Data Mining Project Report

Group 9

Joana Lorenz	Joris Bertens	Luis Riveros	Shabbir Momin
M2080412	M20180423	M20180753	M20180418

1. Introduction

In modern age, companies possess huge amounts of customer data, stored over several decades**,** that hold important insights about customers and their behaviour. Mining these large amounts of data in order to identify customer groups and their specific needs and wants, allows companies to divide their corporate resources more effectively and make better Marketing decisions to ultimately increase profit.

For this reason, a team of Data Analytic Consultants was asked to develop a customer segmentation for an insurance company in Portugal to enable the Marketing Department to better understand the different customer profiles. Two data sets containing information on 10.290 customers were provided as a base for this analysis. The data sets included demographic customer information as well as information on paid premiums in five different lines of business.

In order to develop interpretable customer segments, we, the Data Analytic Consultants, followed a typical Data Mining process: First, the data set was preprocessed in order to assure consistency, accuracy and completeness. This step included the cleaning of the data, handling of missing values and the treatment of outliers. Furthermore, dimensionality was reduced on the data set and data was transformed where this was assumed to lead to a more effective analysis. Next, clustering analyses were performed on the clean data set. After evaluation of multiple approaches, the customers were segmented on their demographic and behavioural information separately and clusters were then joined. The clusters were named comprehensively and specified through their most important characteristics in order for the Marketing Department and other stakeholders to properly understand the segments. Finally, suggestions on the Marketing strategy for each customer segment were given.

2. Project Overview

Multiple decisions had to be taken at different points of the project. For example, some variables were only transformed for the clustering analysis because we wanted to try different approaches. In this report, we grouped all decisions into the appropriate logical categories of the KDD/Data Mining process, in order to facilitate the understanding of the project approach. In the project code, however, these decisions could be located at a later or earlier point.

All concepts and techniques as well as information on their practical implementation were developed based on content from the Data Mining lectures at NOVA IMS in 2018/19 as well as from the book ‘Data Mining - Concepts and Techniques’ by Jiawei Han, Micheline Kamber and Jian Pei (2012). Thus, all theoretical content in this project can be referenced to these two sources.

Before going into more detail on the process, *Figure 1* summarizes decisions take during the preprocessing and cluster analysis. It should give a brief overview of considered alternative approaches, with crossed-out statements standing for options not pursued.

Data Preprocessing	**Data Cleaning**	**Inconsistencies**	**Correction of grammatical errors**	-
			Drop variable Birthday Year	
			Premiums in LOB: Household: convert 0 to NaN	
		Outliers	~~Removal using IQR method~~	
			~~Removal using Z-score~~	
			Removal using visual observation	
		Null Values	**Null Values in Premiums**	**Fill null values with 0**
			Null Values in other variables	~~Impute with global constant~~
				~~Impute with most probable value (KNN)~~
				Drop data points
	Data Transformation	**Attributes**	**Create new variable ‚Total Premiums'**	-
			First Policy Year → Membership Duration	
		Normalization, discretization of continuous variables	**Engage**	~~Binning~~
				~~Min Max Scaler~~
				Standard Scaler
			LOB	~~Standard Scaler~~
				Min Max Scaler
		Transformation of categorical values	**Educational Degree & Has Children**	~~Categorization by Educational Degree & Has Children~~
			Educational Degree	~~One Hot Encoding~~
				~~Transform to numerical values~~
				Categorization by 'low' vs 'high'
	Data Reduction	**Dimensionality Reduction**	**Engage**	~~Geographic Living Area~~ → no relevance
				Educational Degree
				Has Children
				~~Membership Duration~~ → no relevance
				~~Customer Monetary Value~~ → highly correlated
				Total Premiums
				Gross Monthly Salary
				Claims Rate
			LOB	**PCA for Visualization**
		~~Numerosity Reduction~~	**-**	-
Cluster Analysis	~~Combined clustering on Engage and LOB~~	~~KPrototypes~~		
	Separate clustering on Engage (demographic) and LOB (behavioural)	**Engage**	~~Separate clustering on continuous and categorical variables~~	~~KModes on categorical variables~~
				~~All algorithms on continuous variables~~
			Combined clustering on continuous and categorical variables	**KPrototypes**
		LOB	~~Clustering on PCA's~~	-
			Clustering on normalized data	~~Expectation Maximization~~
				~~Mean Shift~~
				~~DB Scan~~
				~~Hierarchical Clustering~~
				KMeans
				-
	Measurement of cluster quality	**Silhouette Score**	-	-
		Visual observation of distributions	-	-

Figure 1: Project Decision Overview

3. Data Preprocessing

Data is only useful if it satisfies certain requirements. There are many factors that comprise the data quality, including accuracy, completeness, consistency, believability and interpretability. During data preprocessing, we transformed the data set in order to make it most useful for the upcoming analyses.

3.1 Data Cleaning

In order to be able to treat data consistently and avoid error or misinterpretation, we pursued data cleaning on both the LOB and ENGAGE table with the objective of a complete, accurate and consistent data set. Through data cleaning, 1.7 % of data points were removed.

Inconsistencies

In order to guarantee consistency in our data, small adjustments such as the correction of spelling mistakes in the variable names were made. Also, we noticed that 'Premiums in LOB: Household' was the only Premium variable that showed 0 instead of NaN values. We assumed this to be an inconsistency and for now transformed the 0 values to NaN in order to handle missing values uniformly for all Premiums later on.

Next, we noticed that in ~20 % of cases, First Policy Year was an earlier date than the Birthday Year, which is logically impossible. Thus, we felt that we could not trust the variables and needed to adjust or remove one or both of them. We considered imputing the variables for the affected data points. However, without any proof of the trustworthiness of any of the two variables for all other data points, we feared this would only introduce further error to the data set.

We then decided to drop one of the variables and treat the remaining one with caution. We decided to drop Birthday Year based on two factors. First, its strong correlation with Gross Monthly Salary, since correlated variables would not be useful in our later analysis in any case and they both seemed to explain the same characteristic. Second, we assumed that First Policy Year was a more reliable variable since it is most probably given by the system itself, whereas Birthday Year is provided manually and may thus be entered wrongly on purpose or based on human error.

Outliers

Next, we took steps in order to remove potential outliers from our data set. While outliers can be highly interesting for certain analyses, the aim for the customer segmentation was to find representative groups of people within the great majority of the customers, rather than observing the exceptional cases. Furthermore, since some clustering algorithms are highly influenced by outliers, we decided to trade a small percentage of deviating data points (outliers) for a more accurate and reliable analysis.

One important consideration we took was the distinction between outliers and extreme values. While outliers are usually points that strongly deviate from the rest of the data and follow a different underlying structure, extreme values rather represent the minimum and maximum ranges of the data points along a certain variable. We thus considered them to be relevant to our analysis.

We considered three options for outlier removal: the IQR and Z-score methods, which are based on the relationship of data points to all other data points, as well as 'manual' removal by visual observation of the distribution. The IQR method, which removes outliers based on a distance threshold to the first or third quartile, would have removed a large amount of our data and may have sacrificed extreme values. The Z-score method removes data points based on their relationship to the mean and standard deviation of the data set. This way, we removed much less outliers than with the IQR method, but we still felt that extreme values were overly removed.

Thus, we decided to remove outliers based on visual observation of the distribution in each variable in order to be sure to keep extreme values within the data and only remove true outliers. An example of this procedure can be seen in *Figure 2* and *Figure 3.* During this process we found some obvious errors, represented by illogical values such as a membership duration of -50 000, as well as values that could be true, but deviated so strongly from all others that they would have skewed later models. After removing outliers based on our consideration, we had decreased the amount of data points by 1.4 %, which we considered acceptable.

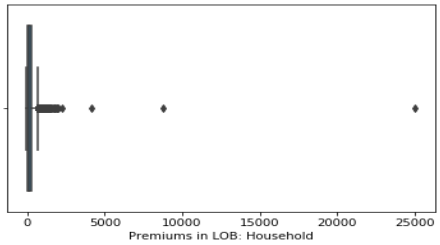

Figure 2: Premiums in LOB Household before outlier removal

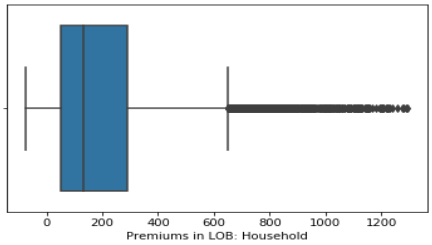

Figure 3: Premiums in LOB: Household after outlier removal

Missing Values

Next, in order to guarantee a complete data set and since null values could not be processed by the clustering algorithms, we treated missing values in our data. We considered two different kinds of null values. The null values in all five premiums were considered to mean that a customer really does not pay any premiums in this specific line of business. This assumption is based on the fact that this information must be known to the insurance company and thus, NaN values must mean 'inexistent'/'no payments', instead of 'unkown'. Thus, these null values were set to 0.

Customer Identity	0
Premiums in LOB: Motor	34
Premiums in LOB: Household	60
Premiums in LOB: Health	43
Premiums in LOB: Life	104
Premiums in LOB: Work Compensations	86
Educational Degree	17
Gross Monthly Salary	36
Geographic Living Area	1
Has Children (Y=1)	21
Customer Monetary Value	0
Claims Rate	0
Membership Duration	30

Figure 4: Missing values in all variables before treatment

We then considered all other null values - information that each customer inherently possesses and can thus only be missing by error or fault. Most probably, this information was not provided by the customers or was overseen during the sign-up process. Options we considered for these data points were the imputing with measures of central tendency, such as the mean or median, or using the most probable value to fill in the missing fields, e.g. by using the k-nearest neighbor approximator.

Generally, the imputing of values is preferred over deletion, in order not to lose information. However, we found very few null values in relation to the data set size, the maximum count of all null values for a single variable being 36 (~0.33% of the data points), after we converted the null values in the LOBs into 0. Since all described methods for null value removal can influence and bias the data, we decided to take an efficient approach and remove the data points with null values. This way, we renounced a few data points but could be sure that our data set would not be distorted through imputation.

3.2 Data Transformation

Attribute Transformation

During data transformation, we constructed a new variable that we found interesting: Total Premiums, which represents the sum of all Premiums in LOB per customer. We found this measure useful, as it is one of the measures that represents the value of each customer for the insurance company.

Then, we adjusted First Policy Year to Membership Year by deducting the First Policy Year from the Year the data stems from (2016). The rationale here was that minimum and maximum values of First Policy Year would be very close together since the thousand and hundred values are the same or very close together for each data point. Re-calculating the variable to Membership Year shows the true value difference between customers.

Value Transformation

We considered adjustments that could be made to the categorical variable Educational Degree. This attribute was specifically considered since it is ordinal with multiple potential values. We saw a risk of the algorithm not properly valuing the rank among the four options and wanted to make sure that the variable was correctly interpreted by the algorithm and results could be properly analysed.

Since values of the Educational Degree are ordinal, i.e. have an order, we did not consider to use one-hot encoding, which diminishes any rank. We considered transforming the four categories to numeric values (0,⅓,⅔,1). However, the distance between degrees is not necessarily equal. Thus, instead we made use of concept hierarchies by generalizing the Educational Degrees to 'low education' (Basic, High School) and 'high education' (B.Sc./M.Sc., PhD.). This way, we had a binary variable with two relatively equal sized groups, which we found more useful for the clustering process.

For continuous variables in Engage, we considered discretization through binning, i.e. assigning data points into different categories, based on the range the variable value falls into. This could transform variables Gross Monthly Salary, Claims Rate and Total Premiums into categorical variables with values such as 'low', 'medium', 'high' and would lead to a uniform-type clustering on the Engage variables. However, binning would have led to a (significant) information loss, with the gravity depending on the amount of categories chosen for each data point. Thus, we refrained from binning continuous variables in Engage.

Normalization

Next, since each variable has a different range of possible values and variables with larger ranges and higher absolute values could bias the clustering algorithms, we decided to normalize our data. Options we considered were the Standard Scaler (Z-Scaler) and the MinMax Scaler. The former normalizes data based on the mean and standard deviation along each variable, the latter linearly transforms data to a smaller range [0,1].

As opposed to the Standard Scaler, the MinMax Scaler does not assume any particular distribution of the data and preserves the original relationship among data points. However, an advantage of the Standard Scaler is, that it takes the dispersion of the data into account, which is something that we are interested in during clustering. For the behavioural customer data (LOB) we found the MinMax Scaler to give us more clear, interpretable clusters for the different clustering algorithms. For the demographic clustering (Engage), however, the StandardScaler worked better. Thus, we used the two different scaling methods on our LOB ang Engage data.

3.3 Data Reduction

Dimensionality Reduction

We did not pursue numerosity reduction on our efficiently-sized data sets, but performed dimensionality reduction in order to avoid the curse of dimensionality and an inappropriate bias, e.g. towards correlated variables, during clustering. In order to remove redundant or irrelevant variables, we took into account the correlation coefficients, the importance of each feature in the Principal Components and the difference in clustering results when variables were added or removed from the clustering algorithms.

Engage
Geographic Living Area and Membership Duration did not seem to have any impact on clustering and were equally distributed over values of all other variables. Thus, we did not assume these variables to be of importance for the customer segmentation and excluded them from the clustering. Additionally, we excluded Customer Monetary Value, due to its strong correlation with Claims Rate, since the inclusion of variables with strong correlation and thus essentially the same meaning, can give biased clustering results.

LOB
Since we were only given five dimensions for customer behaviour, we were highly interested in keeping the majority, in the best case all, variables. Some correlations could be found, especially between Premiums in LOB: Motor and the other Premiums. However, none of these seemed strong enough to support the decision of fully dropping a variable. Principal Component Analysis (PCA) could transfer the variables into an uncorrelated state and reduce dimensionality but could also lead to a partial loss of the original variation within the data. Since we did not have too many variables that needed to be reduced for the cluster analysis, we performed PCA on all five normalized Premiums mainly in order to visualize the clustering results in a three-dimensional space. However, we also used the three first PCA's during clustering in order to see the difference of clustering on PCA's versus clustering on the five normalized variables. The results will be covered in the next section.

4. Cluster Analysis

For the clustering analysis, we first considered clustering on the joined data set of LOB and Engage using KPrototypes. However, this approach could lead to a loss of insights into the structure of the customer base, since each demographically similar group (Engage) of customers could be divided into subgroups, that show different insurance behaviour (LOB). For this reason a separate clustering of a) Engage and b) LOB was chosen to be most applicable. Afterwards these were merged via a cross tabulation.

Measuring of cluster quality

In order to differentiate an appropriate from an inappropriate clustering, we used three methods to evaluate each resulting segmentation:

1. *Visually, subjectively observing the resulting distribution* of the defining variables per cluster. Our aim was to achieve a clear separation of the achieved clusters over as many variables as possible.
2. *Observing cluster centroids*. A good segmentation would be defined by differing cluster centers/mean values. If two clusters show very similar means across many variables, it can be concluded that they are too close together and should not represent different segments
3. *Analysis of the Silhouette Score*. The Silhouette analysis can be used to study the separation distance between the resulting clusters and is included in the sci-kit learn package. The Silhouette Score measure has a range of [-1,1], where an average value close to +1 means the samples are far away from the neighbouring clusters, as is desired. The Silhouette Graph shows the average Silhouette Score (Y-Axis) for different numbers of clusters (X-Axis).

4.1 Engage

For the clustering of the demographical customer information, we considered clustering on continuous and categorical variables separately and then combine results using a cross table OR cluster on both continuous and categorical variables simultaneously.

1. Clustering on continuous and categorical separately

The advantages we saw in this approach were the possible exploitation of all available clustering algorithms for continuous variables: Hierarchical, KMeans, Mean-Shift, Expectation Maximization and DBScan. For the categorical variables, we used KModes. Results were combined using a cross table.

However, after reducing dimensionality on the Engage data, we only had few variables left, especially for the categorical clustering, which means that clusters were rather a very simple categorization among the two variables, than a true clustering. Furthermore, clustering only on parts of the data that all together represent a customer, did not seem appropriate and we felt we were not using all available information in an appropriate combination during clustering. Lastly, the interpretation of the cross table is never trivial and can lead to misinterpretation or an inappropriate joining of adjacent groups, which is a source of error we would rather avoid, if possible.

2. Clustering on continuous and categorical variables using KPrototypes

This approach allowed us to make full use of the breadth of our data by considering both continuous and categorical variables in KPrototypes - a clustering process that combines KMeans and KModes and can take raw categorical values as an input. In order to find the appropriate number of clusters, we used the KMeans Elbow Graph, that shows within-cluster variance for different numbers of clusters, and the Silhouette Score graph as described above. The optimum amount of clusters should exhibit a Silhouette Score as high as possible and a 'Elbow' in the Elbow Graph – signifying a stagnating decrease in within-cluster variance. In this case, we found this point at four clusters.

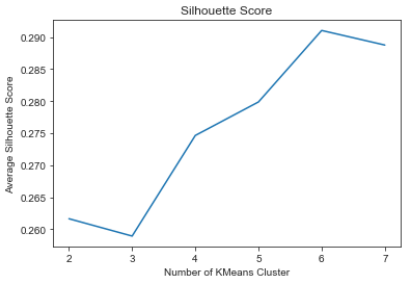

Figure 5: Silhouette Analysis

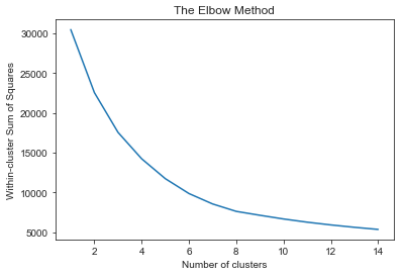

Figure 6: Elbow Graph

When clustering with 3,4 and 5 clusters in order to find the final number of clusters, the distribution of variables in the centroids changed as follows: Using four clusters, which was the most probable optimal amount, we received two groups that were similar on all characteristics, except for Total Premiums and a slightly differing claims rate. Otherwise, all cluster centroids were very diverse on all variables. The distribution plots in F*igure 7* show that clusters that behave similarly on a specific variable were different on others, which is a desired outcome.

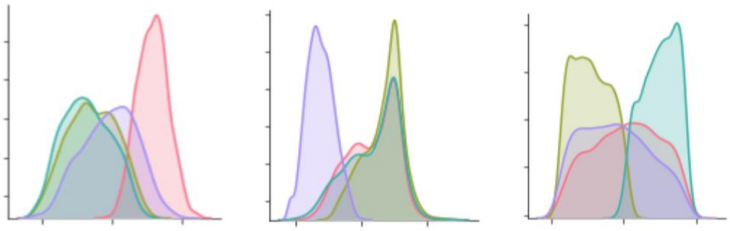

Figure 7: Distributions of Gross Monthly Salary, Claims Rate, Total Premiums along 4 clusters

	Gross Monthly Salary	Claims Rate	Total Premiums	Has Children	Educational Degree	Cluster Size
0	3650.631970	0.793751	6133.614613	No Children	High Education	2690
1	1953.256007	0.899921	3268.259541	Has Children	Low Education	2289
2	1801.305219	0.800113	8868.285081	Has Children	Low Education	2395
3	2445.195613	0.286030	5293.521719	Has Children	High Education	2781

Figure 8: Centroid values for 4 clusters

Using three clusters, there was no separation in distributions of Total Premiums over the different clusters, whatsoever.

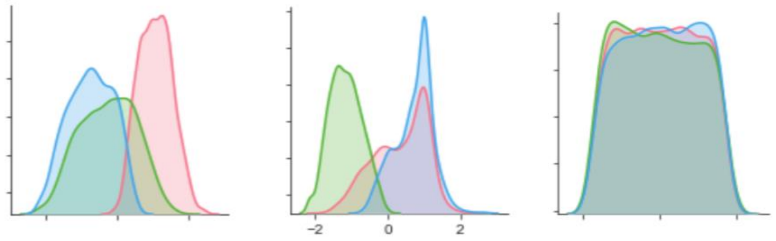

Figure 9: Distributions of Gross Monthly Salary, Claims Rate, Total Premiums along 3 clusters

	Gross Monthly Salary	Claims Rate	Total Premiums	Has Children	Educational Degree	Cluster Size
0	3571.000000	0.793214	5889.527040	No Children	High Education	3189
1	2265.486862	0.313093	5789.642262	Has Children	High Education	3311
2	1782.892476	0.913986	6016.440167	Has Children	Low Education	3655

Figure 10: Centroid values for 3 clusters

Using five clusters, previous segments seemed to be broken down into more subgroups with differing Total Premiums (Cluster 3,4). However, the segmentation of a cluster with a low Claims Rate into a higher and lower Total Premiums group did not make much sense to us. The Total Premiums variable is very helpful in cases where Claims Rate is high. This is because a high Claims Rate combined with a high total spend in Premiums could still make an overall profitable customer. With a low Claims Rate, however, we did not feel the need to further divide a cluster.

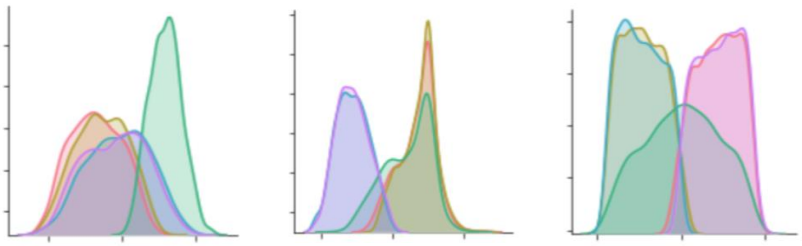

Figure 11: Distributions of Gross Monthly Salary, Claims Rate, Total Premiums

	Gross Monthly Salary	Claims Rate	Total Premiums	Has Children	Educational Degree	Cluster Size
0	1832.607212	0.907563	8504.208606	Has Children	Low Education	2052
1	1978.430279	0.927221	3239.568456	Has Children	High Education	2008
2	3675.875313	0.820513	6036.897110	No Children	High Education	2398
3	2445.338616	0.327512	3158.403180	Has Children	High Education	1893
4	2343.927384	0.329867	8608.797151	Has Children	High Education	1804

Figure 12: Centroid values for 5 clusters

Thus, since we had already achieved a separation of profitable customers with high Claims Rate and much less profitable customers with high Claims Rate and we further achieved a good separation along all desired variables, we ultimately chose four clusters for KPrototypes.

4.2 LOB

For the behavioural segmentation (LOB), we considered five clustering algorithms for continuous variables: KMeans, Hierarchical Agglomerative, DBScan, Mean-Shift and Expectation-Maximization clustering.

The least desirable clustering results were achieved using DBScan and Mean-Shift clustering. For DBScan, the most crucial factor for the success of the clustering algorithm is a good estimation of *epsilon* and the *MinPts*, which define the radius and minimum number of points within that radius in order for a point to be a core point within a dense region. For Mean Shift, the parameter to be defined is the bandwidth *h*, defining the neighborhood of each data point. For DBScan, we used the K-Nearest-Neighbor algorithm in order to see the average distance of the *k* nearest neighbors (Y-axis) to any point in the data set (X-axis), ordered by decreasing average distance. The 'Elbow' of the resulting graph defines a theoretically good epsilon for the given *K*. For Mean-Shift, we used the bandwidth estimator to estimate an appropriate *h*. Additionally, we tried ranges of the parameters around the estimated 'optimals'.

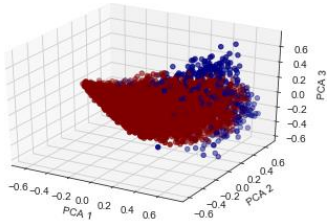

Figure 13: Clustering Result using DBScan

However, both algorithms failed to find useful cluster: in the best case, DBScan found one large cluster and noise, while Mean Shift identified three clusters of greatly differing size. We believe this clustering result is due to the fact that both algorithms are based on the density of data points in space: DBScan partitions the data points into clusters of dense regions with less dense regions in between, while Mean-Shift gradually moves the centroid to the most dense point in a region and identifies clusters around it. Looking at the PCA representation of our data, the given data points seem to be very close together in space along the most important vectors. Thus, it seems very hard or even impossible to separate meaningful clusters based on density.

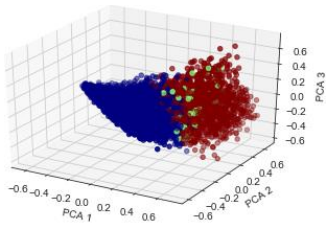

Figure 14: Clustering Result using Mean-Shift

KMeans, Expectation Maximization and Hierarchical Clustering all required the specification of the *K* expected clusters. As for Engage clustering, in order to find an appropriate number of clusters, we used the KMeans Elbow Plot, the Silhouette Score Graph as well as the Dendogram based on Hierarchical Clustering. Even though these methods are usually specific for a certain clustering algorithm, we assumed that the resulting number of clusters would be generally suited for the data set. All three methods tended towards 3 clusters for LOB. Specifically the Silhouette Score seemed to derease strongly above this number of clusters. However, we still tried clustering on k = [3,4,5] in order to see which segmentation was the cleanest and made most sense to us.

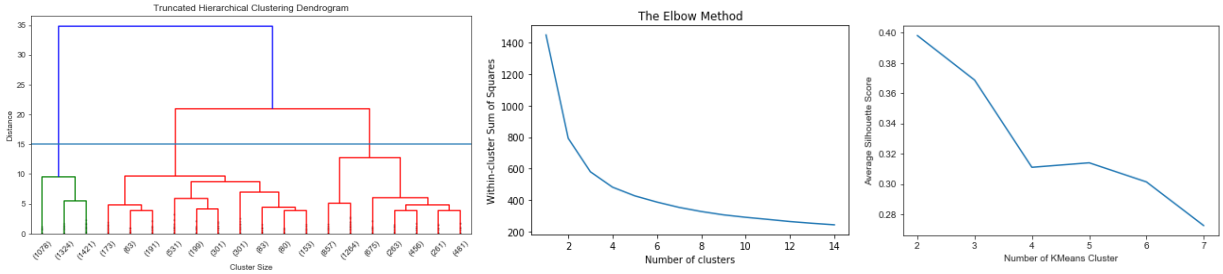

Figure 15: Dendogram, Elbow Graph, Silhouette Score Graph

KMeans, Hierarchical Clustering and Expectation Maximization clustering all returned clusters with similar tendencies in the centroids - meaning a cluster with high Motor expenses, a cluster with high Health expenses and a cluster with high Household, Life and Work expenses. When clustering with

more than three clusters, we received another cluster with high Motor expenses, but very similar tendencies to the other Motor-related cluster. Clustering with five clusters resulted in even less useful segments. Since we were more interested in having groups that are clearly separable by their general behaviour than having exact centroid values for each subgroup, we considered three clusters to be appropriate.

	Premiums in LOB: Motor	Premiums in LOB: Household	Premiums in LOB: Health	Premiums in LOB: Life	Premiums in LOB: Work Compensations	N Individuals
0	433.753498	73.925981	108.968469	15.333924	14.708629	4123
1	241.483401	182.128924	226.614125	40.744761	40.413672	4202
2	114.407292	545.398386	164.044825	101.486849	98.810656	1830

Figure 16: KMeans centroids using 3 clusters

	Premiums in LOB: Motor	Premiums in LOB: Household	Premiums in LOB: Health	Premiums in LOB: Life	Premiums in LOB: Work Compensations	N Individuals
0	322.487831	151.026951	170.918990	31.656779	31.104077	3432
1	183.925345	215.089055	262.428727	48.721118	48.252534	2382
2	467.443670	48.977953	89.775145	10.007116	10.157681	2760
3	112.808166	570.016698	153.037894	106.020474	102.057230	1581

Figure 17: KMeans centroids using 5 clusters

Out of the three clustering methods, Expectation Maximization had the lowest Silhouette Score and returned one cluster of a much smaller size than the other two. Hierarchical Clustering and KMeans returned very similar results, with KMeans still showing the highest Silhouette Score as well as, in our eyes, the most clear separation of the clusters along the different variables. Hierarchical clustering is generally especially useful when a natural hierarchy of the data is assumed and should be represented in the clustering. Since we did not assume such a structure of higher- and lower-level groups within our behavioural data, we did not depend on hierarchical clustering. Thus, based on the named factors, we chose KMeans clustering with K=3. In order to validate our choice with another objective measure, we also calculated the Calinski Harabaz score, which calculates the ratio between within-cluster and between-cluster dispersion and also showed the highest (=best) value for Kmeans.

For comparison reasons, we also tried clustering on PCA's in KMeans Clustering. The resulting clusters were very similar but since we feared we could loose variation within the data when using PCA's on a already low amount of variables, we decided to stick to using the normalized data without PCA analysis.

5. Resulting Clusters

Out of the different methods used for clustering the KMeans for LOB and K-Prototypes for Engage were chosen to be the most useful ways of clustering. Final clusters were created by combining the two of them in a cross table.

5.1 Engage

Clustering on Engage resulted in four clusters, that were named and characterized in the following. In order to estimate the profitability of each cluster we calculated Estimated Net Profit using (Total Premiums) * (1-Claims Rate). This does not take into account any additional Marketing or other costs that the insurance company has for an individual customer but gives a rough indication on the overall profitability of the customer segment.

	Gross Monthly Salary	Claims Rate	Total Premiums	Has Children	Educational Degree	Cluster Size	Cluster Name	Est. Net Profit
0	3650.631970	0.793751	6133.614613	No Children	High Education	2690	Wealthy, childless	1265.052314
1	1953.256007	0.899921	3268.259541	Has Children	Low Education	2289	Unprofitable lower class	327.082960
2	1801.305219	0.800113	8868.285081	Has Children	Low Education	2395	Lower class families	1772.657251
3	2445.195613	0.286030	5293.521719	Has Children	High Education	2781	Profitable families	3779.414617

Figure 19: Final Engage cluster centroids

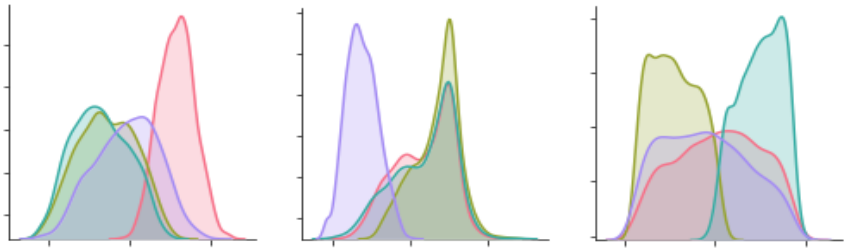

Figure 18: Distribution of Gross Monthly Salary, Claims Rate, Total Premiums for each cluster along 4 clusters

Cluster #0 of ENGAGE *'Wealthy and Childless'* (RED):

High educated	High claims rate	**No** Children	High income	Middle total Premiums

The individuals in this cluster are highly educated, childless and have a much higher income than the others. Given the correlation between age and salary, these people are probably one of the oldest segments and given the high income, could be particularly career-focused. Relatively high paid premiums are offset by a relatively high Claims Rate, making them one of the less profitable clusters for the insurance company.

Cluster #1 of ENGAGE '*Unprofitable Lower Class*' (Light GREEN):

Low educated	High claims rate	Children	Low income	Low total Premiums

This cluster represents low income families. Expected is that those people are more in labour jobs and are raising their children in a less stable financial situation. They show the highest mean Claims Rate together with the lowest total spend on Premiums, which makes them by far the least profitable cluster for the insurance company.

Cluster #2 of ENGAGE '*Lower Class Families*' (BLUE):

Low educated	Middle claims rate	Children	Low income	High total Premiums

This cluster demographically holds similar individuals as cluster #2. However, in comparison to the previous cluster, these individuals seem to be much more safety-oriented, with the highest amount of paid premiums. In combination with a lower claims rate than cluster #2, the lower class families are one of the most profitable customer segments.

Cluster #3 of ENGAGE '*Profitable Families*' (PURPLE):

High educated	Low claims rate	Children	Middle income	Middle total Premiums

This cluster consists of people with a high education, a middle income and with children. It seems to be the cluster of responsible middle class families, who pay a good amount of premiums and show **very** low claims, making them by far the most valuable cluster for the insurance company.

5.1 LOB

Clustering on LOB variables using Kmeans resulted in three clusters as described in the following.

	Premiums in LOB: Motor	Premiums in LOB: Household	Premiums in LOB: Health	Premiums in LOB: Life	Premiums in LOB: Work Compensations	Cluster Size	Cluster Name
0	433.753498	73.925981	108.968469	15.333924	14.708629	4123	Motor Fans
1	241.483401	182.128924	226.614125	40.744761	40.413672	4202	Homey, safe
2	114.407292	545.398386	164.044825	101.486849	98.810656	1830	Health first

Figure 20: Final LOB cluster centroids

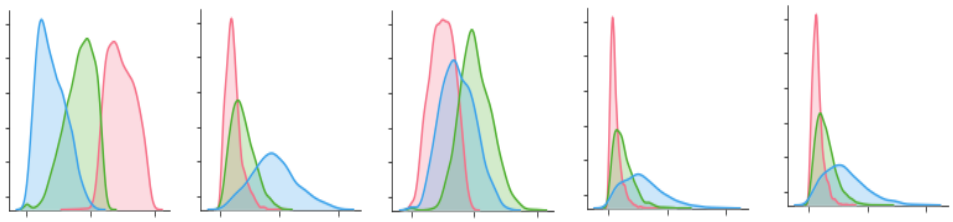

Figure 21: Premiums in LOB: Motor, Household, Health, Life, Work Compensations

Cluster 0 of LOB '*Motor Fans*' (RED):

High Motor	Low Household	Low Health	Low Life	Low Work Comp.

These people seem to spend the lowest amount of money on all LOB's except for Motor, where they have the highest spend. These could either be people who have a lot of accidents, being forced to high premiums or people with very expensive cars that need to be secured through high premiums.

Cluster 1 of LOB '*Health First*' (GREEN):

Middle Motor	Middle Household	High Health	Middle Life	Middle Work Comp.

Individuals in this cluster generally spend middle amounts on all business lines but exhibit a strong focus on health expenses. These could be people that put a strong focus on their health by getting a good insurance with additional benefits and/or people that need good health insurance because of repeated or severe illness.

Cluster 2 of LOB '*Homey, safe*' (BLUE):

Low Motor	Very High Household	Middle Health	High Life	High Work Comp.

A much smaller cluster, these people are generally very well insured, especially on the premiums Household, Life and Work Compensations. This could indicate a focus on family and rather working class individuals that need to be well insured on Life and Work Compensations.

5.3 Combining Clusters

Combining the ENGAGE and LOB clusters results in twelve different subclusters, which we considered too many to form meaningful customer profiles. In order to reduce clusters, we combined customer segments that we found could logically be grouped. The ENGAGE clusters were used as a baseline for customer profiles in order to see how these different people tended to behave in their LOB premiums spendings. Similar clusters or clusters of specific interest were then merged to form six final clusters.

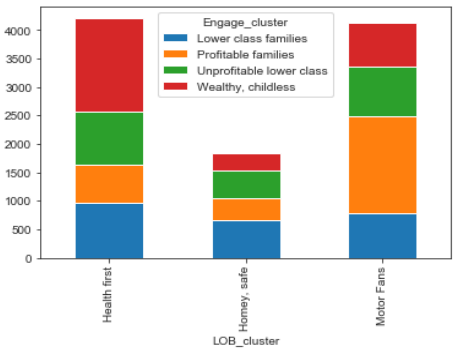

Figure 22: Distribution of Engage clusters within LOB clusters

Merging Process

Three main observations helped us in the merging process:

1. Wealthy, childless individuals are demographically much different from all other clusters.
2. 'Unprofitable families' are much more unprofitable for the insurance company than all other clusters.
3. Lower class families and Midd families seem to behave much similarly within the LOB's.

First, we created a single cluster for all unprofitable lower class individuals (1). Even though they exhibit much similar behaviour to the ‘Lower class families’, as could be expected from the similar demographic profile, we did not want to group them with other profitable groups of customers. This is because these people might actually cost the insurance company money and thus should be treated separately from all customers that are valuable to the company.

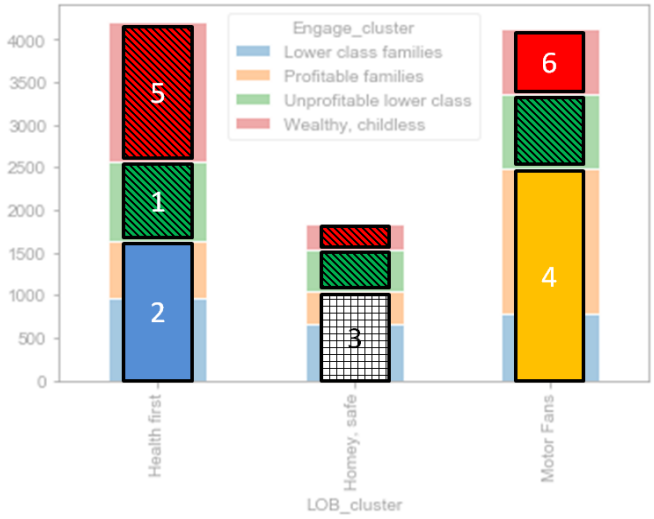

Figure 23: Merging process based on the cross tabulation

Next, ‘Lower class families’ and profitable families were grouped within all LOB clusters (2,3,4). These two groups are both profitable, similar in their Claims Rate, have families and only differ much on their income and education, which is why we think they can be grouped for Marketing purposes. It could generally be observed that the higher education, higher income families tended towards the ‘Motor Fans’ LOB cluster, while the ‘Lower class families’ tended more towards the ‘Health first’ cluster. This could be based on lower income families seeking more security in health insurance since there is less financial back-up in case of an emergency.

Next, we found wealthy, childless people to be mostly present in the ‘Health first’ LOB segment (5). To this cluster we decided to add the group of wealthy, childless people in the ‘Homey, safe’ LOB segment. We reasoned that his cluster differs from the other clusters in the ‘Homey, safe’ segment and as it is a small group of individuals, we considered it reasonable to transfer it to the large wealthy, childless cluster in the ‘Health first’ LOB cluster.

Finally, we left wealthy, childless people in the ‘Motor Fans’ LOB cluster as one small but interesting group (6). This is because the wealthy, childless cluster demographically strongly differs from the family clusters and we did not consider it reasonable to group these individuals. Furthermore, the ‘Motor Fans’ cluster exhibits much different behaviour from the ‘Health first’ cluster, which is why we also did not group all wealthy, childless individuals together.

6. Resulting Customer Segments

Merging Engage and LOB customer clusters resulted in six final clusters, that were provided to the insurance company. *Figure 25* gives an overview of the mean values over the considered variables, the cluster size as well as the estimated net profit, which gives an indication of the average profitability based on the Claims Rate and Total Premiums. In the following, the customer segments are qualitatively characterized in order to be interpretable by all involved stakeholders.

Cluster	Premiums in LOB: Motor	Premiums in LOB: Household	Premiums in LOB: Health	Premiums in LOB: Life	Premiums in LOB: Work Comp.	Educational Degree	Gross Monthly Salary	Has Children (Y=1)	Claims Rate	Total Premiums	Est. Net Profit	Cluster Size
Health-oriented families	248.71	187.14	218.54	38.98	40.48	2.49	2050.49	0.87	0.61	7264.2	2548.11	1641
Lower class families	114.07	570.49	159.75	100.22	100.16	1.76	1570.69	0.72	0.56	7610.61	3003.93	1052
Middle-class families	439.22	69.56	105.98	14.53	13.97	2.77	2455.49	0.9	0.45	6457.48	3076.49	2483
Unprofitable families	284.63	206.86	167.54	47.96	44.88	2.36	1953.26	0.89	0.9	3268.26	272.07	2289
Wealthy Motor Fans	413.23	90.42	120.7	18.41	17.68	2.9	3516.3	0.55	0.86	6278.65	958.13	759
Wealthy, health-focused	222.58	230.35	225.96	46.96	45.78	2.47	3703.43	0.17	0.77	6076.61	1419.4	1931

Figure 25: Means across variables for each customer segment

#1: 'Middle Class Families' (GREEN)

This cluster is the largest cluster and is defined by seemingly successful middle class families. This cluster combines lower and higher educated individuals with children and exhibits a higher middle class income. These families have low spendings in all insurances except for the Motor LOB, where spendings are exceptionally high. Furthermore, this cluster is the **most profitable,** combining a very low average Claims Rate with a rather high spend in insurance premiums.

It can be assumed that these are relatively young families, probably in white-collar jobs, based on their spendings in Life and Work Compensation premiums. They generally do not need high insurances, based on responsible behaviour with few claims and probably a stable financial situation and thus a good back-up in case of large emergencies. The high Motor expenses could be explained by the families owning several cars in order to drive the children and get to work.

#2: 'Health-oriented families' (BLUE)

This cluster includes lower to middle class families. With one of the highest overall payments in premiums and also a much higher Claims Rate than the 'Middle class families', these people seem to be more reliant on a good health insurance in order to cover potential expenses. They also have medium spendings in all other lines of business. Since higher claims are still offset by high total premium payments, this cluster is still very profitable for the insurance company.

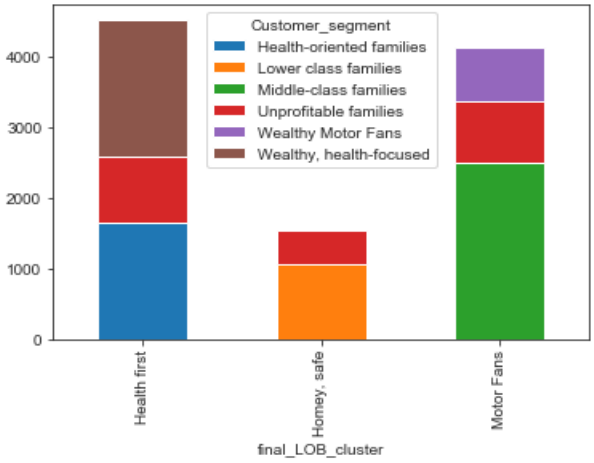

Figure 24: Final Clusters after merging

#3: 'Lower class families' (ORANGE)

With the lowest monthly salary, children and generally a low educational degree, this cluster represents the lower class families within the customer base. These individuals exhibit especially high payments of premiums, specifically in the lines of business Household, Life and Work Compensations. It can be assumed that these parents work in blue-collar jobs and need to be secured against work accidents and other threats to their life, since they probably do not have a strong financial back-up. The extraordinarily high expenses in Household could be explained by a high dependency on their accommodation. Since these people are very security-oriented, paying high premiums and claiming little, they are one of the most profitable customer segments.

#4: 'Unprofitable Families' (RED)

The 'Unprofitable Families' can be defined as the cluster that is of least value for the insurance company. These customers have the same demographic characteristics as the lower class families and behave much similarly to the other family segments - being most present in the Motor and 'Health first' segments. However, paying the lowest total premiums combined with the highest average Claims Rate, the insurance company does not make much profit or is even losing money with these customers. Thus, it is incredibly important to filter them from the valuable customers and treat them separately.

#5: 'Wealthy Motor Fans' (PURPLE)

The 'Wealthy Motor Fans' is the smallest cluster, consisting of wealthy, childless individuals with a very high educational degree. These people have very low spendings in each LOB except for Motor. The individuals in this cluster are probably successful upper class individuals with an expensive hobby: cars. High premiums in this LOB could be justified by either high claims rates through car accidents or the possession of high valued cars from expensive brands (or both). Based on the claims

rate, the former seems very probable. Otherwise, the ‘Motor Fans’ do not seem to seek much safety and probably have a strong financial back-up. Based on the high claims rates, these customers are one of the least profitable customer segments.

#6: ‘Wealthy, health-focused’ (BROWN)

This customer segments also consists of childless upper-class individuals. As the most income-strong group of all, these people are much less focused on Motor, but instead exhibit high Health expenses, followed by Household. These individuals might represent middle-aged, career-focused individuals with expensive property as well as high expectations towards their health-care.

7. Suggested Marketing Approaches

Middle class families

As this is the largest and the seemingly most profitable segment for the insurance company, it is of highest importance to keep these customers satisfied. As this segment is already profitable in their insurance behaviour, we mainly want to keep and further attract these kinds of customers in order to grow this highly valuable segment.

First, we should make sure that these customers feel taken care of by investing into a relationship Marketing approach. By rewarding them with special offers and personalized contract follow-ups, we could strengthen the relationship with the insurance company.

Second, these families could be able to attract other customers of a similar profile and equally profitable behaviour in order to grow this customer segment. On one hand, they might be well connected to other middle class families, for example through their work or the children’s friends. We could make use of word-of-mouth Marketing by offering discounts, if these customers successfully recommend friends to join the company. On the other hand, the children of the middle-class families, living by the example of their parents, have a high chance of becoming profitable customers one day as well and should be early attracted through a personalized approach, for example by specifically sending information on car insurance when the children turn eighteen.

Lower class families and ‘Health-oriented families’

Lower class families as well as ‘Health-oriented families’ are some of the very profitable segments as well. For lower class families specifically, we do not recommend rewarding referrals, since they share the same profile as unprofitable lower class families and could thus easily bring in customers that are highly unprofitable for the insurance company.

Instead, the similarity between lower class and ‘Health-oriented families’ in their Claims Rates, Total Premiums as well as their demographic profile, that only differs in their education and salary, suggests cross-selling as an approach that should be pursued.

Since both the ‘Homey, safe’ and ‘Health first’ LOB segment show strong tendencies towards safety-oriented insurances, we could promote each of the two customer segments the premiums that they are currently not investing into. Thus, the lower class families would receive increased information on health premiums that could complete their all-round family care package, whereas ‘Health-oriented families’ would be made more aware of the Household, Work and Life insurance package that could provide additional coverage for the family.

Unprofitable families

The unprofitable 'Lower class families' are currently not lucrative for the insurance company. Exhibiting high Claims Rates but low paid premiums, these customers are probably not looking for a good insurance coverage, but are seeking coverage of their extensive claims at low rates. For such customers, we recommend to either drive them into a profitable territory or repel them from the company all-together. Since this is a relatively large segment, it should first be tried to increase profit and more extreme measures, such as 'firing' the customer, should only be taken in hopeless cases.

Profit on these customers could be increased by increasing premiums paid or lowering the claims rate. Since these low income, low premium-paying customers are probably financially sensitive, they could be lead towards higher-valued premium packages with temporary discounts, locking them in for the long term. Furthermore, 'good' behaviour could be incentivized by lowering rates slightly based on an exponentially decreasing Claims Rate. Also, different insurance packages could be promoted that are highly interesting for other customers of their profile, the 'Lower class families'. This could for example be a package consisting of Household, Life and Work insurance.

As an approach of last resort, the insurance company could decide to actively *"fire"* strongly unprofitable customers in this segment, if profits cannot be increased. Since a simple de-investment in Marketing campaigns would probably only give such insurance-exploiting customers the feeling of more freedom, an active approach would need to be followed. This could include the strict raising of prices, the lowering of service levels and the restriction of these customers to insurance packages with much lower coverage.

Wealthy Motor Fans

The people in this small but interesting segment are different to all other segments in their demographic characteristics and insurance spendings. Even though high claims rates drive down profit from these customers, they are a highly interesting due to their exceptionally high income and very low paid premiums in most lines of business. Their financial situation hints that there is a high potential for increased premiums, which could balance their increased claims rate and make this segment much more profitable in the future.

Since we can observe that other wealthy customers spend highly on health insurance, this line of business could be specifically interesting to the 'Wealthy Motor Fans', as well. Especially assuming the Motor sport as a hobby, a high-valued health-insurance that is greatly reliable in case of an accident could be promoted as a great addition to the already existing Motor insurance, possibly even sold as a package. Furthermore, a more profitable behaviour could be rewarded by an annual decrease in premiums, if these customers can lower their claims rates.

Wealthy, health-focused

Just like for the 'Wealthy Motor Fans', if we could increase overall premium payments on this high-income, childless, health-focused segment, that exhibits lower claims rates than the ''Wealthy Motor Fans''', it could become a very valuable segment for the insurance company.

Already exhibiting average to high premium payments in all lines of business and a particular focus on health-care, these customers seem to put a strong focus on a good insurance coverage. Knowing that they have high financial funds available, we can target this customer segment by offering them a *prime* insurance – a superior, high-quality service. In this way, to drive up spendings on insurance premiums through the feeling of a superior treatment, customers could be given special attention through personalized contract follow-ups as well as the offering of an expensive insurance package for Health, Household, Life and Work, that offer particularly good coverage as well as strong service levels to these customers.

8. Conclusion

Following the KDD/Data Mining process, from data preprocessing, over the clustering analysis up to the characterization of the clusters, we were able to identify six customer segments, that are clearly separable through their demographical characteristics as well as their behaviour within the different lines of business. Based on the found customer profiles as well as certain assumptions, suggestions regarding the Marketing approach were made, that could now be implemented and evaluated by the insurance company.

Furthermore, the insurance company is now equipped with a clear picture of the different segments that make up its customer base and with a comprehensive, interpretable qualitative and quantitative descriptions of every customer group. Thus, besides the already given Marketing suggestions, the company is enabled to come up with own conclusions and strategies based on previous or upcoming experiences.