Nova University of Lisbon

NOVA IMS Information Management School

Postgraduate Program in Enterprise Data Science & Analytics

Group Project – Unsupervised Learning
Sportify

Data Science and Machine Learning

Flávio Côrte 20231124

Luis Mourisco 20231119

Maria Delgado 20231229

Nuno Machado 20231118

Pedro Reis 20231419

Professors: Carina Albuquerque and Roberto Henriques

Spring Semester 2024

Index

Abstract[1]

As members of the company's data science team, we were challenged to develop an innovative approach to analyse Sportify's data. To achieve this, we worked with three distinct datasets, each corresponding to various aspects of the business (social interactions such as email, app and social media, consumption patterns segregated by different product categories, and demographic such as residence, age, and education).

To dive into this data science project, we followed the CRISP-DM methodology, since we were delivered research questions collected from our business stakeholders. The CRISP-DM methodology helped guide the project through its various stages (including while going back and forward) and ensures that main aspects are adequately addressed.

We made several attempts while at the clustering phase, as we became truly conscious that the way the data was presented to the model would deeply influence the outcome. Therefore, we went back and forward multiple times, in order to present a robust analysis that we could be comfortable with.

Each effort would begin with refreshing statistical information about each variable, such as distribution, as well as correlation and variance. Afterwards, we adopted a trial-and-error method, meant to test the output that would be generated with several types of data treatment. In case we group three highly correlated variables into one having their mean and input this new parameter, would K-Means give us back better-defined clusters?

Succeeding the several tests performed by our team, we decided to choose the approach with better results. We concluded that the optimal number of clusters is three on a digital behavior perspective, plus three on a sales perspective, which makes nine combined clusters. As we had the demographic data of customers available for analysis, we decided to add more detail to the labels of each cluster by putting demographic data into use.

[1] An AI generative tool was merely used to help translate and review the text written in this document.

The nine clusters are, as described in greater depth in chapter 4, the combination of the 3 clusters in Products (Low Spenders, Team Gamers, and Sport Lovers) and in Digital Contact (Ad Clickers, Social Butterflies and App Explorers).

Lastly, in Chapter 5 we presented some creative action plans for each cluster, trying to capture more business for the company (by suggesting actions like targeted campaigns, advertisements, sponsorship, and rewards, amongst others).

1. Data Exploration

To meet the challenge posed by our company, we began by exploring data in each one of the three different datasets that were made available to our team: 1) digital contact, 2) products, 3) demographics. We also gathered with four key partners: the marketing department, data science department, data quality assurance department and business stakeholders. After conducting interviews to have a better understanding of the research questions and data-related topics, we managed to have documentation about the data we were about to work with, as well as observations from the data quality team. Furthermore, business stakeholders gave us two research questions: one for each dataset (digital contact and products), and these were always on the back of our minds while working on this project.

1.1. Digital Contact

Starting by digital contact, this dataset captures digital interactions from Sportify's customers, namely on e-mail, app, and social media. The initial dataset had seven variables, including the customer ID. We checked for duplicates and there were none, therefore we decided to convert the customer ID into the data set index, and thus reducing the number of variables to six.

We discovered that the variable "SM_Shares" (social media shares) had missing values (39) and that it had a non-suitable data type ("float64"). Although missing values represented less than 1% of the total, we agreed that data should be complete, so we set out to find the best solution for this dataset. We first looked at mean (8.4) and then at mode (both 0.0 and 2.0) and given the inferior mode (comparing to the mean), we decided to transform null values into zero. We then converted the variable "SM_Shares" to data type "int64".

Continuing on a deeper statistical analysis, we observed no extreme values (commonly known as outliers), even though the data did not have a Gaussian distribution. We detected skewness above 1 on two variables (e-mail clicks and app clicks), where app clicks won biggest right tail (1.8). As for kurtosis, no values equal to

3 were encounter, and 4 out of 6 were negative (called platykurtic distribution). Therefore, considering skewness and kurtosis, we can say that there is less of a concentration of values around the mean (compared to a normal distribution), which means that extreme values are less likely to occur (less variability in the dataset).

Right at this point, it is possible to tell that the data was high dispersion, which prevents the formation of globular clusters; and with no globular clusters, we have a strong sign that the K-Means may not be the most suitable algorithm for this dataset.

Moving to the correlation between variables, we used Spearman correlation, which gave us a positive and strong correlation (above 0.75) between variables, specifically between e-mail clicks and social media clicks, and also between social media comments and social media shares and social media likes.

For details on the numerical description of variables, please refer to Table 1 on Appendices.

1.2. Products

Similar to what we did with the digital contact dataset, we started with a simple analysis of the products dataset. We found that the customer ID was unique for all 4.000 entries on the data set and decided it would be our index (plus removing a variable that does not contribute to characterize clusters).

We detected several duplicates on the data set. These duplicates were equal in all variables except customer ID: meaning the same amount of money spent on each category of products and same last purchase date, but different customer ID. We discussed whether or not to eliminate the total of duplicates found (510); at the end, we agreed that dropping 13% of our total number of entries having distinct customer IDs would not be reasonable. Besides, the data quality assurance department mentioned that data is reliable since it comes directly from the core system. One of the possible explanations for the duplications could be a promotional campaign, for instance. Since we turned customer ID into an index, we were left with six variables. Despite the variables appearing to be normally distributed (based on having a similar mean and median), skewness values reveal that the Hiking and Running category has a high positive skewness of 4.7, indicating a right-tail elongation in the distribution of data points. Not far behind, the total products variable displays a positive skewness of 1.8, indicating a tendency towards higher values.

About kurtosis, we had four variables with higher than 3 values: Hiking and Running (37.8), Team Games (10.6), Total Products (9.2) and Total Spent (5.5). High values on kurtosis may point out outliers, so we set out to do data visualization looking for a comprehensive understanding of our data points in the above-mentioned variables.

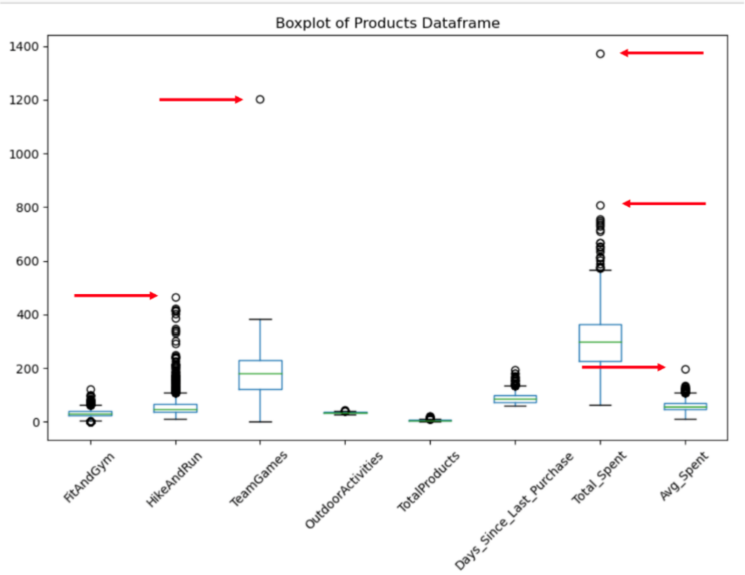


Figure 1 – Example of outliers identified on the products dataset

As we detected the presence of outliers, we were conscious that the K-Means algorithm would not perform well unless we handled the outliers, as it is sensitive to extreme values.

1.3. Demographic

The demographic dataset was meant to complement the characterization of each cluster (that will be determined by digital contact and products bought), therefore playing an important role on the labelling of clusters. Given that the data quality assurance department raised some red flags about this dataset (which made us somehow suspicious), we decided to take a closer look at each one of the variables:

- ‘City’: We detected a typo ("Brimingham" instead of "Birmingham") and 1.981 null values. Our sales team should make an effort to update this information next time a customer purchases goods in our stores.
- ‘Dependents’: A binary variable that should only have values of 0 or 1. We found 79 observations with values of 2. Maybe some large families?
- ‘Education Level’: Some confusion between uppercase/lowercase ("PHD" and "Phd", "high School" and "High School"), since Python is case sensitive.
- ‘Name’: All names have a prefix "Mr" or "Miss" (which we can use to determine customer’s gender).

Afterwards, we looked what the raw data and took some preliminary insights:

- The age range is between 18 and 80 years old (please see table [x] for visual information).
- About 67% of customers are aged between 18 and 35, although the average age in males is higher than females (38 vs 32, respectively).
- Around 9% are “youths” with more than 65 years old.
- About 59% of customers are women.
- 700 customers are based in London, 1.320 in Birmingham, and 1.981 on unknown locations (can it be a group of secret agents?). All 3 ‘locations’ have an approximate average age of 34 to 35.
- About 54% of customers have a high school diploma or less, 26% have a bachelor's degree and 21% a master's or a PhD. A curious fact is that all 5 levels of education have a similar average age, although we were expecting that at least, Master and PhD had a higher average age.
- People with less education are more likely to have dependents than people with higher education (please see figure [3] and [4]).
- Fun fact: We have four customers named Michael Jackson.

2. Preprocessing

As described in the previous chapter, each data set presents unique constraints, requiring different preprocessing procedures before applying the clustering algorithms. Typically, this is one of the most complex and time-consuming aspect on data science and we felt it firsthand.

Next, we will explain the work carried out by our data analysis team.

2.1. Digital Contact

In order to fill in the missing values encountered in “SM_Likes”, we evaluated several options, such as using mean or median; at the end, we decided to use the K-Nearest Neighbors (also known as KNN) algorithm, given that data is not continuously distributed. Please refer to Appendix 1 for in-depth explanation about KNN).

We tested the application of KNN to: 1) all “neighbors”, and 2) only to social media variables. No significative differences were found, making us decide to apply the algorithm to all neighbors. Using the KNN algorithm, we were able to fill in 39 missing values, making our dataset more complete, with six variables having the same number of entries: 4.000 non-null values.

2.2. Products

We noticed that two variable names were inconsistent: “Fitness&Gym” and “Hiking&Running”. We replaced “&” with “_” using snake case convention to ensure readability and ease of use in coding.

As outliers were detected on the data exploration analysis phase, we tested them using several statistic techniques and decided to remove all values with a z-score of four or higher, a total of 64 data points.

We also decided to do some feature engineering by creating three different columns: 1) total_spent (which represents the sum of 'fitness_gym' with 'hiking_running', 'teamgames', and 'outdooractivities'), 2) Avg_Spent which represents the average amount spent per product and 3) Days_Since_Last_Purchase which represents the number of days since the last purchase of the customer. Afterwards we dropped the ‘last_purchase’ feature as it is not used anymore.

After creating the new features, we ran a correlation matrix to check if any of the features were highly correlated which would impact the clustering algorithm. The ‘total spent’ feature has a 0.93 correlation value with ‘Teamgames’ feature and therefore we decided to drop that feature as well.

To be certain of our approach, we decided to generate a boxplot and histograms of our features to visualize their individual distributions.

We verified that the skewness of ‘Hiking_running’ and ‘Total_products’ features were 4.67 and 1.79 respectively, both high values and so we decided to apply a transformation on the data to normalize it. We tested several types of transformations and opted for a log transformation because it performed better, lowering the skewness to near zero.

After the initial cleaning of the data, we proceeded to scale our dataset using ‘Standard Scaler’ so that the data clustering results in better separated clusters. We opted for a standard scaler since we previously removed the outliers in the dataset and verified that the distribution of the features was close to normal.

2.3. Demographic

To address the previously mentioned issues with this dataset, the following work was carried out:

1) City:

- Fixing Typos: A "replace" operation was performed to correct the typo from "Brimingham" to the correct name, "Birmingham".
- Blank Values: For blank values, a new classification "Other location" was created, allowing the data and marketing departments to work on obtaining the real locations of the customers. The decision not to apply KNN (K-Nearest Neighbors) was made because this dataset was not intended for clustering, but rather to enrich the information in other datasets.

2) Dependents:

- Incorrect Values: Although there were 79 cases of "large families" (with a value of 2), a "replace" was performed to change these to 1, as this is a binary column.

3) Name:

- Gender Identification: The "Mr" and "Miss" prefixes were used to create a new column indicating the gender of the customer (male for "Mr" and female for "Miss").

4) Education Level:

- Standardizing Uppercase/Lowercase: The confusion between uppercase and lowercase was corrected to maintain consistency, such as "PHD" and "PhD", "high School" and "High School".

5) Age:

- Creating an Age Column: A new age column was created, representing the difference between 2024 and the year of birth, to make the final analysis more intuitive.

These corrections were crucial for improving the quality of the dataset and enabling more accurate and effective analysis. With these adjustments, marketing and other departments can use the dataset more reliably without worrying about inconsistencies or incorrect values.

3. Modelling

3.1. Perspective: Digital Contacts

Since we encountered highly correlated variables on the dataset, we were curious to know if any data treatment (different inputs) would return different outputs. So, we created three options (option zero, option one and option two). For each option, we applied the Elbow Method and the Silhouette Score to determine the optimal number of clusters.

For all options, we used standard scaler as a preprocessing technique to standardize features, ensuring that features have the same scale, thus preventing certain features from dominating due to their larger magnitudes.

On option zero we kept the dataset as it was, with no deleted columns. On option one, two of the highly correlated social media variables (shares, likes and comments) were deleted and one variable of those variables was kept (that would represent the other two). On option two, the three variables were gathered into one new column (after normalization), keeping all available information with less noise, but more difficult to interpret deeper.

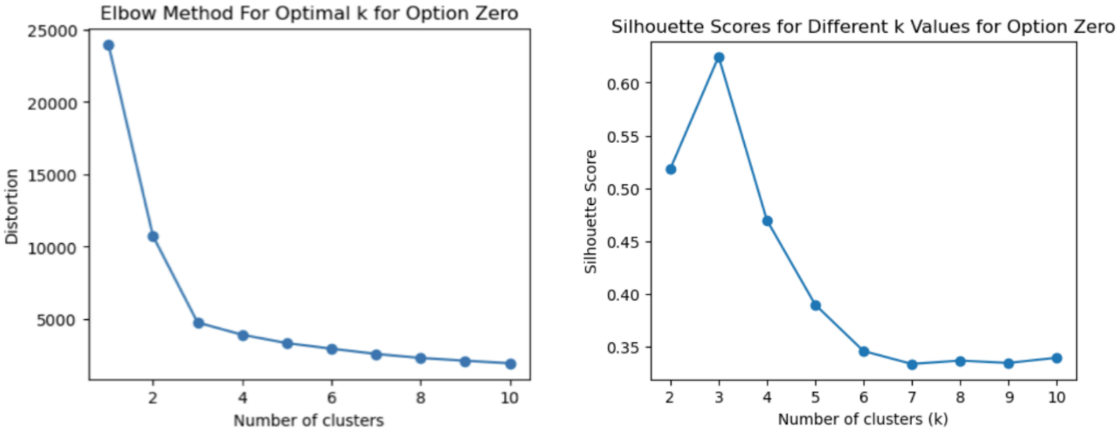


Figure 2 – Digital: Elbow Method and Silhouette Score for Option Zero

In option zero, we concluded that the number of clusters might be three, according to the Elbow Method (corresponding to the first point where the curve starts to stabilize) and supported by a higher Silhouette Score.

On option one, we eliminated two out of three variables (SM_Comments, SM_Likes, SM_Shares) identified to be highly correlated (correlation coefficients above 0.7). The selection of the two variables to be discarded was made by analyzing certain relevant characteristics. From the data exploration, we realized that SM_Shares had missing values, therefore we considered it a better variable to be excluded. While comparing SM_Comments and SM_Likes, we observed that SM_Comments shows a higher correlation with the discarded SM_Shares variable, which led us to decide to keep SM_Comments as representative of social media interactions.

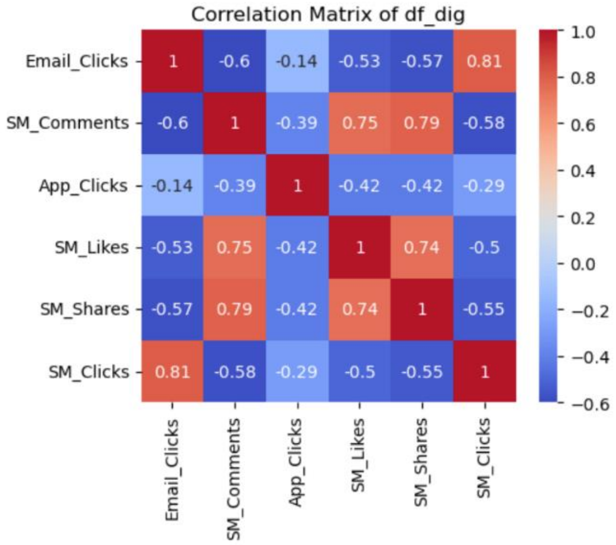


Figure 3 – Heatmap with Digital_Contacts dataset variables

By applying the Elbow Method and the Silhouette Score to the transformed data, the conclusion was identical to that of option zero: the optimal number of clusters is three, confirmed by a higher score on Silhouette.

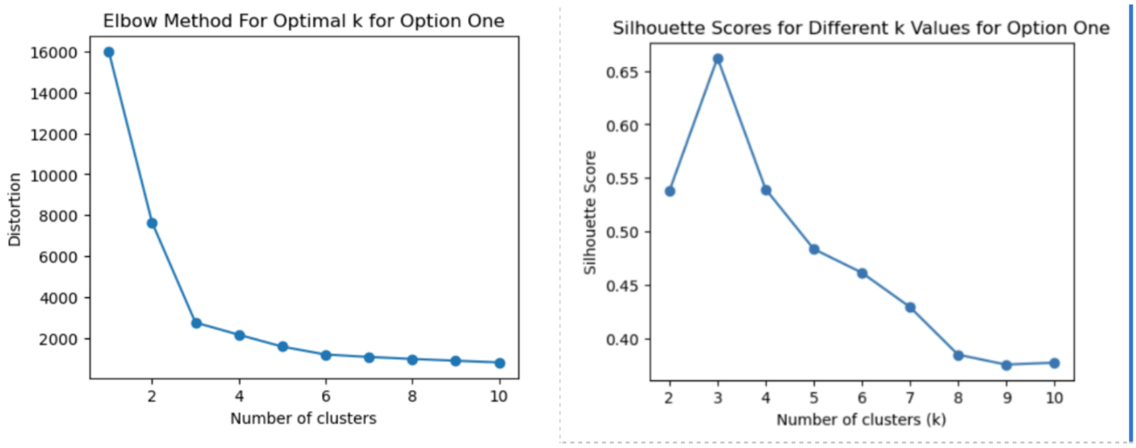


Figure 4 – Digital: Elbow Method and Silhouette Score for Option One

For option two, we tried merging three social media columns into a single one (representative of the social media interactions). Once again, we reached the conclusion that three is the optimal number of clusters.

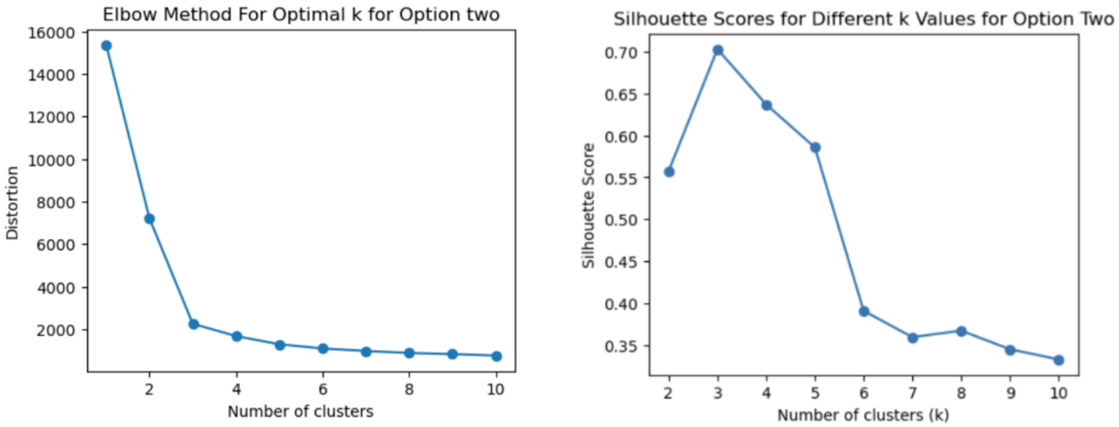


Figure 5 – Digital: Elbow Method and Silhouette Score for Option Two

We concluded that three was the optimal number of clusters, from a digital contact perspective.

Moving to K-Means, we used K-Means++ to get a better representation of the initial centroids.

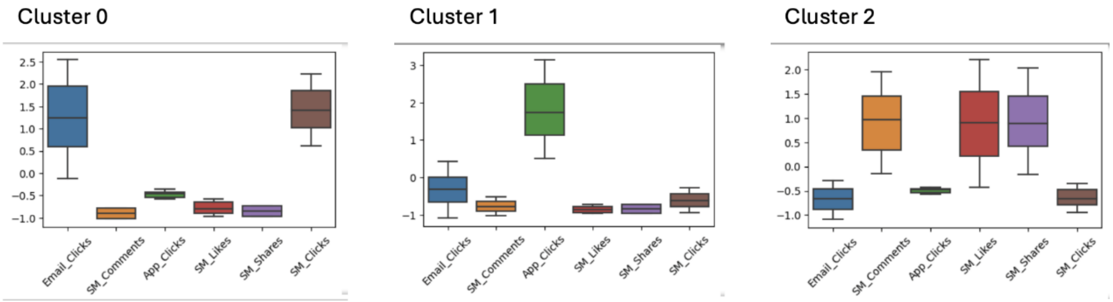


Figure 6 – Clusters' patterns considering each feature of the digital contact dataset

We identified clusters as:

- Cluster 0 – Ad Clickers

- Cluster 1 – App Explorers
- Cluster 2 – Social Butterflies

We will further characterize each cluster on the next chapter (Description of Resulting Clusters).

3.2 Perspective: Products

The first step is to identify which clustering algorithm to use. As we previously dealt with the presence of outliers on the dataset, we were comfortable to use K-Means.

Step two was to identify the optimal number of clusters. In order to do our clustering analysis, we used both the Elbow Method and the Silhouette Score. After generating the plotted graphs, it was a challenge selecting the appropriate number of clusters, since the Elbow Method was showing an elongated curve, making it unclear on what was the optimal number of clusters was. To make a more informed decision, we computed the Silhouette Scores, and the highest score was three clusters, followed by four, and then two. Therefore, we decided to consider k=3 as the optimal number of clusters, although the relative low Silhouette score clearly raised some concerns amongst the team. In this sense, some steps should be taken to improve the number of clusters in a future iteration, such as trying different clustering methods, like DBSCAN or the hierarchical method.

We applied the K-Means algorithm on the data frame with k=3 and 300 iterations.

From the output of the algorithm, we took the cluster labels and assigned them to our initial data frame so we can identify which customers belong to which of our three clusters.

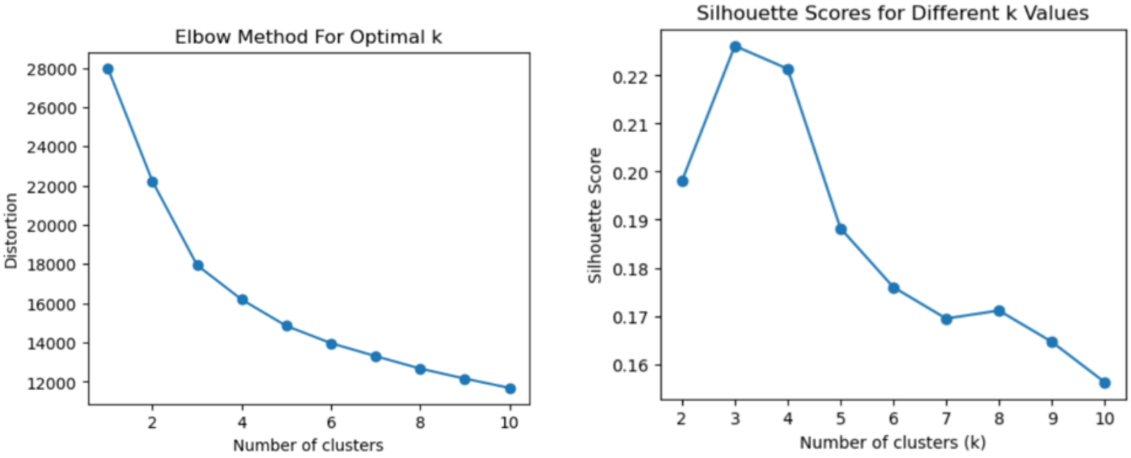


Figure 7 – Products: Elbow Method and Silhouette Score

4. Description of Resulting Clusters

4.1. Digital Contacts

Bearing in mind the research question: "What distinct customer segments exist based on their engagement behaviour across email, mobile app and social media platforms?", we identified the following clusters:

- Cluster 0 – Social Butterflies:

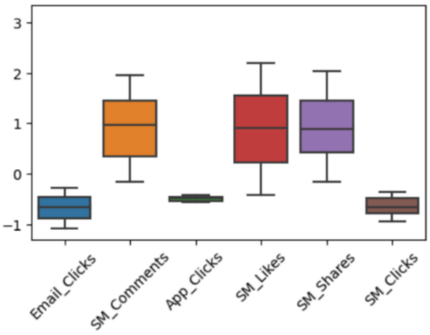


Figure 8 – Cluster 0 visualization of behaviour on each feature of the dataset

Customers present a higher-than-average tendency to interact with features of social media, such as comments, likes and shares, while all other variables sit below average.

- Cluster 1 – Ad Clickers:

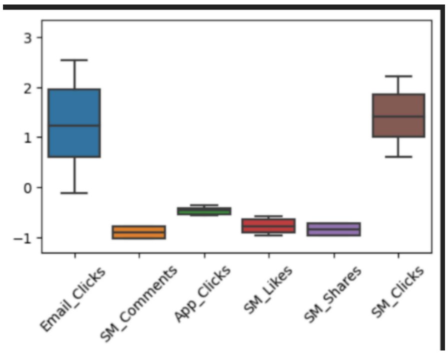


Figure 9 – Cluster 1 visualization of behaviour on each feature of the dataset

Ad Clickers present a higher-than-average tendency to click on ads (coming both from email and social media). Other variables below average.

- Cluster 2 – App Explorers:

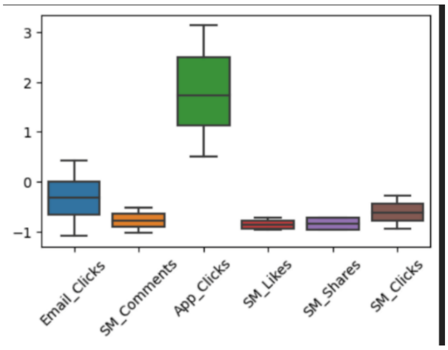


Figure 10 – Cluster 2 visualization of behaviour on each feature of the dataset

The App Explorers have low interaction with social media and do not usually click on our ads. Their preferred way of interacting with Sportify is through the mobile app.

Although the clearly segregated behaviours we observed from clustering our clients, we wanted to observe how well separated the clusters really were. One of the best ways to do so (besides, for example, the silhouette score mentioned above) is by reducing dimensionality and plot the clusters in 2D. Instead of plotting combinations of features, we can use PCA. In this project, PCA was applied to data that was free of outliers and transformed by StandardScaler. The relative weight of each component, up to the number of features, in the explained variance was analyzed, and finally a 2D plot was built with the two principal components.

By applying PCA to our dataset we captured 86.2% of the variance in the data on the two first principal components. The visualization shows a clear separation between clusters that although being what could be described as round, do have a globular shape, appropriate to a k-means clustering.

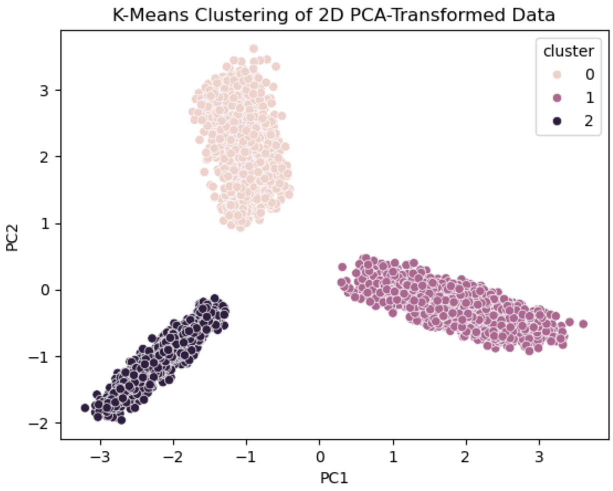


Figure 11 – K-Means clustering of 2D PCA-transformed data (digital contact dataset)

4.2. Products

To characterize clusters originated from the products' perspective, we were guided by the research question: "Can we identify unique customer segments by analysing how they purchase across various product categories and their buying patterns?".

We started by giving labels to each of the three clusters. The distribution of customers between clusters was considered to be homogeneous and significative (each cluster with more than 1.000 customers).

Cluster ID	# of customers	Percentage
0	1114	28,30%
1	1230	31,3%
2	1592	40,4%

Table 1 – Distribution of customers for each cluster

- Cluster 0 – Sport Lovers: Although being the smallest of clusters (1.114 customers), these customers buy more frequently (every 76 days). Like Cluster 1, they spend the most money in team games (218), with a twist – they spend bigger on the rest; however, we observed that they spend more in hiking and running compared to other clusters (78).

- Cluster 1 – Team Gamers: These customers are the ones that on average spend more per product (77.0). However, they buy less frequently than other clusters (100 days between purchases). We can consider their favourite category to be team games (based on the most money spent on: 218). In terms of cluster size, we are referring to 1.230 customers.

- Cluster 2 – Low Spenders: This is the largest cluster in terms of number of customers (1.592). These customers can be characterized by being the ones spending less money on average per product (44.3), although they are more regular customers than Cluster 1 (88 days since last purchase).

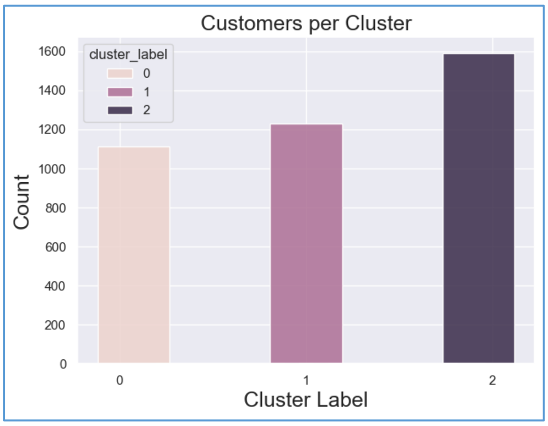


Figure 12 – Visual representation of the magnitude of customers per cluster

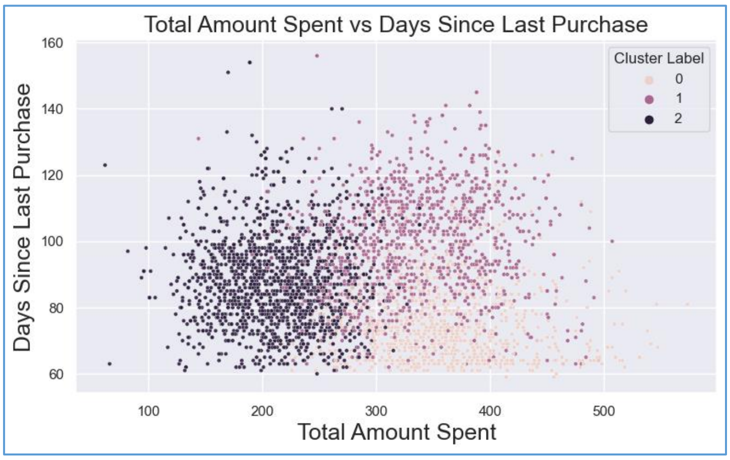


Figure 13 – Scatterplot with total amount spent versus recency for each of the three clusters

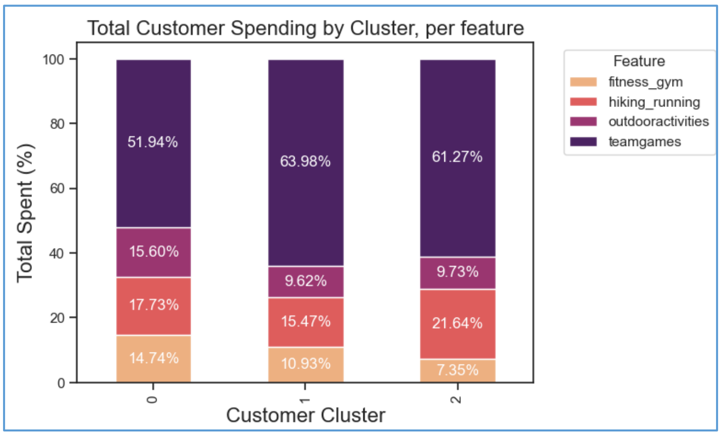


Figure 14 – Customer spending divided by cluster and detailed with categories (or features)

Similarly to the Digital Contact dataset. We performed a PCA analysis here. The first 2 principal components represent only 57% of the variability, indicating an overlap between clusters would be likely. Indeed, when observing the 2D plot there is an area where datapoints are mashed instead of well separated and globular as in Digital Contacts. This less than ideal cluster separation is reinforced by a relatively low silhouette score when compared to the Digital Contact dataset (0.23 vs 0.64). Thus, we consider that the k-means algorithm may have not separated the clusters ideally, and further investigation into other clustering alternatives like DBSCAN, and compare the results may be beneficial in the future.

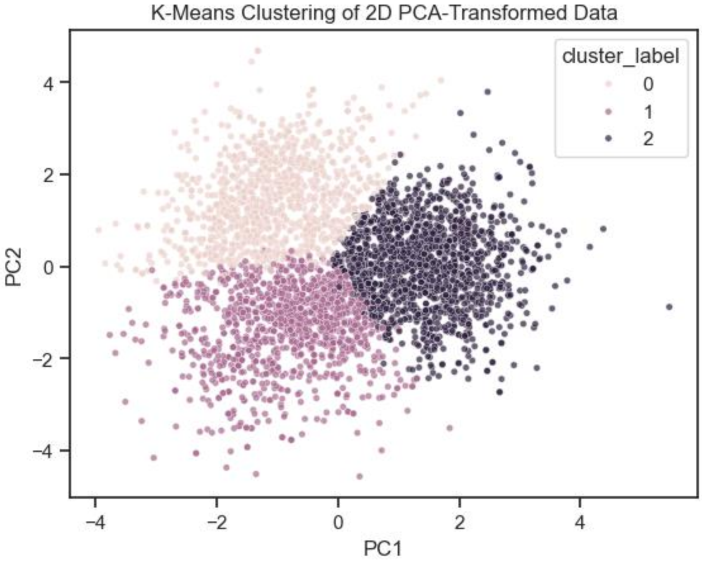


Figure 15 – K-Means clustering of 2D PCA-transformed data (products dataset)

4.3. Merged clusters plus demographics

After successfully reached three clusters from two perspectives (digital contact and products), we set out to merge the two processed datasets into one, concatenating the two cluster labels and resulting in nine distinct clusters.

As soon as we had the nine clusters, we started to look for valuable insights that could help labelling each one of these clusters. For better interpretation, we have segmented the following topics from a digital contact perspective (example: Cluster 0 refers to Cluster 0 from the digital contacts perspective and Cluster 00 refers to the previously mentioned cluster together with Cluster 0 from the products perspective).

4.3.1. Cluster 0 and its children (00, 01, 02)

- In Cluster 0, all variants (00, 01 and 02) are returning a large number of females, meaning that Social Butterflies are mostly women.
- Cluster 0 has an average age of 29 years old (“Early Millennials”).
- When looking at dependents, Cluster 0 has 47% of the total number of dependents. However, if we step back and look at the number of dependents on Cluster 0, we can observe that only 38% of the total of customers on this cluster has dependents.
- Cluster 0 has the biggest concentration of high-level education customers. In Cluster 02, 28% of customers have a master or PhD. Despite that, 53% of the customers have high school or less in Cluster 02.

4.3.2. Cluster 1 and its children (10,11,12)

- Cluster 1 was named App Explorers.
- For Cluster 1, the average age is 38 years old (“Mid Millennials”).
- App Explorers are relatively balanced between females and males (prevalence of males).

4.3.3. Cluster 2 and its children (20,21,22)

- We named Cluster 2 as Ad Clickers.
- In Cluster 2 we observe a predominant trend (applicable to all variants): most of the customers are male. We can assume that the persona that represents Ad Clickers is a male character.
- The average age in Cluster 0 is 42 years old (“Late Millennials”).

5. Action Plan

3, 2, 1... Action!

Now that we have decided on the clusters and studied their characteristics (together with demographic data), we present a suggestion of actions to effectively target each one of our nine clusters.

5.1. Low Spenders Social Butterflies

- These are female customers who better engage with organic content tailored to social media. We suggest the marketing department to work on their strategy to interact with these female customers on social media platforms, to potentially increase sales (since these are low spenders).
- We also suggest offering exclusive discounts for social media followers to increase purchases.

5.2. Low Spenders Ad Clickers

- Since these are low spenders, we can experiment different ad formats and placements to know what works for our customers, and then use the successful pieces to target to more premium customers (such as Sport Lovers Ad Clickers).

5.3. Low Spenders App Explorers

- Here we can offer exclusive in-app promotions to reward app usage and try to boost purchases.
- Take in consideration that these are mid Millennials, so we can adapt the language and imagery in order to resonate better with their life stage, aspirations, and challenges.

5.4. Team Gamers Social Butterflies

- We can consider sponsoring gaming events or tournaments to increase brand visibility and engage with the gaming community.
- Also create gamified social media campaigns or challenges to encourage female customers participation (and, perhaps, virality).

5.5. Team Gamers Ad Clickers

- Develop targeted ad campaigns highlighting gaming-related products.
- Partner with gaming platforms or content creators to promote our brand through sponsored content or ads.

5.6. Team Gamers App Explorers

- A suggestion would be to integrate gaming-related features or functionalities into the mobile app to better appeal to this segment of customers.
- Implement in-app events or challenges to foster community engagement and retention.

5.7. Sport Lovers Social Butterflies

- We suggest developing sports-themed content with focus on female audience.
- Also partner with sports influencers to co-create content or sponsorships.
- We can even use social media to offer exclusive sports-related experiences or merchandise to reward loyal customers.

5.8. Sport Lovers Ad Clickers

- Create targeted ad campaigns featuring sports-related products (with a tendency to prefer hiking and running) to capture the attention of this segment of male customers.
- Consider that this segment is made of late Millennials. We should have a CTA (call-to-action) on our ads that can redirect customers to a frictionless shopping experience, in order to meet the expectations of this target.

5.9. Sport Lovers App Explorers

- Offer exclusive in-app rewards or benefits for sports enthusiasts.
- Implement push notification highlighting upcoming sports events or promotions to keep customers informed and engaged.

6. Conclusion

As aspiring data scientists, we wanted to test several approaches and deliver the best possible result. At the initial phase, we had no trouble analysing the datasets and applying data processing techniques, including KNN. Still, we encountered challenges, such as identifying 500+ duplicates on the products dataset that had unique customer ID. We debated how strange it was to have such a tremendous number of entries that were duplicate on all extend, except for customer ID. Was it a bug from the sales program? Was it a human error from the operator? We did not have answers to these questions, so we decided to ignore this event and move on to modelling.

On the modelling phase, we first used the Elbow Method and the Silhouette Score to determine the optimal number of clusters. We debated most the edgy cases, such as having a linear-like “elbow” that would not help deciding on the optimal number of clusters (for the products dataset). We also discussed what should our approach be to the digital contact dataset, since we now knew that we had three highly correlated features. Were these features impacting on the optimal number of clusters? We were not sure, so we did what a data scientist would do: test hypotheses. After having agreed on the optimal number of clusters, we used K-Means to have data points attributed to their cluster. We proceeded in this way for both datasets: digital contacts and products.

With three clusters each for both digital contacts and products, we set out to characterize each of the clusters, followed by adding the demographic data available to the concatenated clusters (resulting in a total number of nine clusters).

Finally, as we had analysed and labelled the nine resulting clusters, we then reflected on suggestions to present to the readers of this report (mainly directed to the marketing department).

Appendices

1. *KNN Imputer*

According to IBM, "…the k-nearest neighbors (KNN) algorithm is a non-parametric, supervised learning classifier, which uses proximity to make classifications or predictions about the grouping of an individual data point. It is one of the popular and simple classification and regression classifiers used in machine learning today."

It was developed in 1951 by Evelyn Fix and Joseph Hodges, and later expanded by Thomas Cover. Also, according to IBM, KNN is widely used in recommendation systems, pattern recognition, data mining, financial market predictions and intrusion detection.

KNN is a relatively simple and straightforward algorithm. To predict the value or class of a new observation, it calculates the distance between that observation and all others in the training dataset. In this chapter, the most used distance metric is Euclidean, but there are others that we will not explore in this academic work. After calculating the distances, KNN identifies the "k" closest neighbors and assigns the predominant class among these neighbors to the new observation. For regression problems, the prediction is based on the average of the nearest neighbor values.

However, it is also important to say that despite its simplicity, KNN also has some disadvantages, namely i) as the size of the data set grows, it becomes less efficient, compromising the model's performance, ii) the time execution time can be high , especially for large datasets, as each prediction requires distance calculations for all points in the training set, and iii) by "memorizing" the dataset, KNN requires more memory to work, making potentially slower than other supervised algorithms.

Various information consulted (such as StackExchange and scikit-learn) tell us that it is a good practice to apply the KNN Imputer after some data normalization, to prevent variables with larger scales from having more weight in the predictions. Moreover, IBM also suggests that the correct choice of the "k" value and the distance metric are crucial to the success of KNN.

Lastly, KNN is considered a "lazy learner" - it does not build a discriminative model, rather "memorizes" the training data set. This means that KNN does not have training time like other supervised algorithms, such as decision trees or neural networks, as discussed in theoretical classes.

2. *Silhouette Method*

We used the Silhouette score to validate the number of clusters suggested by the elbow method and subsequently considered in KMeans. The Silhouette score (often referred to as the silhouette coefficient in some literature) attempts to measure whether each data group fits into its cluster by comparing it with the nearest neighbouring clusters and evaluating the separation between clusters and the internal cohesion of each cluster. This allows for the assessment of the quality of our clustering.

From our research on the internet, there were two concepts important for calculating the silhouette coefficient that deserve clarification:

- ***Intercluster***: Refers to the differences between clusters and measures how separated groups are from each other. In this sense, the higher the value of "b" (intercluster distance), the further the clusters are from each other, indicating a good separation.
- ***Intracluster***: Refers to the similarity within a cluster and measures how close or related data points are within a specific group. A low value of "a" (intracluster distance) suggests a high internal cohesion of the cluster.

The silhouette coefficient calculation is done for each data point using the following mathematical formula:

$$s = \frac{b - a}{\max(b - a)}$$

where "a" is the intracluster coefficient and "b" is the intercluster coefficient. If "a" is smaller than "b", the silhouette will be positive, indicating that the point is closer to the cluster it belongs to than any other cluster.

The resulting coefficient values vary between -1 and 1 and have the following meaning:

- A value close to 1 indicates that the clusters are dense and well-separated.
- A value close to 0 suggests that the clusters are overlapping or indistinct, or may also indicate the presence of outliers.
- Negative values indicate that clustering may be incorrect, with misallocated points or overlapping clusters, or may also indicate the presence of outliers.

In conclusion regarding the previously presented information, the silhouette score is useful for choosing the optimal number of clusters in algorithms such as k-means, which is the case in our academic work. Therefore, we used the Silhouette score to have greater confidence that the number of clusters we considered for our datasets was correct.

3. PCA

Principal component analysis (PCA) was originally invented in 1901 by an Englishman called Karl Pearson (funny fact: he was the founder of the Department of Applied Statistics at University College London in 1911). PCA is a statistical technique used to reduce the dimensionality of a dataset, keeping as much relevant information as possible. In practice, PCA transforms a large dataset (with many variables) into a new dataset with fewer variables. The last are the linear combinations of the original variables and called principal components ('PCs').

Thus, two of its main purposes are:

1) **Reduction of dimensionality**: PCA is used to reduce the number of variables in a dataset, easing the analysis and visualization, plus simplifying statistical models. It was essentially for this purpose that our group used this technique in this work.

2) **Performance enhancer**: By reducing the number of variables, PCA can improve the performance of machine learning algorithms, especially for large data sets. Making a bridge with our work, would not be the reason for using it, since we have relatively few variables.

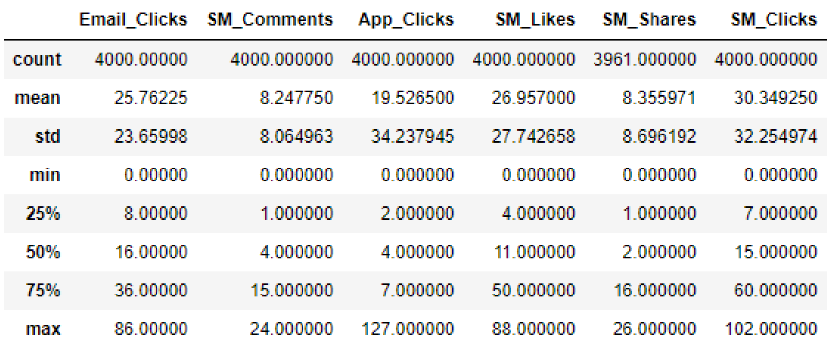
Additionally, before applying PCA, the data must be previously normalized, to ensure that all variables have the same influence. Thereafter, PCA calculates the covariance matrix to understand how the variables correlate with each other. Subsequently, the eigenvalues and eigenvectors of this matrix are found, which represent the direction and magnitude of variation in the data.

Eigenvalues indicate the variance each principal component captures from the original data set.

Eigenvectors represent the directions of these variations. Normally, the main components are chosen in order of explained variance, keeping those with the highest variance. Then, the data can be transformed to this new space defined by the selected CPs.

At the interpretation level, the proportion of variance explained by each principal component helps determine how many components must be maintained to retain a sizeable portion of the information. Another factor is the loading of each main component, which shows how each original variable contributes to each PC.

Finally, the literature used during this paper stated that PCA is widely used in data visualization, preprocessing for machine learning algorithms and data compression, among others. It allows you to visualize complex data in 2D or 3D spaces, reducing the complexity of the dataset while maintaining the most relevant information.



	Email_Clicks	SM_Comments	App_Clicks	SM_Likes	SM_Shares	SM_Clicks
count	4000.00000	4000.000000	4000.000000	4000.000000	3961.000000	4000.000000
mean	25.76225	8.247750	19.526500	26.957000	8.355971	30.349250
std	23.65998	8.064963	34.237945	27.742658	8.696192	32.254974
min	0.00000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	8.00000	1.000000	2.000000	4.000000	1.000000	7.000000
50%	16.00000	4.000000	4.000000	11.000000	2.000000	15.000000
75%	36.00000	15.000000	7.000000	50.000000	16.000000	60.000000
max	86.00000	24.000000	127.000000	88.000000	26.000000	102.000000

Figure 16 – Dataset “Digital_Contact” statistical description

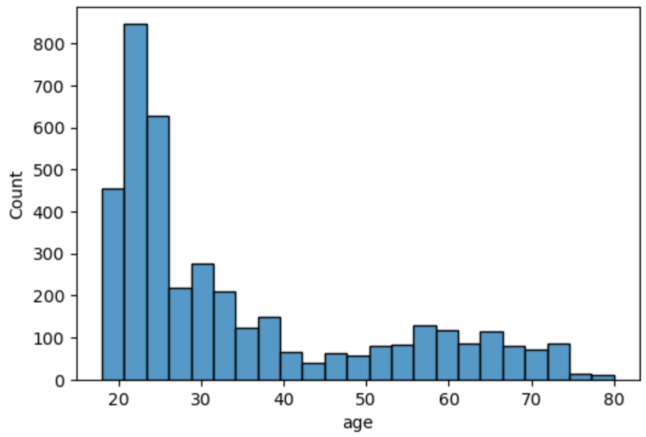


Figure 17 – Age distribution

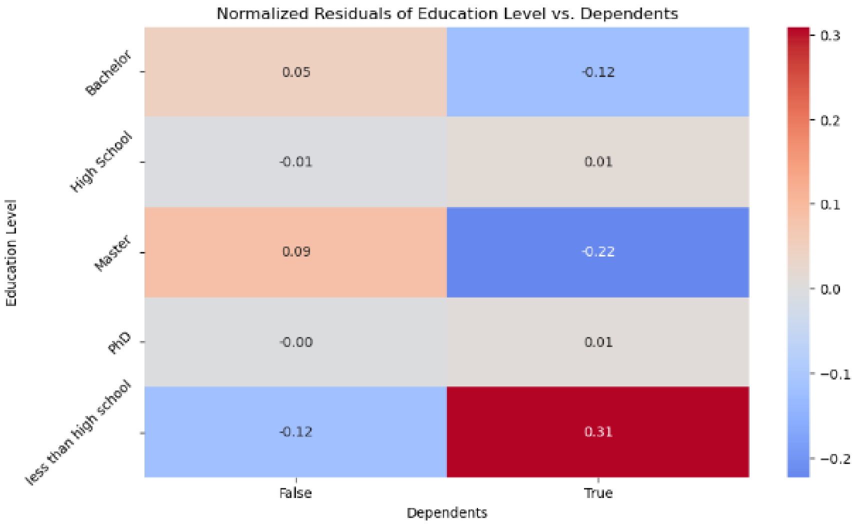


Figure 18 - Heatmap of normalized residuals from the contingency table of education level by dependent status

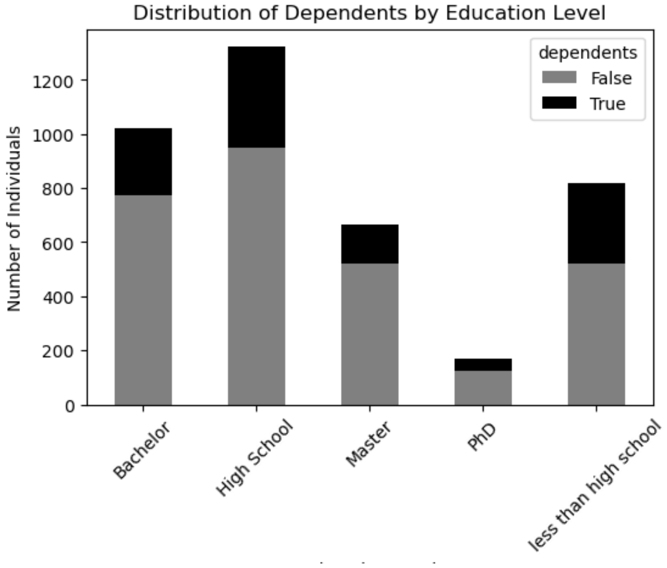


Figure 19 - Distribution of dependents by education level