Customer Segmentation for Credit Card Users

Final Project Report- Credit Card Usage data - Using Clustering algorithms to detect groups of users

Sanjana Gupta
DSBA M2
ESSEC Business School and
Centrale Supélec, France
sanjana.gupta@student-cs.fr

Sarvani Sambaraju
DSBA M2
ESSEC Business School and
Centrale Supélec, France
sarvani.sambaraju@student-cs.fr

Aline Helburg
DSBA M2
ESSEC Business School and
Centrale Supélec, France
aline.helburg@student-cs.fr

Ayush Tankha
DSBA M1
ESSEC Business School and
Centrale Supélec, France
ayush.tankha@student-cs.fr

ABSTRACT

The use of data-driven strategies is becoming increasingly prevalent in business organizations as a means of generating profits. In this research paper, we examine the application of k-means clustering for customer market segmentation in the context of bank data. Our goal is to identify distinct groups of customers within a bank's client base in order to more effectively target marketing efforts and improve customer retention.

To achieve this, we first preprocess the bank transaction data to extract relevant features that can be used to represent customer behavior. These features are then fed into a k-means clustering algorithm, which divides the customers into clusters based on similarities in their feature vectors.

Various metrics were used to validate the effectiveness of our approach. We also provide visualizations of the resulting clusters to gain a better understanding of the characteristics of each cluster and how they differ from one another.

Our results demonstrate that k-means clustering is a valuable tool for customer market segmentation in the context of bank data. Banks can utilize these methods to target users and improve their marketing efforts. In addition, the insights gained from the resulting clusters can inform the development of more personalized products and services, further enhancing the customer experience.

CCS CONCEPTS

• Principal Component Analysis • Customer clustering • Segmentation • K-means clustering

KEYWORDS

Unsupervised Machine Learning, Clustering, Segmentation, PCA

INTRODUCTION

The project presented in this research paper focuses on the use of k-means clustering for customer market segmentation in the context of bank data. Our goal is to identify distinct groups of customers within a bank's client base in order to more effectively target marketing efforts and improve customer retention.

The problem we are trying to solve is the need to find efficient and effective methods for segmenting customers in the banking industry. Traditional approaches to customer segmentation, such as demographic analysis and customer surveys, can be time-consuming and subject to bias. Data-driven approaches, on the other hand, can be more objective and efficient, making them an attractive alternative.

To answer this problem, we seek to answer the following questions: Can k-means clustering be used to identify distinct groups of customers in the context of bank data? How does the performance of k-means clustering compare to other methods of customer segmentation? What are the characteristics of the resulting clusters? How do these clusters make a difference in the real world scenario?

The problem of customer segmentation is important for banks because it can help them improve customer retention and increase profits.

Potential applications of this research include the development of targeted marketing campaigns, the creation of personalized product recommendations, and the optimization of customer service strategies. By better understanding the needs and preferences of different customer segments, banks can make more informed decisions about how to engage with their clients.

1 PROBLEM DEFINITION

The project we are working on involves performing customer segmentation in order to define a marketing strategy that will be tailored to the specific needs and characteristics of different groups of customers. Specifically, we want to focus on segmentation based on credit card usage patterns, as this can provide valuable insights into the behavior and preferences of our customers. The challenge we are facing is how to find the best strategy for each segment of active users, in order to optimize the marketing efforts and achieve the highest possible return on investment.

Segmentation solutions offer a number of benefits, including the ability to provide customized products and services to different groups of customers, which can lead to higher retention rates. This is particularly important in the banking industry, where it is crucial to maintain strong relationships with customers and prevent them from switching to competitors. However, segmentation is also widely used in other fields, where it can be employed to identify opportunities for upselling and to invest in the right marketing channels.

One potential application of our segmentation work is to use machine learning algorithms to launch targeted marketing campaigns that are designed to reach specific segments of customers with high precision. By carefully targeting our marketing efforts, we hope to maximize the conversion rate of these campaigns, which will ultimately help us to achieve our business objectives. The end goal of this project is to provide a basis to develop a marketing strategy that is tailored to the needs and characteristics of our customers, and that leverages the latest advances in machine learning to deliver the best possible results.

There are many different ways to segment credit card customers in machine learning. Some common approaches include:

- **Demographic segmentation:** Dividing customers into groups based on characteristics such as age, gender, income, education level, and location.
- **Behavioral segmentation:** Dividing customers into groups based on their patterns of credit card usage, such as frequency of purchases, average transaction size, and types of products or services purchased.
- **Psychological segmentation:** Dividing customers into groups based on their attitudes, values, and motivations.
- **Value-based segmentation:** Dividing customers into groups based on the value they bring to the company, such as their level of loyalty, profitability, or potential for future growth.

By using machine learning techniques, companies can analyze large amounts of data about their customers and identify patterns and trends that can help them segment their customer base more effectively. This can help them better understand the needs and behaviors of different groups of customers, and tailor their marketing, customer service, and other efforts to better meet those needs.

2 RELATED WORK

There is much relevant research related to customer marketing segmentation using clustering. From [1], [2], [3] we learn the approach of statistical learning methods, including linear regression, classification, resampling methods and machine learning. We use these methods for the visualization of our data. We get further guidance from [4], [5] and [6] where the research allows us to understand the concepts of cluster analysis in R, including methods for analyzing unsupervised data. It provides a guide to principal component analysis in R, including methods for identifying the underlying structure in data sets and a guide to creating effective data visualizations in R, including techniques for producing high-quality plots and charts.

The research in [8] describes the development of a real-time and online system using k-means clustering and SPSS to predict sales for a specific supermarket in annual seasonal cycles. The system, which is an intelligent tool, receives input directly from sales data records and automatically updates segmentation statistics at the end of each business day.

While the work related in [9] we have the use of clustering algorithms for customer segmentation and key account management in the context of a Portuguese wholesaler for food and household supplies. A two-stage approach is proposed to improve the quality of the analysis. The first stage involves filtering key accounts using density-based outlier detection, and the second stage involves applying a Gaussian Mixture Model (GMM) to cluster smaller customers. This approach is aligned with the business implications of key accounts as exceptional customers, and it demonstrates better results compared to using a GMM alone. The conclusion is that using density-based detection followed by a GMM is beneficial for customer segmentation in B2B applications. This shows the different approaches adopted in clustering segmentation of customer population for achieving effective results. The project revolves around similar developments but different approaches for clustering.

We can also look at the applicative approach from [10] which discusses the adoption of data-driven strategies by business organizations to increase profits, with a focus on the use of artificial intelligence and machine learning in developing intelligent tools for market segmentation based on user behavior and geographical distributions. The proposed toolkit uses Principal Component Analysis (PCA) followed by k-mode clustering for segmentation, and includes interactive visualizations and maps. The architecture and decision process of the business intelligence (BI) tool are also discussed, with consideration given to factors such as business goals, size, model, and technology.

3 METHODOLOGY AND TECHNIQUES USED

To address the problem of customer market segmentation in the context of bank data [7], we followed the following steps:

1. *Data collection*: We obtained a dataset of bank transaction data from a large, national bank. The data included a variety of features such as transaction amount, transaction date, merchant category, and geographic location.

2. *Data explorations*: After data collection, it had to be analysed to see the obvious trends and understand the users on a high-level. As examples, we plotted the tenure of customers and their minimum payments. More data explorations are presented in the Python NoteBook.

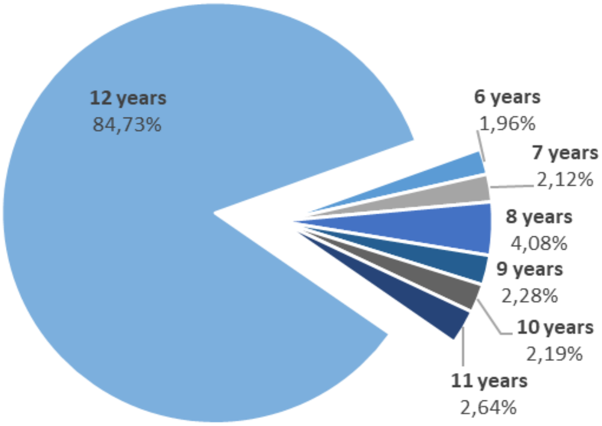

Figure 1. Tenure of customers in years

The plot above reveals a lot of customers have been loyal as a large segment of users have been with the bank for 12 years.

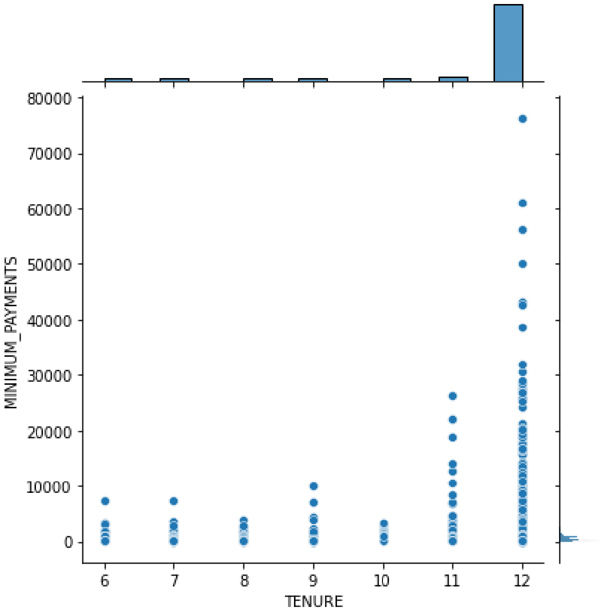

Figure 2. Jointplot between minimum payments and tenure

As we plot tenure against minimum payments, we notice the older customers make the most minimum payments.

3. *Data preprocessing:* Before applying the k-means clustering algorithm, we performed a number of preprocessing steps on the data. This included handling missing values, removing outliers, and scaling the features to a common range.

4. *Feature Engineering*: A new column TOT_TRX was created to give an overall idea about the transactions made by the user (sum of purchases and cash advance). Cash advance and purchases were taken as average values.

5. *Data Visualization*: We use the following exploratory data visualizations used for the project

   a. **Standard Scaler** - used for data visualization that scales the data to have zero mean and unit variance. This transformation can be useful for data that follow a Gaussian distribution or approximately follow a Gaussian distribution. By scaling the data to have zero mean and unit variance, the data is transformed to a standard Gaussian distribution with a mean of 0 and a standard deviation of 1. This can be helpful for data visualization because it can make patterns and trends in the data more visible and easier to interpret.

b. **Min Max Scaler** - used for data visualization that scales the data to a specific range. This transformation scales the data such that the minimum value in the data is transformed to a specific value (usually 0), and the maximum value in the data is transformed to another specific value (usually 1). This can be useful for data visualization because it can make patterns and trends in the data more visible and easier to interpret.

c. **Quartile Transformation Scaler** - used for data visualization that scales the data to the range of the quartiles. This transformation scales the data such that the minimum value in the data is transformed to the first quartile, the median value in the data is transformed to the second quartile, and the maximum value in the data is transformed to the third quartile. This can be useful for data visualization because it can make patterns and trends in the data more visible and easier to interpret. It can also be useful for data that is skewed or has outliers, as the transformation is less sensitive to these extreme values compared to the min max scaler.

6. *K-means clustering*: We applied the k-means clustering algorithm to the preprocessed data in order to identify clusters of similar customers. The k-means algorithm works by initializing K centroids, where K is the number of desired clusters. It then assigns each data point to the closest centroid and updates the centroids to the mean of the points assigned to them. This process is repeated until convergence, resulting in K clusters of data points.

We did not encounter any significant issues with scalability or overfitting in this study. However, one limitation of our approach is that the number of clusters (K) must be specified in advance, which can be challenging in some cases.

In terms of mathematical background, the k-means clustering algorithm is based on the concept of minimizing the sum of squared distances between points in a cluster and their centroid. This is achieved by alternating between two steps: assigning each point to its closest centroid, and updating the centroids to the mean of the points assigned to them. The algorithm is guaranteed to converge to a local minimum of the sum of squared distances, but the global minimum is not necessarily found.

One common challenge when using the k-means clustering algorithm is determining the appropriate number of clusters (K). In our study, we used the elbow method to assist with this task. The elbow method is a heuristic approach that involves fitting the k-means algorithm for a range of values for K and selecting the value that results in the greatest reduction in the sum of squared distances between points and their centroids.

To implement the elbow method, we first fit the k-means algorithm for a range of values for K, from 2 to 5. For each value of K, we calculated the sum of squared distances between points and their centroids, and plotted the resulting values as a function of K. The value of K at which the rate of change in the sum of squared distances begins to slow is considered the "elbow" of the curve, and is taken as the optimal value for K.

The elbow method uses the WCSS (Within Cluster Sum of Squares) value; this value shows the variations we have inside a cluster.

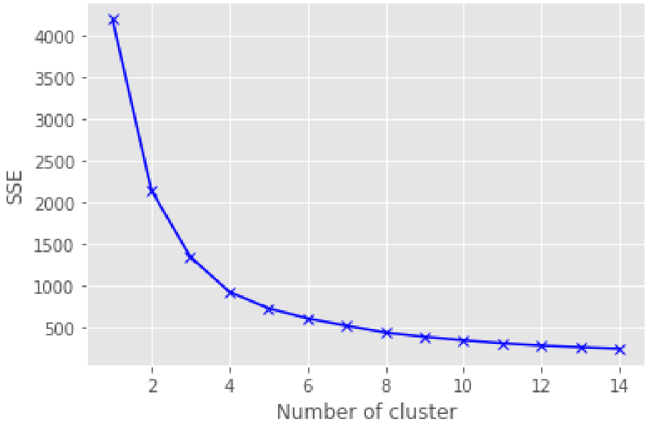

Figure 3. Sum of Squared Errors with respect to the number of clusters

In our study, we found that the elbow occurred between K = 3 and K=4, indicating that 3 or 4 clusters were the optimal numbers of clusters for our data. We used these two values of K to fit the final k-means model, which we then used to perform customer market segmentation.

We did not explore any variations of the k-means clustering algorithm in this study, but there are many variations available in the literature. For example, some variants of k-means allow for the inclusion of constraints or penalties on the centroid locations, which can be useful in certain scenarios. Other variants of k-means include the k-medoids algorithm, which uses medoids (i.e points in the data) rather than means as the centroids, and the k-modes algorithm, which is specifically designed for categorical data.

A Gaussian Mixture Model (GMMs) was used on the dataset assuming there are gaussian distributions in the dataset thus forming clusters. Calculated values of Akaike information criterion (AIC) and Bayesian information criteria (BIC) estimate the number of clusters needed. They have an indirect relationship. A loop iterates over four different covariance types ('full', 'tied', 'diag', 'spherical') and a range of number of clusters (1-20) to perform Gaussian Mixture Model clustering on a normalized dataframe (df_norm). For each combination of covariance type and number of clusters, it uses the GaussianMixture() function from the sklearn.mixture library to fit a clustering model, with the specified number of components and covariance type, and then predict the cluster assignments for each data point using the fit_predict() method.

Then, we conducted a PCA (principal component analysis) to plot the clusters on bi-dimensional scale.

Finally, we used a Python data visualization library - seaborn (sns) to visualize the resulting clusters and gain insights into the characteristics of each cluster.

4 EVALUATION

We evaluated the performance of the elbow method for the choice of the number of clusters using the following metrics: silhouette scores and Davies-Bouldin score. We calculate the silhouette score and Davies-Bouldin score for each combination of covariance type and number of clusters.

Measuring the silhouette score can give an estimate of how an individual is compared (matched) to the cluster assigned.

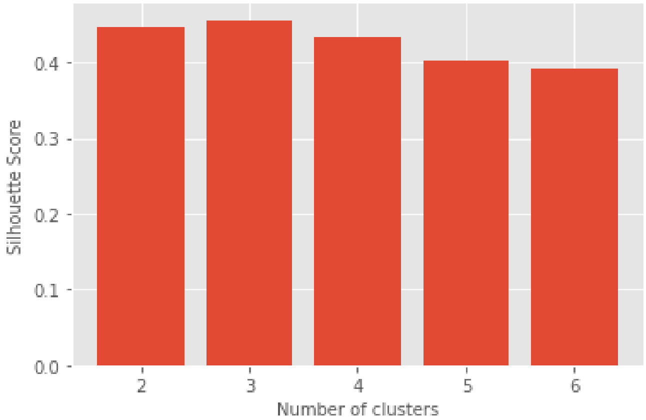

Figure 4. Silhouette score with respect to the number of clusters

The result provided by the silhouette score meets the results obtained using the elbow method: K=3 appears to be the optimal number of clusters.

These metrics showed that the chosen values of K=3 or K=4 are the optimal values that fit the best our model. K=3 appears to have a better silhouette and Davies-Bouldin scores than K=4, however 4 clusters can be used as well since it increases the precision of our analysis without losing accuracy or overfitting the data. The following analysis and conclusions will, thus, be done using 3 and 4 clusters.

5 DISCUSSION

Beside the techniques described in this paper, we implemented other techniques to better fit our model to the data.

A DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm was built to evaluate if the model fits better. A dataset cluster is stored and evaluates the results using two different evaluation metrics, silhouette_score and davies_bouldin_score, from scikit-learn library's metrics module. A check is made to see if the number of unique cluster labels (number of clusters) is greater than or equal to 2, if this condition is met, the values of eps, min_samples, the number of clusters and two evaluation scores (silhouette_score, davies_bouldin_score) are appended to the results dataframe, with the append function and passing the parameters as a dictionary object with keys matching the columns and values as the evaluated values. This code runs the DBSCAN algorithm multiple times with different eps and min_samples parameters, keeping track of the results in the results DataFrame, which at the end should contain all the different results, number of clusters, Silhouette and Davies Bouldin scores, for all the combination of eps, min_samples and the respective number of clusters formed by DBSCAN algorithm. However, we conclude that DBSCAN is not an appropriate method for the dataset.

DBSCAN may not be the optimal method for clustering for our dataset as:

1. DBSCAN is sensitive to the choice of parameters, particularly the radius (eps) and minimum number of points (minPts) required to form a dense region.
2. DBSCAN does not work well with clusters of varying densities or sizes.
3. DBSCAN does not work well with non-linear data distributions.

4. DBSCAN is sensitive to noise and irrelevant data points, it can misclassify noise points as core points and form clusters around them.
5. DBSCAN can also be computationally expensive, especially for large datasets.

Therefore, as observed it is logical that clustering algorithms such as k-means, hierarchical clustering, or Gaussian mixture models might be more appropriate to our dataset.

Furthermore, we used the Davies-Bouldin score to evaluate the optimal number of clusters. The Davies-Bouldin score is a measure of the similarity between each cluster and its most similar cluster, a low score meaning good clustering. The result DataFrame is appended with a new row of results for each combination of covariance type and number of clusters, as long as there are more than 2 clusters.

Table 1. Davies-Bouldin score with respect to the number of clusters

Number of clusters	Davies-Bouldin score
3	4.47
4	3.42
5	3.20

The Davies-Bouldin scores show that K=3 and K=4 are not the optimal numbers of clusters. However, we choose to still conduct the market segmentation with 3 and 4 clusters as these numbers were seen as optimal by the elbow method and the silhouette score.

6 CONCLUSION

The use of k-means clustering allows us to achieve a bank's customer segmentation based on their credit card transaction.

1. Customers' segmentation based on 3 clusters

First, we provided customer segmentation based on 3 different groups.

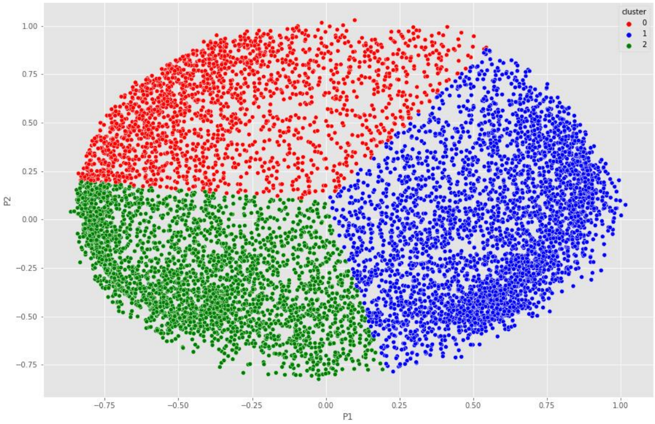

Figure 5: Repartition of the 3 clusters with K=3

With principal component 1: cash balance from low (left) to high (right)
With principal component 2: spending habits from high (bottom) to low (top)

We defined the characteristics and the best strategies to implement for each group.

1. *Group 0*: The customers have a low to median cash balance with a low spending habits. They tend to spend a low amount of money. As of now, they do not bring plus-value to the bank's profit. In order to change this situation, the bank should understand their needs and adapt their strategies in order to increase their return on investment.

2. *Group 1*: The customers in group 1 are in the top half of customers in terms of cash balance - with a high number of individuals on the extreme right of the figure (which indicate a large amount of people with a high balance). Their spending habits are heterogeneous among the scale. The bank should focus on increasing the consumption of low consumer customers while preserving the consumption of the high-spending customers.

3. *Group 2*: Those customers have a lower cash balance compared to the rest of the consumers however they tend to spend more money than the rest of the customers. We see a high density of individuals on the left of the graph - they have an extreme low balance but still a high consumption rate. For this group of customers, the bank should provide support in the financing of their purchases while making sure it does not put into jeopardy the performance and the reliability of the bank.

Then, we built a customers' segmentation based on 4 distinct groups. Four different clusters are relevant in this case as it allows us to describe with better precision the group of customers and avoid the heterogeneity in some groups (group 1 for K=3, for example) which makes it difficult to adapt the bank's strategies.

2. Customers' segmentation based on 4 clusters

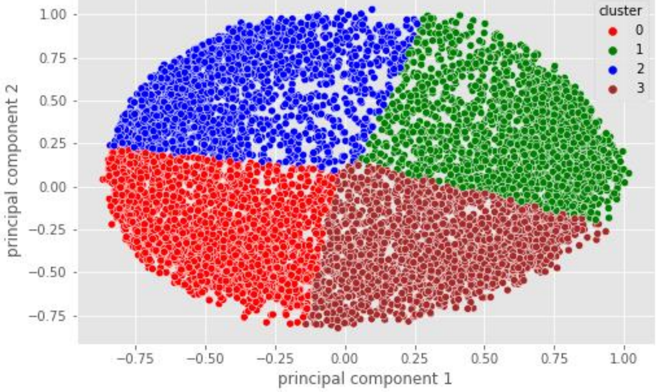

Figure 6: Repartition of the 4 clusters with K=4

With principal component 1: cash balance from low (left) to high (right)
With principal component 2: spending habits from high (bottom) to low (top)

Based on our analysis, we divided the bank's customers in four different groups.

1. *Group 0*: The customers in this group have a low cash balance, however it does not impact their spending habits. They can be described as loyal customers and frequent buyers. The key to increase their retention and their loyalty is to develop the payments in installments especially for the more expensive products. These options will push them to spend money.

2. *Group 1*: The customers in this cluster can be described as wealthy individuals (high cash balance) but they are not spendthrift. They do not use payments in installments and they have a tendency to not be in debt. In order to improve the customer experience and increase the benefits, the bank should keep pushing them to spend money on expensive items but also on less expensive daily items in order to increase the frequency and the recency. As they have an important cash balance, this group of customers can be a great asset to the bank.

3. *Group 2*: This last group of customers represents a low interest - for the moment - for the bank. Their spendings is low in terms of amount and frequency. They do not take advantage of the buying options such as payments in installments. However, they represent a large group of customers with a high tenure rate. In order to increase the return on investment of the bank, this financial institution should provide them with options and offers that encourage them to increase their spending habits.

4. *Group 3*: This group of customers may be considered as "the ideal customers". They show a high loyalty as they have been customers for a significant amount of years. They have a high frequency of purchase while having a good cash balance and they are at low-risk of bankruptcy and debt. The bank should therefore make sure to secure even more the business relationship by personalizing and adapting the relationship and their offers. Why not offer them special perks?

This paper defines a market segmentation of customers based on their behavior in terms of consumption and their available cash balance. The use of PCA and a k-means clustering algorithm allow us to reach a precise segmentation. We were able to compute common characteristics and propose strategies that the financial institution could implement in order to increase their profitability and the customer retention by offering better services and experiences. This analysis can be performed on other components such as the geographical location - the services and needs may differ according to the location.

REFERENCES

[1] James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning: With Applications in R. Springer

[2] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2008. The Elements of Statistical Learning. Springer.

[3] Brett Lantz. 2019. Machine Learning with R. Packt Publishing Ltd

[4] Alboukadel Kassambara. 2017. Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning. Sthda.com.

[5] Alboukadel Kassambara. 2017. Practical Guide to Principal Component Methods in R. Sthda.com.

[6] Alboukadel Kassambara. 2017. R Graphics Essentials for Great Data Visualization. Sthda.com.

[7] Credit Card Dataset for Clustering. https://www.kaggle.com/datasets/arjunbhasin2013/ccdata

[8] Kashwan, K.R. & Velu, C.. (2013). Customer Segmentation Using Clustering and Data Mining Techniques. International Journal of Computer Theory and Engineering. 5. 856-861. 10.7763/IJCTE.2013.V5.811.

[9] Spoor, J.M. Improving customer segmentation via classification of key accounts as outliers. J Market Anal (2022) https://doi.org/10.1057/s41270-022-00185-4

[10] Kamthania, Deepali & Pahwa, Ashish & Madhavan, Srijit. (2018). Market Segmentation Analysis and Visualization Using K-Mode Clustering Algorithm for E-Commerce Business. Journal of Computing and Information Technology. 26. 57-68. 10.20532/cit.2018.1003863.