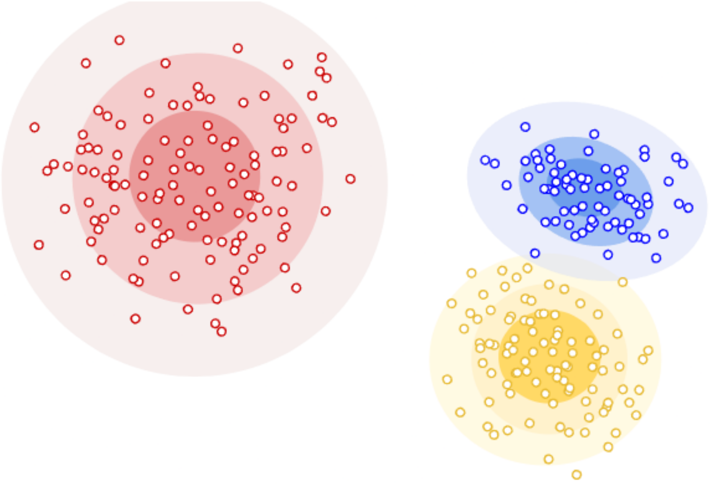
MIF11 - Machine Learning for Quotation Creation



Master’s degree in computer science – Claude Bernard Lyon 1 University

2022-2023

PhD Student:	**Teacher**	**Students**
Amira BEN HADID	Hamamache KHEDDOUCI	Kévin TANG p1501263 David TRAN p1911682 Anh-Kiet VO p1907921

Table of contents

I. Introduction

As part of our Research Introduction project under the supervision of Mr. Hamamache KHEDDOUCI at Claude Bernard Lyon 1 University, we worked on the analysis of a customer dataset to extract relevant and actionable information for a company.

This anonymized dataset was provided by the Lyon-based company Béton Direct, the leading online sales platform specializing in the delivery of fresh concrete and mortar to individuals and small building companies. We thank them for their trust and for allowing us to work with data from a professional database.

Our objective was to understand the underlying structure of the data and identify groups or clusters within the dataset.

To achieve this goal, we followed a multi-step methodology including data extraction, preliminary exploratory analysis, selection of important features, data preprocessing, determining the optimal number of clusters, applying the K-means algorithm, and analyzing the results.

This report will present in detail these different steps, the obtained results, and the main conclusions of our work.

II. Initial data analysis

1. Data extraction

As mentioned earlier, we received a dataset from Béton Direct in the form of a CSV file through our supervisor. We used specific tools and libraries, such as **pandas** and **numpy**, to import and load the data in a suitable format for analysis.
Once the data was extracted, we conducted an initial exploration to better understand it. We examined the different available variables, their types, formats, and meanings. This step allowed us to familiarize ourselves with the data and identify any potential issues such as missing values, format errors, or outliers.
Data extraction is a crucial step in any analysis as it ensures that we have a complete and appropriate dataset for the subsequent stages of our work.

2. Initial state

The database has 26 different features and is made up of 55,117 quotations from Béton Direct customers. customers, of which 9.74% have been finalized.

It can be seen that a large proportion of the quotes have not been finalized, which can be a problem for machine learning algorithms, which need many both positive and negative to learn and operate, and to produce accurate results.

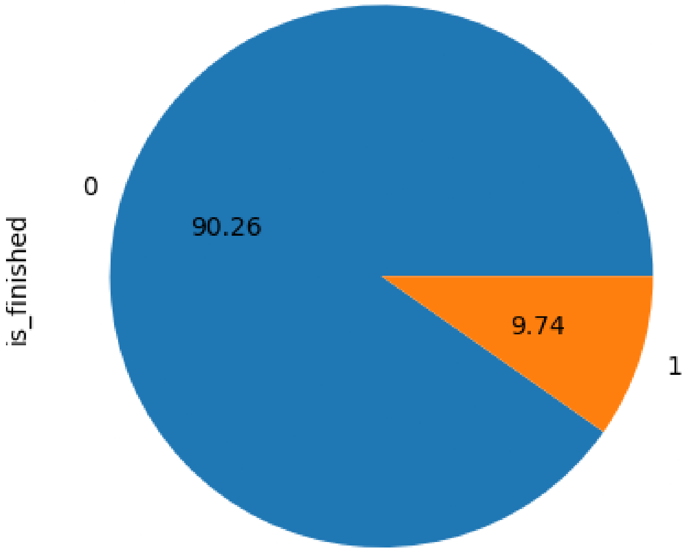


Figure 1 : Proportion diagram of the variable "is_finished"

III. Features selection

With the help of our supervisor, Mrs. Amira BEN HADID, we undertook the task of selecting the most relevant and informative features for our analysis. The objective was to reduce the dimensionality of the data by eliminating redundant or unhelpful features, in order to simplify the model and improve the performance of our analysis. We carefully examined each feature to understand its precise role and determine which ones to select or eliminate.

Feature selection allows us to focus the analysis on the most important characteristics, thereby improving the quality of results, understanding of the studied phenomena, and the efficiency of data analysis. It is an essential step in maximizing the information extracted from the data and making informed decisions.
Here are the features we selected at the end of this step:

Features
is_pro
acquisition_id
work_type_id
volume
handling
is_finished
plant_id
previous_bill_count
client_maturity

Figure 2 : List of selected variables

IV. Data preprocessing

1. Handling missing or outlier values

Cleaning missing or aberrant values is an essential step in our analysis, aiming to ensure the quality of the data used. By eliminating inconsistent or missing values, we reduce potential biases and obtain more reliable and accurate results. This strengthens the credibility of our analysis by allowing us to work with high-quality data. The conclusions we draw are thus more robust, and our interpretations more precise.

To address this step, we first identified missing values in our dataset using detection techniques such as the **isna()** function. This allowed us to identify rows or columns that contain null values.

Next, we took measures to appropriately handle these missing values. We decided to remove entire quotations containing missing values because, in our case, it does not significantly impact the quality of the database or our analysis. In other cases, we could have used an **Imputer** to replace missing values with mean, median, or most frequent values to preserve data integrity.

Additionally, we also identified and handled aberrant values in our data. Studying the data beforehand has been beneficial in this regard because with the knowledge acquired from the company's website, we understood, for example, that volumes exceeding 50 m3 are aberrant values since it is not possible to order such a volume beyond this limit.

2. Normalization of numerical values

After separating the numerical and categorical features, we focused on the numerical values of our database. This step focused on normalizing the numerical variables to eliminate potential biases and facilitate a fair comparison between different features.

To achieve this goal, we used the **StandardScaler** method, a commonly used technique in data analysis. This technique centers the values around the mean and scales them based on the standard deviation. As a result, each feature is adjusted to have a mean of 0 and a standard deviation of 1. This transformation standardizes the variables, making it easier to compare them regardless of their initial value ranges.

Normalizing numerical values is crucial in our study because it eliminates potential distortions due to differences in magnitude between variables. By standardizing the data, we were able to reduce the influence of certain features with larger value ranges on others, avoiding unjustified dominance during analysis. By removing scale differences, we were able to compare and interpret the coefficients, weights, and contributions of different variables more fairly.

3. Encoding categorical values

The data we used includes categorical variables such as gender, geographic region, or product type. However, most machine learning algorithms require data to be in numeric format to process them. There are several encoding techniques, each suitable for specific situations. In our case, we used **OneHotEncoder** from the scikit-learn library. This method creates new binary variables for each category present in the categorical variable. Each observation is represented by 0s and 1s, indicating whether it belongs to a specific category or not.

The one-hot encoding process involves creating new binary variables for each value present in a categorical variable. Each observation is represented by a binary vector, where each element indicates whether the observation belongs to a specific category or not. For example, if we have a "color" variable with categories "red", "green" and "blue" we would create three new variables: "red", "green" and "blue".

Each observation would be marked with a "1" in the corresponding variable for its color and "0" in the other variables.

Color		**Red**	**Green**	**Blue**
Red	->	1	0	0
Green	->	0	1	0
Blue	->	0	0	1

Figure 3 : Example of how OneHotEncoder works

One-hot encoding allowed us to effectively handle categorical variables with multiple distinct categories. It's important to note that other encoding methods exist, such as ordinal encoding, which assigns numerical values to each category based on a specified order. However, in our analysis, we chose to exclusively use one-hot encoding due to the nature of our categorical variables and the need to treat each category independently.

V. Determining the optimal number of clusters

1. Methods used

Once the data is properly prepared, we can now start analyzing the data in-depth.

Determining the optimal number of clusters is a preliminary step before using the K-means algorithm. This step helps us find the number of groups or clusters that maximizes intra-cluster cohesion while minimizing inter-cluster similarity.

Several approaches can be used to determine the optimal number of clusters. We used two widely employed approaches: the **elbow method** and the **silhouette score method**.

The **elbow method** involves running the K-means algorithm with an increasing number of clusters and plotting the sum of squared distances of samples to their nearest cluster center against the number of clusters. The resulting graph takes the shape of a curve resembling a bent arm, hence the name "elbow". The elbow point on the curve represents the optimal number of clusters where adding more clusters does not lead to a significant improvement in intra-cluster cohesion. However, it's important to note that the interpretation of this method can sometimes be subjective.

The **silhouette score method** evaluates the quality of a clustering by calculating the average distance between each sample and the other samples in the same cluster (intra-cluster cohesion) and comparing this distance to the average distance between the sample and the samples in neighboring clusters (inter-cluster similarity). The silhouette coefficient ranges from -1 to 1, where a value close to 1 indicates good separation between clusters, a value close to 0 indicates overlap between clusters, and inter-cluster similarity. This allowed us to identify the optimal number of clusters that provided a balance between cluster separation and intra-cluster cohesion.

2. Results obtained

We obtained results in the form of graphs for these two methods.

The elbow method allowed us to observe the variation of the intra-cluster inertia as a function of the number of clusters. We plotted a curve representing these variations and identified the elbow point, where the increase in clusters no longer results in a significant reduction in intra-cluster inertia. This point suggested an optimal number of clusters for our dataset. As we can see below, the graph indicates an elbow point for a number of 8 clusters.

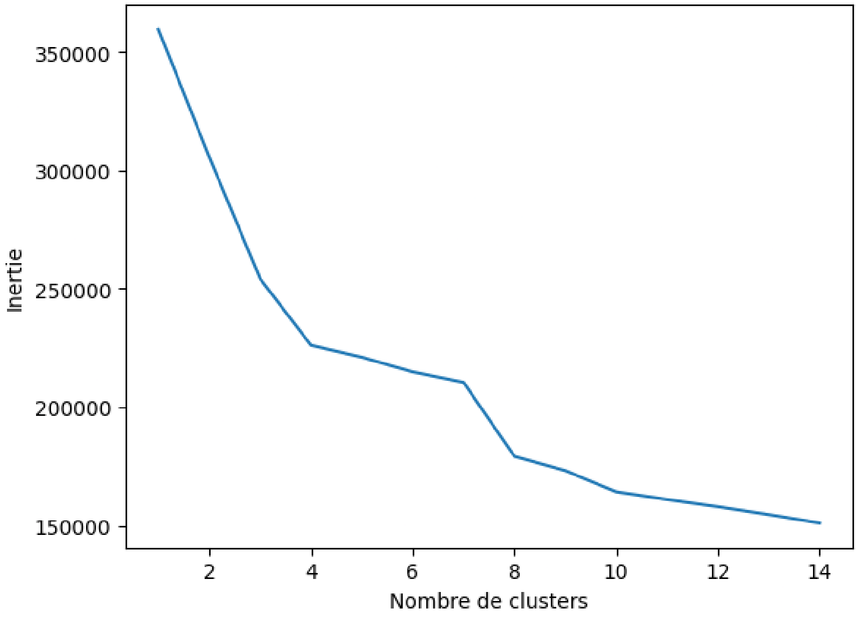


Figure 4 : Use of the elbow method chart

The silhouette coefficient method allowed us to evaluate the quality of the obtained partition for different numbers of clusters. We calculated the silhouette coefficient for each sample, which measures how well it is classified in its own cluster compared to other clusters. We then plotted a graph showing the average silhouette coefficient values for each number of clusters. The number of clusters associated with the highest silhouette coefficient indicates better data separation.

The silhouette score indicates a maximum score obtained for a number of 4 clusters. However, we can notice that the obtained score is 0.23, which is quite low. This indicates relatively weak separation between the clusters and suggests some overlap or overlap between the clusters.

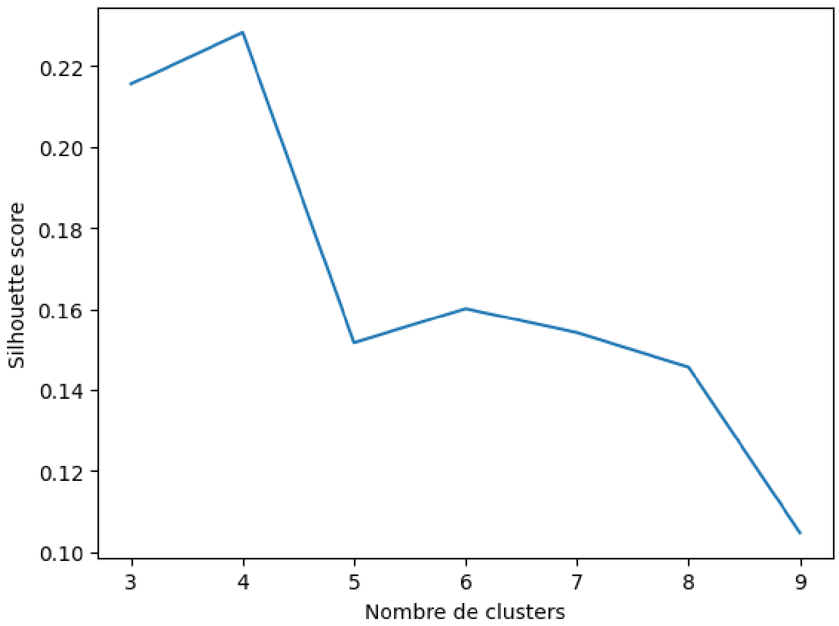


Figure 5 : Silhouette coefficient usage chart

The two methods we used yielded divergent results. Despite the questionable outcome of the silhouette coefficient, we have decided to apply the k-means algorithm using both numbers of clusters to definitively determine the optimal choice. This approach will allow us to evaluate the data structure and identify the most appropriate cluster configuration more accurately.

VI. K-Means algorithm

1. Description of the algorithm

The K-means algorithm is one of the most used methods for data segmentation. It aims to divide a dataset into a predefined number of clusters by minimizing the sum of squared distances between the data points and the cluster centers.

In our study, once we determined the optimal number of clusters using the previous methods, we applied the K-means algorithm to our preprocessed data.

The algorithm starts by randomly initializing the cluster centers and then iterates until convergence is achieved. At each iteration, data points are assigned to the nearest cluster based on Euclidean distance, and the cluster centers are updated by calculating the mean of the points assigned to each cluster.

After running the K-means algorithm, we obtained the following results: the positions of the final cluster centers and the assignment of each data point to a specific cluster. This information allows us to better understand the structure of our data and perform more in-depth analyses on each cluster.

2. Principal Component Analysis (PCA)

At this stage of the analysis, we have reduced the dimensionality of our data from 476 variables to 2 variables in order to construct a 2-dimensional graphical representation. To do this, we used **Principal Component Analysis (PCA),** a widely used statistical technique for dimensionality reduction while preserving the essential information and characteristics of the data.

PCA transforms a set of original variables into a set of new variables called principal components, which are linear combinations of the original variables. Its goal is to select a limited number of principal components that explain most of the total variance. By projecting our original variables onto the first two principal components, we were able to visualize our data in a 2-dimensional space.

It is important to note that PCA does not modify the original data but provides a new representation by rearranging the axes in space. This approach finds broad use in data exploration and visualization to facilitate analysis and interpretation of the data.

3. Graphical Analysis

- With 4 clusters :

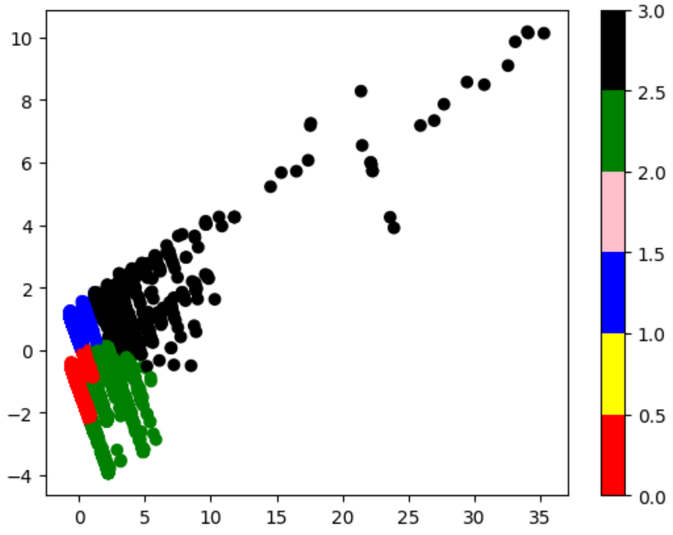


Figure 6 : Graphical representation of 4 clusters in 2 dimensions

We can see variation in the size of clusters, with some being compact while others are more scattered. In particular, the black-colored cluster on this graph contains both a region with data points very close to each other and a region with data points more spread out. This clearly shows a lack of homogeneity within this cluster and indicates that this number of clusters is not a relevant choice.

- With 8 clusters :

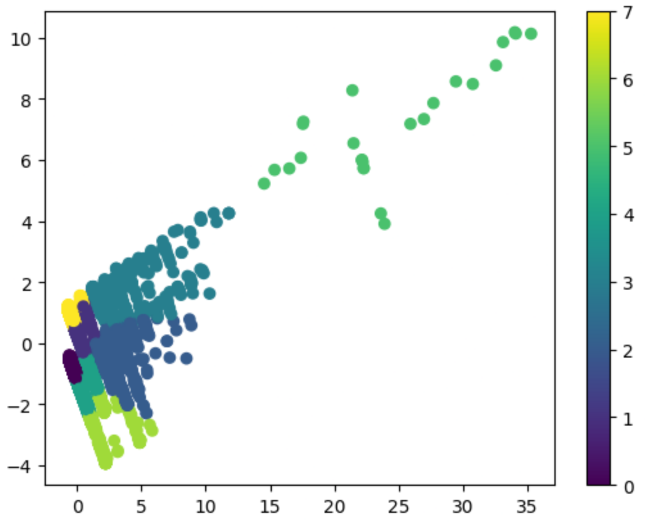


Figure 7 : Graphical representation of 8 clusters in 2 dimensions

Here, it seems that the choice of the number of clusters is more relevant here because the isolated data points are grouped together in a separate cluster, distinct from the more compactly clustered data.

For the rest of the project, we focused on analyzing the 8 clusters. It is now necessary to conduct a more detailed analysis to understand the reasons why the clusters were formed in this way.

VII. Cluster analysis

This step allowed us to better understand the factors contributing to cluster formation and to identify the key characteristics that define each group.
By analyzing the values and proportions of different features, we were able to highlight the similarities and differences between clusters, which helped us interpret the results more accurately and obtain more meaningful information about the data structure.

The in-depth analysis of the clusters enabled us to determine the variables that had the greatest influence on their formation. The results are summarized in the table below, highlighting the key features that contributed to their grouping :

Cluster	**Features**	**Percentage of influence**
1	handling	27.7 %
	acquisition_id	25.3 %
2	work_type_id	23.32 %
	handling	22.46 %
	acquisition_id	21.91 %
3	is_pro	58.71 %

	acquisition_id	14.21 %
4	volume	41 %
	acquisition_id	21.0 %
5	client_maturity	24.1 %
	handling	23.21 %
	acquisition_id	22.56 %
6	previous_bill_count	81.87 %
7	handling	26.36 %
	acquisition_id	26.14 %
8	volume	68.52 %
	acquisition_id	12.52 %

Figure 8 : Table showing the most influential features in each cluster

These percentages provide us with an overall view of the clusters and the characteristics that bring them together. However, they do not indicate the specific values of the variables that contribute to their grouping. For example, in the last cluster, we observe that volume has an influence of 68.52% on its grouping, but we do not know whether it is low or high volumes that play this role.

To clarify this information, we proceeded with the identification and counting of distinct values for each variable in each cluster. This allows us to accurately determine the specific characteristics that bring the data together within each cluster.

The results we obtained allow us to characterize each cluster with a descriptive phrase, summarizing the main features of the grouped data. By analyzing the values and proportions of the different variables in each cluster, we can make meaningful observations that describe the nature and distinctive traits of each data group. These descriptions provide valuable insights into the behaviors or profiles that emerge within each cluster, enabling a better understanding of the trends and structures present in the data.

Cluster	**Description**
1	The **non-do-it-yourself public** who discover Béton Direct by chance and buy **small volumes** of concrete
2	**Resourceful** customers who do their own work and only do **one type of job**
3	**Professionals** who need products **fast**, work in a hurry
4	Customers who buy a **medium volume** of concrete (15-35)
5	**Resourceful** customers who do the work themselves **within 15 days**

6	**Professionals** who buy very **regularly**, the most **loyal** to Béton Direct, **always buy the same product**.
7	**Resourceful** customers who do the work themselves and often buy the **same product**
8	Buyers of **large volumes** of concrete

Figure 9 : Descriptive table of clusters

VIII. Conclusion

1. Summary

In conclusion, our study focused on the analysis and classification of data using the K-means algorithm. We followed a rigorous methodology that included data extraction, feature exploration, selection of relevant variables, handling missing or outlier values, preprocessing numerical values, and encoding categorical values. With the elbow method, we determined that the optimal number of clusters for our dataset was 8. We then applied the K-means algorithm to group the data based on their similarities.

The analysis of the results revealed distinct clusters, each with specific characteristics. This information allowed us to formulate concise and precise descriptions for each cluster, highlighting the key factors that differentiate them.

However, it is important to note that our study has some limitations. The results and conclusions obtained are specific to our dataset and may not be generalizable to other datasets.

2. Opening to the next project

The knowledge gained through this analysis can be used to make informed decisions, improve strategies, and enhance performance for the company Béton Direct.

Furthermore, as part of our study, we conducted a cross-analysis with the work of the group led by Yann and Iliesse, who also explored the same dataset. Their main objective was to predict which customers were most likely to finalize their quotes to create a customer classification for prioritizing customer callbacks.

In this regard, we developed our own approach to classification based on the importance of clusters, considering the characteristics and groupings we identified. We then shared these results with the group of Yann and Iliesse to complement their work and provide a more comprehensive view of customer analysis. Here is the ranking we provided :

Ranking	Clusters
1	Cluster 5
2	Cluster 2
3	Cluster 4
4	Cluster 0

5	Cluster 6
6	Cluster 1
7	Cluster 3
8	Cluster 7

Figure 10 : Ranking of cluster priorities

This collaboration between our two teams has enriched our respective results and opened new perspectives for optimizing customer callbacks. By combining our approaches and leveraging the synergies between our work, we were able to provide more targeted and relevant recommendations for follow-up actions.

IX. Illustration table

X. Appendices

Scikit-learn library:
https://scikit-learn.org/stable/index.html

Machine Learnia YouTube account:
https://www.youtube.com/@MachineLearnia

Mohamad KANAAN, Analysis of customer behavior on an online merchant site, Lyon, Claude Bernard University Lyon 1 (doctoral thesis in computer science), 142 p.