Master Degree in Data Science and Advanced Analytics

Major in Data Science

Customer Segmentation: Insurance Company

Data Mining

Teaching Staff:

Fernando Lucas Bação

João Fonseca

David Silva

Group AF:

Filipa Alves, m20210662

Helena Oliveira, r20181121

J. Daniel Conde, m20210656

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

January 2022

INDEX

1. Business Understanding

This project aims to cluster together the costumers of an insurance company. Often, companies will have great amounts of data on their costumers and no way to describe them. Clustering analysis aims to group together costumers that have characteristics in common in such a way that firms can model behavior. In doing this, it is possible to learn from the data to inform business decisions. Companies can estimate price elasticities – how much clients are willing to pay for a given product -, they can predict purchases based on external factors, make smarter use of digital marketing tools and get a clearer picture of who is buying their products.

The latter is of paramount importance. In knowing who the clients are, the firm can know what they buy. They can know how much of it they buy and combine that information with whatever social data they have, thus building a model of their various client profiles. If a given service is being bought mostly by a given type of customer – that is, a cluster – the firm will not need to waste resources by trying to sell it to uninterested targets.

The main objective of the current analysis is exactly that. To know the client, what they want, and how the firm can sell it to them. In sum, this will be achieved by clustering the clients into groups and creating marketing campaigns that specifically target their profile. With the use of historical data on the clients, we hope to create clusters that reflect the overall insurance-buying population.

2. Data Understanding

In this project, the chosen approach was CRISP-DM. Following the Business Understanding, the next step is to proceed to Data Understanding. In this phase, all the features and their correspondent values within the data frame created are analyzed. The starting point was to look at central tendency measures. These can be found in **Table A2.1** in Appendix. Features such as '*FirstYearPol*', '*PremMotor*' or '*PremLife*' appear to have outliers, since their maximum value is disproportionately large when compared to the mean and the 3rd quartile.

The data had 389 missing values. These included *numpy* nan values, but also other ways of displaying a lack of information such as an empty space (" ") and an underscore ("_"), among others . Each record became represented by the Customer ID. This value is unique to each customer and can be used to identify them. Duplicated rows were also checked and eliminated (3 rows in total).

Upon inspection of the contents of each variable,*'EducDeg'* was split into '*EducDeg_cat*' and *'EducDeg_ord'*. The former contains the non-numeric portion of the information (Basic, High School, BSc/MSc and PhD), while the latter represented the numeric scale (1 – 4). Lastly, types of feature fields were then changed to more appropriate type objects. Inspection of modes in the categorical variables revealed that the vast majority of the clients have Children; the most frequent educational level is Bachelor/Masters, and the least frequent one is PhD. The most inhabited area is the 4th one, although no further information was provided to analyze its meaning.

Lastly, correlation analysis showed that – surprisingly – '*PremMotor*' is negatively correlated with almost every feature. The same happens for other variables as well, namely '*CustMonVal*' and '*ClaimsRate*'.

3. Data Preparation

Data preparation is a very important step in every clustering project. In fact, the quality of a model is highly constrained by the quality of the data used.

3.1. Coherence Check

Before analysing the data, it necessary to check if the records make sense in the given context. The cases presented in **Table A3.1** in the Appendix were checked, to see if there were clients with inconsistent information. In every case described, null values were kept, as that was a part of the missing values treatment to be done later.

When analysing coherence rule 2, it was concluded that 1997 records violated this condition, which represents 20% of data. Deleting nearly 2000 records was not a possibility, so this inconsistency needed further attention. '*FirstPolYear*' is most likely inserted by the company. On the other hand, 'BirthYear' is likely inserted by the clients, and is therefore more fallible. Assuming that the records were mistakes in data entry, the *'BirthYear'* column was deleted. As these records were erroneous, there was no telling if the rest of the data in the columns was accurate. After that, the variables '*EducDeg_Cat*' and '*EducDeg*' were dropped as they were redundant and gave the exact same information as '*EducDeg_Ord*'. After applying all the coherence rules, 0.0097% of the data was deleted.

3.2. Outliers

Having the descriptive statistics performed before in mind, it is known that most of the variables have outliers. Considering that algorithms like K-Prototypes or K-means are sensitive to outliers, various outliers detection methods were used to temporarily remove outliers from the dataset. After clustering, the outliers would be reintroduced, in order to find out where the outliers would be, so that a reasonable and well-rounded marketing campaign can be applied.

3.2.1. Manual Filter

The first step consisted in the most direct approach – manual limitation. If a datapoint seems to have a behaviour too detached from the norm within a certain dimension, it is considered an outlier. Histograms and box plots of each numerical variable (**Figures A3.1** and **A3.2** in the Appendix) were created to evaluate the distribution of the variables. The points that displayed awkward behaviour (**Table A3.1**) and did not properly represent the distribution were removed.

3.2.2. IQR Method

The IQR method was tested but with the standard value (1.5 * IQR, where IQR represents the inter-quantile range), too much data was being removed. Even though different values were tested, only with much higher values than 1.5 did the results seem to be plausible. It seemed that the essence of the method was lost, so it was not used. In fact, the IQR method is best when applied to normally distributed data, so it was not the best option in this case.

3.2.3. DBSCAN

Manual filtering of outliers usually does not detect all the noise. A further and more robust analysis was applied in order to spot the observations which could not be identified as outliers just by looking

at the boxplots and histograms. The chosen algorithm was DBSCAN, a density-based unsupervised clustering algorithm that can also be used to detect outliers - points that are neither core points nor border points are considered outliers and are assigned a label of -1. It is important to note that the records with missing values were not accounted for this algorithm. Two major parameters were established, the epsilon (maximum distance between two points for one to be considered as in the neighborhood) and *min_samples* (minimum number of samples in a neighborhood for a point to be considered as a core point).

By applying *NearestNeighbors* on the dataset considering the metric features, with 20 neighbors, the distance from each point to its closest neighbor was calculated. Then by plotting the array containing the distance to the sorted closest neighbors' points and the indexes also sorted, it is possible to determine a plausible epsilon value (300), where the change is most pronounced – **Figure A3.3** in the Appendix. The *min_samples* parameter was defined to be twice the number of features considered in DBSCAN. This is a common practice in this algorithm.

48 records were identified as outliers and so were removed. At this point 2.33% of the dataset were disregarded for the clusters' identification. The manual filtration was done once again to remove visible outliers.

3.2.4. Multi-dimensional outliers

The data did not show visible multi-dimensional outliers, so no points were removed from the data (**Figure A3.4** in the Appendix).

3.3. Scale the data

In order to make all variables comparable since their scales are very different from each other, the *Min-Max* method was applied. As the algorithms that are going to be used deal with distances, this step is crucial since it avoids variables with larger scales to take more weight and importance on the cluster solution. The *Min-Max* has the advantage in comparison to the Standard scaler of being more interpretable since the minimum and maximum value of each variable is known and always the same, also the latter one should only be applied to features normally distributed, which is not the case on most of our variables. The Robust Scaler was out of the equation since it seems that the clearest outliers were removed.

This was done before the missing values treatment since *KNN* (highly sensitive to scale differences) was used for missing's imputation, so it is essential to scale the data before applying this method. The outliers' dataset was also scaled using the scaler applied to the data without outliers so that they can be reintroduced on the clusters formed at the end of the project.

3.4. Missing values

Throughout the treatment of the missing values, a careful examination of their type was taken. We concluded that all of them except the Premiums variables were *MCAR* (missing completely at random) – there is no reason for their missing, therefore their imputation with models or mode was plausible and would not bias the results.

Firstly, we checked if any record had a high number of missing values. Since the observations with the highest number of 'missing' had only 3 features with this issue, we decided not to drop them. Also,

there were no features with more than 20% of missing values, not even close, so the removal was not an option.

'PremMotor', 'PremHealth', 'PremLife' and 'PremWork' missing values (33; 42; 103 and 86 respectively) were **imputed with 0**, since as this is an information the company controls, we assumed the missing of this information implies the absence of such insurances - being *MNAR* (missing not at random). As 'GeoLivArea' only had one missing value, we considered it was not worth to apply a more complex model than the **mode**. Therefore, value 4 was imputed.

The missing values from 'FirstPolYear' and 'MonthSal' (30 and 36 respectively) were imputed using KNImputer tool from sklearn with 5 as the number of neighbours and weights as 'distance' (closer neighbours of a data point will have a greater influence on the prediction than neighbours further away). Only 'PremMotor', 'PremHousehold', 'PremWork' and 'PremLife' were considered to train the model since these were the most correlated variables (looking at the spearman correlation matrix) with 'MonthSal' ('FirstPolYear' is nearly uncorrelated with every variable)

To impute the missing values from 'EducDeg_ord' KNN was applied using KNeighborsClassifier() from sklearn, splitting the data into train and validation. The f1score was not very promising (0.61) so oversampling the minority class using SMOTE was tried, however no improvements were found. Therefore, the imputed values were only 2 and 3. The Logistic Regression was used to fill missing values on 'Children' (21), after a train_test_split to check for overfitting. The F1 score for the train and validation set was 0.81. Therefore, there is no evidence of overfitting, it seems it is predicting very well the feature Children. Both 1 and 0 were imputed.

3.5. Feature Engineering

To conduce a more precise customer segmentation, new variables were created (**Table A3.3** in Appendixes), that came from transformations of the variables in the dataset, so that they might be useful to analyse and form the clusters. These new variables are introduced both in the dataset without outliers and in the outliers' dataset with slight differences.

One hot encoding was applied to the categorical variables '*EducDeg_ord*' and '*GeoLivArea*' since they are nominal variables, and it might be more useful to deal with binary variables when interpreting clusters. PCA was not used since its main disadvantage is its difficulty on interpretation, which is crucial in order to identify the clusters formed.

3.6. Redo outliers check

At this phase the outliers were checked once again, since with the creation of new variables, some extreme observations could arise. In fact, some of the newest variables ('annual_profit', 'perc_inc_household', 'perc_inc_life', 'perc_inc_motor' and 'perc_inc_work'), outliers were found by analysing the box-plot (**Figure A3.5** in Appendixes).

In fact, after removing these new outliers, globally 2.60 % of the data were removed due to the presence of extreme values. These were added to the data frame where outliers were stored.

The scaling had to be done also after the feature engineering step since some of the newest features were not in the appropriate scale. The same was done for the outliers' dataset.

3.7. Feature Selection

Two perspectives were developed, a **social** one in which it is evaluated family aspects, salary and educational level, and for the **product perspective,** 'log_household', 'PremHealth', 'log_life', 'log_work', 'annual_profit' were chosen since by analyzing the spearman correlation matrix (**Figure A3.6** in the Appendix), these variables are not very correlated among themselves but still are also not independent. The latter perspective is due to evaluate consumer habits amongst the clients of the insurance company.

By applying the chi-squared test, it was noticed that 'EducOrd_Deg' and 'Children' were not much correlated so they could cooperate on the same perspective (**Figure A3.7** in Appendixes). By analyzing a pairwise relationship (**Figure A3.8** in Appendixes) between the metric features against the binary feature 'Children', 'MonthSal' seems to be slightly related with this dummy variable. It seems that as the monthly salary increases, there is a more predominant proportion of people with children.

4. Modeling

4.1. K-Prototypes – Social Perspective

The K-prototypes algorithm was used to compute the clusters from the Personal Perspective. This can handle categorical and numeric data, being an improvement of both k-means and k-modes in that sense.

In order to select the k initial prototypes, there are two methods available: 'Huang' and 'Cao'. The first one will select the first k distinct records from the data set as initial 'centroids' and then assign the most frequent categories equally to the initial k-modes. The 'Cao' approach selects prototypes for each data object based on the density of the data point and the dissimilarity value. We choose to apply Huang although the two methods performed very similar results.

"The similarity measure on numeric attributes is the square Euclidean distance whereas the similarity measure on categorical attributes is the number of mismatches between objects and cluster prototypes. There is also a parameter 'weight' γ [1[1]] that is introduced to avoid favouring either type of attribute" [1].

We ran the algorithm from 1 to 20 clusters in order to plot each clustering cost, defined as the sum distance of all points to their respective cluster centroids. By analysing the Elbow plot (**Figure A4.1** in Appendixes), 3 clusters seem to be the appropriate number of groups to form.

By analysing the centroids of each cluster, it is possible to conclude that the first one is, generally, composed by clients with Children, a low-medium level of Education and a lower salary. Cluster 2 constituted of customers without Children, with a medium-high level of Education and high salary and cluster 3 is constituted of customers with Children, a high level of Education but a low salary.

Other Features were also considered to interpret the clusters. As expected, 'LivGeoArea' seemed not to have a great difference among the clusters. In fact, this variable was not accounted to form the

[1] If γ=0, clustering only depends on numeric attributes, i.e., locations of the objects.

clusters from the social perspective since no information about the meaning of the variable was provided.

Cluster Labels	Children	EducDeg_ord	MonthSal
1	0.895	1.951	0.354
2	0.047	2.595	0.722
3	0.966	2.926	0.434

Table 4.1: Clusters description from social perspective

The pie charts in **Figure A4.2** in Appendixes provide a better understanding of how Education level is different from cluster to cluster than by computing the average value, present on Table 4.1, since this is an ordinal variable and so the difference between 1 and 2 is not the same as 2 and 3.

3478 clients were identified as belonging to cluster 1, 2535 to cluster 2 and 4011 to cluster 3. This does make sense since it is common that the majority of insurance customers are of medium social class with children.

4.2. Hierarchical Clustering on K-Medoids – Product perspective

To cluster the product perspective, a combination of the K-Medoids and Hierarchical Clustering algorithms was used. Only metric features were used for this part, as the algorithms are based on distance. Binary features would appear too important as all variables are scaled from 0 to 1. K-Medoids was computed, using many clusters and posteriorly hierarchical clustering is performed, using the k-medoids centroids as input data points. This procedure is done with the main goal of allowing the creation of clusters with different forms, since the K-means algorithm by itself tends to form spherical clusters. Besides that, hierarchical clustering does not support large amounts of data that well and by inputting the centroids of the clusters, instead of the original points, we allow for a more efficient computation.

The first thing to do was computing the K-Medoids algorithm, which was chosen instead of K-means once it is more robust outliers. The main difference between these two algorithms is that k-medoids choses the closest point from the centroid of the cluster (medoid) as center of the cluster (being therefore more robust to outliers), while k-means uses centroids (the average point of each feature). K-means will try to minimize the sum of the squared errors, while k-medoids minimizes the sum of dissimilarities between points of a cluster and the respective medoid. The initialization method performed was k-medoids ++ in order to overcome the possibility of initializing the centroids very close to each other. This means that the seeds are more likely to be initialized far apart from each other, providing a possible better final solution.

K-Medoids was run with 1 to 15 clusters, in order to create an Elbow plot. Even though the goal of running the k-medoids wasn't using it as a final form of clustering, the elbow plot (**Figure A4.3** in Appendixes) was useful to confirm the number of clusters posteriorly chosen with hierarchical clustering. By analysing it it's possible to conclude that the optimum number of clusters is 3.

After the analysis of the elbow plot, a K-medoids clustering with 50 clusters was run, in order to use the results in Hierarchical clustering.

After calculating the centroids of the 50 K-medoids clusters, R^2 was computed for hierarchical clustering with various types of linkage and various numbers of clusters (**Figure A4.4** in Appendixes), in order to decide which of them is the most appropriate, where it was concluded that Ward's linkage was the one with the highest R^2 in almost every situation, being then the chosen linkage method.

Subsequently, Hierarchical Clustering is run, using the previously computed centroids of the 50 clusters given by the k-medoids algorithm. The dendrogram can be seen in **Figure A4.5** in Appendixes, where, based on the Euclidean distance between the clusters, it was chosen to keep 4 clusters.

By analyzing the clusters' centroids of the product perspective, it is possible to conclude that cluster 0 is the one with highest value of premiums in LOB Household and the highest annual profit. They have the second highest value in LOB Work Compensations, which means this cluster can very likely represent successful people with ages around 40. Cluster 1 is the one with higher value of premiums in LOB Health and the second with lowest values in LOB Life, which might mean that older people or sick people can be represented by it. Concerning cluster 2, this is the cluster with lowest values in every premium, so it's possible that it represents the base standard consumers. Lastly, cluster 3 has the highest values of premiums in LOB Life and LOB Work Compensation. Even though '*year_cust*' has not been used in the creation of clusters, it is known that cluster 3 is the one where the clients with the longest tenure (which means they are on the company for longer) are, so this might be the case of wealthy families around 50 years old, since '*log_household*' is the highest.

Product Labels	**log_household**	**PremHealth**	**log_life**	**log_work**	**annual_profit**	**year_cust**
0	0.640208	0.358629	0.545108	0.353136	0.443570	0.413
1	0.296229	0.499325	0.302099	0.246825	0.265611	0.512
2	0.146804	0.270305	0.128790	0.118261	0.279727	0.498
3	0.311641	0.399991	0.701849	0.450682	0.223594	0.522

Table 4.2: Product's Perspective Final Centroids

784 clients were identified as belonging to cluster 0, 4302 to cluster 1, 4363 to cluster 2 and 575 to cluster 3. A further visual analysis of the cluster is provided in **Figure A4.6** in Appendixes.

The SOM algorithm was also attempted for the products' perspective, however since the clusters formed did not have meaningful interpretations, the results from the k-medoids and hierarchical were kept.

4.3. Merging Perspectives

In the last stage of the analysis, a semi-final combined clustering solution of 12 clusters (3 clusters from the personal perspective * 4 clusters of the products perspective) was obtained. However, by analyzing **Table A4.1** and **Figure A4.7** in Appendix, it was clear that some clusters had few individuals, which means that probably there are clusters with very similar characteristics in such a way that there are no advantages in creating 12 different marketing campaigns. So, it was decided to progressively merge the clusters with fewer observations to bigger clusters, since groups with few individuals aren't so meaningful for the marketing campaigns. To know to which big cluster the small one should be merged,

the pairwise distance was used, so that when joining the clusters, the criteria being used is the proximity from each of them to each other. These distances were only based on the metric features from the two perspectives (*EducOrd_Deg* and *Children* would bias the result).

The first clusters to merge were all that had under 200 observations, so {(0,2), (0,3), (3,2), (3,3)}. After that, the clusters with less than 300 observations – {(0,3), (3,3)} – were merged. Lastly, by analyzing the centroids of the remaining clusters, it was decided to merge 2 more – {(1,3), (2,1)} -, since their centroids were very similar. The final clusters' size can be seen in **Figure A4.8** and **Table A4.2** in Appendix.

At this point, it's considered that every cluster has a relevant size to considerate doing a marketing campaign for each one of them. The final centroids are:

Product Labels	**log_household**	**PremHealth**	**log_life**	**log_work**	**annual_ profit**	**Children**	**EducDeg_ord**	**MonthSal**
1	0.640	0.359	0.545	0.353	0.444	0.639	1.818	0.345
4	0.304	0.4780	0.307	0.255	0.267	0.927	2.361	0.368
5	0.283	0.534	0.293	0.232	0.264	0.025	2.513	0.732
8	0.158	0.323	0.143	0.128	0.279	0.102	2.984	0.700
9	0.145	0.2596	0.126	0.116	0.280	0.981	2.766	0.460
10	0.312	0.400	0.702	0.451	0.224	0.645	1.837	0.356

Table 4.3: Final cluster centroids

It was considered that every cluster had different enough centroids to be considered for a marketing campaign. **Table A4.3** in the Appendix is a summary of the interpretation of each final group of clients. The figures **A4.9**, **A4.10** and **A4.11** in the Appendix, helped on the interpretation of the clusters.

4.4. Cluster Profiling – Marketing Approaches

The cluster analysis provides a good picture of the client base, their habits and characteristics. It is therefore easier to create ways to approach clients with similar characteristics in the marketplace.

If the firm wants to expand on Household products – the cluster with the highest profit margin – they should invest in nuclear family marketing - **Cluster 1**. If marketing is guided towards selling the idea of family safety and security at home, it will draw in people with similar values; customers that behave similarly. The marketing could be done online or through most TV channels. The firm could launch a promotion where for a few months you could get extra coverage for the same price, and it would increase at half the duration of the contract (eg: 6 months fire coverage at a cheaper rate – Safe House campaign).

To expand on Health products, the best direction would be the sector of the older demographic with high incomes and education levels - **Cluster 5**. The best route to expand on health insurance is to update the coverage to include more treatments and exams. This would also slightly increase the monthly payments. TV and radio advertisements would be ideal for this cluster since it seems it is composed of older people. A campaign to attract new customers, could revolve around a free check-up appointment at a partner clinic.

Cluster 4 could be approached using the same methods as cluster 5, the difference being the lower package coverage, since the former has lower income.

The low education and low salary cluster - **Cluster 10** - is the one with the highest purchases of life insurance and work compensation insurance. These are usually lower-income people, who might work in dangerous or unstable occupations. The best way to market these products would either be directly through the employer, or to offer a discount of 10% on the bundle package of Life-Work insurance, by the typical TV routes-since lower education. Once this cluster is the best client for both, they could be offered a package deal with pre-selected terms and coverage.

Cluster 8 is made up of high income, high education people that yield the firm average profits. They spend an average amount in every aspect, indicating they might not have many specific needs. New clients could be motivated by targeted ads for young professionals with higher education, as a hassle-free product that keeps them covered. One way to incentivize consumption in this cluster would be to offer the client a 20% discount on the bundling of a second product, upon purchase of a first one.

The same applied to **Cluster 9**. The people in this cluster are similar but vary in two aspects. Their salaries are lower – so the targeted packages would be cheaper – and they have children – family packages on health insurance could be advantageous. The best route to increase this cluster is through targeted ads with an emphasis on Household products (the most profitable package) or through employers. The firm can also give a discount for the inclusion of children in health packages, such as a lower price for the first months if the contract exceeds X months.

4.5. Multidimensionality visualization

In order to have a clearer and visual assessment of the clusters formed, TSNE and UMAP (2 and 3 dimensional) were developed - Figures **A4.12**, **A4.13** and **A4.14** in the Appendix. These techniques are tools to visualize high-dimensional data so that it is possible to better understand if the clusters formed seem visually well separated. As proposed in reference [2], a PCA was conducted in order to reduce firstly the number of dimensions a bit, so that TSNE visualization can avoid some noise.

4.6. Outliers reintroduced

The scaled outliers were then reintroduced so that unusual clients have also a marketing campaign.

For that, a Decision Tree Classifier was trained in the data without outliers used to form the clusters. The KNN algorithm may have also been used, however, as clusters were formed also based on categorical variables, a decision tree is more suitable since KNN must only be used with metric features. A train test split was executed in order to control overfitting and access the quality of the model. The target in this model was the '*merged_labels'* so that the predictions obtained when applying the model to the outliers' dataset were the labels of the clusters. The f1-score on the training and validation dataset was remarkably similar, 0.82 so that it is possible to infer the quality of the model developed to label the outliers on the cluster labels. 97 customers were assigned to cluster 1, 65 to cluster 10, 54 to cluster 5, 22 to cluster 4, 8 to cluster 9 and only 1 client to cluster 8.

5. Conclusion

Marketing relies heavily on knowing which products are being sold to whom. The purpose is to split the customers in such a way that they can clearly be told apart from one another, and along meaningful dimensions.

Having split the clients into 6 clusters, the graphical analysis shows that these clusters are well separated along clear lines. This appears to be concise proof that the clustering was effective in its purpose. The client database was separated using factors such as income, consumption, education and children. We believe these factors are the most important ones to separate along.

Income is crucial from a marketing standpoint, as it allows the pricing expert to decide which products to sell. In the current scenario, it allows product managers to settle on the coverage to apply to each cluster in a way that meets their needs and increases profits.

However, insurance companies are split into different areas. There are a lot of product managers, who need to decide where they can position themselves in different income levels. That is why the clustering also considers consumer purchasing habits. Knowledge of these two dimensions lets each product manager know which consumers are typically interested in the product they are selling, and their approximate income.

Lastly, the decision of what marketing campaign to implement is informed by two other factors of importance. When selling something it is important to adjust our speech. Both in vocabulary and in message. Both in platform and frequency. When selling to parents – clusters with more children – salespeople tend to pull at the heart strings. It is easier to sell products such as insurance if one can call invoke the pathos. The sales strategy is informed by this fact. Additionally, the level of education matters in sales. Especially in online sales. The vehicle of information for knowledge workers and more educated people is typically more steered towards targeted ads and technological real estate.

In sum, the clustering was made along three dimensions. Two of them were built to inform the Product and the Pricing. The last one was for the Placement. Chapter 4.4 settles the Promotion.

In sum, the clustering was made along three dimensions. Two of them were built to inform the **Product** and the **Pricing**. The last one was for the **Place**. Chapter 4.4 settles the **Promotion**.[2]

[2] 4 P's of Marketing

6. References

[1] Huang, Z. (1997, February). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st pacific-asia conference on knowledge discovery and data mining,(PAKDD)* (pp. 21-34).

[2] Website scikit-learn.org, article 'Sklearn Manifold TSNE', visited on 05th of January 2022 https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

[3] Website www.analyticsvidhya.com, article 'How Dbscan Clustering Works', visited on 19th of December 2022

[4] Website towardsdatascience.com, article 'Cluster Analysis Create Visualize and Interpret Customer Segments', visited on 31st of December of 2022

[5] Website zachary-a-zazueta.medium.com, article 'K-prototypes Clustering for When You're Clustering Continuous and Categorical Data', visited on 9th of December 2021

[6] Website www.math.le.ac.uk, article 'K-means and K-medoids', visited on 2nd of January 2022

[7] Website investopedia.com, 'The 4 Ps'

Practical Classes' GitHub repository:
https://github.com/helenado/Data-Mining-21-22.git

Project's GitHub repository:
https://github.com/helenado/DataMining_project_master.git

7. Appendix

7.1. Tables

7.1.1. Data Understanding

	count	mean	std	min	25%	50%	75%	max
FirstPolYear	10266	1991.062634	511.267913	1974	1980	1986	1992	53784
BirthYear	10279	1968.007783	19.709476	1028	1953	1968	1983	2001
MonthSal	10260	2506.667057	1157.449634	333	1706	2501.5	3290.25	55215
GeoLivArea	10295	2.709859	1.266291	1	1	3	4	4
Children	10275	0.706764	0.455268	0	0	1	1	1
CustMonVal	10296	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296	0.742772	2.916964	0	0.39	0.72	0.98	256.2
PremMotor	10262	300.470252	211.914997	-4.11	190.59	298.61	408.3	11604.42
PremHou-sehold	10296	210.431192	352.595984	-75	49.45	132.8	290.05	25048.8
PremHealth	10253	171.580833	296.405976	-2.11	111.8	162.81	219.82	28272
PremLife	10192	41.855782	47.480632	-7	9.89	25.56	57.79	398.3
PremWork	10210	41.277514	51.513572	-12	10.67	25.67	56.79	1988.7

Table A2.1: Central tendency measures, min and max Table

7.1.2. Data Preparation

Coherence Number	Rule	Explanation
1	*EducDeg_cat* == 'PhD' & *EducDeg_ord* ==4 *EducDeg_cat* == 'Basic' & *EducDeg_ord* == 1 *EducDeg_cat* == 'High School' & *EducDeg_ord* == 2 *EducDeg_cat* == 'BSc/MSc' & *EducDeg_ord* == 3	Make sure that all the records with a kind of education in column *EducDeg_cat* have the supposed ordinal category (in *EducDeg_ord* column)
2	*FirstPolYear* < *BirthYear*	People can't be clients before they were born
3	*FirstPolYear* >= 2016	As the year of the database is 2016, the first contract year should be before 2016.
4	*FirstPolYear* <= 2016 - 120	If a person is client since 1896, she probably won't be alive by now

Table A7.1: Coherence Check Rules

MonthSal > 10000	*ClaimsRate* > 4	*PremHealth* > 800
CustMonVal < -1250	*PremMotor* > 2000	*PremLife* > 200
CustMonVal > 1500	*PremHousehold* > 1600	*PremWork* > 400

Table A7.2: Manual Filter Rules

Name	Condition	Explanation
year_cust	2016 - FirstPolYear[3]	Number of years since the first policy year
Total_Premiums	PremMotor + PremHousehold + PremHealth + PremLife + PremWork	Sum of the positive premiums
annual_profit	CustMonVal / year_cust	Annual profit of the premiums
log_custMon	log(CustMonVal + 1)[4]	Logarithm of Customer Mon. Value
log_household	log(PremHousehold + 1)	Logarithm of household premiums
log_life	log(PremLife + 1)	Logarithm of life premiums
log_work	log(PremWork + 1)	Logarithm of work compensation premiums
Annual_claims_rate	ClaimsRate / 2	Annual Claims Rate
AnnualSal	AnnualSal * 12	Annual Salary
perc_inc_health	If (AnnualSal = 0) then 0 Else PremHealth / AnnualSal	Percentage of the annual salary that is spent in household premiums
perc_inc_household	If (AnnualSal = 0) then 0 Else PremHousehold / AnnualSal	Percentage of the annual salary that is spent in household premiums
perc_inc_life	If (AnnualSal = 0) then 0 Else PremLife / AnnualSal	Percentage of the annual salary that is spent in life premiums
perc_inc_motor	If (AnnualSal = 0) then 0 Else PremMotor / AnnualSal	Percentage of the annual salary that is spent in motor premiums
perc_inc_work	If (AnnualSal = 0) then 0 Else PremWork / AnnualSal	Percentage of the annual salary that is spent in work compensation premiums

Table A7.3: Feature Engeneering

Note: the creation of the variables was done using the logarithmic transformation to reduce the skewness of the data. To create the variables with this transformation, 1 is added to all values of the variable, so that there's no infinite values, caused by the variable being equal to 0. Because the data is already scaled, all the values are between 0 and 1 (in the dataset without outliers) so there aren't problems caused by negative values. In the outliers' dataset, only one variable had negative values ('*CustMonVal*'), so a different transformation was done (subtract the minimum of the variable and sum one), in order to avoid null values.

[3] At this point, the variables are already scaled, so there's a need to scale the number 2016. As the scaling in the outliers' dataset is done with the scaler applied to the data without outliers, 2016 is scaled the same way both for the outliers' dataset and for the dataset without outliers.

[4] In the outliers' dataset, the transformation was log(CustMonVal – min(CustMonVal) + 1)

7.1.3. Modeling

Social Labels	1	2	3
Product Labels			
0	507	150	127
1	1481	1533	1288
2	1126	735	2502
3	364	117	94

Table A7.1: Number of records in each initial cluster

Social Labels	1	2	3
Product Labels			
0	784	NaN	NaN
1	2769	1533	NaN
2	NaN	735	3628
3	575	NaN	NaN

Table A7.2: Number of records in each final cluster

Cluster	Cluster Name	Most bought product	Profit Level	Education	Salary Level
1	Professionals in their 40s	Household	High	Low	Low
4	Elderly mid-lower class	Na	Medium	Medium	Low
5	Wealthy educated elderly	Health	Medium	Medium – High	High
8	Average Consumption, high-earning and childless	Na	Medium	High	High
9	Average Consumption, educated with children	Na	Medium	High	Medium
10	Low-education life and health, less profit	Life, Work	Low	Low	Low

Table A7.3: Centroids of the final clustering solution

7.2. Figures

7.2.1. Data Understanding

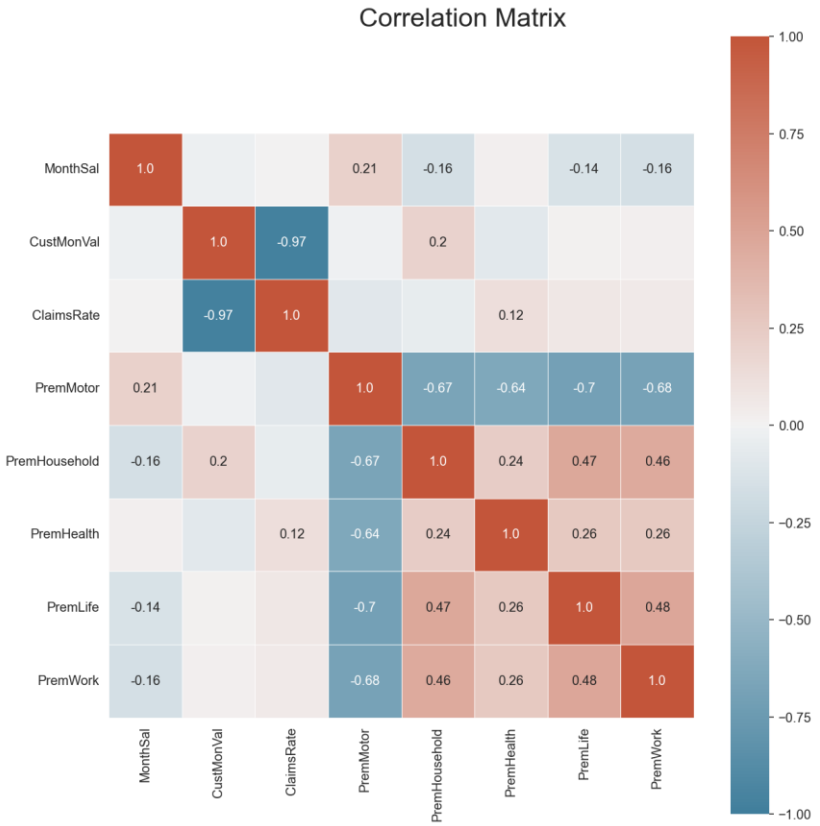

Figure A2.1: Initial correlation matrix

7.2.2. Data Preparation

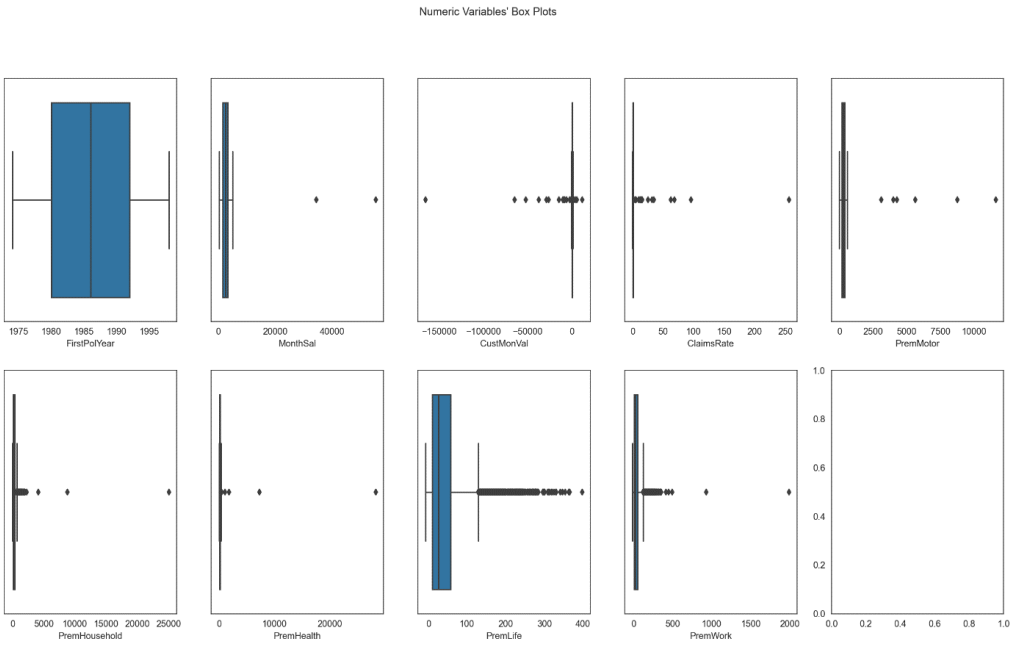

Figure A3.1: Box-plots before Manual Filtering

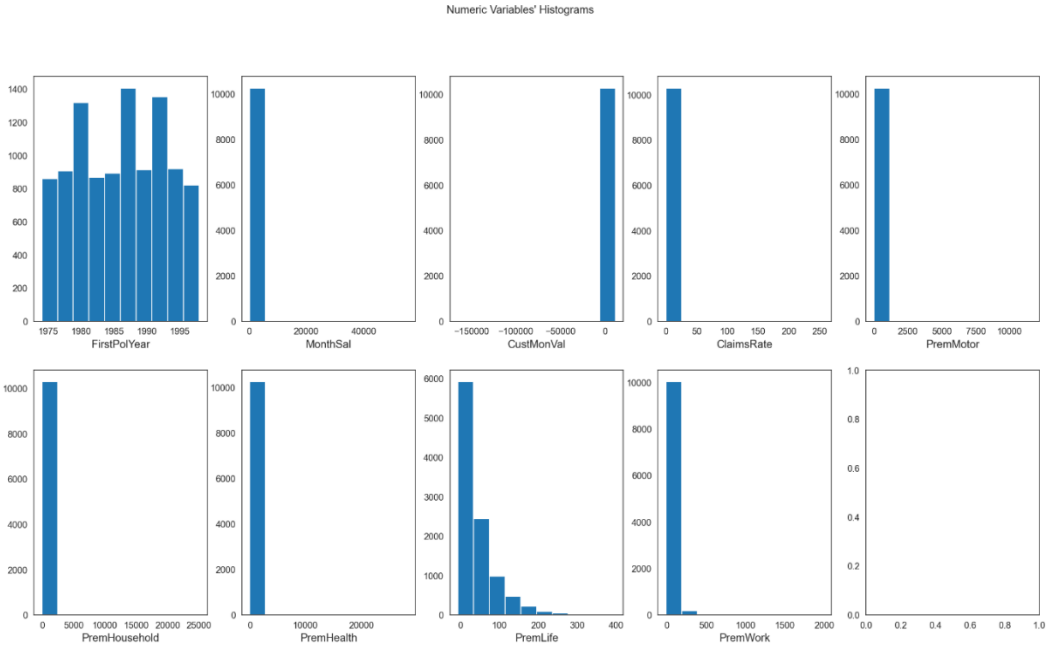

Figure A3.2: Histograms before Manual Filtering

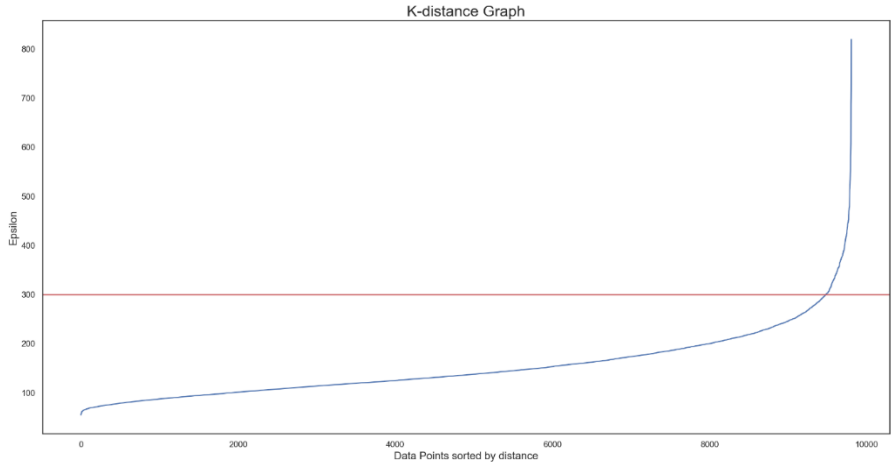

Figure A3.3: Optimal epsilon value – DBSCAN

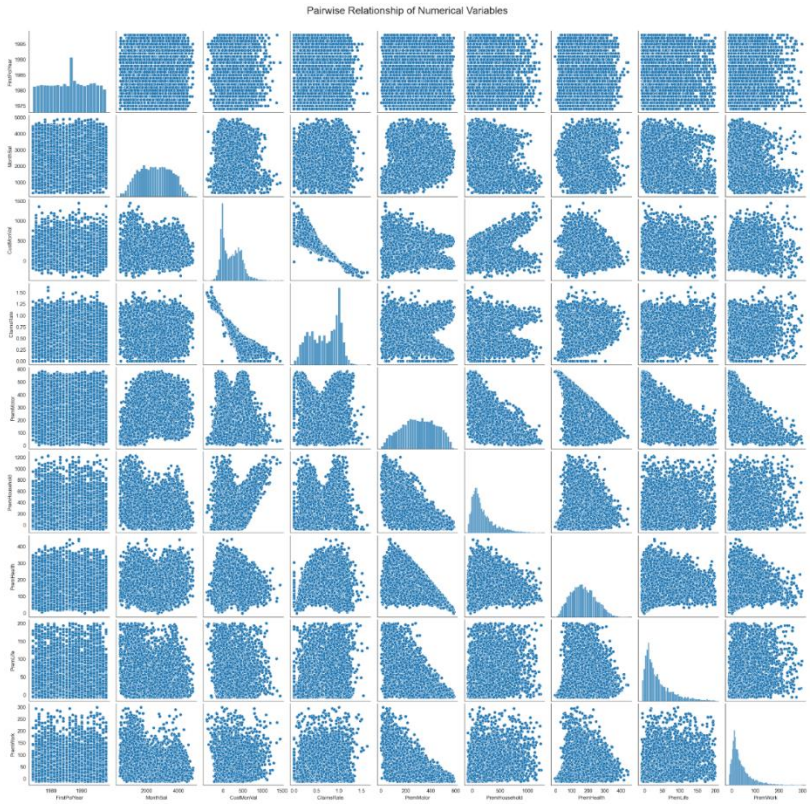

Figure A3.4: Pairwise relationship – multi-dimensional outliers

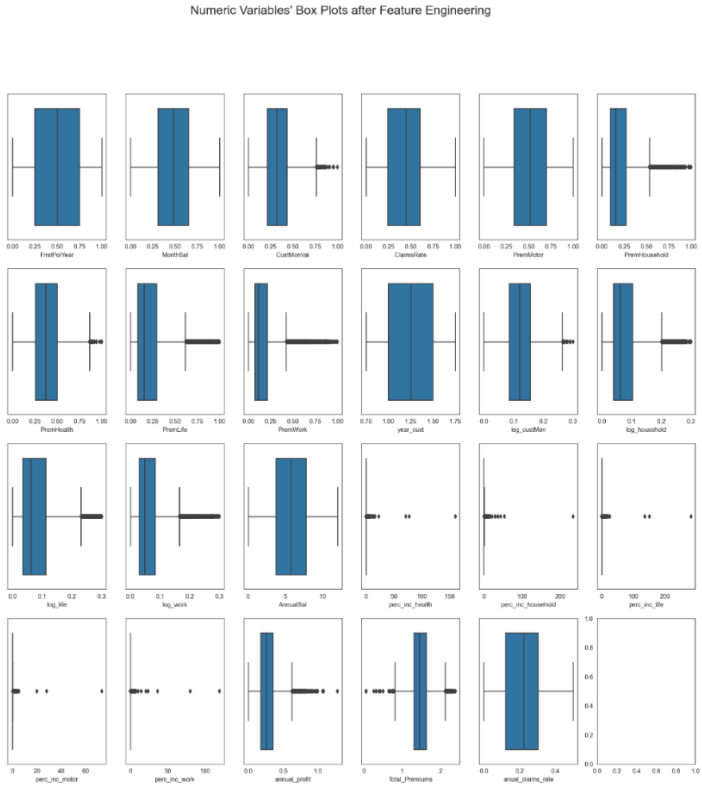

Figure A3.5: Redo outliers check – Box-Plot

(*annual_profit* <=1; *perc_inc_household* <= 100; *perc_inc_life*<=100; *perc_inc_motor* <= 40; *perc_inc_work* <= 50)

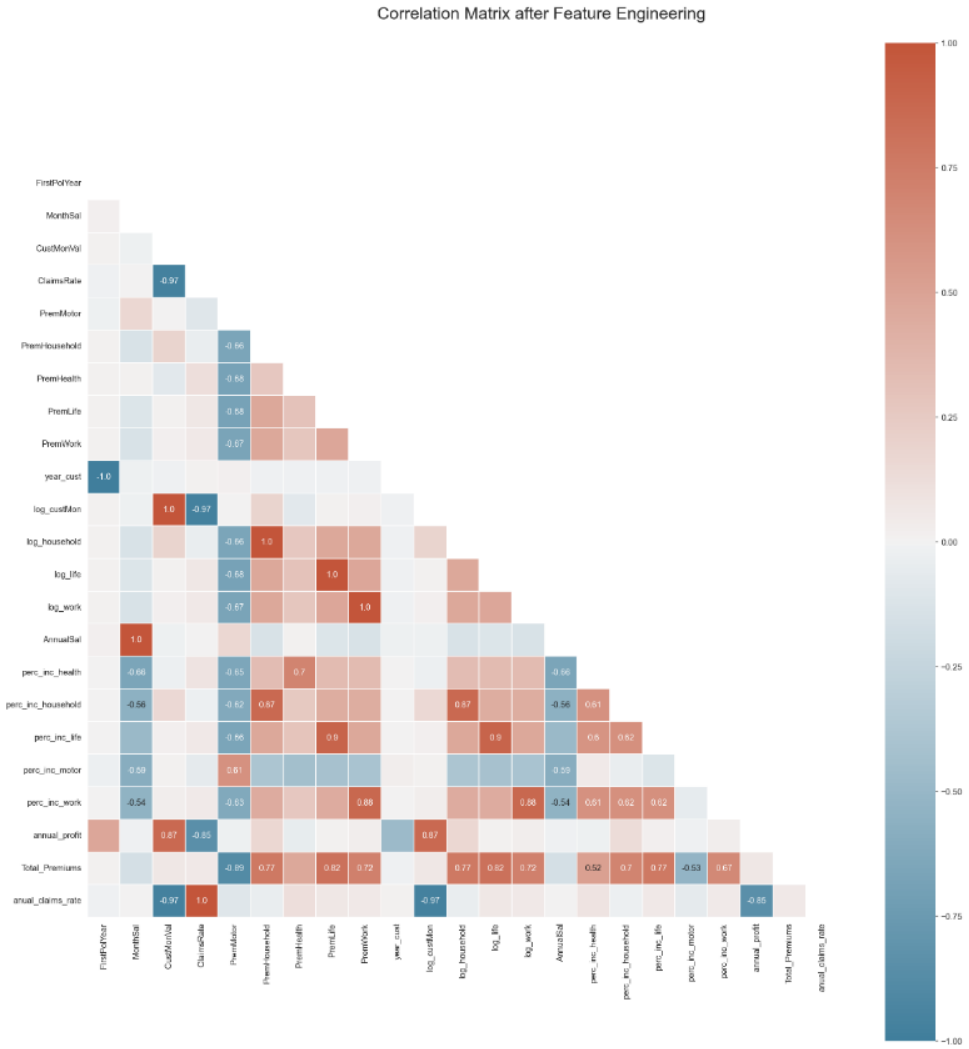

Figure A3.6: Spearman's correlation Matrix after feature engineering

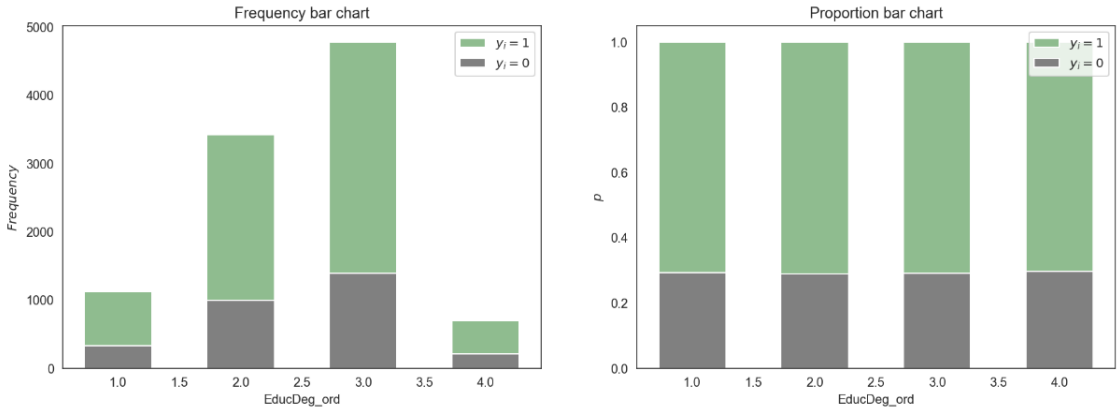

Figure A3.7: Bar plot Education vs Children

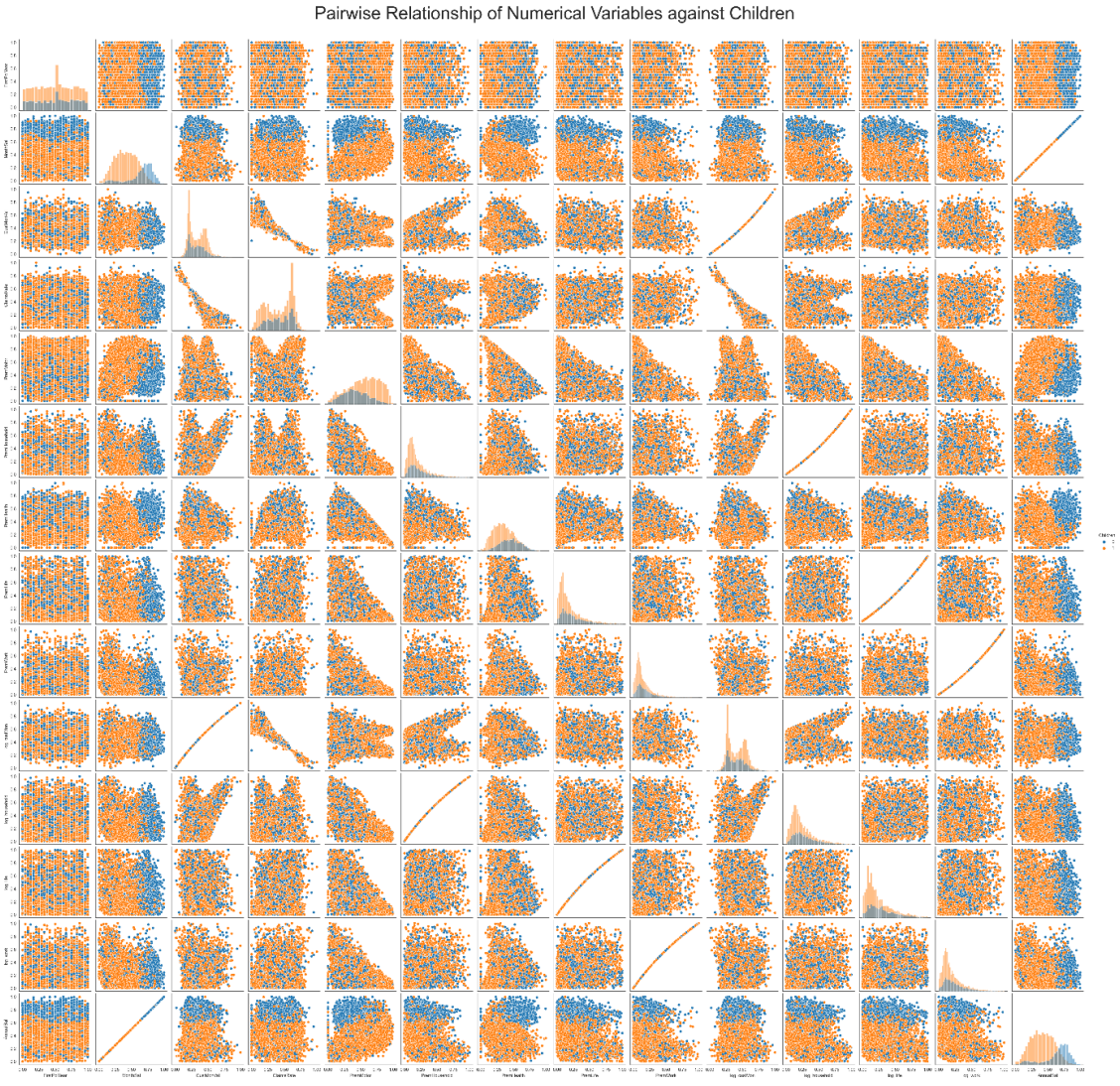

Figure A3.8: Pairwise plot – Metric features vs Children

7.2.3. Modeling

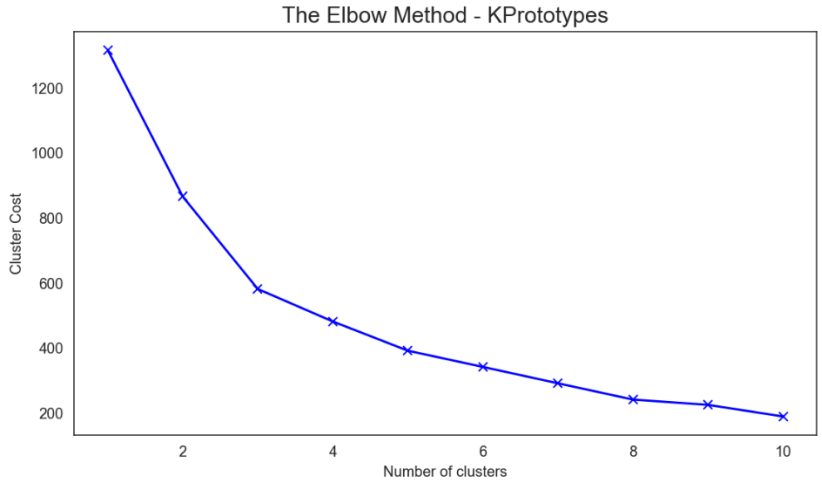

Figure A4.1: Elbow plot for the social perspective

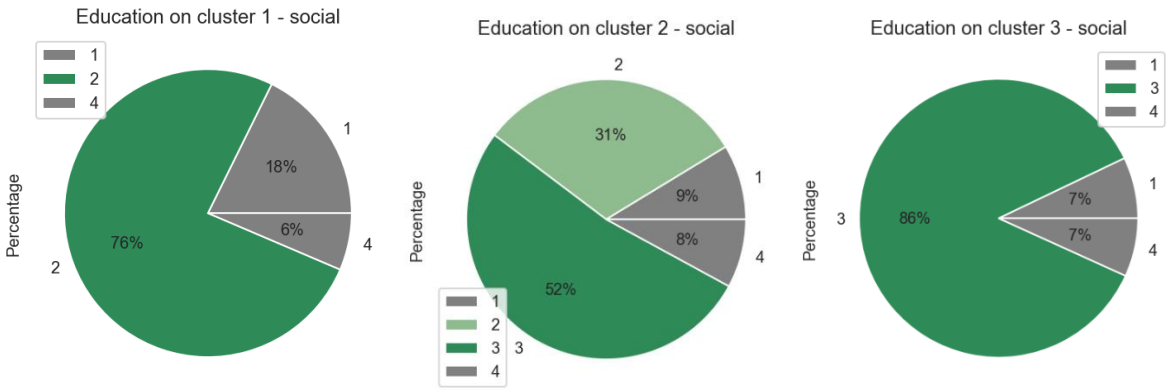

Figure A4.2: Pie Charts for each cluster on Education Level.

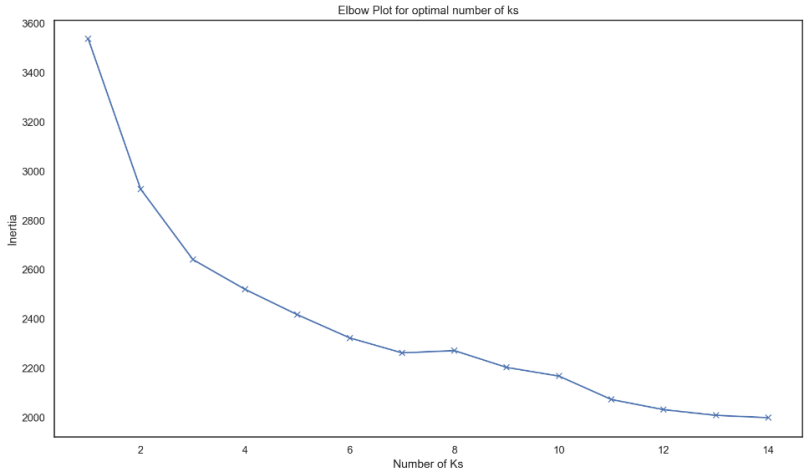

Figure A7.3: Elbow Method for K-Means

Figure A4.4: R^2 plot for different linkage criterion.

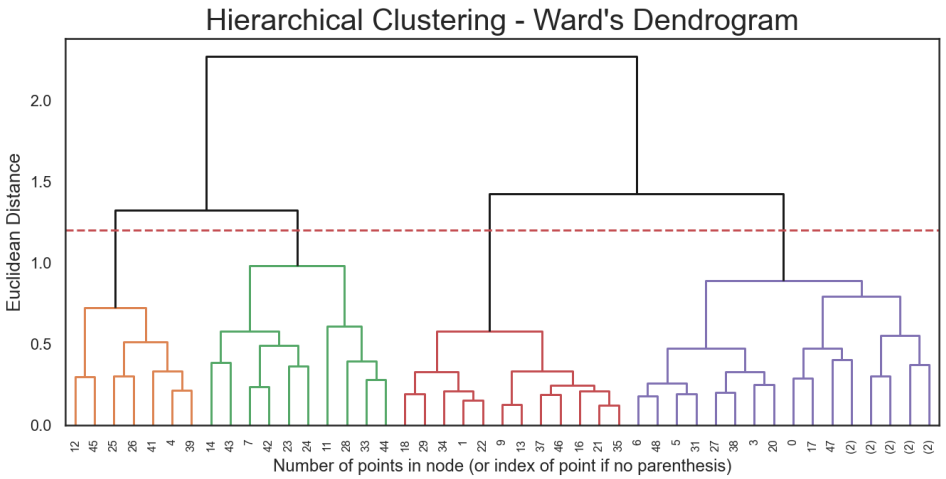

Figure A4.5: Dendrogram

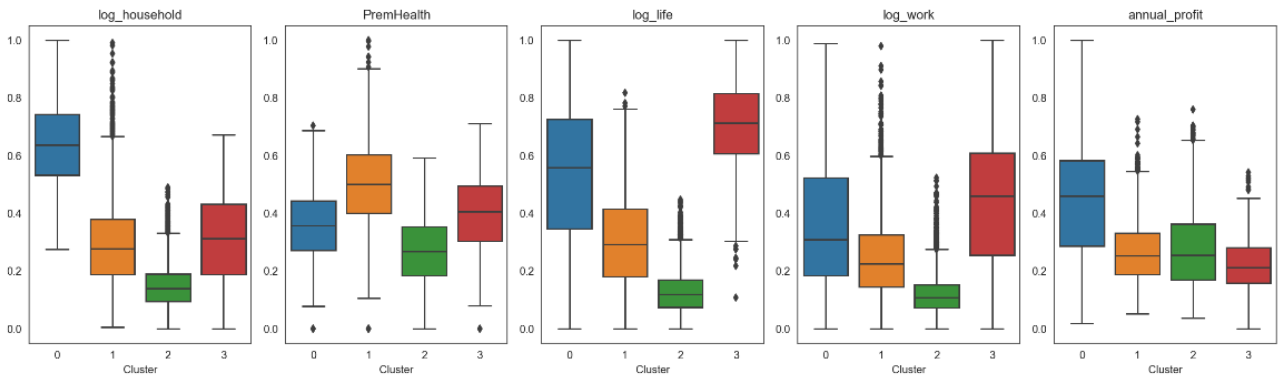

Figure A4.6: Box – plots for the variables in each cluster – Product Perspective

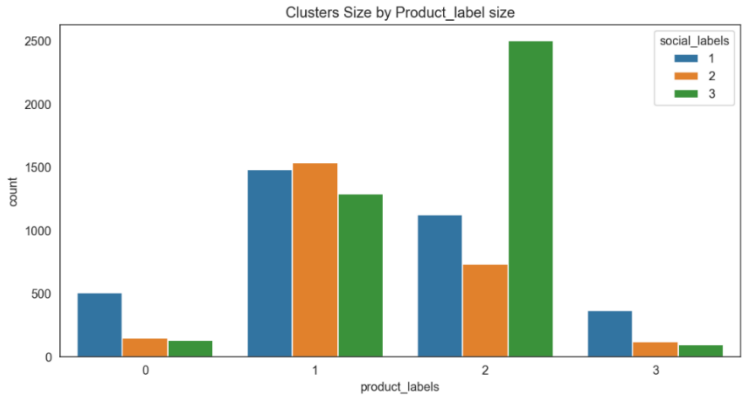

Figure A4.7: Initial clusters' size by Product

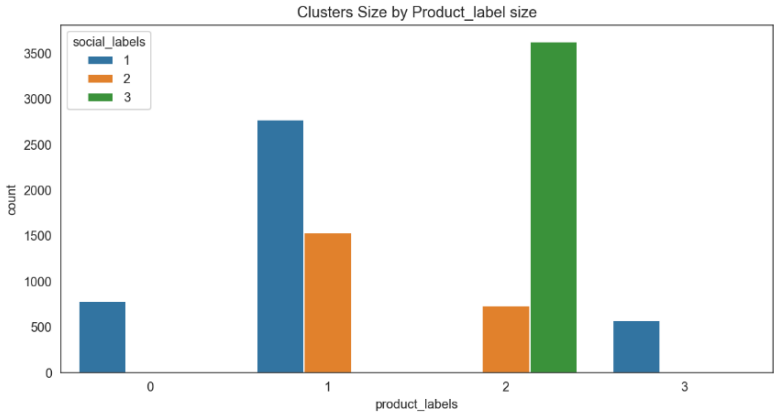

Figure A4.8: Final clusters' size by Product

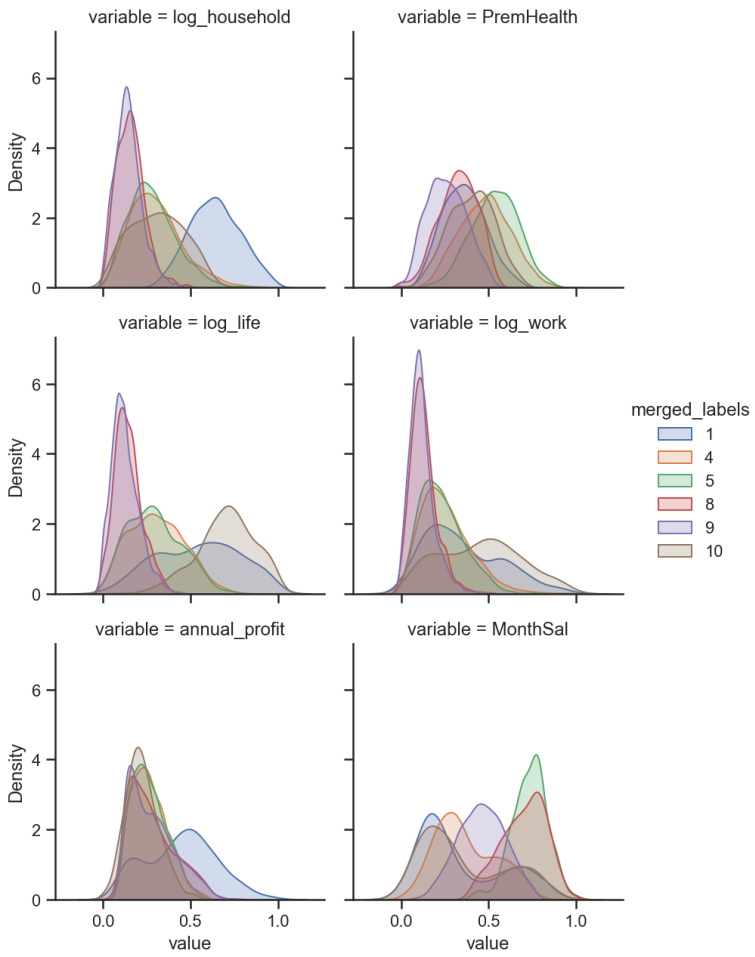

Figure **A4.9**: Cluster Analysis: Feature Distribution

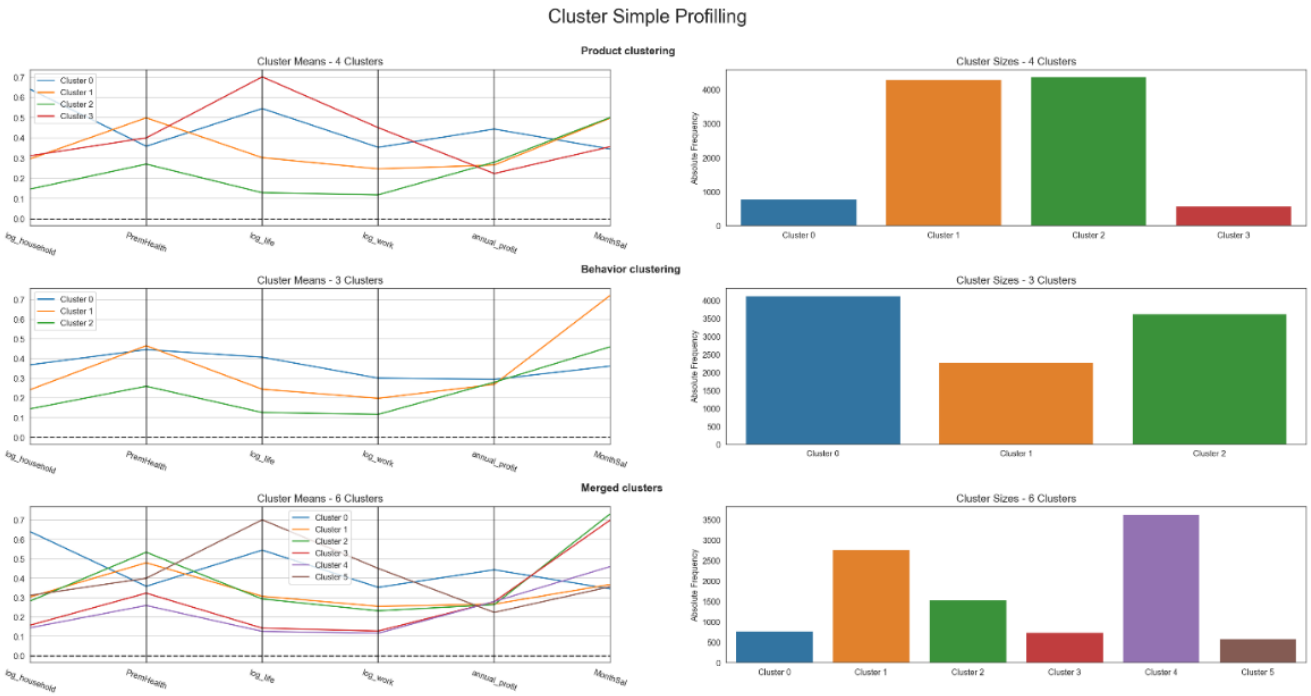

Figure **A4.10**: Cluster Analysis – Cluster Profiling

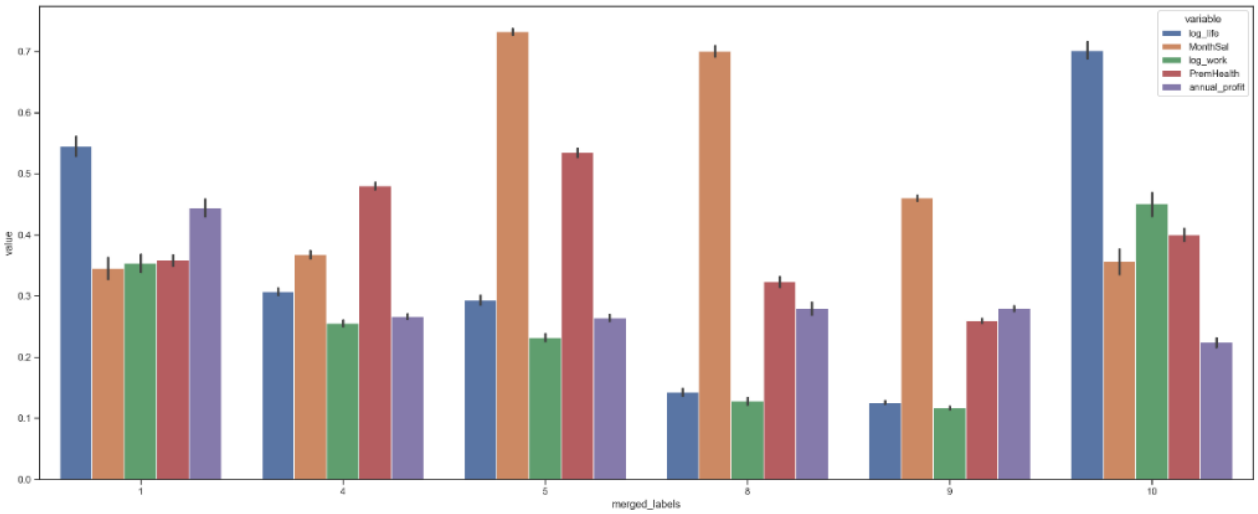

Figure **A4.11**: Cluster Analysis: Bar plot

Note: The bar height represents the average value of the feature. Thus, since the '*log_life*' corresponding bar differs greatly from cluster to cluster, it is possible to conclude that this feature is very important in the creation of the cluster.

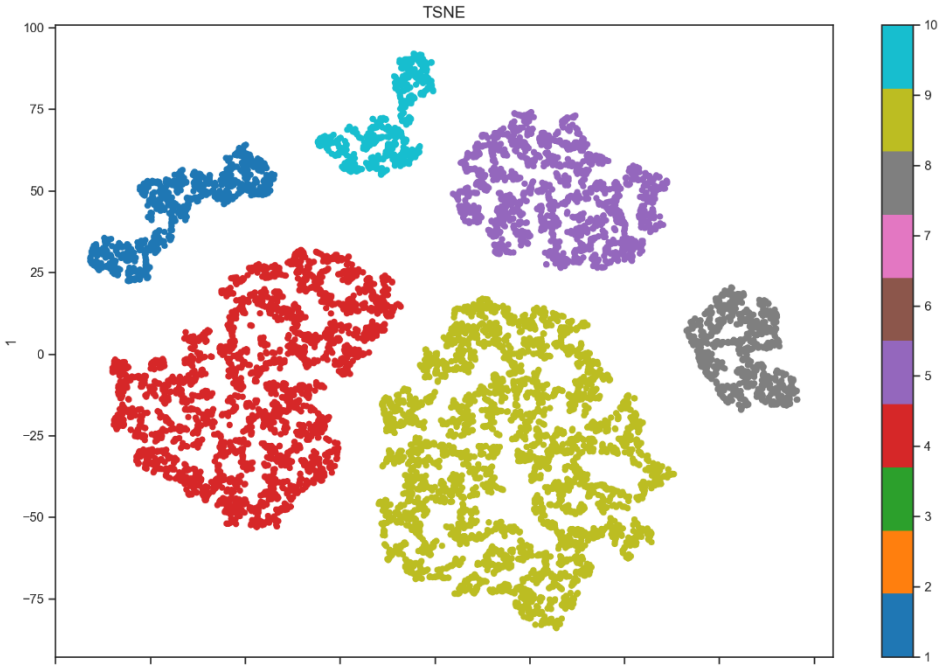

Figure A4.12: TSNE

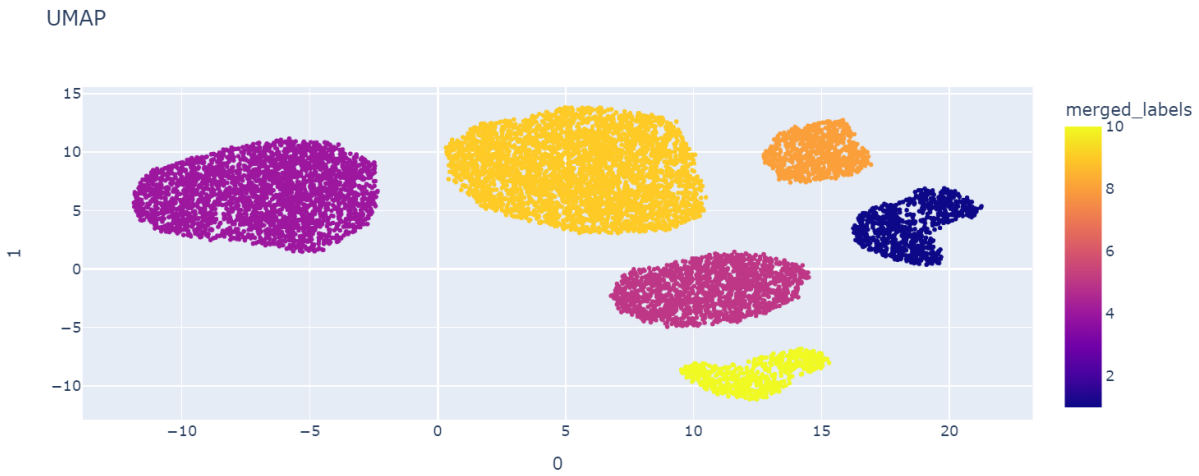

Figure **A4.13**: UMAP 2D

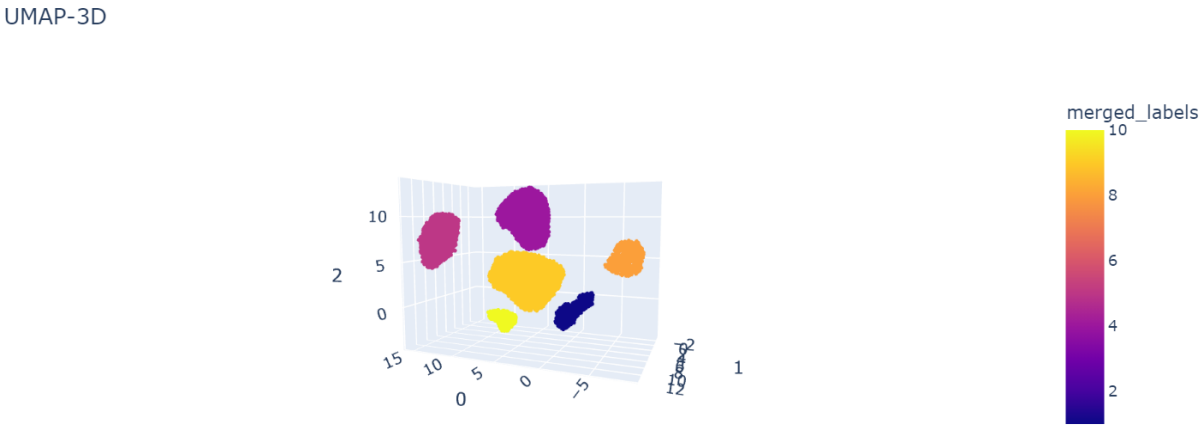

Figure **A4.14**: UMAP 3d