Master's Degree Program in Data Science and Advanced Analytics

Data Mining

Data Mining PROJECT REPORT

Group 28

Mohamed Ettaher Ben Slama	20221039
Skander Chaabini	20221041
Bruna Vieira Duarte	20210669

Academic Year

2022/2023

NOVA Information Management School

Universidade NOVA de Lisboa

Table of Content

Introduction

Market prominence and diversification increase competition, which makes it harder for businesses to succeed. Businesses must instead concentrate on creating more individualized marketing tactics through consumer segmentation in order to reach a wider market segment to succeed. In this project, the case of, A2Z Insurance, an insurance company that operates in Portugal will be handled. In order to better understand and meet the needs of its customers, it is important for A2Z Insurance to distinguish and identify the different profiles of customers it serves. The clustering made in this project is mainly to help build solid marketing strategies targeting each type of customer.

1. Data Exploration

It is very essential to first comprehend the data and get as many insights as possible before preprocessing and clustering. Data quality, correlations, and patterns between the variables should be examined and visualized, as well as identifying anomalies in the data.

1.1 Data Review

The insurance company's dataset is composed of 10296 observations and 14 variables all of them are floats except for one that is an object. The variables reflect different information about the

customers describing them individually (birth year, education level, monthly salary) and in relation to

the company (customer monetary values, the different premium they're paying for…)

1.2 Data Analysis Exploration

It was essential to make a further analysis about each feature to collect insights .Thus as a first step a descriptive table was implemented to see the statistical data about each feature (see **Figure 1** Annexes).

After that ,Missing values were checked and it was discovered that the data contains missing values that will be dealt with later (see **Figure 2** Annexes).The next step was to check the duplicate values since it is essential that no record is duplicated not to bias the interpretations .No duplicate value was found in the data .

Finally it was necessary to check the features distribution.For the metric features histograms were implemented for each feature .(see **Figure 3** Annexes) .It was noticed that the data suffers from skewness ,a challenge that was treated later .For the categorical features a count plot for each feature was implemented (see **Figure 4** Annexes)

1.3 Coherence check.

The first coherence check was to inspect if there are clients with firstpolyear after 2016 which is the date of extraction of the dataset . One record was discovered with such a problem.
The second coherence check was to check if there were clients born before 1900 which is quite impossible. One record was discovered with such a problem.
The third coherence check was to check if there were minors in the dataset which means clients born after 1998 .116 clients were minors. So it was necessary to inspect these records with count plots and histograms for only those records. It was discovered that the maximum birth year is 2001 thus the minimum age is 16, the education is 'basic' for all of them which is logical (No BSc/MSc or Ph.D. found), and the majority of them have children which is a strong motivation to have insurance.

The legal age to have insurance in Portugal is 18. However, someone who is 16 or 17 can have insurance with parental approval as they can also marry and drive motorcycles at this age with the approval of their parents. It was decided to accept those values since they can be possible.
The fourth check was to see if there were clients with birth years bigger than the firstpolyear . It was surprising that there were 1997 records (20% of the data) with such a problem. Since it was not possible to drop this huge number of records it was decided to deal with this problem later by dropping one of the columns.
The fifth coherence check was to check if the customer ID has duplicate values or not. It was discovered that all records have unique customers ID.

1.4 Outliers Check

The next step was performing an outliers check. Outliers are data points that fall significantly outside the range of the rest of the data. They can be caused by measurement errors or errors in data entry, or they may represent genuinely unusual observations. Outliers can enormously affect the model's performance. Thus, the need to deal with them carefully.

For this matter, box plots were created to identify points outside the whiskers and assess the situation. More explanation is present in **Figure** 5 (Annexes).

It was clear that there were many outliers in all variables.After that, it was interesting to check a pairwise plot to see if there are outliers in two dimensions.It was discovered that there are a lot of outliers in the pairwise plot too (See **Figure** 6 Annexes)

2. Data Preprocessing

Cleaning the data and its preparation for Clustering is the purpose of the data preprocessing step. This might include filling data gaps from the data exploratory analysis step and dealing with missing values, identifying the important features, transforming the data, and creating new relevant features when it is best suited. This is a crucial phase, as the clustering solutions presented will depend on this step enormously.

2.1 Droping Impossile Values

It was necessary before doing anything to drop the impossible values because keeping them will make the analysis and clustering biased

2.2 Treating Missing values

This step consists of finding a solution to the missing values. The data contained 389 missing values in different variables. So it is obvious that it was crucial to deal with each variable at a time and find the most suitable solution to deal with it.

First, the premiums variable that contained missing values was filled with zero because it was noticed that the only prem variable without missing values is PremHousehold is the only premium that contained zeros.

Then It was interesting to check if there were clients with 0 value in all the premiums and It was discovered that 13 records were false clients. As a result, these records were eliminated.

Secondly, GeoLivArea was filled with the mode since there is only one missing value.

Then, the rest of the values (FirstPolYear, BirthYear, EducDeg, MonthSal, Children) were filled using KNN imputer. After careful analysis, No pattern was found to help fill these values so KNN imputer seemed like the most objective method. Of Course, before using it the categorical data were transformed to numerical and the data was scaled using the MinMax scaler. After filling in all the missing values the scaling was reversed.

2.3 Outliers Treatment

As mentioned in the first part, there are outliers to deal with. However, It is essential to note that determining whether a value is an outlier that should be removed or not is very subjective. While there are certainly valid reasons for throwing away outliers if they are the result of a computer glitch or a human error, eliminating every extreme value is not always a good idea.
So it was noticed that the data contained skewed distribution so after some research it was opted for dealing with outliers in two phases first deleting some very obvious outliers to make the data more normal and then making a normalization using one of the transformation methods that we will discuss later than perform further checks for outliers.
The method chosen for the first outlier removal was a manual filter based on the box plot.
After the first outlier removal, most of the distributions had better shapes.(See **Figure7&8** Annexes)
The second outliers treatment was based on density rather than one-dimensional observations. The method used was first the Local outlier Factor from Sklearn [1] which is an algorithm for identifying anomalies or outliers in data. It is based on the idea that an outlier is a point that is significantly different from its neighbors in the feature space. Then the DBSCAN was used to eliminate the rest of the outliers, an algorithm in which noise points are considered to be outliers because they do not belong to any cluster. After the removal of the outliers 160 rows were removed and 98% of the data was kept.

2.4 Feature Engineering

- After exploring the data and taking a look at each feature alone and the relationship of each feature several actions were decided :
- Creating a new categorical variable containing information about the generation (Baby_Boomers, Generation_X, Millennials, Generation_Z)of the customer because can be extremely helpful in the Marketing Strategy
- Since the premium variables are calculated yearly we chose to transform the monthly salary into a yearly salary
- FirstPolYear was changed to a period of inscription calculated in the number of years as a client
- The variable CustMonVal was changed in a way to calculate the value of the customer throughout the whole time he was a customer by an average yearly value that calculates each value approximately

2.5 Feature Selection

A correlation matrix was implemented to see if there were features that are highly correlated so it will be possible to keep only one . After checking the correlation matrix (See **Figure 9** Annexes)

It was discovered that yearly_salary and birth year are highly correlated so the birth year was dropped since the generation feature was created.
Also, yearly_cust_value is highly correlated with customerval so we will drop it.
ClaimsRate is also highly correlated with yearly_cust_value so it was dropped.
The only other interesting thing is that premMotor is negatively correlated with all other premiums but since the other premiums are not highly correlated together it was decided not to group them together.

2.6 Scaling

Previously to any scaling process, the datasets were separated into categorical and numerical data.
For this dataset, different approaches were tried :

- MinMax scaler
- Power transformation
- Robust scaler

After trying the three methods they were compared to each other and the power transformer was chosen because the shape of the data is the best with its transformation. (See **Figure10** Annexes)
The power transformer class has several different power transformation methods available, including the yeo-johnson method. The yeo-johnson method is a power transformation that is similar to the box-cox method, but it can handle both positive and negative data values.[2]
It was clear that the distribution in the power transformer using yeo-johnson method gave better results.In fact, the data looks very close to normal in most variables[3]

2.7 PCA Check

To access what are the most important variables in our dataset a principal component analysis was performed. After that a correlation matrix containing the 4 principal components that explain 80% of the data and the variable we selected (See **Figure 11** Annexes). It was understood after this check that Premiums features are highly correlated with the first Principle component especially the motor premium and Value features Which are related to the client and his income and are correlated with the other three components. This made the perspective that can be used clearer: Premium features and Value Features.

2.8 One Hot Encoding

One-hot encoding is a way of representing categorical variables as numerical data by creating a new binary (0 or 1) column for each possible category.
This technique was used on 'EducDeg', 'GeoLivArea', and 'Generation' variables.
By finishing this Step the Data preprocessing was finished and the data was ready to implement Clustering Models.

3. Clustering

It was difficult to immediately determine which cluster algorithm would work best for the data in the clustering phase. Therefore, different clustering methods were tried and the results were compared in order to choose the most fitting method for the data. The R^2 value served as the main basis for comparison, but the clusters were also visualized and profiled to ensure their accuracy and meaningfulness before a final decision was made.

3.1 K-means

The first applied algorithm was the K-means. To begin, the Elbow and Silhouette methods were applied in order to get the optimal number of clusters. The two approaches settled on K=2. After executing the K-means, several visualizations were carried out to understand the distribution of the variables. The two clusters had a balanced number of observations, they were mainly divergent in the premium variables with no difference in the personal data values. The K-means algorithm gave an R^2=0.28.

3.2 K-means with perspectives

The data was separated into Premium features and Personal Value features. An R^2 plot was applied on the two datasets and resulted in optimal K=3 for both of them. This led into 9 clusters after combining the two perspectives; a number of clusters that is considered high with an imbalanced distribution. To reduce the number of clusters, a hierarchical clustering was implemented on the 9 clusters in order to merge the most similar ones. The dendrogram gave an optimal number of 3 clusters, merging the clusters with the same perspective (Premiums). Thus the second perspective of Personal Value was no longer taken into account, meaning that this model is not efficient with the final data. The R^2 of the K-means with perspectives algorithm was 0.41.

3.3 Hierarchical Clustering

The first step in building clusters with the hierarchical algorithm is to properly set the parameters. This process began by selecting the appropriate linkage method through the construction of clusters with various linkage methods and a range of cluster numbers. The ward linkage method ultimately proved to be the most effective. A dendrogram was then created giving an optimal number of 3 clusters. The hierarchical clustering with 3 clustering resulted in an R^2=0.34.

3.4 DBSCAN Clustering

An attempt to use density-based clustering on the data was made applying the DBSCAN algorithm. Starting by finding the best epsilon parameter, which gave an eps=1.7. The model resulted in 2 clusters; a cluster containing all the data points and a second for outliers. The DBSCAN gave the lowest R^2=0.007. Given the results of the clusters and the R^2, it is clear that density-based clustering is not suited for this dataset and therefore no other density-based algorithms were tried.

3.5 Self Organizing Maps

The training of Self Organizing Maps was done through two phases: the unfolding phase, and the fine-tuning phase.
In the unfolding phase, the algorithm started with a high radius=17 and ran through 100 iterations, in order to spread the units in the region of the input space. Followed by the fine-tuning phase, in which the radius was reduced to 4 and also the learning rate was decreased, in order to reduce the quantization error, and center the units in the areas where the density of patterns is highest.
Finally, the U-matrix and the Hits map were plotted in order to visualize the potential clusters in a 2-dimensional space.

- **K-Means on top of SOM units:**

K-means algorithm was applied on SOM units with K=2. Each unit from the 2500 neurons of the output space was clustered within these two clusters. Then each data point in the input space was matched with its Best Matching Unit from the neurons, and therefore classed within its final cluster. The K-means on top of SOM units gave an R^2=0.27.

- **Hierarchical clustering on top of SOM units:**

Hierarchical clustering was applied on SOM units with K=3. Each unit from the 2500 neurons of the output space was clustered within these three clusters. Then each data point in the input space was matched with its Best Matching Unit from the neurons, and therefore classed within its final cluster. The hierarchical clustering on top of SOM units gave an R^2=0.37.

3.6 K-Means Hierarchical Hybrid

First step was implementing K-means with K=50, then applying hierarchical clustering on the 50 clusters with ward linkage and euclidean distance. The dendrogram gave an optimal number of 3 clusters. This algorithm gave an R^2=0.35.

3.7 Gaussian Mixture Model

Gaussian clustering, also known as Gaussian mixture modeling, is a method of cluster analysis in which a dataset is modeled as a mixture of multiple Gaussian distributions.
Before applying the GMM, the BIC / AIC approach was used to find the optimal number of clusters. After plotting the BIC / AIC curves, it gave an optimal K=3. The GMM algorithm gave an R^2=0.33.

3.8 K-Prototype Model

The K-prototypes model is a clustering algorithm that is used to cluster mixed data types (both numeric and categorical variables). It is a combination of the K-means and K-modes clustering algorithms.
The first step was to initial the data by dividing the numeric and categorical features and then integrate them into a matrix. To check the optimal number of clusters, the cost function was plotted, giving an optimal K=3. The first K-prototype algorithm gave an R^2=0.40. However, after checking the distribution of the variables within the cluster, it was noticed that 3 variables were not affecting the clusters.
Therefore an optimal k-prototype was tried without the ‘YearsAsCust’, ‘Yearly_cust_value‘, and the 'GeoLivArea' features. This final K-Prototype algorithm has obtained the highest R^2=0.52.

4. Clusters Analysis & Profiling

In this final part, the clusters were visualized in a two-dimensional space, then analyzed, giving a customer profile for each cluster.

4.1 t-SNE

t-Distributed Stochastic Neighbor Embedding is a visualization method for high dimensional data, it reduces the dimensions in a way that can be visualized. (See Figure in annexes)

4.2 UMAP

Uniform Manifold Approximation and Projection is a machine learning algorithm for visualizing high-dimensional data, similar to t-SNE. However UMAP is faster than t-SNE and is better at preserving the global structure of the data by minimizing the intra-cluster distance and maximizing the extra-cluster distance. (See Figure in annexes)

4.3 Decision Tree

To be able to interpret the cluster solutions, it is very beneficial to visualize a decision tree, and check the rules that the algorithm used to label each cluster. It can also be used to predict the cluster of new observations.
After modeling the decision tree It is estimated that in average,the decision tree is able to predict 91.27% of the customers correctly which is a good indicator that the clustering was satisfying and followed a reasonable reasoning that was understood thanks to the decision tree.(See **Figure 19** Annexes)
Based on the results provided by the decision tree, it was understood that the 'PremMotor' and the 'Yearly_Salary' were the most important features and that the clustering model depends highly on them.

4.4 Customer Profiling

After visualizing the global distribution of the clusters and understanding the most important features, this part is about characterizing the clusters and profiling each cluster.
Based on the multiple visualizations (See Figures x to x), the profiling table below was created.

Low Income Cluster	Middle Income Cluster	High income Cluster
Mean Salary= 1426 **Generation**: Millennials+GenZ **Education:** HighSchool **Premium :** Balanced	**Mean Salary=** 2462 **Generation:** GenX **Education:** BSc/MSc **Premium :** Motor Premium	**Mean Salary=** 3311 **Generation:** Baby Boomers **Education:** BSc/MSc **Premium :** Health Premium

5. Marketing Strategies & Conclusion

Marketing Strategies

High Income :
This cluster has money but apart from the health premium they are not putting a great amount of money into the insurance. Thus the strategy with this cluster should be very specific. Discounts and special offers might not be very effective since money is not really an issue but rather that the insurance should build trust: Make sure the person trusts the premiums offered and believes in its quality. This can be achieved through good customer service, transparent business practices, and positive reviews and testimonials from satisfied customers. Since the majority of this cluster is Baby boomers the marketing campaign should be more personal and involves door-to-door visits or sending signed letters . Indeed such actions will have a greater effect on this cluster rather than digital campaigns.
Middle Income :
This class is very challenging since they have money but they would rather use it elsewhere. Also what is special about this cluster is that its majority is the generation and quite well-educated. Thus they might be more ready for digital campaigns . Yet another interesting factor that should be taken into consideration is that they put a lot of money into Motorpremium.
The best strategy with this category is Product bundling in a way that the insurance sent them emails to present them the opportunity to change their current premiums to a new package containing the motor premium as a principal part of the package and other premiums added to the package with a reasonable discount to push them to buy.
Low income :
Low-income cluster is also a very critical cluster since although they do not have the most money they are the most category investing in insurance premiums. In other words, they are the cash cow of the insurance, and handling them the right way is very important. Since the majority are millennials and Genz the campaign for this cluster must be digital since such a category would be using social media. So it would be wise to make sponsored targeted advertisements for this category with discounts because the main challenge with this category is making them loyal to the insurance since they will follow always the cheapest offer. So a digital presence, discounts, and presenting long-term deals is the best approach that can be followed with them.

Conclusion

The quality of the data was a big challenge in this project. In fact, with the incoherent values and the missing values it was essential to take subjective decisions that can affect the analysis . However, it was very interesting to discover the most important variables via clustering first and the decision trees later.

The marketing strategies presented earlier should be a good start . Yet, focusing on collecting better-quality data should be the next step to do . In addition trial and error and monitoring those strategies and fixing benchmarks and KPIs for them is necessary.

References

Introduction to Machine Learning Third Edition Ethem Alpaydın The MIT Press Cambridge, Massachusetts London, England

Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems

https://github.com/joaopfonseca/Data-Mining-22-23

https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.LocalOutlierFactor.html [1]
https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PowerTransformer.html[2]
https://www.ucd.ie/ecomodel/Resources/QQplots_WebVersion.html#:~:text=The%20normal%20distribution%20is%20symmetric,deviate%20from%20the%20straight%20line[3]

Annexes

	count	mean	std	min	25%	50%	75%	max
CustID	10296.0	5148.500000	2972.343520	1.00	2574.75	5148.50	7722.2500	10296.00
FirstPolYear	10266.0	1991.062634	511.267913	1974.00	1980.00	1986.00	1992.0000	53784.00
BirthYear	10279.0	1968.007783	19.709476	1028.00	1953.00	1968.00	1983.0000	2001.00
MonthSal	10260.0	2506.667057	1157.449634	333.00	1706.00	2501.50	3290.2500	55215.00
GeoLivArea	10295.0	2.709859	1.266291	1.00	1.00	3.00	4.0000	4.00
Children	10275.0	0.706764	0.455268	0.00	0.00	1.00	1.0000	1.00
CustMonVal	10296.0	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	0.742772	2.916964	0.00	0.39	0.72	0.9800	256.20
PremMotor	10262.0	300.470252	211.914997	-4.11	190.59	298.61	408.3000	11604.42
PremHousehold	10296.0	210.431192	352.595984	-75.00	49.45	132.80	290.0500	25048.80
PremHealth	10253.0	171.580833	296.405976	-2.11	111.80	162.81	219.8200	28272.00
PremLife	10192.0	41.855782	47.480632	-7.00	9.89	25.56	57.7900	398.30

Figure1:Data Statistical Description

```
Out[14]: CustID            0
         FirstPolYear     30
         BirthYear        17
         EducDeg           0
         MonthSal         36
         GeoLivArea        1
         Children         21
         CustMonVal        0
         ClaimsRate        0
         PremMotor        34
         PremHousehold     0
         PremHealth       43
         PremLife        104
         PremWork         86
         dtype: int64
```

Figure2:Missing Values

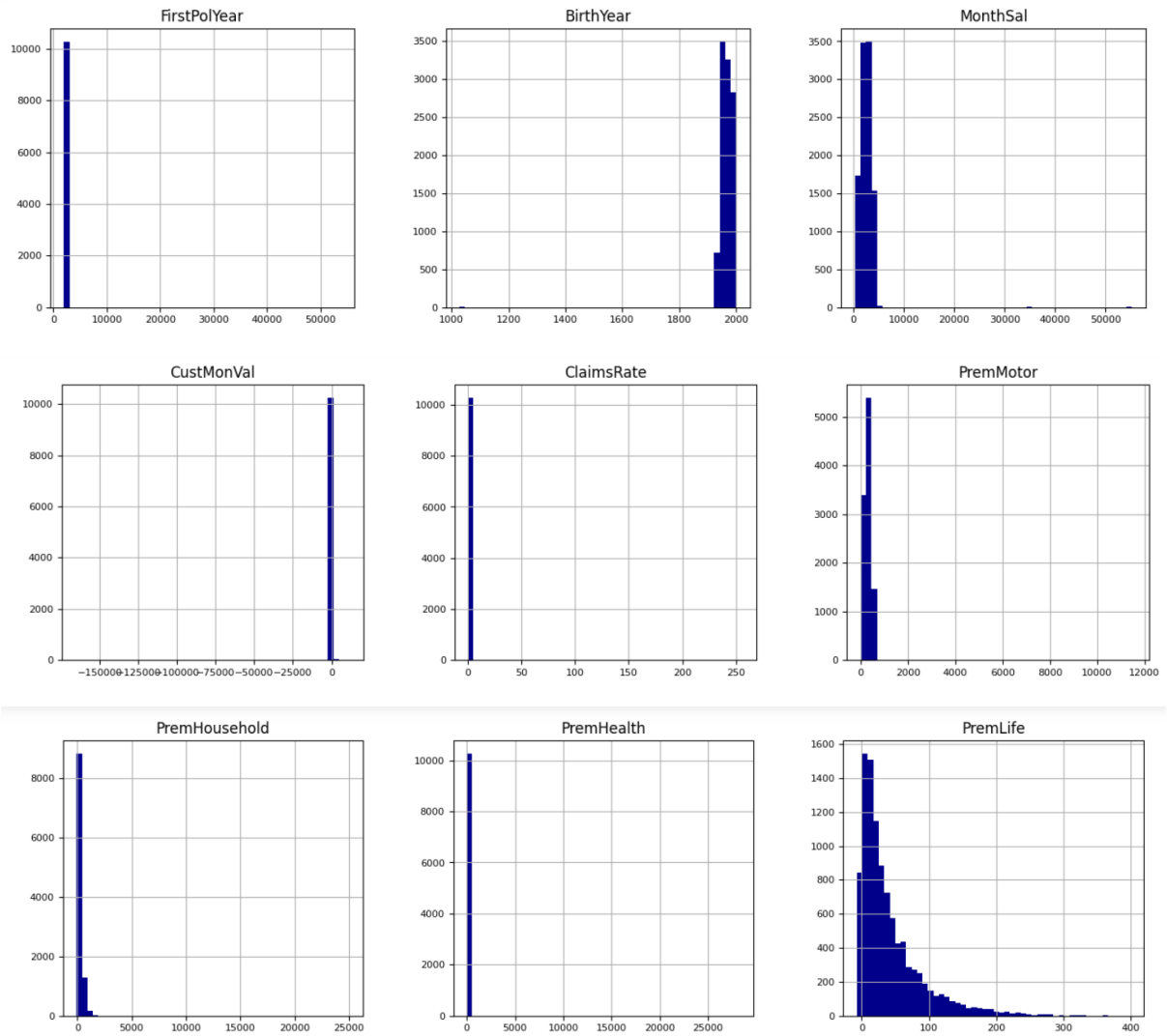

Figure 3: Histograms for metric features

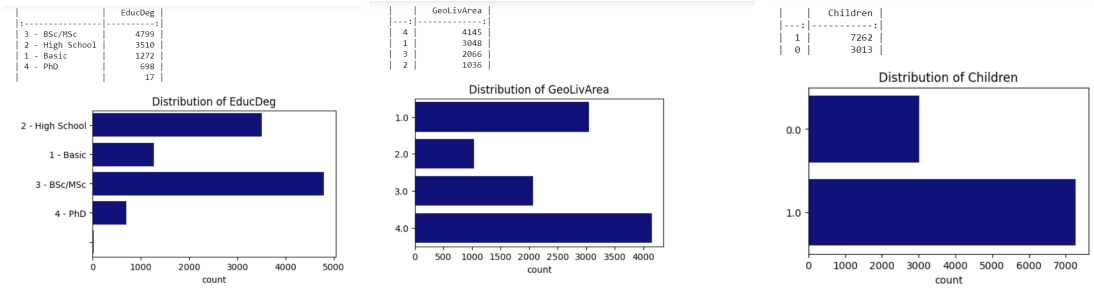

Figure 4:Countplots for categorical Features

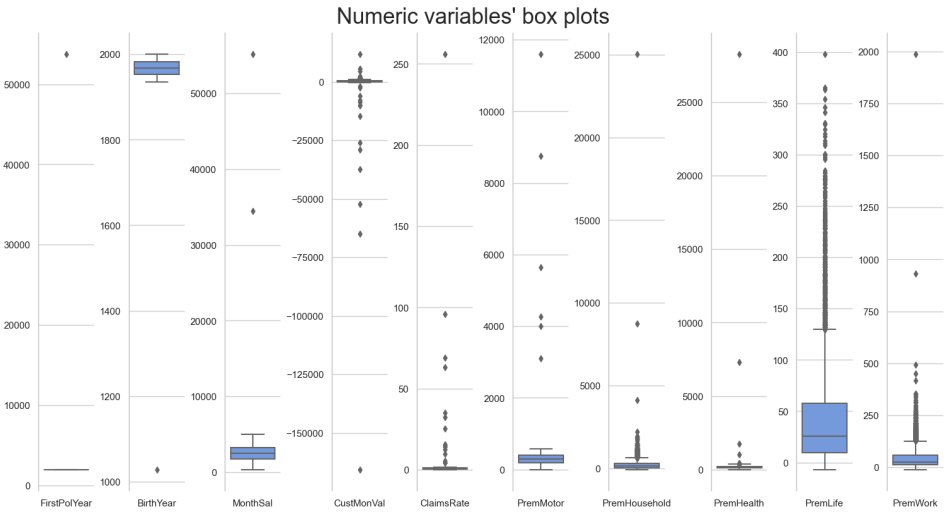

Figure 5 : Numeric Variables Box Plots

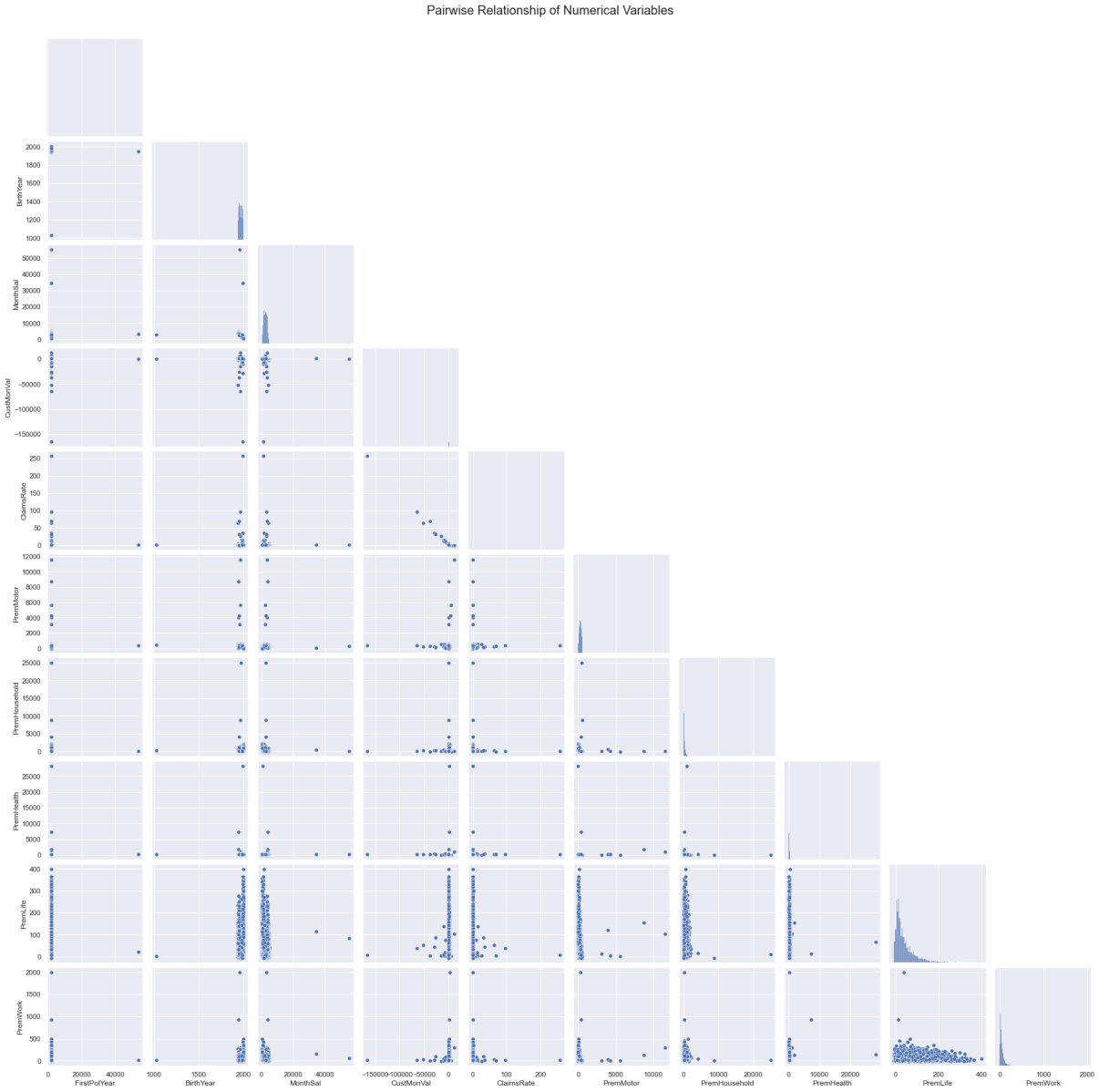

Figure 6: Pairwise Relationship between numerical variables

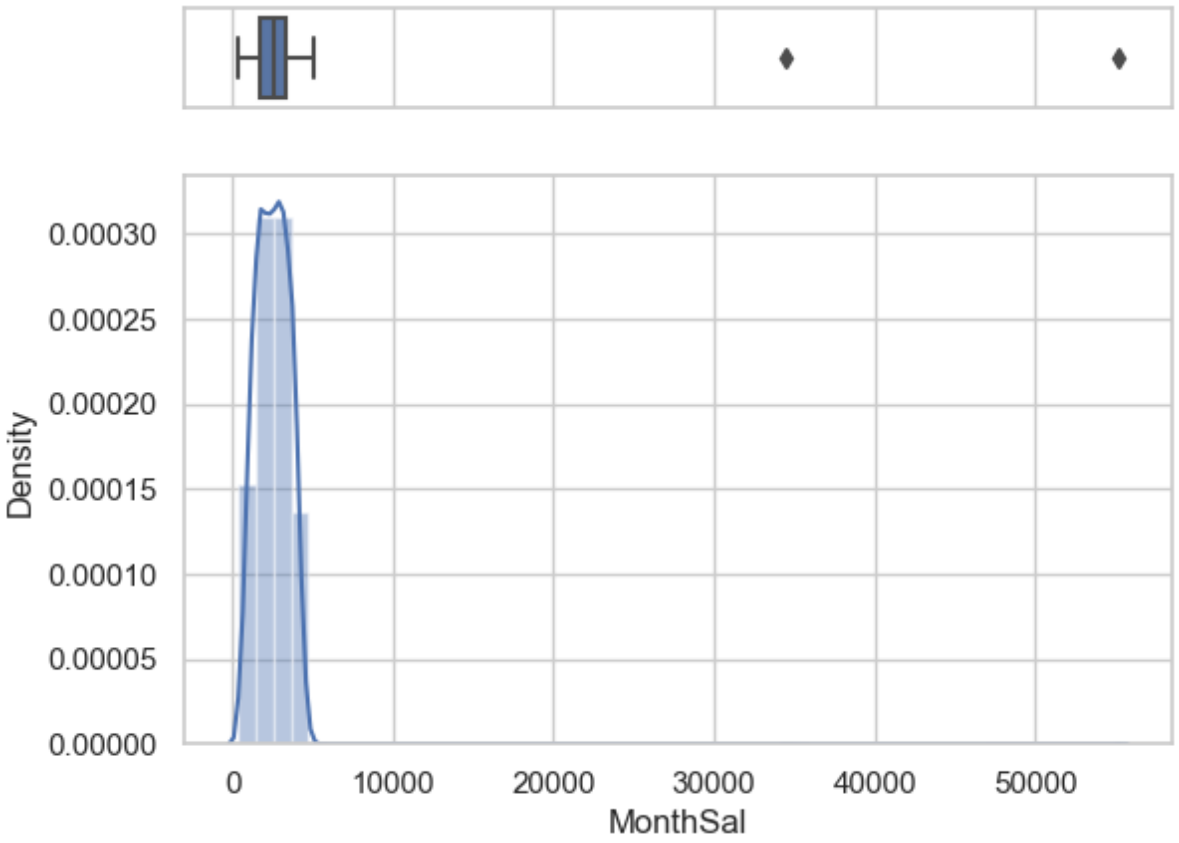

Figure 7 : Monthly salary Distribution before first Outliers Treatment

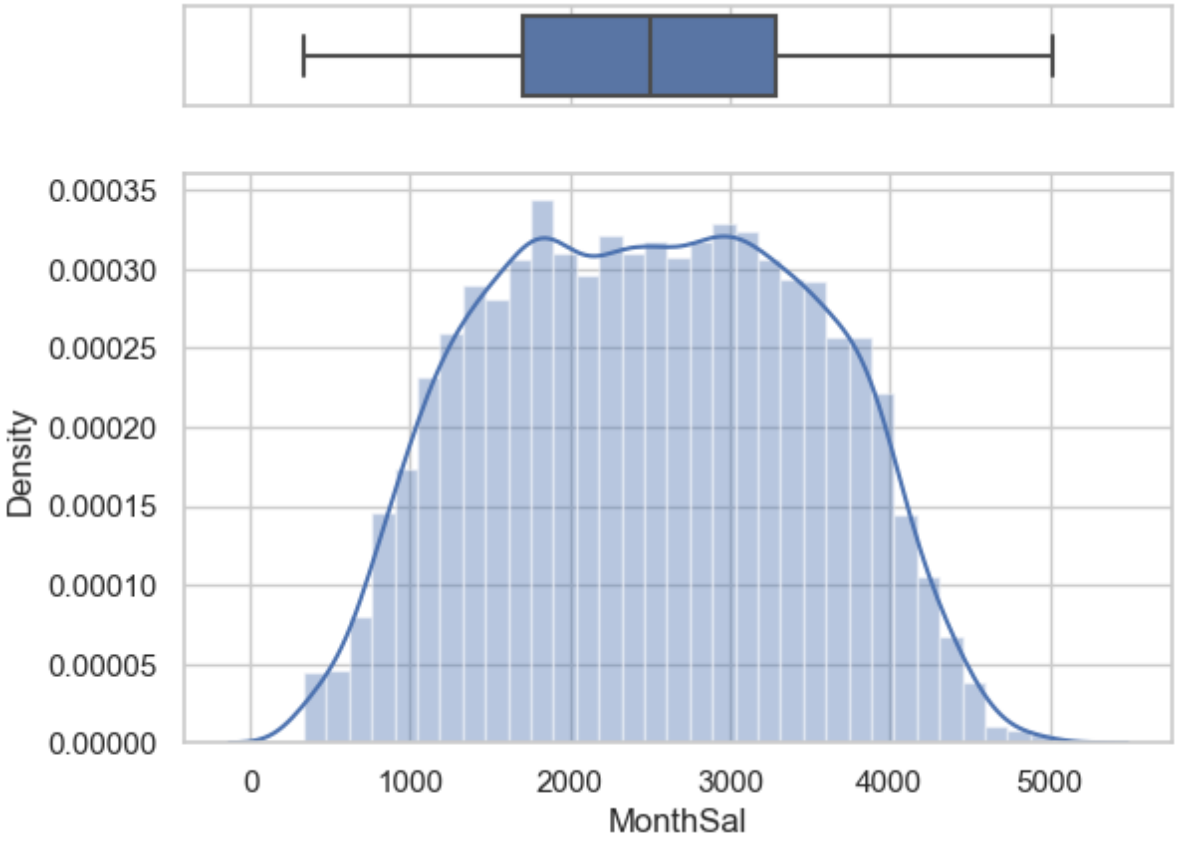

Figure 8 : Monthly salary Distribution after first Outliers Treatment

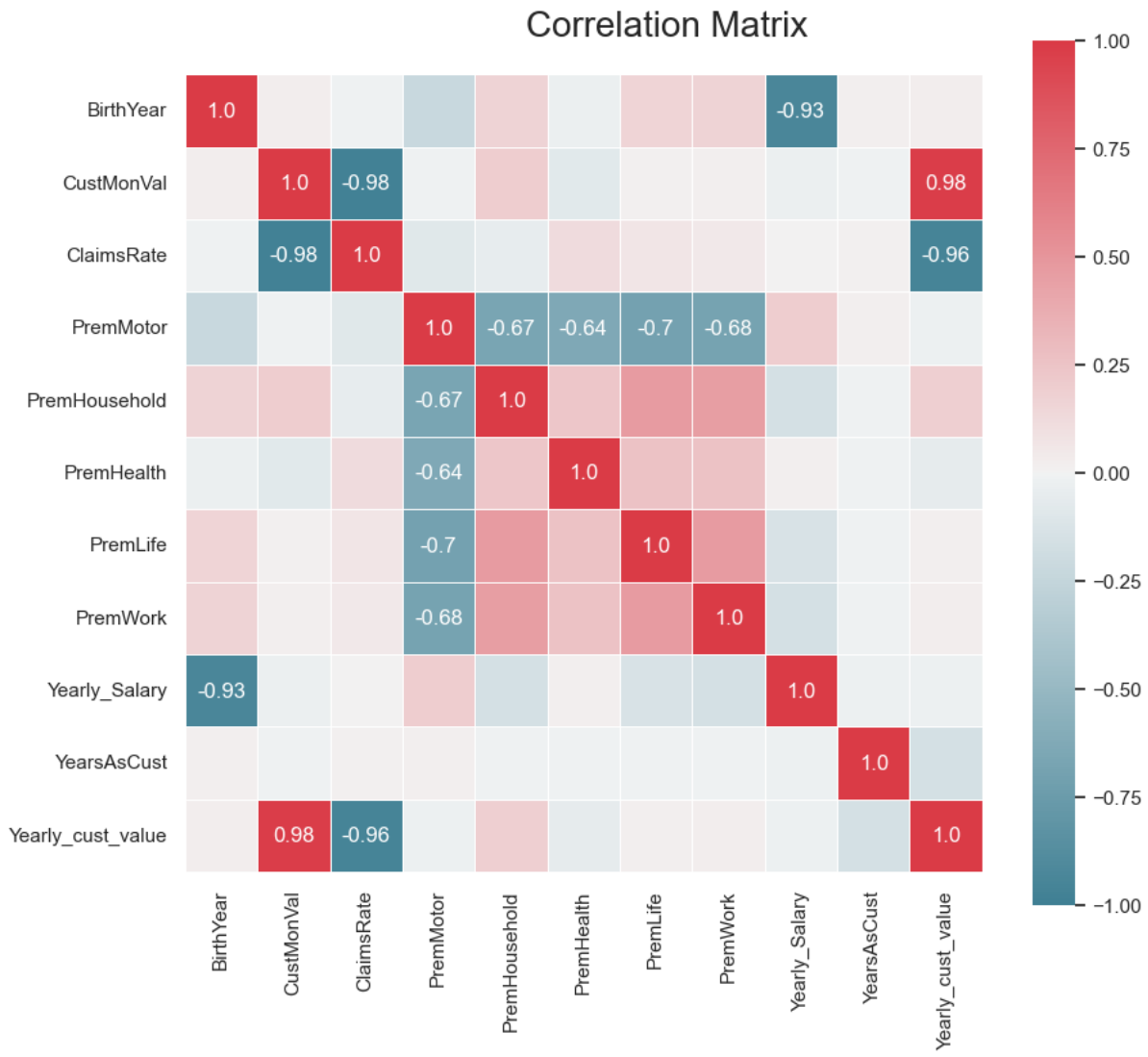

Figure 9 : Variables Correlation Matrix

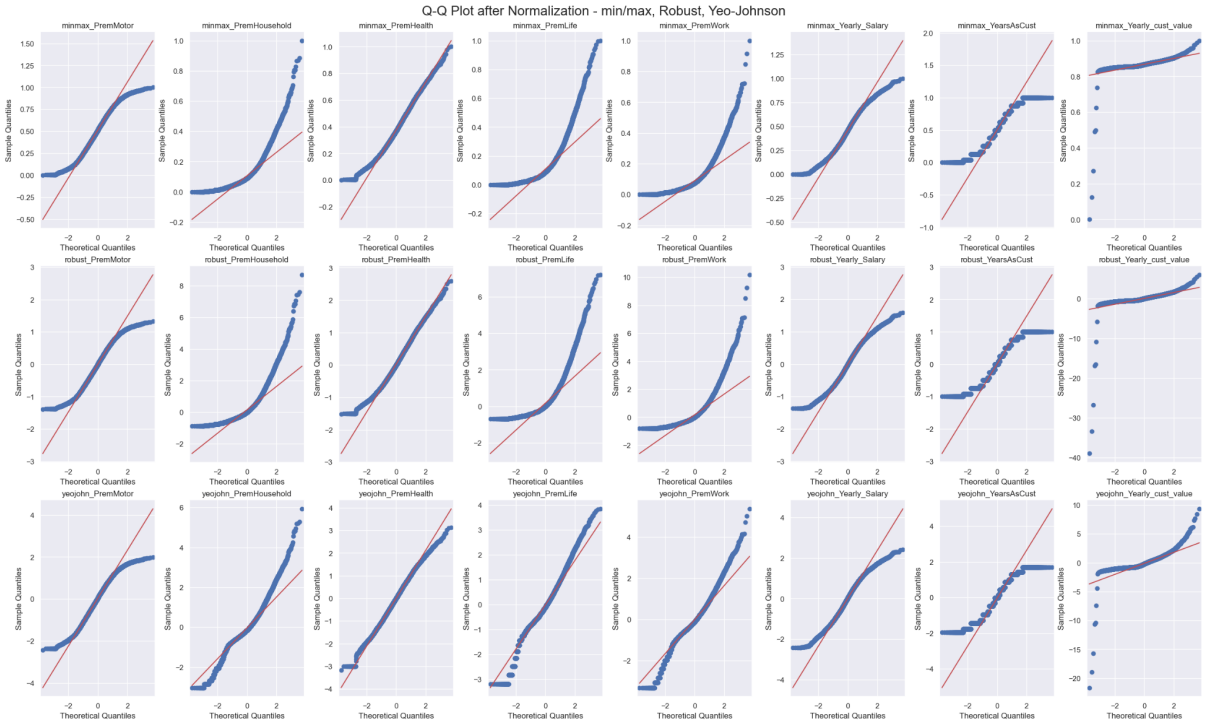

Figure 10 : Scaling Methods Comparaision

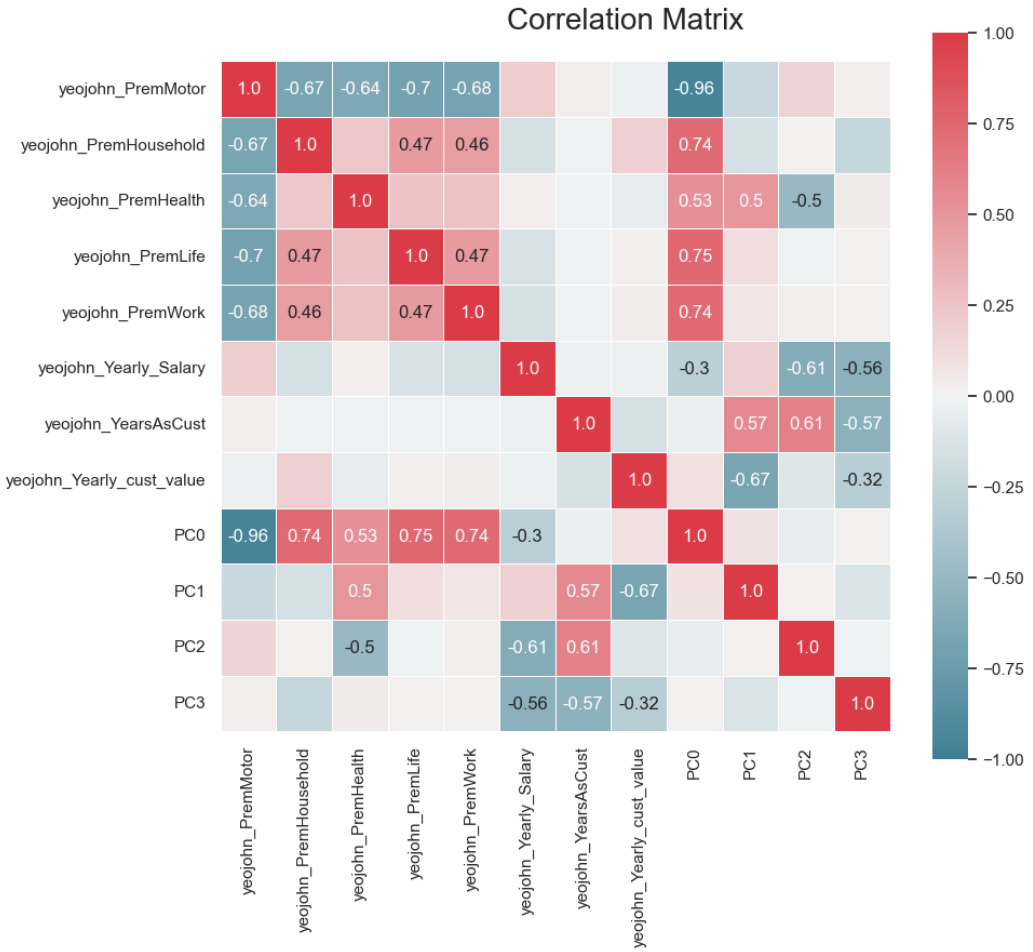

Figure 11 : Corelation Matrix Including Principle Components

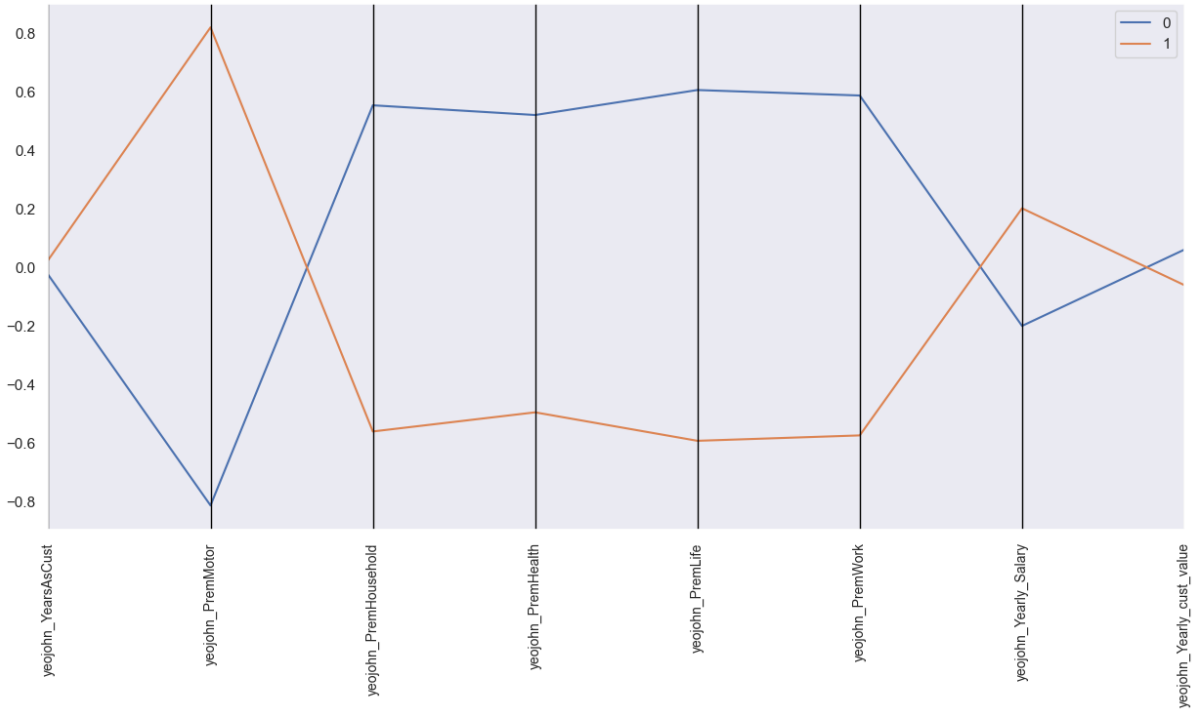

Figure 12: K-means clusters variables distribution

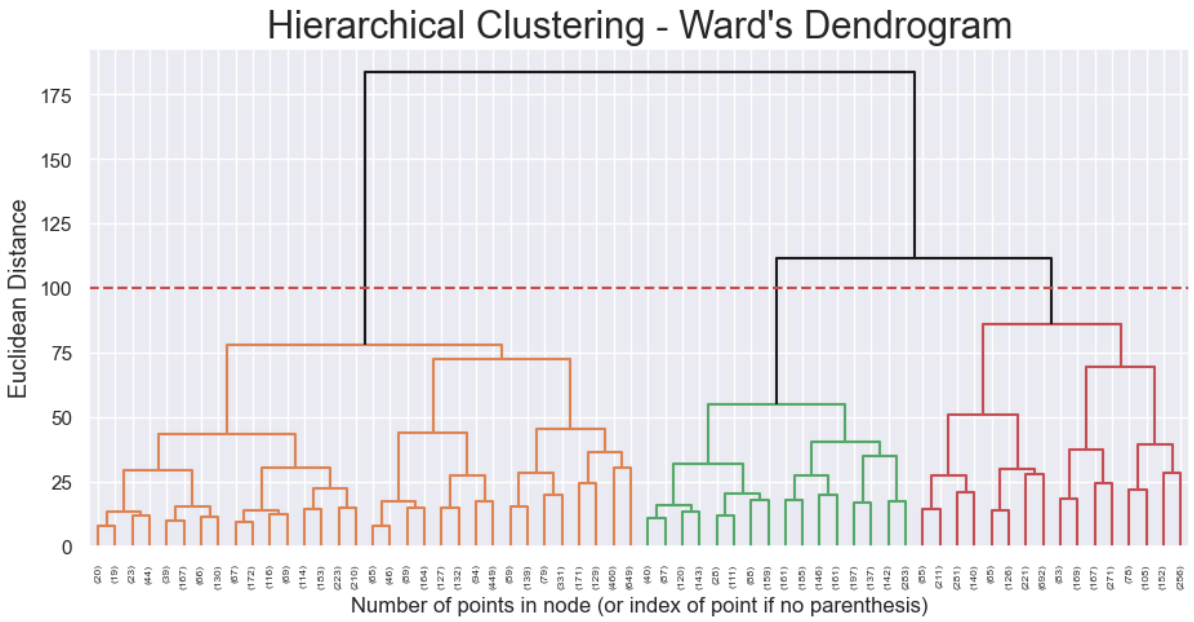

Figure 13: Hierarchical clustering dendrogram

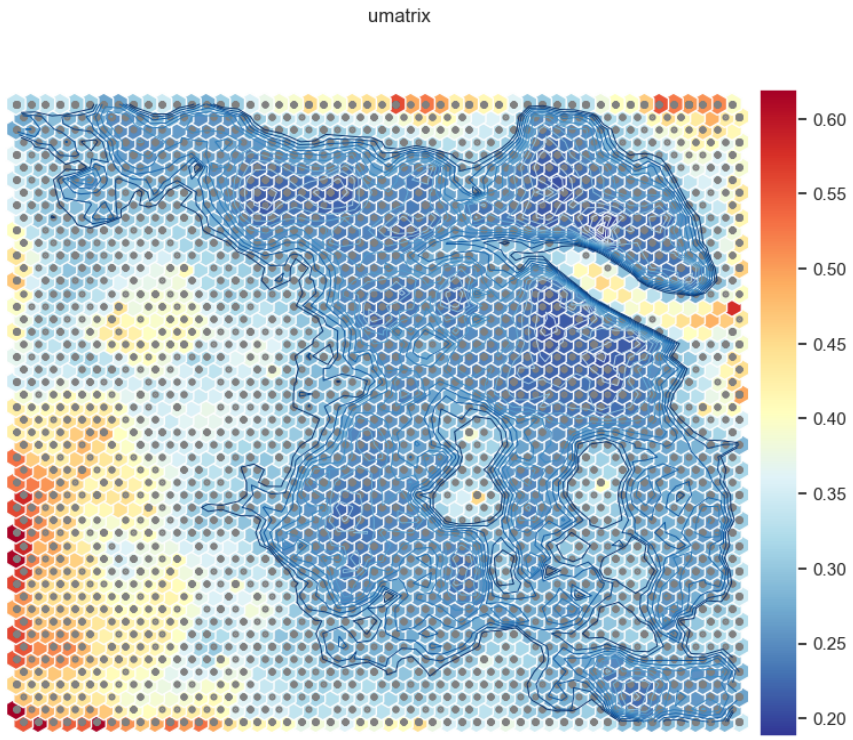

Figure 14: U-matrix of the SOM

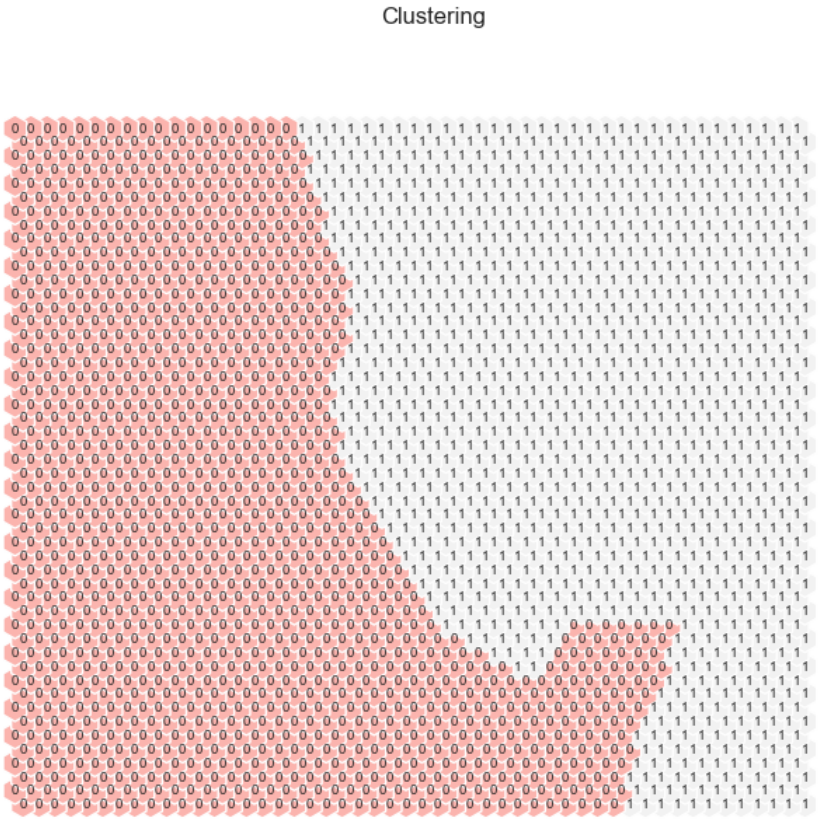

Figure 15: Kmeans on SOM units

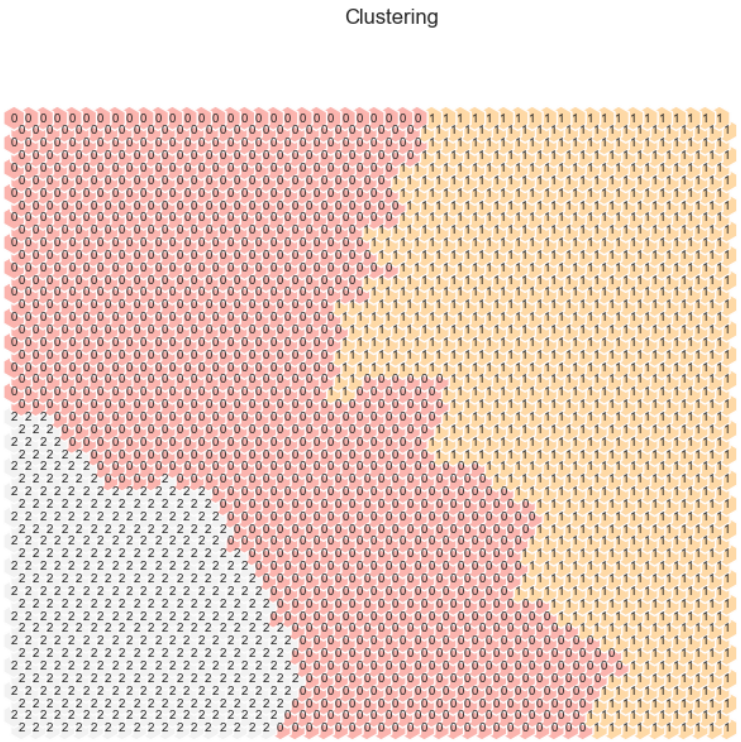

Figure 16: Hierarchical clustering on SOM units

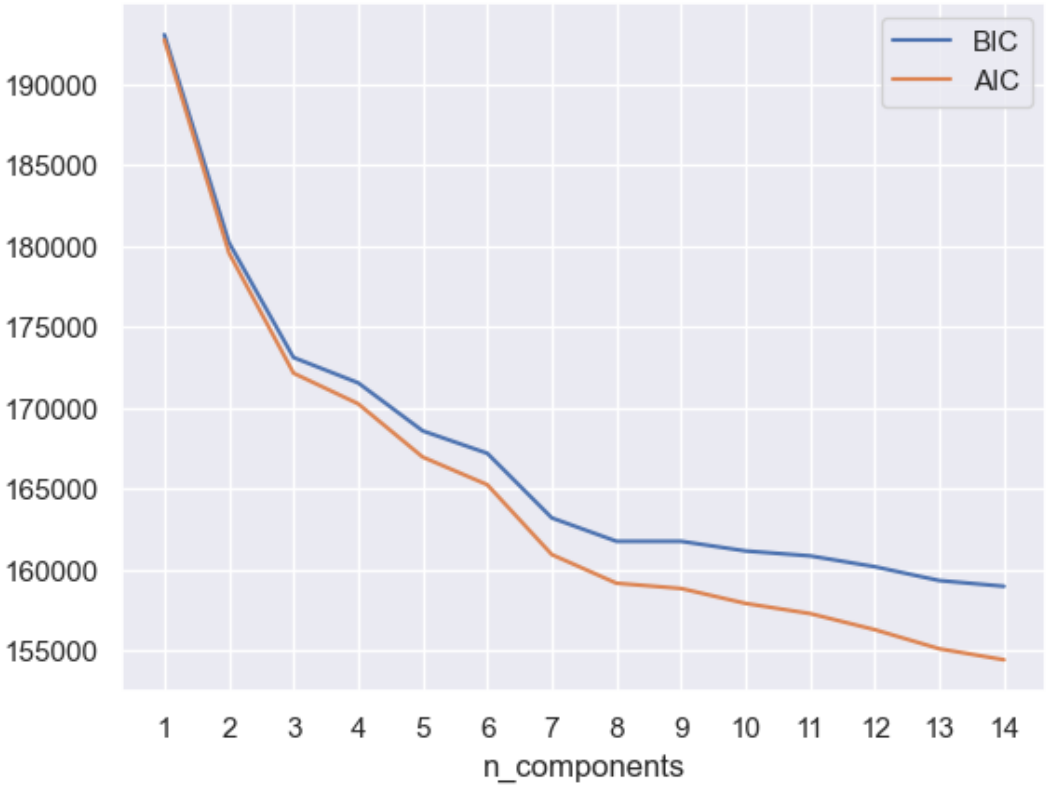

Figure 17: BIC/AIC Curve for the GMM

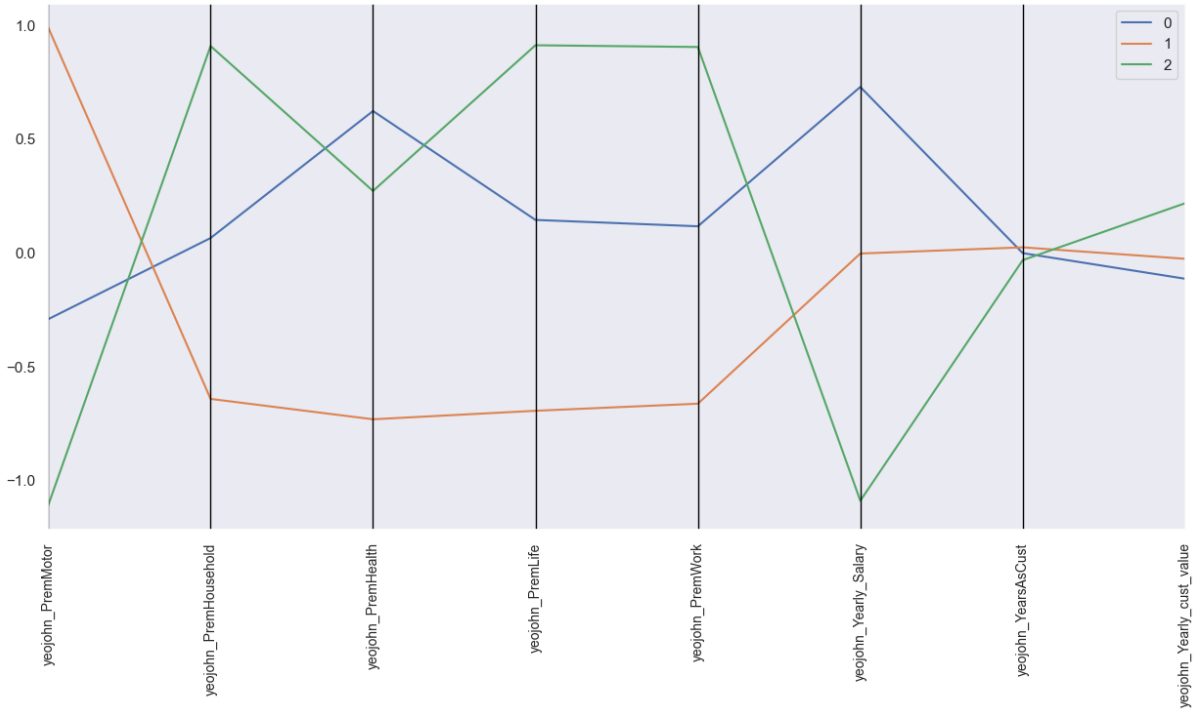

Figure 18: K-prototype clusters variables distribution

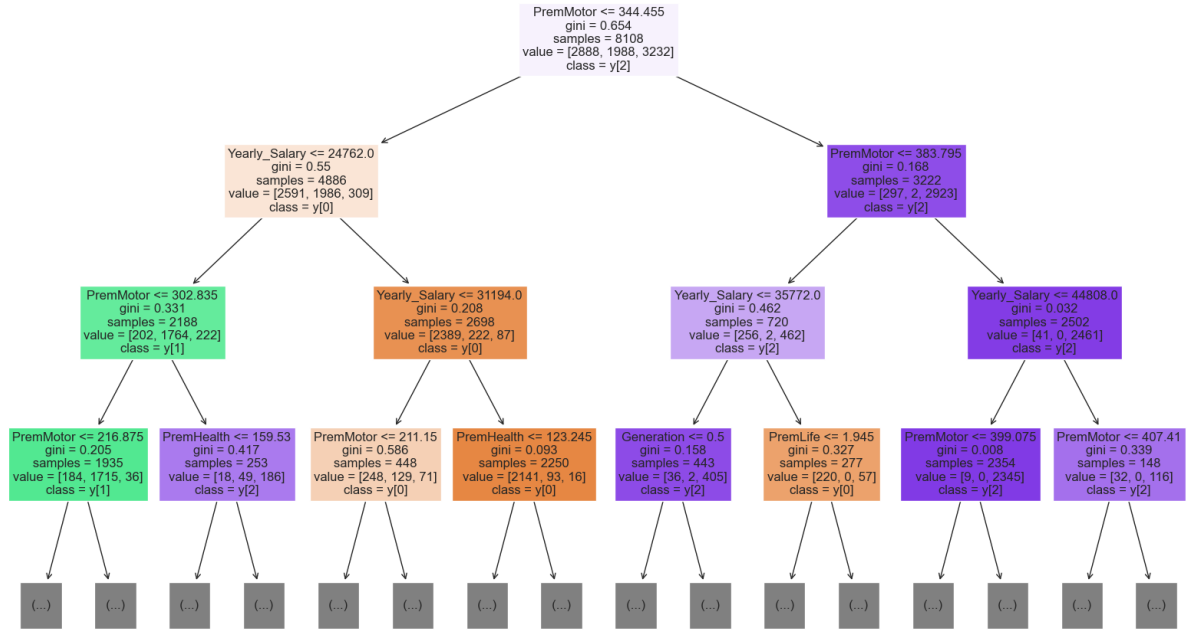

Figure 19: Decision Tree

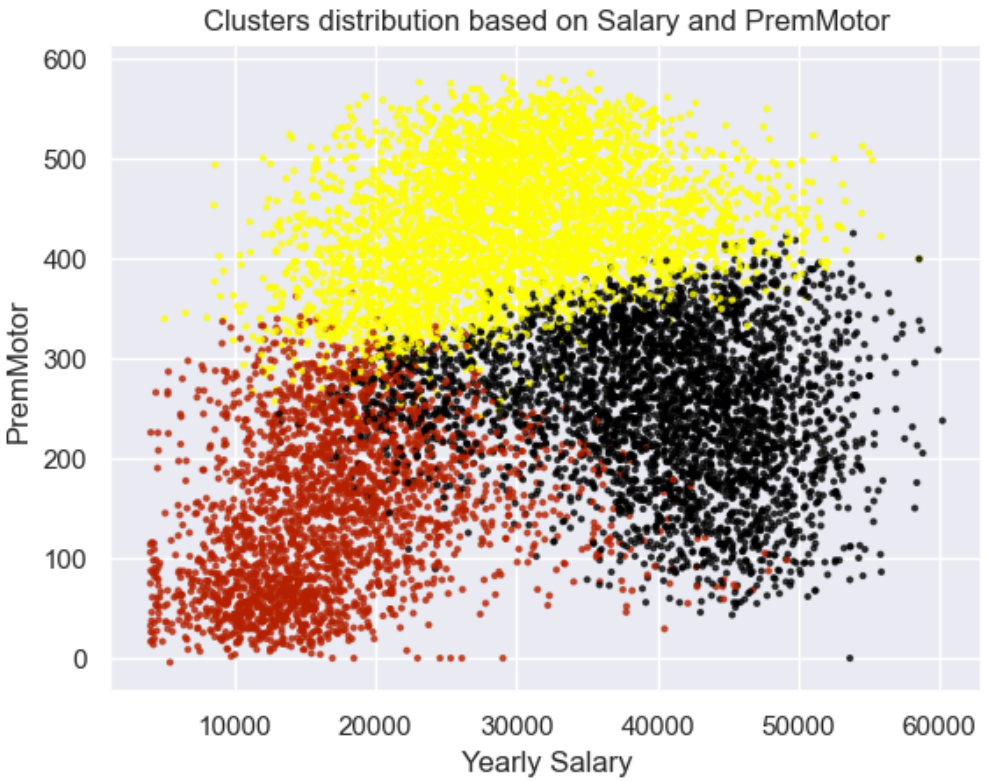

Figure 20: 2-dimensional distribution of the clusters on the most important features

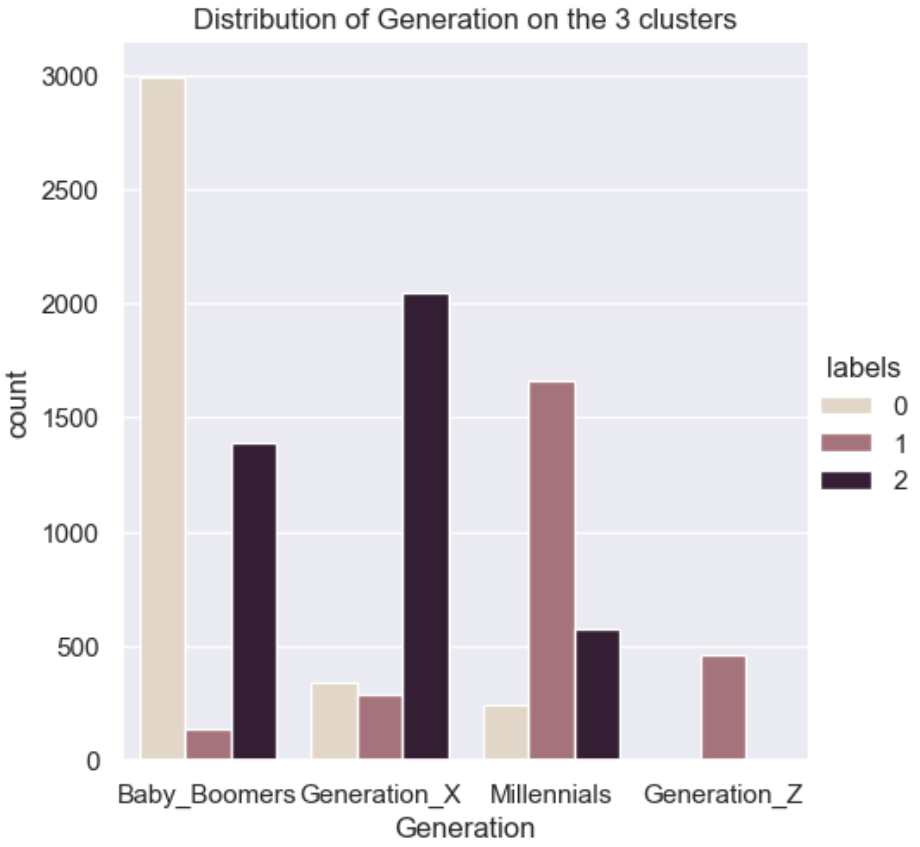

Figure 21: Distribution of Generations between Clusters

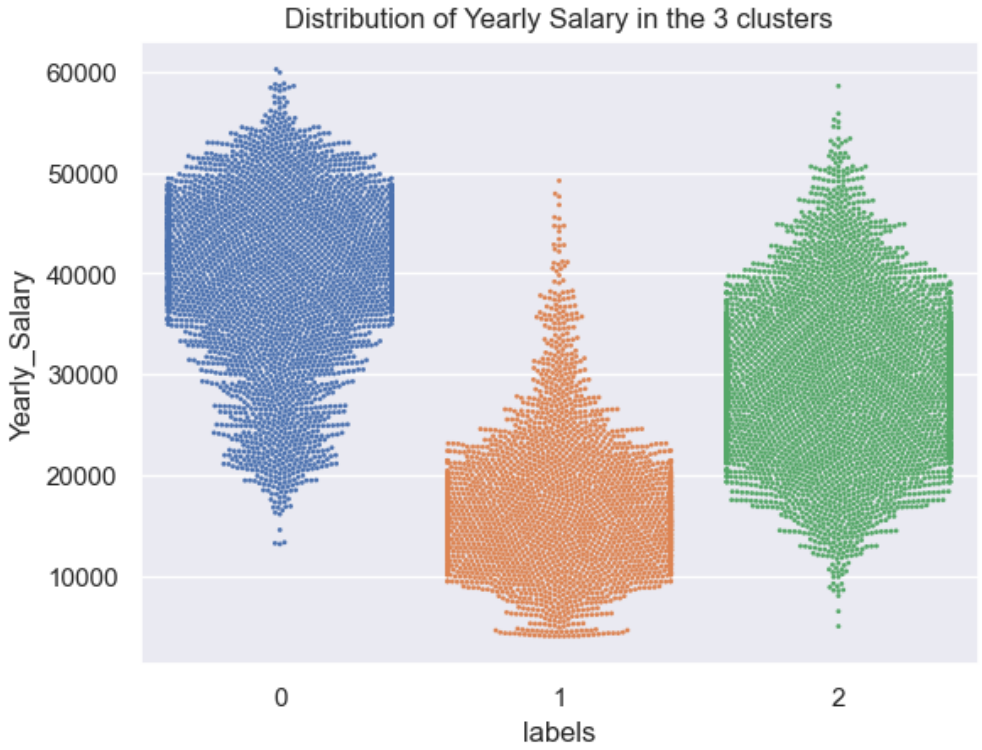

Figure 22: Distribution Of Salries between Clusters

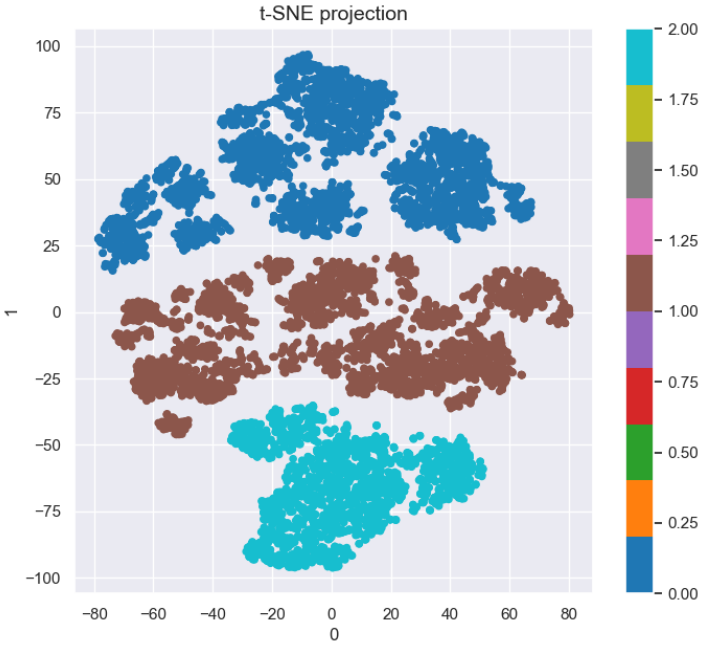

Figure 23: T-SNE Map Visualisation

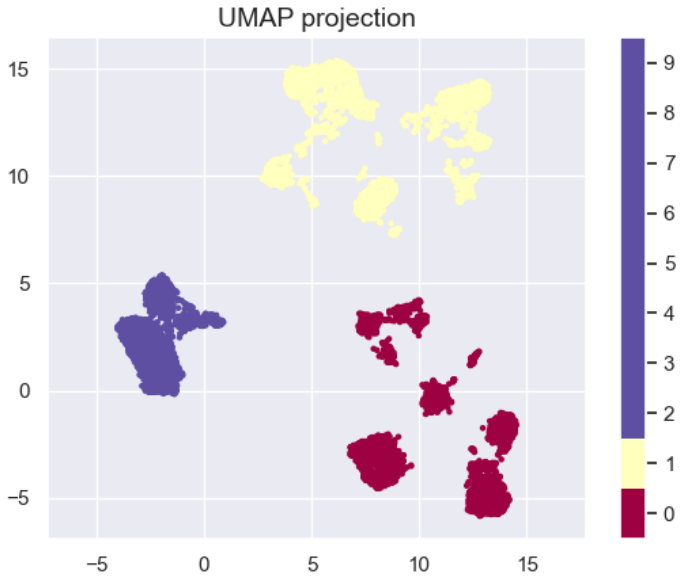

Figure 24: UMAP visualization