NOVA IMS Information Management School

Data Mining Project

MASTER'S DEGREE PROGRAM IN

DATA SCIENCE AND ADVANCED ANALYTICS

CUSTOMER INSURANCE SEGMENTATION

Group AP

Fernandes Hugo, number: 20210682

Figueira Martim, number: 20210686

Parenti Alberto, number: 20211304

01/2022

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

Contents

1 Introduction

The following project aims to develop a Customer Segmentation for a fictional insurance company and to provide a clear and proven profiling of the 10296 customers in the data set, in order to be delivered to the Marketing Department for further development. First of all, we started by exploring the data set, getting a better grasp of its behaviour and possible incoherences. Then, it continued by pre-processing the given data: finding and removing the outliers and filling the missing values. After these steps, we looked over all variables and we discussed the goal of our clustering, defining which segmentations would make sense to do. The first segmentation clusters were achieved by using the combination of the K-means algorithm and the hierarchical clustering. The clusters from the second segmentation were achieved by using the K-Prototypes. At the end, we got the final clusters crossing the results from both segmentations. Finally, we elaborated succinct marketing approaches hinged on our final clusters.

2 Exploratory analysis

2.1 Original Dataset Exploration

Starting by doing some exploratory analysis to our data set, we checked for its size, number of features, and we did some descriptive analysis for both numeric and categorical features. At this point, we were able to see that there are some missing values in almost all features and there are some incoherent values and possible outliers. To make this analysis more complete, we made checked the scatter plot between all pairs of features, histograms(fig 2) and boxplots(fig 3) to all the numerical features and we figuered out that there are some outliers. To conclude the initial analysis, we checked the correlation matrix(fig 4), where *CustMonVal* and *ClaimsRate* have a -0.99 correlation, so they are providing the same information. Hence, we chose not to drop this features, but in the costumer segmentation only one of them will be used.

2.2 Data Incoherence

As we previously noticed, there are some incoherent values in our data set: there is a client with a *FirstPolYear* equal to 53784 when the data set year is 2016 and a client with a *BirthYear* equal to 1028 (988 years old). These are obviously typos so we will automatically point them as outliers by applying a threshold.

3 Data Pre-processing

3.1 Outliers detection

Outliers are observation points distant from other elements, more weighted than others. Since the following clustering algorithms are sensible to outliers, it is been decided to ensemble different methods to remove them from the initial data set. Also, outliers in lower dimensions may not be considered as outliers anymore in higher dimensions. So, it is preferable to use more than one method to detect outliers. The detected outliers, after the clustering get done, will be reintroduced with the goal of finding out which clusters these points join.

3.1.1 Z-score

The first method tested was the Z-score, which represents the number of standard deviations away from the mean that a certain data point is. It has to be defined the threshold value for the Z-score to classify whereas a point is an outlier or not.[1] In this case, it is been considered as outlier an observation with the Z-score outside the interval -3 and 3. This method resulted to classify 4.30% of the elements, out of the whole data set, as outliers.

3.1.2 LOF

After, it is been tried the Local Outlier Factor, a density-based-unsupervised algorithm for outlier data handling. In summary, it works by comparing the density of a given point to its neighbours and hence determining whether that data point can be considered normal or as an outlier.[2] In the same circumstances as before, 0.90% of the entire data frame is considered an outlier.

3.1.3 Isolation Forest

The third method tried is the isolation forest. It aims to isolate observations of the dataset by selecting a random feature and then choosing a split value between the maximum and the minimum of the selected one.[3] It assumes that since outliers are more rare and less frequent than the rest of the dataset, they are easier to isolate. By opposite, the normal observations would require more partitions to be identified rather than an outlier data point.[4] Through isolation forest 1.94% of the observations is considered as an outlier.

3.1.4 DBScan

Moreover, it is been tested DBScan, a unsupervised density-based clustering algorithm used also for outliers detection. It is possible to use it for the second purpose because points that do

not belong to any cluster get their own class: -1.[5] In this dataset, the class -1 is represented by the all 9987 observations of the dataset without the records with missing values. The reason of this result is that there are no clear clusters since the data is very sparse.

3.1.5 Ensembling

Finally, the first three methods have been ensembled and an observation is been considered as an outlier in case at least two out of the three algorithms classified it like that. Hence, 26 observations have been removed.

3.1.6 Manual filtering

After having filled the missing values in the dataset, it is been computed a filtering selecting manual thresholds in order to remove the outliers in excess. The values have been selected analysing visually the histograms and box plots of the numerical variables. The following table shows the chosen thresholds:

PremHousehold	PremMotor	CustMonVal	MonthSal	FirstPolYear	BirthYear
< 2000	< 2000	> -1000	<20000	< 2016	> 1900

Table 1: thresholds for manual filtering

19 observations have been removed using manual filtering. Hence, considering the whole dataset of 10296 records, 45 values, equal to 0,44% of it, have been considered an outlier.

3.2 Dealing with Null Values

In order to avoid any possible bias from the outliers, we removed them so we could apply data imputers to fill the missing values. The data imputing changes whether the considered feature is discrete or continuous. The metric features have been treated using iterative imputation, where each feature is modeled as a function of the other features. Each variable is imputed sequentially, one after the other, allowing prior imputed values to be used as part of a model in predicting subsequent features.[6] On the other hand, for the non-metric variables: *EducDeg*, *GeoLivArea*, *Children*, the missing values have been filled training three decision trees with the data from each categorical variable and then used that framed data to obtain them.

4 Segmentation

Having all the pre-processing done, we start thinking about how we would cluster our data. After discussed which set of variables would make sense be considered together we ended up deciding that the best approach would be apply 2 segmentations to our data in order to get 2 different informations from it. The first segmentation regards the engagement/value of our costumers putting all the premium variables with *CustrMonVal* variable. This decision was based on which we believe for Marketing Department give best insights about which are the fields that the company earns more as well as the differences of money spent in fields. As we figured out in exploration analysis phase *CustMonVal* is highly correlated with *ClaimsRate*, so to avoid bringing redundancy to our clustering we kept just *CustMonVal*. This choice was supported by the fact that as it is a lifetime value would make more sense to keep it than a variable that was based on the last 2 years. The second segmentation regards the bio-demographic information of our costumers. We decided to add the variable Salary to this segmentation as it's a personal attribute.

The last step before implementing the cluster algorithms was deciding whether to standardize or not the data. Although for the first segmentation the data was all in the same currency, it did not have the same scale, given that there were negative values and really high values. For the second segmentation , the only numerical variables were in different measures. Thus, we standardized all data using standard MinMax Scaler.

5 Clustering

Regarding the clustering analysis, we implemented 2 algorithms: k-means, k-prototypes and hierarchical with the seeds from k-means, the selection of the best method will be explained in the next chapter.

5.1 Clustering for Premium Segmentation

In this segmentation we used the following features: *CustMonVal*, *ClaimsRate*, *PremMotor*, *PremHousehold*, *PremHealth*, *PremLife* and *PremWork*.
Since they are only numerical features in this segmentation and we are able to work with distance-based clustering methods. Firstly, we decided to apply k-means with k=3 and k=4, but then because of the scarce quality of our clustering we use an higher k (k=10). Then, with the final k-means centroids, we apply hierarchical algorithm in order to find the best fitting solution.

5.2 Clustering for Bio-Demographic Segmentation

On the other hand, in the bio-demographic segmentation, we used *BirthYear*, *EducDeg*, *MonthSal*, *GeoLivArea* and *Children* as variables.
As we had some numerical and categorical features, we had to apply K-Prototypes, an algorithm that handles both data types. According to the elbow plot, (fig 6) the optimal number of clusters is 3.

5.3 Merging Clusters

Having both segmentation done, the next step was to cross the segmentations' results. We might check the size of the resulting cluster in (fig 7) and conclude that all clusters seem to have a good size except clusters (0,0) and (2,0) that have less then 300 customers, which for us, doesn't seem the be a be a size big enough to keep, so we had so shift the customers from this clusters to the closest one. The customers from cluster (0,0) went to (0,2) and from (2,0) went to (2,2). We can see the updated size in (fig 8). Let's define a number to each cluster, representing this in (table 5).

6 Customers Profiling

Regarding the costumers profiling, considering both segmentations and the merged clusters, we will show a table with the summarized values, giving a quick description of each cluster.

6.1 Premium Profiles

For premium segmentation, the following table summarizes our clusters behaviour.

Premium Clusters	**Avg CustMonVal (€)**	**Avg Prem Motor (€)**	**Avg Prem Household (€)**	**Avg Prem Health (€)**	**Avg Prem Life (€)**	**Avg Prem Work (€)**
1	353	98	595	156	113	103
2	211	419	83	120	17	17
3	179	219	220	230	47	48

Table 2: Premium Clusters Profiles

- 1: Costumers with a high costumer monetary value are also the ones that spend the most in Household premiums;

- 2: Costumers that spend the most in Motor premiums;

- 3: Costumers with lower costumer monetary value spend more in Motor, Household and Health.

6.2 Bio-demo Profiles

For bio-demo segmentation the next table summarizes our clusters bio-demo behavior.

Bio-Demo Clusters	**Avg Salary (€)**	**Avg Age**	**Have Children?**	**MostCom EducDeg**	**MostCom LivArea**
1	3615	68	No	BSc/MSc	4
2	1627	32	Yes	High School	4
3	2460	47	Yes	BSc/MSc	1

Table 3: Bio-Demo Clusters Profiles

- 1: Older costumers with a higher monthly salary, higher education degree and live with no children. The majority seems to live in area 4;

- 2: Youngest costumers with a lower monthly salary, lower education degree and have children. The majority seems to live in area 4;

- 3: Costumers with ages and salary between the values from the previous groups, the majority have children, a higher education degree and live in area 1.

6.3 Final Profiles

Looking at the final clustering label, we have 7 different profiles that combine the previous groups information, and they are summarized in the next table.

Clusters	Avg Age	Avg Salary (€)	Avg CustMonVal (€)	Have Children?	MostCom EducDeg	Liv Area
1	23	1197	334	Yes	Basic	4
2	66	3517	202	No	BSc/MSc	4
3	40	2043	220	Yes	High School	4
4	48	2516	210	Yes	BSc/MSc	1
5	70	3670	202	No	BSc/MSc	4
6	32	1619	180	Yes	High School	4
7	44	2312	217	Yes	BSc/MSc	1

Clusters	Avg PremMotor (€)	Avg PremHousehold (€)	Avg PremHealth (€)	Avg PremLife (€)	Avg PremWork (€)
1	85	602	151	121	113
2	391	101	139	20	20
3	404	97	128	19	19
4	436	70	110	14	14
5	207	257	230	51	49
6	205	235	232	50	54
7	231	257	206	50	50

Table 4: Final Clusters Profiles

- 1: Most valuable costumers, young adults(mean age is 23) with lower salary(around 1200€).This costumers spent the most in Household premiums(around 600€), have a lower education degree, live most in area 4 and have children;

- 2: Older people(mean age is 66) with high salary(around 3500€) that spent the most in Motor premiums(around 390€) and spend some in Household(around and Health premiums. This costumers have a higher education degree, live most in area 4 and don't have children;

- 3: Adult people(mean age is 40) with medium salary that spend the most in Motor premiums and spend some in Household and Health premiums. This costumers have a lower education degree, live most in area 4 and have children;

- 4: Adult people(mean age is 48) with medium salary that spend the most in Motor premiums and spend some in Household and Health premiums. This costumers have a higher education degree, live most in area 1 and have children;

- 5: Older people(mean age is 70) with high salary that spend the most in Motor , Household and Health premiums. This costumers have a higher education degree, live most in area 4 and don't have children;

- 6: Less valuable costumers, adult people(mean age is 32) with low salary that spend the most in Motor, Household and Health premiums. This costumers have a lower education degree, live most in area 4 and have children;

- 7: Adult people(mean age is 44) with medium salary that spend the most in Motor, Household and Health premiums. These costumers have a higher education degree, live most in area 1 and have children.

7 Marketing Solutions

There are a lot of possible approaches for the Marketing Department to explore, but analysing the customer profiling, we can follow two different approaches. The first consists of, looking at the profile, define the type(s) of premiums where the customers spent most money and trying to find a good solution in order to give some discounts in these topics. For example, in the cluster 1, the customers spent around €602 in Household, so they could receive a discount in this area. The department could also find what type of customers the company needs more at the moment or in the future and develop some advertising and a marketing approach focused on attracting new customers.

8 Final Clusters Visualization

8.1 Merging Outliers

After defining possibles marketing solutions, the final goal is achieved, so we might merge the outliers with the data set. From the 45 outliers, 1 was associated to cluster 2, another to cluster 4 and the rest to cluster 3.

8.2 Data Visualization

With the all the customers belonging to a cluster, it was time to do some data visualization. The next graph (fig 1) shows a two-dimensional projection from the clusters and in the

attachments, in (fig9) we might see a three-dimensional graph. By making use of UMAP, we were able to get these plots.

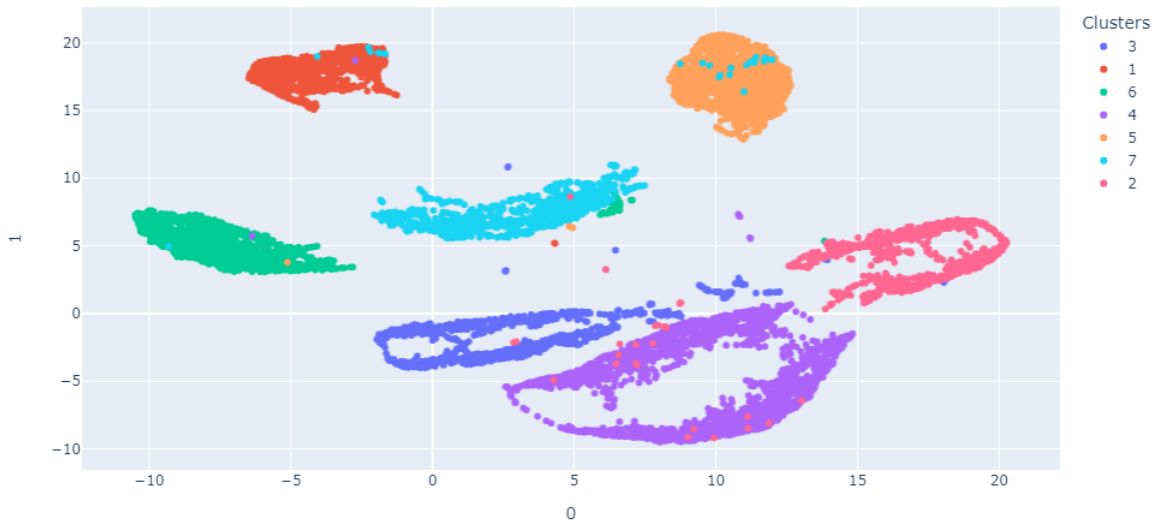

Figure 1: Data 2D Representation

There are some points that in the projections are closer to other clusters than the respective clusters. This might happen because they are outliers or they might be customers that have an extreme behaviour inside their respective cluster.

9 References

1. W. Iden, *Z-Score and How It's Used to Determine an Outlier*, https://urly.it/3h0r4, 2021

2. A. Mahbubul, *Anomaly detection with Local Outlier Factor (LOF)*, https://.it/3h0r-, 2020

3. *Sklearn Documentation for Isolation Forest Algorithm*, https://urly.it/3h0sj, 2007-2021

4. E. Lewinson, *Outlier detection and isolation forest*, https://urly.it/3h1zf, 2018

5. Ernst, *Outlier detection: DBSCAN*, https://urly.it/3h0sy, 2019

6. J. Brownlee, *Iterative Imputation for Missing Values in Machine Learning-Tutorial*, https://urly.it/3h1zh, 2020

10 Attachment

10.1 Graphs

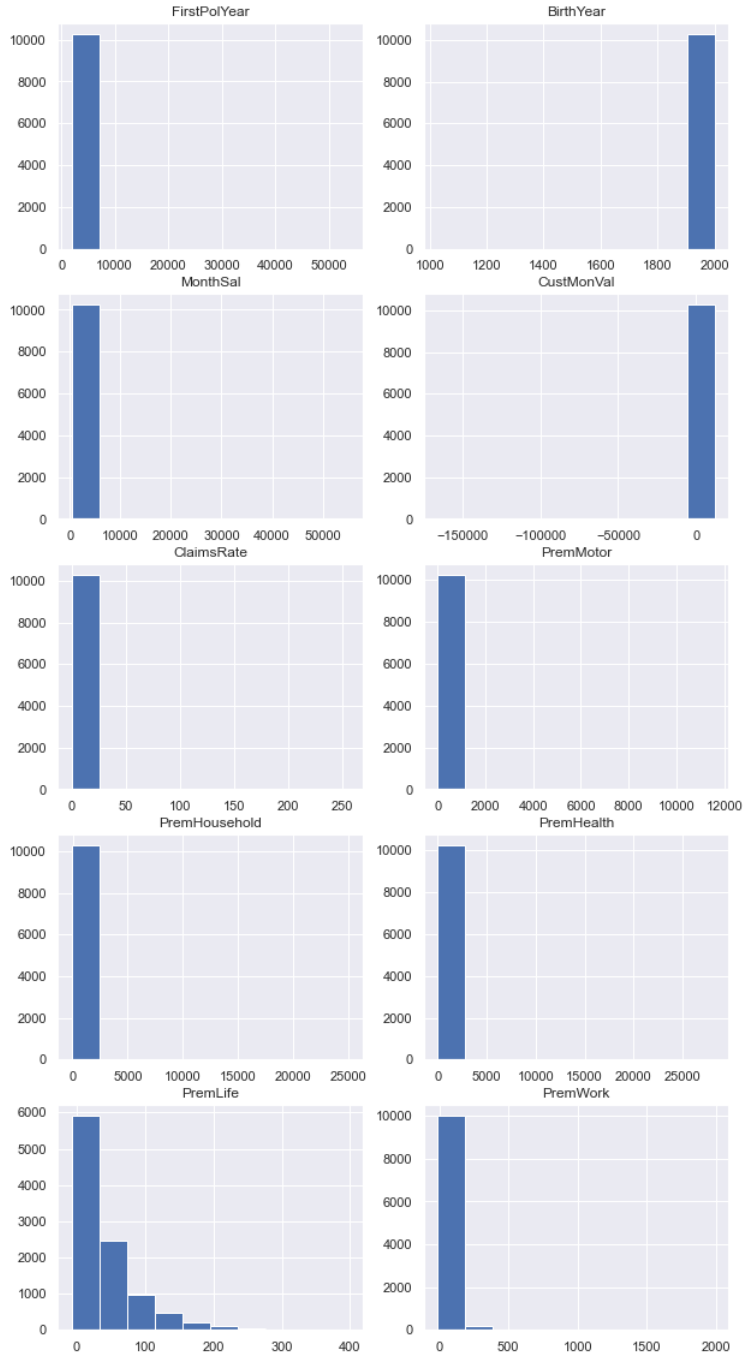

Figure 2: Numeric Features Histograms

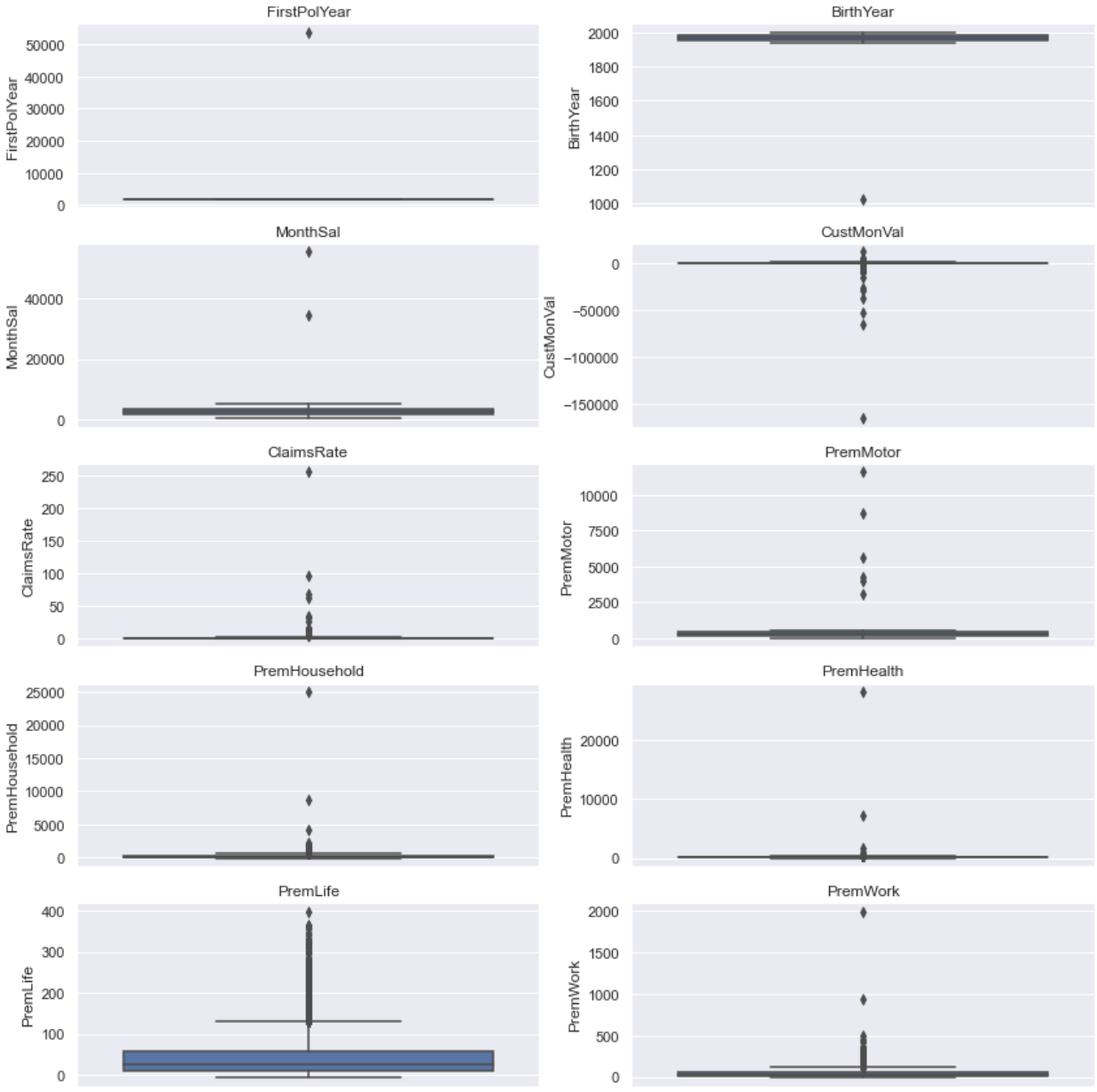

Figure 3: Numeric Features Boxplots

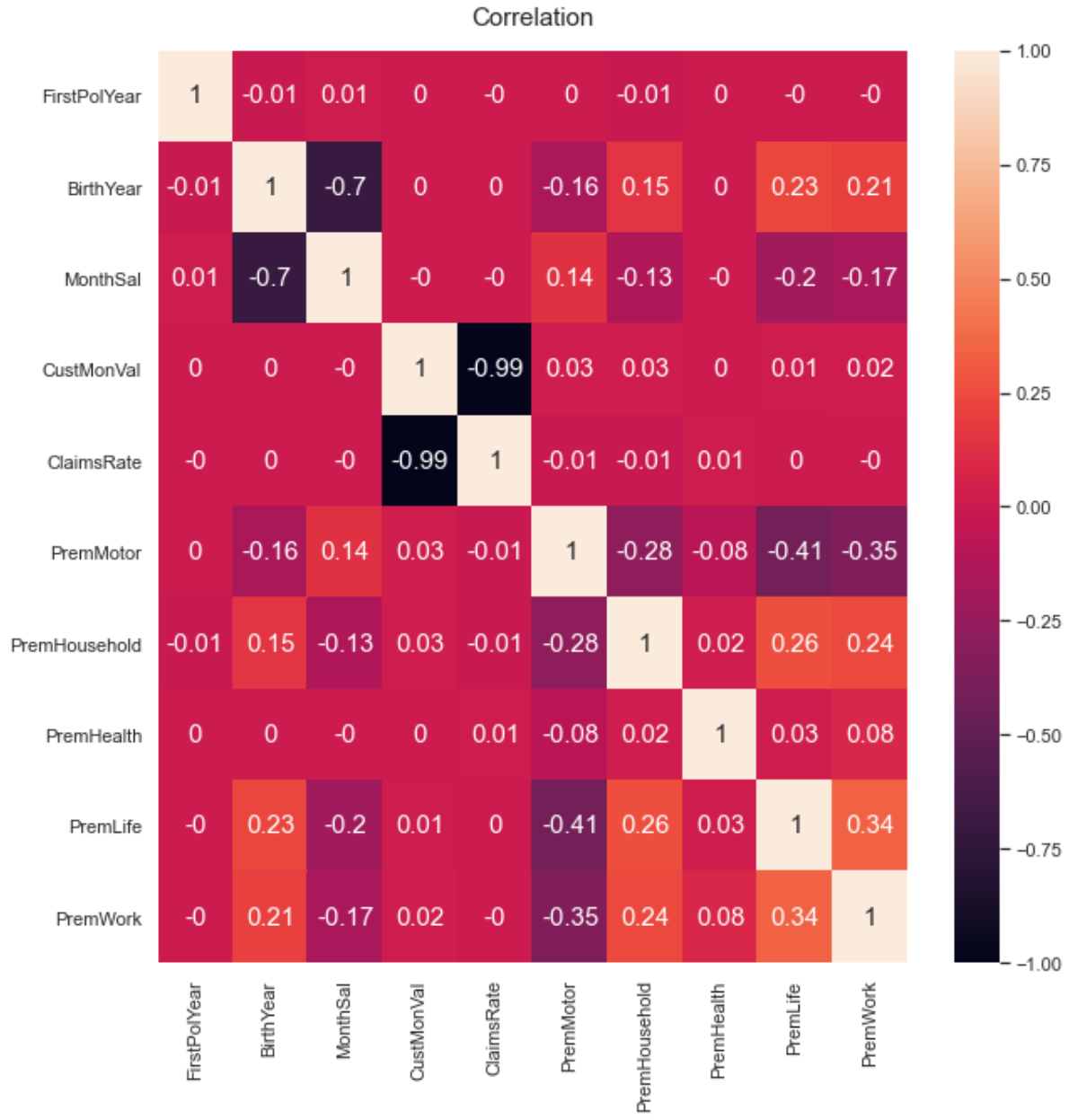

Figure 4: Numeric Features Correlation Matrix

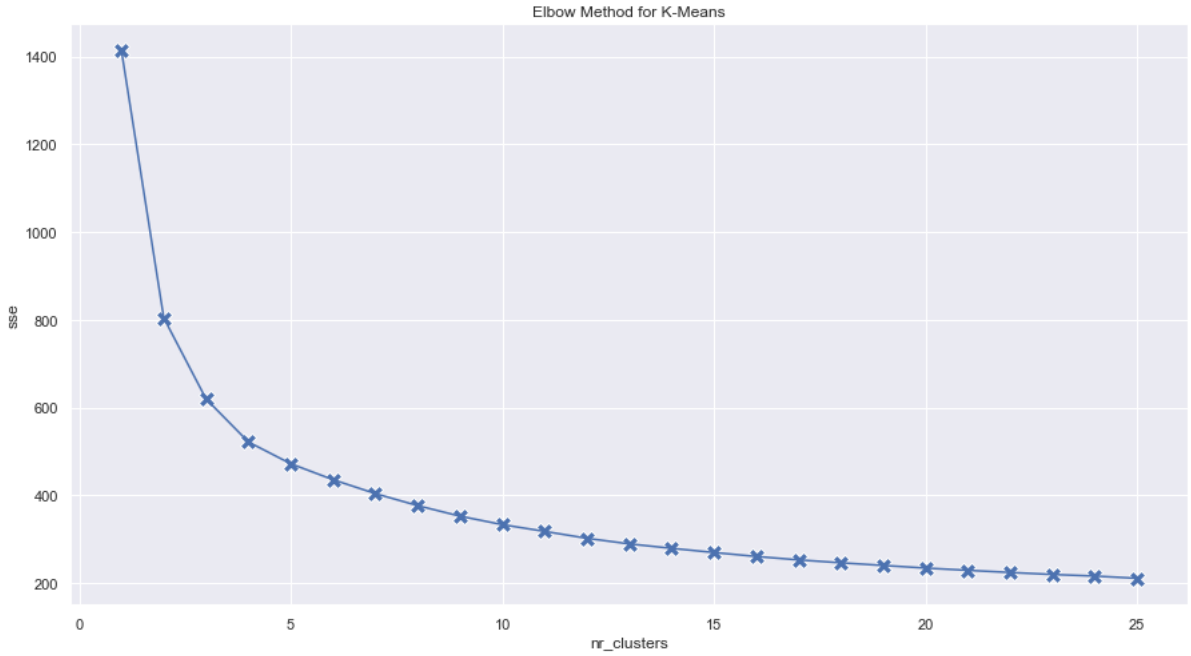

Figure 5: Elbow Plot for K-Means

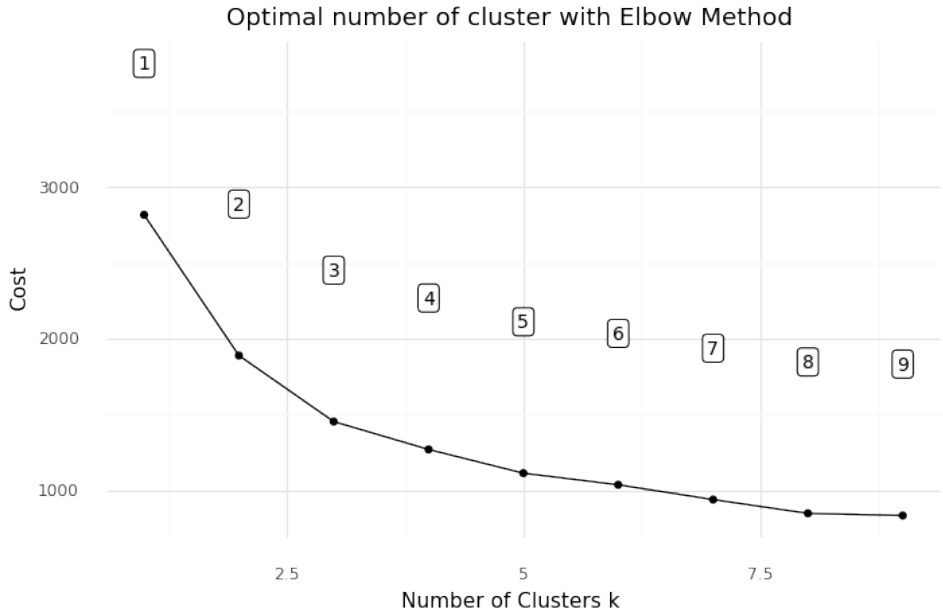

Figure 6: Elbow Plot for K-Prototypes

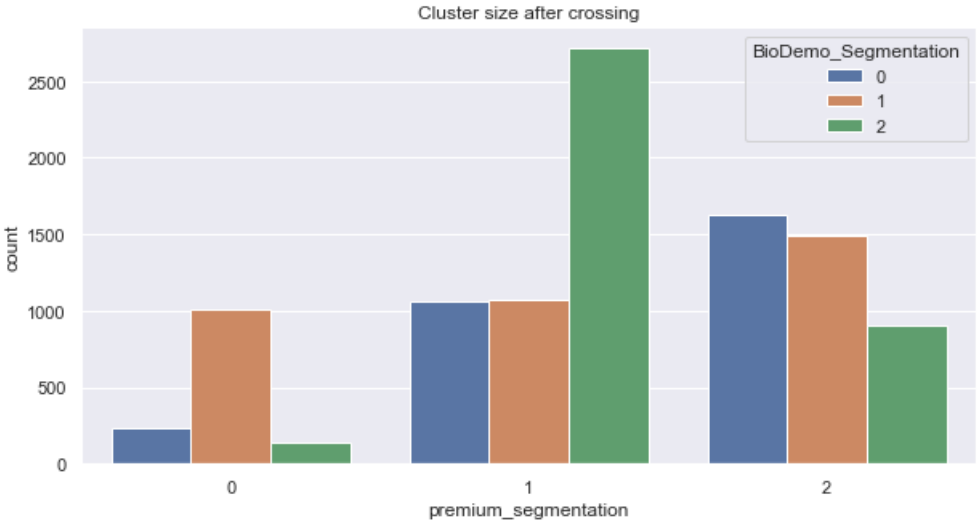

Figure 7: Clusters Size after Crossing

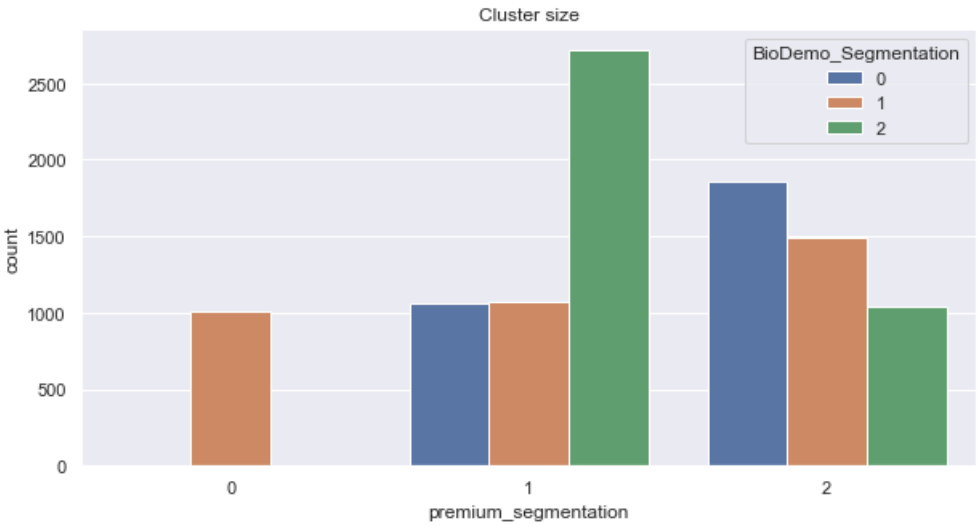

Figure 8: Clusters Size after Shifting Smallest Clusters

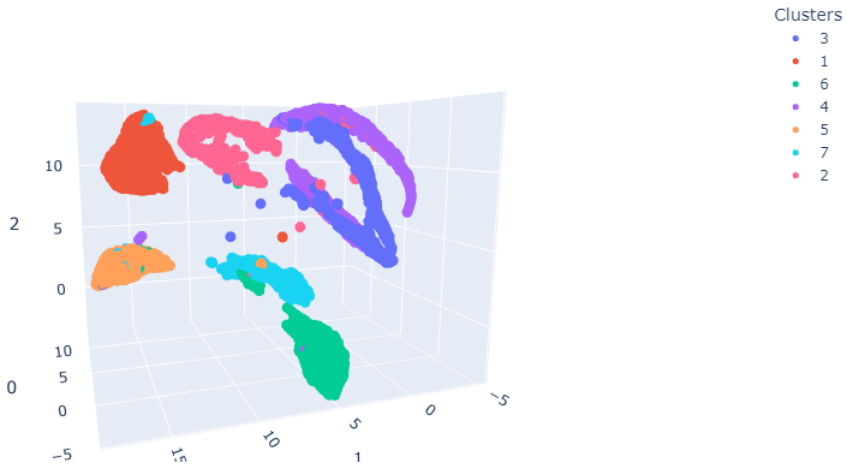

Figure 9: Data 3D Representation

10.2 Tables

Premium Segmentation Label	Bio-Demographic Segmentation Label	Final Cluster	Cluster Size
0	1	1	1008
1	0	2	1062
1	1	3	1074
1	2	4	2714
2	0	5	1862
2	1	6	1491
2	2	7	1040

Table 5: Table with the Merged Clusters Label and Size