Machine Learning Group 18

Baran Can Çelik (20232067), Carlos Lourenço (20232020), Kida Aly (20231491), Priyá Dessai (20232053) and Raysa Rocha (20232051).

Table of Contents

Abstract **4**
1 Exploration **5**
1.1 Digital Contact **5**
1.1.1 Statistical Exploration 5
1.1.2 Visual Exploration 5
1.2 Products **6**
1.2.1 Statistical Exploration 6
1.3 Demographic **7**
1.3.1 Statistical Exploration 7
2 Preprocessing **8**
2.1 Digital Contact **8**
2.1.1 Handling missing values 8
2.1.2 Normalizing the data 8
2.2 Products **8**
2.2.1 Replacing Last_Purchase column 8
2.2.2 Exploring and removing outliers 8
2.2.3 Normalizing Data 9
2.3 Demographic **9**
2.3.1 Data transformation 9
3 Modelling **10**
3.1 Digital Contact **10**
3.1.1 Defining the number of clusters 10
3.1.2 Training the model with K-Means 10
3.1.3 Applying K-means after performing PCA 13
3.1.4 Clusters with DBSCAN 14
3.1.5 Applying t-SNE 16
3.1.6 Applying DBSCAN after performing t-SNE 16
3.2 Products **17**
3.2.1 Defining the number of clusters 17
3.2.2 Training the model with K-Means 18
3.2.3 Applying K-means after performing PCA 19
3.2.4 Applying DBSCAN 21
3.2.5 Applying t-SNE 22
3.2.6 Applying K-means after performing t-SNE 22
4 Description of Resulting Clusters **23**
4.1 Results from Digital Contact clustering **23**
4.2 Results from Products clustering **24**

4.3 Results from processing clustering 24
4.4 Our Customers Overview: 24
5 Action Plan 29
6 Conclusion 30
References 33
7 Annexes 34
7.1 KNN Imputer 34
7.2 Silhouette Method 34
7.3 PCA 35
7.4 t-SNE 35
7.5 DBSCAN 36

Abstract

In this project, we analyze customer interactions and behaviors within Sportify, a company specializing in sports products and gear. Our analysis is based on three primary datasets: Digital Contact, Products, and Demographics. The Digital Contact dataset captures customer interactions through the company's app, email, and social media channels. The Products dataset provides information on Sportify's product offerings, while the Demographics dataset includes customer demographic details such as customer name, education level, dependency.

Our analysis begins with data pre-processing to ensure data quality, including cleaning to address discrepancies and removing outliers. Additionally, we improve the Products dataset by adding a new variable to better identify the number of days since last purchase, and we enhance the Demographics dataset by introducing new variables that enrich the customer profiles such as gender and age.

Next, we employ unsupervised clustering algorithms, including K-Means and DBSCAN, to segment the customer base based on their interactions on Digital platforms, purchasing behavior, and frequency of purchases. To enhance our analysis and reduce dimensionality, we integrate these methods with techniques like Principal Component Analysis (PCA) and t-Distributed Stochastic Neighbor Embedding (t-SNE).

Through comprehensive evaluation and comparison of various clustering algorithms, we could define our final clusters and their unique characteristics. This analysis allowed us to optimize our clustering methodology, identifying valuable customer segments that can guide targeted marketing strategies and business decisions for Sportify.

1 Exploration

We began our data exploration by examining three datasets: digital interactions, sports product spending and demographic information. The initial steps included reviewing variable descriptions, identifying patterns, data types, and assessing the presence of missing values. Overall, the objective was to determine which variables were pertinent for the analysis and which ones could be omitted to ensure the accuracy of data analysis. This process likely involved exploring each dataset to understand its structure and content, ultimately preparing the data for further analysis and modelling.

1.1 Digital Contact

1.1.1 Statistical Exploration

Starting the file Digital_Contact.csv with the main descriptive statistics for all the (numeric) variables.

	count	mean	std	min	25%	50%	75%	max
Email_Clicks	4000.0	25.762250	23.659980	0.0	8.0	16.0	36.0	86.0
App_Clicks	4000.0	19.526500	34.237945	0.0	2.0	4.0	7.0	127.0
SM_Comments	4000.0	8.247750	8.064963	0.0	1.0	4.0	15.0	24.0
SM_Likes	4000.0	26.957000	27.742658	0.0	4.0	11.0	50.0	88.0
SM_Shares	3961.0	8.355971	8.696192	0.0	1.0	2.0	16.0	26.0
SM_Clicks	4000.0	30.349250	32.254974	0.0	7.0	15.0	60.0	102.0

Figure 1 – Main descriptive statistics for Digital Contact dataset.

Some observations were noticeable based on the statistical summary:

- “SM_Shares” had a high standard deviation, indicating that the data points are more spread out.
- “App_Clicks” also exhibited a high standard deviation compared to the mean, suggesting an abnormal distribution with even more spread-out data points.
- “Email_Clicks” has higher first and second quartiles, indicating a different distribution pattern compared to the other variables.

1.1.2 Visual Exploration

1.1.2.1 Analyzing outliers

Using boxplot charts we could determine outliers. Through Histplots it was considered a more detailed observation for Email_Clicks and App_Clicks that were presenting a higher deviation of outliers.

For the outliers, we considered that given our dataset with 4,000 observations, by removing 854 or even 171 data points it wouldn’t be ideal. This reduction in data size could significantly impact the validity of our statistical tests and hide our ability to draw meaningful conclusions from the data. For this reason, we have decided to keep the outliers in the columns App_Clicks and Email_Clicks.

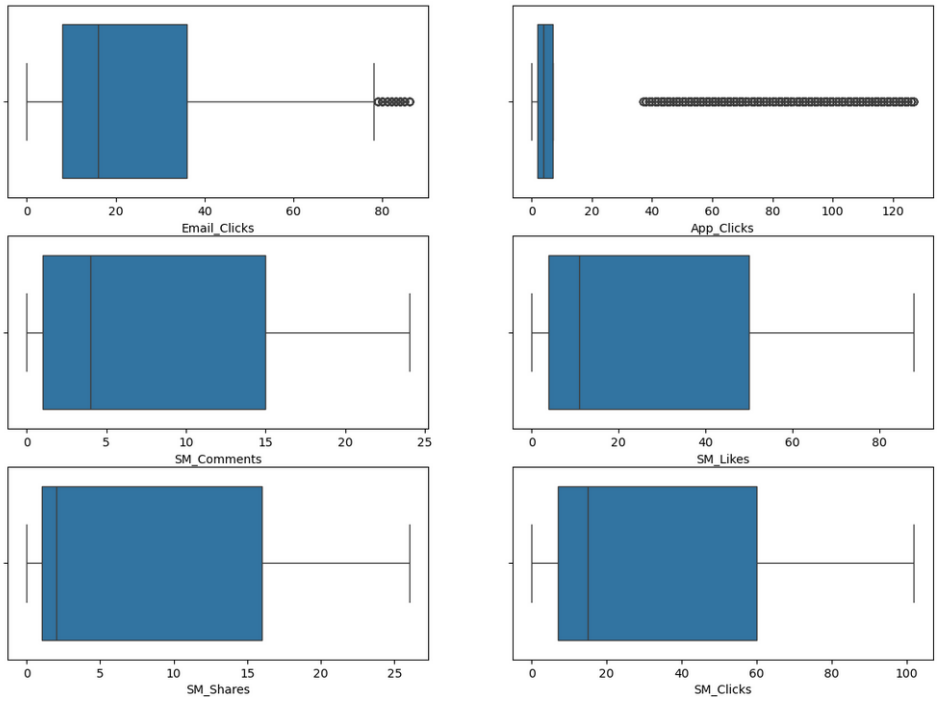


Figure 2 – Digital Contact Box Plot

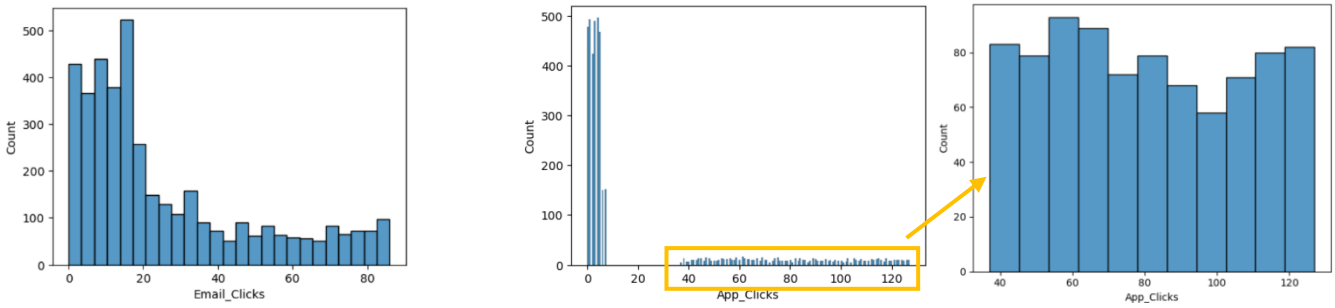


Figure 3 - Histplots for Email Clicks (left) and App Clicks (right)

Email Clicks presented right skewed distribution and zooming App_Clicks values are similar to each other (between 60 and 90), since a pattern on these data was identified, the data was kept.

1.2 Products

1.2.1 Statistical Exploration

	count	mean	min	25%	50%	75%	max	std
Fitness&Gym	4000.0	32.02025	0.0	25.0	31.0	40.0	123.0	14.469284
Hiking&Running	4000.0	55.84475	9.0	38.0	47.0	66.0	464.0	33.119109
Last_Purchase	4000	2024-01-30 18:50:02.399999744	2023-10-15 00:00:00	2024-01-20 00:00:00	2024-02-01 00:00:00	2024-02-13 00:00:00	2024-02-29 00:00:00	NaN
TeamGames	4000.0	176.085	0.0	121.0	180.0	230.0	1203.0	69.964738
OutdoorActivities	4000.0	33.60275	28.0	32.0	33.0	35.0	42.0	1.711483
TotalProducts	4000.0	5.3295	2.0	4.0	5.0	6.0	20.0	1.515257

Figure 4 – Main descriptive statistics for Products dataset.

Analyzing the data presented above we can verify:

- **Fitness&Gym** - Has a considerable difference between the 3rd quartile and the maximum. It might have outliers.

- **Hiking&Running** - Has a high standard deviation (meaning that data points are more spread out) and has a considerable difference between the 3rd quartile and the maximum, meaning that we might have outliers too.
- **TeamGames** - Has the higher mean, standard deviation, 1st, 2nd and 3rd quartiles and the higher maximum value (between columns). It might have outliers too.
- **OutdoorActivities** - Has the lower standard deviation compared to the other columns. It means that data tend to be close to the mean (values vary from 28 to 42).

As shown above, we have a Last_Purchase column showing the date of each customer's last purchase. To improve our analysis, we will introduce a new column named "Days_Since_Last_P" in Preprocessing Chapter, to calculate the number of days since the last purchase. Afterwards, we will identify and analyze any outliers in the "Days_Since_Last_P" column to gain further insights.

1.3 Demographic

1.3.1 Statistical Exploration

Starting the file Demographic.txt with the main descriptive statistics for categorical variables.

	name	education_level	City
count	4000	4000	2019
unique	3892	7	3
top	Mr Michael Jackson	high school	Birmingham
freq	4	1280	1284

Figure 5 – Categorical variables in Demographic data.

Checking the levels/possible values in the variables "Education Level" and “City”

```
education_level
high school              1280
Bachelor                 1023
less than high school     818
Master                    666
PhD                       150
High School                42
PHD                        21
Name: count, dtype: int64
```

```
City
Birmingham    1284
London         700
Brimingham      35
Name: count, dtype: int64
```

- We have 2 different values for "high school" and for "PhD". We need to replace those values for a unique one;
- We have 2 different values for “City” and a lot of missing data.

2 Preprocessing

2.1 Digital Contact

2.1.1 Handling missing values

Since there was missing data in the variable SM_Shares we have tried to apply two methods: filling with 0 and the KNN imputer.

Comparison between different methods:

- Leaving missing values as is: The mean is 8.36 and std is 8.7
- Filling with Zeros: The mean becomes 8.27 while std 8.69.
- Filling with KNN Imputer: The mean becomes 8.36 while std 8.69

After comparing these methods for handling missing data, filling with zeros and using the KNN Imputer (which fills in missing values by referencing similar data points, preserving the local structure of the dataset), we observed minimal differences in statistical properties. Given the simplicity and comparable performance, we decided to fill missing values with zeros as our preferred approach.

2.1.2 Normalizing the data

We have used *MinMaxScaler()* method to normalize our data, in order to have a common scale and range before clustering.

2.2 Products

2.2.1 Replacing Last_Purchase column

As mentioned before, we'll replace the "Last_Purchase" column with "Days_Since_Last_P", which represents the number of days since each customer's last purchase.

2.2.2 Exploring and removing outliers

After transforming the data, we'll explore and remove some outliers to refine the dataset for further analysis.

Using boxplot charts we could verify that there are outliers in every column, as shown below.

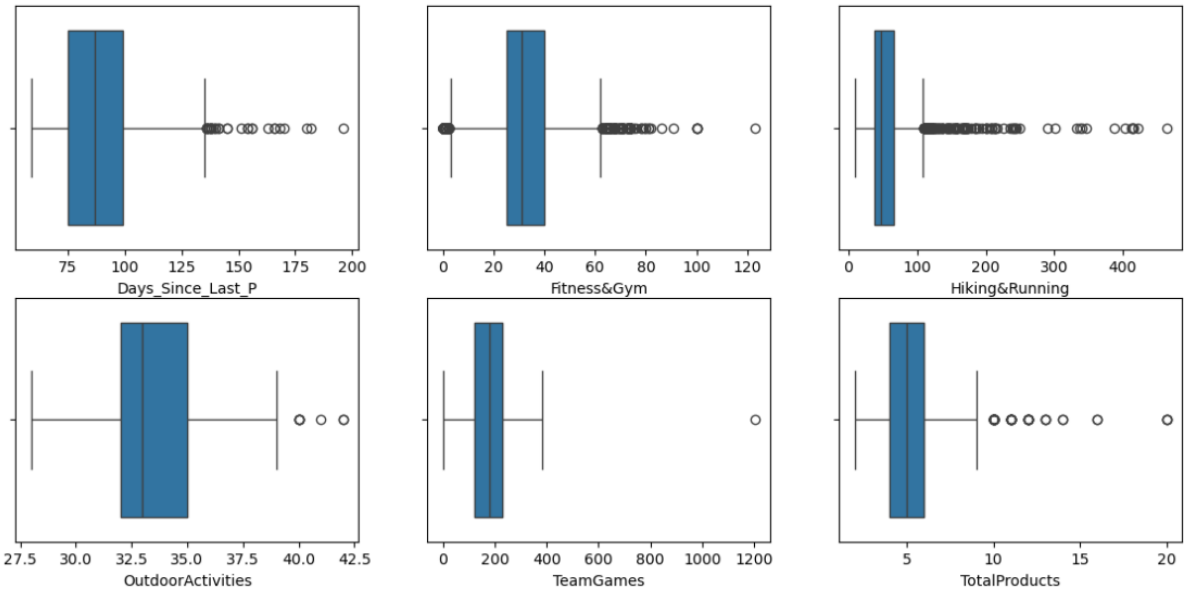


Figure 6 – Transformed Products dataset boxplots.

To minimise data loss while addressing these outliers, we'll strategically remove only the data points that deviate the most from the overall trend.

Given this, we propose removing outliers that deviate significantly from metrics like Fitness&Gym (values above 100) and TeamGames (values above 1000) to ensure maximal data preservation.

After verification, we have observed the removal of 6 rows from Products dataset, resulting in 3994 final rows.

2.2.3 Normalizing Data

We have used *MinMaxScaler()* method to normalize our data, in order to have a common scale and range before clustering.

2.3 Demographic

2.3.1 Data transformation

1. For the missing values in “City”, since it lacked sufficient data for meaningful analysis, we decided to remove them.
2. For the "PHD" and "high school" values in “education_level”, we correct them:

```
education_level
High School              1322
Bachelor                 1023
Less Than High School     818
Master                    666
PhD                       171
```

3. For a better analysis, we decided to create 2 new variables:
    a. “age” from "birth_year" column;
    b. “gender” from “name” column, once we verified it uses "Miss" and "Mr".

From the transformation, we achieved the following results:

Cust_ID	name	education_level	dependents	age	gender
4	Mr Daniel Spencer	Master	1	22	Male
5	Miss Abigail Garcia	High School	0	30	Female
6	Miss Laura Williams	PhD	1	19	Female
9	Mr Justin Hamilton	High School	0	21	Male
10	Mr Steven Vaughn	Less Than High School	0	22	Male

Figure 7 - Demographic after Data Transformation.

3 Modelling

3.1 Digital Contact

3.1.1 Defining the number of clusters

In Digital Contact, the Elbow Method and the Silhouette Method were applied to determine the optimal number of clusters.

Based on the Elbow Method plot, it seems that the optimal number of clusters falls between 3 and 4. Utilizing the Silhouette Score method, knowing that a higher score indicates a well-defined clusters, we will proceed with $k = 3$.

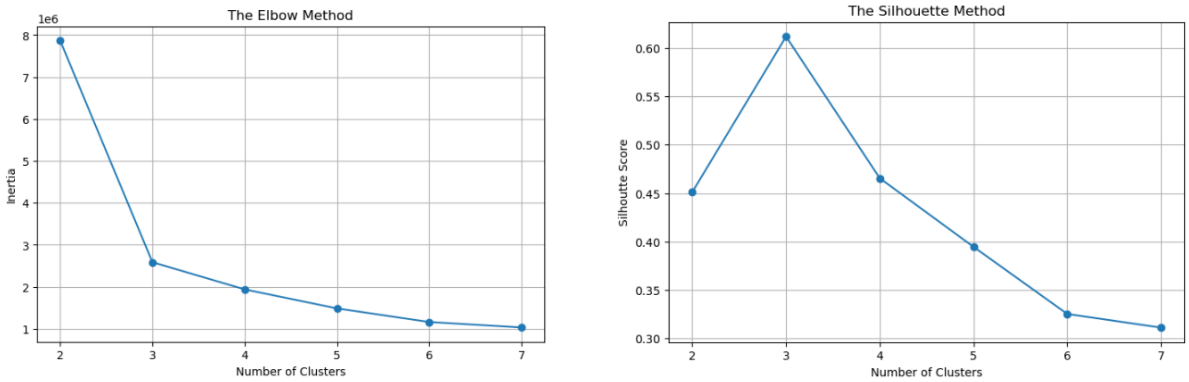


Figure 8 - Elbow Method (left) where number of clusters considered was 3 and Silhouette Method (right) where we confirm 3 clusters is the best approach.

For this cluster analysis several methods were applied: K-means, DBSCAN, K-means with PCA, and DBSCAN with t-SNE.

3.1.2 Training the model with K-Means

Applying the K-means where $k = 3$,

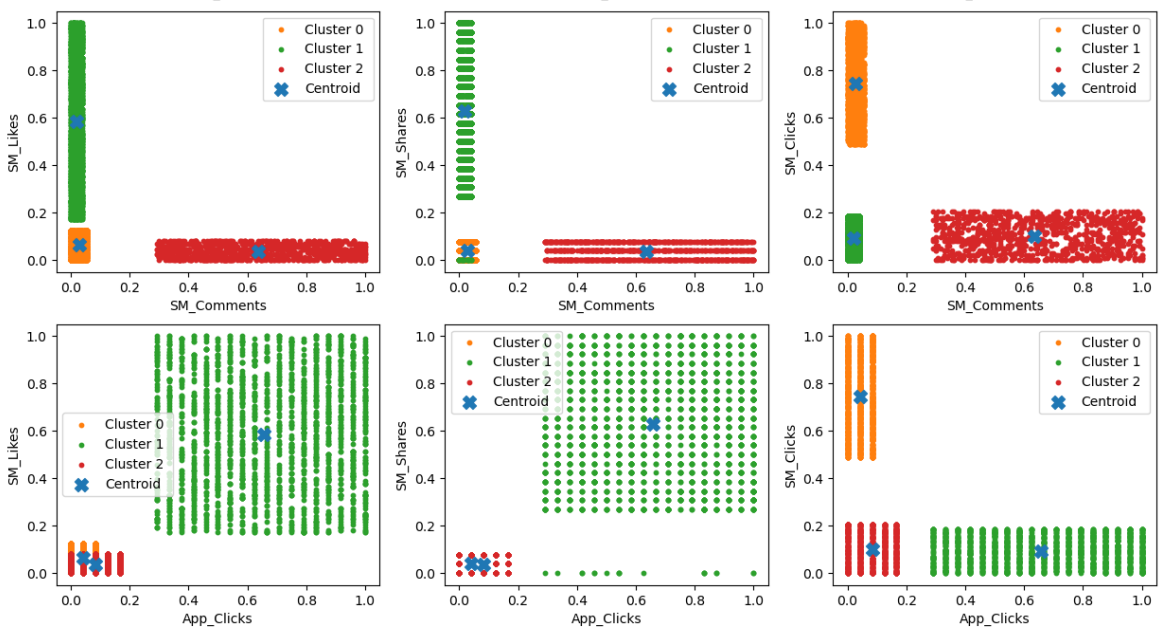
label_kmeans	0	1	2
Email_Clicks	55.628319	9.682081	18.124122
App_Clicks	3.501207	2.483973	80.827869
SM_Comments	0.998391	15.781923	2.010539
SM_Likes	5.539019	51.545455	3.339578
SM_Shares	1.003218	16.301629	0.970726
SM_Clicks	76.072405	9.521282	10.210773

Figure 9 - Average Feature Values by Cluster (K-means Labels) in Digital Contact

After defining the clusters, three groups were identified:

- **Cluster 0:**
  - The customers interacting more via Email_Clicks (55.6) and SM_Clicks (76.1)
  - The customers interacting moderately via SM_Shares (1.0), App_Clicks (3.5) and SM_Likes (5.5)
  - The customers interacting less via SM_Comments (0.99)
- **Cluster 1:**
  - The customers interacting though Social Media actions such as SM_Comments (15.8), SM_Shares (16.3) and SM_Likes (51.5)
  - The customers interacting moderately on: None
  - The customers interacting less on: App_Clicks (2.5), SM_Clicks (9.5) and Email_Clicks (9.7)
- **Cluster 2:**
  - The customers who have interacted more on: App_Clicks (80.8)
  - The customers who have interacted moderately on: SM_Comments (2.0), SM_Clicks (10.2) and Email_Clicks (18.1)
  - The customers who have interacted less on: SM_Shares (1.0) and SM_Likes (3.3).

Creating scatter plots for all SM variables in Digital Contact:



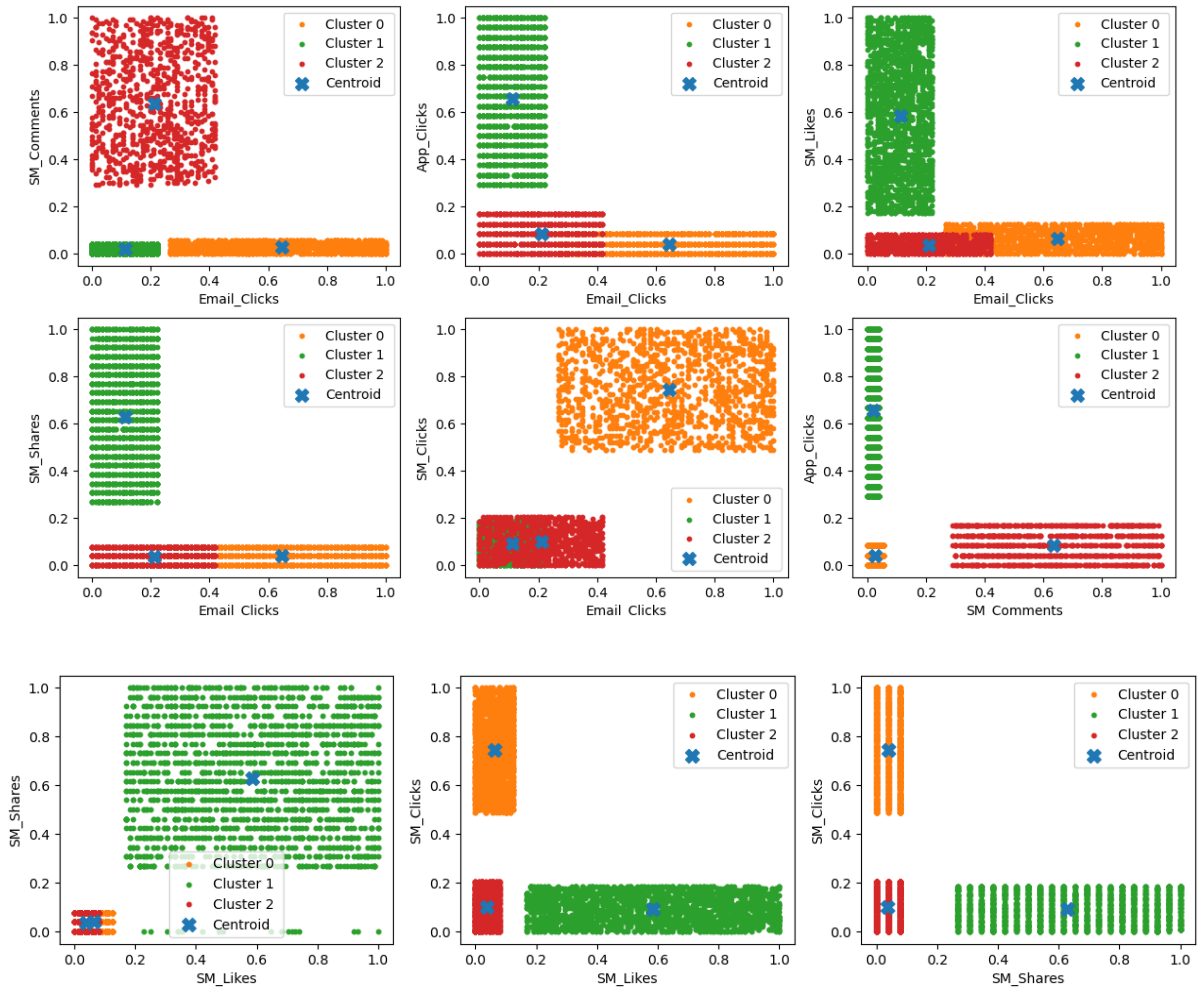


Figure 10 - Scatter plots for all variables in Digital Contact

From the **Figure 9** and **Figure 10**, we can conclude the following.

- **Cluster 0** - Comprising individuals who primarily engage in passive consumption by clicking on social media and email content, this group shows minimal activity in terms of sharing, commenting, and app usage. Referred to as **"Curious-Viewers"** they exhibit a tendency to explore content without active participation.
- **Cluster 1** - Characterized by individuals who actively engage with content by liking, sharing, and commenting frequently, this group demonstrates a strong influence on social media platforms. Despite moderate app usage, their primary focus remains on social interaction. This group, labeled as **"Influencers"** plays a significant role in shaping online discourse.
- **Cluster 2** - Encompassing individuals who predominantly utilize the app while showing limited activity on social media, this group exhibits minimal engagement in terms of sharing and commenting. Referred to as **"App users"** their behavior suggests a preference for app-centric interactions over social media engagement.

3.1.3 Applying K-means after performing PCA

3.1.3.1 Performing and fitting the PCA model to the scaled dataset.

Obtaining the proportion of the digital_df_prod_scaled variance explained by a single principal component.

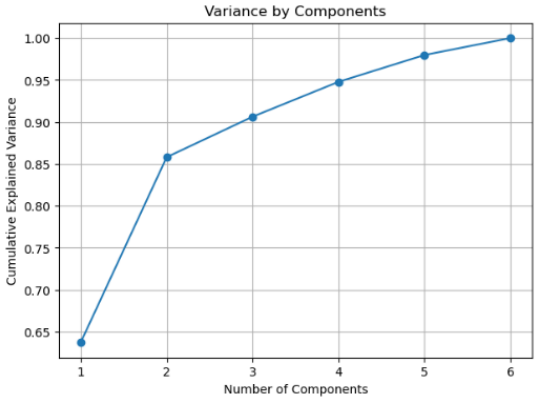
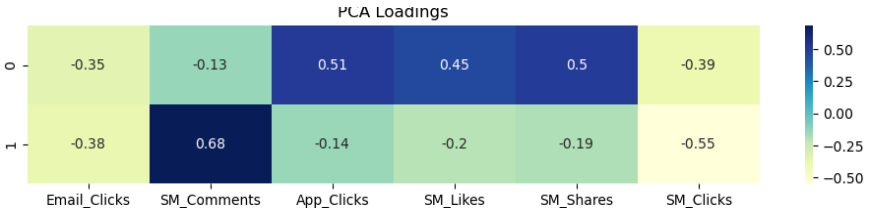


Figure 11 – Variance by Components

From the plot, we understand that the optimal number of components in PCA that involves a balance between preserving enough information from the original dataset while reducing dimensionality is **2** (Cumulative Explained Variance = 85%).

3.1.3.2 Performing PCA with dimensionality reduction and visualizing K-Means with PCA.



	Email_Clicks	SM_Comments	App_Clicks	SM_Likes	SM_Shares	SM_Clicks
0	-0.349401	-0.134724	0.507828	0.450827	0.496040	-0.390613
1	-0.377073	0.682894	-0.138325	-0.201002	-0.185393	-0.545495

Figure 12 - PCA Loadings

We can verify that:

- **Component 1 (Row 0):**
  - App_Clicks, SM_Likes and SM_Shares - The positive impact of App_Clicks, SM_Likes, and SM_Shares on Component 1 suggests a correlation between them. **Customers who frequently interact with the company's App, engage with Social Media content by liking posts and sharing content from Sportify.** These behaviours collectively reflect a comprehensive approach to digital interaction, where customers who exhibit one type of engagement are likely to demonstrate similar patterns across other channels.
- **Component 2 (Row 1):**
  - It suggests that customer engagement through comments on social media posts SM_Comments plays a crucial role in defining a particular clustering pattern. **This indicates that customers who actively engage by commenting on social media content may exhibit distinct behaviours or preferences that differentiate them from other customer segments.** On the

other hand, actions such as Email_Clicks, App_Clicks, SM_Likes, SM_Shares, and SM_Clicks appear to have less influence on this clustering pattern.

3.1.3.3 Clusters from PCA-KMeans

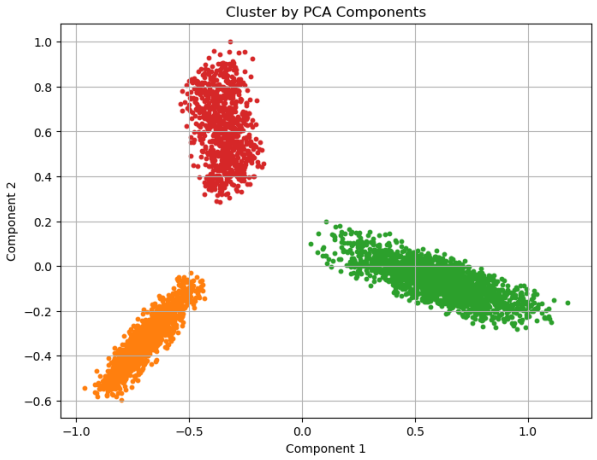


Figure 13 - Assignments between K-means and K-means with PCA.

We can verify that:

- The **Orange** clusters represent the customers who exhibit minimal engagement with email, app interactions, and social media activities.
- The **Green** clusters represent the customers who frequently interact with the company's App, engage with Social Media content by liking posts and sharing content from Sportify.
- The **Red** clusters represent the customers who actively engage by commenting on social media content may exhibit distinct behaviours or preferences that differentiate them from other customer segments.

3.1.4 Clusters with DBSCAN

Unlike K-Means, DBSCAN does not require pre-defining the number of clusters before application. However, to ensure accurate results, it is crucial to determine the optimal values for **eps** and **min_samples**.

- Min_Samples – Generally twice the data dimensionality. Our data has 6 dimensions, then 2 * 6 columns = 12;
- Epsilon (Eps) – Calculate the distance between each point and its closest neighbor using Nearest Neighbors. Then, sort the distances, plot them, and identify the maximum value at the graph's curvature, which defines our optimum Eps.

3.1.4.1 *Calculating the K-nearest neighbour (K-NN) distances for each data point in digital_df_prod_scaled dataset and then plotting the sorted K-NN distances to define Eps*

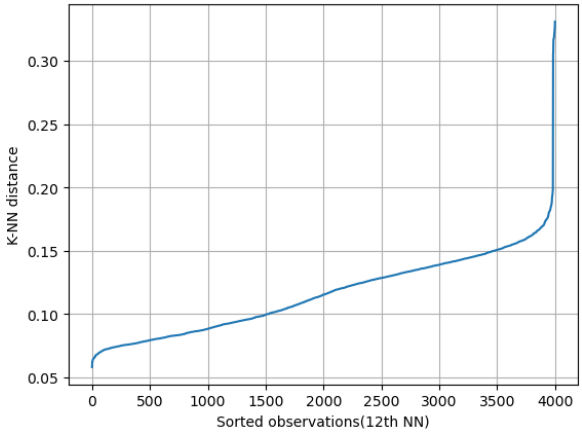


Figure 14 - Nearest Neighbors Distance Plot

The best point on a Nearest Neighbors Distance Plot is where the plot exhibits a noticeable "elbow" or abrupt change in slope. In our case, we identified that an **Eps value of 0.15** or greater represents the optimal point for DBSCAN clustering.

3.1.4.2 *Visualizing the clustering results from DBSCAN algorithm.*

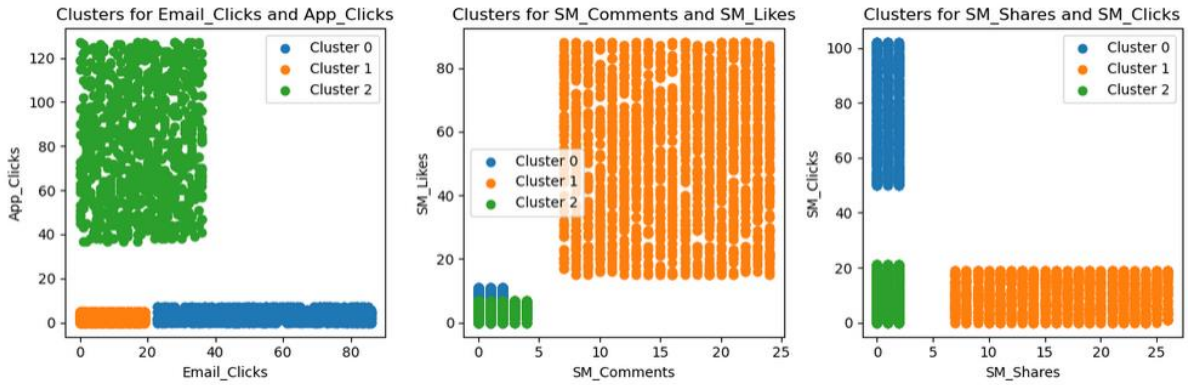


Figure 15 - Clustering results from DBSCAN.

- **Cluster 0**:
  - Represents users who engage primarily through clicks on Email and Social Media platforms (Email_Clicks and SM_Clicks), yet show minimal activity on the company's App and lack interaction such as sharing or liking posts on Social Media.

- **Cluster 1:**
  - Indicates users with the highest level of engagement in terms of Likes, Comments, and Shares on Social Media. However, they exhibit lower interest in using the company's App and are less inclined to click on ads across both Email and Social Media platforms.

- **Cluster 2:**
  - This cluster identifies users who primarily utilize the company's App but demonstrate limited interaction via Email, compared to Cluster 0. Interestingly, they show no activity on Social Media platforms, suggesting a possible lack of engagement with such channels in their daily routines.

3.1.5 Applying t-SNE

- Perplexity is a critical hyperparameter in t-SNE, regulating the number of neighbours each point considers.
- By assessing the KL (Kullback-Leibler) divergence metric across different perplexity values, we can determine the most suitable perplexity, this value is best when close to 1;
- Lower KL divergence indicates superior results, signifying better preservation of the data's structure.

Computing KL divergence for t-SNE using different perplexity values.

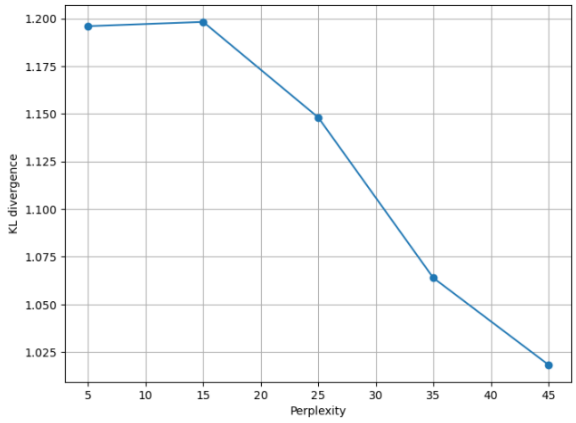


Figure 16 - KL Divergence vs. Perplexity in t-SNE Embedding

This plot helps in determining the optimal perplexity value for t-SNE embedding, where lower KL divergence indicates better alignment with the original data distribution. Upon analysis, we've selected perplexity = 50, resulting in a KL divergence of 0.9898, indicating a reasonably good fit.

Scatter plot visualization of the digital data using t-SNE:

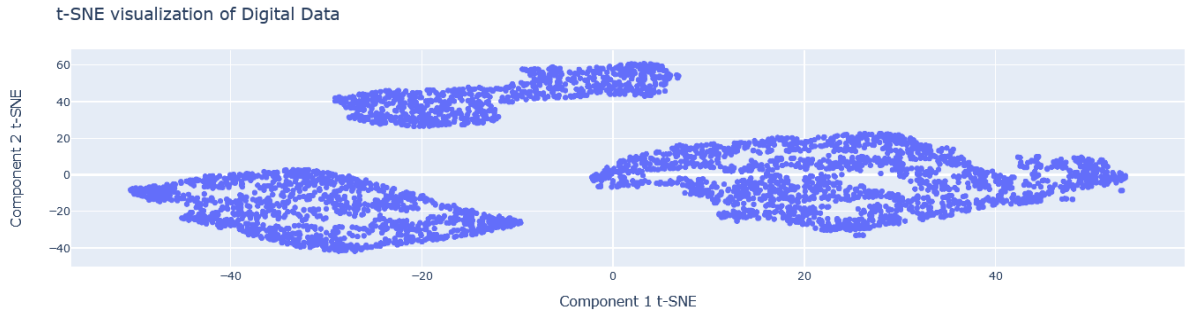


Figure 17 - t-SNE Visualization of Digital Data

The t-SNE visualization reveals more distinct clusters compared to the K-means algorithm. This suggests that t-SNE has effectively captured the underlying structure of the data and visualized it in a lower-dimensional space. The clearer separation between clusters indicates that t-SNE may be better suited for capturing complex relationships and patterns within the data.

3.1.6 Applying DBSCAN after performing t-SNE

Defining Min_Samples and Epsilon (Eps):

- Min_Samples - 2 *Data dimension = 2 * 2 Components = 4

- Epsilon:

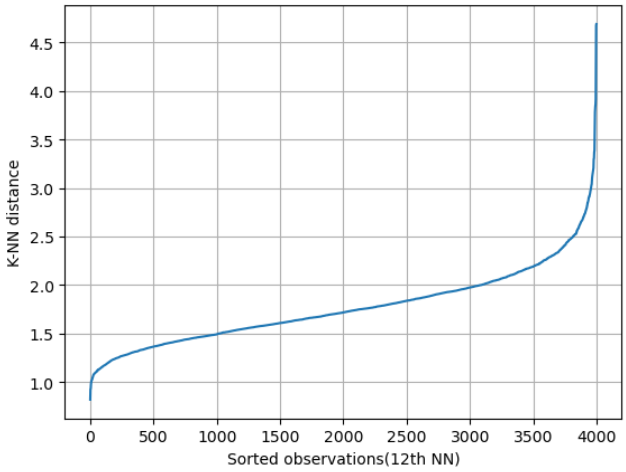
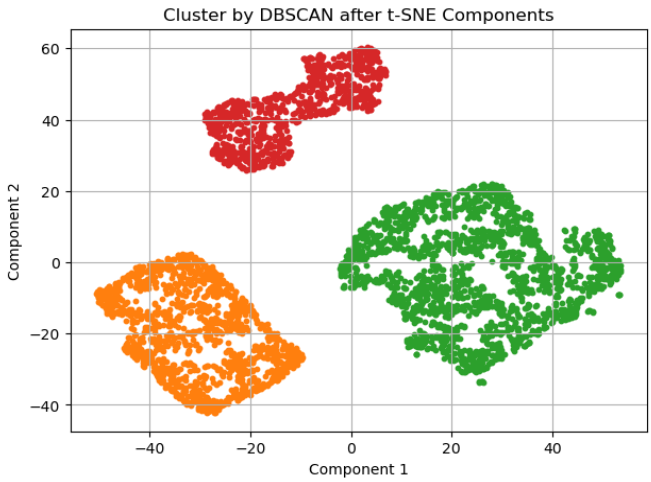


Figure 18 - Nearest Neighbors Distance Plot

From the plot we identified that an **Eps value of 2.5** or greater represents the optimal point for DBSCAN clustering.

After evaluating various Eps values and their corresponding scores, we determined that the optimal configuration is achieved when **Eps is set to 3.61 and the minimum number of samples is set to 4.**

Generating a scatter plot to visualize the clusters identified by DBSCAN after performing (t-SNE) and its clusters.



label_tsne_dbscan	0	1	2
Email_Clicks	55.628319	9.682081	18.124122
SM_Comments	0.998391	15.781923	2.010539
App_Clicks	3.501207	2.483973	80.827869
SM_Likes	5.539019	51.545455	3.339578
SM_Shares	1.003218	16.301629	0.970726
SM_Clicks	76.072405	9.521282	10.210773

Figure 19 - Cluster by DBSCAN after t-SNE Components

Based on the **Figure 19**, it appears that the t-SNE visualization produces a comparable outcome to the K-means algorithm post-PCA, with similar cluster distributions and mean values across clusters.

3.2 Products

3.2.1 Defining the number of clusters

In the same way, the Elbow Method and the Silhouette Method were applied to determine the optimal number of clusters.

Based on the Elbow Method plot, it seems that the optimal number of clusters falls between 3 and 4. Utilizing the Silhouette Score method, knowing that a higher score indicates a well-defined clusters, we will proceed with $k = 4$.

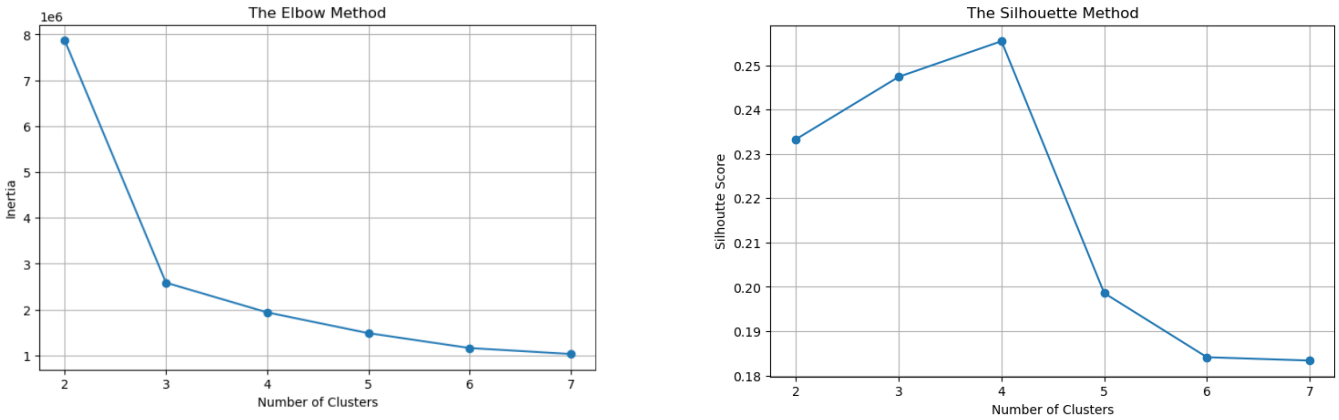


Figure 20 - Elbow Method (left) where number of clusters considered was 3 or 4 and Silhouette Method (right) where we confirm 4 clusters is the best approach.

For this cluster analysis several methods we will try to apply: K-means, K-means with PCA, and K-means with t-SNE and all have achieved the different silhouette scores.

3.2.2 Training the model with K-Means

Applying K-means where $k = 4$,

label_kmeans	0	1	2	3
Fitness&Gym	32.593827	51.300885	15.447891	33.160518
Hiking&Running	40.229012	82.231858	71.962779	53.204387
TeamGames	109.511728	234.578761	202.358561	228.421735
OutdoorActivities	33.264815	33.228319	35.259305	33.032901
TotalProducts	4.865432	6.525664	5.851117	4.991027
Days_Since_Last_P	86.738889	76.998230	75.000000	103.516451

Figure 21 - Average Feature Values by Cluster (K-means Labels) in Products

- **Cluster 0 (Low Spenders):**
    - This cluster **exhibits relatively low values across all categories compared to the other clusters.** These customers tend to focus their spending on TeamGames products, indicating a potential interest in group-oriented recreational activities. They also have fewer total products purchased and a moderate number of days since their last purchase. We believe this segment may represent customers with limited disposable income or those who prioritise other expenses over recreational activities and sports-related purchases.
- **Cluster 1 (Sport Lovers):**
    - Customers in this cluster demonstrate the highest average spending in Fitness&Gym, Hiking&Running, and TeamGames compared to the other clusters. They also have a relatively

high total number of products purchased and a moderate number of days since their last purchase. **This segment likely comprises individuals who seem to be sports enthusiasts.**

- **Cluster 2 (Outdoor Lovers):**
    - This cluster shows moderate spending across Hiking&Running and TeamGames. They demonstrate a consistent interest in outdoor pursuits. Customers in this segment also exhibit a higher total number of products purchased compared to Clusters 0 (Low Spenders) and 3 (Occasional Customers). **Their preference for outdoor experiences implies a potential affinity for nature-related adventures such as hiking, camping, or outdoor sports.**
- **Cluster 3 (Occasional Customers):**
    - Customers in this cluster exhibit moderately low spending across all categories, similar to Cluster 0, but scoring the 2nd highest on TeamGames. They have the highest number of days since their last purchase. **This segment may include occasional leisure shoppers who make sporadic purchases without a strong commitment to fitness or outdoor activities.**

Plotting K-means clusters below,

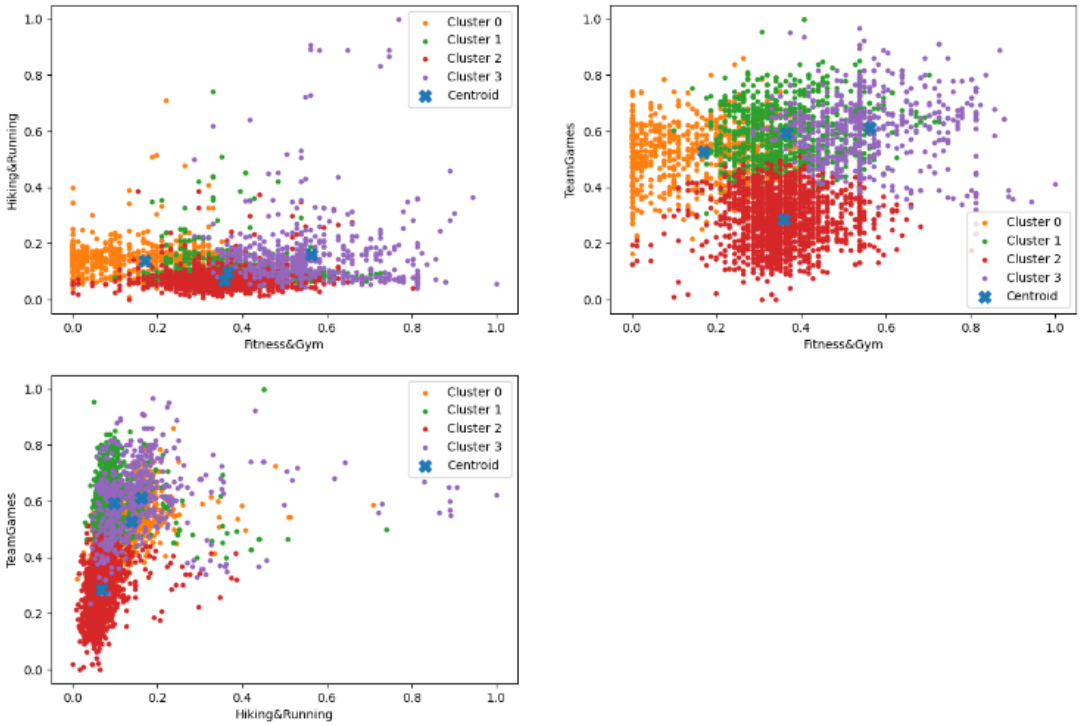


Figure 22 - Plots generated from the clustering process using the K-means algorithm.

3.2.3 Applying K-means after performing PCA

3.2.3.1 Performing and fitting the PCA model to the scaled dataset

Obtaining the proportion of the *product_df_prod_scaled* variance explained by a single principal component.

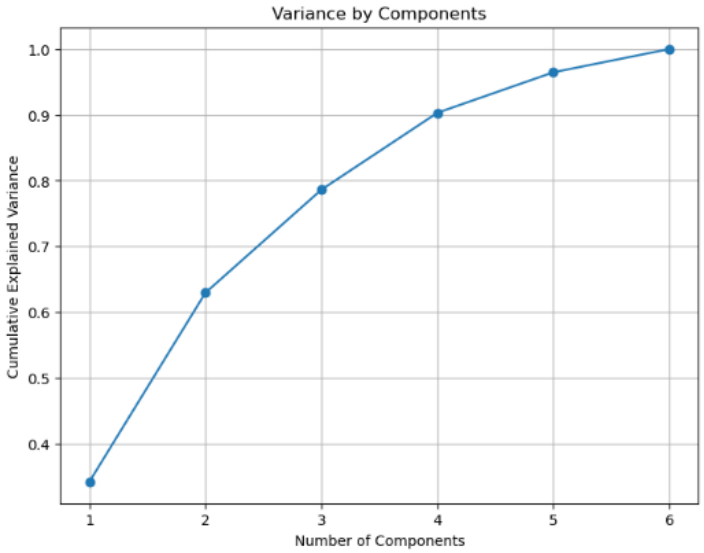


Figure 23 - Variance by Components

From the plot, we understand that the optimal number of components in PCA that involves a balance between preserving enough information from the original dataset while reducing dimensionality is **3** (Cumulative Explained Variance = 78%).

3.2.3.2 Performing PCA with dimensionality reduction and visualizing K-Means with PCA

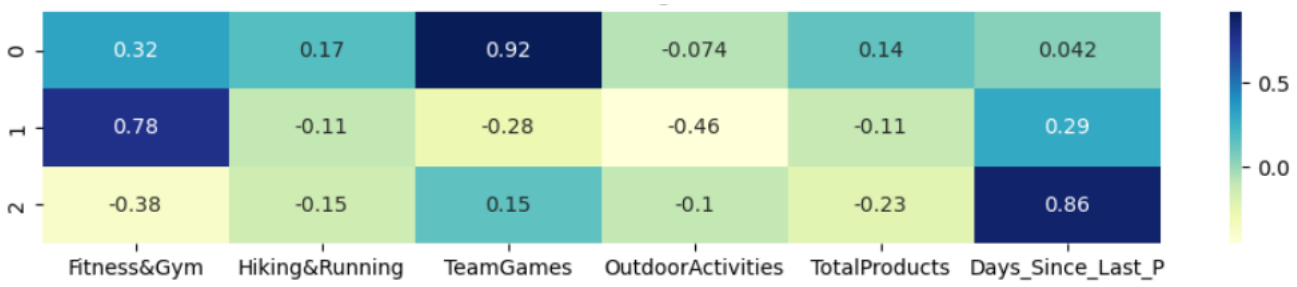


Figure 24 - PCA Loadings

We can verify that:

- **Component 1:**
  - It mainly reflects spending patterns related to **TeamGames**, with a strong positive association. It also has some association with Fitness&Gym and Hiking&Running. OutdoorActivities has a weak negative association with this component, **suggesting that customers who spend more on TeamGames tend to spend less on OutdoorActivities**.

- **Component 2:**
  - It mainly reflects spending patterns related to **Fitness&Gym**, with a strong positive association. It shows negative associations with spending on Hiking&Running, TeamGames, and OutdoorActivities, **indicating that customers who spend more on Fitness&Gym tend to spend less in these other categories**.

- **Component 3:**
  - It mainly reflects days passed since last Purchase - **Days_Since_Last_P** with a strong positive association. It also has some association with TeamGames. Plus, Fitness&Gym, Hiking&Running

and OutdoorActivities have a weak negative association with this component, **suggesting that customers who has spent more on TeamGames tend to spend less on Fitness&Gym, Hiking&Running and OutdoorActivities, but compared to component 1 they tend to wait longer to re-make Purchases**.

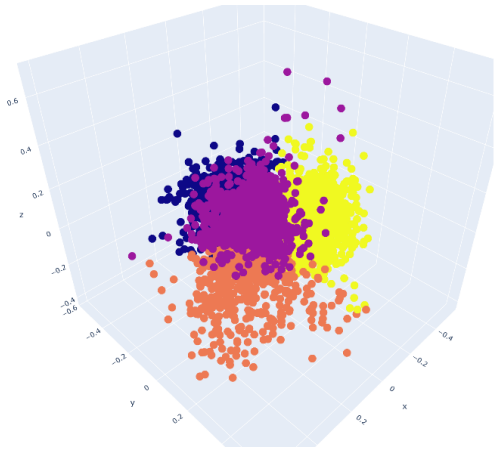


Figure 25 - 3D visualization of clusters obtained through PCA-based K-means clustering.

3.2.4 Applying DBSCAN

Defining Min_Samples and Epsilon (Eps):

- Min_Samples - 2 *Data dimension = 2 * 6 columns = 12
- Epsilon:

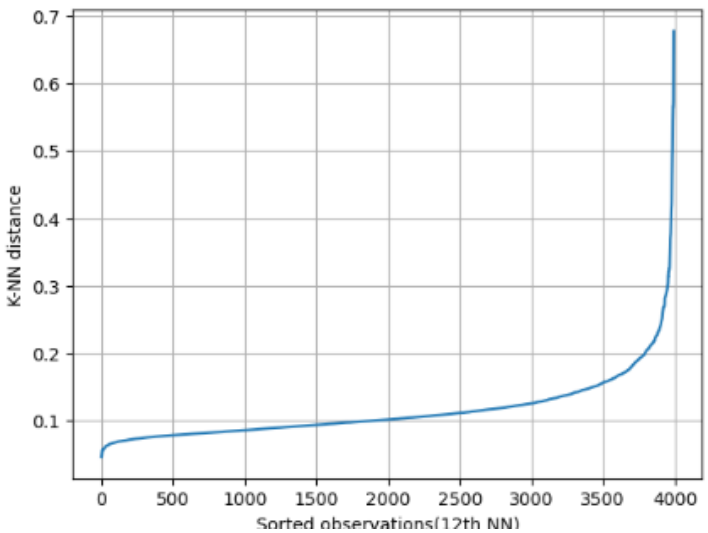


Figure 26 - Nearest Neighbors Distance Plot

From the plot, we have identified an **Eps value of 0.2** or greater represents the optimal point for DBSCAN clustering.

After completing multiple iterations for DBSCAN algorithm, it became evident that it produced only one cluster containing all data points, indicating a bad performance. To address this issue, we will first utilize t-SNE and assess its results, considering that DBSCAN may encounter difficulties with clusters of similar density.

```
label_dbscan
bad    3994
Name: count, dtype: int64
```

Figure 27 - DBSCAN Clustering Result: Single Cluster Detected

3.2.5 Applying t-SNE

Computing KL divergence for t-SNE using different perplexity values.

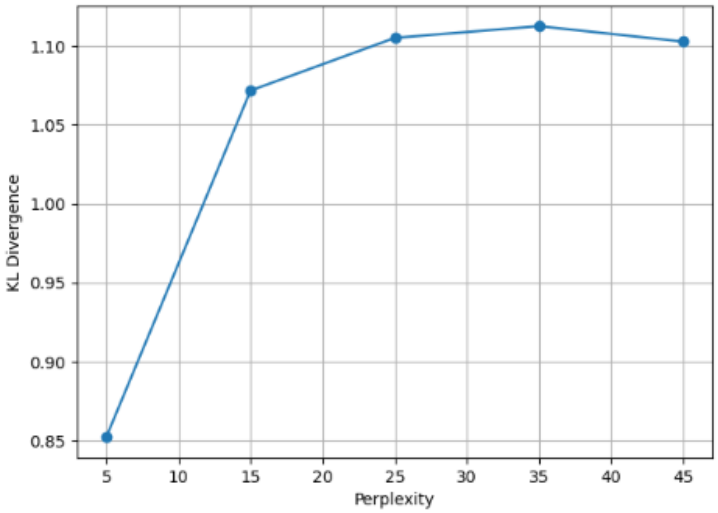


Figure 28 - KL Divergence vs. Perplexity in t-SNE.

Upon analysis from the plot, we've selected perplexity = 5, resulting in a KL divergence of 0. 8511, indicating a reasonably good fit.

Scatter plot visualization of the Products data using t-SNE:

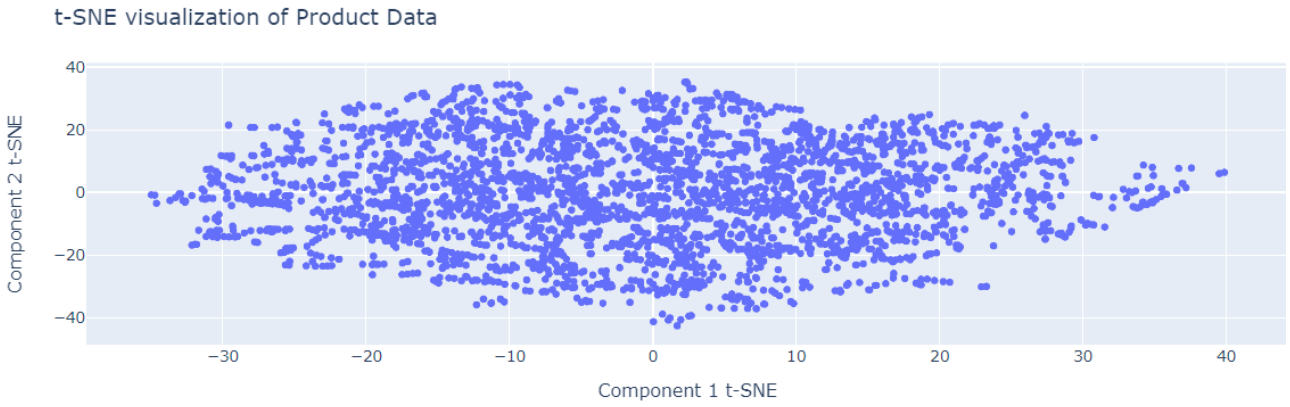
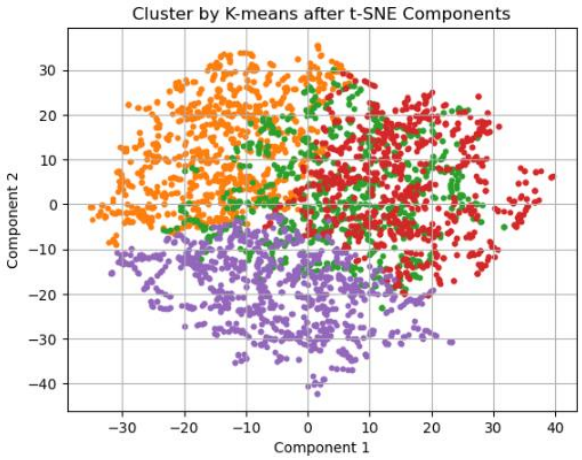


Figure 29 - t-SNE Visualization of Products

Given the unclear definition observed in the t-SNE scatter plot, further exploration and refinement of clustering techniques are warranted to better capture the underlying patterns in the data. Therefore, we will apply K-means after t-SNE to see if we can improve our clustering analysis.

3.2.6 Applying K-means after performing t-SNE

We have applied t-SNE and combined its result with K-means.



label_tsne_kmeans	0	1	2	3
Fitness&Gym	31.699630	45.181818	33.437949	18.017154
Hiking&Running	39.906654	69.516913	51.638974	64.286579
TeamGames	109.560074	197.105708	234.346667	170.191726
OutdoorActivities	32.557301	33.872093	32.980513	35.103935
TotalProducts	4.712569	6.060254	5.012308	5.622603
Days_Since_Last_P	87.177449	79.000000	103.713846	82.856710

Figure 30 – Clustering with k-means after t-SNE transformation.

Based on the **Figure 30**, we can verify:

- **Cluster 0** (Low Spenders) – These are the customers who don't buy often and tend to spend less, labelled as "Low Spenders," similar to those in K-means Cluster 0.
- **Cluster 1** (Sport Lovers) – This group comprises customers passionate about sports who make frequent purchases, labelled as "Sport Lovers," similar to K-means Cluster 1.
- **Cluster 2** (Occasional Customers) - Customers here buy sporadically, favoring TeamGames products (with the highest number), labelled as "Occasional Customers", similar to K-means Cluster 3.
- **Cluster 3** (Outdoor Lovers**)** – Customers in this cluster purchase frequently, showing a preference for outdoor activities like Hiking&Running and OutdoorActivities, and less interest in Fitness&Gym products. They're labelled as "Outdoor Lovers," similar to K-means Cluster 2.

4 Description of Resulting Clusters

4.1 Results from Digital Contact clustering

From the results, we can verify that the clustering methods—K-means, DBSCAN, K-means with PCA, and DBSCAN with t-SNE—all achieved a consistent silhouette score of around 0.612, as shown below. Hence, regardless of the specific method chosen, the resulting clusters will be similar.

Clustering method	Silhouette score
K-means	0.612
DBSCAN	0.612
K-means with pca	0.612
DBSCAN with t-SNE	0.612

Figure 31 - All clustering methods tested have given same silhouette.

Our final clusters for Digital Contact will be the ones defined in subchapter 3.1.2 (pg. 12).

4.2 Results from Products clustering

Based on the obtained silhouette scores, K-means alone (0.2554) appears to be the most suitable approach for our clustering analysis, closely followed by K-means with PCA (0.2551), indicating similar performance. Although DBSCAN exhibited the highest score, it resulted in a single cluster encompassing all observations, which makes it ineligible for our purposes due to the lack of meaningful cluster differentiation. Detailed results for each method are provided below.

Clustering Method	Silhouette Score
K-Means	0.2554
K-Means with PCA	0.2551
K-Means with T-SNE	0.1637
DBSCAN	not eligible

Figure 32 - All clustering methods tested have given different silhouettes.

Our final clusters for Products will be the ones defined in subchapter 3.2.2 (pg. 23).

4.3 Results from processing clustering

By further analysis based on digital and product information we observe the following:

Cluster_digital_label	Cluster_product_label	Count	Mean_age	Most_freq_gender	Most_freq_Education
Influencers	Occasional Customers	775	28.630968	Female	Bachelor
Curious-Viewers	Occasional Customers	506	41.069170	Male	High School
Influencers	Low Spenders	478	29.073222	Female	Less Than High School
	Outdoor Lovers	379	29.506596	Female	High School
App users	Occasional Customers	341	37.730205	Male	Bachelor
Curious-Viewers	Low Spenders	306	42.447712	Male	High School
	Outdoor Lovers	268	42.820896	Male	High School
Influencers	Sport Lovers	267	27.820225	Female	Bachelor
App users	Low Spenders	210	38.842857	Male	High School
	Outdoor Lovers	168	36.458333	Male	High School
Curious-Viewers	Sport Lovers	162	39.981481	Male	Bachelor
App users	Sport Lovers	134	37.910448	Male	Bachelor

Figure 33 - Profile of customers based in all files provided.

4.4 Our Customers Overview:

- **Influencers && Occasional Customers** - These female customers (consisting of our majority of customers) with Bachelors degree are likely individuals at their late 20's, who have the potential to influence others purchasing decisions but make purchases sporadically themselves. Despite their influence, they may not prioritize spending on products regularly.

- **Curious-Viewers && Occasional Customers** - These male customers at their 40's exhibit a curious behaviour towards products or content but do not make frequent purchases. Their engagement level suggests interest, but they may not have the financial means or inclination to buy regularly.

- **Influencers && Low Spenders** - These female customers (not holding a very high education level) have influence potential but tend to spend less on purchases. Despite their ability to influence others, they may have limited spending capacity or preferences for budget-friendly options.
- **Influencers && Outdoor Lovers** - These female customers, categorized as influencers, exhibit a preference for outdoor activities or products despite having a High School education. They may have influence potential but also have a passion for outdoor pursuits, reflecting a diverse range of interests.
- **App users && Occasional Customers** - These male customers, identified as app users, primarily engage with the Company's app but demonstrate sporadic purchasing behaviour. Despite having a Bachelor's degree, they make occasional purchases, indicating a lack of consistent spending habits.
- **Curious-Viewers && Low Spenders** - These male customers, apart from being the oldest, exhibit curiosity towards the Company but tend to spend less on purchases. With a High School education, they make occasional purchases, suggesting limited spending capacity or other priorities.
- **Curious-Viewers && Outdoor Lovers** - These male customers, being the oldest (same as the previous) and categorized as curious-viewers, exhibit a passion for outdoor activities or products despite having a High School education.
- **Influencers && Sport Lovers** - These youngest female customers, identified as influencers, exhibit influence potential and also demonstrate a passion for sports-related products or activities. With a Bachelor's degree, they actively engage in sports-related pursuits, reflecting both personal interest and influence potential.
- **App users && Low Spenders** - These male customers primarily engage with Company's app but tend to spend less on purchases. With a High School education, they demonstrate limited spending habits.
- **App users && Outdoor Lovers** - These male customers, identified as app users, exhibit a preference for outdoor activities or products despite having a High School education.
- **Curious-Viewers && Sport Lovers** - These male customers demonstrate curiosity towards products and also exhibit a passion for sports-related activities or products. With a Bachelor's degree, they actively engage in both exploring products and sports-related pursuits.
- **App users && Sport Lovers** - These male customers (consisting of our minority of customers) primarily engage with mobile apps and also exhibit a strong affinity for sports-related products or activities. With a Bachelor's degree, they actively engage in sports-related pursuits while utilizing mobile apps

Further relation between our digital clusters (from K-means) and gender.

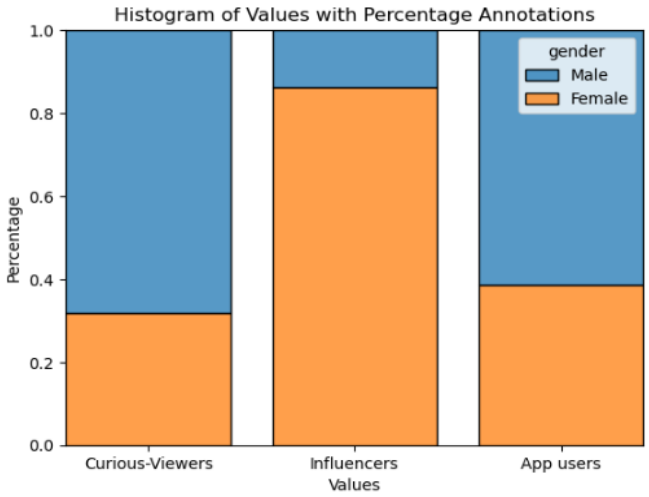


Figure 34 - Profile of customers based in all files provided.

We can observe in terms digital cluster and age:

- **Curious Viewers:** Predominantly male (68%), with an older age profile.
- **Influencers:** Mostly female (86%), characterised by a younger age group.
- **App users:** Skewed towards males (61%), with a mid-age demographic.

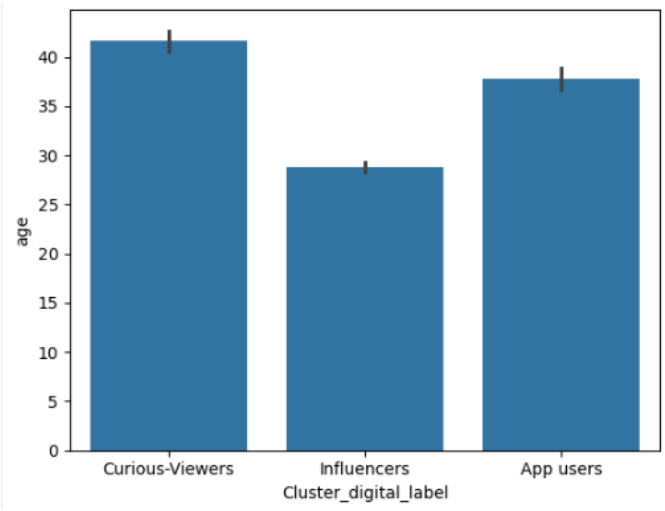


Figure 35 - Profile of customers based in all files provided.

We can check that:

- The Curious Viewers are the older;
- The Influencers are the youngers;
- The App users are the mid age.

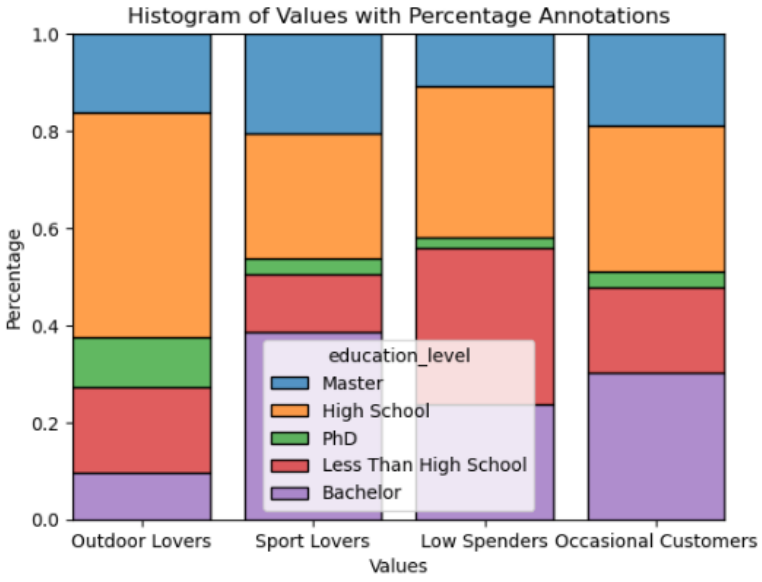


Figure 36 - Profile of customers based in all files provided.

From the plot, we can verify:

- **Outdoor Lovers** - This cluster demonstrates a higher percentage of individuals with a High School education level and PhD qualifications. High School graduates may have limited disposable income for gym memberships, leading them to prefer outdoor activities for social gatherings and leisure. Similarly, individuals with PhDs, often burdened with research commitments, may opt for outdoor pursuits like running or hiking to alleviate stress.

- **Sport Lovers** - This segment, characterized by higher spending, primarily comprises individuals with Bachelor's and Master's degrees, likely indicating higher income levels. The prevalence of Bachelor's degree holders suggests a correlation between higher education and increased spending on fitness products. Moreover, individuals with higher education levels may prioritize physical health and wellness, driving their expenditure on such products.

- **Low Spenders** - Individuals in this cluster predominantly have educational backgrounds below high school or high school diplomas, suggesting lower income levels. The lack of higher education may contribute to lower income and, consequently, reluctance to spend on fitness products.

- **Occasional Customers** - This group, consisting mainly of individuals with Bachelor's and Master's degrees, displays a preference for Team Games over individual sports activities. Despite their higher education levels, they exhibit sporadic spending behaviour, possibly due to a lower interest in sports-related products or other financial priorities.

Comparing the product label with age:

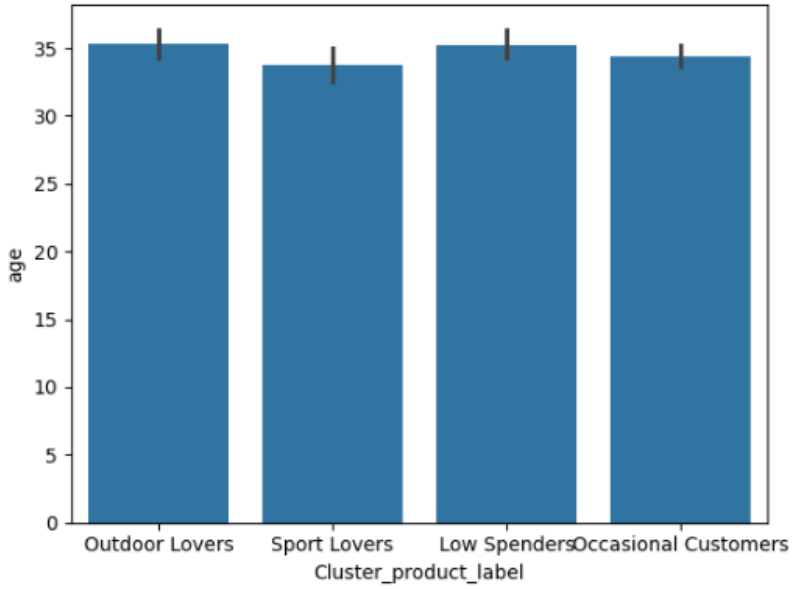


Figure 37 - Profile of product label with age.

- Sport Lovers are the youngest;
- Outdoor Lovers are the oldest;
- Occasional Customers are the 2nd youngest.

Comparing the product label with dependents:

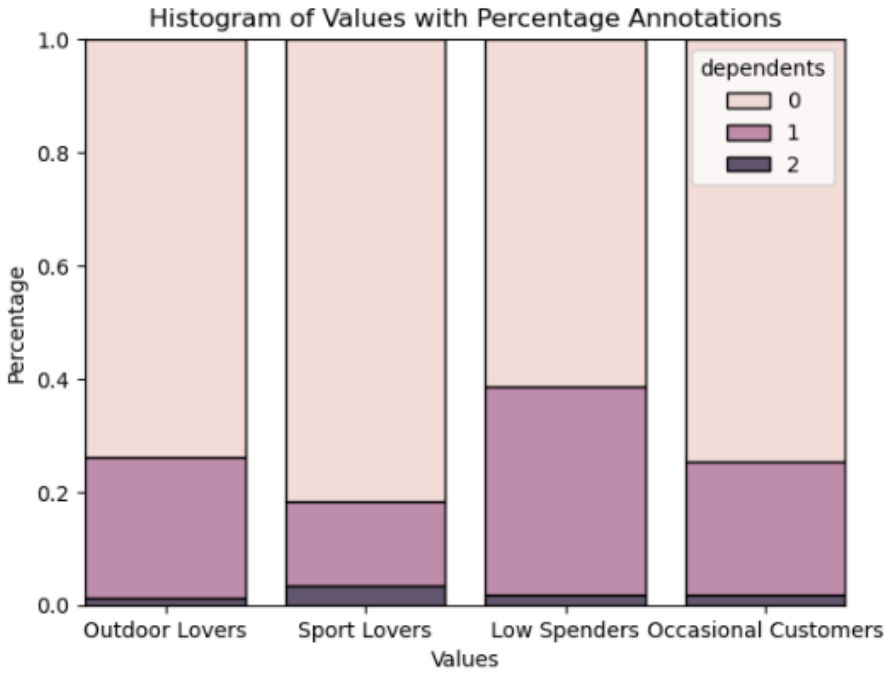


Figure 38 - Profile of product label with dependent.

- The customers with no dependents are the ones more interested in sports products, also their percentage is higher for every category. Their percentage is higher in Sport Lovers category;
- The customers with 1 dependent have a higher percentage in Low Spenders category.

5 Action Plan

Drawing upon valuable insights gained from our customer segmentation analysis, Sportify has the goal to develop a strategic action plan that effectively targets and engages diverse customer segments.

By tailoring specific initiatives to the unique characteristics and preferences of each segment, Sportify is able to optimize customer engagement and drive business growth.

Influencers (Youngest Females):

- Partnering with influencers to create trendy and engaging content that showcases our products in a fun and relatable way, leveraging their youthful energy to connect with a younger audience.
- Giving influencers early access to our new products and encouraging them to share their experiences with their followers. This creates a sense of exclusivity and drives excitement around our brand.
- Depending on their income and/or buying habits, we can foster a sense of community among low-spending influencers through interactive forums or dedicated social media groups.
- For Outdoor Lovers in this segment, organizing adventure challenges or contests that would align with this segment's outdoor interests would be ideal.
- Offering prizes or incentives for participation to drive engagement and brand loyalty.

Sport Lovers (High spenders):

- Creating a loyalty program with exclusive discounts, early access to new products, and special rewards events. This can lead them to keep purchasing and feel valued.
- Offering VIP experiences like meet-and-greets with athletes, access to training camps, or tickets to exclusive sporting events. This taps into their passion and creates a sense of exclusivity.
- We can also use their purchase history to recommend personalized training gear or other relevant items. This caters to their specific needs and increases the chances of them buying.

App Users (Highly Engaged):

- Implementing gamification elements in our app, like points, badges, and leaderboards, to motivate users and keep them coming back.
- Integrating personalized training programs into our app based on user preferences and goals. This adds value and encourages continued app usage among people in this segment.
- We could also use push notifications strategically to offer relevant workout tips, product recommendations based on in-app activity, or exclusive deals for app users. This personalizes the experience through the app engagement.

Curious-Viewers:

- We intend to include easy ways for those people to interact such as likes, polls, quizzes within the content.

- Reminding viewers of interesting content through retargeting ads on social media and email as they only make clicks on links and/ads through Social Media and Email.
- Creating targeted emails based on their interests gained from content interaction.

Low Spenders:

- Our initial aim is to promote budget-friendly product options.
- We also introduce the instalment payment options to make purchases more accessible and affordable for low spenders.
- We can also offer exclusive discounts or incentives to encourage more frequent purchases.

Outdoor Lovers:

- Highlighting outdoor-themed product collections for camping, hiking, and outdoor activities.
- Organizing outdoor challenges or contests to engage and excite outdoor enthusiasts.
- Collaborate with influencers passionate about outdoor adventures to create compelling content and endorsements.

Occasional Customers:

- We implement targeted promotions and limited-time offers to encourage more frequent purchases.
- Providing personalized recommendations based on past interactions and preferences.
- Offering loyalty rewards or incentives for repeat purchases to increase customer retention.
- Enhancing communication channels to keep occasional customers informed about new products and promotions.

6 Conclusion

The goal of this project was to create a market segmentation strategy for Sportify, a company specializing in sports products and gear. Using unsupervised Machine Learning models, we aimed to identify distinct customer clusters and propose an action plan to guide marketing efforts.

Our work began with data pre-processing and cleaning, and other actions including removing outliers and irrelevant variables, while adding new ones like age, gender, and days since the last purchase. This groundwork was crucial for ensuring accurate cluster formation and generating meaningful insights.

We have explored various clustering algorithms, discovering that there's no one-size-fits-all solution. Each algorithm has its strengths and weaknesses, depending on the type of data.

Later on, to improve visualization and better understand the clusters, we have used dimensionality reduction techniques like PCA and t-SNE.

For the “Digital Contact” dataset, the clusters appeared more distinct, with greater separation between them. However, for the “Products” dataset, the clusters were less defined and appeared closer together. We employed the silhouette score to assess the performance of our clustering algorithms, but we have found that it can

sometimes be misleading. For example, in our product clustering task, DBSCAN received the highest silhouette score, indicating strong performance. However, in reality, DBSCAN produced only one large cluster, which wasn't the desired outcome.

These observations highlighted the importance of carefully choosing the right algorithm and metric for each specific dataset and understanding their limitations.

Ultimately, we identified 12 unique customer groups with varying characteristics.

1. **Influencers && Occasional Customers (Female, Late 20s, Bachelor's Degree):** These customers have influence potential but make sporadic purchases themselves, despite their ability to influence others' buying decisions.
2. **Curious-Viewers && Occasional Customers (Male, 40s):** They show interest in products or content but do not buy frequently due to financial constraints or other priorities.
3. **Influencers && Low Spenders (Female, Lower Education):** Despite influence potential, they tend to spend less, possibly due to limited financial capacity.
4. **Influencers && Outdoor Lovers (Female, High School Education):** These customers have influence potential and a passion for outdoor activities.
5. **App Users && Occasional Customers (Male, Bachelor's Degree):** They primarily engage with the app but make occasional purchases, indicating inconsistent spending habits.
6. **Curious-Viewers && Low Spenders (Male, Oldest Age Group, High School Education):** They show curiosity but spend less, possibly due to limited finances or other priorities.
7. **Curious-Viewers && Outdoor Lovers (Male, Oldest Age Group, High School Education):** They have a passion for outdoor activities
8. **Influencers && Sport Lovers (Female, Youngest Age Group, Bachelor's Degree):** These customers have influence potential and spend a lot on sports-related products or activities.
9. **App Users && Low Spenders (Male, High School Education):** They primarily engage with the app but demonstrate limited spending habits.
10. **App Users && Outdoor Lovers (Male, High School Education):** They prefer outdoor activities or products and engage with the app.
11. **Curious-Viewers && Sport Lovers (Male, Bachelor's Degree):** They are curious about products and have a passion for sports-related activities.
12. **App Users && Sport Lovers (Male, Bachelor's Degree):** This minority group primarily engages with mobile apps and they consume a lot on sports-related products.

To maximize impact, our recommended strategy focuses on engaging more with these key segments: "App Users", "Influencers" and "Sports Lovers".

By concentrating on these key groups and acknowledging that different clustering algorithms may suit different datasets, we can increase revenue, improve customer engagement, and build a stronger brand presence through targeted marketing and partnerships with influencers.

References

Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow. O'Reilly Media.

Larose, Daniel T., Larose, Chantal D. (2015). Data Mining and Predictive Analytics. 2nd Edition. WILEY.

Müller, A. C., & Guido, S. (2017). Introduction to Machine Learning with Python. O'Reilly Media.

scikit-learn.org. (n.d.). sklearn.impute.KNNImputer. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.impute.KNNImputer.html

scikit-learn.org. (n.d.). sklearn.metrics.silhouette_score. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html#:~:text=The%20Silhouette%20Coefficient%20is%20calculated,is%20not%20a%20part%20of

scikit-learn.org. (n.d.). sklearn.cluster.DBSCAN. Retrieved from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

7 Annexes

7.1 KNN Imputer

The KNN Imputer, or k-Nearest Neighbors Imputer, is a machine learning technique used to fill in missing values in a dataset based on the values of similar data points (neighbors).

KNN Imputer estimates missing values by identifying the k nearest neighbors (similar data points) for each sample with missing values. It calculates distances between the sample with missing values and its neighbors based on available feature values. Missing values are then imputed (filled in) by taking a weighted average of corresponding values from the nearest neighbors. A very simple example of use of K-NN Imputer would be as follows.

```
>>> import numpy as np
>>> from sklearn.impute import KNNImputer
>>> X = [[1, 2, np.nan], [3, 4, 3], [np.nan, 6, 5], [8, 8, 7]]
>>> imputer = KNNImputer(n_neighbors=2)
>>> imputer.fit_transform(X)
array([[1. , 2. , 4. ],
       [3. , 4. , 3. ],
       [5.5, 6. , 5. ],
       [8. , 8. , 7. ]])
```

Overall, the KNN Imputer is a powerful tool for handling missing data in machine learning and data preprocessing tasks, offering flexibility and effectiveness in a wide range of scenarios.

7.2 Silhouette Method

The silhouette method is a technique to evaluate the quality of clustering results. It measures how similar an object is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a high score indicates that the object is well-matched to its cluster and poorly matched to neighboring clusters, while a low or negative score suggests possible misclassification.

Steps to implement it:

1. Create Clusters: Use a clustering algorithm to group the data.
2. Calculate Silhouette Scores: Compute the silhouette score for each data point.

After that we can visualize the Results, analyze the silhouette scores to assess cluster quality.

A very handy example of the application of Silhouette Score on K-Means would be as follows:

```
>>> from sklearn.datasets import make_blobs
>>> from sklearn.cluster import KMeans
>>> from sklearn.metrics import silhouette_score
>>> X, y = make_blobs(random_state=42)
>>> kmeans = KMeans(n_clusters=2, random_state=42)
>>> silhouette_score(X, kmeans.fit_predict(X))
0.49...
```

Overall, the Silhouette Method provides a valuable tool for understanding the effectiveness of your clustering and helps you refine your approach to achieve optimal cluster separation.

7.3 PCA

Principal Component Analysis (PCA) is a dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional form while preserving the most important information. It does this by finding the directions (principal components) that maximize the variance in the data.

Steps to implement PCA:

1. Data Scaling: PCA relies on data variance, so it's crucial to scale the features to ensure they contribute equally.
2. Fit PCA: Determine the principal components based on the data.
3. Transform Data: Project the data onto the principal components to reduce its dimensionality.

After that we can examine the explained variance to understand how much information is retained in the lower-dimensional data.

Here's how you can implement PCA in Python using the scikit-learn library.

```
>>> import numpy as np
>>> from sklearn.decomposition import PCA
>>> X = np.array([[-1, -1], [-2, -1], [-3, -2], [1, 1], [2, 1], [3, 2]])
>>> pca = PCA(n_components=2)
>>> pca.fit(X)
PCA(n_components=2)
>>> print(pca.explained_variance_ratio_)
[0.9924... 0.0075...]
>>> print(pca.singular_values_)
[6.30061... 0.54980...]
```

Overall, PCA is a powerful tool for simplifying complex data, making it easier to visualize, analyze, and use in various applications.

7.4 t-SNE

t-SNE (t-Distributed Stochastic Neighbor Embedding) is a powerful non-linear dimensionality reduction technique commonly used for data visualization. It excels at transforming high-dimensional data into lower dimensions (typically 2D or 3D) while preserving local relationships and complex structures. Unlike linear methods like PCA, t-SNE is effective for revealing intricate patterns and clusters within complex datasets.

Steps to Implement t-SNE:

- Setting t-SNE Parameters: Determine the number of components (n_components) and select an appropriate perplexity value.
- Fitting and Transforming Data: Apply t-SNE to transform the high-dimensional data into a lower-dimensional representation.

How to choose the optimum perplexity?

Choosing the right perplexity for t-SNE is crucial for visualizing your data effectively. Perplexity determines how t-SNE balances local and global structures within your dataset, influencing the way relationships among data points are modeled and how clusters are identified.

Perplexity can be seen as a measure of data complexity, controlling the level of detail t-SNE considers when defining the "neighborhood" of each data point. Higher perplexity values prioritize distant neighbors, capturing broader patterns, while lower values focus more on local relationships. Common perplexity values range from 5 to 50, which generally work well across different datasets.

To find the optimal perplexity, experiment with different values to minimize KL divergence effectively. After applying t-SNE with various perplexities, visually inspect the results to identify a balance of well-separated clusters and meaningful patterns, indicating successful dimensionality reduction with low KL divergence.

Here we presented an example of implementing t-SNE:

```
>>> import numpy as np
>>> from sklearn.manifold import TSNE
>>> X = np.array([[0, 0, 0], [0, 1, 1], [1, 0, 1], [1, 1, 1]])
>>> X_embedded = TSNE(n_components=2, learning_rate='auto',
...                   init='random', perplexity=3).fit_transform(X)
>>> X_embedded.shape
(4, 2)
```

Example 2: **Figure 25** could be a very understanding example of use of t-SNE.

Overall, t-SNE is a powerful tool for visualizing the hidden structures within high-dimensional data, offering valuable insights during exploratory data analysis.

7.5 DBSCAN

DBSCAN is a popular clustering algorithm that identifies clusters based on density. It is useful for finding clusters of arbitrary shape, identifying outliers, and handling noisy data. DBSCAN groups data points into clusters based on density. It identifies "core points" (high-density areas), "border points" (on the edge of high-density areas), and "noise points" (isolated from clusters). The algorithm does this by using two main parameters: eps (the maximum distance for considering points as neighbors) and min_samples (the minimum number of points required to form a cluster).

Here we presented an example of implementing DBSCAN:

```
>>> from sklearn.cluster import DBSCAN
>>> import numpy as np
>>> X = np.array([[1, 2], [2, 2], [2, 3],
...               [8, 7], [8, 8], [25, 80]])
>>> clustering = DBSCAN(eps=3, min_samples=2).fit(X)
>>> clustering.labels_
array([ 0,  0,  0,  1,  1, -1])
>>> clustering
DBSCAN(eps=3, min_samples=2)
```

Overall, DBSCAN is a powerful tool for data exploration and clustering tasks, particularly when dealing with datasets with varying densities, non-spherical clusters, and potential noise points.