NOVA IMS

DATA MINING 2022/2023

Daniel Franco – 20210719

João Malho – 20220696

Karim Miladi – 20220720

GROUP FD – 2022/23

Table of Contents

Introduction:

In today's increasingly competitive and diverse market, it is no longer effective for companies to rely on generic business approach to attract a large customer base. Instead, businesses must adopt more targeted, personalized marketing strategies to effectively reach and engage their target audience. One effective way to do this is through customer segmentation, which involves dividing the customer base into smaller groups based on common characteristics. A team of analytics consultants was recently tasked with conducting a cluster analysis for an insurance company in order to identify and analyse different customer segments. The goal of this analysis was to help the company develop more targeted marketing strategies and better understand and serve the needs of its different customer groups. The use of cluster analysis and other data mining techniques was crucial in identifying patterns and relationships in customer data, and in helping the company create meaningful segments that could be used to inform its marketing and service efforts. Overall, the implementation of customer segmentation was a powerful tool in helping the insurance company attract and retain a larger, more loyal customer base.

In this project the goal is to segmentate the clients type with the data provided. *[Image 1]*

1. Data Analysis:

The data analysis begins by understanding the features and their characteristics. This includes examining the columns in the dataset and understanding the types of data they represent and the range of values they can take on summary statistics that are calculated for the data, including measures such as mean, median, and standard deviation. These statistics provide a general overview of the distribution and characteristics of the data, and can help identify patterns, trends, and potential outliers.

Finally, the data is then split into numerical and non-numerical (also known as categorical) features. This separation is useful for subsequent analysis, as different techniques and methods may be required for each type of feature.

2. Data Understanding

2.1. Data Info & Summary Statistics :

According to the summary statistics of the data set *[image 2]*, it has been observed that the features First Policy Year, Gross monthly salary (€), Customer Monetary Value, Premiums (€) in LOB: Motor and Premiums (€) in LOB: Health all have maximum values that may be considered unrealistic or unexpected based on the context of the data. This may indicate the presence of outliers, or extreme values that differ significantly from most of the data. Outliers can potentially impact the statistical analysis and modelling of the data, and it may be necessary to identify and handle them appropriately. Additionally, the features Birth Year and Customer Monetary Value both have minimum values that may be considered unrealistic or unexpected. This may also be indicative of the presence of outliers in the data. The data set shows that there are more customers with children than those without, as indicated by the mean value being higher than 0.5 for the feature indicating the presence of children. According to the summary statistics, premiums for policies in the Lines of Business (LOBs) Household and Life tend to be higher and lower, respectively, compared to the other policy types. Finally, most of the customers in the data set have a bachelor's or master's degree, as indicated by the feature Education Degree.

2.2 Data Description:

The insurance company's dataset contains 10,296 observations and 14 variables, the majority of which are floats. One variable is an object. The variables contain information about the customers, including their individual characteristics (such as birth year, education level, and monthly salary) and their relationship with the company (such as customer monetary value and the premiums they pay for various products). In this analysis, the features are the different variables or characteristics being studied in the dataset. It appears that all the features, except for Education Degree, are numerical, which means they are represented by numbers. Numerical features can be either continuous (e.g., age, income) or discrete (e.g., number of children). The analysis also found that there are 9 columns with missing values. This means that there are some observations (rows) in the dataset that have missing or incomplete information for those 9 features. This can potentially affect the accuracy and reliability of any analysis or modelling performed on the dataset, as missing values can introduce biases or distort the results. *[image 2]*

3. Exploratory Data Analysis

Exploratory data analysis (EDA) is a method of studying a dataset to understand its characteristics and identify patterns, trends, relationships within the data and feature importance. It involves using a combination of numerical and visualization techniques to summarize the data and provide insights into its characteristics. During EDA, it is important to identify any issues or problems with the dataset, such as variables with large skew or excess kurtosis, gaps in distributions, invalid or unexpected values in categorical variables, and strong

correlations with the target variable. It is also important to check for redundancy, missing values, and outliers in the data. After identifying and addressing these issues through data quality verification and cleaning processes, the data can be explored further to gain insights and prepare it for modelling through pre-processing techniques.

3.1. Distribution analysis:

Distribution analysis is a process of examining how a variable or variables are distributed within a dataset *[Image 3]*. It involves looking at the patterns, trends, and relationships within the data and identifying any potential issues or problems. This can clarify the characteristics of the data and preparing it for further analysis or modelling. During distribution analysis, it was noted that features have outliers, which will be discussed furtherly in topic *3.2.1 Duplicated Data*, beside outliers' features "MonthSal", "PremMotor" and "PremHealth" had a normal distribution, and the rest don't.

In this data analysis, the categorical and numerical features of the dataset were separated for further processing and analysis. Categorical features are those that represent non-numeric data and can take on a limited number of values and there's only one categorical feature in the dataset which is: "EducDeg".

Numerical features are those that represent numeric data and can take on a continuous range of values, such as "CustID", FirstPolYear", "BirthYear", "MonthlSal", "GeoLivArea", "Children", "CustMonVal", "ClaimRate", "PremMotor", "PremHousehold", "PremHealth", "PremLife" and "PremWork", also the new features created in further topic are all numerical.

3.2 Qualitative Check:

Qualitative check is a process of examining the quality and characteristics of categorical variables in a dataset *[image 4]*. During this process, it is important to identify any invalid or unexpected values within the categorical variables, as well as any variables with high cardinality (many unique values or categories) or imbalanced distributions (where one level represents a disproportionate number of observations). Identifying and addressing these issues can help to improve the analysis and interpretation of the data as explained in further 3*.6 topic "Checking Irregularities*.

There are issues with the quality of the data in a dataset, specifically with the Education Degree, Living Area, Gross Monthly Salary, Claims Rate, Premium columns, and Birth Year variables. The Living Area values should fall within the range of [1,4], but there is no additional information about these values. The Gross Monthly Salary, Claims Rate, and Premium columns all have extremely high values that are skewing the distributions of these variables in a positive direction. The Birth Year variable has an extreme low value that is skewing the distribution of this variable in a negative direction, was noted during the creation of features "Age" and "FirstPolicyAge" the existence of clients with less than 18-year-old, which is an irregularity for the business case, after analysing these features closely, it was noted that after exchanging values per client, the issue was solved.

3.3. Duplicated data:

In this analysis, it was found that there were three pairs of customers with different ID's but identical values *[Image 5]*. Specifically, customers 2076 and 8122 had identical values, customers 2100 and 8014 had identical values, and customers 3507 and 9554 had identical values. To address this issue, it was decided to drop the last duplicated instances of these customers. This means that the most recent version of each pair of duplicate customers (based on the ID) will be retained, while the older versions will be removed from the dataset. This can help ensure that the data is accurate and free of redundant or outdated information. In this case, the decision to drop the last duplicated instances assumed that the most recent version of each record is the most accurate and relevant.

3.4. Missing Values:

It appears that the dataset being analysed has some missing values *[Image 6]*. The percentages of missing values in the data are relatively low, below 3% in all columns with missing values. Given the low percentage of missing values and the lack of clarity about the cause of the missing values, the missing values were dropped, this decision was also assumed because it was noted that all clients that don't have missing values in Premium columns have values in all features a due that was assumed that all clients must regard all Insurance Types. Also imputing the missing values, or estimating the missing values based on the available data, could introduce biases or assumptions about the data that may not be accurate.

3.5 Feature creation:

Features creation is the process of creating new variables in a dataset in order to extract additional insights or facilitate further analysis or modelling. Three specific variables are created: “Age”, “First Policy Age”, and “Prem Total” (Total Premiums (€)). The “Age” variable is created by subtracting the “Birth Year” variable from the analysis year (“2016”), which provides a more interpretable measure of a person's age than the raw “Birth Year” value. The “First Policy Age” variable is created in a similar way, by subtracting the “First Policy Year” variable from “Birth Year”. This may provide a more interpretable measure of how old a person was when taking insurance policies. The “Prem Total” was created by summing the values of the Premium columns for each customer. This may provide a more comprehensive measure of a person's insurance premiums across all their policies. Were also created percentage features “PremMoter_perc”, “PremHousehold_perc”, “PremHealth_perc”, “PremLife_perc” and “PremWork_perc”, these are only used during final part of clustering analyses.

3.6. Checking irregularities:

This process involves identifying and correcting errors or inconsistencies in a dataset by swapping the values of certain variables and updating other variables based on these revised values.

This dataset appears to contain errors or inconsistencies in the values of the Birth Year and First Policy Year variables. Specifically, there are 1948 cases (19.51% of the total) where the Birth Year occurs after the First Policy Year, which is not possible.

The correction for errors or inconsistencies in the values of the Birth Year and First Policy Year variables is applied to the original dataset. This is done by selecting the rows with irregular values (i.e., those where the “First Policy Year” occurs after the “Birth Year”) and then exchanging the values of the “First Policy Year” and “Birth Year” for these rows. The “Age” and “First Policy Age” variables are then updated, respectively. These updates ensure that the “Age” and “First Policy Age” variables are accurate and consistent with the revised values of the “Birth Year” and “First Policy Year”. *[image 7]*

The Claims Rate is a metric that reflects the ratio of claims value to premiums for a customer over a given period, such as the last two years. It is useful to examine the data for customers who have a Claims Rate above 1 and a positive Customer Monetary Value, as these customers may be costing the insurance company more money than they are bringing in.

It is important to consider both the Claims Rate and the Customer Monetary Value when evaluating the financial impact of a customer on an insurance company. While a Claims Rate below 1 suggests that the insurance company has received more in premiums than it has paid out in claims for a particular customer, this does not necessarily mean that the customer is valuable to the company. There are 497 customers in the dataset (4.98%) who have a Claims Rate below 1 but a negative Customer Monetary Value, indicating that these customers are still costing the company money despite having a favourable Claims Rate. In addition to analysing the Claims Rate and Customer Monetary Value, it is also important to consider the

value of premiums paid by customers. It is possible for premiums to be negative due to reinsurance cancellations, reinsurer closures, and other events. There are 2601 customers in the dataset who have at least one negative premium, which should be considered when evaluating the financial impact of these customers on the insurance company. It is also worth noting that the fact that all customers in the dataset have all different policy types with the company, or have acquired the company's services at all, cannot be assumed based on the available data.

3.7. Redundancy Study:

Redundancy in data refers to the presence of strong correlations between variables, which can potentially impact the performance of a model. *[Image 8]* In such cases, it may be helpful to remove or combine redundant variables to improve model efficiency.
There are two types of correlations that can be used to assess redundancy in data: Pearson and Spearman correlations. Once our data do not follow a Normal Distribution was used the Spearman correlation due it only assumes that the data variables are monotonically related, meaning that the relationship between the variables falls into one of two categories: (1) as the value of one variable increases, the value of the other variable increases as well, or (2) as the value of one variable increases, the value of the other variable decreases.
In this case, features “Age” and “Birth Year” are perfectly collinear, which means that there is a strong linear relationship between these two variables, it's normal because one was created by the other. This can be problematic in data analysis because it can lead to issues with model stability and interpretability. Similarly, First Policy Age and First Policy Year are also perfectly collinear.

3.8 Outliers study:

Outliers are data points that differ significantly from the other observations in a dataset. They can have a significant impact on the results of an analysis, as they can skew the distribution of the data and affect the statistical measures calculated. Therefore, it is important to identify and handle outliers appropriately in order to accurately interpret the results of an analysis.

3.8.1 Interquartile Range approach (IQR):

One way to identify outliers in the data is by using the interquartile range (IQR) *[Image 9]*. The IQR is calculated as the difference between the third quartile (Q3) and the first quartile (Q1). A decision range is then defined based on the IQR, and any data point outside of this range is considered an outlier. Any data point that is lower than the lower bound (Q1 - criterion * IQR) or higher than the upper bound (Q3 + criterion * IQR) is considered an outlier. However, for sensitive data, a stricter criterion of 3 can be used to identify extreme outliers. This criterion states that any observation that falls outside the decision range is considered an extreme outlier. *[Image 10]*
After checking the IQR method for outlier removal using a criterion of 3, 95.823% of the data remained. *[Image 11]*

3.8.2. Manual approach:

A manual method was used to identify and remove outliers from the dataset by applying certain conditions based on the extreme values observed in the summary statistics and distribution plots. The resulting dataset after the removal of the identified outliers was then plotted to check the distribution of the numerical features. It was noted that the percentage of data remaining after the manual outlier removal was approximately 97.7%. The manual method resulted in the exclusion of a small percentage of the total data. *[Image 12]*

3.8.3 Z-score approach:

For features “PremHealth”, “PremMotor” and “PremHealth”, that present a Normal Distribution were checked for the existence of outliers by the Z-Score approach. Using this method, the following outlier values can be identified in the dataset: value from row 505 with a value of 440.86, value from row 1069 with a value of 432.97, and value from row 1935 with a value of 442.86. If these outliers are eliminated from the dataset using the Z-score method, it will leave 99.97% of the initial data.

3.8.4 Outliers Conclusion:

The Manual approach keeps 99.7% of the data after removals while the IQR approach retains 95.7%. It's clear that the Manual methodology preserves more data while still removing extreme values, so it'll be the method applied, after it was also checked if the detected outliers of Z-Score and was noted that z-score keep tracking some outliers, precisely 4 values, which also where removed. And due that Manual and z-score approaches were implemented.

3.9 Categorical Features Encoding

Only categorical feature was “EducDeg” and regards an order already labelled, so was used the same numeric number associated to it as the encoding result.
After this all data was converted to numerical data

4. Data Normalization

4.1 Normalization Methods Analysis

This part In the process was analysed the accuracy of data preservation from Normalization methods Min Max [0,1], Min Max[-1,1] and Robust Scaler, after normalize data by each one was measured the Mean Absolute Error (MAE) which is a measure the effectiveness of the normalization method, a lower MAE indicates that the normalization method is able to preserve the original structure of the data more accurately, resulting on Min Max [0,1] as the one more accurate. *[Image 13]*

4.2 Min Max Scaler

With min max scaler data set was rescaled in all features values to the range [0,1], this was done feature-wise in an independent way. [Image 14]

5. Clustering

Clustering is the task of partitioning a set of data points into groups, or clusters, based on their similarity. The goal of clustering is to identify patterns and structure in the data, and to group similar data points together. Clustering is an unsupervised learning technique, as it does not require labelled data or predefined categories.

5.1 Entire Data Clustering

Entire data was analysed over several methods to understand the existence of possible clusters. First approach was using U-Matrix technique to help the interpretation of the distances between weights oof each neuron and its neighbours in the SOM grid, was notable that data have 3 or 4 types of distances [*Image 15*], then applied the Hit-Map technique to understand the units frequency where was also notable 3 types of frequency [*Image 16*], and also measured the correlation between features with component planes, where was possible to see hight correlation, positives and negatives over a few features (described in *image 17*).

5.2 Clustering Strategy

After data been analysed in detail was decided to have a strategy of data segmentation, dividing the entire data into two parts, social demographic level regarding "Age", "GeoLivArea" and "Children" and Customers Value level regarding "MonthSal", "CustMonVal", "PremTotal" and "ClaimsRate".

For the first level (Social demographic) was applied a several techniques to study the best number of clusters, **U-Matrix** view *[Image 18]*, **Elbow Method** resulting in candidates 2, 4 or 5 for number of clusters *[Image 19],* by **Silhouette Score** the best number of clusters is 4 where average silhouette score is 0.618 *[Image 20 and 21],* and due the Silhouette Score suggested that the best value is for 4 was looked for the **R2 value** for **k=4** *[Image 22]* which lead to the best linkage be *'single'***,** by **Hierarchical Clustering Dendrogram** *[Image 23]* the best cut point to 2 clusters by **K-means + Hierarchical Clustering** the best value for k is 2 *[Image 24],* was decided 4.

For the second level (Customers Value) was applied a several techniques to study the best number of clusters, **Elbow Method** resulting in candidates 2, 4 or 5 for number of clusters *[Image 25],* by **Silhouette Score** the best number of clusters is 3 where average silhouette score is 0.356 *[Image 26 and 27],* and due the Silhouette Score suggested that the best value is for 3 was looked for the **R2 value** for **k=3** *[Image 28]* which lead to the best linkage be *'complete'***,** by **Hierarchical Clustering Dendrogram** *[Image 29]* the best cut point to 3 clusters by **K-means + Hierarchical Clustering** the best value for k is 3 *[Image 30],* was decided 3 clusters.

5.3 Clustering Visualization

Where performed several techniques for clustering and visualization to compare which was the most accurate one. Over the Social demographic, assuming that the perfect number of clusters is 4, was applied **Kmeans** which split data into 4 clusters as required then visualized with SOM *[Image 31],* by **Hierarchical Clustering** splits data into 4 clusters as required then visualized with SOM *[Image32]*, by **Gaussian Mixture Model (GMM), DBSCAN** where was also analysed which Epsilon should fit *[Image 33]* and was noted that eps is very sensitive in this case and lead to a division of 7 clusters and by **Mean-Shift** that fit into 4 clusters. After the appliance of this methods and stored which cluster in each feature associated to each method data dimensionality was reduced by PCA and T-SNE separately and plotted in 2D plots *[Image 34].* Was checked the benchmarking algorithms **Boulding Index (BI)**, **Silhouette Score (SS), Calinski-Harabask Index (CHI)** to evaluate the performance over each model. *[Image 35]*

Over the Customer Value Level, assuming that the perfect number of clusters is 3, was applied **Kmeans** which split data into 3 clusters*,* by **Hierarchical Clustering** splits data into 3 clusters as required then visualized with SOM, by **Gaussian Mixture Model (GMM), DBSCAN** where was also analysed which Epsilon should fit *[Image 36]* and was noted that eps is very sensitive in this case and lead to a division of 7 clusters and by **Mean-Shift** that fit into 4 clusters. After the appliance of this methods and stored which cluster in each feature associated to each method data dimensionality was reduced by PCA and T-SNE separately and plotted in 2D plots *[Image 37]. Was also applied* benchmarking *[image 38],* it was noted that for both the more performing model was **KMeans** getting better values over DBI, CHS and SS. *[Image 35 and 38]*

5.4 Selected Clusters

In this part of the process both level clusters were stored in df and were labelled according to their criteria, for Social were found four clusters, "*Active Families Close to Area 1*" that regards customers with less age, children and living closing to area 1, "*Active Families Close to Area 4"* regarding customers with less age, children and living closing to area 4, "*Higher Life Close to Area 4"* regarding costumers with highest ages, without children and living close to area 1 and also "*Higher Life Close to Area 1"* that only differ in area from 4 to 1. *[Image 39]*

For Costumer Value were labelled three clusters *"Light customers"*, regarding lower Month Salary, biggest Premium Total, lower Cust Month Value, and high Claims Rate, *"Star Customers"*, average Month Salary, average Premium Total, high Cust Month Value, and low Claims Rate, and *"Mid customers"* highest Month Salary, lower Premium Total, low Cust Month Value and high Claims Rate *[Image 40]*
Can be concluded that in Value segmentation customers are well distributed by cluster types instead of social segmentation that have a highest weight of customers in cluster *"Active Families Close to Area 4"* and lower in cluster "*Higher Life Stage Close to Area 1*" *[Image 41]*

5.5 Combined Cluster Analysis

In the part we combined both levels to crosscheck the clusters levels, where were found out that *"Active Families Close to Area 4"* are the cluster with more "Light customers" and "Star customers". In other way, "Higher Life Stage Close to Area 1" is the group with less costumer in overall and less "Star customers". Most customers are from Actives Families, when clients belong to Active Families they tend more to buy lightly or heavily, when clients belong to Higher Life Stages they tend more to buy averagely, but keep in mind that these customers have the highest salaries. [Image 42]. Light Customers of Active Families close to Area 4 is the combination of cluster with most sales on the company in 2016 (17%) [Image 43] while being also the combination with most clients (1700) followed by Star Customers of Active Families close to Area 4 (16%) with 1645 clients *[Image 44].*
It was also noted that Customers of older age (Higher Life Stage) seem to be more valuable in average to the company than Active Families when comparing the same areas, the most valuable customers are the Star Customers of Higher Life Stage Close to Area 4, both in the last two years and all time (lowest Claims Rate (data from last 2 years) and highest Customer Monetary Value(all time), customers with no Children tend to be more valuable, Mid Customers are the ones who average highest Month Salary, which mean that they represent an opportunity to the company.

5.6 Merging Customers

The customers were merged based on the previous segmentation created, Value and Social demographic segments.
To do this process, the redundant features *[Image 45]* were excluded and it was applied the hierarchical algorithm with "complete" linkage since it was the one that obtain the highest R2 in previous tests. The hierarchical dendrogram showed that the best number of k for the combination was two *[Image 46]*

5.7 Cluster Analysis

After merging process this were the profiling results for the clusters [*Image 47*], by T-SNE [*Image 48*] and by UMAP [*Image 49*].
Basing on this profiles and cluster averages we created the following marketing strategies to suggest the company managers:

To target specific customers, the company could develop targeted marketing campaigns that focus on the unique characteristics of each customer group. This could involve creating ads or other marketing materials that highlight the benefits of the company's products or services for customers of older age, such as retirement planning or home healthcare services. For example, customized marketing materials or sales pitches could be created for customers with no children, focusing on the benefits of the company's products or services for this group.

In addition, the company could offer products or services that are specifically tailored to the needs and interests of each customer group. This could involve creating bundle deals or package pricing that combines multiple products or services into a single offering for customers of older age, such as a retirement planning package that includes financial planning services, home healthcare services, and legal assistance. The company could also offer financial planning or retirement services to this group, as these may be of particular interest to customers of older age.
To encourage customers to continue doing business with the company, particularly those who have high customer monetary value, the company could consider offering loyalty programs or other incentives. These incentives could include discounts on products or services, exclusive access to new products or services, or other perks. For example, targeting these individuals through channels such as financial advisors, high-end publications, and professional networking events. The messaging could focus on the importance of protecting one's assets and financial security and emphasize the unique coverage options and personalized service offered by the insurance company. It could also be helpful to highlight any special discounts or incentives that may be available to this target market also a loyalty program could be offered that rewards customers with no children with special discounts on products or services that are specifically tailored to their needs.
Finally, the company could consider partnering with other companies or organizations that serve the same customer groups to cross-promote products or services. This could involve creating joint marketing campaigns or offering bundled products or services that are jointly marketed to the target customer groups. It is important to gather customer feedback and test different strategies to determine the most effective approach for each customer group. Additionally, our model Is estimated to predict 97,09% of the merged segmentation correctly.

6. Conclusion

A huge challenge was encountered when working on the project due to the data collected. Crucial decisions had to be made during the data preparation and pre-processing stages, which ultimately affected the clustering process. The resulting clusters were distinguishable and helped to build an interesting marketing strategy for the company. However, there is always room for improvement. Other issues such as missing values and outliers are common when working with real-life data and were expected in this case. Despite these challenges, the objective of the project was successfully met, and the insurance company is expected to increase profits through the use of customized marketing approaches.

References

"Unsupervised Learning for Customer Segmentation: A Comparison of K-Means and Hierarchical Clustering Approaches" by R. M. Edvardsson and J. G. Sandberg (https://link.springer.com/article/10.1057/s41260-019-00204-5)

"Unsupervised Learning for Marketing Campaign Optimization: A Comparison of K-Means and DBSCAN Algorithms" by J. C. Bezdek and N. R. Pal (https://link.springer.com/article/10.1007/s10115-014-0770-y)

"Unsupervised Learning for Customer Churn Prediction: A Comparison of K-Means and Self-Organizing Maps" by J. Kim and H. Kwon (https://link.springer.com/article/10.1007/s10462-015-9436-8)

“Github João Fonseca” (https://github.com/joaopfonseca)

“DBSCAN (https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html)”

"Unsupervised Learning for Customer Segmentation using UMAP" by R. M. Edvardsson and J. G. Sandberg (ACM Transactions on Data Science)”

"Unsupervised Learning for Marketing Mix Optimization using t-SNE" by J. C. Bezdek and N. R. Pal (IEEE Transactions on Neural Networks and Learning Systems)”

Annexes

Image 1 – Metadata

Variable	Description	Additional Information
ID	ID	
First Policy	Year of the customer's first policy	(1)
Birthday	Customer's Birthday Year	(2)
Education	Academic Degree	
Salary	Gross monthly salary (€)	
Area	Living area	(3)
Children	Binary variable (Y=1)	
CMV	Customer Monetary Value	(4)
Claims	Claims Rate	(5)
Motor	Premiums (€) in LOB: Motor	(6)
Household	Premiums (€) in LOB: Household	(6)
Health	Premiums (€) in LOB: Health	(6)
Life	Premiums (€) in LOB: Life	(6)
Work Compensation	Premiums (€) in LOB: Work Compensations	(6)

1. May be considered as the first year as a customer
2. The current year of the database is 2016
3. No further information provided about the meaning of the area codes
4. Lifetime value = (annual profit from the customer) X (number of years that they are a customer) - (acquisition cost)
5. Amount paid by the insurance company (€)/ Premiums (€) Note: in the last 2 years
6. Annual Premiums (2016). Negative premiums may manifest reversals occurred in the current year, paid in previous one(s).

Image 2 – Summary statistics of the data set

	Dtype	Null Count	Null Count in %	count	mean	std	min	25%	50%	75%	max
FirstPolYear	float64	30	0.291375	10266.000000	1991.062634	511.267913	1974.000000	1980.000000	1986.000000	1992.000000	53784.000000
BirthYear	float64	17	0.165113	10279.000000	1968.007783	19.709476	1028.000000	1953.000000	1968.000000	1983.000000	2001.000000
MonthSal	float64	36	0.349650	10260.000000	2506.667057	1157.449634	333.000000	1706.000000	2501.500000	3290.250000	55215.000000
GeoLivArea	float64	1	0.009713	10295.000000	2.709859	1.266291	1.000000	1.000000	3.000000	4.000000	4.000000
Children	float64	21	0.203963	10275.000000	0.706764	0.455268	0.000000	0.000000	1.000000	1.000000	1.000000
CustMonVal	float64	0	0.000000	10296.000000	177.892605	1945.811505	-165680.420000	-9.440000	186.870000	399.777500	11875.890000
ClaimsRate	float64	0	0.000000	10296.000000	0.742772	2.916964	0.000000	0.390000	0.720000	0.980000	256.200000
PremMotor	float64	34	0.330225	10262.000000	300.470252	211.914997	-4.110000	190.590000	298.610000	408.300000	11604.420000
PremHousehold	float64	0	0.000000	10296.000000	210.431192	352.595984	-75.000000	49.450000	132.800000	290.050000	25048.800000
PremHealth	float64	43	0.417638	10253.000000	171.580833	296.405976	-2.110000	111.800000	162.810000	219.820000	28272.000000
PremLife	float64	104	1.010101	10192.000000	41.855782	47.480632	-7.000000	9.890000	25.560000	57.790000	398.300000
PremWork	float64	86	0.835276	10210.000000	41.277514	51.513572	-12.000000	10.670000	25.670000	56.790000	1988.700000

```
Train set has 13 features with data from 10296 Customers
```

	Dtype	Null Count	Null Count in %	count	unique	top	freq
EducDeg	object	0	0.000000	10296	5	3 - BSc/MSc	4799

Image 3 – Distribution analysis

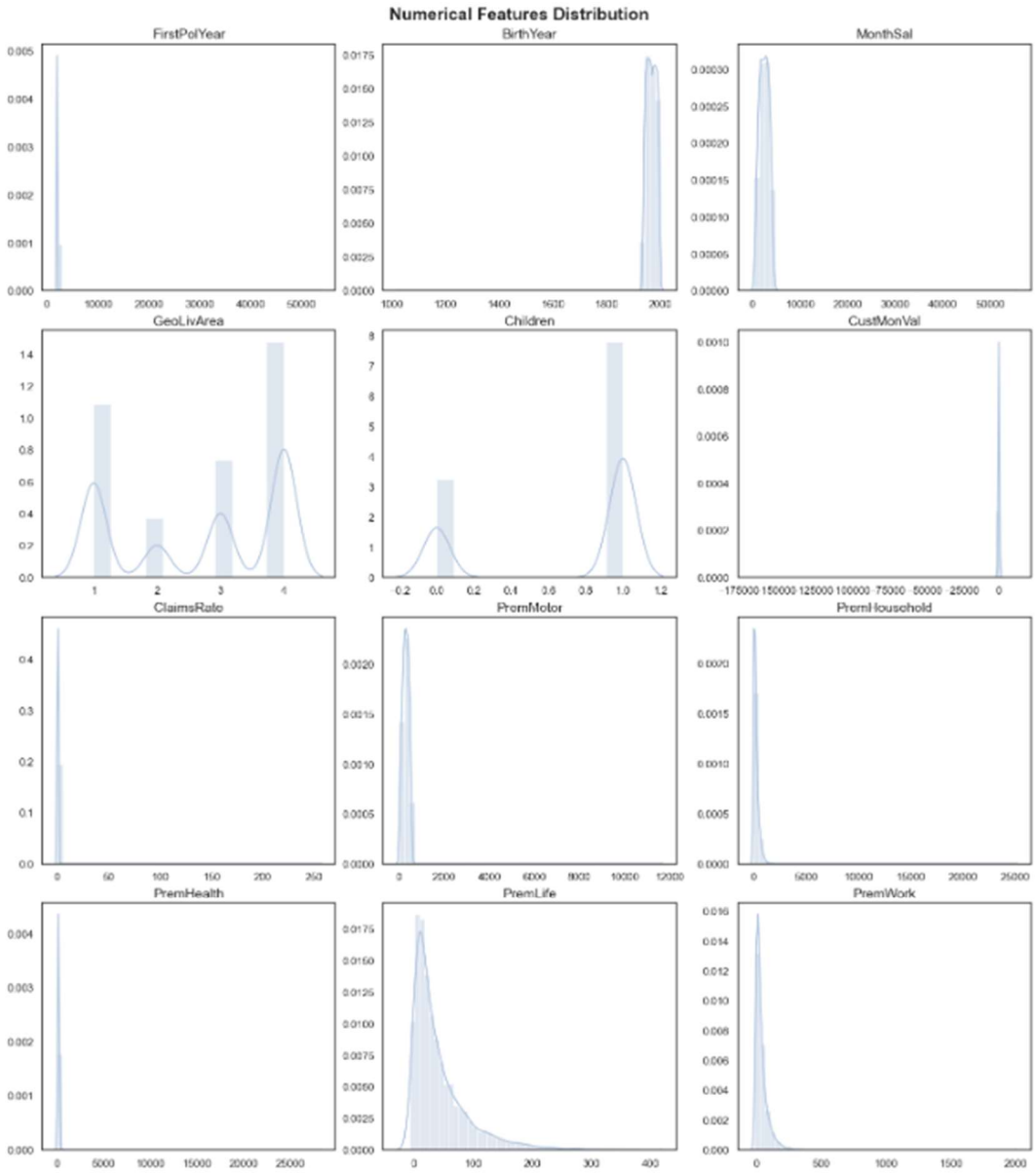


Image 4 - Qualitative check

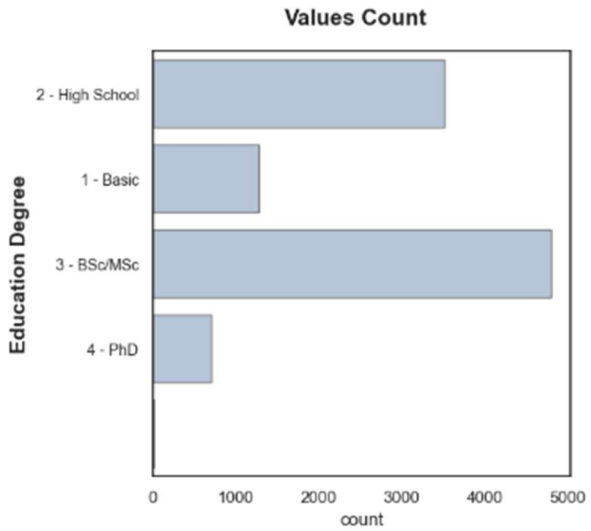


Image 5 – Three pairs of customers with different ID's but identical values

CustID	FirstPolYear	BirthYear	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife
2076	1977.0	1974.0	2 - High School	2204.0	4.0	1.0	-22.11	1.00	214.93	88.90	266.94	39.23
2100	1987.0	1987.0	2 - High School	1912.0	4.0	1.0	290.61	0.58	202.37	177.25	306.39	63.90
3507	1986.0	1952.0	2 - High School	3900.0	4.0	0.0	-119.35	1.10	163.03	481.75	224.82	94.35

CustID	FirstPolYear	BirthYear	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife
8014	1987.0	1987.0	2 - High School	1912.0	4.0	1.0	290.61	0.58	202.37	177.25	306.39	63.90
8122	1977.0	1974.0	2 - High School	2204.0	4.0	1.0	-22.11	1.00	214.93	88.90	266.94	39.23
9554	1986.0	1952.0	2 - High School	3900.0	4.0	0.0	-119.35	1.10	163.03	481.75	224.82	94.35

Image 6 - missing values.

	Feature	N° of missings	% of missings	Above threshold
PremLife	PremLife	104	0.01	False
PremWork	PremWork	86	0.01	False
PremHealth	PremHealth	43	0.00	False
MonthSal	MonthSal	36	0.00	False
PremMotor	PremMotor	34	0.00	False
FirstPolYear	FirstPolYear	30	0.00	False
Children	Children	21	0.00	False
BirthYear	BirthYear	17	0.00	False
GeoLivArea	GeoLivArea	1	0.00	False

- Seems that only numerical features have missing values, but the categorical column (**Education Degree**) can possibly have strange character values
- Also the percentages of missing values on the data are low **(bellow 5% in all columns with NAN) so they could be dropped**
- Imputing data on Premium columns would **assume that the Customers with missing data have these insurances policies contracted, which might not be the case (lack of information in this matter)**
- Due to **lack of clarity and low percentage of the data, the missing values will be dropped**

Image 7 – Exchange of feature "First Policy Year" with "Birth Year"

Some values when First Policy Year occurs priorly to Birth Year:

Age	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
BirthYear	12	34	68	93	124	126	127	143	113	127	111	122	94	81	87	91	73	68	53	47	39	42	28	25	14	3	3

Age	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41
Children	7	23	55	72	89	80	100	106	84	104	95	97	89	78	78	86	68	65	49	44	37	41	24	24	13	2	3

Some Values if this costumers had their First Policy Year switched with Birth Year:

Age	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
BirthYear	1	11	12	18	23	34	45	45	50	71	80	67	91	85	92	95	121	135	113	139	154	163	151	112	40

Age	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Children	1	9	10	13	18	24	32	39	41	55	66	53	69	69	82	74	101	116	92	115	130	144	127	99	34

Image 8 – Spearman Correlation Heatmap

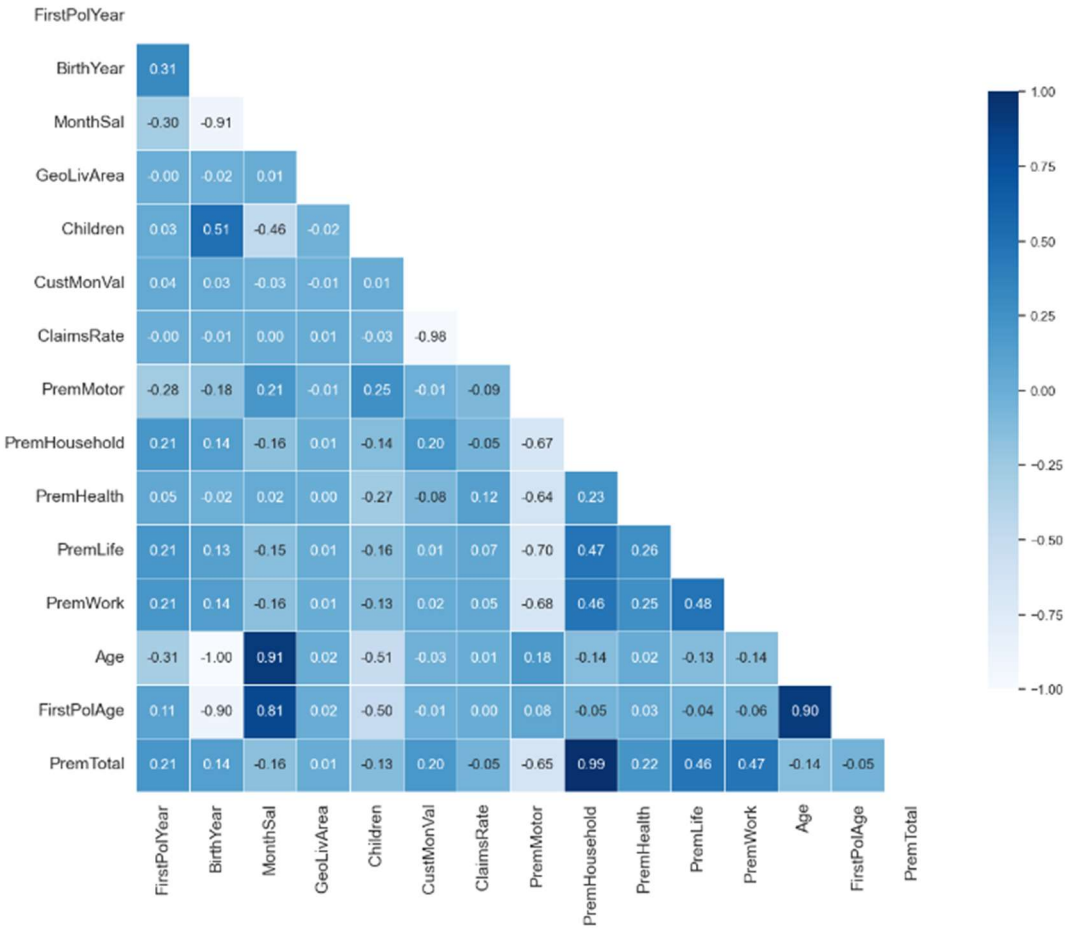


Image 9 - Interquartile Range Method (IQR)

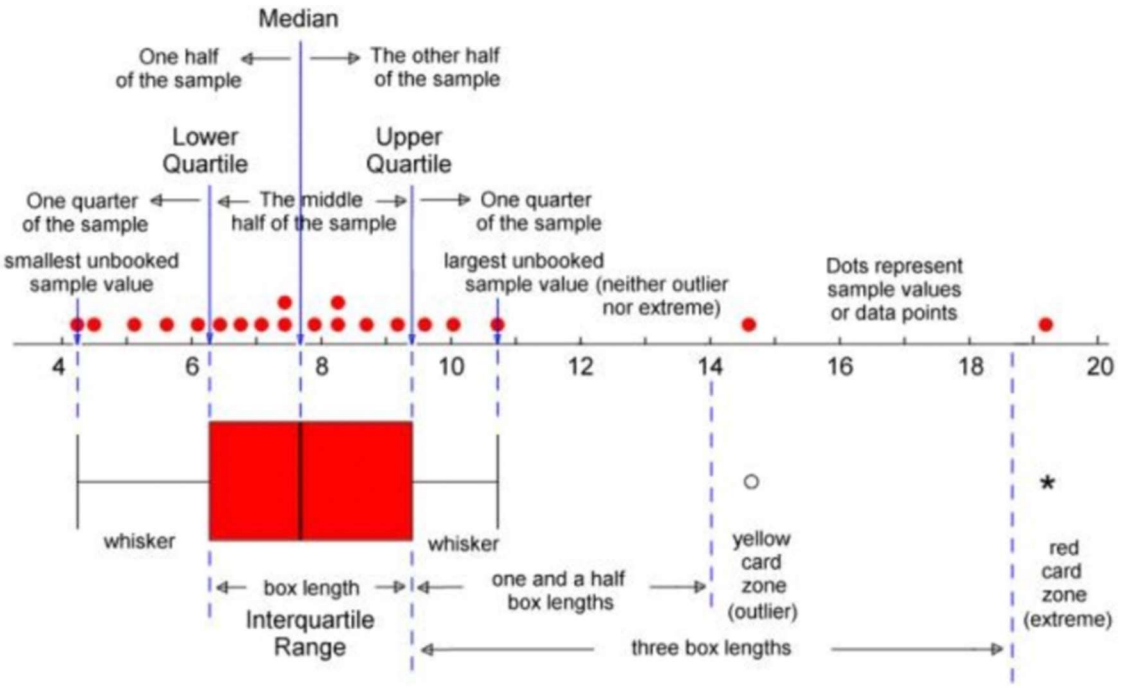


Image 10 - Interquartile Range boxplots on data

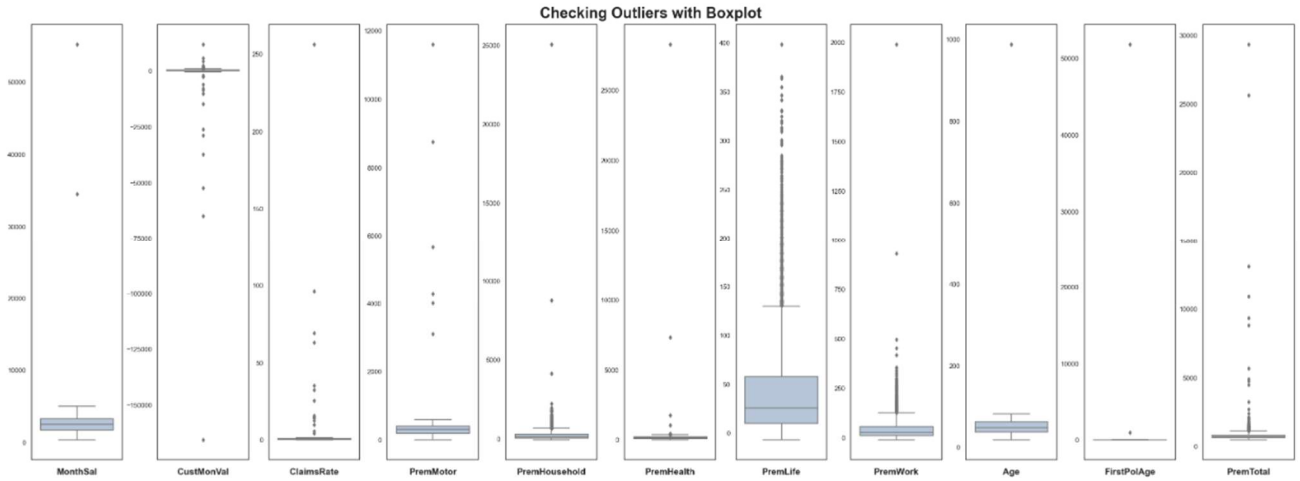
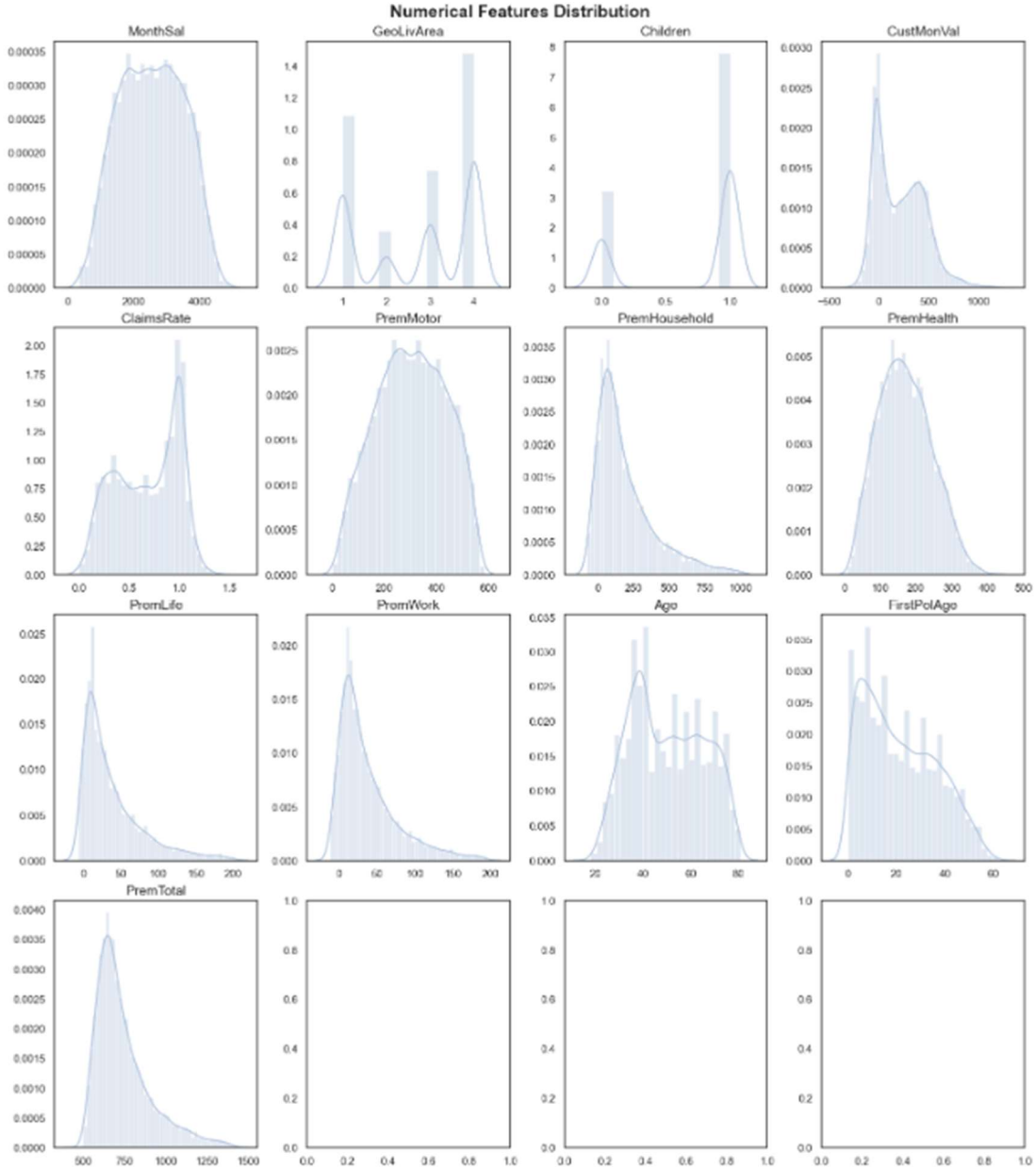


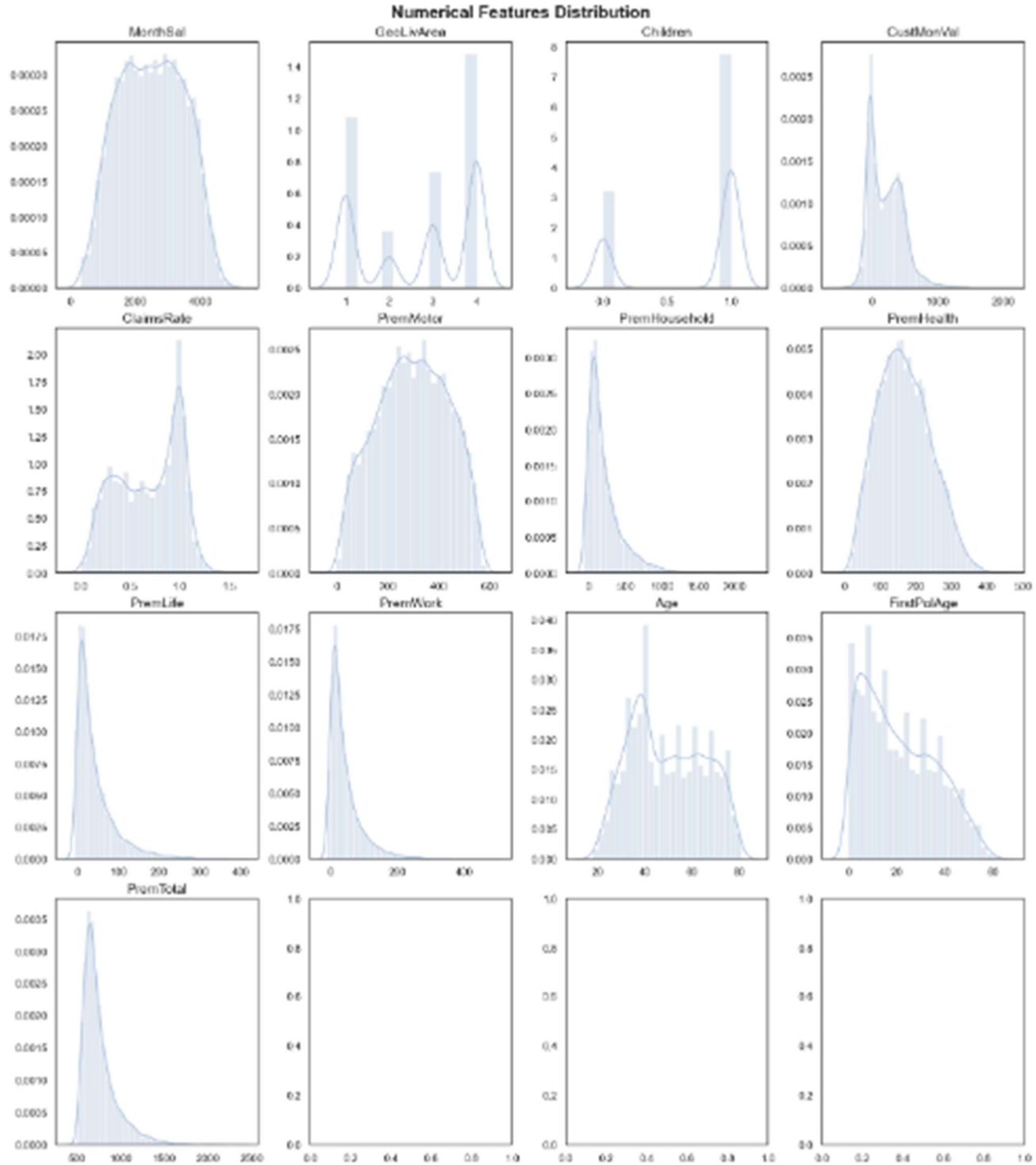
Image 11 – Data distribution after IQR Outliers exclusion



Percentage of remaining data after outlier removal with IQR method:

95.823 % of remaining data after IQR method outlier removal, when criterion is 3

Image 12 – Data distribution after Manual Outliers exclusion




```
After excluding the outliers manually, the dataset will remain with 99.7 % of its original Customers

Number and Percentage (of total data) of Outliers per Feature:

There is 1 Customer (0.01% of total data) with Age above 115
There are 2 Customers (0.02% of total data) with MonthSal above 30000
There are 14 Customers (0.14% of total data) with CustMonVal bellow -2000
There are 14 Customers (0.14% of total data) with ClaimsRate above 4
There are 6 Customers (0.06% of total data) with PremMotor above 2000
There are 3 Customers (0.03% of total data) with PremHousehold above 4000
There are 2 Customers (0.02% of total data) with PremHealth above 5000
There are 2 Customers (0.02% of total data) with PremWork above 750
```

Image 13 – Mean absolute error per scaler method

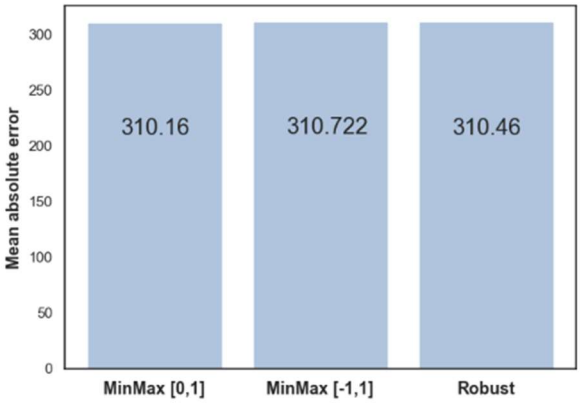
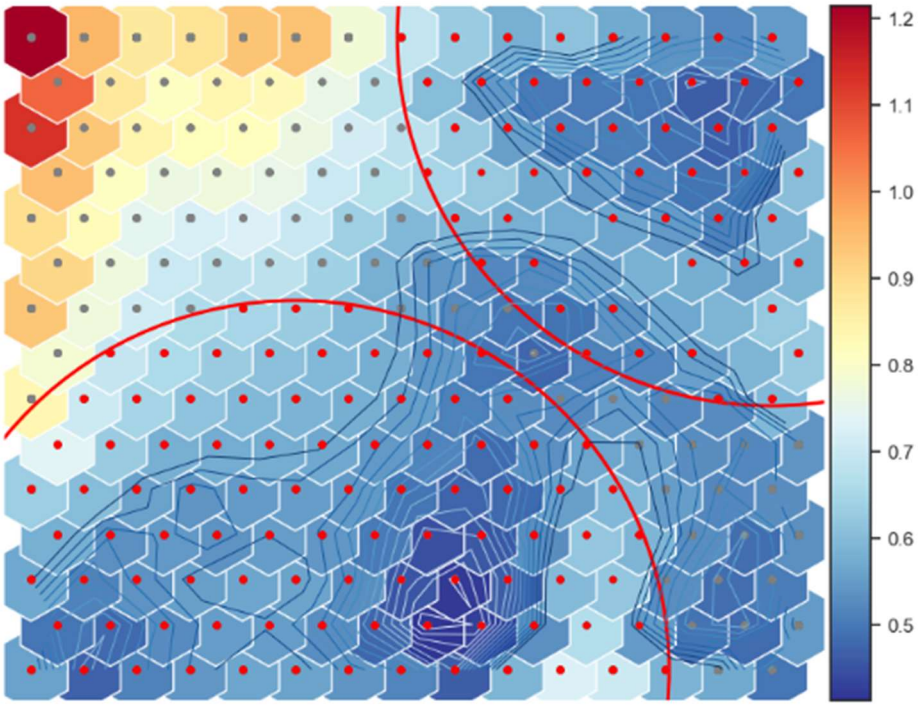


Image 14 – Data scaled by Min Max

	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife	PremWork
count	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000	9951.000000
mean	0.491274	0.461490	0.571165	0.705758	0.278951	0.439294	0.508255	0.123714	0.406659	0.120911	0.105140
std	0.265301	0.210569	0.422327	0.455724	0.101904	0.204461	0.233596	0.103179	0.176308	0.117543	0.092950
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.333333	0.291382	0.000000	0.000000	0.189744	0.251613	0.329247	0.054986	0.272120	0.041673	0.044794
50%	0.666667	0.462244	0.666667	1.000000	0.266800	0.464516	0.511411	0.090875	0.393744	0.080336	0.075756
75%	0.666667	0.631186	1.000000	1.000000	0.349829	0.632258	0.695739	0.160740	0.529148	0.160128	0.135922
max	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000

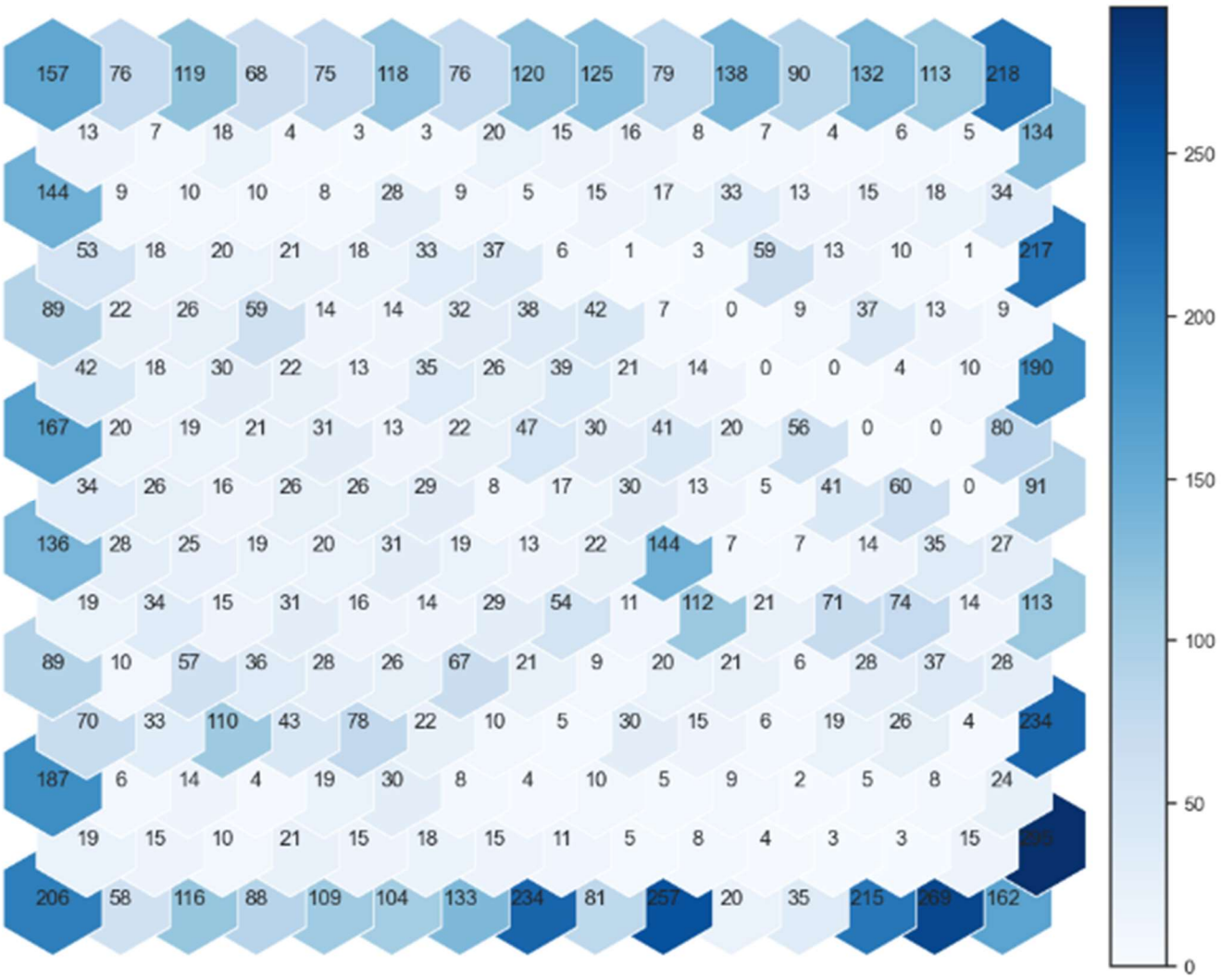
Image 15 – U-Matrix of entire data



Interpretation

- Is notable that data have 3 or 4 type of data distances:
  - In lower right corner, compounded by elements with height distances (Red and Yellow)
  - Lower right corner, top mid and top left corner are 3 groups compounded with less distances (Darker Blue)
  - Between the last two groups are the rest of data that a median distance (light blue)

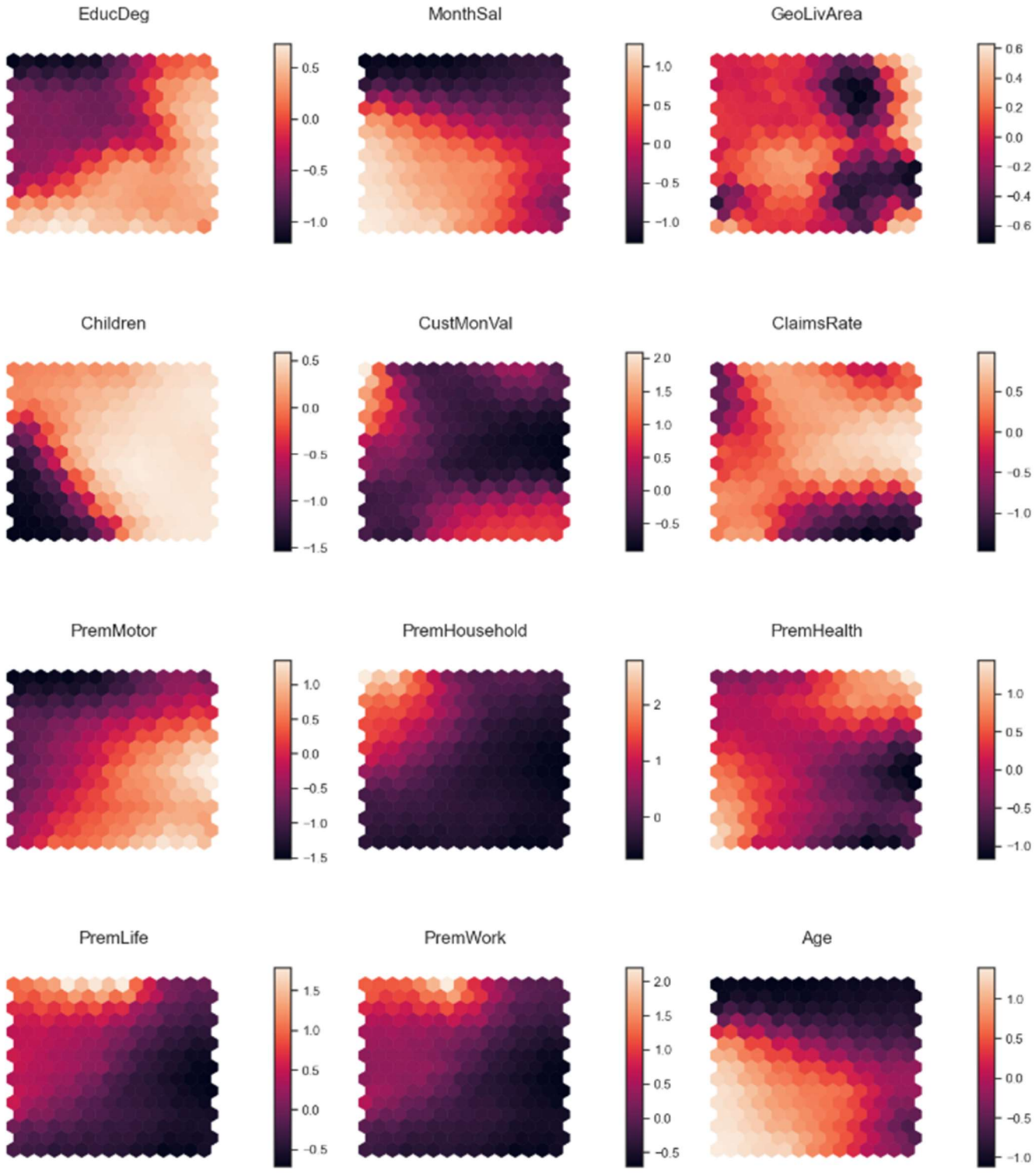
Image 16 – Hit map from entire data



Interpretation

- Is notable that data have 3 or 4 type of data frequency:
  - In border, is notable that exist elements with height frequencies (Darker Blue)
  - In middle, is understandable that makority of elements with lower frequencies (Lighter Blue and White)

Image 17 – Component planes from entire data



Interpretation

- With hit map is possible to check some correlation between features:
  - PremLife and PremWork are hight positivly correlated
  - MonthSal and Age are hight positivly correlated
  - CustMonVal and ClaimsRate are negative correlated
  - PremMotor and PremHousehold are negative correlated

Image 18 – Component planes from entire data

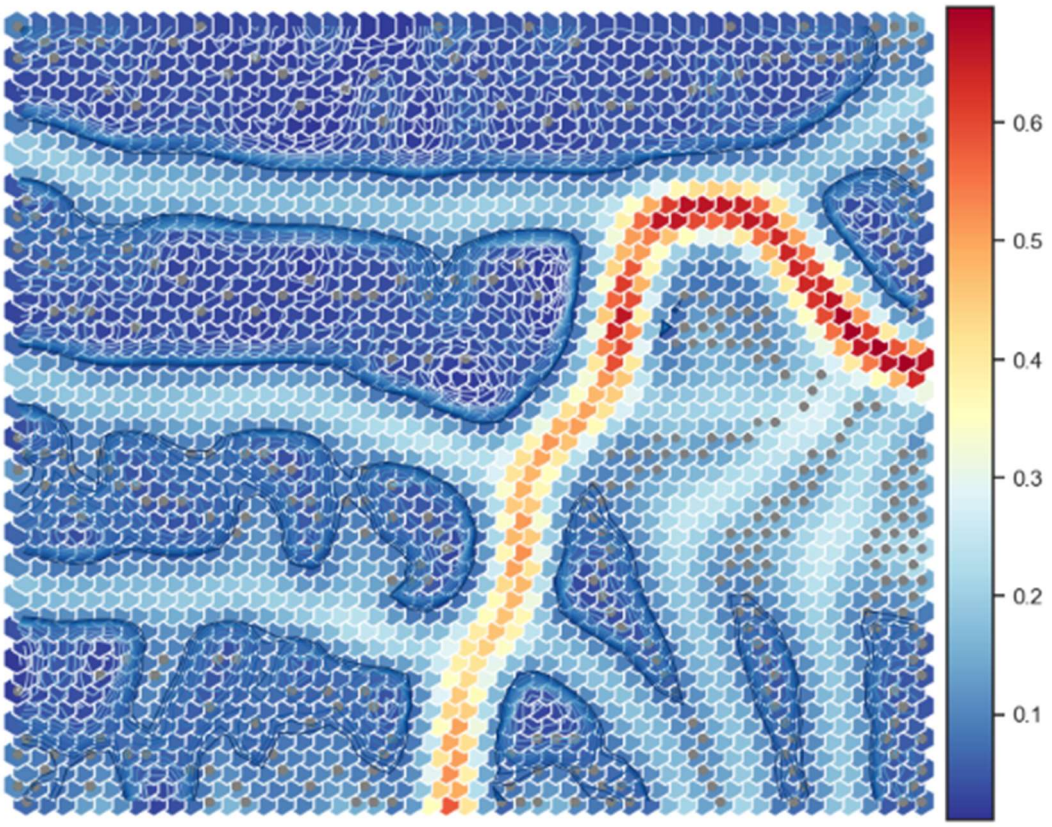
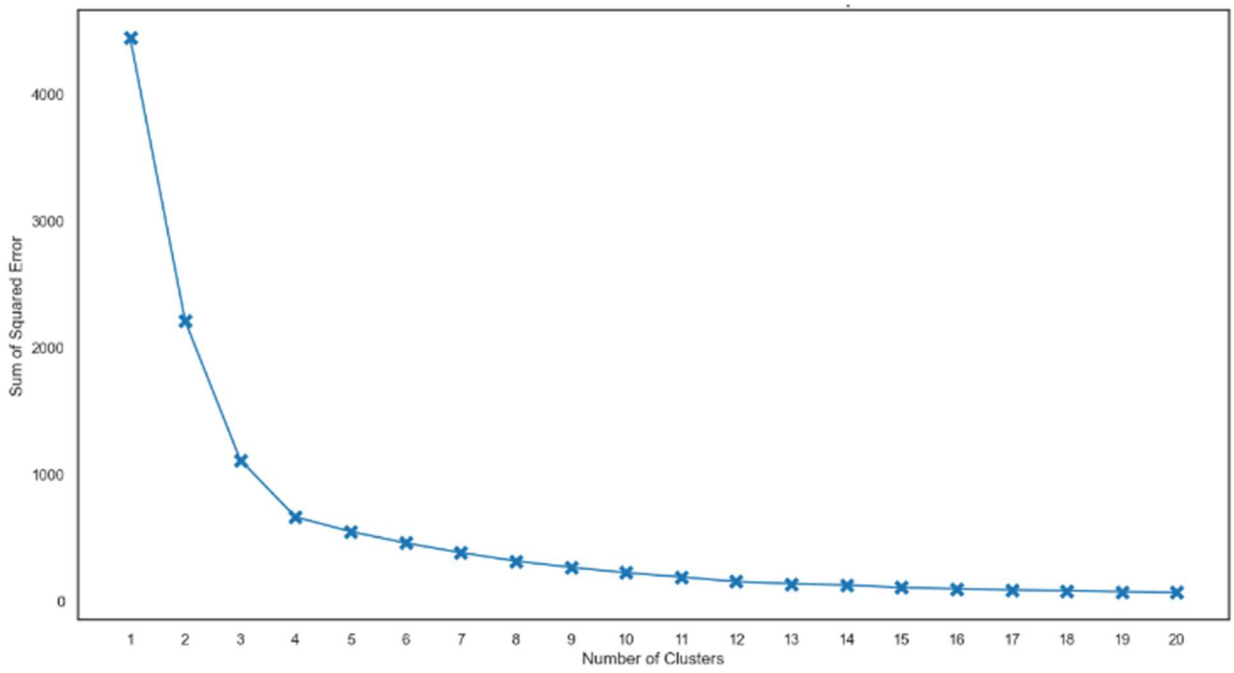


Image 19 – Elbow Method for Socialdemographic Level




```
Total error: 69.42

Best candidates for number of clusters: 3, 4 and 5;
```

Image 20 – Silhouette Score for Socialdemographic Level

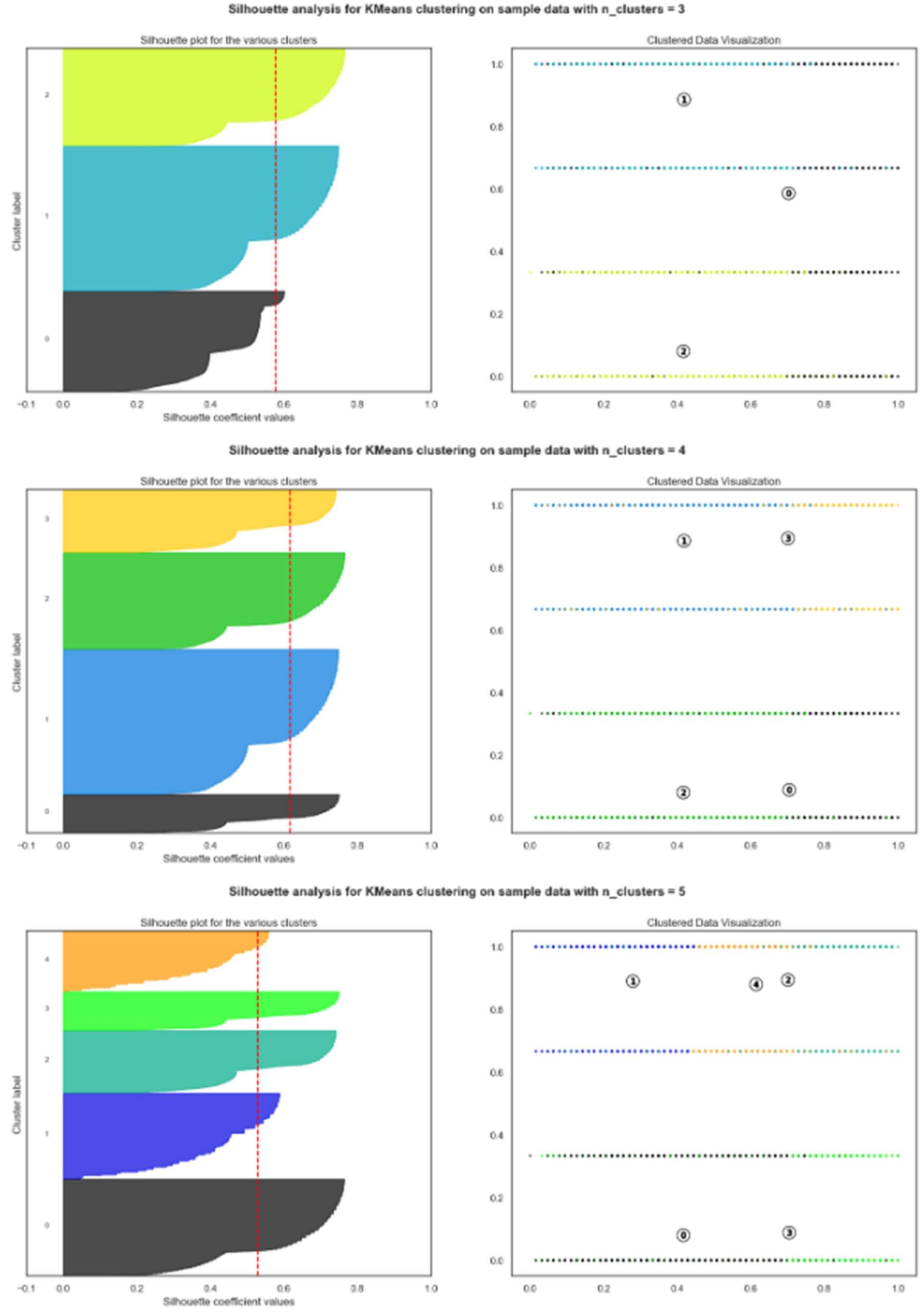
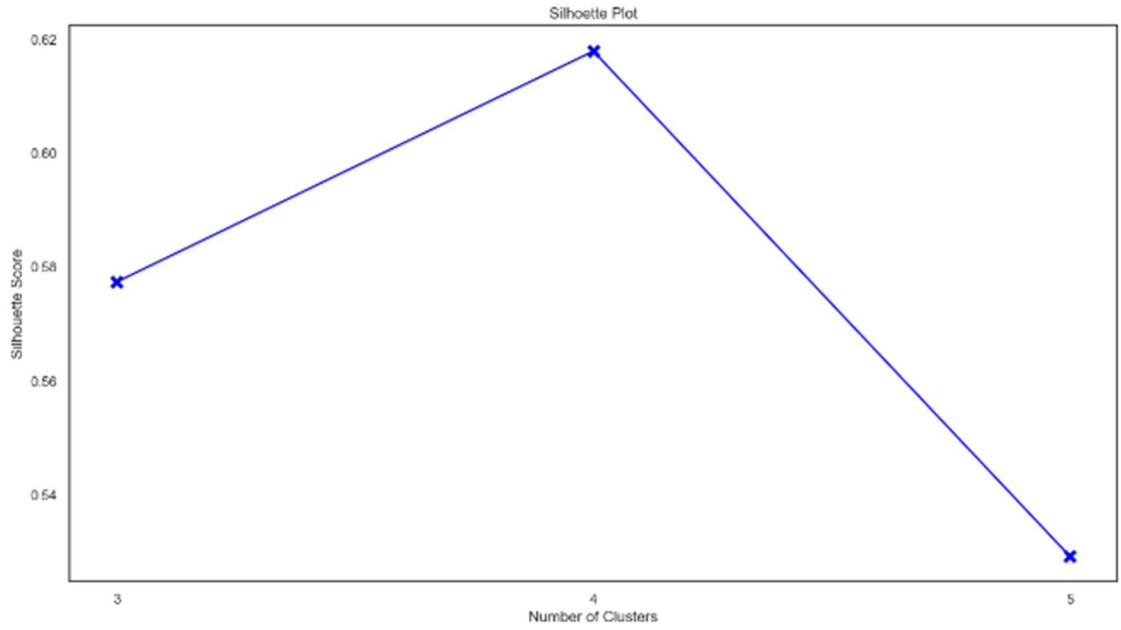


Image 21 – Silhouette Score for Socialdemographic Level




```
For n_clusters = 3 The average silhouette score is : 0.577
For n_clusters = 4 The average silhouette score is : 0.618
For n_clusters = 5 The average silhouette score is : 0.529
```

Image 22 – R2 plot for various hierarchical methods

Image 23 – Hierarchical Clustering Dendrogram

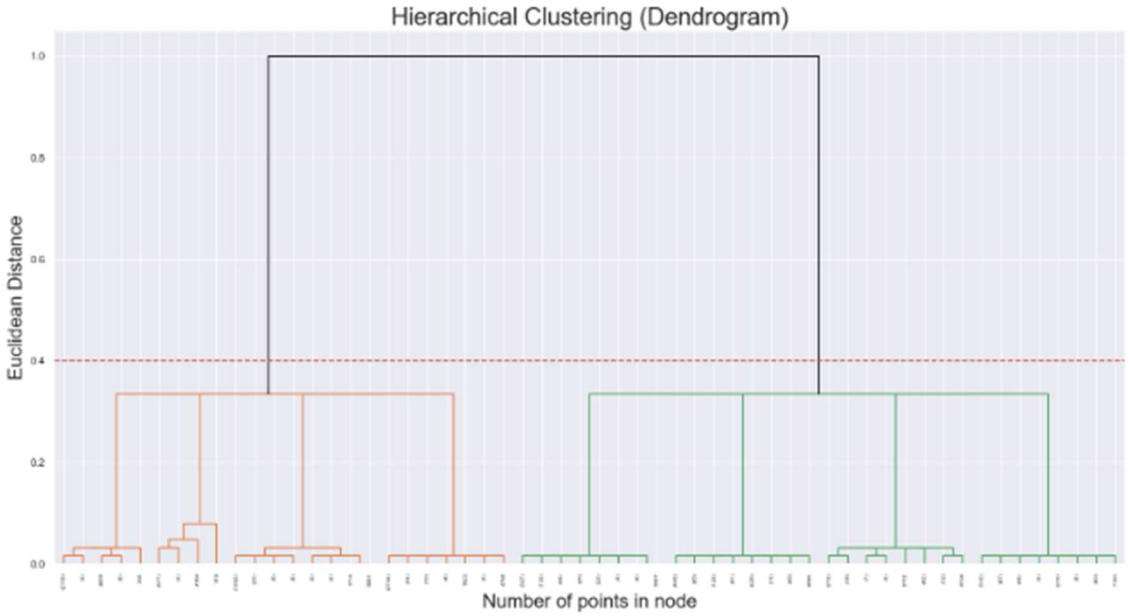


Image 24 – Hierarchical Clustering with Kmeans

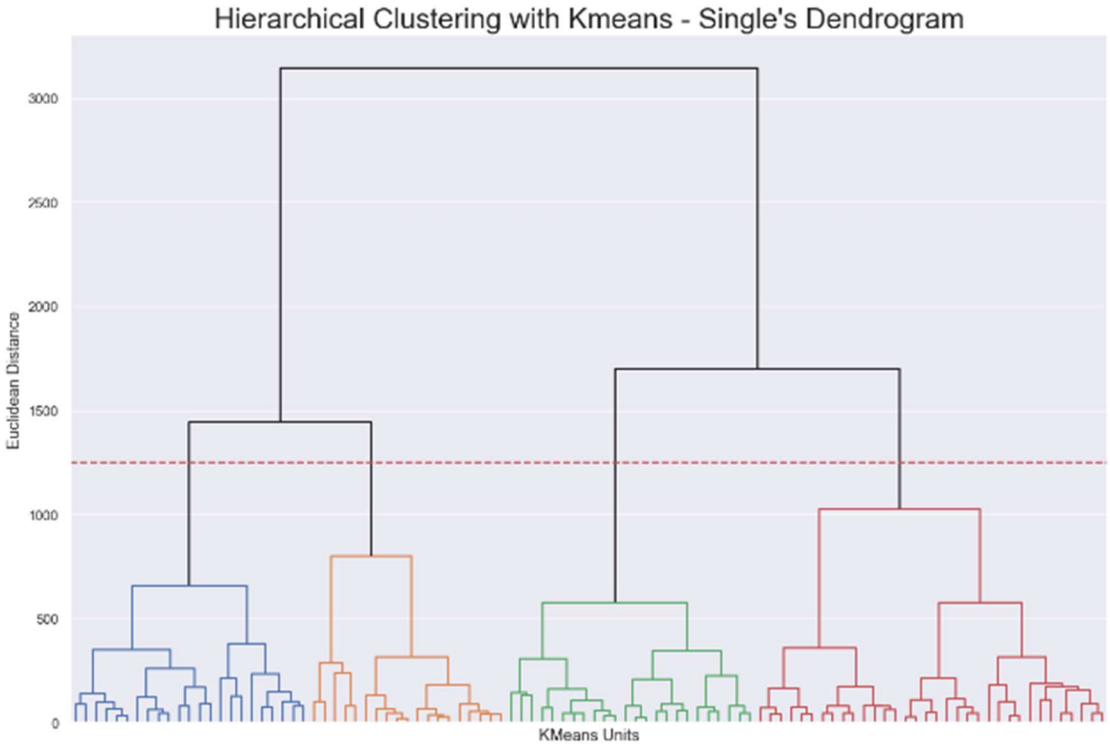
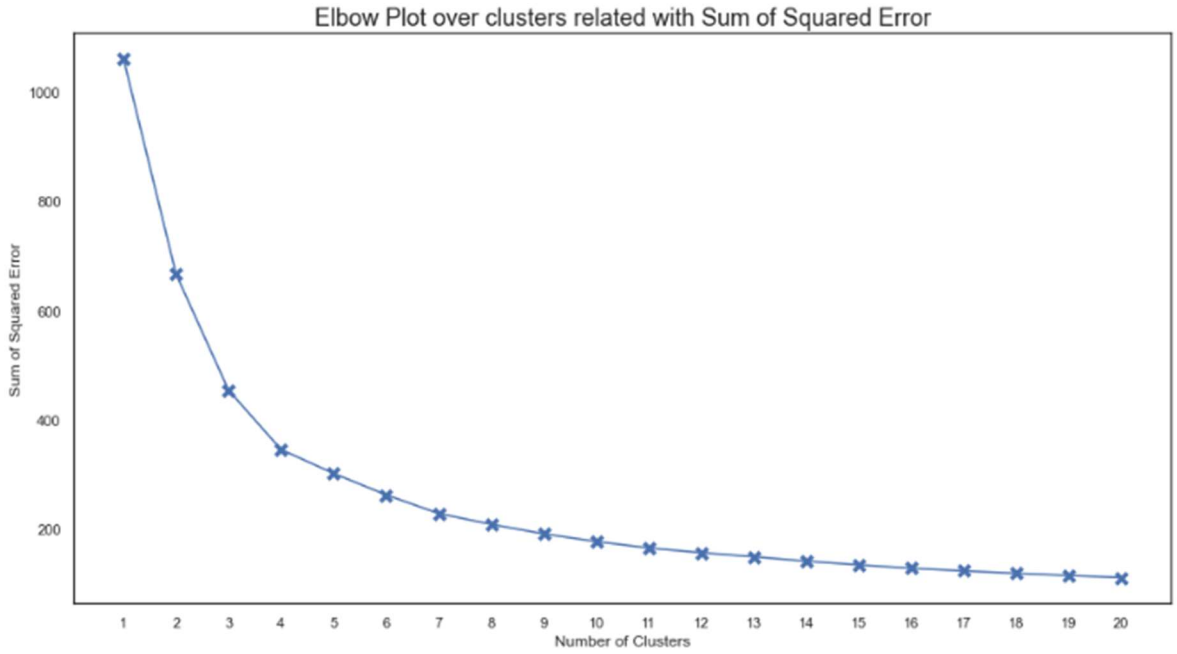


Image 25 – Elbow Method for Customer Value Level




```
Total error: 110.81

Best candidates for number of clusters: 3, 4 and 5;
```

Image 26 – Silhouette Score for Customer Value Level

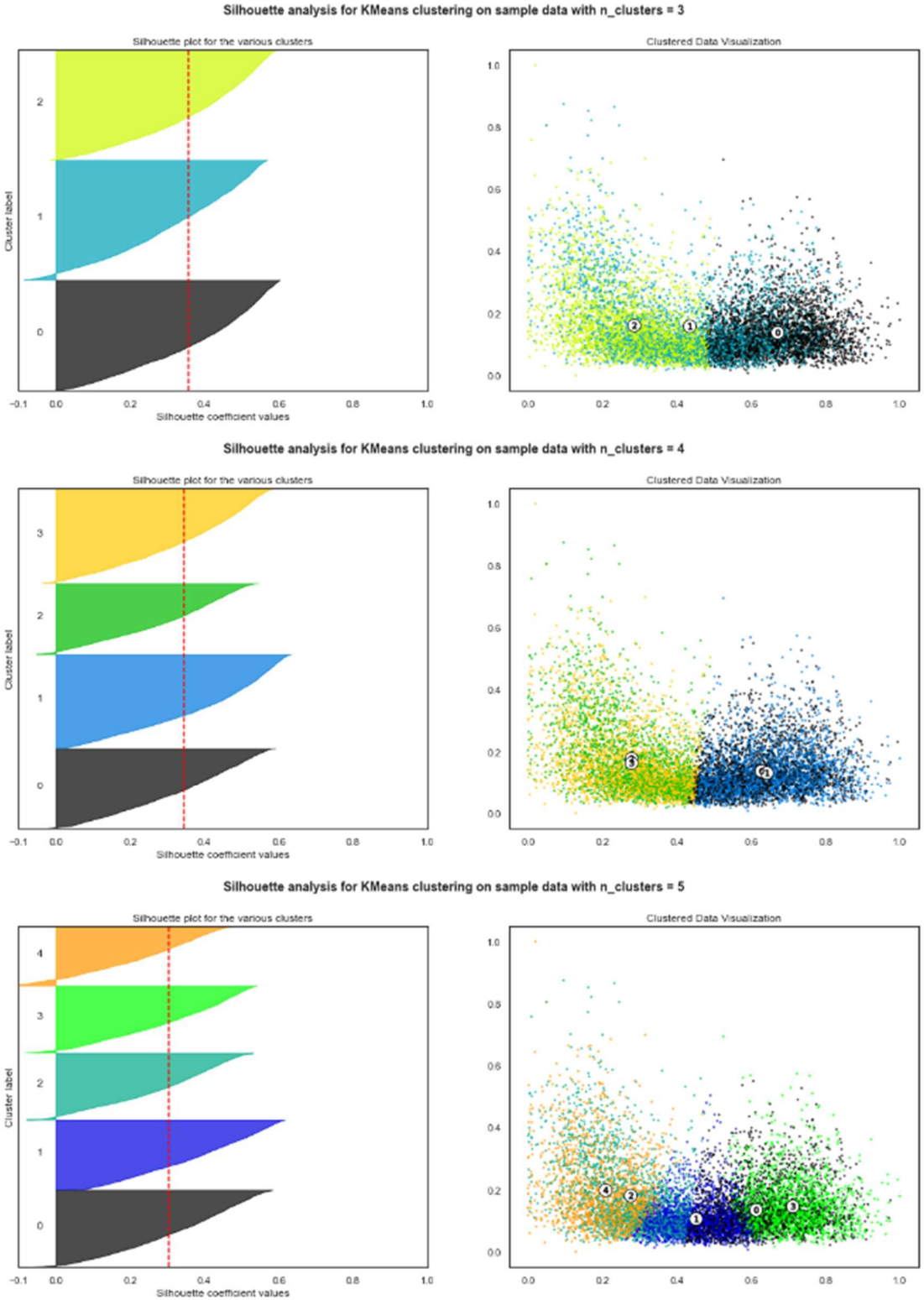
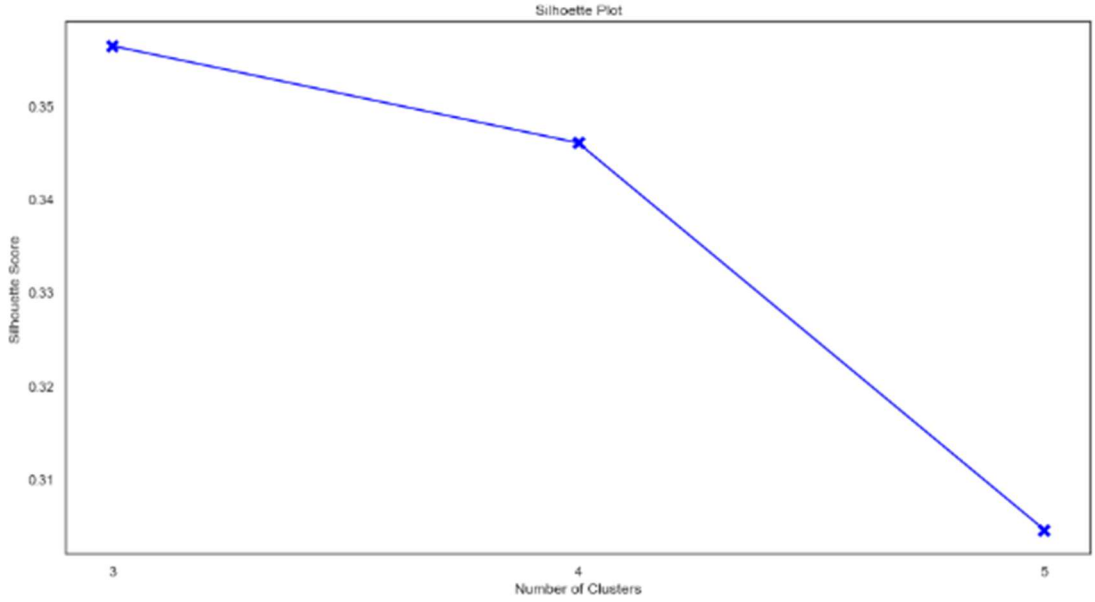


Image 27 – Silhouette Score for Customer Value Level




```
For n_clusters = 3 The average silhouette score is : 0.356
For n_clusters = 4 The average silhouette score is : 0.346
For n_clusters = 5 The average silhouette score is : 0.305
```

Image 28 – R2 plot for various hierarchical methods

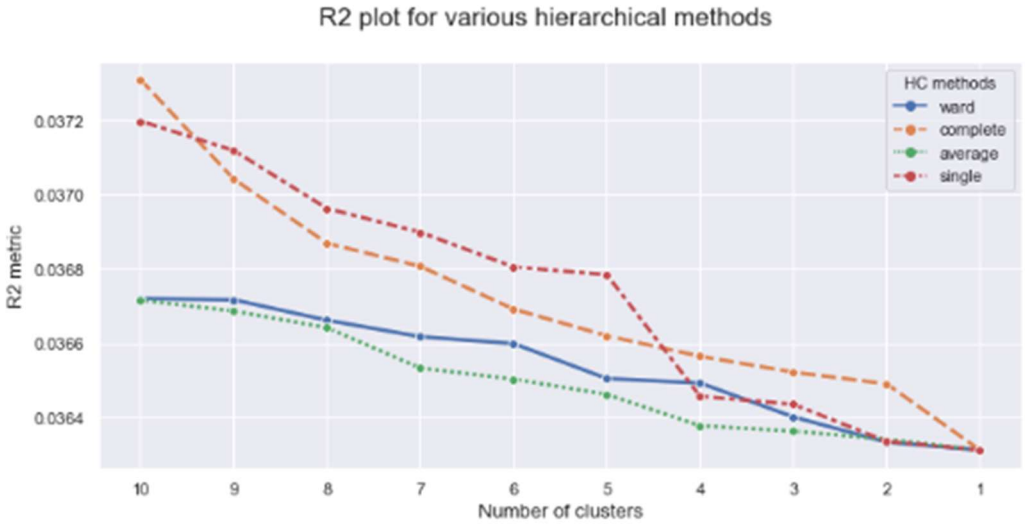


Image 29 – Hierarchical Clustering Dendrogram

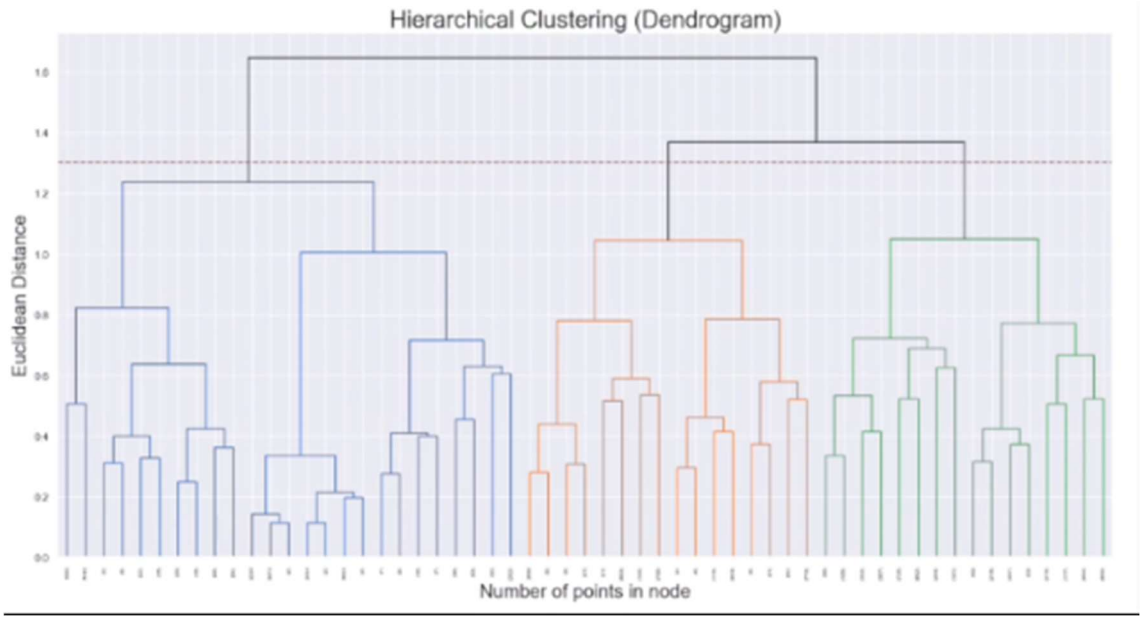


Image 30 – Hierarchical Clustering with Kmeans

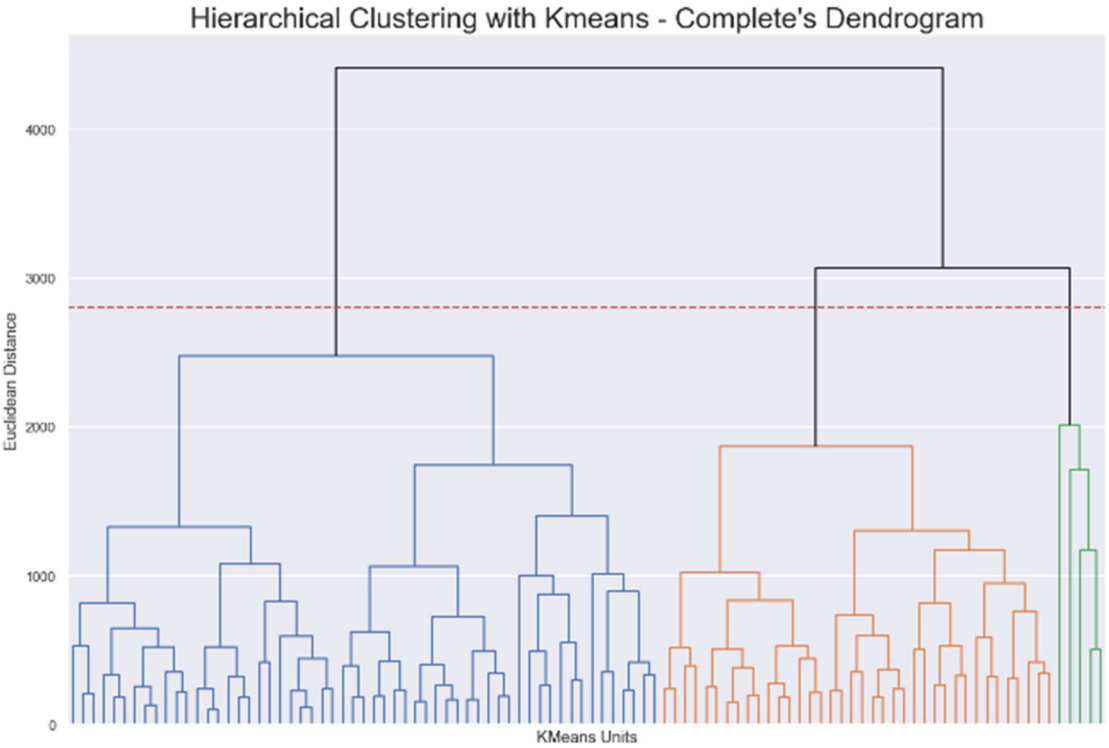


Image 31 –SOM + Kmeans for Social Demographic Segmentation

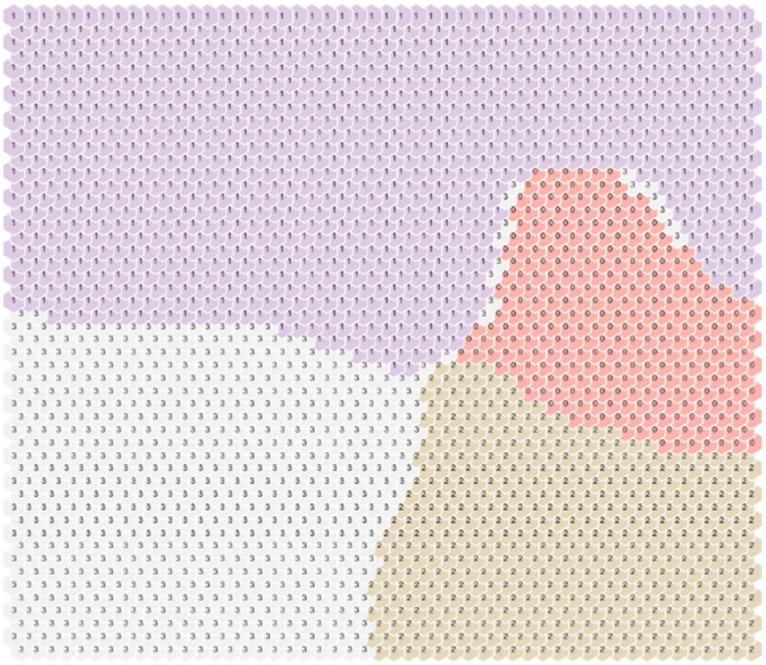


Image 32 –SOM + Hierarchical for Social Demographic Segmentation

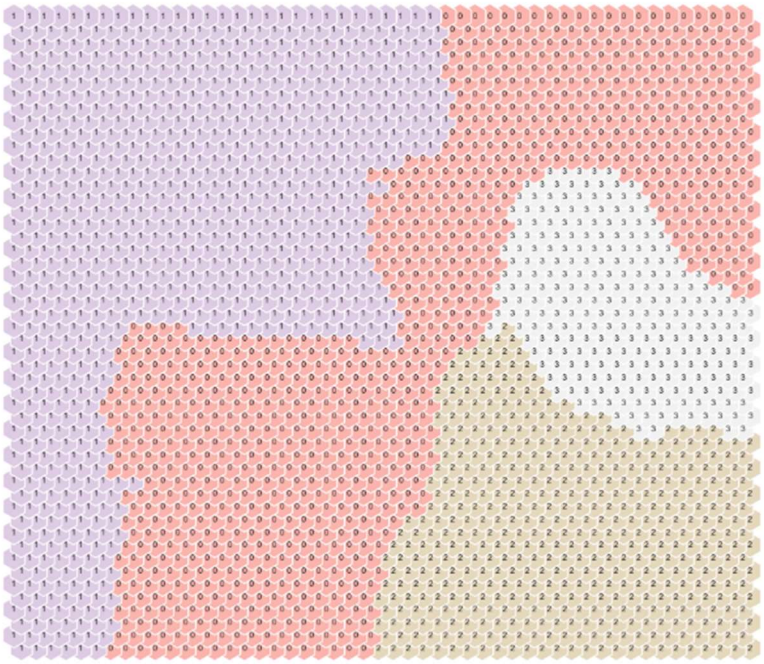


Image 33 –Best number of Epsilon/Radius

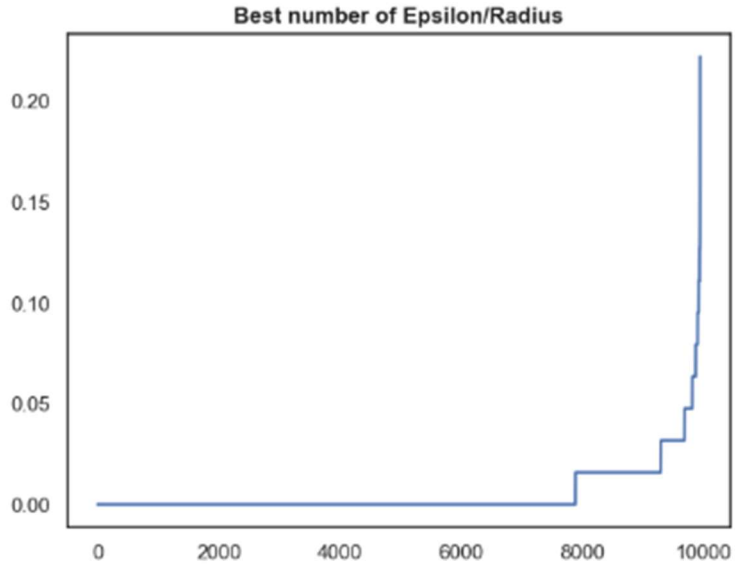


Image 34 –Visual Comparison of Models

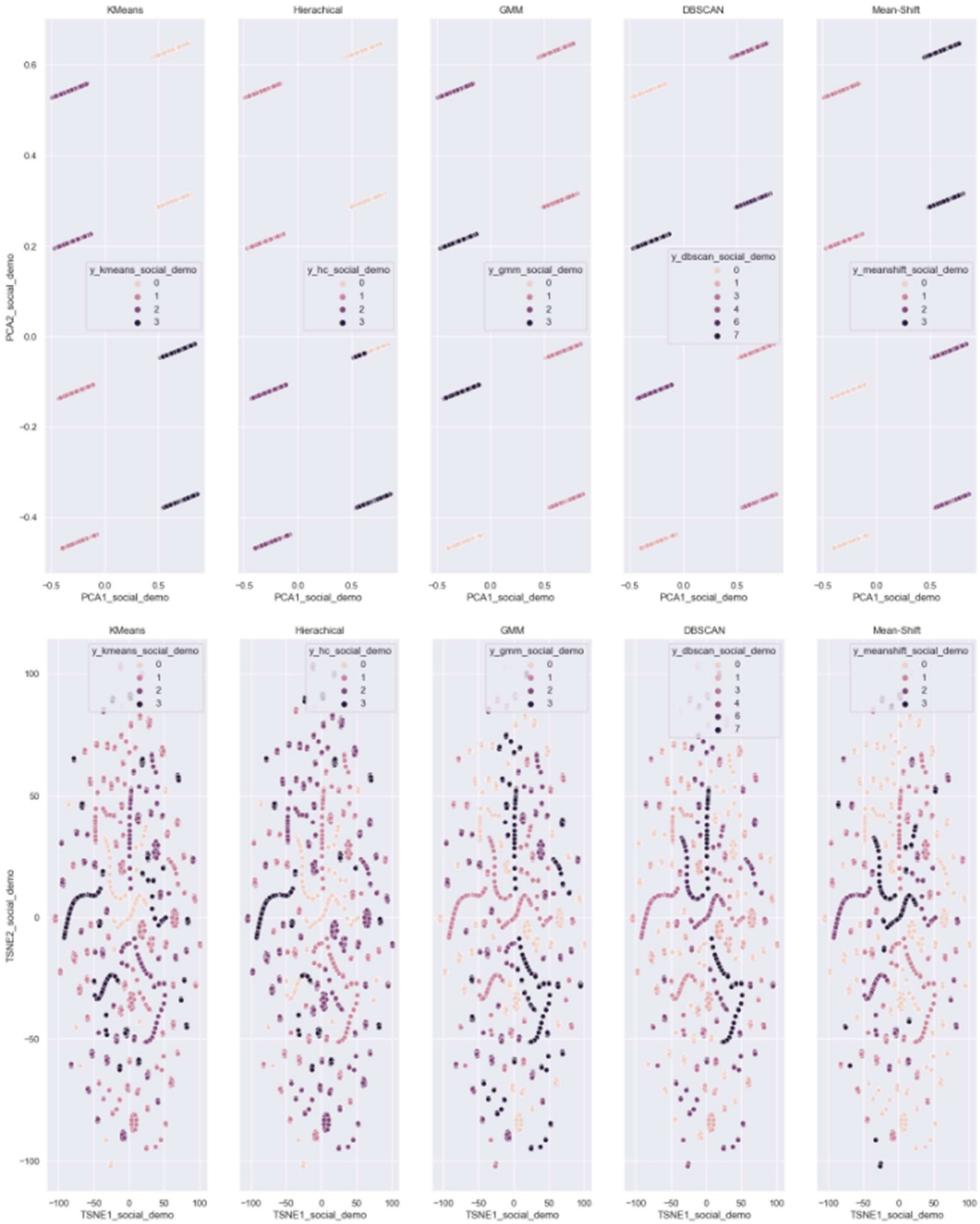


Image 35 – Benchmarking Algorithms

	DBI	CHS	SS
K-Means	0.593312	18870.192429	0.618007
Hierarchical	0.679066	16624.068388	0.581965
Gaussian	0.790177	11394.086045	0.493166
DBSCAN	0.857849	10075.721383	0.532640
Mean-Shift	0.593312	18870.192429	0.618007

Image 36 – Best number of Epsilon/Radius

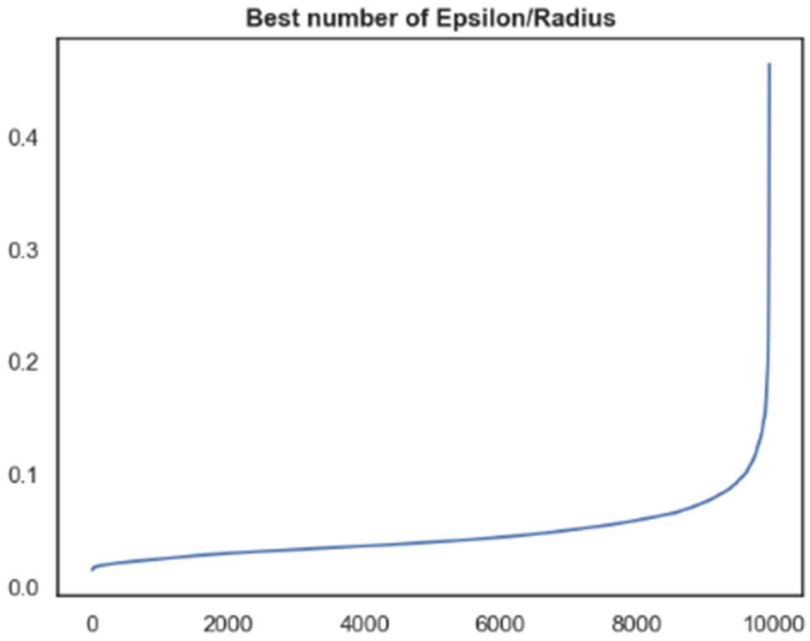


Image 37 –Visual Comparison of Models

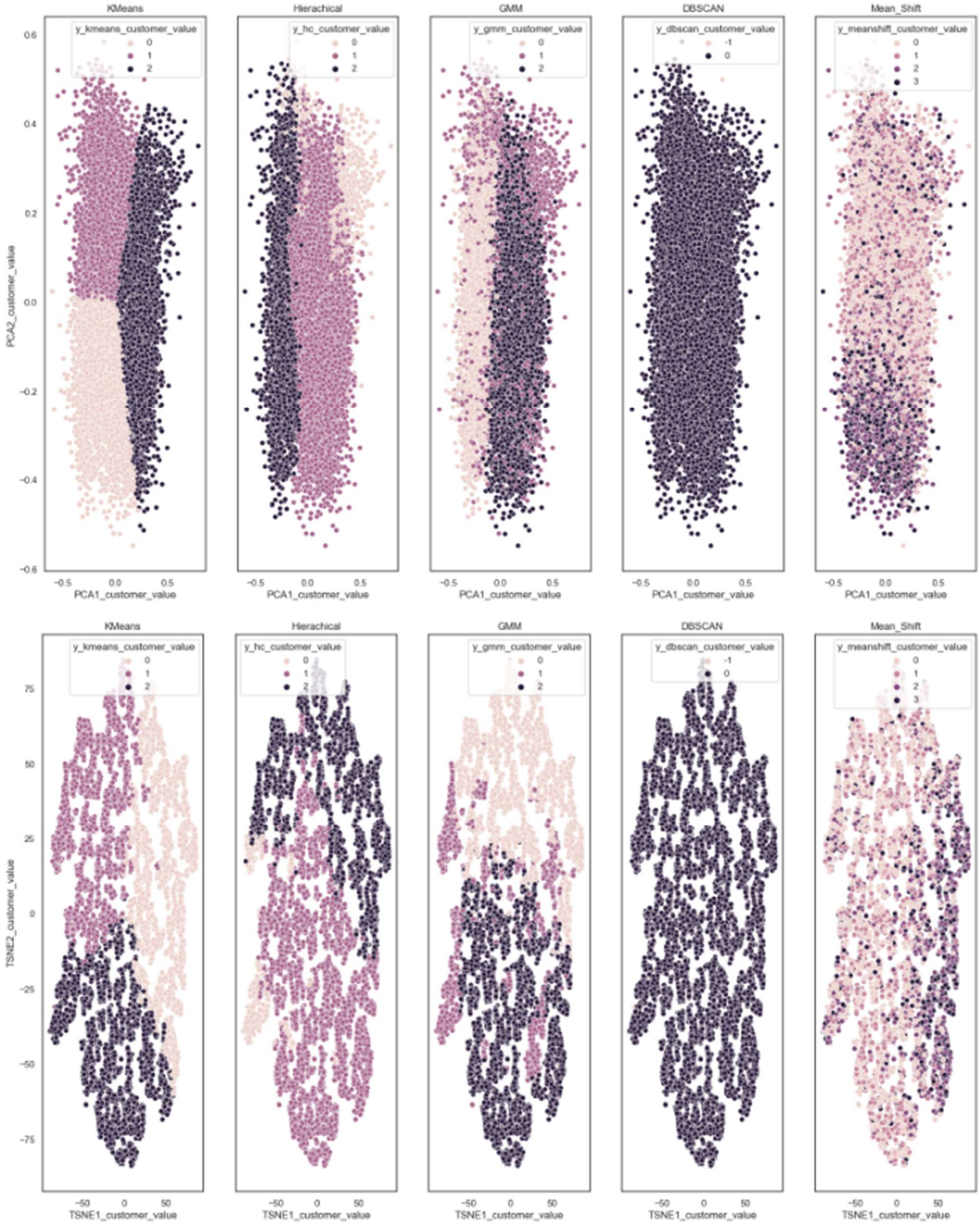


Image 38 – Benchmarking Algorithms

	DBI	CHS	SS
K-Means	0.920403	6650.671016	0.356366
Hierarchical	1.427981	2881.237610	0.235956
Gaussian	1.988099	2657.292652	0.255285
DBSCAN	0.386437	5.895035	0.478273
Mean-Shift	128.717244	325.934002	-0.037817

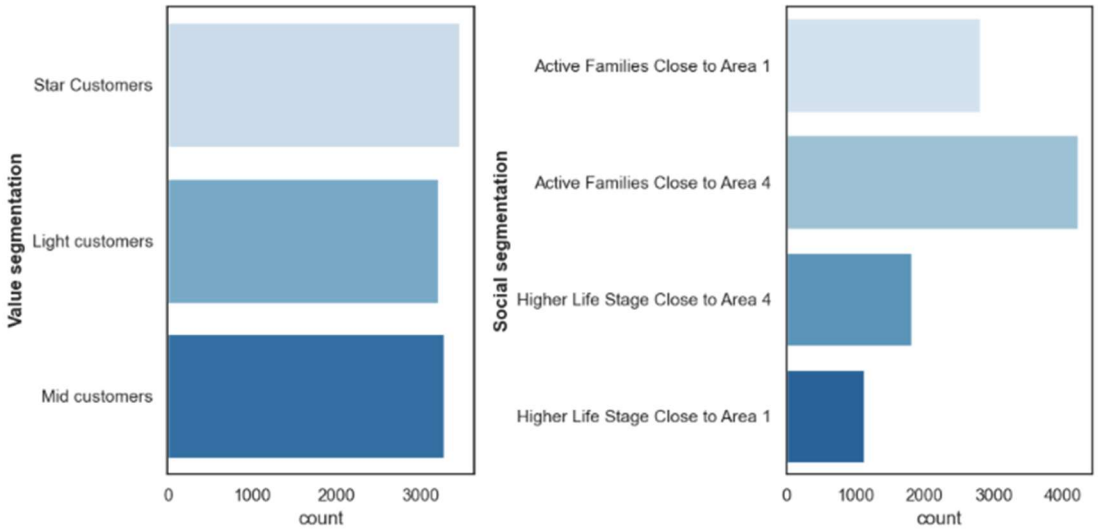
Image 39 – Clusters of Socialdemographic Level

Social_segmentation	Age	GeoLivArea	Children
Active Families Close to Area 1	44.195939	1.242608	1.0
Active Families Close to Area 4	44.258776	3.660816	1.0
Higher Life Stage Close to Area 4	62.160862	3.683250	0.0
Higher Life Stage Close to Area 1	62.307417	1.266309	0.0

Image 40 – Clusters of Customer Value Level

Value_segmentation	MonthSal	PremTotal	CustMonVal	ClaimsRate
Light customers	1664.553344	772.184781	47.663956	0.908880
Star Customers	2345.480474	768.329242	497.945904	0.316960
Mid customers	3471.318390	728.562342	94.503968	0.841086

Image 41 – Clusters of Customer Value Level



5*Image 41 – Clusters Combined*

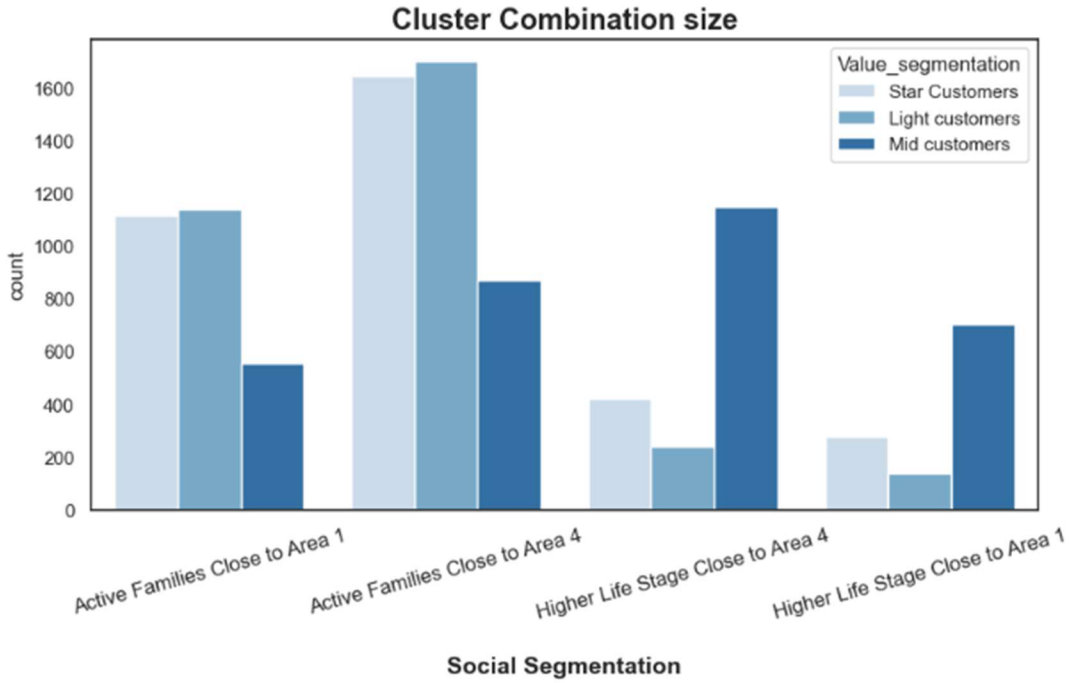


Image 42 – Clusters Combined

Social_segmentation	PremTotal	Prem_perc
Active Families Close to Area 4	1293549.02	17.0
Active Families Close to Area 1	862036.15	11.0
Higher Life Stage Close to Area 4	206988.23	3.0
Higher Life Stage Close to Area 1	120000.67	2.0

Image 43 – Clusters Combined

Social_segmentation	PremTotal	Prem_perc
Active Families Close to Area 4	1227579.16	16.0
Active Families Close to Area 1	832800.09	11.0
Higher Life Stage Close to Area 4	357961.58	5.0
Higher Life Stage Close to Area 1	237773.36	3.0

Image 44 – Clusters Combined

Social_segmentation	CustMonVal	ClaimsRate	PremTotal
Higher Life Stage Close to Area 1	563.537292	0.320542	858.387581
Higher Life Stage Close to Area 4	555.460523	0.319691	850.265036
Active Families Close to Area 4	484.804444	0.313240	746.248729
Active Families Close to Area 1	479.306077	0.320530	747.576382

Image 45 – Correlation spearman heatmap

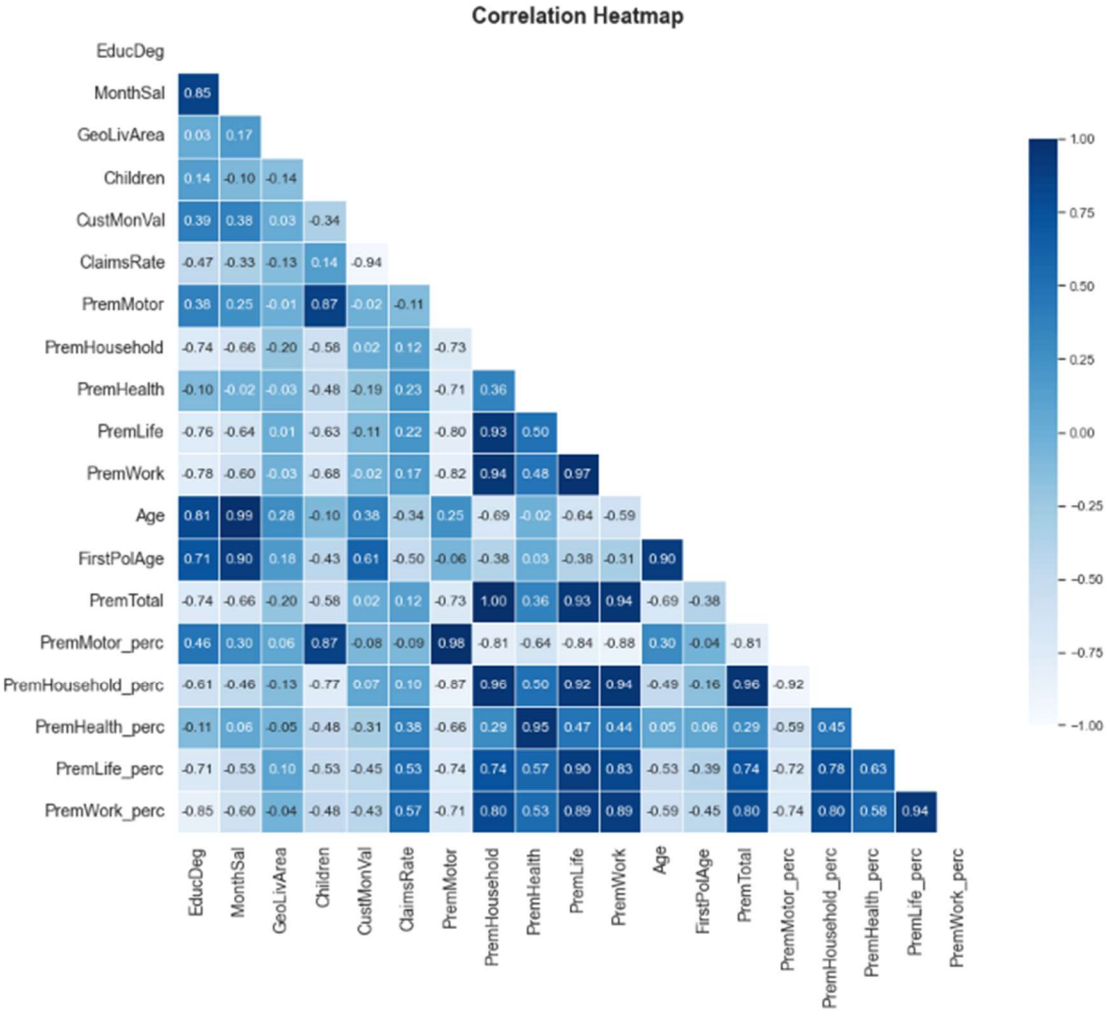


Image 46 – Cluster merge Dendrogram

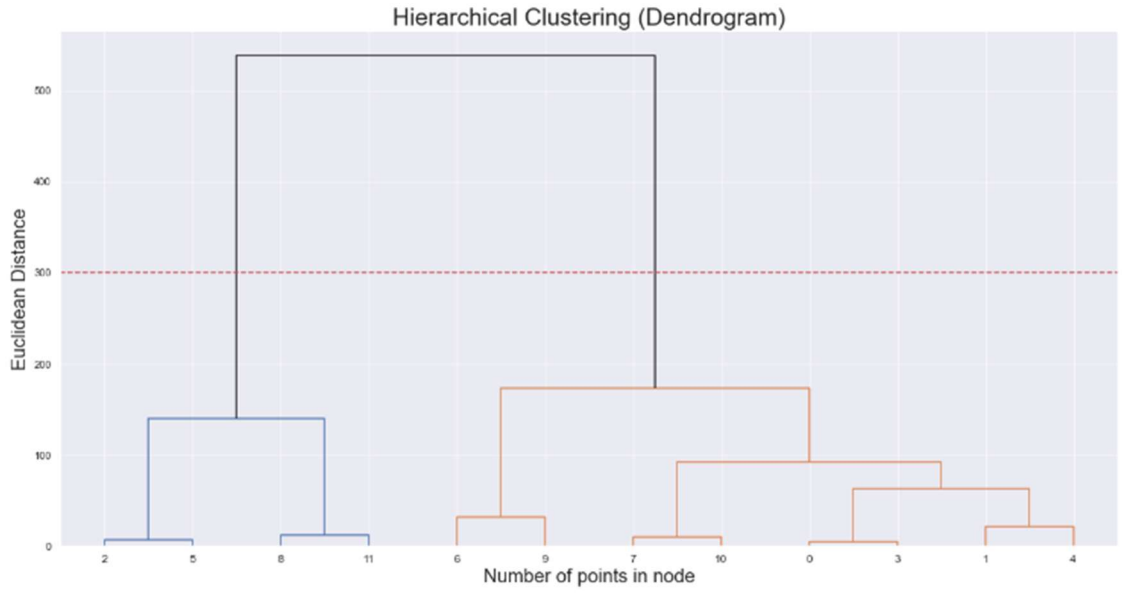


Image 47 – Cluster profile

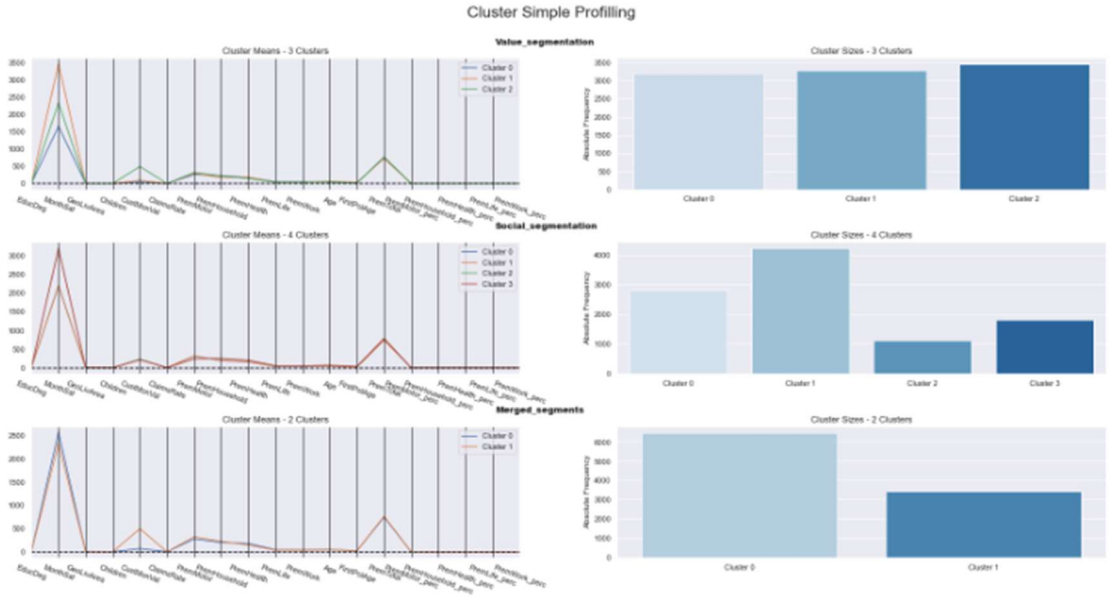


Image 48 – Cluster profile T-SNE

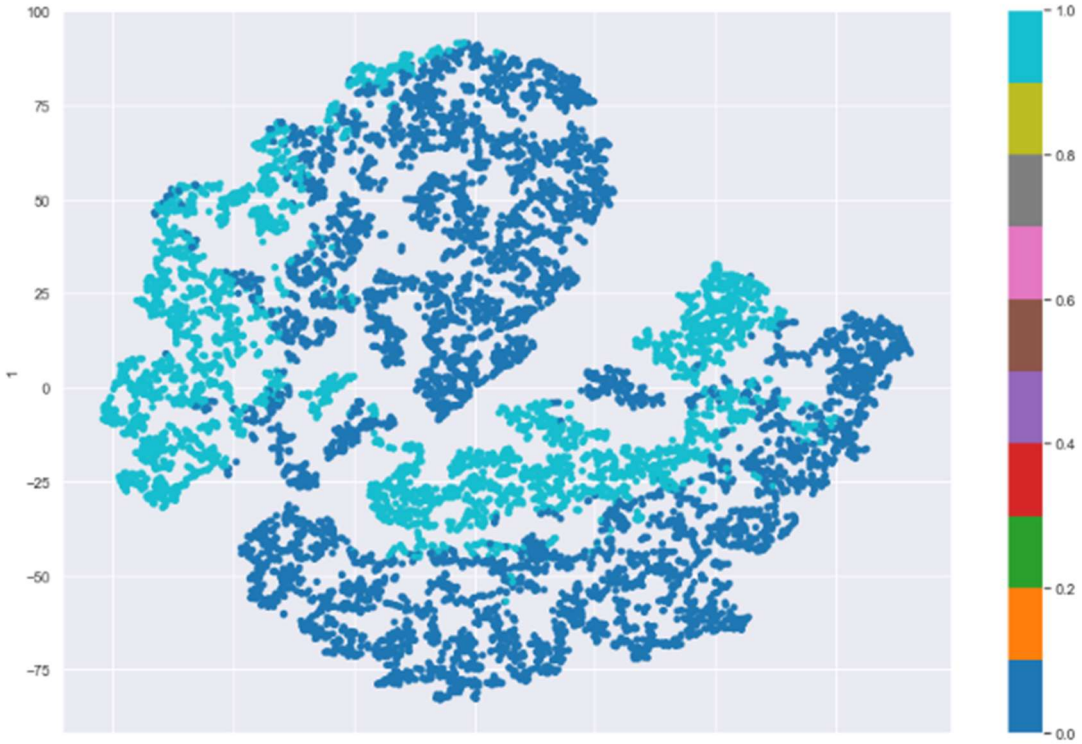


Image 49 – Cluster profile UMAP