SUPERMARKET NOCHE

Unsupervised learning

Executive Summary

Four clusters were obtained that achieve good performance, with elements similar to each other and different from those of other clusters.

The interpretation of each cluster, based on the results obtained, is as follows:

Cluster 0 has a demographic of highly educated, single males who are likely to be working professionals.

Cluster 1 has a demographic of low-educated, married individuals who are mostly unemployed.

Cluster 2 has a demographic of highly educated, single males who are likely to be self-employed entrepreneurs.

Cluster 3 has a demographic of low-educated, married individuals who are likely to be employed in medium-sized cities.

The recommendations for the Noche supermarket marketing team are as follows:

1-Consider the types and interests particular to each of the clusters

2-Consider clusters 0 and 2 as possible premium audiences, and depending on spending, offer special offers intrinsic to the particular taste of each of the groups.

3-Make more affordable offers for audiences 1 and 3, as well as offers on products related to couples, or more commonly consumed by couples.

5-Consider future analyses, with more experimentation at the time of segmentation, and possible corrections in the data taken as mentioned in the quality problems section.

Table of Contents

Content

Introduction

The purpose of this report is to develop a tool for the marketing team of Supermarket Noche that allows customers to be grouped according to certain similar characteristics in order to identify needs in certain products or services and from this, give a competitive advantage in the market. The information that was analyzed comes from the database of 2000 customers, obtained from the supermarket discount system. The characteristics included in the Database regarding clients are the identifier, sex, marital status, age, educational level, annual income in dollars, occupation and the size of the city where they live.

To achieve this objective, different procedures were carried out that allowed the understanding of the data, the quality problems and the correct preparation of the same, modeling with the appropriate method and finally testing the results obtained. With this process, the contents seen in the Department of Technology for Data Exploitation of the Universidad Tecnológica Nacional - Facultad Regional Concepción del Uruguay were put to the test, so it is important to highlight that our analysis was limited to the available data and the tools used. in the chair

Regarding modeling x-means was used to first find the optimal number of clusters. Within x-means, k-means was selected as the unsupervised learning method to find cluster centroids. Other approaches were made considering a certain subjectivity about target audience campaigns, thus obtaining larger segmentations.

Subsequently, we present the conclusions that can be drawn from the results of the model, and the recommendations for the marketing team.

The report presents this introduction in the first instance in order to contextualize the reader about the subject, the problem and the procedures carried out to achieve the stated objective.

Discussion

Data understanding

The source of the data for this project is a file named "supermercado_noche.csv", which is housed within a Google Drive folder that was provided by the professor.

Attached to this file there is also a table 1 which briefly explains each attribute.

Variable	Type	Values	Description
ID	Numeric	Integer	Unique identifier per client
Sex	Nominal	{0,1}	Biological sex of the client. 0 = male; 1 = female.
Marital status	Nominal	{0,1}	Civil status. 0 = single / 1 = not single.
Age	Numeric	Integer	Age in years, calculated as the current year minus the year of birth at the time of dataset creation.
Education	Nominal	{0,1,2,3}	Client's highest educational level 0=no education / 1=high school / 2=university / 3=postgraduate degree
Income	Numeric	Real	Estimated annual income in dollars, according to the client.
Occupation	Nominal	{0,1,2}	Type of client's employment 0=unemployed / 1=employed / 2=self-employed.
Settlement size	Nominal	{0,1,2}	Size of the city where the customer lives. 0=small / 1=medium / 2=large

Table 1-Description of variables

As we can see, the dataset provides information about Noche supermarket customers, including their unique ID, sex, marital status, age, education level, estimated income, type of employment and size of t the city where they live. This information can be used to design personalized marketing campaigns.

The dataset consists of 2000 observations.

The attributes sex and marital status were loaded in rapidminer as binomial type. In turn, the attributes education, occupation, and settlement size were loaded as polynomials. The other attributes have the type specified in table 1, and their units are [years] and [$] accordingly.

For the following histograms, where applicable the number of bins was determined using the "Sturge's Rule", due to the number of observations. In this case 12 is the number of bins.

There are no duplicate rows.

Sex

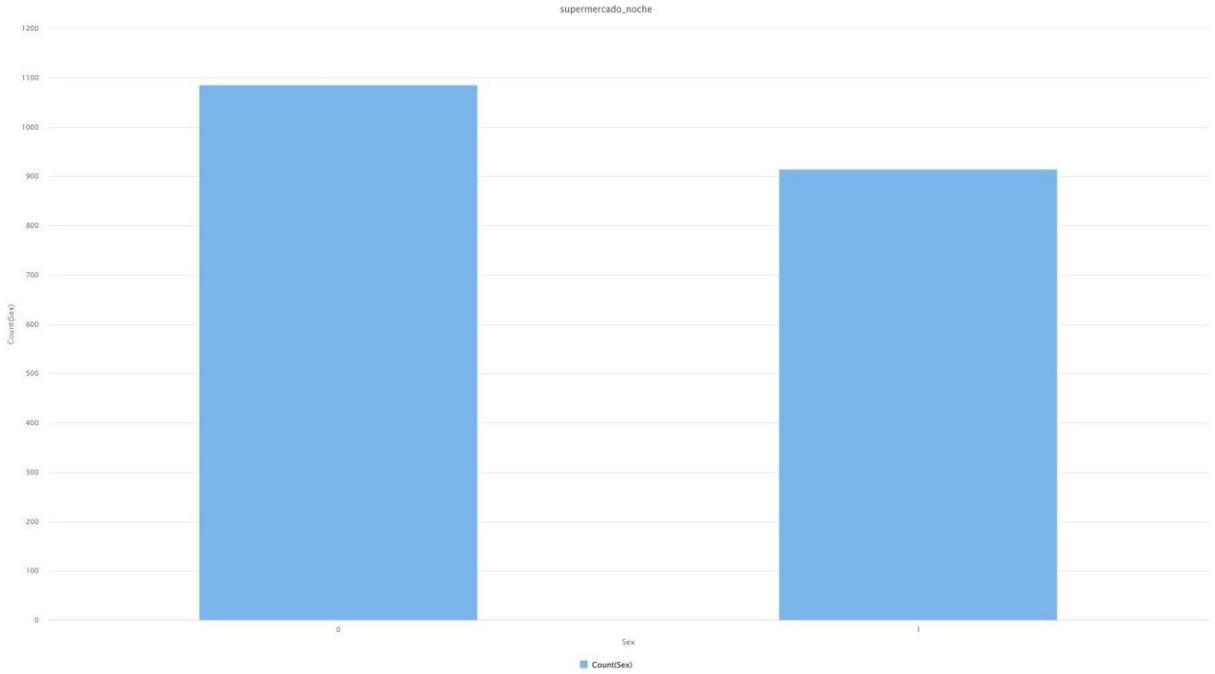

Fig 1- Bar column for Sex

As can be seen in Fig 1, there are more men (0) than women (1). In total there are 1086 males and 914 females.

There are no missing values in the Sex variable.

Marital Status

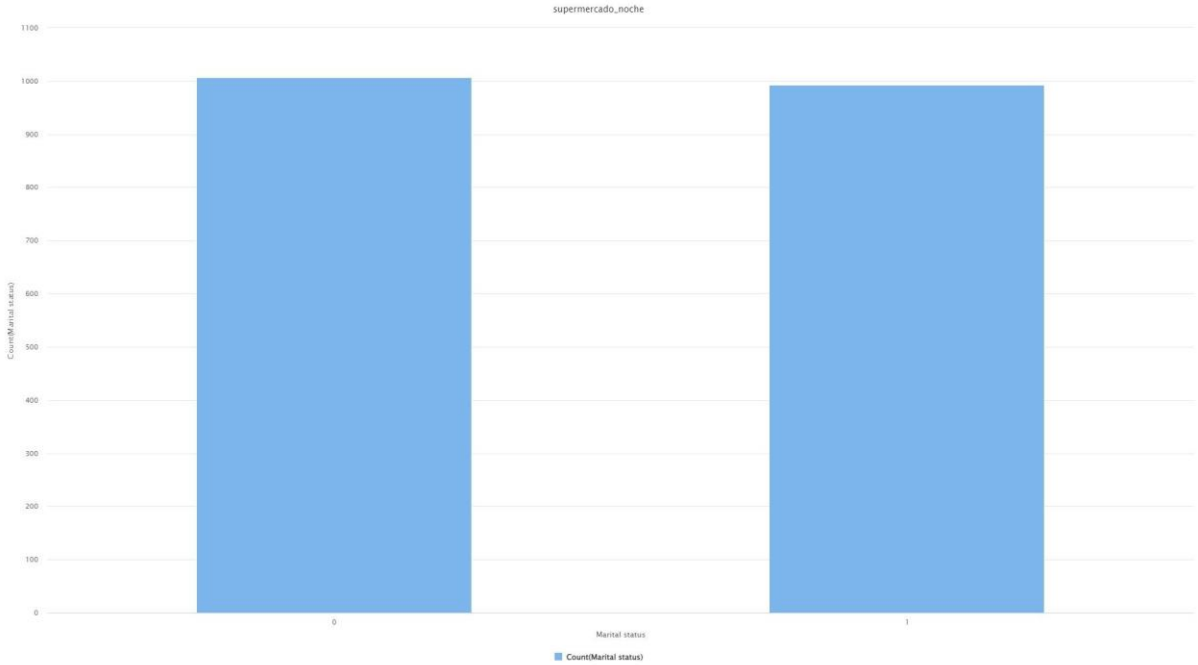

Fig 2- Column Bar plot for marital status

As illustrated in Fig 2, the number of single (0) and non-single (1), is quite close, being 1007 and 993 observations, respectively.

No missing values in this case.

Age

Name	Type	Missing	Min	Max	Average	Deviation
Age	Integer	0	18	76	35.91	11.719

Table 2-Age table

From table 2:

The variable is a discrete numeric variable represented as an integer.

Missing values for the age variable are not present.

The minimum value for age is 18 years, and the maximum value is 76 years.

The average or mean of age is 36 years.

The standard deviation of the age variable is 12 years.

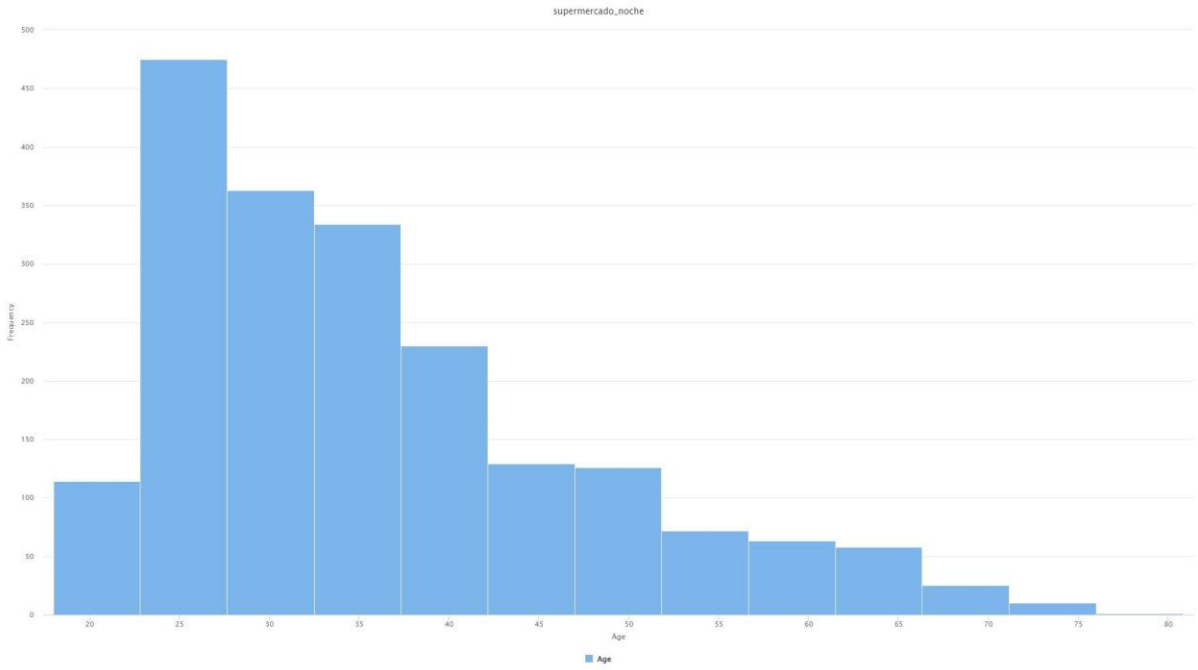

Fig 3-Age histogram

Looking at Fig 3, the histogram appears to have a skewed right, unimodal distribution with no gap of bins. There is one peak with 475 observations between the ages of 23 and 27. The data has a range of 4 years.

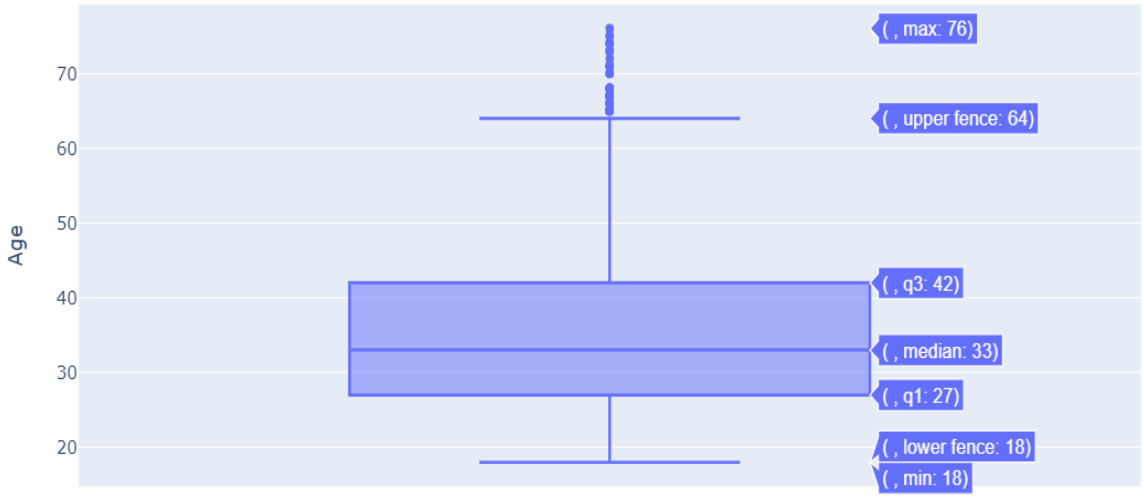

Fig 4-Age BoxPlot

Analyzing fig 4 we can say that the distribution of the variable is skewed to the right because the median 33 years is closer to the lower quartile 27 years than the upper quartile 42 years, and the maximum value is far from the upper quartile and the minimum value from the lower quartile. There appear to be 11 outliers from about 63-64 years of age.

Education

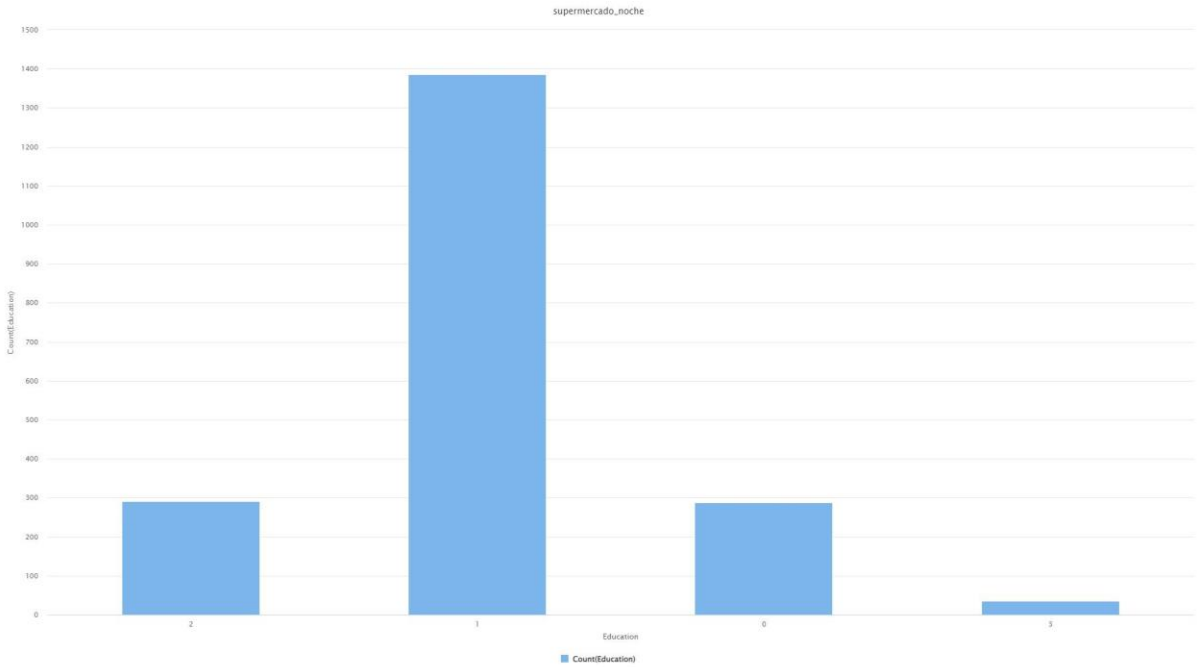

Fig 5-Column bar for Education

Fig. 5 shows that the educational level most frequently reached in general is high school with 1386 observations. In second place is university, almost tied with no education with 291 and 287 observations respectively. The rarest case is that someone has achieved a graduate degree, with only 36 observations.

Education is not missing any values.

Income

Name	Type	Missing	Min	Max	Average	Deviation
income	Real	0	35832	309364	120954.419	38108.8

Table 3-newspaper

From table 3:

The variable is a continuous numeric variable represented as a real number.

There is no missing value.

The minimum income is $35832, and the maximum is $309364.

The average is $120954.4.

The standard deviation is $38108.8.

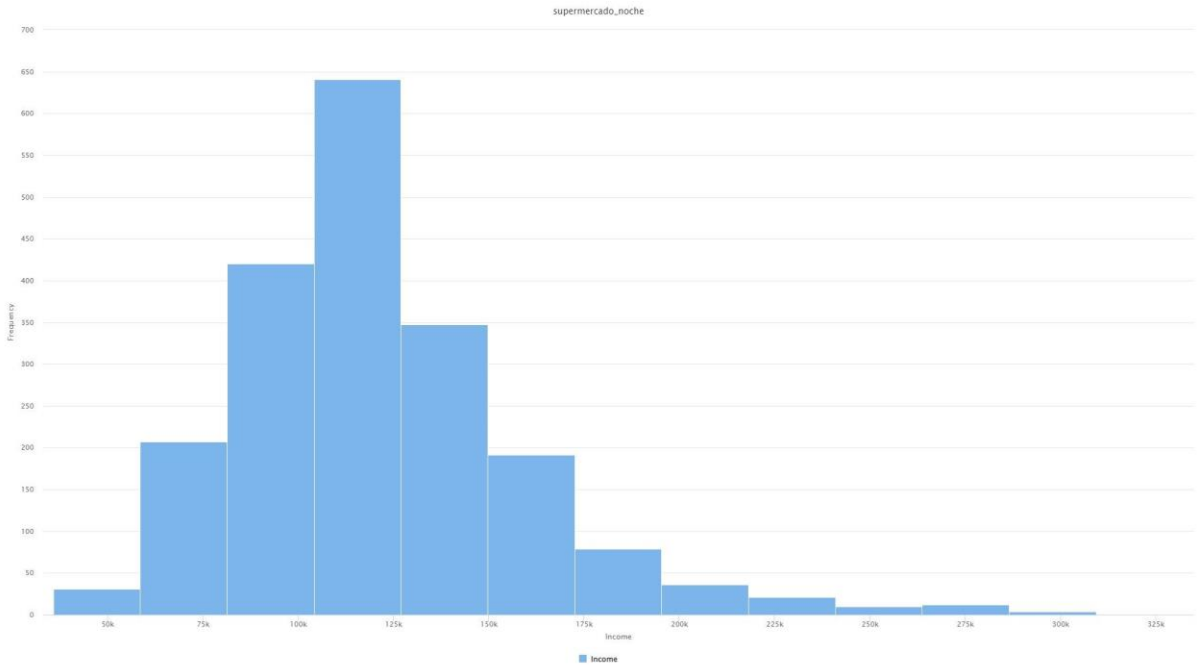

Fig 6-Income Histogram

In fig 6 we can see that the histogram has a symmetric, unimodal distribution without a gap. There is one peak with 641 observations between $104215 and $127009. The data has a range of $273532.

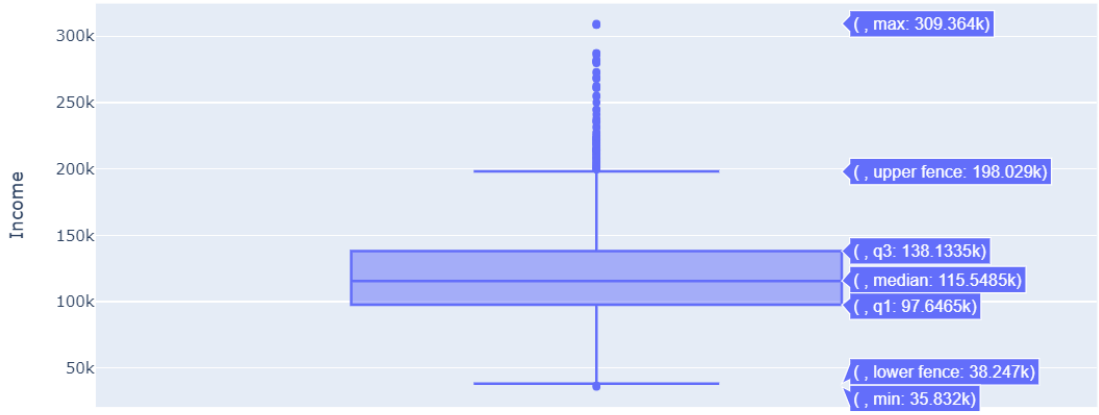

Fig 7-Income BoxPlot

Fig 7 provides information on income. Analyzing it, it can be concluded that the median $115548, is closer to the lower quartile $97629 than the upper quartile $138194 and the maximum value is far from the upper quartile and the minimum value from the lower quartile. There are many outliers starting at approximately $200,000 and one before about $40000.

Occupation

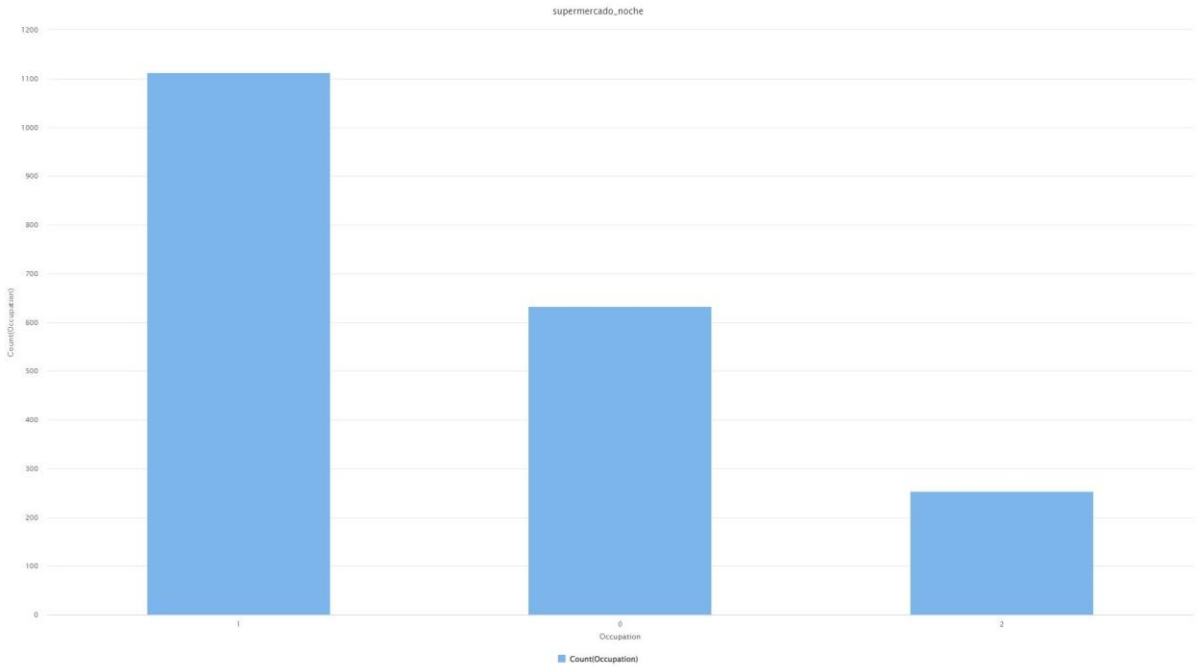

Fig 8- Bar column for Occupation

From Fig. 8 we can deduce that the majority of customers, 1113 are employees, 663 are unemployed and 254 are self-employed.

No missing values.

Settlement size

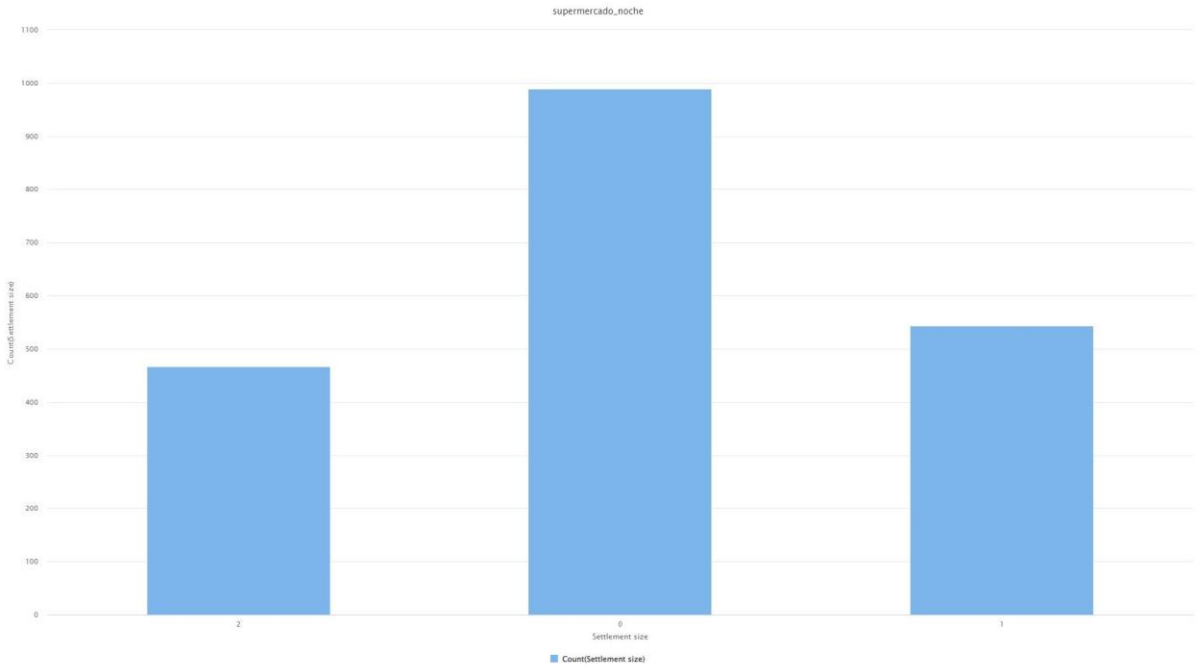

Fig 9-Bar column for Settlement size

According to Fig 9, 989 clients live in a small city, 544 in a medium-sized city, and 467 in a large city.

No missing values for this variable.

Quality problems

1 -The "Sex" attribute in the dataset may not be inclusive enough as it only includes the binary options of male and female, which may not be relevant in all countries and cultures, potentially resulting in the loss of certain segments of the population that identify as non-binary or outside of the male/female binary and, therefore, may not accurately represent the diversity present in the stakeholder country. In general, for unsupervised learning tasks, it's better to use variables that are more informative and less redundant, as they can improve the accuracy of the resulting models.

2-Marital status could provide us with more information to identify well differentiated segments if instead of being binomial it was polynomial.

3- The age may not correspond to the actual age of the client as it is calculated at the time of dataset creation.

4- Settlement size does not clarify the metric used to identify the difference between a small, medium, and large city, if it is at the client's discretion, it could become an ambiguous attribute.

5- Quite a few outliers were found during data understanding in the age and income variables.

Overall, the data quality is good.

Data preparation

As discovered during the understanding, the dataset provided is of very good quality in general. For this study we will assume that items 1 to 4 of the data quality do not influence the population/region for which the stackholders are interested.

To mitigate the fifth quality problem, a data treatment will be performed to ensure a better clustering quality because the k-means algorithm is sensitive to outliers because such objects are far away from the majority of the data, and thus, when assigned to a cluster, they can dramatically distort the mean value of the cluster. This inadvertently affects the assignment of other objects to clusters (Jiawei Jan, 2012).

First the ID was removed to avoid problems during normalization.

The normalization technique to be used only works with numerical data, so does the k-means method which will be used in modeling, it can be applied only when the mean of a set of objects is defined, this is not the case for the data we were working with, in which nominal attributes are found. For a dirtier set you could use one-hot encoding, with which nominal attribute values are correctly mapped to their respective numeric values, but here the dataset was properly assigned to a numeric value (0,1,2, etc) so it was enough to reload it in Rapidminer and assign another type of attribute (integer), to the variables Sex, Marital status, Education, Income, Occupation, Settlement size.

Then the decision of normalizing the dataset was taken. All this can be seen in Fig 11. Normalization is a technique used to rescale numeric variables to a common scale, usually between 0 and 1 or -1 and 1, to remove any inherent differences in scale or range between the variables.

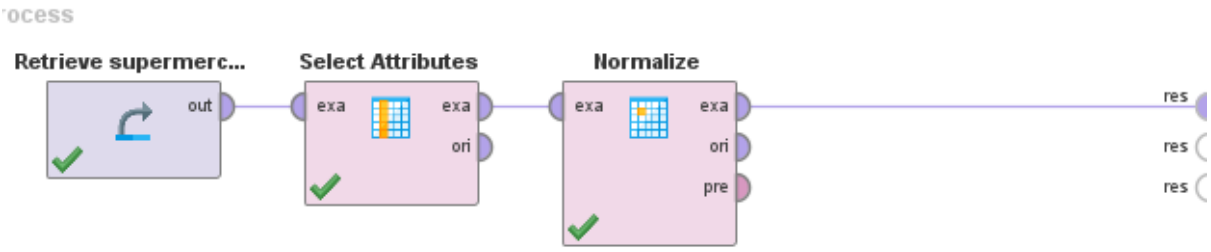

Fig 1- Data Preparation, in order type changes, ID discarded, normalization.

The following factors were considered for doing the normalization:

-The granularity (which is dropped during normalization) of the age variable will not be important. Because we have a medium sized dataset with many observations, the loss of information due to normalization is negligible.

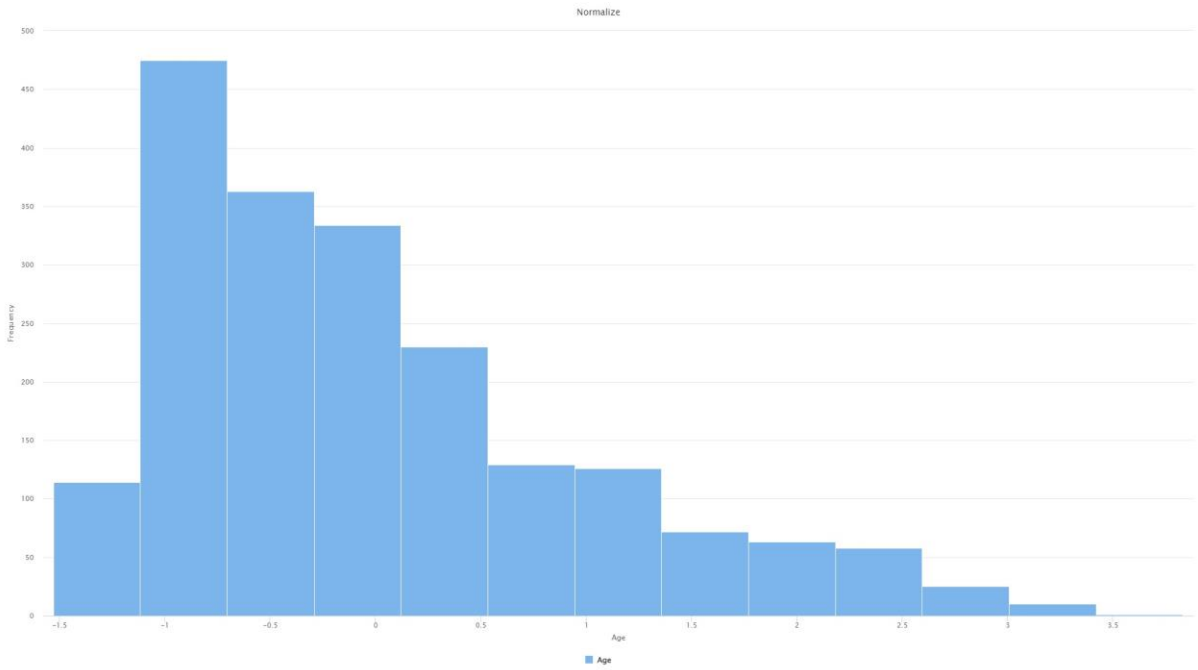

Fig 2-Histogram for Age after normalization

Fig 12 shows that the statistical properties of the normalized data are similar to those of the original data, because of this the loss of information due to normalization is likely to be low.

-It is a good idea to normalize because Age and Income have different ranges and units, Age has a range of 18 to 76 years and Income has a range of $35832 to $309364. Normalizing these variables can help to eliminate the effects of these differences in scale and units, making the data more comparable and

easier to analyze. Variables with a very large range of values compared to others, it may dominate the distance calculation in a clustering algorithm and cause other variables to be ignored. Normalizing the feature can help avoid this problem.

-Normalization can help to avoid having the outlier dominate the distance calculations between data points and skew the clustering results.

-For nominal variables Sex, Marital status, Education, Occupation and Settlement size, there is no inherent scale or range for these values in the same way as there is for continuous numerical variables.

Z-transformation was selected as the method because it is the standard method of RapidMiner. Also, we don’t have many outliers and z-transformation is good enough to rescaler the dataset. In this case, we don’t handle outliers, i.e. it's not necessary to consider outliers in the data’s composition because the distance of data when we plot (see figure 20, Section modeling) is highly near and it isn't considered an outlier.

Modeling

K-means may be a good option for this dataset because we seek to identifying well-differentiated segments of clients, where different demographics can be identified. These segments can help the Noche supermarket's marketing team design targeted marketing campaigns for each group, which can increase campaign effectiveness and customer satisfaction, thereby helping the supermarket gain a competitive advantage in the market. Additionally, K-means is easy to implement and interpret, making it a good choice for marketing teams with less experience in data analysis.

First aproach

“Holistically, K-means suffers from the following limitations:

-K-means is slow and it scales poorly with respect to the time it takes to complete each iteration.

-The number of clusters ‘K’ have to be pre-determined and supplied by the user.

-When confined to run with a fixed value of K, it empirically finds worse local optima than when it can dynamically alter K.

The solution for the first two problems and a partial remedy for the third is X-means. This method of efficient estimation of the number of clusters” (Gupta, 2021)

Because the number of clusters into which the clients should be divided is unknown, X-means will be used in RapidMiner.

Fig 13 shows the process performed.

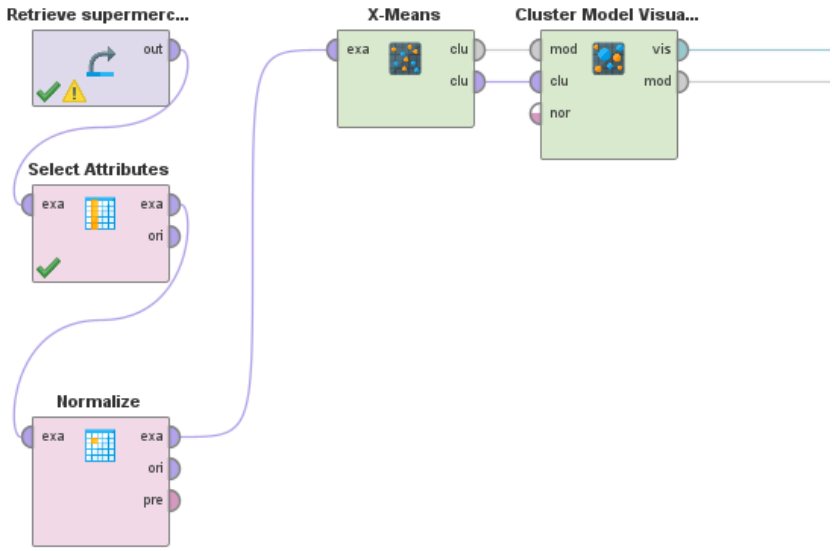

Fig 3- X-means clustering in RapidMiner with his corresponding visualzation

- k-min : 2
- k-max : 10
- determine Good star value: true
- measure types : NumericalMeasures
- numerical measure: EuclideanDistance
- clustering algorithm: KMeans
- max runs: 1000
- max optimization steps: 1000

Since the number of clusters is unknown, we selected the minimum possible value (2) and a maximum value of 10, between which x-means can return the optimal segmentation size in this range, on the other hand because it is a medium sized dataset, max runs was set to 1000 and max optimization steps 1000, to ensure a good segmentation.

The results obtained were as follows:

Two clusters were obtained.

-Cluster 0: 878 items

-Cluster 1: 1122 items

Atribute	**cluster_0**	**cluster_1**
Sex	-0.67030269476706	0.5245327682758162
Marital status	-0.43026056963280773	0.3366923174131966
Age	0.3646288626707982	-0.28533345938053517
Education	0.014500384984265606	-0.011347003579487884
Income	0.6551463897117311	-0.5126724867797683
Occupation	0.6891297810138954	-0.5392655505616804
Settlement size	0.7445407226763137	-0.5826263409178276

Table 1-Centroid table

The centroid table shows the average values for each attribute in each of the clusters generated by the X-means algorithm, one could understand it as the prototypical customer of a given cluster, for this case it could be extracted some general insights:

cluster_0:

The negative value for sex and marital status indicates a higher proportion of single men.

The age for this cluster falls in the bin between 37-42 years. People in this cluster have generally attained a high school level of education so far.

Their income ranges from $127009 to $149803.

Employees, due to the value obtained.

The result for settlement size falls between medium and large cities. This suggests that cluster_0 has a demographic of single, middle-aged males with a high school education attained, employees with slightly higher than average incomes, and living in a medium to large city.

cluster_1:

Positive values for sex and marital status indicate a higher proportion of non-single women in this cluster.

In the age group 32 to 37 years

People in this cluster have generally attained a high school level of education so far.

Their income ranges from $104215 to $127009.

A large proportion of unemployed.

The result of small cities for settlement size.

The cluster_1 demographics suggest that this cluster has a demographic of non-single females, entering middle age (generally younger than those in the previous cluster) with a high school education attained, unemployed with median income, and living in a small sized city.

This is complicated to see in Fig. 14 because we worked with normalized values.

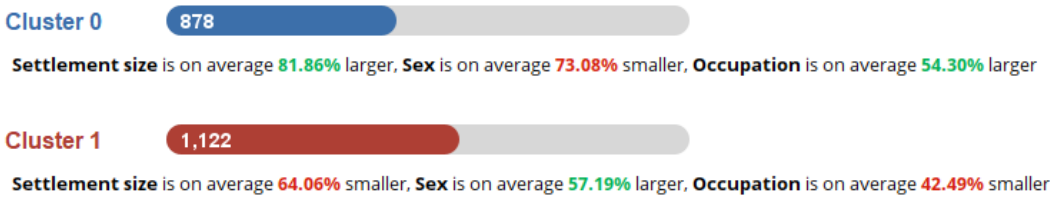

Fig 4- overview of the clusters

Fig 15 indicates that the main differentiators in these clusters were sex, settlement size, and occupation.

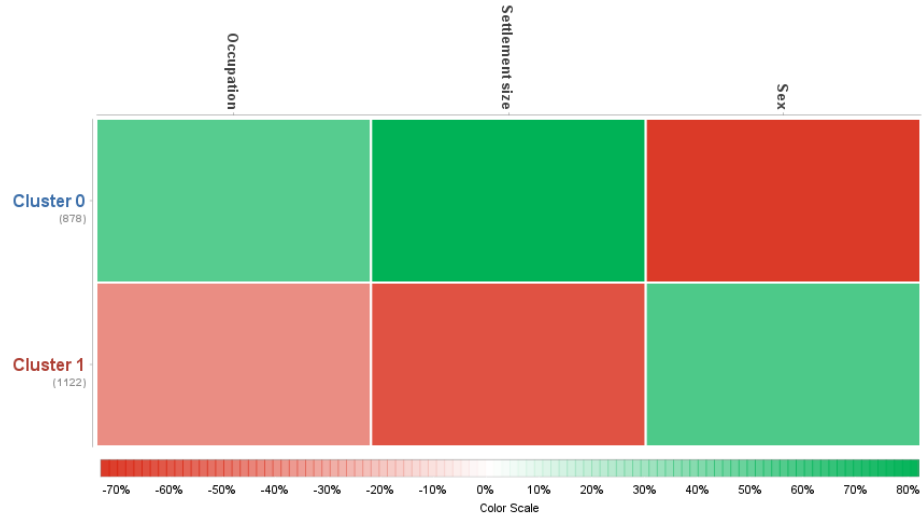

Fig 5-Heatmap of the clusters

And Figure 16 allows us to observe that indeed, the greatest differences between the clusters are between these three categories, and education is where they differ the least.

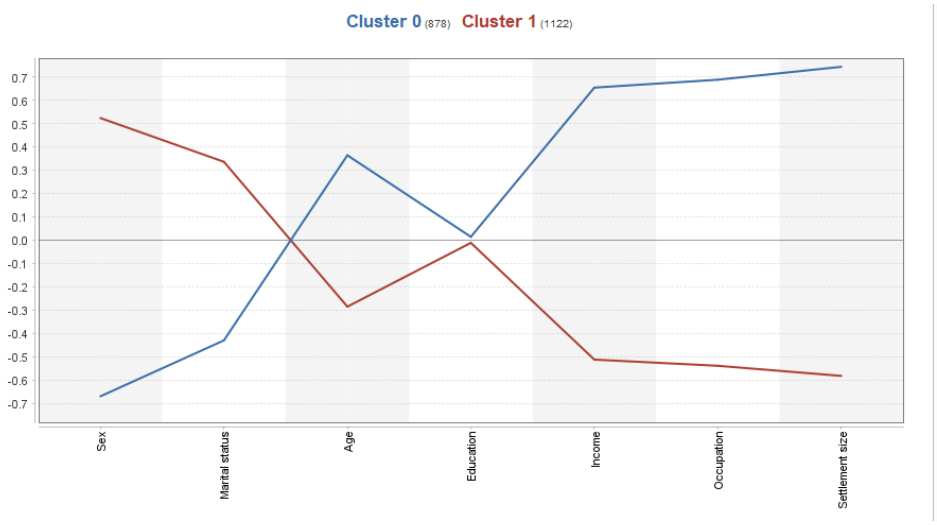

Fig 6-Centroid Chart

Validation and feedback for this first aproach

The performance of the model, was evaluated using the Cluster distance performance process as shown in Fig 16, evaluating it for the criteria:

-Avg. within centroid distance

-Davies Bouldin

Because the first measures how close the data points within each cluster are to the centroid of that cluster, and the second measures the similarity between clusters. Better results for these two metrics generally indicate better clusters.

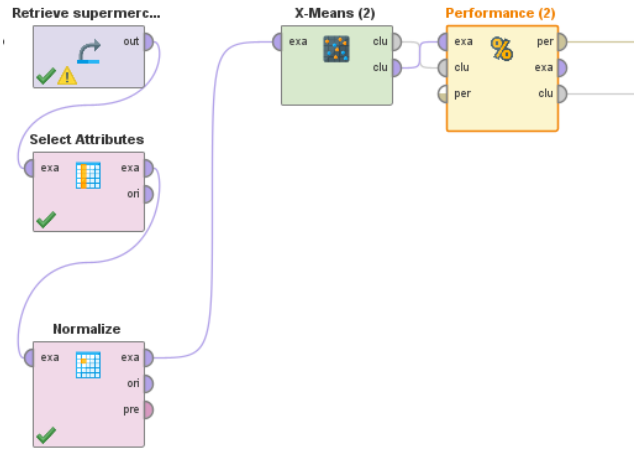

Fig 7-Persormance process

Avg. within centroid distance: -5.255

Avg. within centroid distance_cluster_0: -6.042

Avg. within centroid distance_cluster_1: -4.638

Davies Bouldin: -1.640

Although there is a clear separation between the clusters, the experimental results obtained with X-means still show some mixtures between the variables. Additionally, while the achieved segmentation is good, the marketing team could benefit from a larger number of clusters that would allow for even more specific advertising. Therefore, to address this concern, a second approach will be adopted that involves selecting a larger number of

clusters using techniques that strike a balance between cluster quantity and quality.

Second aproach

In a second approach, python was used as a comparison method. The Elbow technique was used together with k-means and the same parameters used before in the RapidMiner approach to determine the optimal number of clusters from the 'inertia' of the clusters.

The elbow technique consists of repeating a loop over the method chosen to cluster the data, and each step increases by one cluster. Then a plot of the 'inertias' is made and the point where the curvature of the graph ends is seen. Inertia is a Sum of squared distances of samples to their closest cluster center.

Here is the Elbow technique chart:

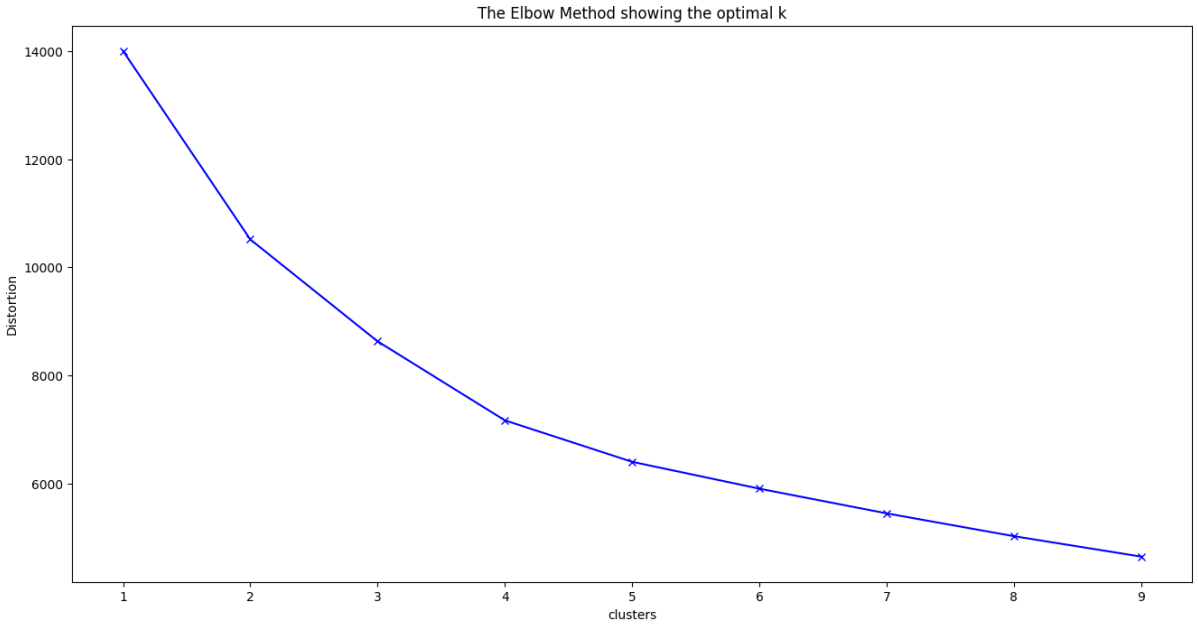

Fig 8-Elbow technique chart

It can be verified that there are two possible cut points, when there are two and four clusters. This is because the largest drop occurs in two clusters (with a 25% drop), but they are too few to discriminate between groups of data. At 4 (there is a 50% total drop) it is the last visible drop on the chart. That's because from 5 clusters forward we don't have significant variations in the drop value comparing them with these two values.

A correlation matrix was created between the data to take into account the fields with greater affinity and thus have a clearer view of the clusters.

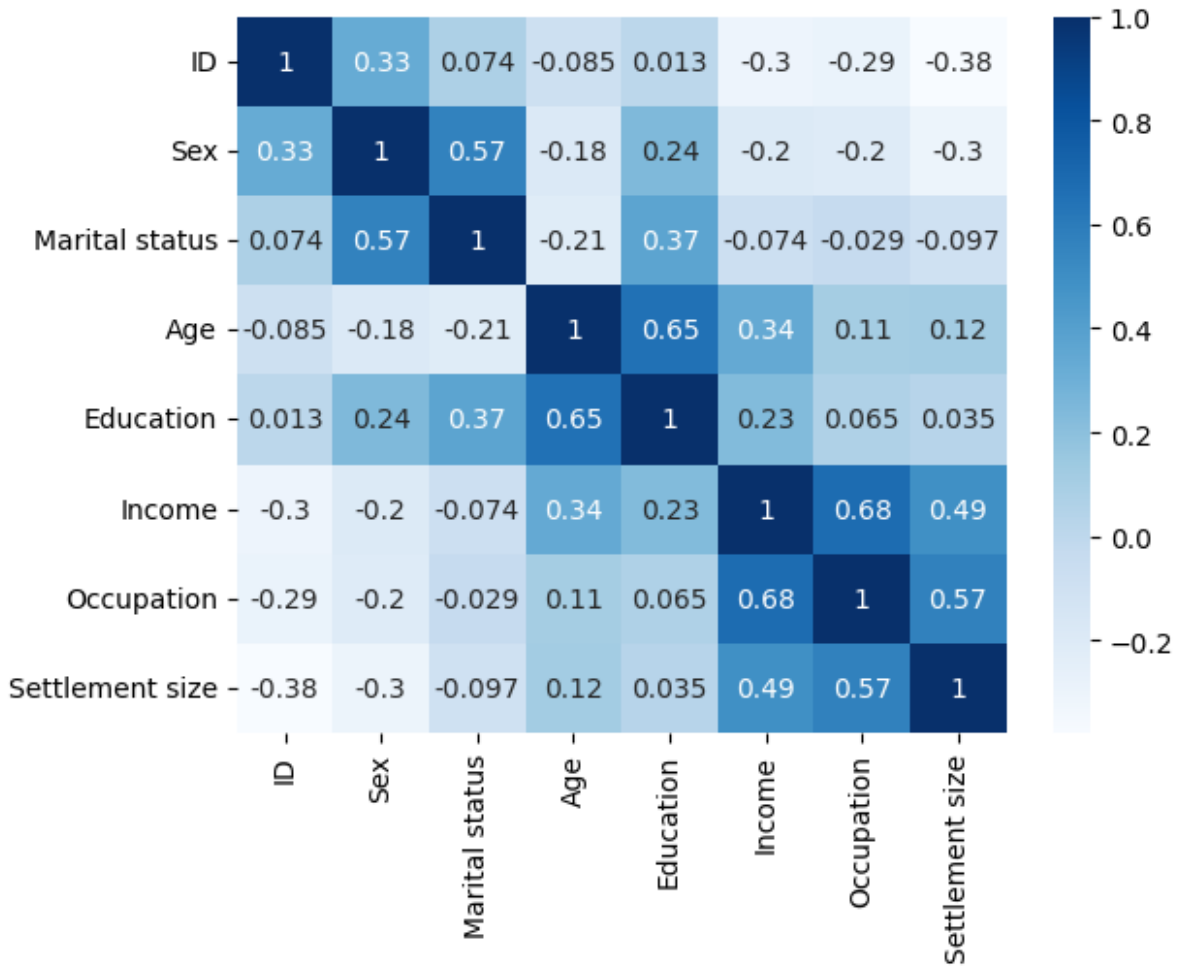

Fig 9-Correlation matrix

Looking ahead to the graphs, unfortunately many fields are nominal and little is done to determine their relationship with the clusters created by their values.

So, here are some sample graphs, and a separate graph of distributed values for age and income as they are the only non-nominal fields, you can get a clearer view of the distribution of clusters:

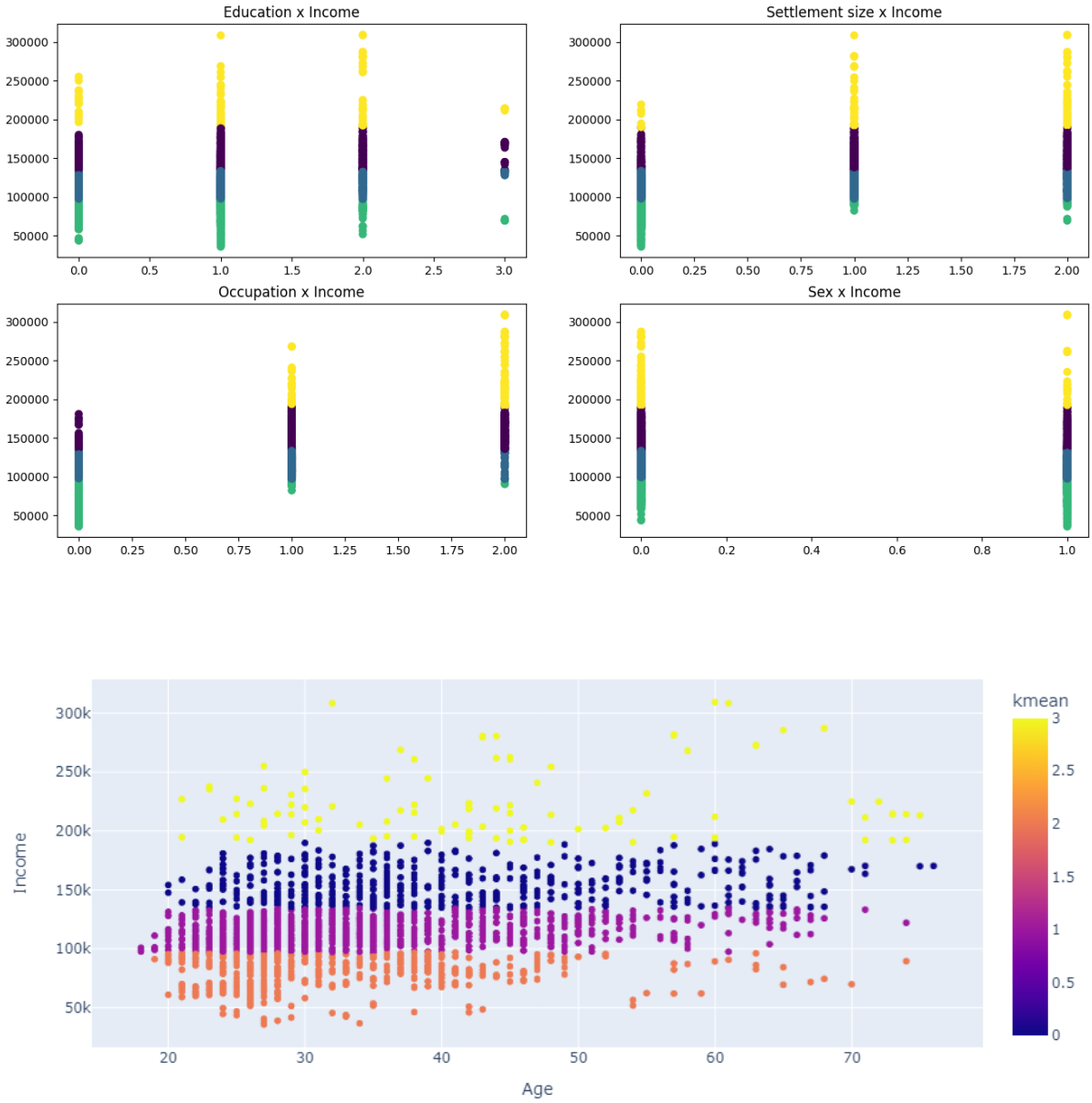

Fig 10-Scatter plot of the clusters

Verification/performance

This approach has more computational use than the first one, but it has a very interesting segmentation, mainly in relation to income. Here's the average for each field in each cluster:

Atribute	**cluster_0**	**cluster_1**	**Cluster_2**	**Cluster_3**
count	457	483	105	955
Sex	0,30	0,54	0,29	0,51
Marital status	0,44	0,51	0,40	0,53
Age	41	32	45	34
Age_min	20	19	21	18
Age_max	76	74	75	74
Education	1,19	0,89	1,34	1,00
Income	$154.151	$79.086	$225.125	$114.791
Income_min	$134.433	$35.832	$189.896	$96.952
Income_max	$189.166	$96.840	$309.364	$134.377
Occupation	1,25	0,15	1,77	0,83
Settlement size	1,24	0,16	1,47	0,71

Table 5- Centroid table

We can see here that the factor that differentiates homogeneously is the income, since each cluster starts and ends one after the other when observing the maximum and minimum values. We can also see that the age value does not interfere much, as it is closer to the peak value of the age histogram (Section Data understanding, Age).

Clusters 0: has a higher density of men, with higher education than high school, single, living in medium and large cities. Cluster 0 has an average of $154,000 and is mostly employed.

Clusters 1: on the other hand, presents a mixture of men and women, with education equal to or less than high school and married. Cluster 1 has the majority living in a small town, unemployed, with an income of $79,000.

Clusters 2: have a higher density of men, with higher education than high school, single, living in medium and large cities. It has income of $225,000 and most are self-employed.

Clusters 3: presents a mixture of men and women, with education equal to or less than high school and married. Cluster 3 has a majority living in a medium-sized city, employed and with an income of $114,000.

Conclusions

In general, first approach is quite computationally convenient, and has a good segmentation that could be used in a broader marketing campaign.

However, for the use of market intelligence, the second approach seems more interesting. This is because one of the most common ways of performing market segmentation is how much they spend with the company, in this case, how much they have in annual revenue (Gustavo Gomes, “3 ways to classify customers to guide strategies”). This is because it is possible to make a correlation with the Pareto Principle, and thus create differentiated marketing campaigns for each level of audience, or even special campaigns for 'premium' audiences.
So in order each cluster seems to represent:

-Cluster 0 has a demographic of highly educated, single males who are likely to be working professionals.

-Cluster 1 has a demographic of low-educated, married individuals who are mostly unemployed.

-Cluster 2 has a demographic of highly educated, single males who are likely to be self-employed entrepreneurs.

-Cluster 3 has a demographic of low-educated, married individuals who are likely to be employed in medium-sized cities.

Recommendations

1-Consider the types and interests particular to each of the clusters

2-Consider clusters 0 and 2 as possible premium audiences, and depending on spending, offer special offers intrinsic to the particular taste of each of the groups.

3-Make more affordable offers for audiences 1 and 3, as well as offers on products related to couples, or more commonly consumed by couples.

4-Consider future data of average expenditure made in the market and the number of children. Thus, it is possible to send even more targeted offers to audiences with a higher propensity for acceptance.

5-Consider the more in-depth specifications for each data explained in the 'Quality Issues' section

References

Çankaya, M. F. (2022, March 22). *engineering.teknasyon.com*. Retrieved from How To Normalize Your Unsupervised Data For Clustering Methods: https://engineering.teknasyon.com/how-to-normalize-your-unsupervised-data-for-clustering-methods-9389298d20d5

Ergen, B. (2022, Nov 25). *Linkedin*. Retrieved from How to detect the strongest outliers with Local Outlier Factor ?: https://www.linkedin.com/pulse/how-determine-strongest-outliers-local-outlier-factor-b%C3%BCnyamin-ergen?trk=public_post

Gareth James, D. W. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer.

Gomez, G. (n.d.). *Blog do Agendor*. Retrieved from 3 formas de classificar clientes para direcionar estratégias: https://www.agendor.com.br/blog/formas-de-classificar-clientes/

Gupta, A. (2021, May 27). *medium.com*. Retrieved from X-Means — A Complement to the K-Means Clustering Algorithm: https://medium.com/geekculture/x-means-algorithm-a-complement-to-the-k-means-algorithm-b087ae88cf88

Jeffrey W. Lockhart, P. (2022). *Gender, Sex, and the Constraints of Machine Learning Methods*. Department of Sociology, University of Chicago.

Jiawei Jan, M. K. (2012). *Data mining concepts and techniques*. 225Wyman Street,Waltham, MA 02451, USA: Elsevier Inc.

Larson, B. N. (n.d.). *Proceedings of the First Workshop on Ethics in Natural Language Processing*. 686 Cherry St. MC 0165, Atlanta, GA 30363 USA.

Markus M. Breunig, H.-P. K. (2000). LOF: Identifying Density-Based Local Outliers.

Data

The data for this project “supermercado_noche.csv” was given by the teacher, if needed you should contact him.