Data Science and Machine Learning

Sportify Customer Segmentation Project

Group 16:

Bruna Luisa do Amaral da Silva – 20231102

Miguel António Oliveira Figueiredo – 20231436

Paulo Manuel Gonçalves Oliveira Valente da Cruz – 20231980

Pedro Miguel Albuquerque Gomes – 20231126

Simona Costache - 20231983

April 2024

Abstract

This study analyzed customer data from Sportify, focusing on digital interactions, consumption patterns, and demographics to identify distinct customer clusters for targeted marketing strategies. Merging various datasets allowed for comprehensive demographic and behavioral analyses. Techniques such as data preprocessing, transformation, and clustering were used to explore and segment customer bases effectively. Key techniques included handling missing data, outlier detection, and normalization, alongside principal component analysis (PCA) and K-means clustering for scalability and efficient segmentation. The results identified three primary customer clusters based on digital engagement and purchasing behavior, critical for developing tailored marketing strategies. Concluding insights emphasize the importance of segment-specific approaches in enhancing customer engagement and product marketing within Sportify's diverse consumer base.

1. Introduction

“Sportify” is a popular destination for sports enthusiasts, offering a diverse range of high-quality sports products. With a vast database at our disposal, we have started by conducting an analysis of customer data, examining digital interactions, consumption patterns, and demographic insights.

The demographic dataset provides insights into age, city, dependents and education level.

The products database shows preferences, behaviours and spending habits across different product categories.

Finally, the digital interactions dataset shows the customer engagement across online platforms, including email, social media, and mobile apps.

As a progressive company with growing expertise in data analytics and machine learning, we seized an opportunity for business development. Our mission was to identify actionable customer segments to enhance marketing effectiveness and drive Sportify's overall success. Using unsupervised machine learning methods and clustering techniques, our group designed a comprehensive marketing plan targeting segments with the greatest growth potential. By aligning our strategies with the unique traits and preferences of each segment, we aimed to elevate the customer experience and generate positive outcomes for Sportify.

2. Data Exploration

2.1 Basic Exploration

Before starting the data exploration, we decided to merge the three datasets into a single dataframe using the Customer ID as the index to ensure that data manipulation and retrieval operations are correctly aligned with each individual customer.

The first step was to do a Basic Exploration of our dataframe. It consists of 4000 rows and 17 columns. 1 is a datetime variable, 1 is float, 12 are integers and 3 are objects. Additionally, we could infer through the info() method that all columns are nearly complete with no missing values except for City - **1981**, corresponding to **49,5%** - and SM_Shares – 39, corresponding to 0,975%.

2.2 Statistical Exploration

2.2.1 Numerical Variables

When exploring the numerical features of the dataset, it's important to address missing values before embarking on any modelling to ensure the accuracy of the results. Here are some key observations:

'birth_year' indicates that the average customer age is approximately 35, with a mean of around 1989.

'dependents' feature reveals that the maximum number of dependents is 2. However, this is most likely an error, as the maximum expected value is 1 according to the data description.

'Fitness&Gym', 'Hiking&Running', 'TeamGames', and 'OutdoorActivities', the total spending varies. Notably, 'TeamGames' exhibits a higher mean and median compared to the other categories, suggesting a potentially greater interest or investment from customers in this area.

On average, customers spend approximately 176 monetary units on Team Games products, compared to an average of 111 monetary units on all other products, highlighting the significant revenue generated by Team Games.

It's noteworthy that some customers are not digitally interacting with the company, and there are also customers who are not purchasing Team Games products, despite their high revenue generation.

Addressing these observations will be essential for refining the dataset and preparing it for further analysis and modelling.

2.2.2 Categorical Variables

We analysed categorical features within our dataframe and we noticed the following:

- ‘education_level’: it's observed that around one-third of customers have high school level, suggesting a substantial portion of the customer base may share a similar educational profile.
- ‘name’ contains 3892 unique names out of a total of 4000, indicating there are 106 duplicated names. The most frequently occurring name is "Mr Michael Jackson," which appears 4 times. With further analysis we concluded that 90 names were repeated 2 times, 5 repeated 3 times and 2 for times. Also, comparing the birth year, the city and the education level, we concluded that, although the names were the same, it was referring to different persons. In this way, all the names were kept.

Regarding the 'City' variable, there are only 3 unique values, with "Birmingham" being the most frequent. Additionally, nearly half of the customers lack city information, potentially impacting location-based analysis and targeting strategies.

2.2.3 Datetime Variables

Date Feature Description:

- Purchases range from October 15, 2023, to February 29, 2024, covering approximately five months.
- 25% of purchases occurred before January 20, 2024.
- Half of the purchases were made on or before February 1, 2024.
- 75% of purchases occurred before February 13, 2024.

2.2.4 Kurtosis and Skewness Analysis

2.2.4.1 Kurtosis

Hiking&Running with 37.84**, TeamGames** with 10.60 and **TotalProducts** with 9.24 display high kurtosis values which suggests a distribution with very heavy tails and a sharply peaked center compared to a normal distribution. This indicates that there are potentially significant outliers or extreme values present in the data.

Understanding and addressing these patterns are crucial for accurate analysis and modelling, as they can influence decision-making processes and strategic initiatives within the dataset.

2.2.4.2 Skewness

Dependents with 1.35**, Email_Clicks** with 1.10**, App_Clicks** with 1.82**, SM_Likes** with 0.78, **SM_Shares** with 0.63**, SM_Clicks** with 0.95**, Hiking&Running** with 4.67**, TeamGames** with 0.74 and **TotalProducts** with 1.79, display a high positive skewness which indicates a concentration of lower values with longer tails towards higher values, suggesting predominantly right-skewed distributions and potentially revealing distinct customer behaviours and spending patterns within the dataset.

The negative skewness for **Birth_Year** with -1.05 indicates a left-skewed distribution, suggesting a longer left tail and a concentration of customer birth years towards the later years.

2.3 Visual Exploration

2.3.1 Distribution Analysis

First, with Distribution Analysis we visualized how data was distributed across various categories and numeric ranges as displayed in the following images:

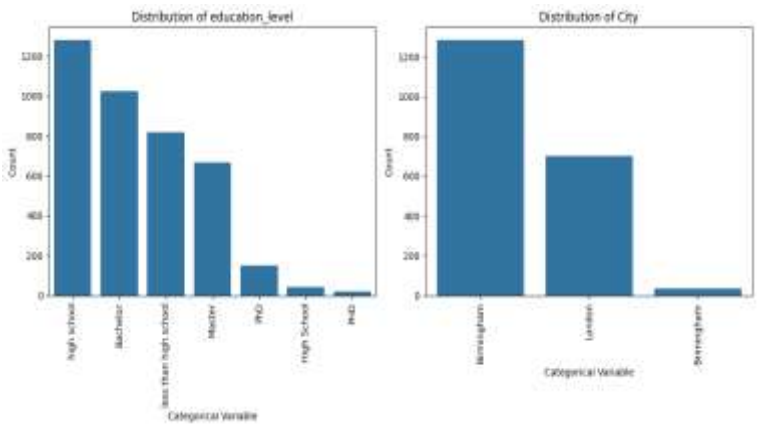


Figure 1 Categorical features distribution

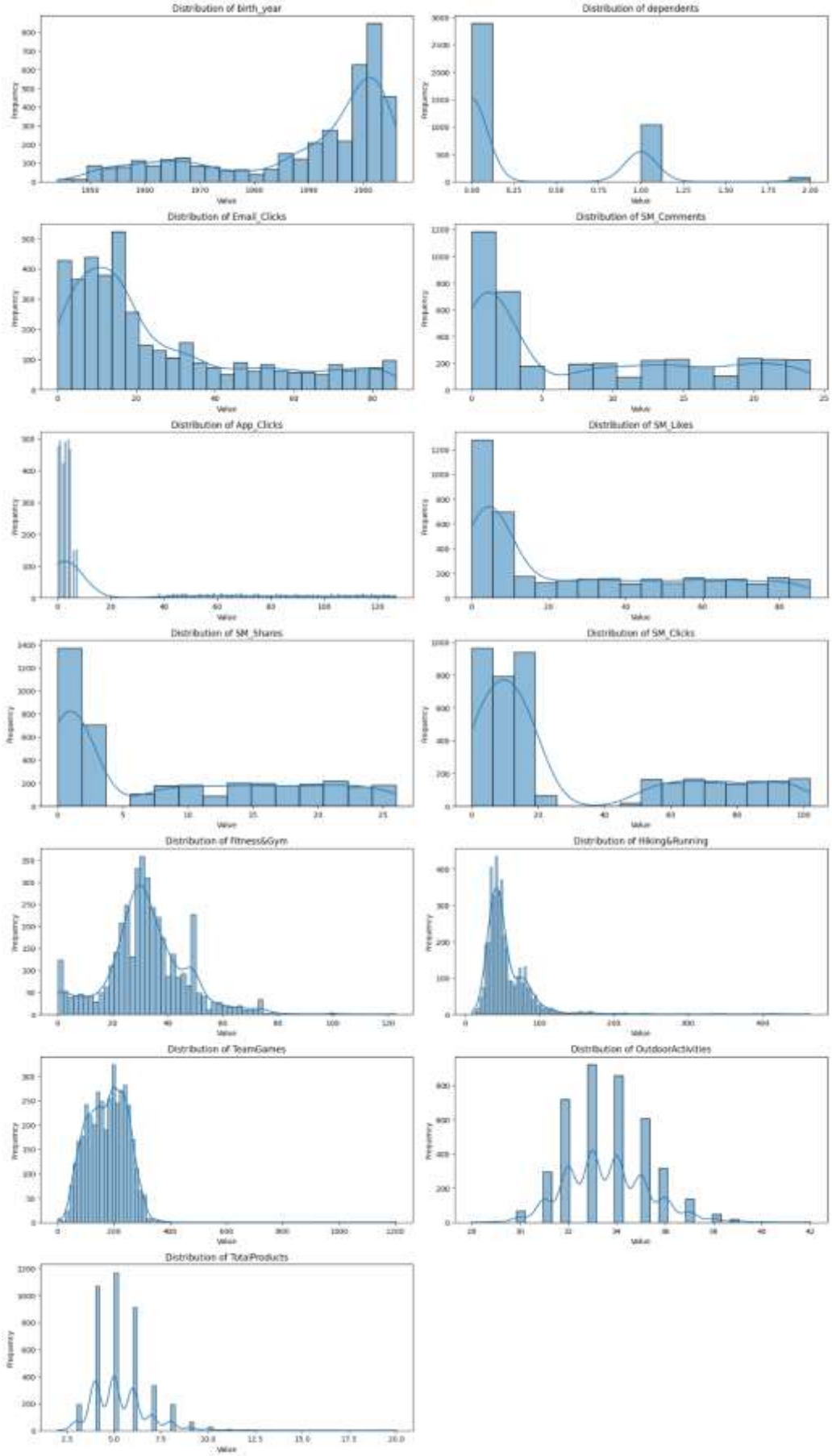


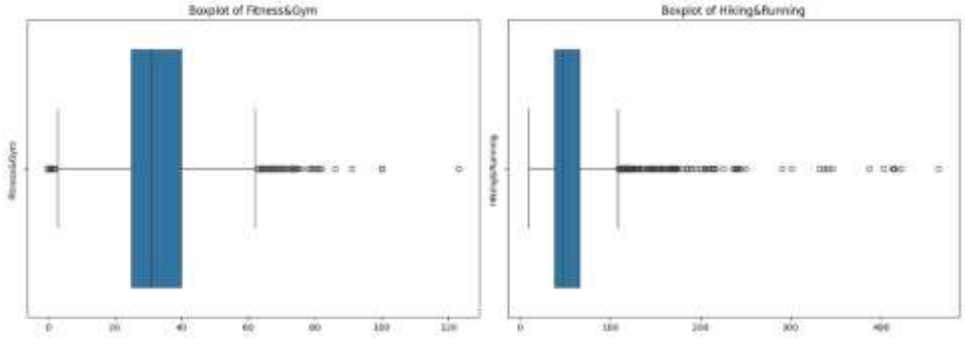
Figure 2 Numerical features distribution

Considering the above charts, the following analysis can be done:

- **Distribution of birth_year**: The histogram shows a distribution that over several decades, so a large variety of ages among our customers, with a higher frequency of younger individuals compared to older ones.
- **Distribution of dependents**: The histogram shows that many individuals have zero dependents, and very few have one or two dependents.
- **Distribution of Email_Clicks, SM_Comments, App_Clicks, SM_Likes, SM_Shares, SM_Clicks**: These histograms appear to show various levels of engagement, with most showing a highly left-oriented distribution. This means that many individuals have low engagement, with few individuals showing very high engagement levels.
- **Distribution of Fitness&Gym, Hiking&Running, TeamGames, OutdoorActivities**: These histograms have a high variability. Some show a normal distribution, while others have a high pick. For instance, the distribution of Hiking&Running shows a very long right tail, representing extreme values or outliers.
- **Distribution of TotalProducts**: This histogram indicates that most individuals buy a relatively low number of total products, with the number of products per individual decreasing as the quantity increases.

2.3.2 Boxplot

Boxplots were employed to illustrate the spread and outliers in the numerical data.



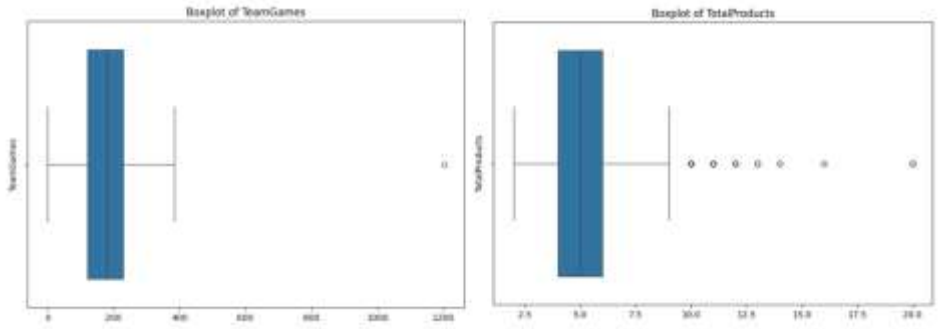


Figure 3 Boxplots of the features identified with outliers.

2.3.3 Heatmap

A heatmap of Spearman correlation coefficients provided insights into pairwise relationships between numerical features, revealing potential associations.

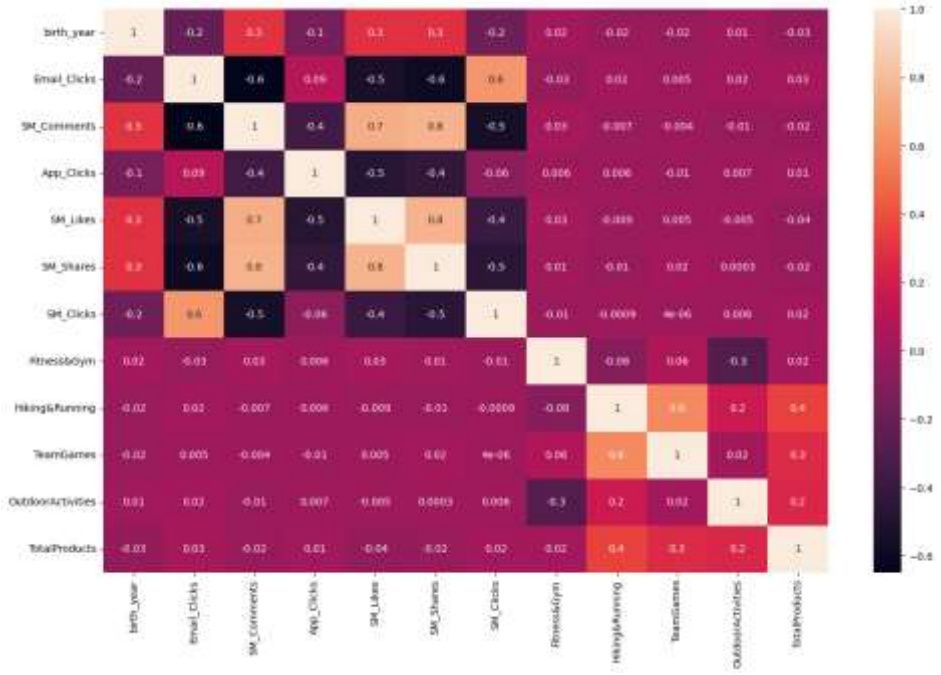


Figure 4 Heatmap of Spearman.

Although we identify some correlation within the features of each of the datasets digital and products, we concluded that there are no perfect positive or negative correlations between any variables to perform a reduction of the features.

2.3.4 Pairplots

- Making use of the pairplots, we explored spending behaviours across different product categories by analysing the total products purchased. Pairwise relationships among variables in specific sub-categories were also examined through pair plots.

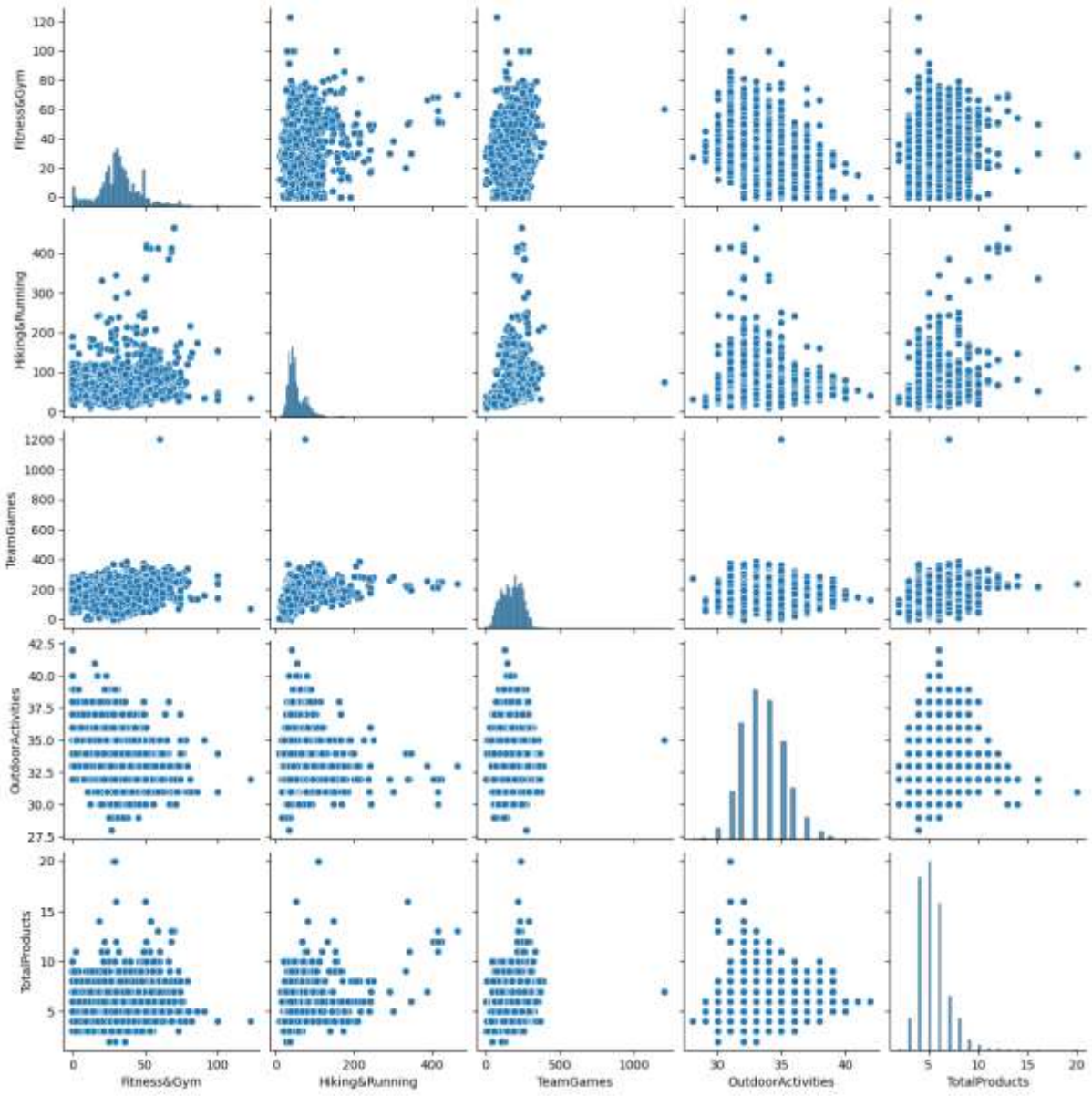


Figure 5 Pairplots for the products dataset.

TotalProducts seems to have a positive correlation with **'TeamGames'** as the scatter plot between them shows a pattern that might indicate that teams who play more games might also purchase more products. Also, we noticed that there may be a correlation between **'HikingRunning'** and **'OutdoorActivities'** as seen in the scatter plot that compares

these two variables. The points seem to show a pattern from lower left to upper right, indicating that as participation in hiking and running increases, so might involvement in outdoor activities.

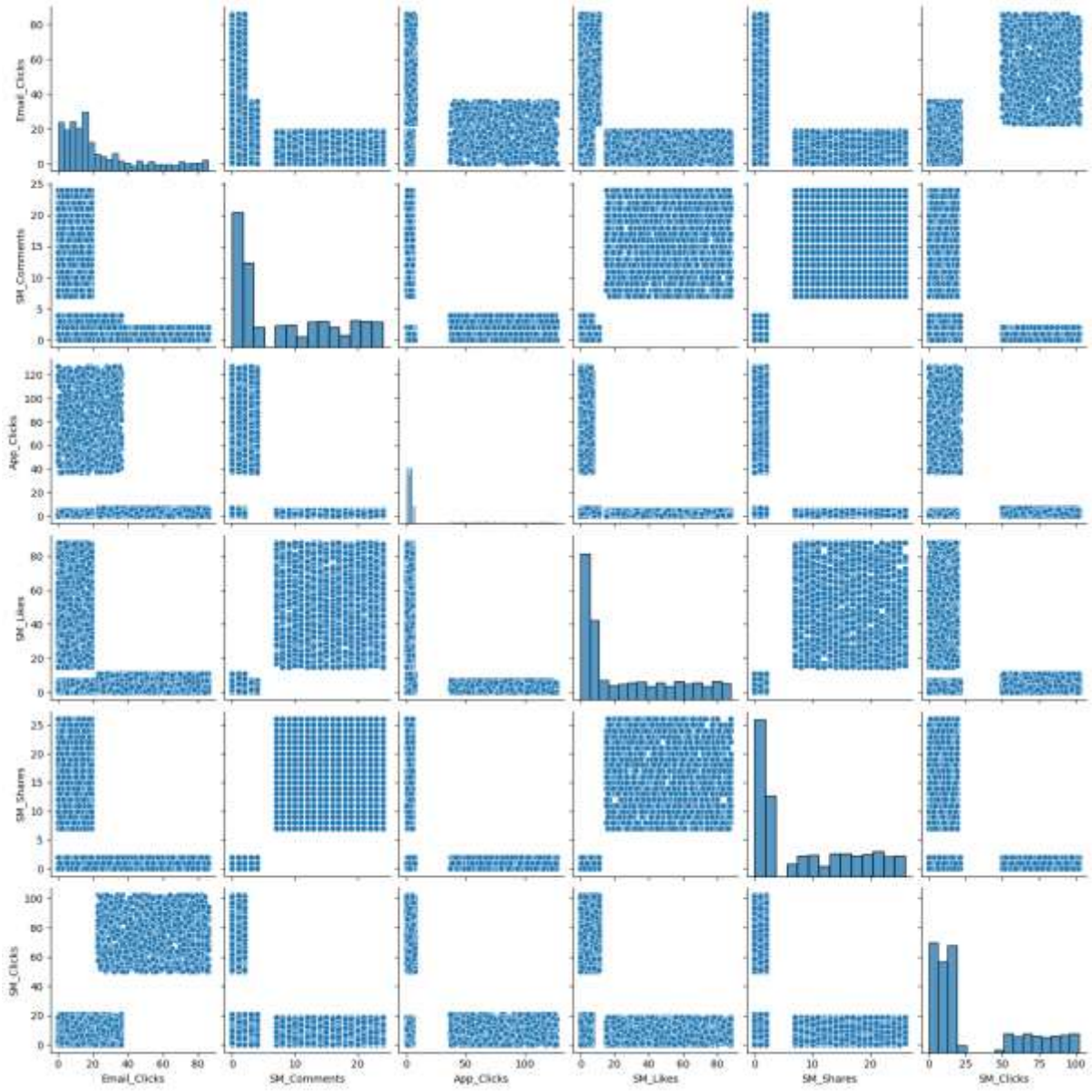


Figure 6 Pairplots for the digital dataset.

We see that most of the scatterplots show a high concentration of points near the origin (0,0) and some plots show lacking areas indicating fewer users with high engagement levels. The plots indicate varying levels of digital engagement among the individuals. Individuals with high **email clicks** don't necessarily have high **social media comments, likes, or shares**.

3. Preprocessing

3.1 Outlier Removal Process

We used boxplots to identify outliers within our dataset that captures customer activities and purchases. By setting thresholds for what we consider extreme values, we were able to flag these outliers in categories such as 'FitnessGym', 'HikingRunning', 'TeamGames', and 'TotalProducts'. After sorting the outliers, we proceeded to remove these atypical data points using the drop() method, with the help of loc[ ] method.

3.2 Inconsistencies

In this chapter, we tackled data inconsistencies by cleaning up our dataset. We normalized the 'education_level' entries by ensuring uniform capitalization and abbreviations. Then, we corrected a common misspelling in the 'City' column ('Brimingham' to 'Birmingham'), and lastly, we addressed a possible data entry error in the 'dependents' column by replacing incorrect values (2 for 1).

3.3 Missing Values

The missing values previously identified in ‘City’ and ‘SM_Shares’, were updated based on different methods:

- For the ‘City’ variable, the missing values were replaced by a new value called ‘unknown’
- For the ‘SM_Shares’ variable we opted to use a k-Nearest Neighbors algorithm to predict and fill in the missing 'SM_Shares' data based on similar data points

3.4 Create New Variables

In this step we focused on transforming existing data into a format more suitable for modeling and creating new features that could enhance our cluster analysis.

- **‘Age’**: From the birth_year provided in the Demographic dataset, we derived the age of customers. This continuous variable was calculated as the difference between the

current year and the birth year, providing a direct measure of customer age which is more useful than birth year for segmentation purposes.

- **'Gender'**: From the ‘name’ variable we’ve created a new ‘Male’ variable by assigning the prefix ‘Mr’ to 1 and ‘Miss’ to 0. From there, we mapped 'F' for female (0) and 'M' for male (1), allowing for more intuitive interpretation of the gender data.
- **‘Recency'**: Measures the number of days since the last purchase of customers, allowing us to track recent customer activity. We do this by converting the 'Last_Purchase' dates to a datetime format and calculating the difference between the current date and these dates.

3.5 Categorical Encoding

In this step, we are transforming categorical data into a format that can be provided to machine learning models.

- **Ordinal Encoding**: We have converted the '**education_level**' and **‘age’** column into an ordinal variable. This means we have replaced the categorical descriptions with an ordered numeric scale, where each level of education is assigned to a unique integer based on its perceived order of importance or level.
- **One-Hot Encoding**: It involves creating a new binary column for each category of the **'City'** variable, where the presence of a category is marked by a 1 and absence by a 0.
- **Binning:** We have defined age bins, which are ranges of ages, to categorize the continuous **'Age'** data into discrete groups such as '18-24 Years', '25-34 Years and so on. We then created a new variable, 'age_group_encoded', which maps these categorical age groups to a numeric code (Ordinal Encoding).

3.6 Power Transform

The primary objective of this step was to normalize the distribution of **'Hiking&Running'** variable. Square root, cube root, and seventh root transformations were systematically applied, and their effectiveness was evaluated by observing changes in kurtosis.

3.7 Data Reduction

In the process of data reduction, we opted to eliminate the variables name, birth_year, City (almost 50% of missing values), and last_purchase from our dataset, as they did not contribute additional value. Some variables were solely utilized for testing purposes and did not factor into identifying or characterizing the final clusters.

3.8 Scaling

This step involves scaling features from different segments of the dataset using the Min-Max Scaler to standardize the values between 0 and 1. This normalization process ensures uniformity across various features.

3.9 PCA

Principal Component Analysis (PCA) is applied to the scaled dataset to potentially reduce dimensionality. We initiate PCA to capture 99% of the data's variance, transform the dataset accordingly, and assess the variance explained by each component. However, due to the limited number of variables, we conclude that PCA may not significantly benefit our analysis and choose not to proceed with it to maintain clear interpretation of our data.

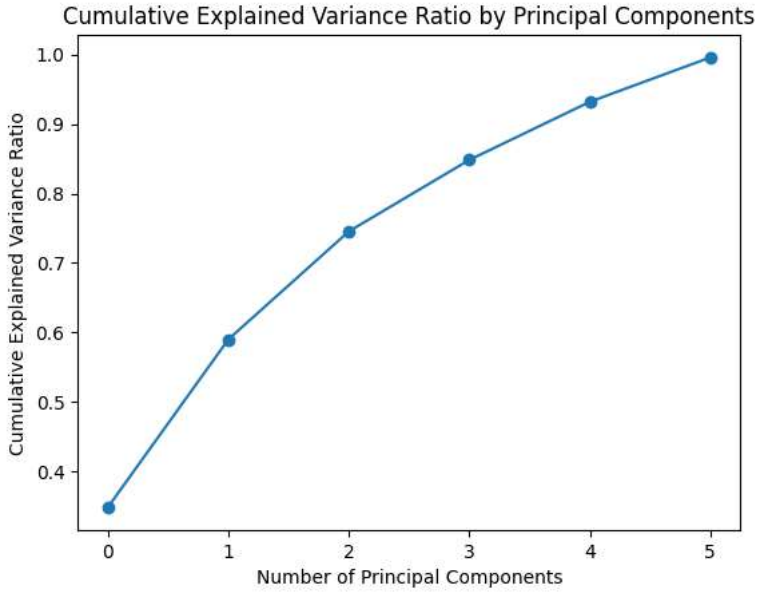


Figure 7 PCA analysis.

4. Modelling

4.1 Selection of Clustering Algorithm

For this project, we chose the K-Means clustering algorithm due to its efficiency and effectiveness in identifying distinct groups within large datasets. K-Means is particularly suitable for our datasets as it performs well with numerical data and is computationally less intensive, allowing for scalability.

K-Means is known for its simplicity and speed in convergence. Given our need to segment customers based on their spending patterns and digital engagement, K-Means was an ideal choice.

4.2 Determining the Number of Clusters

Determining the optimal number of clusters is crucial for effective segmentation. We utilized the Elbow Method, a popular heuristic used in K-Means to determine the number of clusters by plotting the sum of squared distances from each point to its assigned center (inertia).

- **Elbow Method**: We plotted the inertia for different numbers of clusters, ranging from 1 to 10, and looked for the "elbow point" where the rate of decrease sharply shifts. This point represents a balance between having too many or too few clusters, each providing diminishing returns in reducing the sum of squared distances.

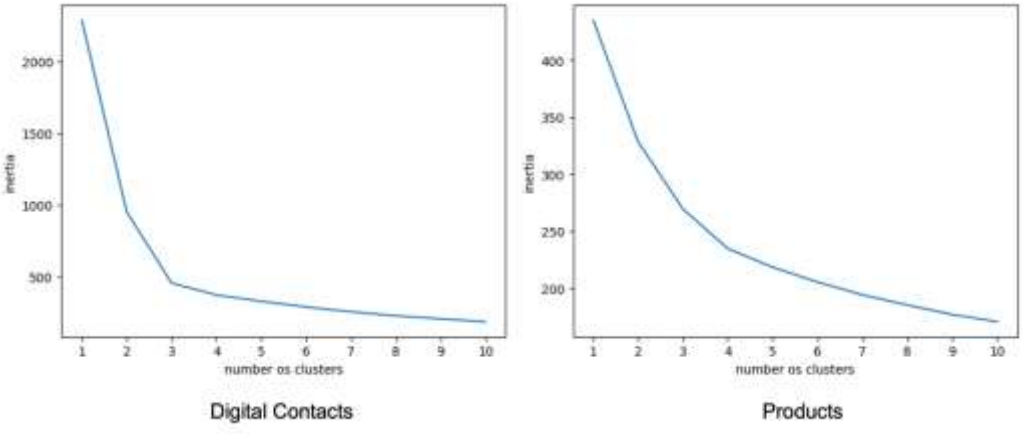


Figure 7 Inertia charts.

- **Dendrogram**: A dendrogram was constructed using a subset of 1000 samples to visualize and assess the data's cluster structure. The random_state parameter was set

to 100 to ensure the reproducibility of the sample selection process, allowing for consistent results across multiple runs of the algorithm. The dendrogram provided visual means to determine the potential number of clusters and acts as a confirmation of the Elbow Method.

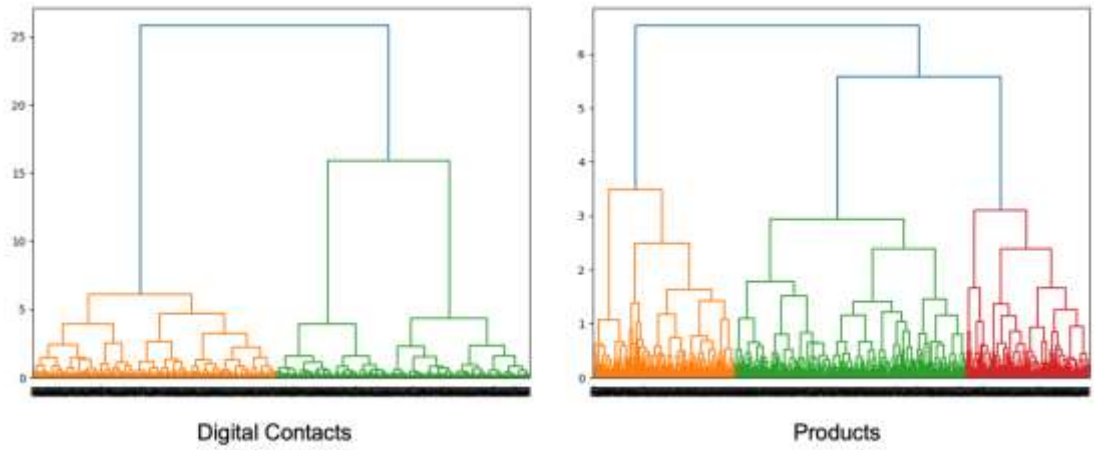


Figure 8 Dendrograms.

- **Silhouette Score**: To validate our choice, we also calculated the silhouette scores for the potential number of clusters. This measure helped confirm the cluster cohesion and separation, ensuring that the segments are well-defined and distinct.

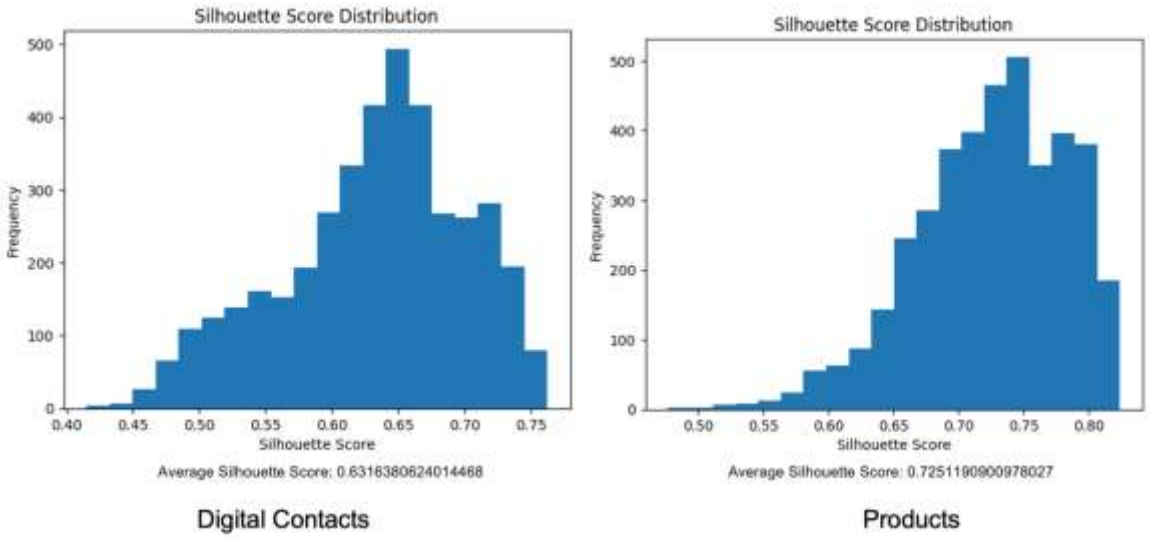


Figure 9 Silhouette Chart and Score.

Based on the 3 previous methods before, we concluded that we would execute K-Means with K = 2 and 3 for the Digital Contacts, and with K = 3 and 4 for the Products.

4.3 Implementation of K-Means Clustering

After establishing the appropriate number of clusters, we proceeded with the K-Means implementation.

- **Feature Set**: We prepared separate feature sets for each dataset—Digital Contact and Products—ensuring that only relevant and pre-processed features were included in the clustering process.
- **Algorithm Parameters**: For the Digital Dataset, K-Means was configured with 2 and 3 clusters based on the results obtained in the previous step. For the Product Dataset, K-Means was configured with 3 and 4 clusters.
- **Execution**: The clustering was executed on each dataset independently, allowing us to segment customers based on digital engagement and product preferences.

5. Description of Resulting Clusters

5.1 Overview of Clusters

Each dataset analysed through K-Means clustering produced distinct segments of customers. Here, we delve into the specifics of these clusters, exploring their unique characteristics and potential implications for business strategies.

5.2 Digital Contact Clusters

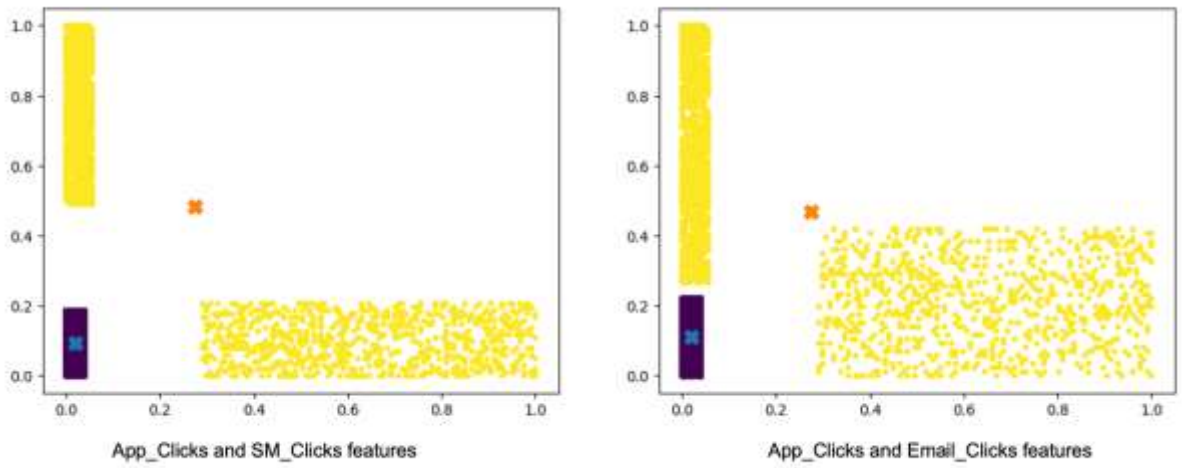


Figure 10 Scatterplot for K=2 (Digital Content).

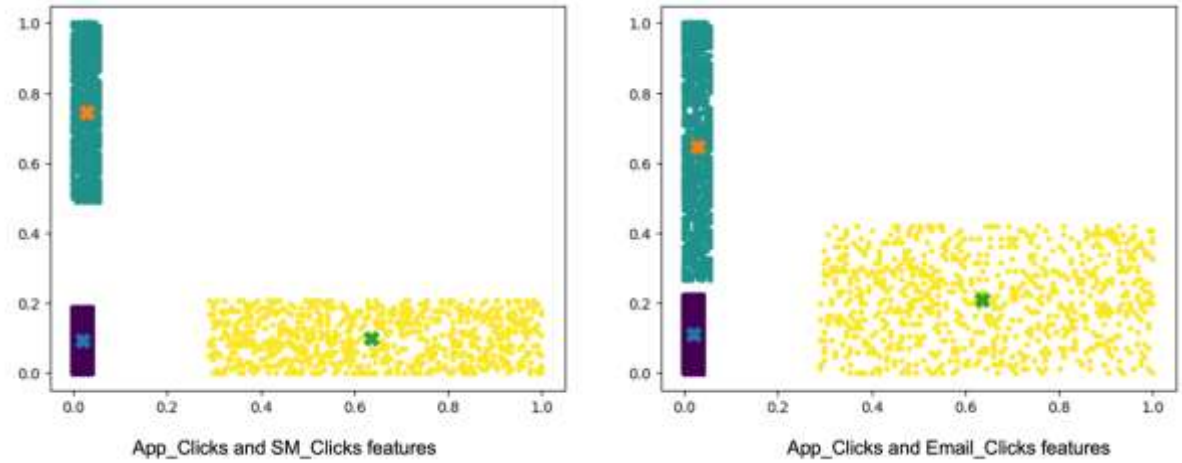


Figure 11 Scatterplot for K=3 (Digital Content).

- **Cluster Characteristics**:
  Looking to the mean for the features per cluster, we can start to understand the distinct behaviour of the costumers according to the cluster they belong to:

  - **Cluster 0:** The customers in this cluster engage the most through SM_Comments, very highly through SM_Shares and SM_Likes, very little through Emai_Clicks and SM_Clicks and show almost no engagement through App_Clicks . This cluster is composed of 1903 customers.

- **Cluster 1:** These are the customers who engage the least through App_clicks and engage the most through SM_Clicks and Email_Clicks. The cluster is composed of 1243 individuals.

- **Cluster 2:** This cluster is composed of 854 customers and can be considered the cluster that groups individuals who engage the most through App_Clicks, moderately through Email_Clicks and almost don't engage through the rest.

Looking at the Euclidean distance between the clusters we can conclude that:

- The nearest cluster to cluster 0 is Cluster 2 (1.17)
- The nearest cluster to cluster 1 is Cluster 2 (0.98)
- The nearest cluster to cluster 2 is cluster 1 (0.98)
- Cluster 0 and Cluster 1 are the farthest away (1.31)

In this way, Cluster 0 and Cluster 1 are the farthest apart, while Cluster 2 is the nearest to both. This suggests that there is less overlap between the digital engagements of Cluster 0 and Cluster 1, while Cluster 2 may share some similarities with both.

"What distinct customer segments exist based on their engagement behaviour across email, mobile app, and social media platforms?"

We identified 3 different customer segments:

- **Cluster 0:** Customers that engage more with social media through likes, comments and shares.
- **Cluster 1:** Customers who engage more through links/ads received via emails and social media.
- **Cluster 2:** Customers that have a high engagement through the Sportify app.

5.3 Products Clusters

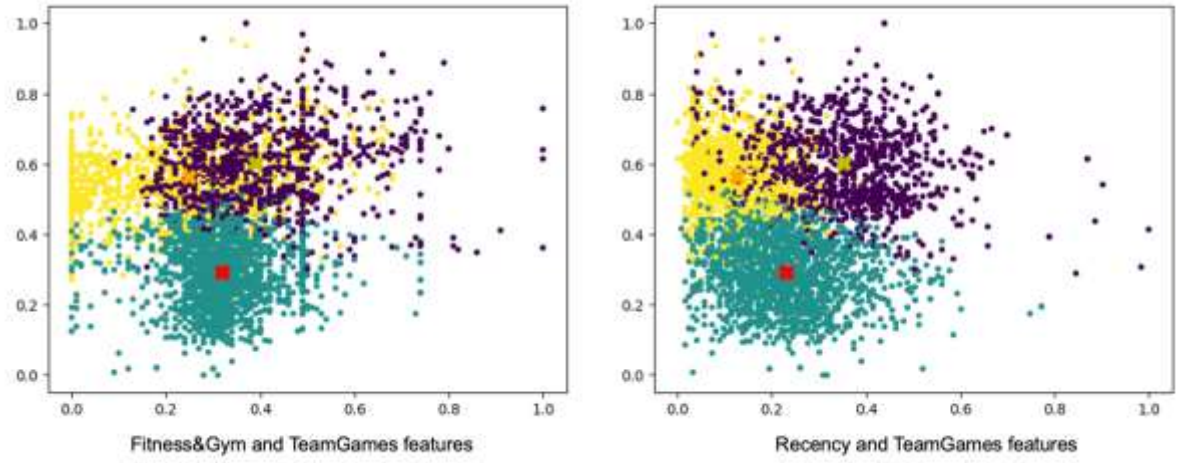


Figure 12 Scatterplot for K=3 (Products).

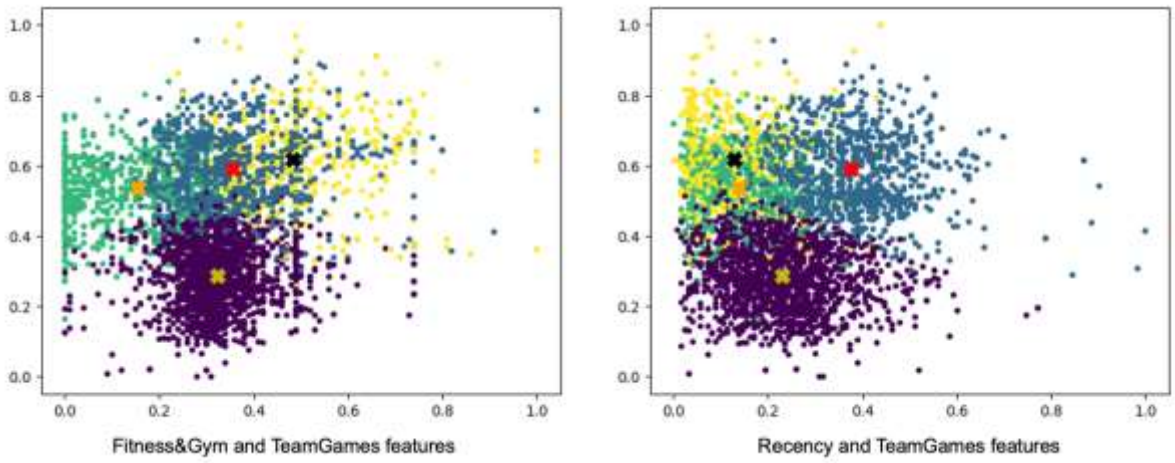


Figure 13 Scatterplot for K=4 (Products).

- **Cluster Characteristics**:
  Based on the provided data, the characteristics that define each cluster are as follows:

  - **Cluster 0:** This cluster exhibits higher values for 'TeamGames' compared to the other clusters. It also has reasonably high values for 'Fitness&Gym' and 'Hiking&Running'. The 'Recency' value is higher compared to the other clusters indicating older purchases.

- **Cluster 1:** The most pronounced characteristic in this cluster is that it has the lowest value for 'TeamGames', 'Hiking&Running' and ‘TotalProducts’ indicating that they probably are the customers that spend less.

- **Cluster 2:** This cluster shows the highest values for Hiking&Running' and 'OutdoorActivities', indicating a significant engagement in outdoor and hiking/running activities. The 'TotalProducts' value is also relatively high, suggesting a substantial level of purchasing activity. 'Recency' values are lower, suggesting less time since the last purchase.

Looking at the Euclidean distance between the clusters we can conclude that:

- The nearest cluster to cluster 0 is Cluster 2 (0.33)
- The nearest cluster to cluster 1 is Cluster 0 (0.35)
- The nearest cluster to cluster 2 is cluster 0 (0.33)
- Cluster 1 and Cluster 2 are the farthest away (0.38)

In this way, Cluster 1 and Cluster 2 are the farthest apart, while Cluster 0 is the nearest to both. This suggests that there is less overlap between the spending behaviors of Cluster 1 and Cluster 2, while Cluster 0 may share some similarities with both.

Additionally, we analysed the average spending per cluster, and concluded that Cluster 1 customers are the ones that in average spend less (216.69 units) and Cluster 0 and Cluster 2 customers spend about the same in average (358.09 and 358.27 units, respectively).

"Can we identify unique customer segments by analysing how they purchase across various product categories and their buying patterns?"

We identified 3 different customer segments:

- **Cluster 0:** Customers that spend more in TeamGames and the ones that made a less recent purchase.
- **Cluster 1:** Customers that spend less overall without having a strong buying category.
- **Cluster 2:** Categories that stand out are Hiking & Running and Outdoor Activities. These are also the customers that buy more products and more recently.

5.4 Integrated Cluster Analysis

After defining the clusters for both Digital Contacts and Products, we are now crossing that information with the corresponding demographics.

For Digital Contacts:

- **Cluster 0**: 86% are female. About half of the customers are on the group of 18-24 years old. 73% don't have dependents.
- **Cluster 1**: Almost 70% are male. 91% in the age groups of 18-34 years old (young adults) and 55-80 years old (seniors). 72% don't have dependents.
- **Cluster 2**: 61% of them male. About 60% of the customers have between 18-34 years of age. 72% don't have dependents.

We don't see any relevant information to be extracted by the education level feature.

For Products:

- **Cluster 0**: 61% are female. About 60% of the customers have between 18-34 years of age. 65% don't have dependents.
- **Cluster 1**: There are no relevant differences in the gender. About 67% of the customers have between 18-34 years of age. 74% don't have dependents.
- **Cluster 2**: 59% are female. About 65% of the customers have between 18-34 years of age. 77% don't have dependents. 26% of the customer have a Master or PhD degree.

Aggregating this data with the cluster information in 5.2 and 5.3, we have:

Digital Contacts:

Cluster 0: Customers that engage more with social media through likes, comments and shares. 86% are female. About half of the customers are on the group of 18-24 years old. 73% don't have dependents.

Cluster 1: Customers who engage more through links/ads received via emails and social media. Almost 70% are male. 91% in the age groups of 18-34 years old (young adults) and 55-80 years old (seniors). 72% don't have dependents.

Cluster 2: Customers that have a high engagement through the Sportify app. 61% of them male. About 60% of the customers have between 18-34 years of age. 72% don't have dependents.

Products:

Cluster 0: Customers that spend more in TeamGames and the ones that made a less recent purchase. 61% are female. About 60% of the customers have between 18-34 years of age. 65% don't have dependents.

Cluster 1: Customers that spend less overall without having a strong buying category. There are no relevant differences in the gender. About 67% of the customers have between 18-34 years of age. 74% don't have dependents.

Cluster 2: Categories that stand out are Hiking & Running and Outdoor Activities. These are also the customers that buy more products and more recently. 59% are female. About 65% of the customers have between 18-34 years of age. 77% don't have dependents. 26% of the customer have a Master or PhD degree.

5.5 DBSCAN Algorithm

We also employed the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to conduct a comparison with the previous K-Means clustering algorithm. This unsupervised clustering technique, capable of identifying clusters based on dense regions within data, allowed to contrast the results obtained from DBSCAN with K-Means, exploring the differences in clustering outcomes and understanding the distinctive characteristics of each approach.

On the Digital_Contact dataset, DBSCAN and K-Means produced similar clustering results, indicating consistency and reliability in the methodologies used. However, when examining the Products dataset, DBSCAN's outcomes differed significantly from K-Means, highlighting the need to carefully match clustering algorithms to the specific characteristics of each dataset for accurate and relevant analyses. This emphasizes the importance of evaluating the suitability of clustering algorithms to ensure the relevance and accuracy of the conducted clustering analyses.

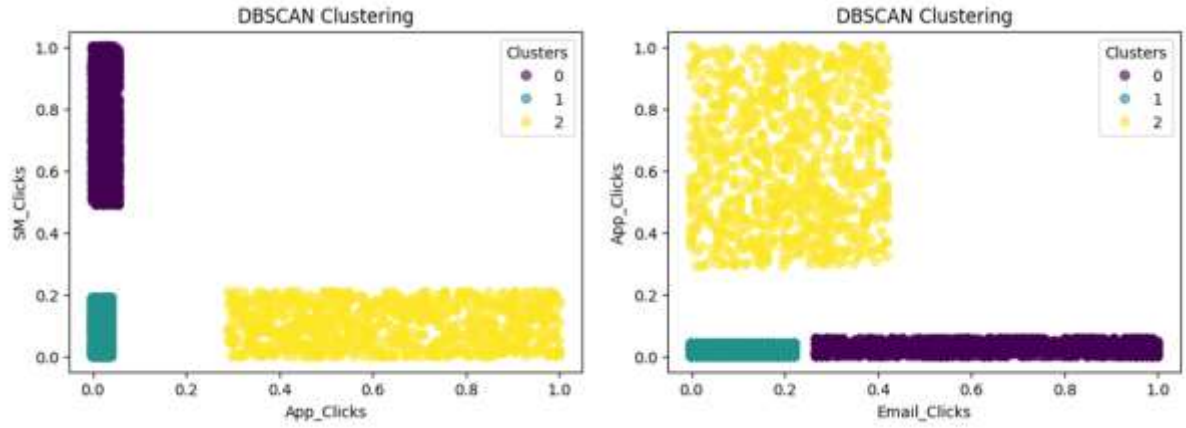


Figure 14 Scatterplots using DBSCAN (Digital_Contacts)

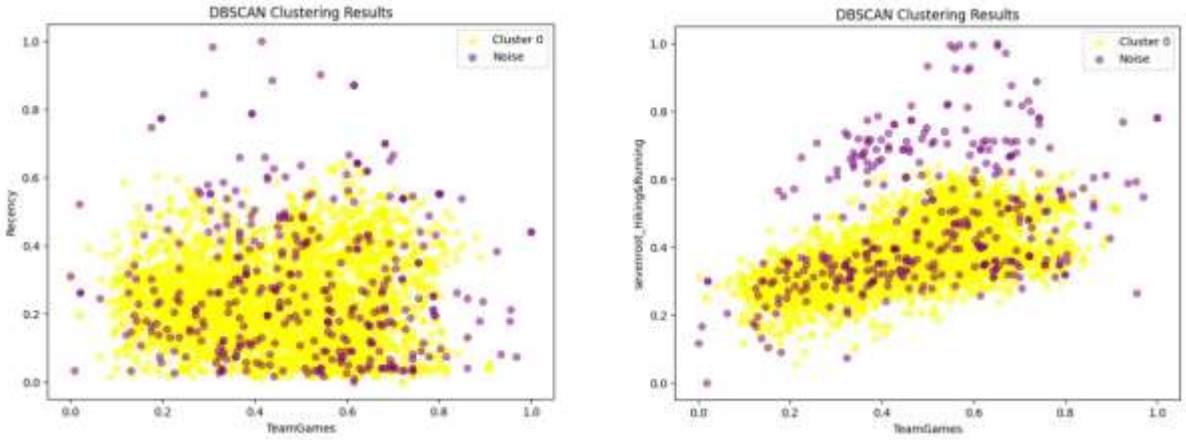


Figure 15 Scatterplots using DBSCAN (Products)

6. Action Plan

Through a deep understanding of each segment's distinct traits, we crafted a customized marketing strategy to resonate with the unique preferences of every group.

6.1 Digital_Contacts

Cluster 0: Social Media Engagers

The action plan for Social Media Engagers includes the development of highly shareable and engaging content to encourage likes, comments, and shares. Additionally, partnering with influencers who match the demographic profile of young, female users without dependents will boost brand visibility. Furthermore, implementing social-media loyalty programs that reward comments and shares will directly benefit users and enhance engagement.

Cluster 1: Email and Social Media Clickers

The proposed plan for Email and Social Media Clickers involves personalized email content to sustain interest and engagement, insightful cross-platform retargeting to seamlessly connect email engagement with social media interactions, and the utilization of data insights to refine content and maximize engagement and return on investment (ROI).

Cluster 2: App Engagers

The strategy for App Engagers entails running exclusive promotions through the Sportify app to boost engagement and encourage frequent app usage. Regular app feature updates tailored to this cluster's preferences will enhance user experience and satisfaction. Additionally, organizing in-app events or challenges aims to drive deeper user interaction with the app.

6.2 Products

Cluster 0: Team Games Enthusiasts

The plan for Team Games Enthusiasts involves organizing local events or online tournaments to foster community building and drive product purchases related to team sports. Additionally, personalized marketing messages will highlight fitness & gym and hiking & running products relevant to this cluster, and feedback mechanisms will be implemented to gather insights on recently purchased items, ensuring continuous improvement of future offerings.

Cluster 1: Low-Spend Generalists

To attract Low-Spend Generalists, promotional discounts and bundled offers will be utilized to boost spending and transition customers to more engaged segments. Engagement surveys will be conducted to understand their preferences better and uncover new interests for targeted marketing strategies. Providing educational content on product benefits and uses aims to stimulate interest and potentially drive increased spending within this demographic.

Cluster 2: Outdoor Activity and Frequent Buyers

To cater to the Outdoor Activity and Frequent Buyers cluster, the strategy involves introducing premium outdoor and hiking products that emphasize quality and durability, justifying a higher price point. Additionally, loyalty programs will be enhanced to offer tiered rewards that escalate with purchase frequency and volume. Emphasizing eco-friendly products and company practices in communications aligns with the active and likely sustainability-oriented values of this customer cluster.

7. Conclusion

Through the use of analytical methods, including unsupervised clustering algorithms, we successfully identified distinct customer clusters, yielding valuable insights into Sportify customer segments. With this knowledge, we crafted a tailored marketing action plan to meet the needs of each specific customer segment, thereby facilitating the expansion of this sports business.

Annexes

Annex1 – KNN Imputer

KNN Imputer, or K-Nearest Neighbours Imputer, addresses missing data within datasets by employing the KNN algorithm. It estimates missing values by using the feature values of the k nearest neighbours to the data point lacking information. The method calculates the distance between data points and imputes missing values based on the computed mean of the nearest neighbours features. This technique is straightforward, suitable for both numerical and categorical data, and maintains the structural integrity of the original data.

Annex2 – Silhouette Method

Silhouette Analysis measures the cohesiveness of clusters generated by algorithms like K-Means. Silhouette scores, which range from -1 to 1, indicate the proximity of data points to their cluster compared to neighbouring clusters. Scores closer to 1 suggest that points are well-separated from other clusters, while scores near 0 indicate proximity to a neighbouring cluster. Negative values flag potential misassignments to incorrect clusters. Calculating the average silhouette score for a cluster helps assess whether its points are distinct from those of other clusters.

Annex3 – DBSCAN

DBSCAN, or Density-Based Spatial Clustering of Applications with Noise, is an unsupervised clustering algorithm that forms clusters based on regions of high density. Unique to DBSCAN is its capacity to identify outlier points, which are labelled as noise and not included in any cluster. Clusters are defined by a core point, which is determined based on a pre-set minimum number of neighbours and a radius (epsilon value). The algorithm assesses all points within this radius to decide cluster memberships. Unlike K-Means, DBSCAN does not assume a specific cluster shape, making it versatile across varied data shapes.

Annex4 – PCA

PCA, or Principal Component Analysis, is a statistical tool used to analyse and simplify datasets with many dimensions or features per observation. Its primary goal is to reduce the

number of dimensions while preserving as much information as possible. This simplifies interpretation by facilitating the visualization of intricate multidimensional data, making data sets easier to manage and explore. In PCA, the original variables are transformed into new variables known as principal components. These components are linear combinations of the original variables, structured to maximize data variance along each component, capturing significant data patterns and variations effectively. Although PCA is advantageous for reducing data and mitigating noise, it does have downsides, such as reduced interpretability since principal components represent a mix of original variables, and potential information loss which can slightly alter the analysis.

References

https://scikit-learn.org/stable/modules/impute.html#knnimpute

https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html

https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html