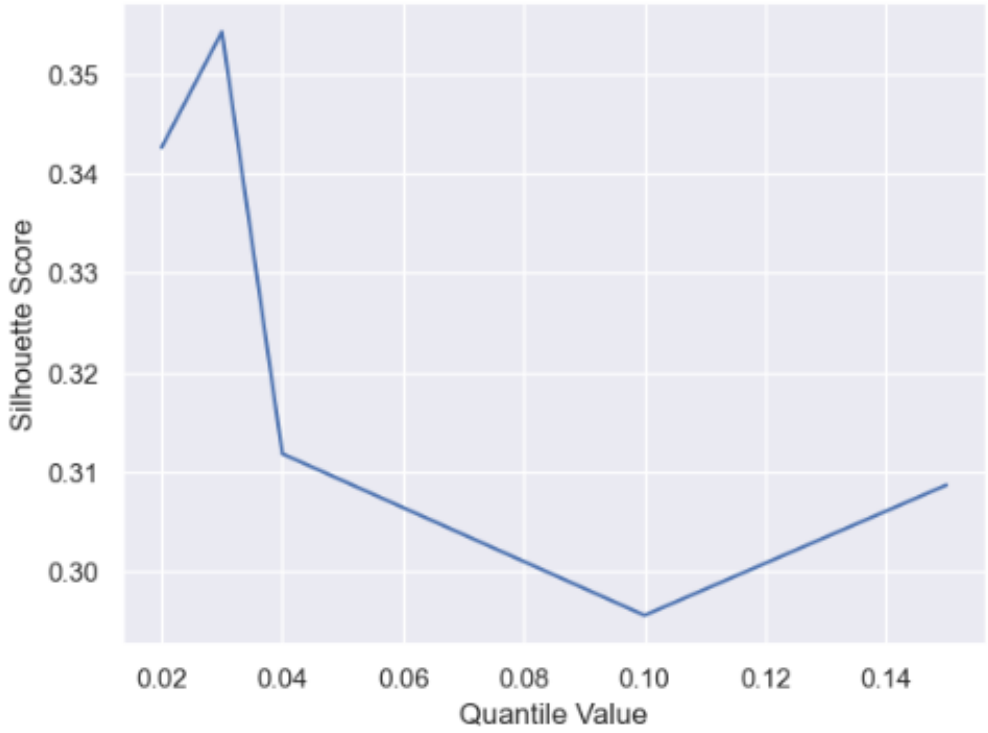

Figure 67 - Silhouette Score for the Quantile of the Mean-Shift Algorithm

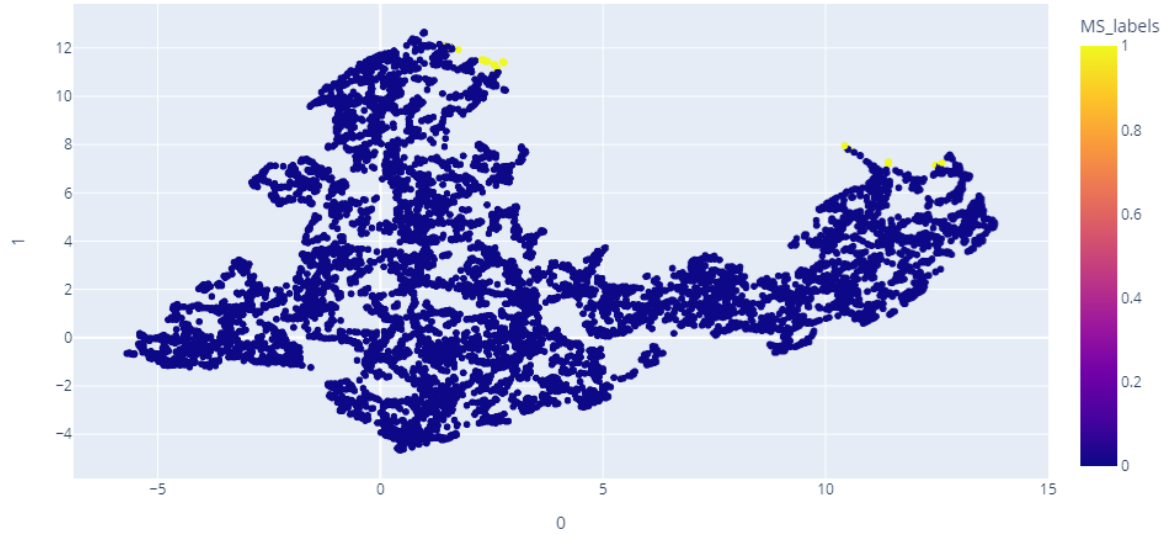

Figure 68 - 2D view of Mean Shift for Insurance Perspective

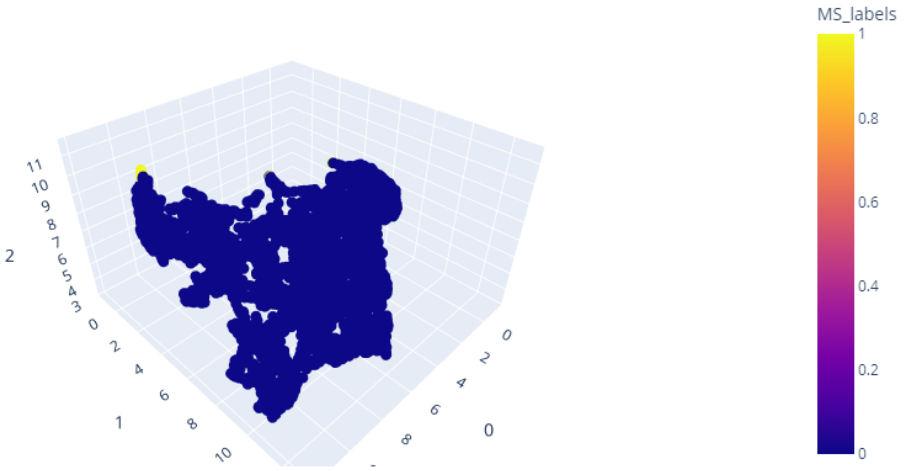

Figure 69 - 3D View of Mean Shift for Insurance Perspective

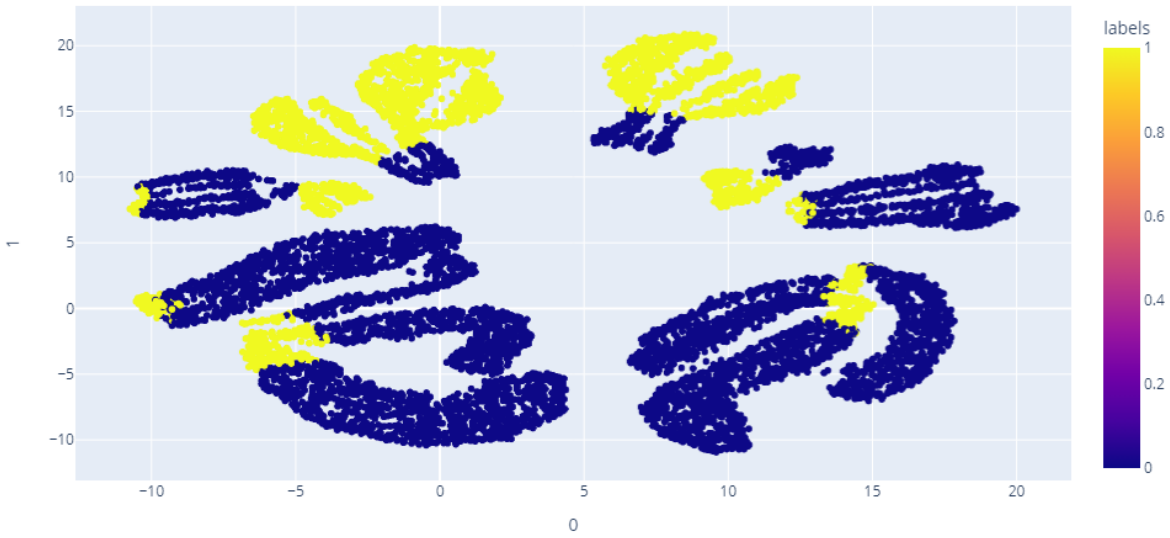

Figure 70 60 - 2D view of Mean Shift for Customer Perspective

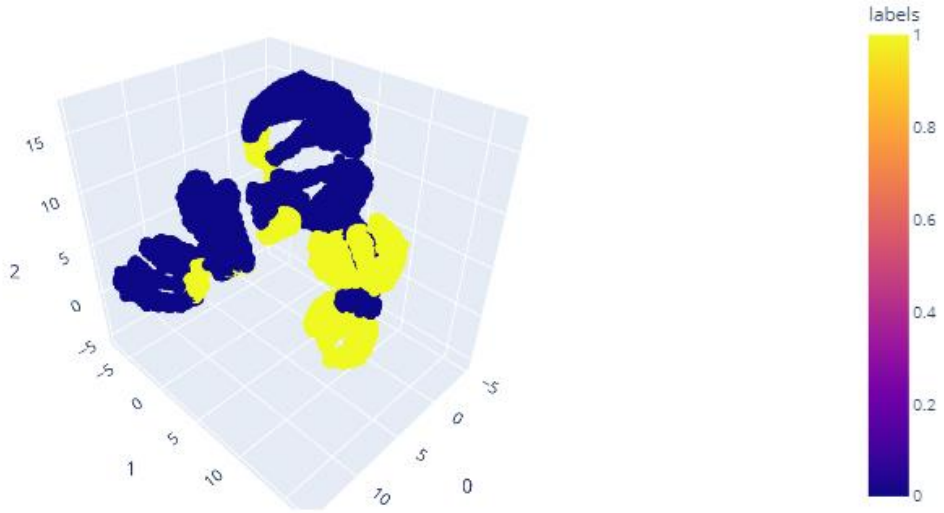

Figure 7161 - 3D View of Mean Shift for Custumer Perspective

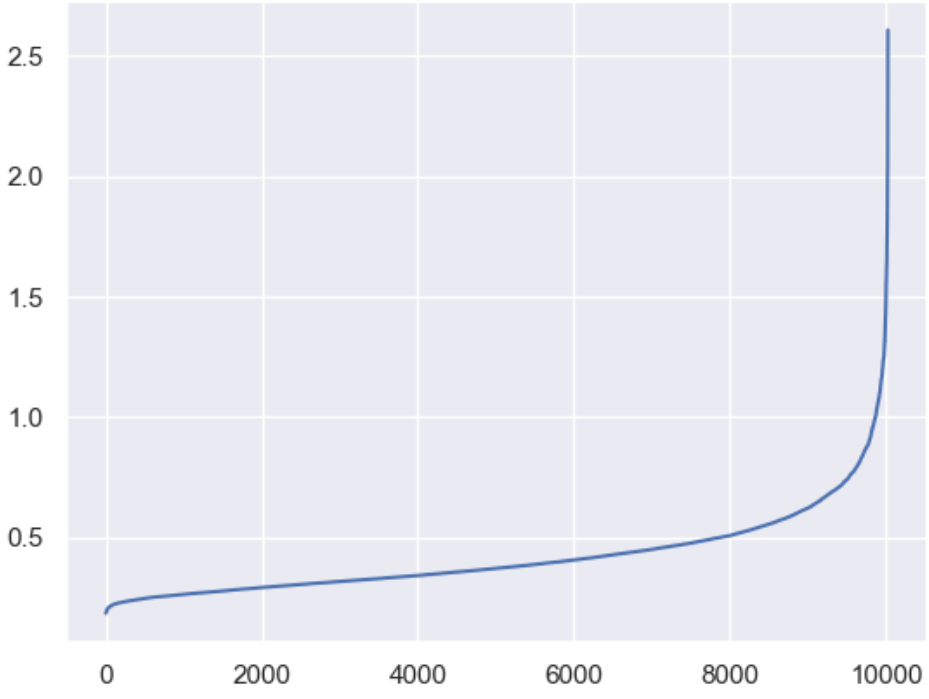

Figure 72 - K-distance graph for Insurance

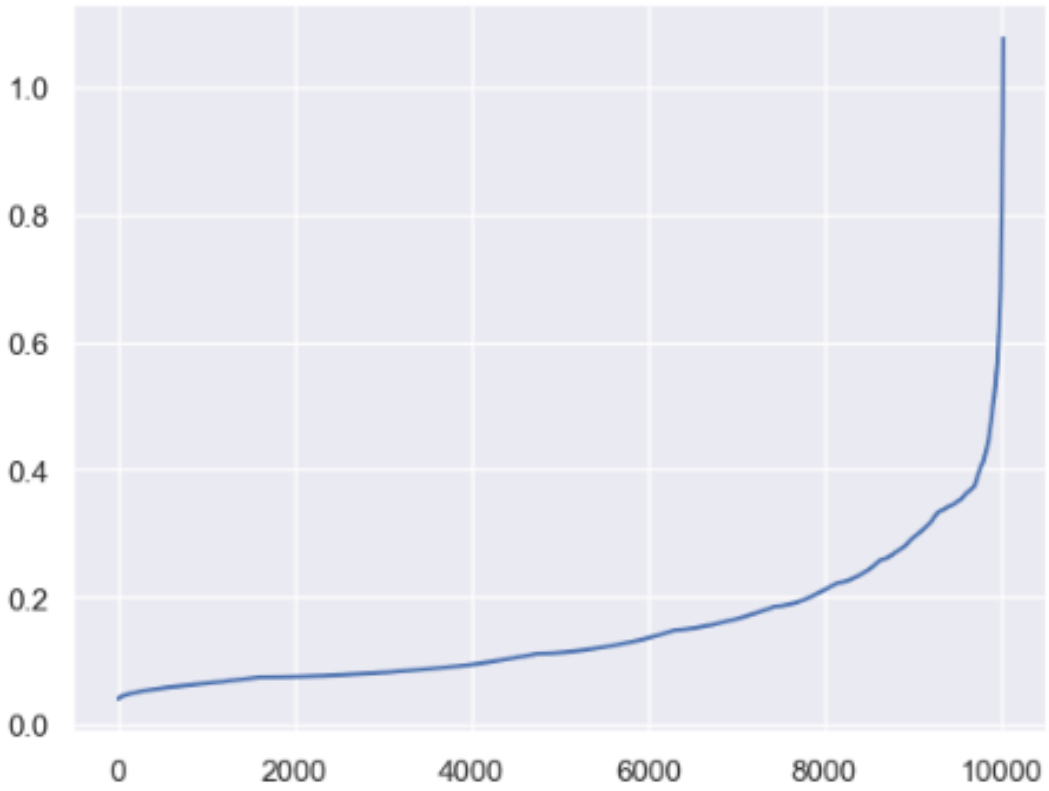

Figure 7362 - K-distance graph for Customer

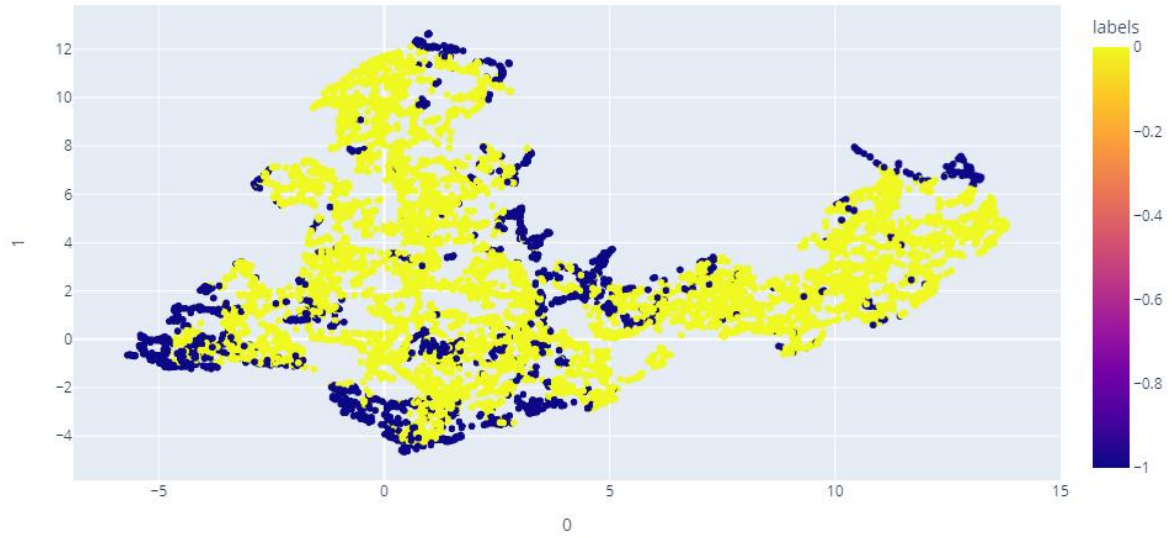

Figure 74 - 2D UMap for DBScan for Insurance

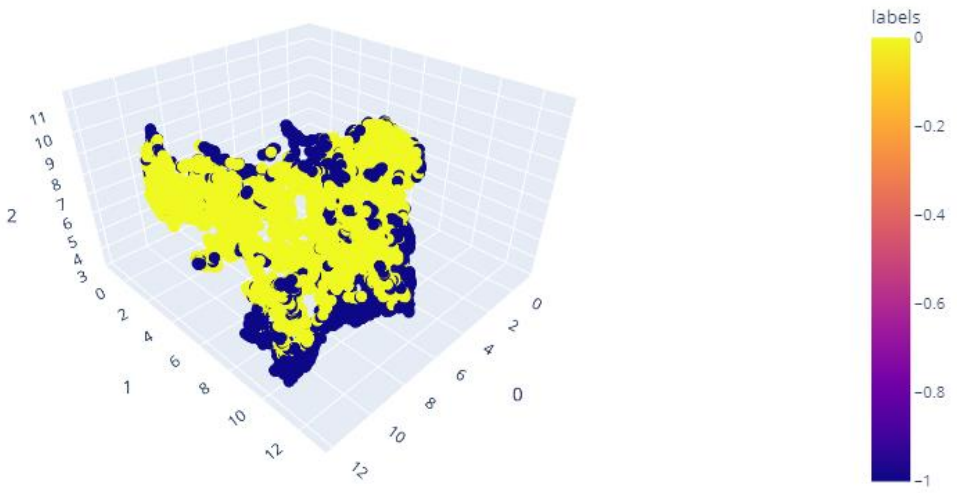

Figure 75 - 3D UMap for DBScan for Insurance

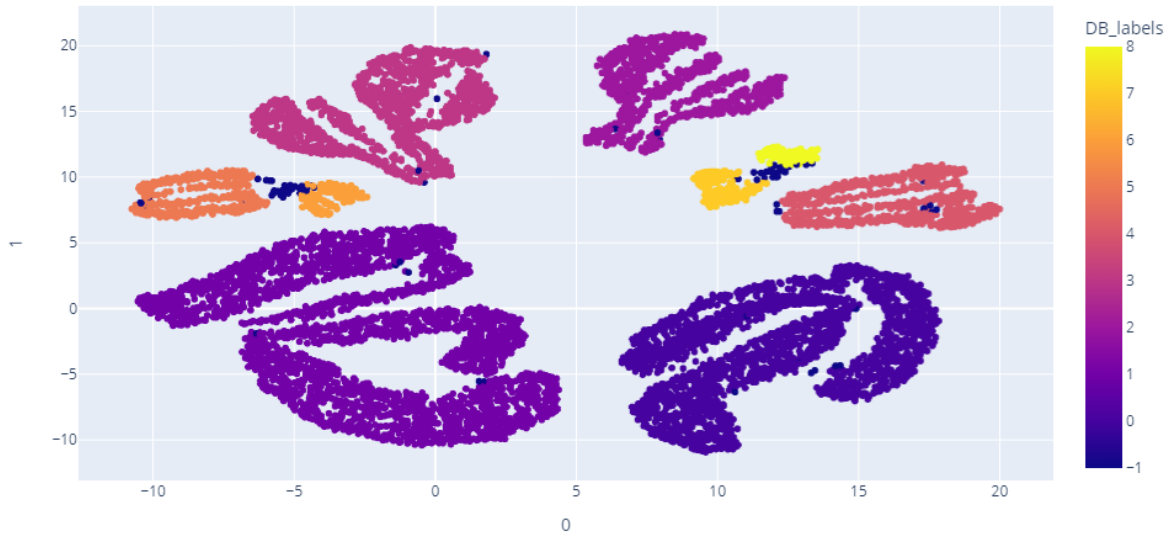

Figure 7663 - 2D UMap for DBScan for Customers

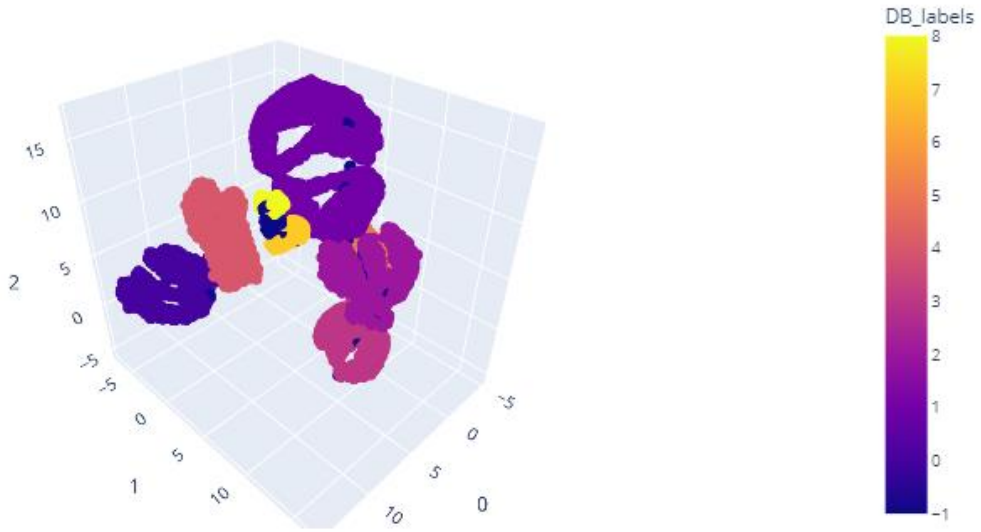

Figure 7764 - 3D UMap for DBScan for Customer

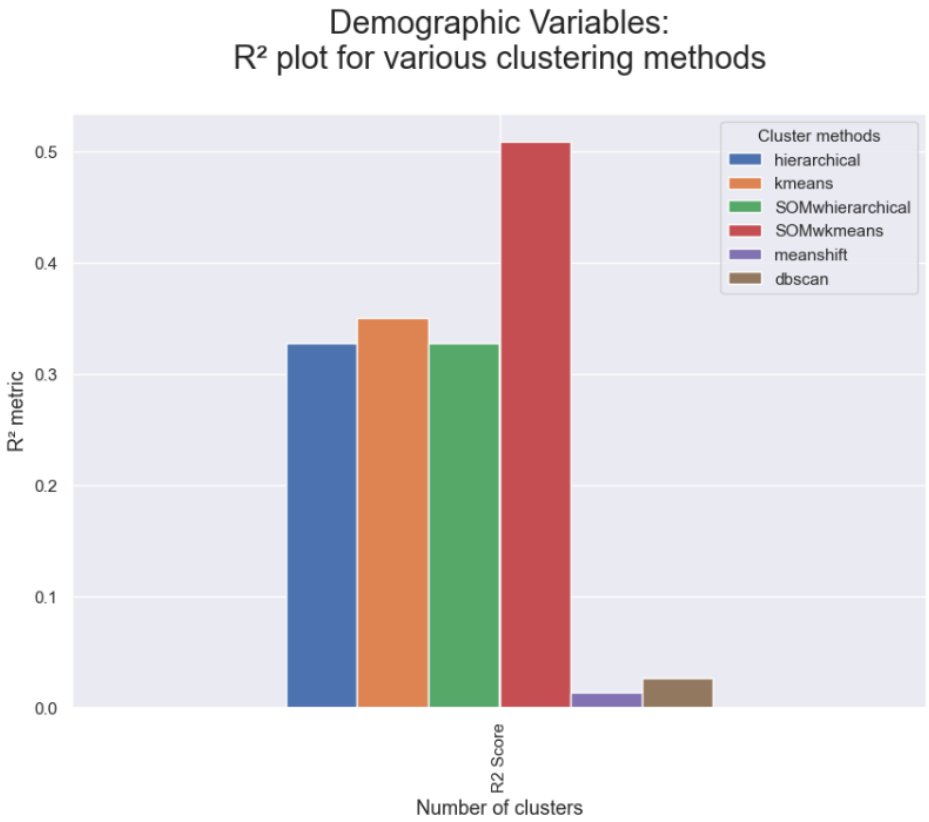

Figure 78 - R-squared score for Insurance

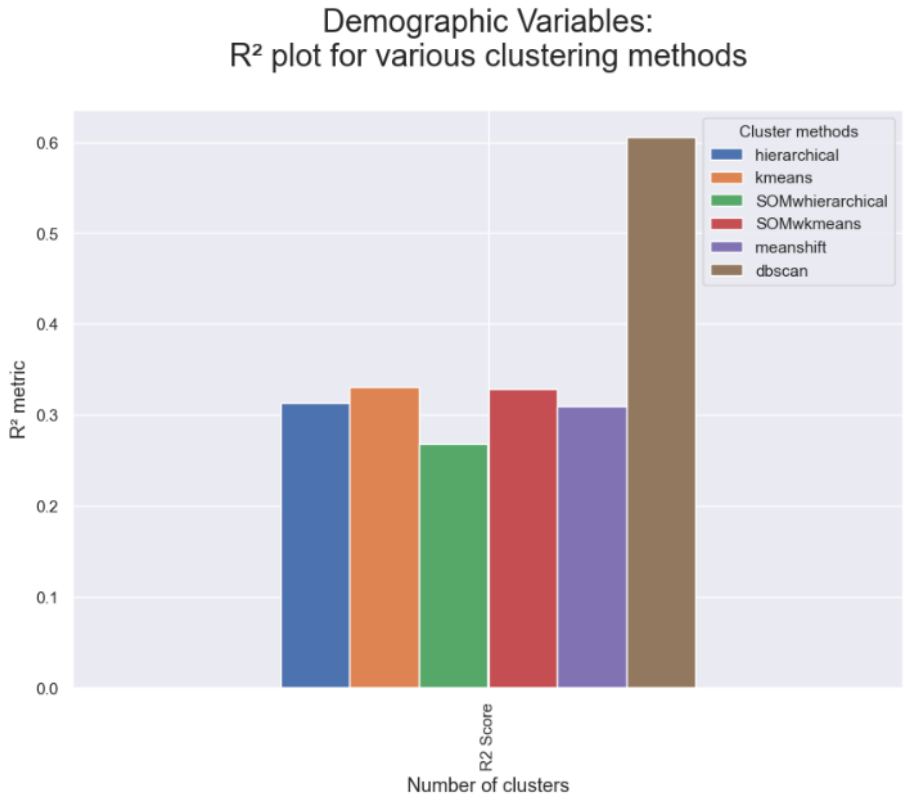

Figure 79 - R-squared score for Customers

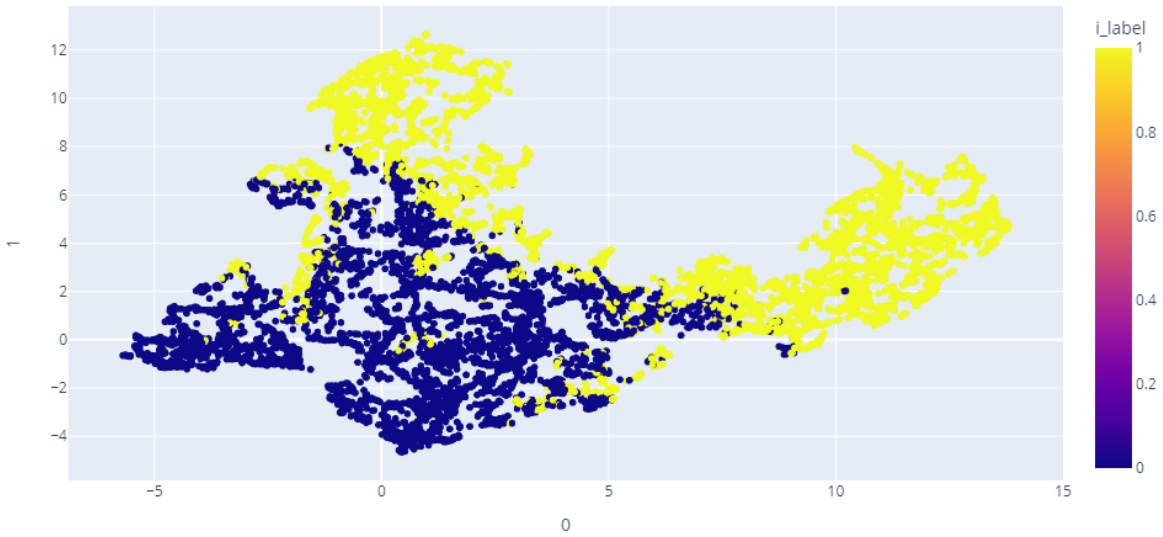

Figure 80 -2D Umap for Insurance

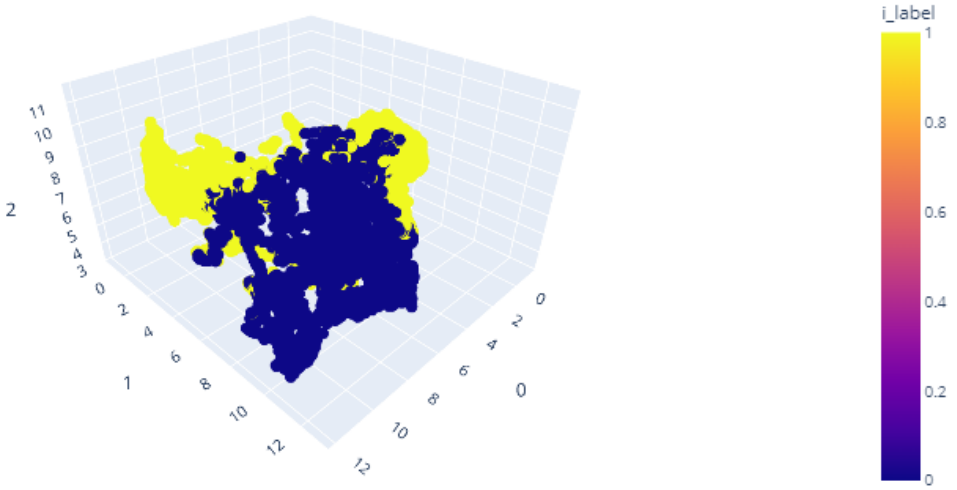

Figure 81 - 3D UMap for Insurance

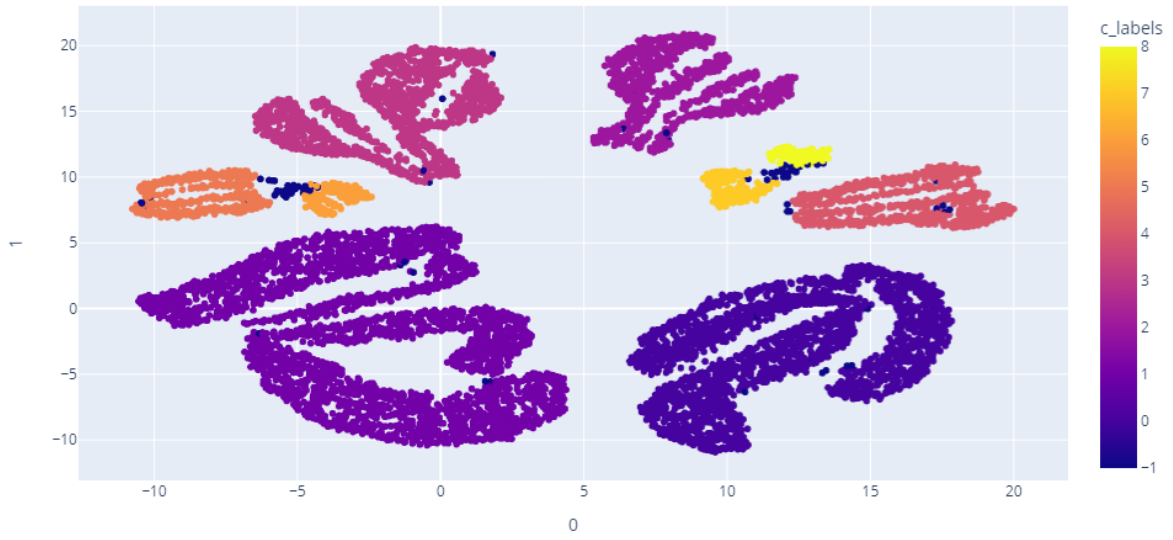

Figure 82 -2D Umap for Customers

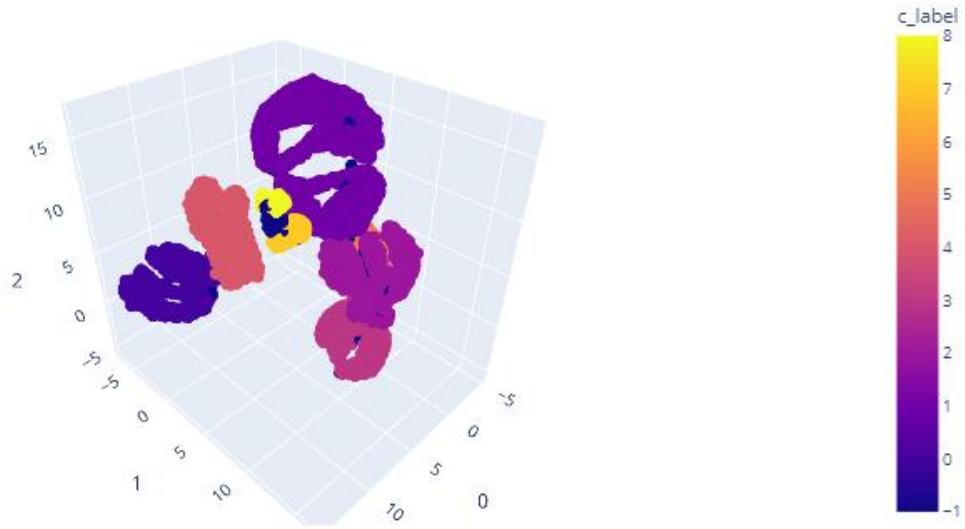

Figure 82 - 3D UMap for Customers

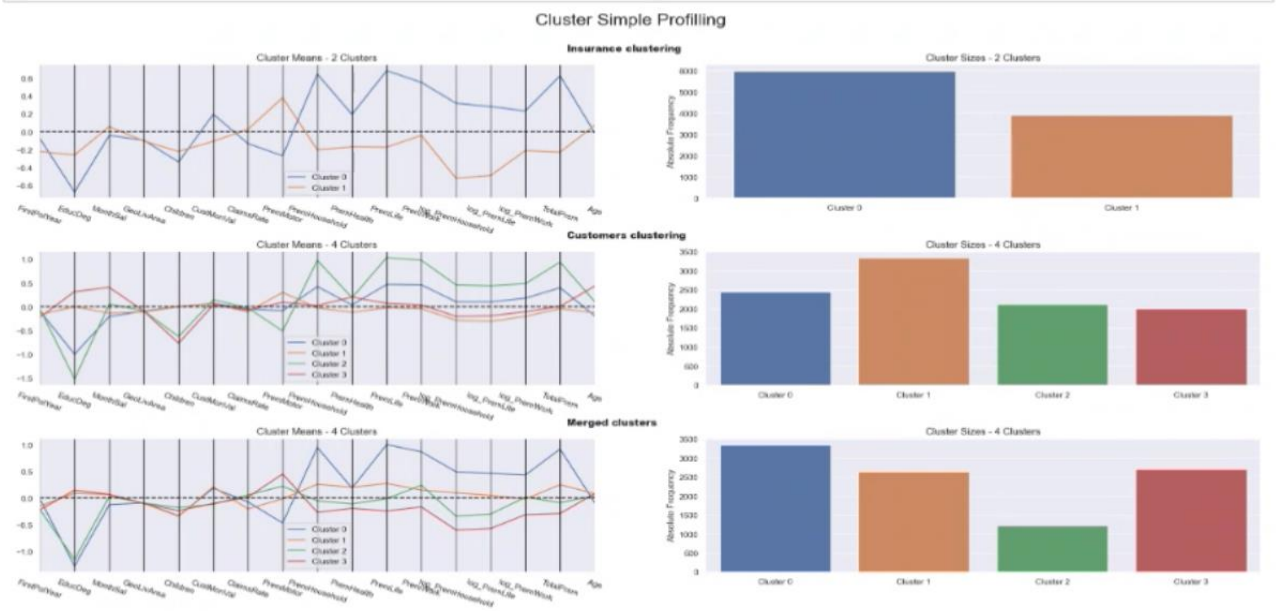

Figure 83 – Customers Simple Profiling

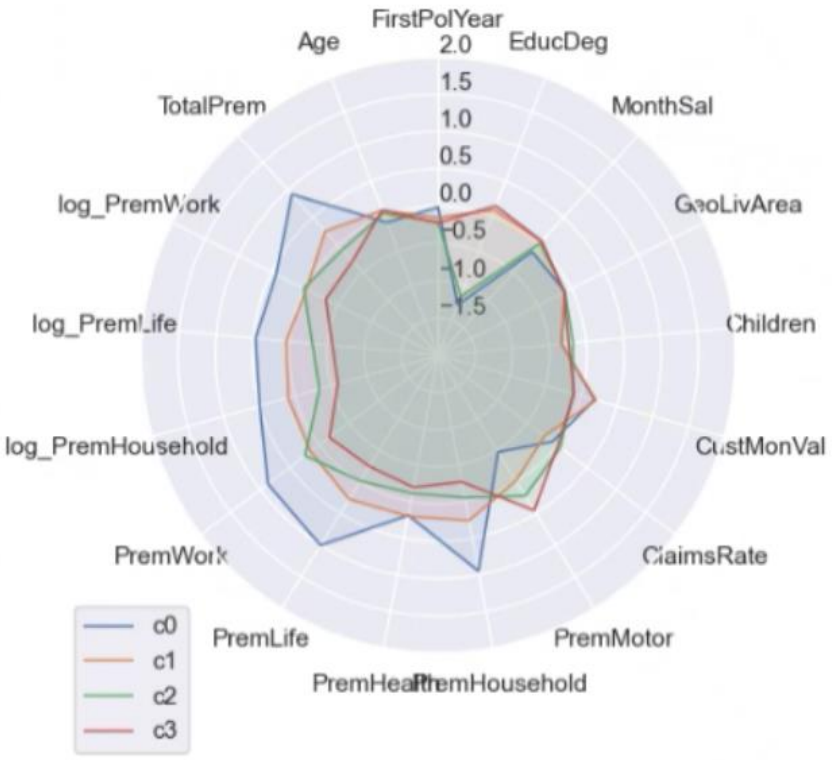

Figure 84 – Customers Profiling – Radar Plott