NOVA IMS Information Management School

Descriptive Methods of Data Mining

ABC Insurance: Customer Segmentation with K-Means Clustering

Academic Year 2021/2022

1st Semester

Cátia Simas	M20210827
Paula Eveling	M20210331
Zukiswa Mdingi	M20210408

TABLE OF CONTENTS

INTRODUCTION **3**

BUSINESS UNDERSTANDING **3**

2.1 Business Objectives 3

2.2 Situation Assessment 4

2.3 Determine Data Mining Goals 5

DATA UNDERSTANDING **5**

3.1 Data Collection 5

3.2 Exploratory Data Analysis 6

3.3 Data Quality Verification 9

DATA PREPARATION **9**

4.1 Data Selection 9

4.2 Data Cleaning 9

4.3 Data Construction 10

4.4 Data Formatting 10

4.5 Remove Data 10

4.6 Data Normalization 10

4.7 Outliers Preparation 11

MODELLING **11**

5.1 Principal Component Analysis 11

5.2 K-Means 13

5.3 Cluster Interpretation 16

EVALUATION **18**

DEPLOYMENT **18**

CONCLUSION **19**

REFERENCES **20**

APPENDIX **21**

10.1 Appendix A: Description of the Variables 21

10.2 Appendix B: Project Plan 23

10.3 Appendix C: Clustering Output tables 24

1. INTRODUCTION

Customers are different. Their preferences, buying power, and other characteristics can differ over a wide spectrum (Onnen, Heiko. 2021). While mass marketing tactics are still able to get results, the assumption that simply everyone will be interested in buying what a company is selling is a time-consuming, inefficient, and expensive strategy (Upland Software). Common characteristics in customer segments can guide how a company markets to individual segments and what products it promotes to them, tailoring the marketing efforts to various audiences' subsets. This process is called Customer Segmentation.

Customer segmentation is the process of dividing customers and prospects into different groups, each of which share characteristics and behave similarly within a segment but look and behave differently across multiple segments. Customers can be segmented based on demographics (age, gender, income), according to the characteristics that influences purchasing decisions, or grouped based on where they are located, among others.

With that said, this project is concerned with developing different customer profiles for the Marketing Department of an insurance company called ABC Insurance Limited, using the data mining clustering algorithm K-means in Python following the CRISP-DM (Cross Industry Standard Process for Data Mining) model. This tool allows a Customer Segmentation such that it will be possible for the marketing department of the company to better understand all the different customer profiles.

2. BUSINESS UNDERSTANDING

2.1 Business Objectives

Background

ABC Insurance Limited is an insurance company that offers its clients short-term and long-term insurance including motor, household, health, life, and work insurance. The company has a vast client base, which is very diversified from having young and elderly customers with different education qualifications and ranging monthly salaries.

Due to this vast client base, the marketing department is currently facing a challenge: how to attract new clients and ensure customer retention in a cost-effective way. If the marketing department could have a better understanding of their clients' profiles, they would be able to market effectively to try and to achieve their mandate.

Business Objectives

The main objective of this project is to develop a customer segmentation in such a way that it will be possible for the marketing department to understand all the different customer profiles and tailor its strategies accordingly. This will allow the company to scale while retaining customers and increasing their monetary value by offering different products.

Business Success Criteria

For the project to be considered a success, a maximum of six customer profiles must be identified and its characteristics well described. This information should provide insights on how to market to each group based on their similarities.

As a result of a more data-driven marketing approach, the department expects to increase the retention rate of customers with high monetary value by 10% and increase the acquisition of customers by at least 15% in the next two quarters.

2.2 Situation Assessment

Inventory of Resources

To implement this project, a SAS dataset containing 10,296 observations and fourteen variables was provided for analysis. In addition, two stakeholders with strong business domain were assigned to answer eventual questions and provide insights to the investigation. In terms of technical administrators, the team is composed of three members: a mathematician, data scientist and a data miner. The technology available for the execution of this project is Jupyter Notebook and Python as a programming language.

Requirements

- The deliverables, written report and Jupyter Notebook, must be submitted to the stakeholders for evaluation by January 16th, 2022, in Moodle.
- A walkthrough section scheduled with the stakeholders on January 29th.
- Expected to define, describe and explain the chosen clusters.
- The project must comply with the General Data Protection Regulation (GDPR).

Assumptions

- First policy year cannot be greater than 2016.
- The birthday year cannot be greater than the first-year policy.
- Not all premiums can be null in the same row.
- Will probably need to have discussions with the stakeholders to get clarity on uncertainties we might have about the data.
- Will need to assess the results of the clustering to ensure that the selected clusters and reasoning we suggest make sense in the real-world.

Risks and Contingencies

No major risks are foreseen that would result in the complete failure of the project. The time constraints of the consultants working on this project could be a fact which could result in some delays in the set project timelines. These time constraints are from the uncertainties that might emerge in the data preparation and clustering activities, if unforeseen circumstances do arise then unplanned consultations with the experts might be required. To plan for this type of risk time management is very important, therefore more time will be scheduled for the data preparation and clustering activities. Time will also be set aside should the need arise with the experts on any data issues experienced.

Cost and Benefits

Costs:

- The operating costs of the project.
- Might cost more for the insurance company to do marketing for different groups.

Benefits:

- Benefits to better understand the type of customers as a result from the data understanding activity.
- To successfully achieve the specified data mining goals.
- To successfully achieve the business goals.
- The insurance company will have a better understanding of their client base and can find ideas on how to market to different groups formed from the customer segmentation.

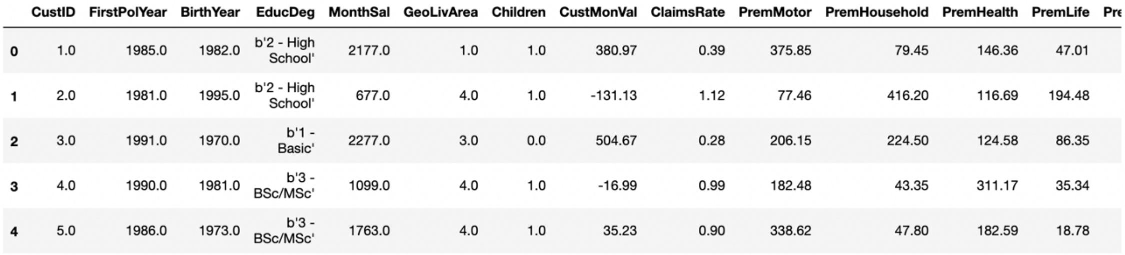
2.3 Determine Data Mining Goals

- Produce a Python source code (Jupyter notebook or .py files).
- To use the correct clustering method and be able to select the most optimal number of clusters.
- Develop a Customer Segmentation.
- To be able to justify why we chose the different approaches used in this project.
- To be able to define, describe and explain the clusters you chose.
- What would be the ideal customer for this company?
- Specify criteria for model assessment.

3. DATA UNDERSTANDING

3.1 Data Collection

The dataset used in this analysis was provided by the stakeholders in October 2021. It contains data from ABC's customers between 1974 and 2016, with 10,296 observations and fourteen variables, where eleven of them (variables) are numerical and three are categorical, as shown on Figure 1. A detailed description of each variable and its format can be found on Appendix A: Description of the Variables.



	CustID	FirstPolYear	BirthYear	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife	Pre
0	1.0	1985.0	1982.0	b'2 - High School'	2177.0	1.0	1.0	380.97	0.39	375.85	79.45	146.36	47.01	
1	2.0	1981.0	1995.0	b'2 - High School'	677.0	4.0	1.0	-131.13	1.12	77.46	416.20	116.69	194.48	
2	3.0	1991.0	1970.0	b'1 - Basic'	2277.0	3.0	0.0	504.67	0.28	206.15	224.50	124.58	86.35	
3	4.0	1990.0	1981.0	b'3 - BSc/MSc'	1099.0	4.0	1.0	-16.99	0.99	182.48	43.35	311.17	35.34	
4	5.0	1986.0	1973.0	b'3 - BSc/MSc'	1763.0	4.0	1.0	35.23	0.90	338.62	47.80	182.59	18.78	

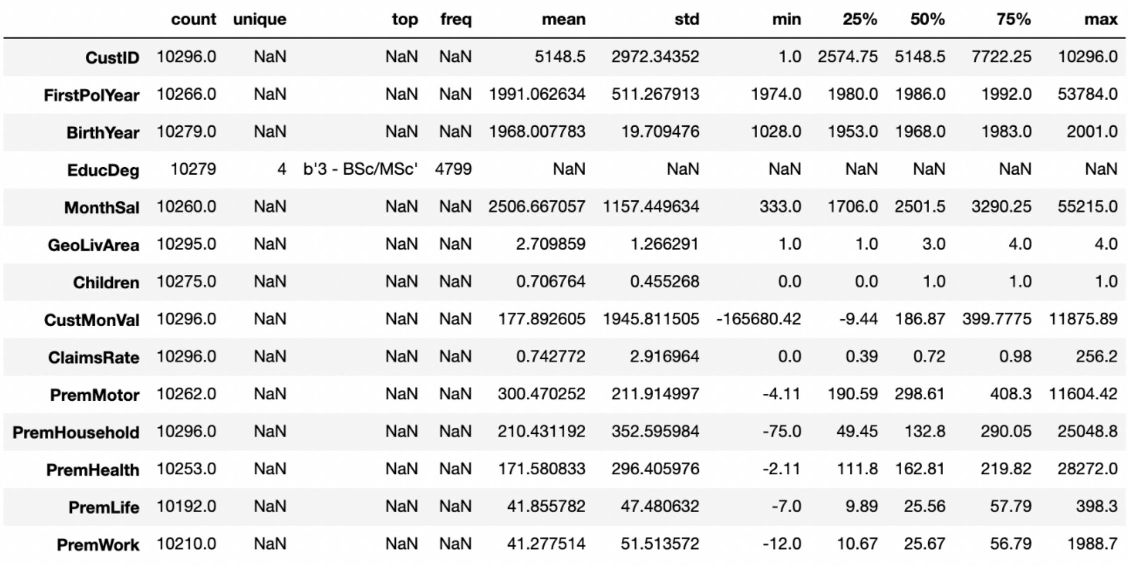
Figure 1: First five rows of the dataset

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an approach to analyzing datasets to summarize their main characteristics often using statistical graphs and other data visualization methods. Through the process of EDA, one can get a better understanding of the dataset variables and the relationship between them, making it easier to discover patterns, detect outliers and anomalies, and test underlying assumptions.

Statistical Description

The descriptive statistics table, shown on figure 2, indicates that the dataset contains outliers, missing values, incorrect inputs, and that most of the variables may not be normally distributed. For instance, the average Premium Health (PremHealth) is equal to 171.58, while its standard deviation is 296.40, meaning that the data is very dispersed. The table also shows that the third percentile (75%) is equal to 219.82, while the maximum value observed is 28272, which clearly indicates that there is at least one severe outlier or incorrect value.



	count	unique	top	freq	mean	std	min	25%	50%	75%	max
CustID	10296.0	NaN	NaN	NaN	5148.5	2972.34352	1.0	2574.75	5148.5	7722.25	10296.0
FirstPolYear	10266.0	NaN	NaN	NaN	1991.062634	511.267913	1974.0	1980.0	1986.0	1992.0	53784.0
BirthYear	10279.0	NaN	NaN	NaN	1968.007783	19.709476	1028.0	1953.0	1968.0	1983.0	2001.0
EducDeg	10279	4	b'3 - BSc/MSc'	4799	NaN	NaN	NaN	NaN	NaN	NaN	NaN
MonthSal	10260.0	NaN	NaN	NaN	2506.667057	1157.449634	333.0	1706.0	2501.5	3290.25	55215.0
GeoLivArea	10295.0	NaN	NaN	NaN	2.709859	1.266291	1.0	1.0	3.0	4.0	4.0
Children	10275.0	NaN	NaN	NaN	0.706764	0.455268	0.0	0.0	1.0	1.0	1.0
CustMonVal	10296.0	NaN	NaN	NaN	177.892605	1945.811505	-165680.42	-9.44	186.87	399.7775	11875.89
ClaimsRate	10296.0	NaN	NaN	NaN	0.742772	2.916964	0.0	0.39	0.72	0.98	256.2
PremMotor	10262.0	NaN	NaN	NaN	300.470252	211.914997	-4.11	190.59	298.61	408.3	11604.42
PremHousehold	10296.0	NaN	NaN	NaN	210.431192	352.595984	-75.0	49.45	132.8	290.05	25048.8
PremHealth	10253.0	NaN	NaN	NaN	171.580833	296.405976	-2.11	111.8	162.81	219.82	28272.0
PremLife	10192.0	NaN	NaN	NaN	41.855782	47.480632	-7.0	9.89	25.56	57.79	398.3
PremWork	10210.0	NaN	NaN	NaN	41.277514	51.513572	-12.0	10.67	25.67	56.79	1988.7

Figure 2: Descriptive Statistics Table

Each variable was analysed individually, in terms of frequency, distribution, and count, and the results are presented on Appendix 1: A description of the variables. Although several underlying assumptions can be made through the statistical description of the dataset, graphs such as histograms and boxplots, can make it easier to visualise such assumptions.

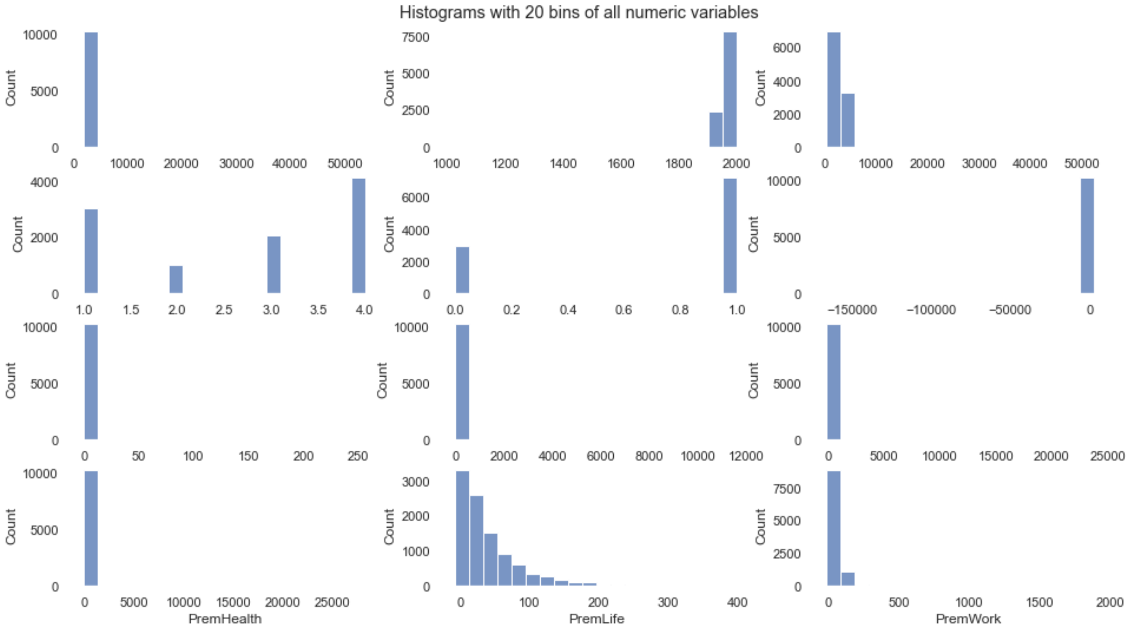


Figure 3: Histogram of the Variables

The histograms for the different variables were plotted in order to understand their distribution and behaviour. However, many of the histograms obtained are not clear. The reason for this is most likely due to the existence of outliers, as observed on the descriptive statistics table.

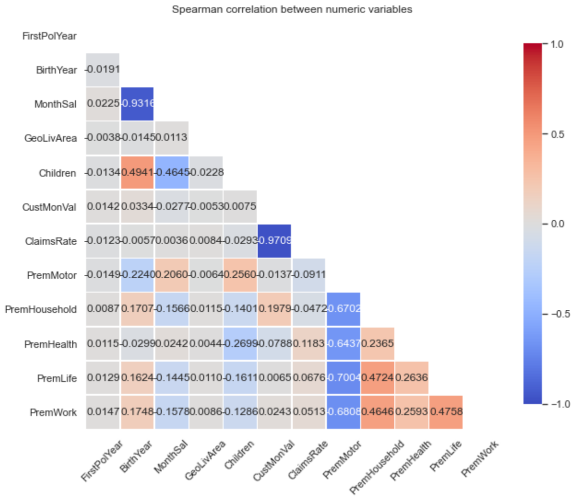


Figure 4: Spearman Correlation Matrix

The Spearman Correlation Matrix (figure 4) indicates a strong negative correlation of -0.97 between the claims rate and customer monetary value, meaning that the higher the claim rate is, the lowest is the customer's value for the company.

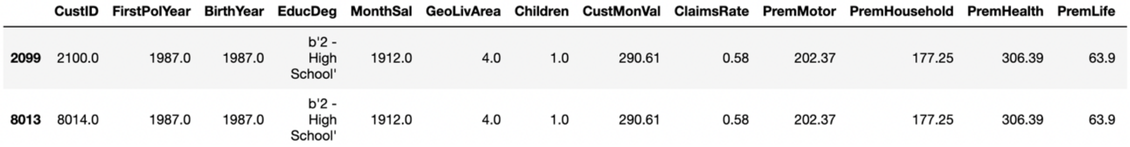
Another interesting correlation is the monthly salary and the year of birth (-0.93 coefficient), indicating that, the older the customers are, the higher are their salaries. This information can be helpful to determine which variables should be kept in the modelling phase.

Missing Values

The dataset contains a total of 309 rows with missing values, accounting for 3% of the total observations. Considering the high representativity of the rows with missing values, removing them altogether is not a viable solution as it could impact the results of this project. In order to understand the cause of missing values, a meeting with the stakeholders was carried out. As a result, it was identified that premiums missing values happen when customers have never acquired the type of insurance. For instance, if a household premium is null, it means that the customer does not have a household insurance. The remaining null values were caused by human errors.

Redundancy

Initially, the redundancy inspection returned zero duplicate rows. However, when the customer id is removed from the analysis, three observations with the exact same values can be identified, as shown on figure 5. It is possible that these customers were added into the system twice by accident. Although their representativity is not significant, cases like these might occur more frequently in the future, and this algorithm must be prepared to handle such situations. For this reason, duplicated rows will be treated.



	CustID	FirstPolYear	BirthYear	EducDeg	MonthSal	GeoLivArea	Children	CustMonVal	ClaimsRate	PremMotor	PremHousehold	PremHealth	PremLife
2099	2100.0	1987.0	1987.0	b'2 - High School'	1912.0	4.0	1.0	290.61	0.58	202.37	177.25	306.39	63.9
8013	8014.0	1987.0	1987.0	b'2 - High School'	1912.0	4.0	1.0	290.61	0.58	202.37	177.25	306.39	63.9

Figure 5: Duplicated rows example

Errors and Incorrect Values

Two types of incorrect issues were found in the dataset. Firstly, there is one row in the dataset where the FirstPolYear is greater than 2016. Secondly, there were 1997 rows where the FirstPolYear is greater than the BirthYear. According to the stakeholders, this happened as a result of human errors, where the years were swapped when added into the system.

Outliers

As seen on the histograms, all columns, except GeoLivArea and Children, contain severe outliers. Some of the outliers seen in the graphs could fall away once the data is cleaned, however, not all outliers might disappear after this process. If one should find that there are still outliers after cleaning the data, it might have to consider grouping the outliers and put them in a cluster of their own.

3.3 Data Quality Verification

Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data quality measures the condition of a dataset to serve its specific purpose based on factors such as accuracy, integrity, completeness, consistency, validity, among others. The data quality assessment is one of the most important steps of any data science project, as poor-data quality could lead to wrong decisions causing lost sales opportunities, for example.

Although there are some data quality issues that need to be addressed as seen in the summary above, the dataset does not include any serious data quality issues that would result in not being able to continue to try and achieve the recommended goals and plans. Based on the detailed analysis of the variables above, it is possible to say that the quality of the dataset provided by the stakeholders is relatively good, since there are no significant structural issues in it. Nevertheless, a proper data preparation plan will need to be established to deal with outliers and possible errors on the dataset.

4. DATA PREPARATION

Data preparation is the process of cleaning and transforming raw data prior to processing and analysis and includes removing columns that will not be needed, check missing values, and change the format of the variables when applicable. Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data preprocessing is a proven method of resolving such issues.

4.1 Data Selection

Due to the purpose of our study, the following variables were selected for this study: FirstPolYear, BirthYear, EducDeg, MonthSal, GeoLivArea, Children, CustMonVal, ClaimsRate, all Premium’s. These variables were selected because they describe customers characteristics or behaviour and can provide valuable insights to the analysis.

4.2 Data Cleaning

Missing Values

Considering the representativity of rows with missing values (3%), two strategies were applied to treat these observations. Firstly, a zero value was assigned to all premium variables with missing values, indicating that the customer has never acquired that type of insurance. These observations represent 2.6% of the dataset. Secondly, when analysing the remaining missing values, it was identified that they represent only 0.89% of the data. In that case, they were removed from the analysis. So, now the dataset contains 10,204 observations.

Redundancy

Since there are only three pairs of possible duplicate rows, one duplicate from each pair was removed from the dataset.

Errors and Incorrect Values

To treat errors and incorrect values, the following decisions were made:

- Birth year cannot be higher than the first-year policy
- FirstPolYear: remove if the value is greater than 50,000
- Year of Birth: remove if the value is less than 1,200
- MonthSal: remove if the value is greater than 6,000
- CustMonVal: remove if the value is less than -2,000 and greater than 2,000
- ClaimsRate: remove if the value is greater than 2
- PremMotor: remove if the value is greater than 2,000
- PremHouseHold: remove if the value is greater than 4,000
- PremHealth: remove if value is greater than 20,000

In total, thirty-two observations were lost in this step. The cumulative lost observations account for 1.23% of the initial observations. The database is now left with 10,169 observations.

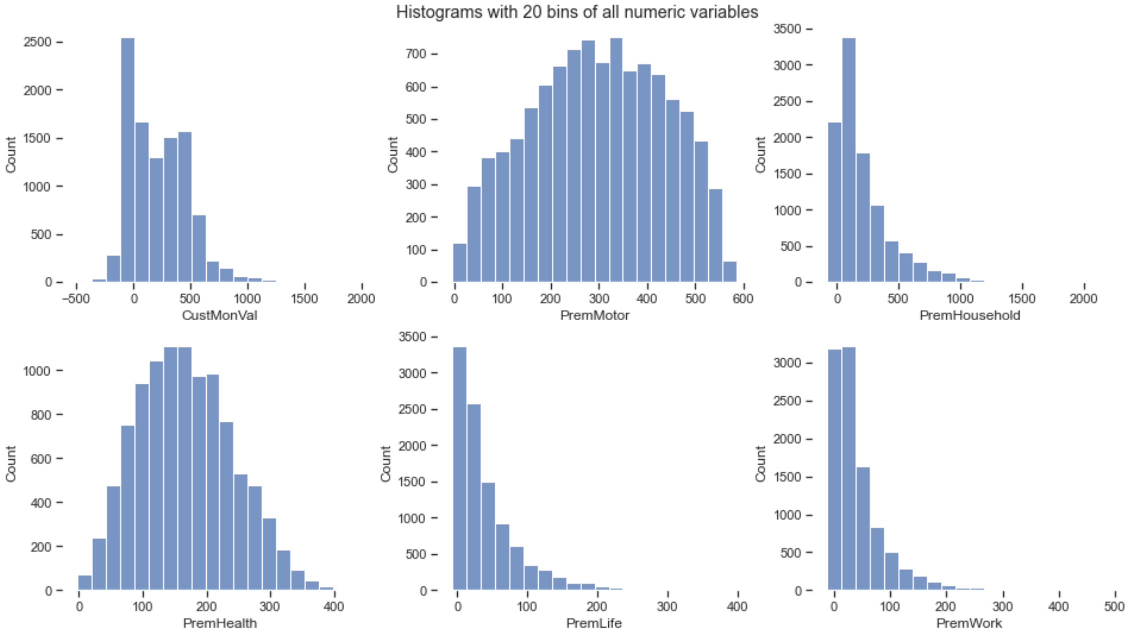


Figure 6: Histogram after removing severe outliers/incorrect values

4.3 Data Construction

Since not much value can be derived from the BirthYear variable as it is in the dataset, it was therefore decided to construct a new variable called ‘CustomerAge’. This is a categorical variable and its values were binned prior to the modelling phase, as follows: 0-17, 18-24, 25-34, 35-44, 45-54, 55-64, 65+.

4.4 Data Formatting

Machine learning models require all input and output variables to be numeric. In this project, only one categorical variable is formatted as such, EducDeg, meaning that it must be encoded prior to the modelling phase. In addition, the variables GeoLivArea and the age bins are also categorical but present only numeric values. Considering that k-means is a distance based algorithm, these variables were encoded as well. The method chosen for treating these variables was the one hot encoding, that is a representation of categorical variables as binary vectors.

4.5 Remove Data

For this project, there are certain variables that will not be used in the analysis, either because they have a high correlation with another variable, or because they don't represent customer's characteristics or behavior. Therefore, the following variables were excluded: CustID, BirthYear, FirstPolYear, CustAge, ClaimsRate.

4.6 Outliers Preparation

After the cleaning of the data, it can be seen on the boxplot below (figure 7) that there are quite a few outliers left in the household, health, work, and life premiums variables even after data cleaning. Because of these outliers that persist, it was decided to cluster these outliers and put them separately from the rest of the data. As a result, two datasets were created. The dataset X contains 8,840 observations without any outliers, while the dataset Y contains 1,329 observations with the outliers observed on the variables customer monetary value, premium household, premium health, premium life, and premium work.

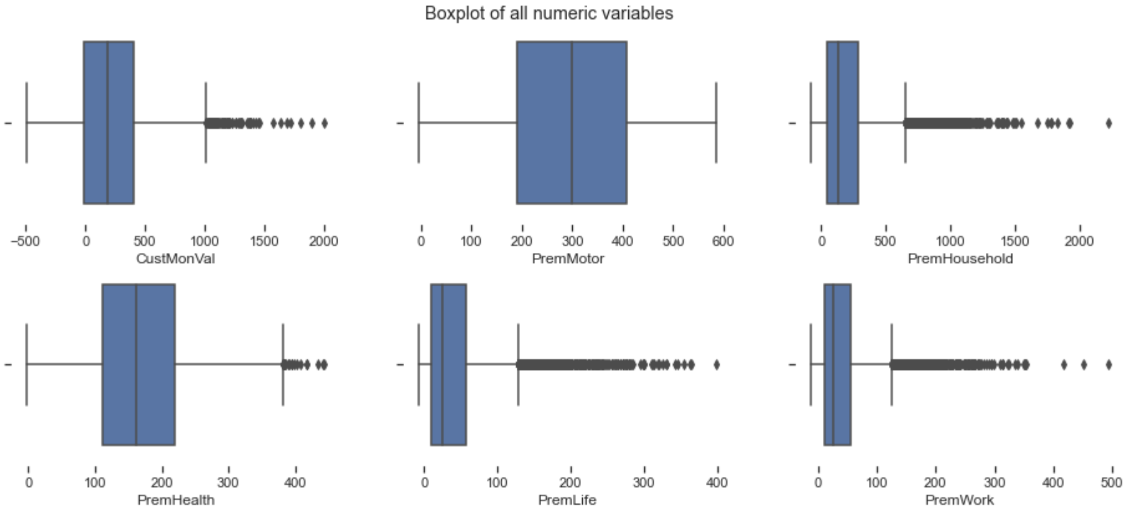


Figure 7: Boxplot after preprocessing

4.7 Data Normalization

The preprocessed data contains attributes with a mixture of scales for various quantities such as salary and premiums. Most machine learning algorithms, like k-means, expect or are more effective if the data attributes have the same scale. With that said, the preprocessed dataset was normalized using the min max scaler technique, which refers to rescaling real valued numeric attributes into the range 0 and 1.

5. MODELLING

5.1 Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure that can be considered as a projection method to conserve the maximum amount of information from the initial dimensions. This technique is usually used for dimensionality reduction, but it can also be used to visualize data, observe trends, clusters and outliers (Tam, Adrian, 2021). This method helps to understand the correlation between the variables and observations and if all variables are indeed relevant for the analysis. A PCA was applied to each of the datasets (with and without outliers) in order to reduce the dimensionality of the dataset, while retaining as much variance in data as possible.

Below is a visualization of the scatter plots of the first two principal components by the "customer monetary value". These graphical displays offer a good partial approximation to the systematic information contained in the data. The plots show that both datasets, without outliers and with outliers only, present a similar pattern, however, it is important to note that scatter plots only take into account two dimensions, which gives a limited view of the multivariate phenomenon.

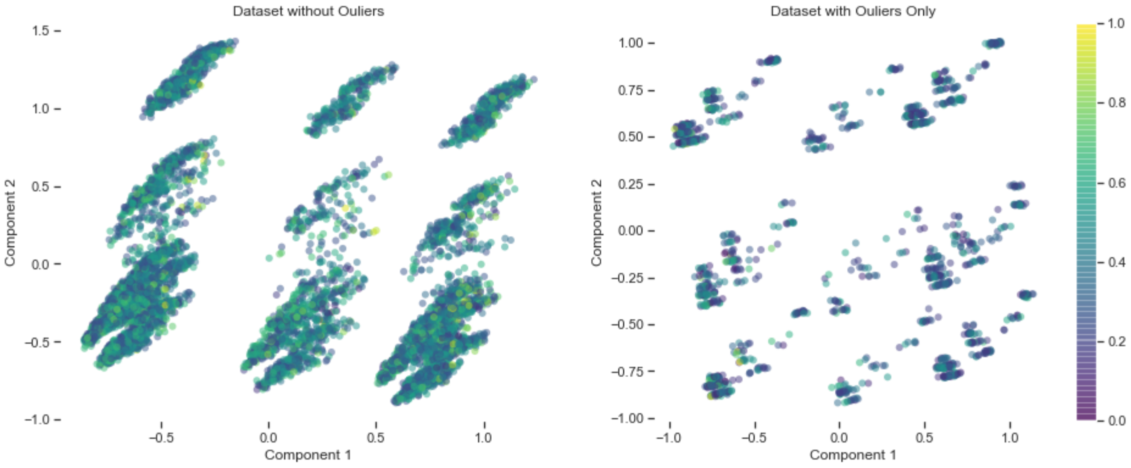


Figure 8: Principal Component

In terms of dimensionality reduction of the database without outliers, the graph below (figure 9) shows the cumulative percentage of the variation accounted for by each principal component. The ideal curve of a cumulative variance explained plot should be steep, bending at an “elbow” and flattening out after that. The elbow indicates the cutting-off point, which in this case indicates that just the principal components 1 to 10 are enough to describe the data. Similarly, the dataset with only outliers presented a cutting-off point around the tenth component. Thus, it is plausible to retain 10 principle components for both datasets, which explains 90.8% and 92.8% of the cumulative variance explained respectively.

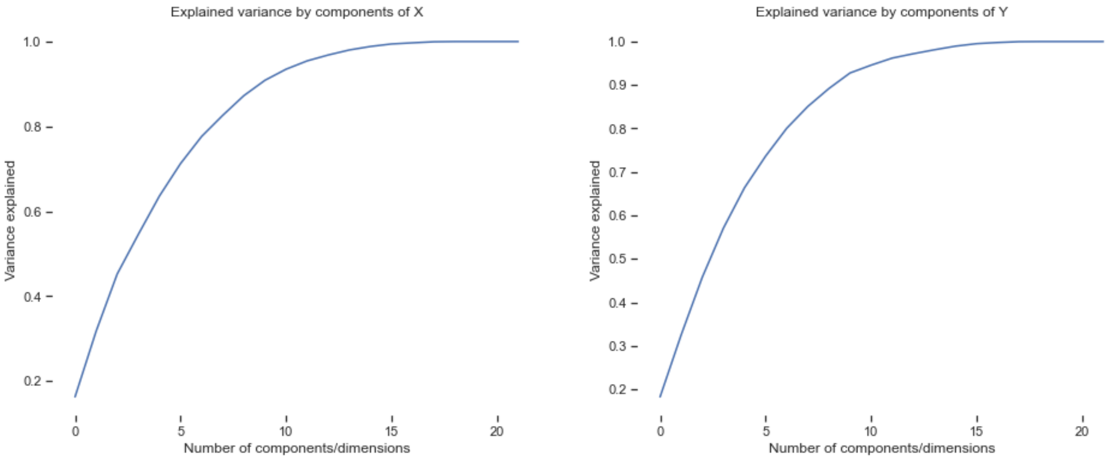


Figure 9: Explained Variance by Components

5.2 K-Means

K-means is a hard partitioning algorithm, meaning that each data point falls into only one partition. According to Garbade, Dr. Michael J (2018), the algorithm identifies k numbers of centroids, being k a number defined by the researcher, and then it allocates every data point to the nearest cluster, while keeping the centroids as small as possible. One of the advantages of this method is that it is simple to implement, and it scales to large datasets. On the other hand, it requires domain knowledge to define k (number of clusters).

Elbow method

A range of 1 to 20 was chosen to determine the number of clusters (k) and the within-cluster sum of squares (WCSS) was computed. The number of clusters (k) should be selected where there is a curve in the graph, as it indicates that from there the WCSS stops rapidly decreasing. It is expected that the plot looks like an arm with a clear elbow. Unfortunately, one does not always have such clearly clustered data. This means that the elbow may not be clear and sharp, as it can be observed on figure 9. In such an ambiguous case, the Silhouette Method can be a more appropriate method.

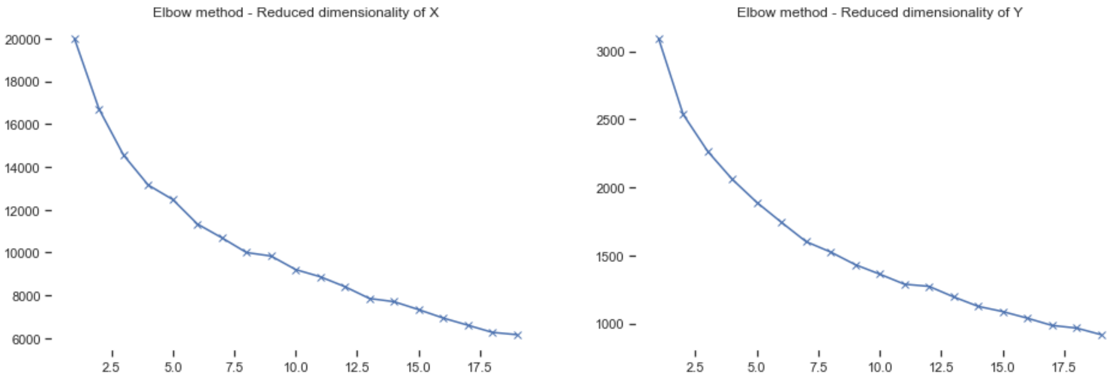


Figure 10: Elbow method - Dimensionality Reduction

Silhouette method

This method measures how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). Because dissimilarity can only be measured in more than 1 partition, a range for k of 2 to 20 was chosen and k-means was applied for each one. One can observe that, on the left graph is reaching a high value for k = 4, indicating that the optimal number of clusters for the dataset without outliers is 4. The same process was applied with the dataset with outliers only, where it can be observed that, where k=3, there is the first value in growth, indicating that this is a reasonable number of clusters and it is aligned with the business success criteria. It should be remembered that the choice of the number of clusters is very subjective and depends on each analyst's interpretation, which is one of the disadvantages of this method.

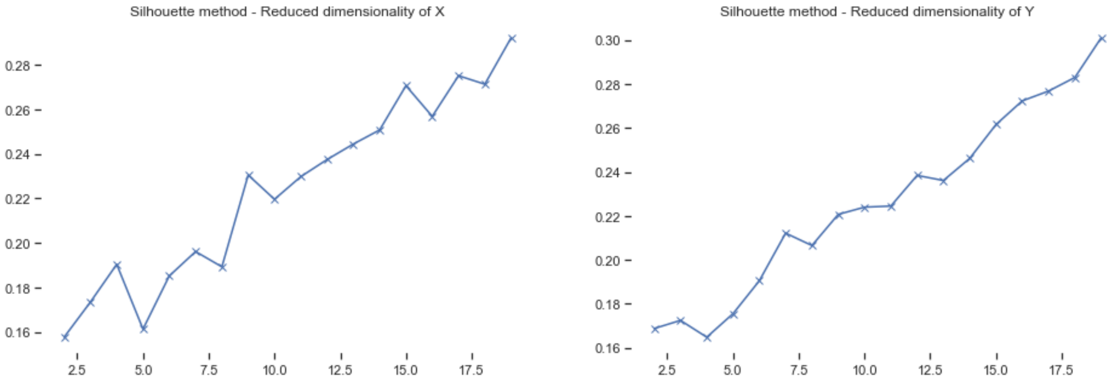


Figure 10: Silhouette method - Dimensionality Reduction

Cardinality versus Magnitude

In addition to the silhouette and elbow methods, the cardinality and magnitude of each dataset must be analysed in order to confirm that the decision about the number of clusters is as accurate as possible. A higher cardinality tends to result in a higher magnitude. Clusters are anomalous when cardinality does not correlate with magnitude. As it can be seen for the dataset with outliers only the clusters seem to be less correlated with the magnitude.

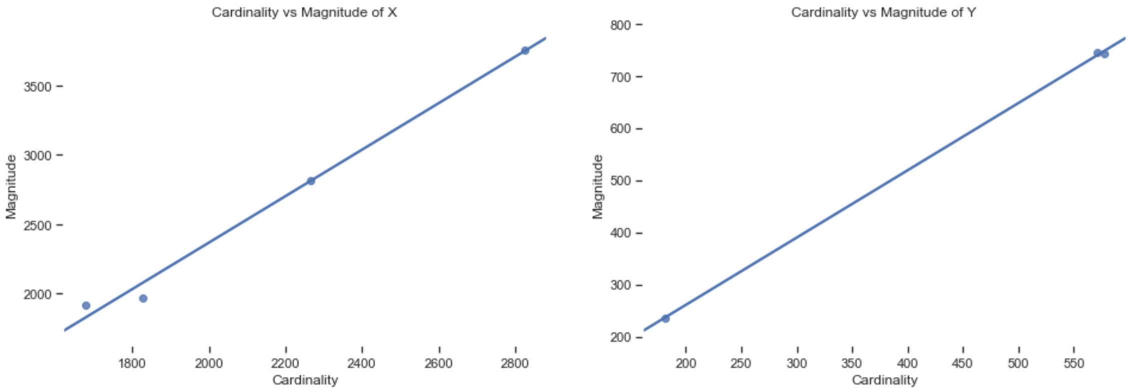


Figure 11: Cardinality versus Magnitude

Once the number of clusters is chosen and analysed, the scatterplot of the principal components is plotted by cluster to see if there are any distinctive patterns. The grey circles represent each centroid and each cluster is represented in a different colour. As one can observe, the plot is not very homogeneous as the green and yellow clusters are mixed together. As mentioned before, since the scatter plot takes into account two dimensions only, it gives a limited view of the multivariate phenomenon.

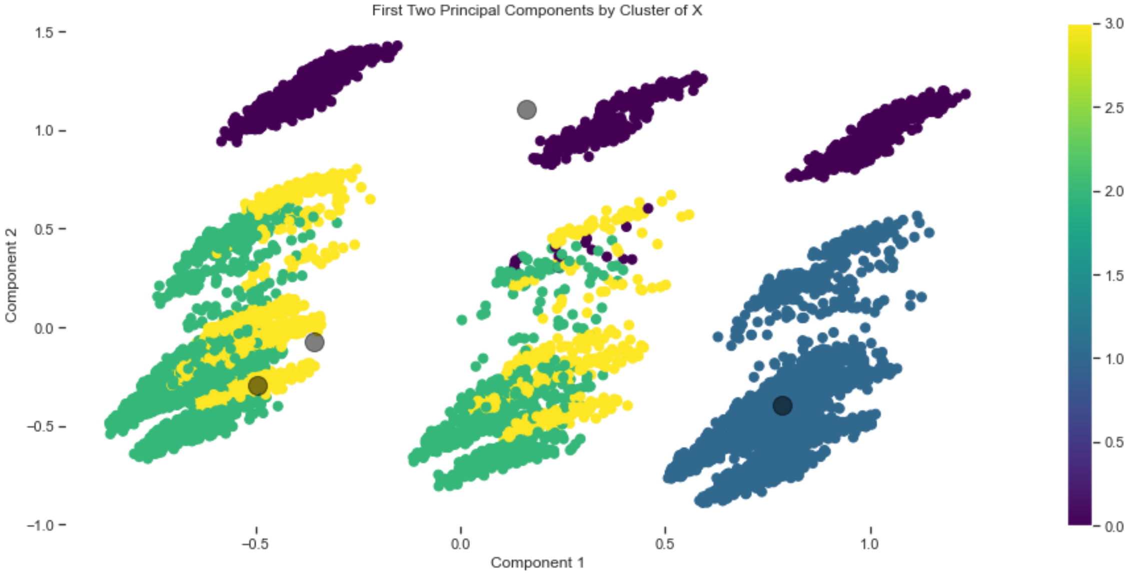


Figure 12: Clustering - Dataset without outliers

The cluster with outliers only, on the other hand, is more homogenous where each cluster can be visualized clearly in the scatter plot.

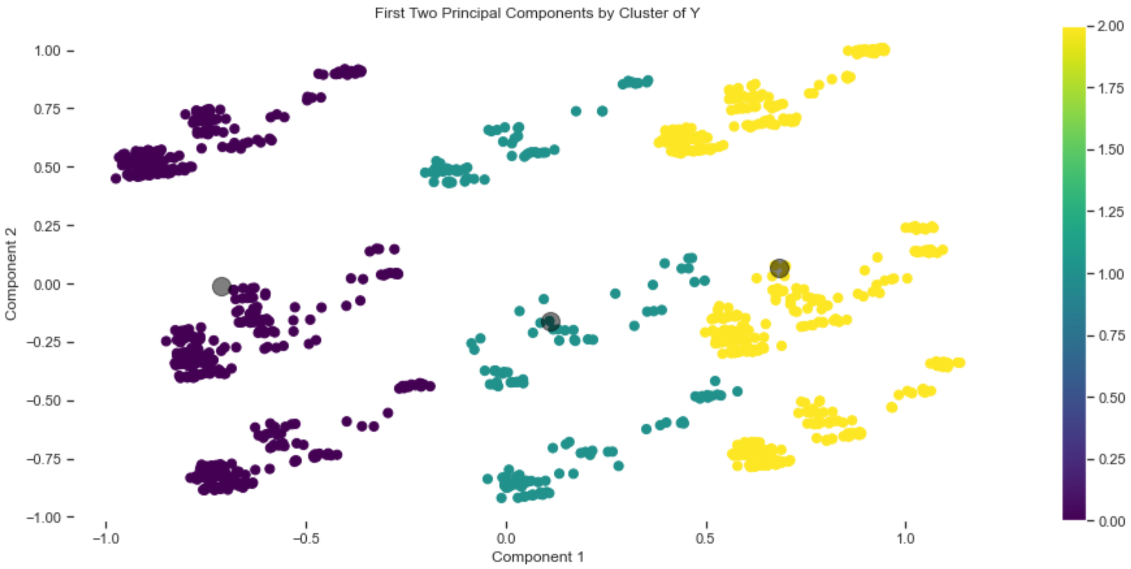


Figure 13: Clustering - Dataset with outliers only

5.3 Cluster Interpretation

Cluster 0 - Seniors

The cluster 0 have the customers with the highest monthly salary (3,758.33) and are considered to be of retirement age (+65; 99%). This cluster has the lowest value in terms of customer monetary value (186.49), this can be expected for clients in this age group because they usually claim more in terms of insurance specifically when it comes to health insurance this can also be seen from having the highest health premium value (218.93). The cluster also has the lowest value for Motor premium's (258.74) which is also expected for people of old age. This cluster also has the lowest value under children (0.08%), this is expected because it can be assumed that the children of the people in this group are already of working age and are not included under their parent's insurance.

Recommended Strategies:

For this group, it is recommended that ABC Insurance creates, if not existent, specific insurance plans for retirement planning, that includes gold or diamond programs with a higher coverage, for example, so customers can upgrade their plans. This would help the company to increase the group's monetary value while potentially increasing the retention rate.

Cluster 1 - Blue-Collars

This cluster can be considered to be of the working class, as 100% of the customers in this segment have a high school diploma and their ages range from 35-44 (31%). This group pays the highest in premium life (22.51) and work (41.20), while having the second worst monetary value (200.38). In terms of geographical location, most of the customers live in region 4 (41%) and 89% have children.

Recommended Strategies:

This is a group that does not generate much revenue for the company, but still accounts for 32% of ABC's customers. Since they have children, there is potential to increase the premium health in this segment. Affordable yet with good coverage health insurance plans could be an opportunity to be explored in this group.

Cluster 2 - Gold Motor

This segment has the highest customer monetary value (208.40), with the highest premium motor compared to all other segments (364.19). On the other hand, this group has the lowest household (118.46), work (26.70), and health (148.10) premiums values, and it has the lowest salary (2,339.42). Most of them live in geographical location 1 (49.5%) and have children (89%).

Recommended Strategies:

There are a lot of opportunities to attract and retain customers from this segment. But before going straight to defining marketing strategies and product development for this group, it would be interesting to understand why most of them have the highest premium motor while the lowest household, work and health insurances. This would help ABC Insurance to streamline its strategies to be more assertive in its decisions. Nonetheless, since they are interested in insuring their vehicles, they might be interested in automotive related content, which could be an opportunity for ABC to educate customers in order to keep claims rate as low as possible while increasing the retention rate.

Cluster 3 - Region 4

This segment is the second best in terms of monetary value, but there is nothing peculiar that differentiates this group significantly from the others, except for the fact that 100% of them live in region 4. Besides this, this is an average customer: second lowest premium life, household, health, work, with the second highest premium motor and salaries. This cluster consists of educated individuals with Bsc/Msc education (77%).

Recommended Strategies:

There are endless strategies that can be adopted for this group. One of them is to apply digital marketing campaigns targeting specifically well-educated residents of region 4. This can help the company to increase the acquisition of valuable customers.

Outliers Clusters - Diamond

This group represents the ideal customer for ABC Insurance. It consists of all the outliers, and it has the highest monetary value (297.75 and higher) compared to all other clusters. These customers mostly have household (484.16 and higher) and life insurance premiums (94.26 and higher). They har lower monthly salaries (2028.33 and lower) therefore it seems as if this group mostly has basic insurance which is household and life insurance.

Recommended Strategies:

Considering that these customers have the highest monetary value compared to all other groups, it is natural that ABC wants to acquire more customers with the same characteristics.

However, it is important to note that these customers are outliers, so strategies should be implemented with parsimony.

6. EVALUATION

The project was able to develop a customer segmentation and describe the ideal customer for the company using techniques such as principal components analysis and k-means. As a result, five customer segments were presented in this analysis: Seniors, Blue-Collars, Gold Motor, Well-Educated in Region 4, and Diamond; where each one have very unique characteristics that can be used by ABC Insurance Limited to create a more data-driven marketing approach, that can result in the retention rate of customers with high monetary value by 10% and increase the acquisition of customers by at least 15% in the next two quarters.

In order to achieve these goals, it is suggested that the company creates specific marketing campaigns for each group. For instance, the Gold Motor segment has the highest premium motor value. Campaigns for this group can focus on advertising tiered insurance packages, where tier 1 has a low coverage and tier 3 a high coverage, for example. In addition, since they are interested in insuring their vehicles, they might be interested in automotive related content, which could be an opportunity for ABC to educate customers in order to keep claims rate as low as possible while increasing the retention rate.

When considering the outlier clusters, which have the highest monetary value, it is recommended that ABC takes a more conservative approach initially, exploring this segment and its nuances in depth, since their representativity is relatively lower compared to the other segments, meaning that the risk of non-conversion could be higher, which would increase the cost of acquisition significantly. One interesting approach for this segment could be to create a set of tiered products that includes both life and household insurances. Once there is a conversion, the customer is in the database and other products and services can be offered to them in the future, or even an insurance upgrade (based on the tiers) with a higher coverage.

7. DEPLOYMENT

Once the model is approved by the stakeholders, the last step in the data mining process is to deploy the models to a production environment. Deployment is important because it makes the models available to users so that they can use them to create predictions and make data-driven business decisions. The results of the analysis will be presented to the marketing department of ABC Limited so that they can apply different marketing strategies depending on the type of customers they have. In applying these different strategies the company should also try to understand whether they help to increase their performance. Since customer profiles can vary over time and even as a result of the strategies implemented, it is important that the company provides an updated dataset quarterly in order to always have the most up-to-date campaigns and decisions.

Deployment planning

Following the successful outcome of this project, a deployment plan needs to be in place. From the results of the customer segmentation and from the results learnt from the raw dataset the marketing department will not be the only department that will benefit from the results of the project but other departments too. As this project will not end once the deployment is done, constant revisions will need to be done in order to always have accurate results. For this project this is how the deployment will be planned in order to make sure that the information reaches the correct departments.

Stakeholders/Board of directors: Take the stakeholders through the entire project and present the findings and the recommendations. Once approval from the top management is granted and they are satisfied with the results presented, the other affected departments can then be informed.

Marketing department: This is the department that initially requested the customer segmentation, therefore the results of the clustering and the different groups are shared with the department. Recommendation on ideal marketing strategies will be shared with the department based on the characteristics of each group as explained in cluster interpretation.

Policy writers: This department could also benefit from the results of the customer segmentation, they could use the clustering results to try to make adjustments to their policies as well as to create attractive packages for clients, especially clients in clusters two and three.

Data capturers: For this department the results of the project/clustering would not have an impact on them as much. However, because the project will not end after deployment the dataset that will be used going forward will need to be maintained. For this department the data understanding and data preparation results could be shared with them, to highlight the issues experienced such as deplications found in the dataset or so obvious data capturing errors.

8. CONCLUSION

Following the CRISP-DM model, the aim of this project was to do a customer segmentation of the clients of ABC Insurance Limited. The purpose of the customer segmentation was to assist the Marketing Department of the company to better understand their different customer profiles in order for the department to tailor unique marketing strategies for each group of clients accordingly.

To do the customer segmentation, the k-means algorithm was used to carry out the clustering. A dataset of the company's clients was provided by the company. Before the k-means algorithm could be applied to the dataset there were some stages of transformation that the data needed to go through before the clustering algorithm could be applied. The transformation from the raw dataset to the final dataset can be found under the sections data understanding and data preparation in this report. Following the transformation of the dataset, it was split into two parts, the first being the data without outliers and the second was the data consisting of only the outliers that remained even after data cleaning.

The first dataset which is the data without outliers resulted in 4 clusters being formed. From the interpretation of these clusters each one was considered to be unique of the other clusters which was considered a good thing because different marketing strategies could be applied to each unique cluster.

For the outliers dataset three clusters were formed according to the results from the clustering however, after having reviewed the three clusters there was no uniqueness found amongst the clusters therefore these clusters would be presented as a single cluster to the marketing department.

To conclude, following the outcomes of this project, the project can be considered to be successfully completed as it has addressed both the business and the data mining objectives.

9. REFERENCES

1. **Garbade, Dr. Michael J. 2018.** Understanding K-means Clustering in Machine Learning. Available on https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
2. **Onnen, H**. 2021. Elbows and Silhouettes: Hands-on Customer Segmentation in Python. https://towardsdatascience.com/elbows-and-silhouettes-hands-on-customer-segmentation-in-python-66c2e794c552
3. **Qualtics XM**. What is cluster analysis and when should you use it? Available on: https://www.qualtrics.com/uk/experience-management/research/cluster-analysis/
4. **Tam, Adrian. 2021.** Principal Component Analysis for Visualization. Available on: https://machinelearningmastery.com/principal-component-analysis-for-visualization/
5. **Upland Software.** The Importance of Customer Segmentation. Available on: https://uplandsoftware.com/bluevenn/resources/blog/the-importance-of-customer-segmentation/

10.APPENDIX

10.1 Appendix A: Description of the Variables

♦ **Customer ID (CustID),** a numerical variable, has 10,296 unique observations that vary from 1 to 10,296, which is a strong indicator that there are no duplicated customers. This is also a good indicator that no customers were deleted from the dataset, which suggests that there are no missing customers. Since the objective of this project is to segment customers, this variable is irrelevant for the analysis and can therefore be removed.

♦ **Year of the customer's first policy (FirstPolYear),** considered the first year as a customer, is a numerical variable. Its inputs vary from 1974 to 53784, which means that there is at least one incorrect input. After analyzing if there are any values higher than 2016, the current year of the database, it is possible to identify only one row in this condition, which is exactly the year of 53784. In addition to one incorrect row, there are 30 missing values that need to be dealt with accordingly.

♦ **Birth year of the customer(BirthYear**), just as the variables above, this is a numerical column. The minimum year register was 1028, which is certainly a faulty input that needs to be treated properly. Furthermore, the birth year should always be smaller than the fist year of the policy, since purchases cannot be made by or for unborns. There are 1,997 rows that need to receive proper treatment. Moreover, there are 17 missing values in this column.

♦ **Academic degree (EducDeg),** a categorical variable, has only four categories and 17 missing values, which is a good indicator that this column is well distributed and somewhat balanced. In terms of missing values, there are many reasons why a dataset could contain null entries, such as observations that were not recorded or perhaps the customer's education did not fall into any given category. Additionally, k-Means algorithm is not directly applicable to categorical data, which means this variable needs to be encoded.

♦ **Monthly gross salary (MonthSal),** a numerical variable, presented 36 null values. The minimum salary observed is 333 euros while the highest salary is 55,215 euros. Since there is an enormous discrepancy between the observed values, it is important to carefully analyze the distribution of this variable, as there is a high possibility that one of the values, either the minimum or the maximum, is an outlier (the latter is more probable).

♦ **The geographical code (GeoLivArea),** is a categorical variable in a numerical format that represents where the customer lives. The descriptive table suggests that there are only 4 possible outcomes: 1, 2, 3 and 4; and it also indicates that this column is well distributed. Since this is already an encoded categorical variable, it is necessary to analyze if the best encoding approach was applied.

♦ **Children** indicates if the customer has children or not, where 0 is equal to False, meaning that the customer does not have children, and 1 that is equal to True, meaning that the customer does have children. There are 21 missing values in this column.

♦ **Customer monetary value (CustMonVal)** is a numerical variable that is calculated as follows: (annual profit from the customer) x (number of years since a customer) - (acquisition cost). Its values vary from -165680.42 to 11875.89, which could indicate the presence of outliers or errors, especially when the quantiles values are taken into account. There are no null values in this column.

♦ **The claims rate (ClaimsRate)** is the amount paid by the insurance company in the last two years. Its numerical values vary from 0 to 256.2. Given the difference between the minimum and maximum observed values and its quantiles, it is possible to assume that this column is not well distributed, which might indicate the presence of outliers or errors. In addition, there are no missing values in this column.

♦ **Premiums in the Line of Business (LOB) Motor (€) (PremMotor)** is a numerical field that contains 34 missing values. Its values vary from -4.11 to 11604.42, which, again, suggests the presence of outliers or errors (quartille 75% value is equal to 408.3).

♦ **Premiums in the LOB Household (PremHousehold)** is a numerical column. The minimum value observed is -75.00 euros, while the maximum is 25048.80 euros. Taking into account its quantiles, it can be said that this column probably contains outliers or errors that must be treated accordingly. There are no missing values in this column.

♦ **Premiums in the LOB Health (PremHealth)** is a numerical column with 43 missing values. The minimum value observed is -2.11 euros, while the maximum is 28272.0 euros. Its quantiles suggest that the data is not well distributed, and that outliers or errors may be present.

♦ **Premiums in the LOB Life (PremLife),** just as all the insurance variables mentioned above, this numerical variable may contain outliers or errors, as its 75% quantile is equal to 47.23 and the maximum observed value is equal to 174.70. This is the variable that contains the highest number of missing values, totalling 104, which represents only 1.01% of the total observations in the dataset.

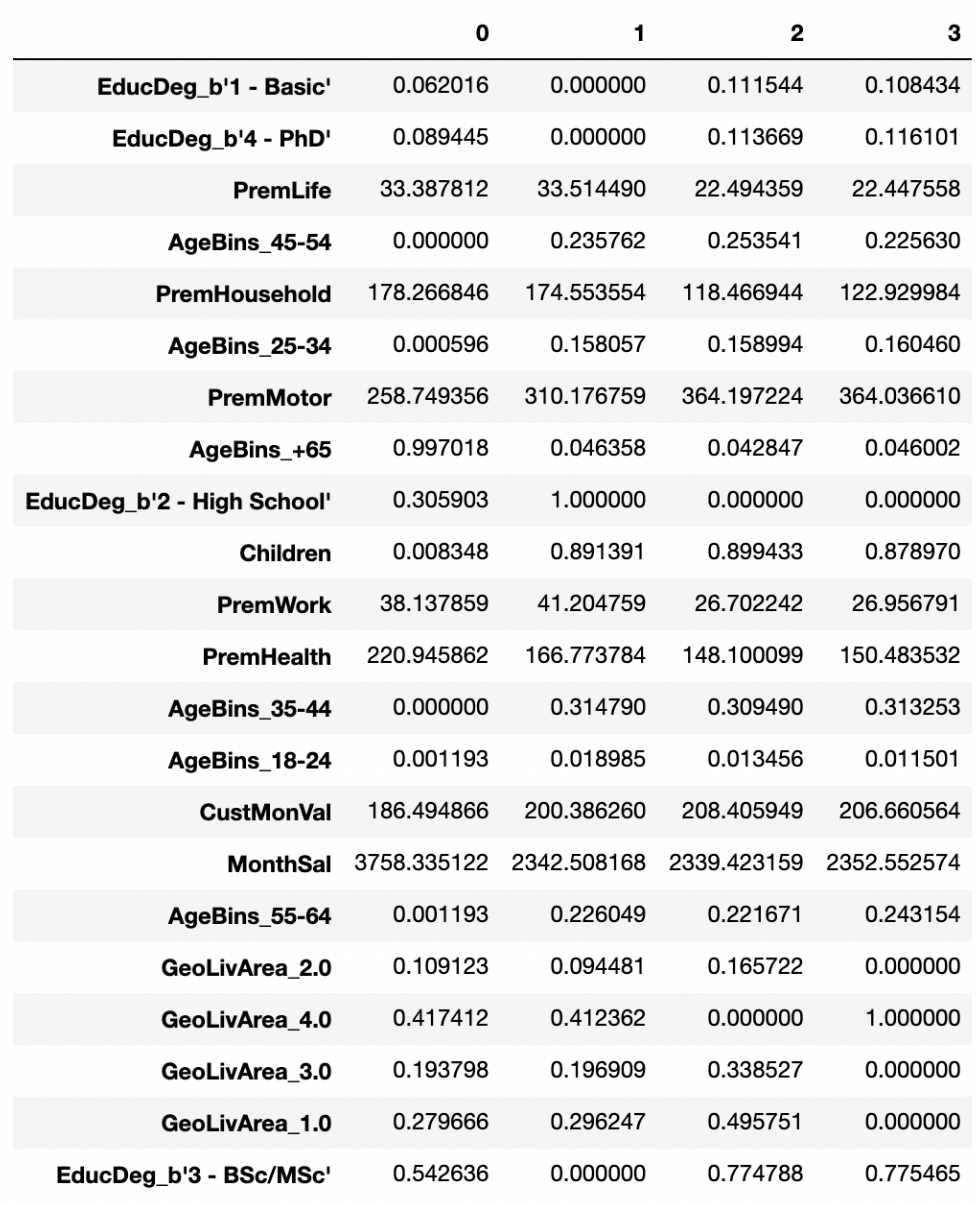
♦ **Premiums in the LOB Work (PremWork),** like the other insurance variables above, this is a numerical variable that does not seem to be well distributed and may contain outliers. Its values vary from -12.0 to 1988.7, while the 75% quantile is equal to 56.79. There are 86 missing values in this column.

10.2 Appendix B: Project Plan

Project Plan Overview		
Phase	**Time Frame**	**Responsibility**
Business Understanding Business objectives Assess situation Data mining goals Project plan	1 week	Project team members
Data Understanding Data collection and description Exploratory data analysis Data quality verification	3 weeks	Project team members
Consultation	Consultation on data quality issues	Project team members and stakeholders
Data Preparation Selecting data Cleaning data Data construction Data formatting Data normalization	8 weeks	Project team members and stakeholders
Data Modeling Principle components K-cluster analysis	4 weeks	Project team members
Consultation	Consultation on data preparation and clustering results	Project team members
Evaluation	1 week	Project team members
Deployment	Continuous	Project team members and stakeholders

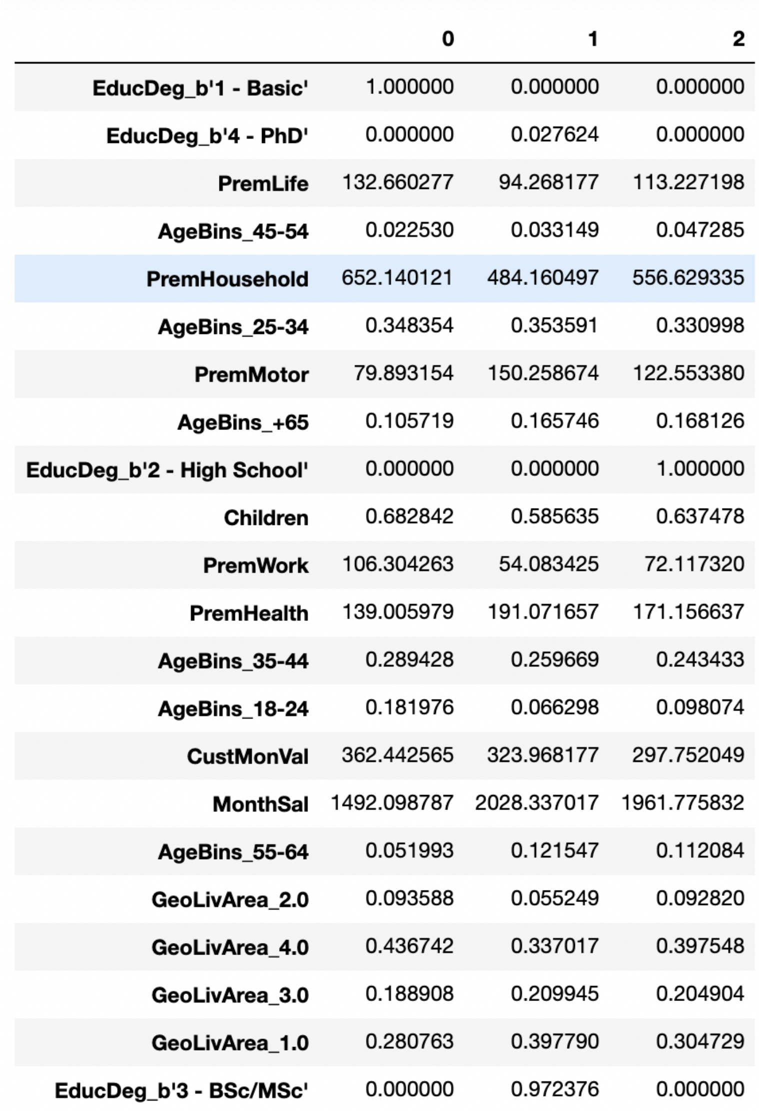
10.3 Appendix C: Clustering Output tables

Clustering Results - Dataset without outliers



	0	1	2	3
EducDeg_b'1 - Basic'	0.062016	0.000000	0.111544	0.108434
EducDeg_b'4 - PhD'	0.089445	0.000000	0.113669	0.116101
PremLife	33.387812	33.514490	22.494359	22.447558
AgeBins_45-54	0.000000	0.235762	0.253541	0.225630
PremHousehold	178.266846	174.553554	118.466944	122.929984
AgeBins_25-34	0.000596	0.158057	0.158994	0.160460
PremMotor	258.749356	310.176759	364.197224	364.036610
AgeBins_+65	0.997018	0.046358	0.042847	0.046002
EducDeg_b'2 - High School'	0.305903	1.000000	0.000000	0.000000
Children	0.008348	0.891391	0.899433	0.878970
PremWork	38.137859	41.204759	26.702242	26.956791
PremHealth	220.945862	166.773784	148.100099	150.483532
AgeBins_35-44	0.000000	0.314790	0.309490	0.313253
AgeBins_18-24	0.001193	0.018985	0.013456	0.011501
CustMonVal	186.494866	200.386260	208.405949	206.660564
MonthSal	3758.335122	2342.508168	2339.423159	2352.552574
AgeBins_55-64	0.001193	0.226049	0.221671	0.243154
GeoLivArea_2.0	0.109123	0.094481	0.165722	0.000000
GeoLivArea_4.0	0.417412	0.412362	0.000000	1.000000
GeoLivArea_3.0	0.193798	0.196909	0.338527	0.000000
GeoLivArea_1.0	0.279666	0.296247	0.495751	0.000000
EducDeg_b'3 - BSc/MSc'	0.542636	0.000000	0.774788	0.775465

Clustering Results - Dataset outliers only



	0	1	2
EducDeg_b'1 - Basic'	1.000000	0.000000	0.000000
EducDeg_b'4 - PhD'	0.000000	0.027624	0.000000
PremLife	132.660277	94.268177	113.227198
AgeBins_45-54	0.022530	0.033149	0.047285
PremHousehold	652.140121	484.160497	556.629335
AgeBins_25-34	0.348354	0.353591	0.330998
PremMotor	79.893154	150.258674	122.553380
AgeBins_+65	0.105719	0.165746	0.168126
EducDeg_b'2 - High School'	0.000000	0.000000	1.000000
Children	0.682842	0.585635	0.637478
PremWork	106.304263	54.083425	72.117320
PremHealth	139.005979	191.071657	171.156637
AgeBins_35-44	0.289428	0.259669	0.243433
AgeBins_18-24	0.181976	0.066298	0.098074
CustMonVal	362.442565	323.968177	297.752049
MonthSal	1492.098787	2028.337017	1961.775832
AgeBins_55-64	0.051993	0.121547	0.112084
GeoLivArea_2.0	0.093588	0.055249	0.092820
GeoLivArea_4.0	0.436742	0.337017	0.397548
GeoLivArea_3.0	0.188908	0.209945	0.204904
GeoLivArea_1.0	0.280763	0.397790	0.304729
EducDeg_b'3 - BSc/MSc'	0.000000	0.972376	0.000000