NOVA IMS Information Management School

Applied Multivariate Data Analysis

Assignment II – Social Media Analysis

Academic Year 2021/2022

1st Semester

Cátia Celestino – M20211398

Joana Ramalho – M20210826

Paula Eveling – M20210331

Professors:

Leonor Bacelar Nicolau

Paulo Gomes

TABLE OF CONTENTS

1. ABSTRACT

Over the past decade, the world has experienced a rapid increase in the use of social media platforms. For businesses, these platforms can be used to engage with customers, find out what people are interested in, and promote campaigns. In sum, social media platforms can be used to attract customers and build loyalty.

With that said, the main objective of this paper is to analyse the performance of two supermarket brands, Auchan or Pingo Doce, on their respective Facebook pages between January 2019 and July 2020. The performance was analysed considering the following key performance indicators: average posts, average reactions, comments, and shares, and finally the percentual of engagement.

In order to compare their results, a Principal Component Analysis was applied, followed by a Hierarchical Clustering and K-Means techniques using the JMP software. The results suggest that the brands have different approaches to handling social media campaigns, where Pingo Doce presents a better performance.

2. KEYWORDS

Principal Component Analysis, Cluster Analysis, K-Means, Hierarchical Clustering, Data Analysis.

3. INTRODUCTION

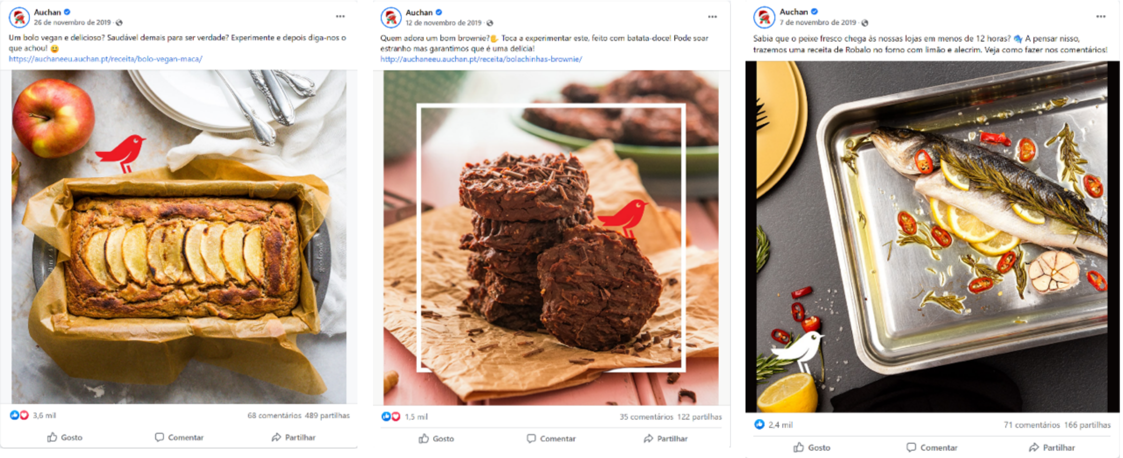
Contextualization

Over the past decade, the world has experienced an explosion in the use of social media platforms. In terms of communication, this explosion not only impacts the way we liaise with friends and family, but also how brands communicate with us. Many retailers and brands are taking advantage of the power social media holds to create additional content to aid a consumer's buying decision (Allen, Ben. 2019). For instance, companies use their platform to promote discounts, post recipe ideas for customer to try, among other strategies.

With the major impact social media has over businesses nowadays, many retailers have strong social media campaigns. The result of each campaign must be thoroughly analysed in order to allow the company to make critical data-driven decisions. In this study, the Facebook performance of two supermarket brands were analysed: Auchan and Pingo Doce.

Auchan

Auchan is one of the world's largest retailers headquartered in Croix, France, with a direct presence in Portugal, among other countries. The figures below exemplify posts from Auchan in November 2019, where its performance was higher in terms of engagement.



Auchan
26 de novembro de 2019 ·

Um bolo vegan e delicioso? Saudável demais para ser verdade? Experimente e depois diga-nos o que achou! 🙂
https://auchaneeu.auchan.pt/receita/bolo-vegan-maca/

Auchan
12 de novembro de 2019 ·

Quem adora um bom brownie?✋ Toca a experimentar este, feito com batata-doce! Pode soar estranho mas garantimos que é uma delícia!
http://auchaneeu.auchan.pt/receita/bolachinhas-brownie/

Auchan
7 de novembro de 2019 ·

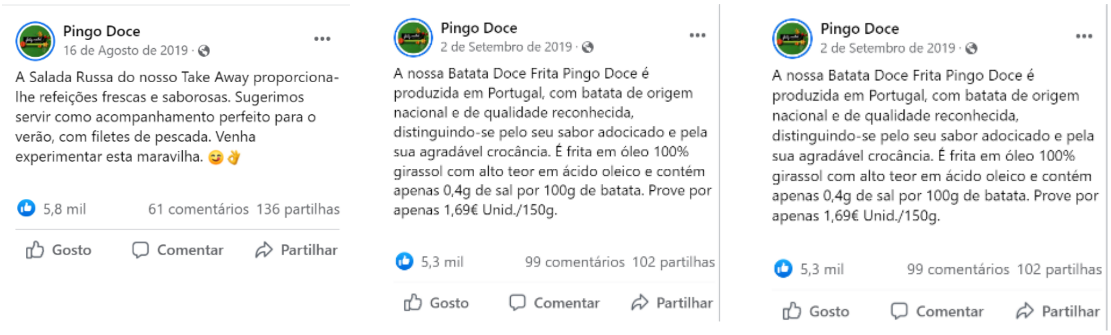
Sabia que o peixe fresco chega às nossas lojas em menos de 12 horas? 🐟 A pensar nisso, trazemos uma receita de Robalo no forno com limão e alecrim. Veja como fazer nos comentários!

Figure 1: Auchan's Facebook Posts

Pingo Doce

Pingo Doce is one of the largest supermarket operators in Portugal with almost 400 stores. The figures below exemplify posts from Pingo Doce in August 2019, where its performance was higher in terms of engagement.

NOVIDADE
PRODUÇÃO BIOLÓGICA,
AO AR LIVRE OU SOLO
OVOS
pingo doce



Pingo Doce
16 de Agosto de 2019 ·

A Salada Russa do nosso Take Away proporciona-lhe refeições frescas e saborosas. Sugerimos servir como acompanhamento perfeito para o verão, com filetes de pescada. Venha experimentar esta maravilha. 😋👌

5,8 mil 61 comentários 136 partilhas

Pingo Doce
2 de Setembro de 2019 ·

A nossa Batata Doce Frita Pingo Doce é produzida em Portugal, com batata de origem nacional e de qualidade reconhecida, distinguindo-se pelo seu sabor adocicado e pela sua agradável crocância. É frita em óleo 100% girassol com alto teor em ácido oleico e contém apenas 0,4g de sal por 100g de batata. Prove por apenas 1,69€ Unid./150g.

5,3 mil 99 comentários 102 partilhas

Pingo Doce
2 de Setembro de 2019 ·

A nossa Batata Doce Frita Pingo Doce é produzida em Portugal, com batata de origem nacional e de qualidade reconhecida, distinguindo-se pelo seu sabor adocicado e pela sua agradável crocância. É frita em óleo 100% girassol com alto teor em ácido oleico e contém apenas 0,4g de sal por 100g de batata. Prove por apenas 1,69€ Unid./150g.

5,3 mil 99 comentários 102 partilhas

Figure 2: Pingo Doce's Facebook Posts

It can be observed that both brands had a higher engagement when posting healthy food related content. However, the data must be analyzed in depth using the proper tools.

Objectives

The main objective of this study is to evaluate the performance of two Facebook Pages; Auchan and Pingo Doce; in terms of engagement, considering the following key performance indicators: number of posts, percentage of engagement, and average reactions, comments, and shares per month. This analysis was carried out in JMP, a statistical software designed for dynamic data visualisation and analytics.

Specific Objectives

The specific objectives of this study are:

- Apply a Principal Component Analysis technique to increase interpretability of the data while minimizing information loss.
- Apply two clustering methods: Hierarchical and K-Means, to segregate groups with similar traits and assign them into clusters.
- Interpret results of both Principal Component Analysis and Clustering to determine internal structure of the data to evaluate the performance of each Facebook Page: Auchan and Pingo Doce.

Limitations

The main limitation faced in this project is that there is no detailed information about the strategies adopted by each brand on a given period, which could clarify the presence of outliers and/or explain an increase or drop in performance so they can be dealt with properly. All the techniques applied in this study are sensitive to outliers, which could impact the results.

In addition, this study does not take into consideration if the engagement is positive or negative for each brand. For example, the increase in comments might be a reflect of a poor customer service, or a call-out as a form of ostracism.

4. METHODOLOGY

The data available for this analysis was collected and provided by the stakeholders in November 2021. The dataset contains 38 observations from Auchan and Pingo Doce's key performance indicators of posts shared between January 2019 and July 2020, where each observation concerns the period over the previous thirty days from the registered date. In order to analyse the performance of each brand, the following techniques were applied:

Methodology:

1. Principal Component Analysis
2. Hierarchical Clustering
3. K-Means
4. Comparison between Hierarchical Clustering and K-Means

Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure that can be considered as a projection method to conserve the maximum amount of information from the initial dimensions. This technique is usually used for dimensionality reduction, but it can also be used to visualize data, observe trends, clusters and outliers (Tam, Adrian, 2021). This method helps to understand the correlation between the variables and observations and if all variables are indeed relevant for the analysis.

Hierarchical Clustering

Hierarchical clustering is a method that groups similar observations into groups called clusters and is commonly used in many fields, such image, text and social media network analysis. One important characteristic of this method is that it does not require the number of clusters in advance, which can be helpful when the business knowledge is limited, such as in this case, where the researchers have no access to the brands to clarify abnormal behaviour in the data.

K-Means

K-means aims to hard-partition n observations into k clusters, where k is a number defined by the researcher. It groups similar results together, as per the example on figure 3. This method requires domain knowledge to define the number of clusters. This is the reason why this technique was the last applied in this study. Both principal component analysis and hierarchical clustering can help to gain insights on the dataset patterns and correlations.

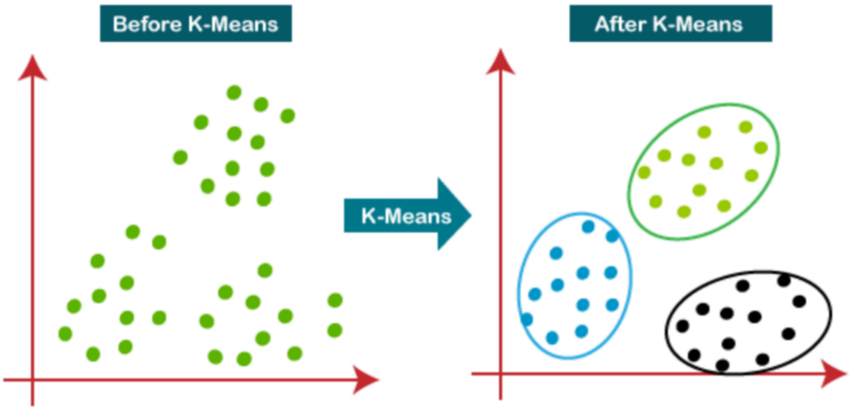


Figure 3: K-Means Example

5. DATA AND METHODS

5.1. Data Description

The dataset provided by the stakeholders contains eleven key performance indicators (KPIs) regarding two Facebook pages: Auchan and Pingo Doce. The data was collected monthly between January 2019 and July 2020, where each observation concerns the period over the previous thirty days from the registered date. The description of each variable can be found on table 1. There are no missing values in this dataset.

Variable	Description
Cod_id	id of the observation
Date	registered date
Page	name of the brand (Auchan or Pingo Doce)
Page_1	id of the Facebook page

Posts per day (average)	number of posts shared per day;
Posts interaction (%)	any action the user performs in relation to the content;
Engagement (%)	action someone takes on each post;
RCS per post (average)	number of reactions, comments and shares per post;
Reactions per post (average)	6 emojis customers can use to respond to a post;
Comments per post (average)	when a comment is made on a piece of content;
Shares per post (average)	total number of times a post was shared;
Likes per post (average)	a type of reaction someone can take on a post;
Reactions - % of RCS	reactions divided by reactions + comments + shares;
Comments - % of RCS	comments divided by reactions + comments + shares;
Shares - % of RCS:	shares divided by reactions + comments + shares;

Table 1: Variables Description

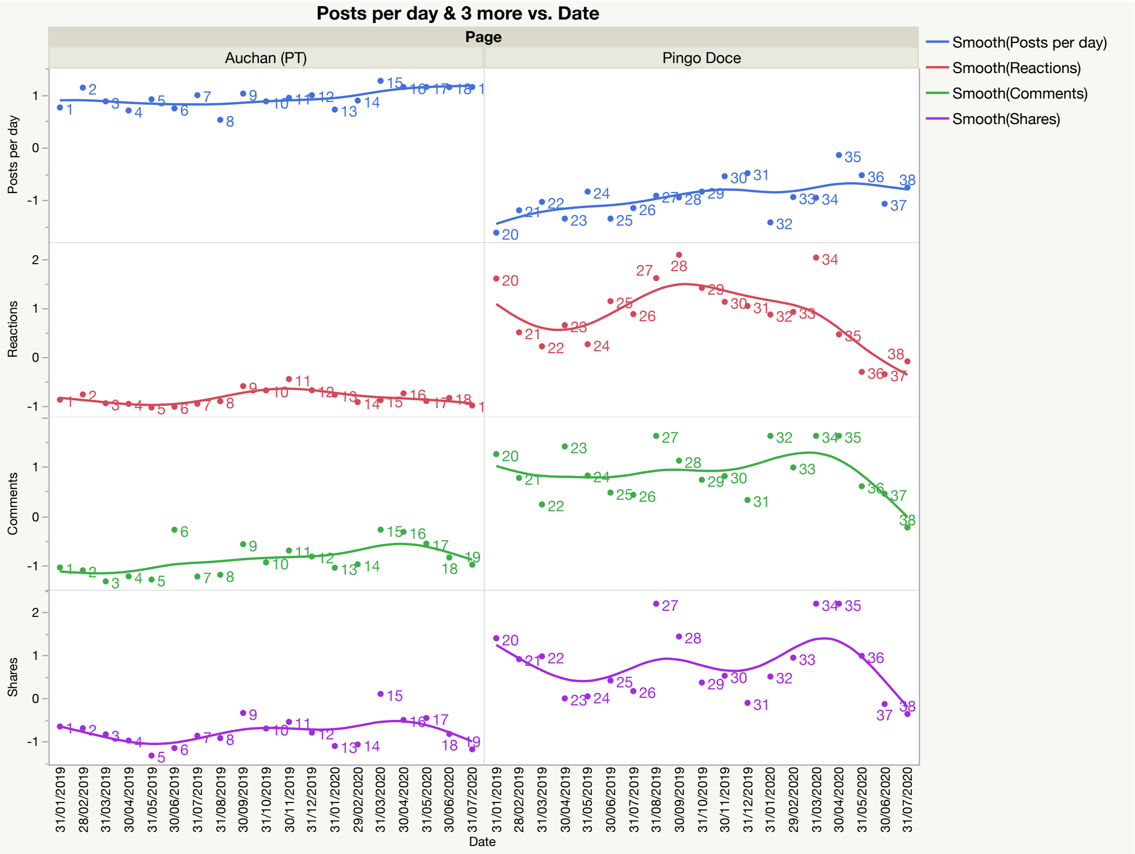


Figure 4 Active Variables Comparison by Brand

The chart above is very helpful to understand the underlying strategy adopted by each brand. It can be observed that Auchan has posted more than Pingo Doce, however, Pingo Doce has had a higher engagement, where the average number of reactions, comments, and shares per post are higher.

5.2 Univariate Exploratory Statistical Analysis

The first step to perform a multivariate analysis is to define active and supplementary variables. Active variables are the ones used to compute the principal component analysis. For the purpose of this project, described on chapter 3, the following variables were used as active: posts per day (average), engagement (%), reactions per post (average), comments per post (average), and shares per post (average).

Supplementary variables, on the other hand, have no influence on the principal components of the analysis but can be taken into account in order to enrich the analysis. They are going to help to interpret the dimensions of variability. In this study, the following variables were used as supplementary: RCS per post, reactions - % of RCS, comments - % of RCS, and shares - % of RCS.

The only KPI that will not be used neither as an active nor a supplementary variable is the Likes per post (average). This variable does not fit into the analysis because the reactions variables already include the number of likes.

The statistical analysis of each active variable can be observed on the table below:

Variable	Minimum	Maximum	Mean	Std Dev
Posts per day (average)	0.6129032	4.1333333	2.008655	0.9343529
Engagement (%)	0.0673897	1.0887783	0.5883565	0.2409053
Reactions per post (average)	266.52381	3248.6286	1256.9768	951.17473
Comments per post (average)	4.8795181	171,63889	39.146438	37.924667
Shares per post (average)	18.083333	533.47222	92.512924	90.774657

Table 2: Statistical Description of the Dataset

The statistical description of the dataset (table 2) indicates that, on average, Auchan and Pingo Doce together share two posts per day, where each post has 1,257 reactions, 39 comments, and 92 shares. However, this information is not helpful when the standard deviation is taken into consideration, as they are nearly the same as their respective means, which is a strong indicator of outliers.

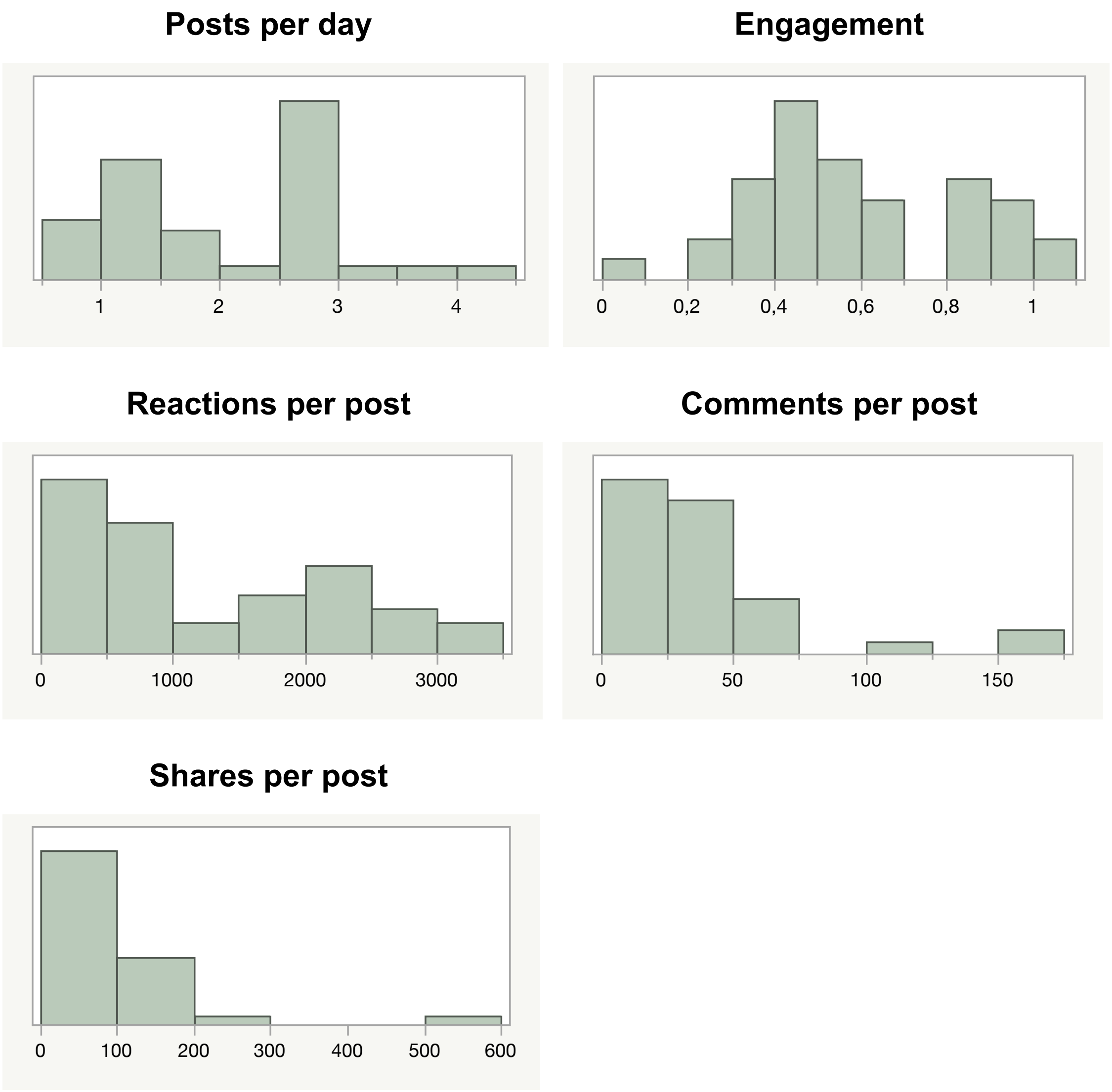


Figure 4: Distributions Charts

The shape of the distributions (figure 4) can be described as non-symmetric. The plots comments and shares per post represent a distribution with highly probable outliers.

5.3 Outlier Analysis

An outlier is a piece of data that is an abnormal distance from other points, in other words, they are unusual values in a dataset. Most of the algorithms are sensitive to outliers, such as the methods applied in this study: hierarchical clustering, k-means, and principal component analysis.

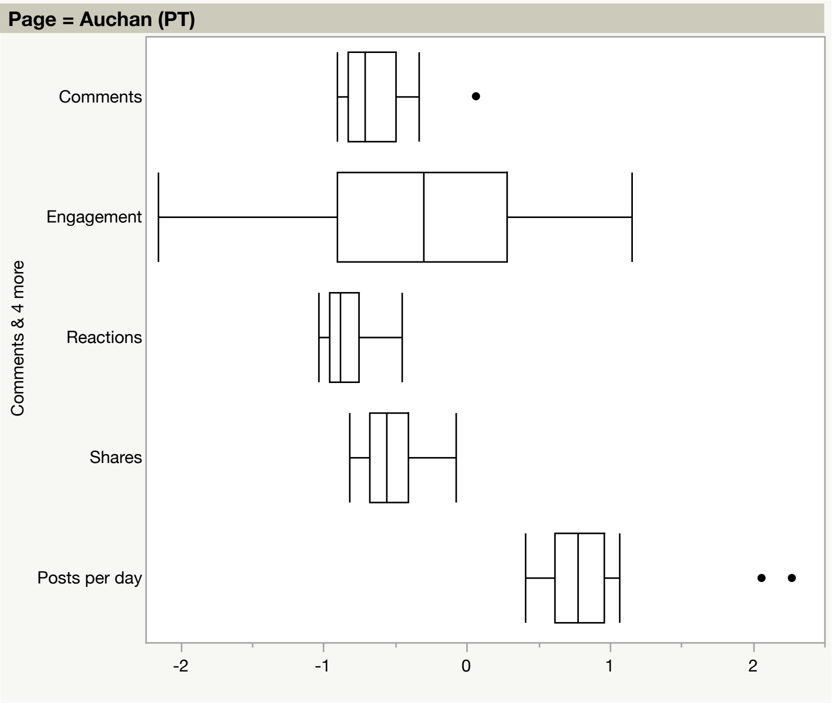


Figure 5: Boxplot Auchan Active Variable

The boxplot with Auchan’s active variables indicates that there are three observations that lie far from the rest in the dataset. The first one shows that the number of comments was significantly higher in March 2019, while the other two indicate an increase in the average posts per day in April and May of the same year.

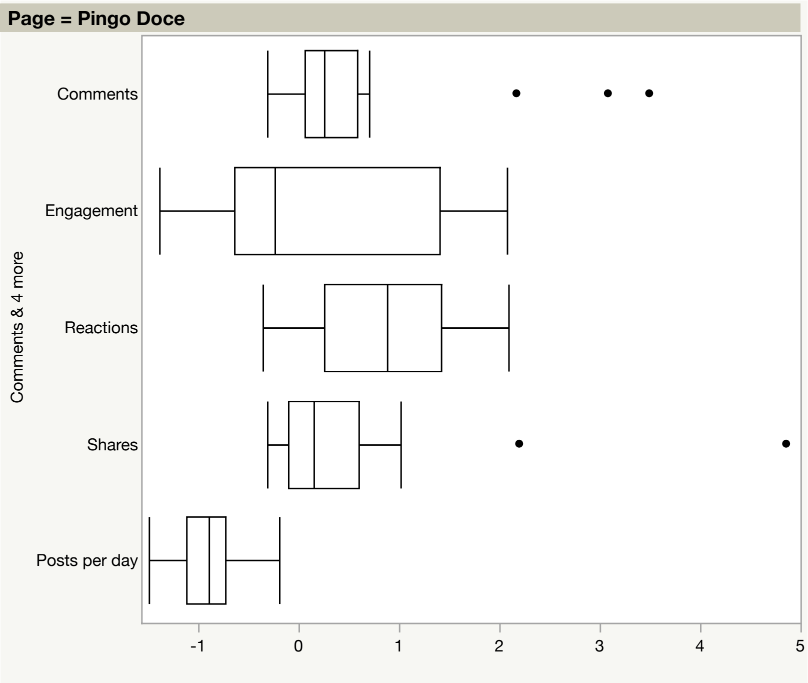


Figure 6: Boxplot Pingo Doce Active Variable

Pingo Doce’s boxplots also show the presence of outliers. In this case, there are three observations in the comments variables that are detached from the others. The increase in comments happened in August 2019, March and April 2020. Similarly, the number of shares increased exponentially in March and April 2020. This behaviour is common when brands run giveaways, for example, where users who comment and share posts get a chance of winning a prize or special discount.

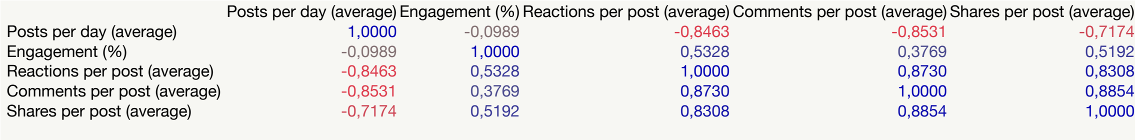
Treatment

There are different techniques to handle outliers, the most common are removing them from the dataset, replacing them with the mean, median or with the closest observation that is not considered an outlier. In this study the latter option was chosen since all outliers are on the far right of the plot, meaning that the results were very positive for the brands. The same procedure was used to treat supplementary variables outliers.

6. PRINCIPAL COMPONENT ANALYSIS

6.1 Metrics and Analysis

A correlation matrix is a table that shows correlation coefficients between variables. The table below shows that the variables reactions, posts, and shares have a strong negative correlation with posts per day, where the correlation coefficient is equal to -0.8463, -0.8531, -0.7174, respectively. On the other hand, the variables comments and share are positively correlated with the variable reactions per post.



	Posts per day (average)	Engagement (%)	Reactions per post (average)	Comments per post (average)	Shares per post (average)
Posts per day (average)	1,0000	-0,0989	-0,8463	-0,8531	-0,7174
Engagement (%)	-0,0989	1,0000	0,5328	0,3769	0,5192
Reactions per post (average)	-0,8463	0,5328	1,0000	0,8730	0,8308
Comments per post (average)	-0,8531	0,3769	0,8730	1,0000	0,8854
Shares per post (average)	-0,7174	0,5192	0,8308	0,8854	1,0000

Table 3: Statistical Description of the Dataset

The Score Plot below (figure 7), shows that the first two principal components account for 93.3% (74.5 + 18.8 = 93.3%) of the total variation in the data. These numbers are also displayed on the axes of the Loading Plot (figure 8) and on the Eigenvalues table (table 3).

Furthermore, the distance to the origin also conveys information. It can be observed that the two brands have very distinctive behaviours. All Auchan observations are negatively correlated to the first principal component, while Pingo Doce is positively correlated to the first principal component, except for observation 38 (July 2020).

On the second principal component, Auchan has a more positive correlation with this component when compared to Pingo Doce. In addition, Auchan's observations seem to be slightly more dispersed compared to Pingo Doce, that has two well defined groups.

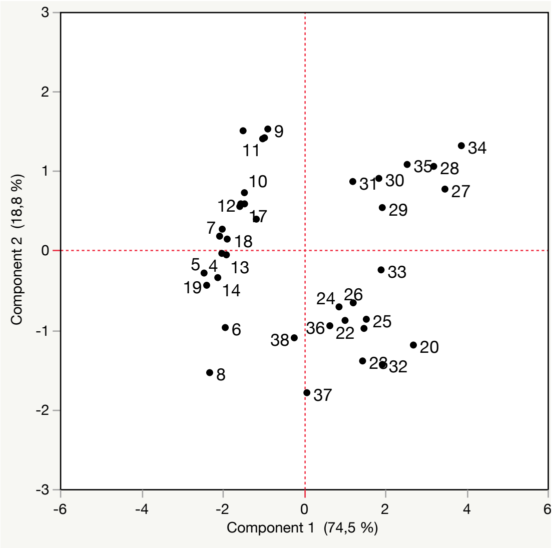


Figure 7: Score Plot

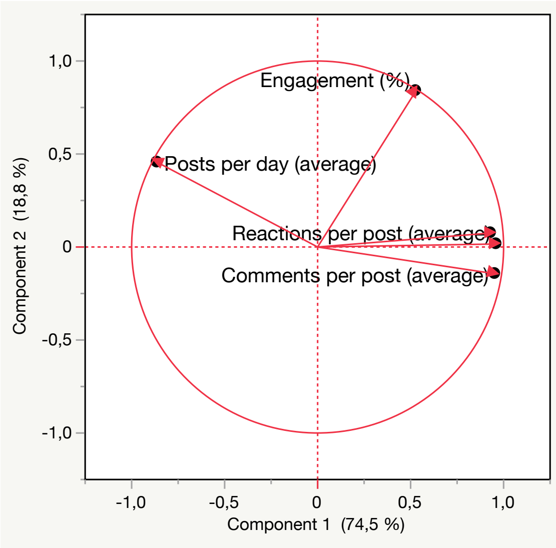


Figure 8: Loading Plot

The Loading Plot (figure 8) presents the correlations between the original variables and the first two principal components. The variables engagement (%), shares per post (average), comments per post (average), and reactions per post (average) are very high positively correlated with the first principal component, while the variable posts per day (average) is negatively correlated with the first principal component.

On the second principal component, the variables posts per day (average), engagement (%), shares per post (average), and reactions per post (average) are positively correlated, while the variable comments per post (average) is negatively correlated with the second principal component.

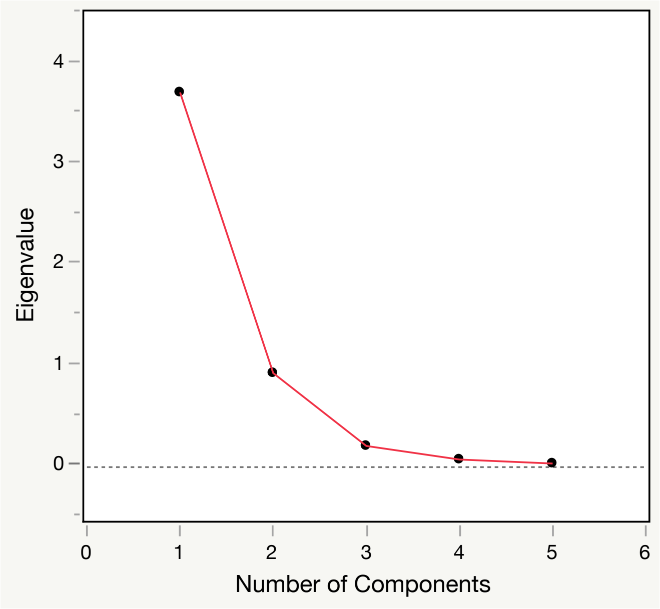
Through the loading plot (figure 8), it is also possible to analyse the angles between the vectors, which tells how the variables correlate with one another. Interestingly, the variable posts per day seems to have low correlation with the RCS variables, which means that, who posts more, does not necessarily get more reactions, comments or shares. On the other hand, the graph indicates a high correlation between reactions, comments, and shares, suggesting that individuals usually engage on all fronts.

6.2 Total Variance Explained

The Eigenvalues table below (table 3) shows the percentage and accumulate percentage of the variation accounted for by each principal component. As it can be observed, the first two principal components account for 93.2% of the variation of the data.

Number	Eigenvalue	% of variance	Cumulative %
1	3,7236	74,471	66,879
2	0.9386	18,772	93,243
3	0,2156	4,311	97,554
4	0,0810	1,619	99,174
5	0.0413	0,826	100,000

Table 3: Eigenvalues



The ideal curve of a scree plot should be steep, bending at an "elbow" and flattening out after that. The elbow indicates the cutting-off point, which in this case indicates that just the principal components 1 and 2 are enough to describe the data.

Figure 9: Scree Plot

6.3 Partial Contribution of Each Variable to the Principal Components

On table 4 below, it can be observed the partial contribution and squared cosines of each variable. These are the most important indicators to analyse when deciding which variables are useful for the interpretation of the data.

Partial Contribution

The larger the value of the contribution, the more the variable contributes to the component. A useful heuristic is to base the interpretation of a component on the observations whose contribution is larger than the average contribution. The variables posts per day (average) and engagement (%), highlighted in orange on the table 4, have a contribution below the average for the first principal component. However, posts per day value is very close to the average. The engagement variable is very important to the second principal component. Therefore, based on the contribution analysis, all variables are relevant for the analysis.

Squared Cosines

The squared cosine shows the importance of a component for a given variable. It indicates the contribution of a component to the squared distance of the observation to the origin. Just as the contribution analysis, variables with larger values contribute a relatively large portion to the total distance and therefore are important for the analysis. The sum of the squared cosines of each variable is above 0.88, meaning that all variables are well represented on the factorial plan.

	CTR1	COS^2 1	CTR2	COS^2 2
Posts per day (average)	**19.95**	**0,74**	22,07	0,21
Engagement (%)	**7.51**	**0,28**	75,09	0,70
Comments per post (avg)	24,81	0,92	0,03	0,00
Shares per post (average)	24,44	0,91	2,21	0,02
Likes per post (average)	23,29	0,87	0,60	0,01

Table 4: Partial Contribution and Squared Cosines of the Variables

6.4 Partial Contribution of Each Observation to the Principal Components

Similarly to the variables interpretation, the partial contribution of each observation to the principal component was analysed. Since the dataset contains 38 rows, only the most relevant results will be displayed in this paper, as per below.

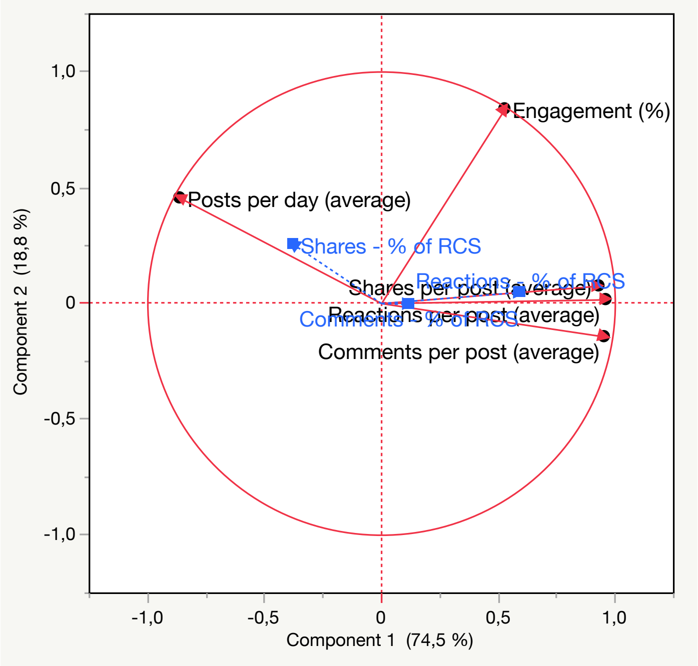
ID	Brand	CTR1	CTR2
16	Auchan	1,61957058	**6,52695211**
27	Pingo Doce	8,75795543	1,71511459
34	Pingo Doce	**10,9137454**	5,00088077

Table 5: Partial Contribution of the Observations

The contribution of the row ID 34 is nearly 11%, which can be considered a large contribution considering that there are 38 observations in the dataset. As seen on the outlier analysis, this observation presented an outstanding performance, indicating that Pingo Doce had launched a very successful strategy or campaign that led to a spike in reactions, comments and shares. Even after treating the outliers, this observation remains very relevant to the analysis.

6.5 Supplementary Variables

The supplementary variables, highlighted in blue in figure 10, have no influence on the principal component. They are going to help to interpret the dimensions of variability. Its position on the factorial plane allows to visualise the relationship of the variable with the set of active variables via the factorial axes.



The active variables reactions, comments and shares (average) seem to have a stronger influence on the supplementary variables reactions and comments % RCS. The variable shares % of RCS, on the contrary, is more correlated to the number of posts per day.

Figure 10: Loading Plot with Supplementary Variables

6.6 Results

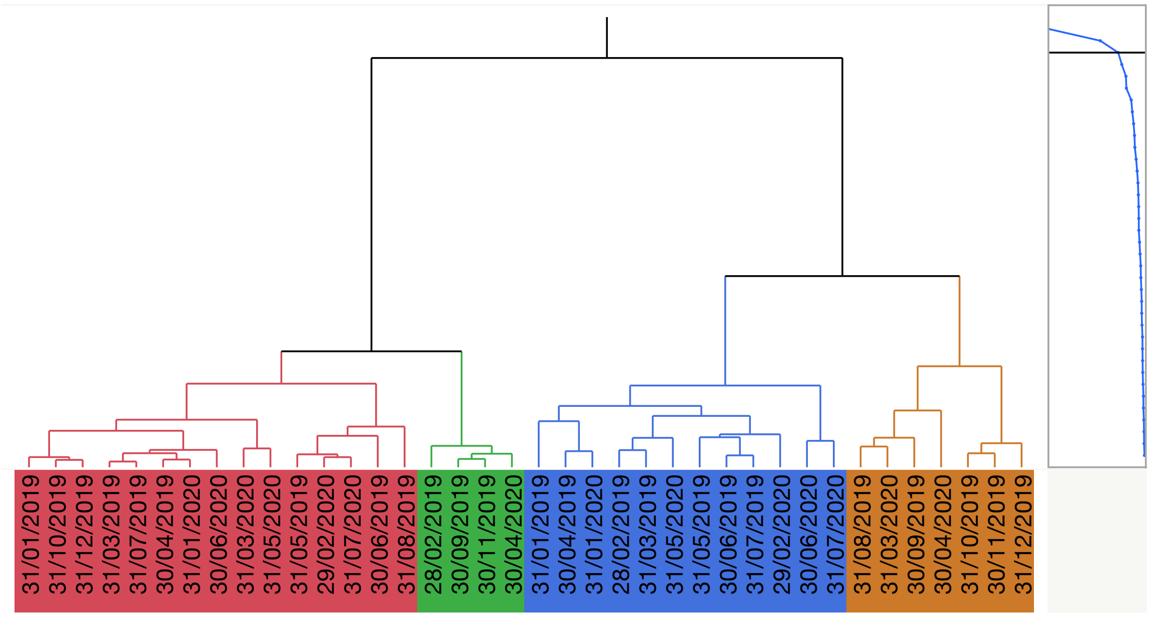
The ideal number of components in this analysis is two, because the two first principal components account for 93.3% of the cumulative proportion of the total inertia. In addition, when deciding the number of components, it is important to consider the partial contribution of each variable to the inertia and its respective squared cosines. The first principal component accounts for only 7.5% of the variability of the variable engagement, while the second component accounts for 75%. The squared cosines have a similar behaviour. To conclude, two components were chosen in this analysis.

7. CLUSTER ANALYSIS

Cluster analysis is a statistical method for processing data. It works by organising items into groups, or clusters, on the basis of how closely associated they are (Qualtics XM). In this study, the clustering methods were applied to the original variables because the number of observations is already educed. When clustering is applied using PCA, information is lost. Both clustering analysis consider the standardized variables.

7.1 Hierarchical Clustering

The Hierarchical Clustering method used in this study was the Ward's. This method joins clusters to maximize the likelihood at each level of the hierarchy. According to the JMP documentation, It tends to join clusters with a small number of observations and is biased toward producing clusters with approximately the same number of observations. It is also very sensitive to outliers.



→ Auchan Observations → Pingo Doce Observations

Figure 11: Hierarchical Clustering Dendrogram

The dendrogram generated by JMP, indicates that the optimal number of clusters is equal to four. As expected, there are no clusters that include both brands: the red and green clusters are related to Auchan's observations, while the blue and orange clusters are Pingo Doce's.

This outcome indicates that their performance on Facebook is significantly distinct. This difference could be a result of the social media strategies adopted by each brand in terms of periodicity, content and promotions. Another possibility is that they have different audiences that require different approaches, or the audience responds differently to the campaigns.

Auchan

When comparing the key performance indicators, one can conclude that the different clusters are related the engagement rate. In February, September, November 2019 and April 2021, the engagement rate was significantly higher compared to the other months. This means that strategies adopted during these months generated good results to the brand in terms of engagement.

Pingo Doce

Regarding Pingo Doce's clusters, a similar behavior can be observed, where the clusters on the right have a higher engagement rate. In this case, a seasonal effect could be attributed to the better performance experienced, as they happened between August and December 2019, a period that includes important events such as, Black Friday and Christmas celebrations. In spite of seasonality, the strategy adopted by the brand during this period kept users more engaged.

7.2 K-means

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

The number of clusters (k) selected in this analysis was four, just as the hierarchical clustering, so the results can be compared, and the robustness of the analysis can be determined.

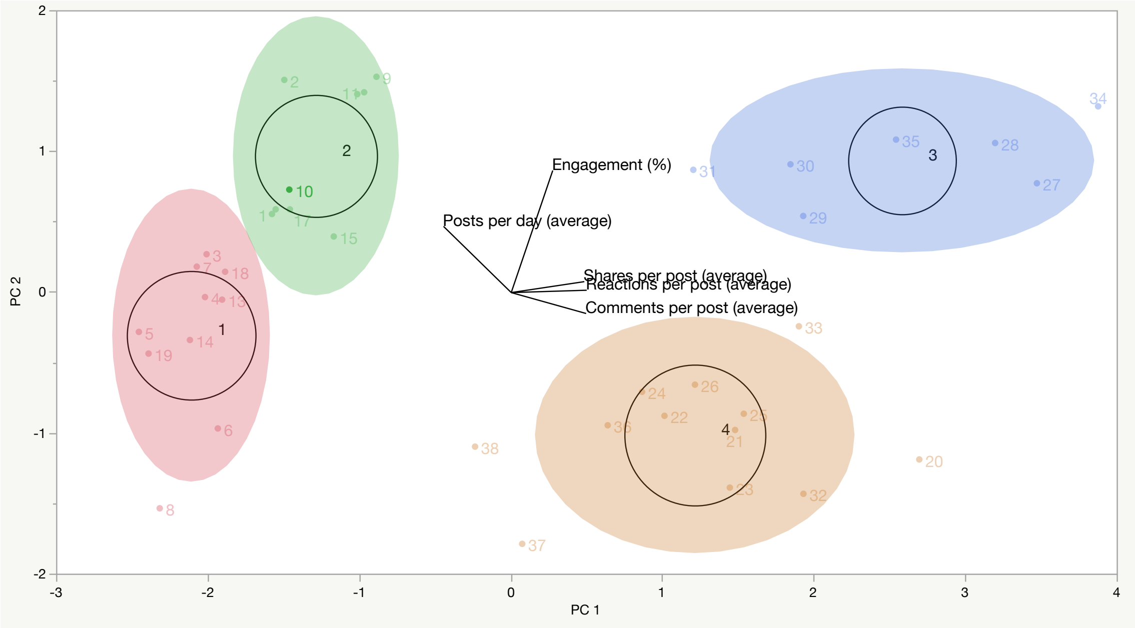


Figure 12: K-Means Clusters Biplot

In these results, K-means clusters data based on the initial partition that was specified. The cluster in red and green are Auchan's observations, while the blue and red are Pingo Doce's. One can observe that the green cluster has a great convergency of the average distance from centroid, which means the correlation satisfies the criteria.

7.3 Results

In order to compare the results of the hierarchical clustering and k-means, the result of each analysis was saved into the dataset and then tabulated.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Hierarchical Clustering	15	4	12	7
K-means	10	9	7	12

Table 6: Partial Contribution and Squared Cosines of the Variables

The output table 6 shows that both methods encountered the same results for Pingo Doce's clusters: observations ids 201, 383, 227, 409, 253, 279, and 305 were grouped together in both analysis indicating a strong similarity between them.

Auchan's results on the other hand, presented different outputs. There are five observations that were placed in different clusters: 64, 158, 356, 418, and 236. This means that the similarity between them might not be as strong as desired. This difference does not come as surprise as the data was moderately more dispersed on the score plot presented on the principal component analysis.

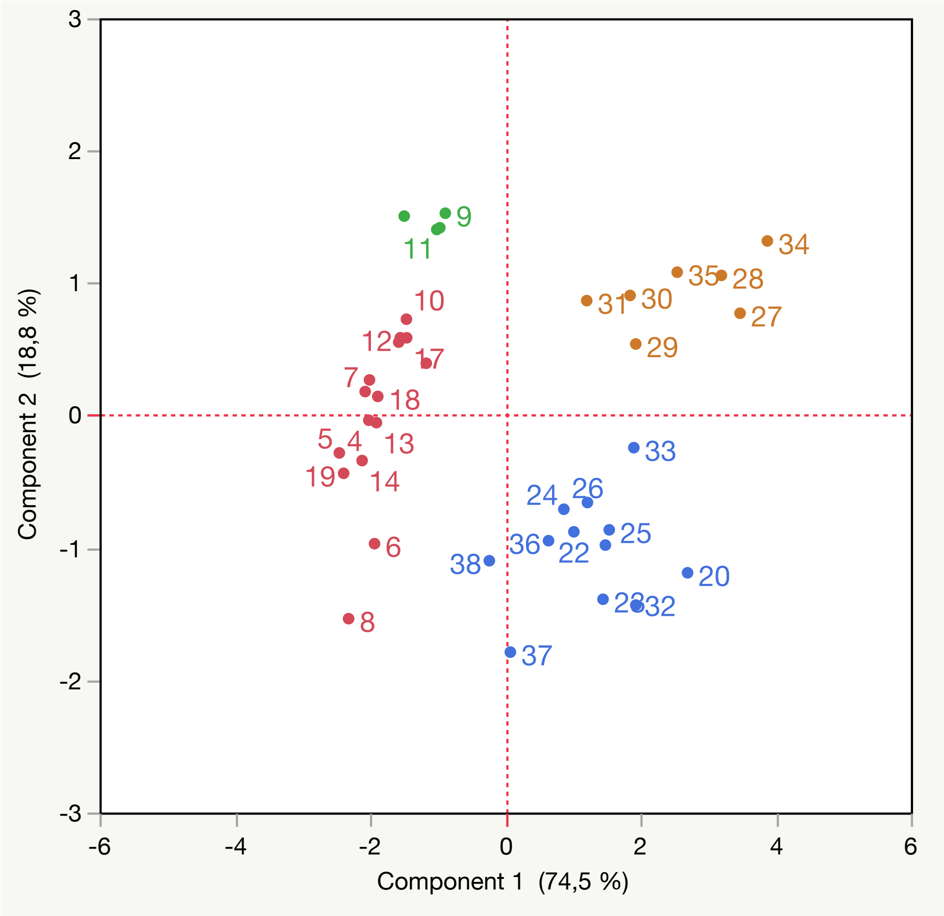


Figure 13: Score Plot with Clusters

Once the cluster results were saved into the dataset, JMP re-labeled the score plot considering the results obtained. Pingo Doce's clusters, in blue and orange, are very clear and well defined, as stated before. While Auchan's has had different results in the analysis.

8. CONCLUSION

Social media has multiple positive impacts on businesses in terms of brand recognition and consumer engagement, allowing companies to build a close relationship with their customers. Most brands have been investing in social media campaigns and their results must be measured to drive the decision-making process.

In this study, the data from Auchan and Pingo Doce's Facebook pages were analysed. A set of techniques were applied to identify data patterns and group similar observations together, as such, principal component analysis, hierarchical clustering, and k-means. The objective was to identify which brand had a higher engagement and which factors could have led to this outcome.

As a result, it was identified that, even though they are part of the same industry and share similar content (i.e., healthy food recipes), both brands have very distinctive social media results, and very likely, different core strategies and audiences. Pingo Doce's followers tend to engage more compared to Auchan's. It has also two very clear groups: one that seems to be end-of-year seasonal and another for the remaining months.

Auchan's clusters, on the other hand, are not well-defined as the hierarchical clustering and k-means presented slightly different results. This outcome did not come as a surprise, as the data was more spread out in the score plot analysed in the principal component analysis.

It is important to highlight though, that the engagement analysed in this study can be either positive or negative. In order to determine if these results are in fact beneficial for the brands, a text mining technique is highly recommended, aligned to domain knowledge.

To conclude, the techniques applied can help to understand similarities and correlation between variables and observations. This, combined with other techniques and business knowledge, can help the data-driven decision-making that uses facts, metrics, and data to guide strategic business decisions that align with its.

9. Bibliography

Allen, Ben. 2019. HOW DOES SOCIAL MEDIA IMPACT THE WAY CONSUMERS SHOP? Available on: https://www.brandbank.com/how-does-social-media-impact-the-way-consumers-shop/

Garbade, Dr. Michael J. 2018. Understanding K-means Clustering in Machine Learning. Available on: https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1

JMP Genomic's User Guide. Hierarchical Clustering Method. Available on: https://www.jmp.com/support/downloads/JMPG101_documentation/Content/JMPGUserGuide/PA_L_PD_0044.htm

Qualtics XM. What is cluster analysis and when should you use it? Available on: https://www.qualtrics.com/uk/experience-management/research/cluster-analysis/

Tam, Adrian. 2021. Principal Component Analysis for Visualization. Available on: https://machinelearningmastery.com/principal-component-analysis-for-visualization/