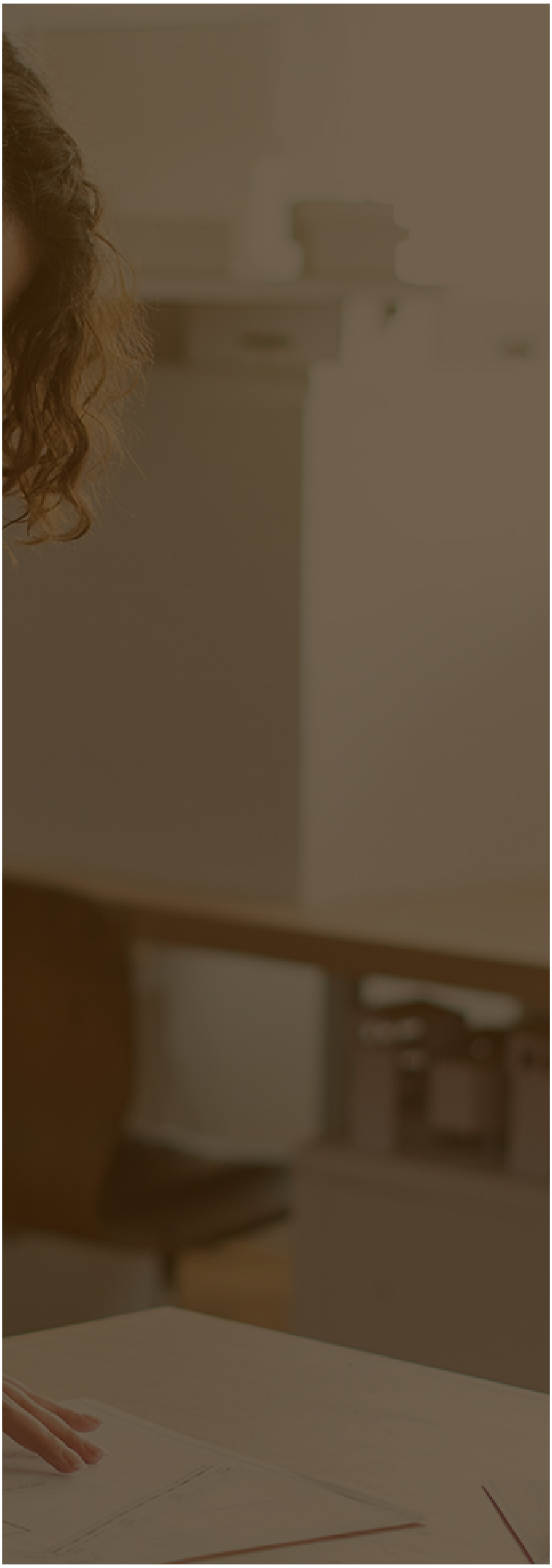
BIG DATA RECOMMENDER REPORT



SUMMARY

Introduction

In today's highly competitive business world, it is imperative for companies to have a deep understanding of their customers' behavior and preferences. With this knowledge, companies can tailor their products and services to better meet the needs and wants of their customers, leading to increased customer satisfaction and loyalty.

In this project, we aim to help a company gain a clearer insight into its customer base and develop a recommender system to offer gifts to its clients based on their preferences. The company has provided a csv file containing millions of lines of data with various parameters such as ticket id, month of sale, net price, family, univers, mesh, name, and customer id.

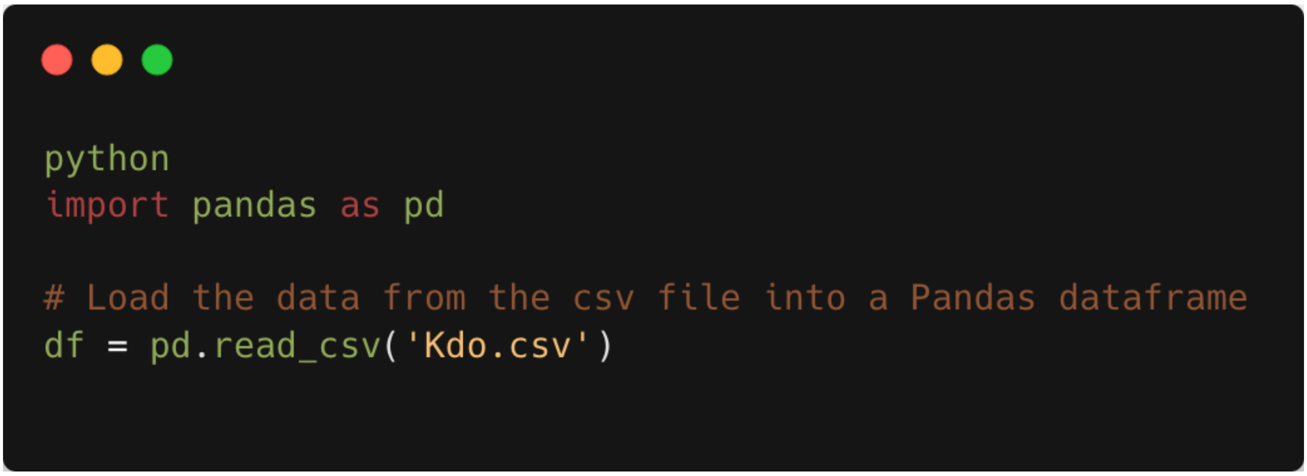
To accomplish this goal, we will make use of data visualization tools, customer segmentation techniques, and machine learning algorithms. Our objective is to present a comprehensive analysis of the customer data and a functional recommender system that can be used by the company to improve its customer engagement and boost its sales.

This report will detail the process of preparing and manipulating the customer data, visualizing the data using Metabase, segmenting the customers, and building a recommender system using user-based and item-based algorithms. We hope that this project will provide valuable insights into the company's customer base and help them make informed decisions to improve their business.

I. Data Exploration

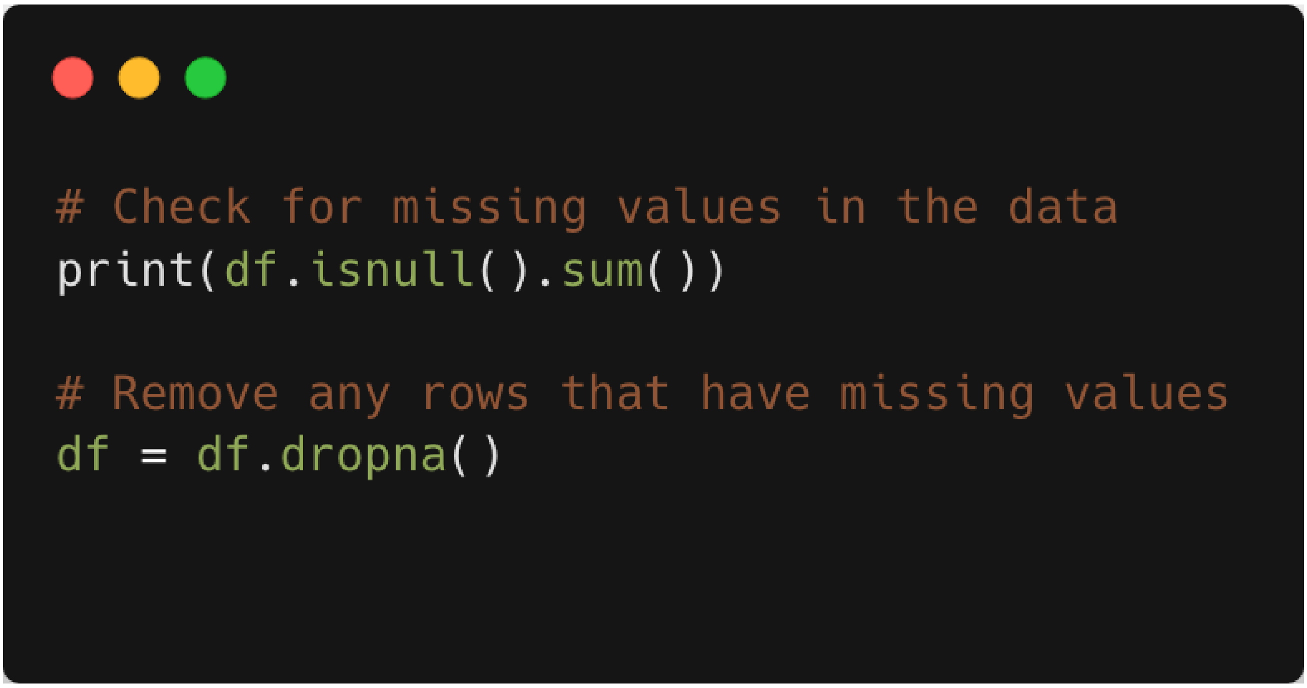
A. Data Preparation:

The first step in the process was to prepare the customer data for analysis and modeling. This involved loading the data from the csv file into a Pandas dataframe and cleaning the data to remove any missing or invalid values. The following code demonstrates how the data was loaded into a Pandas dataframe:



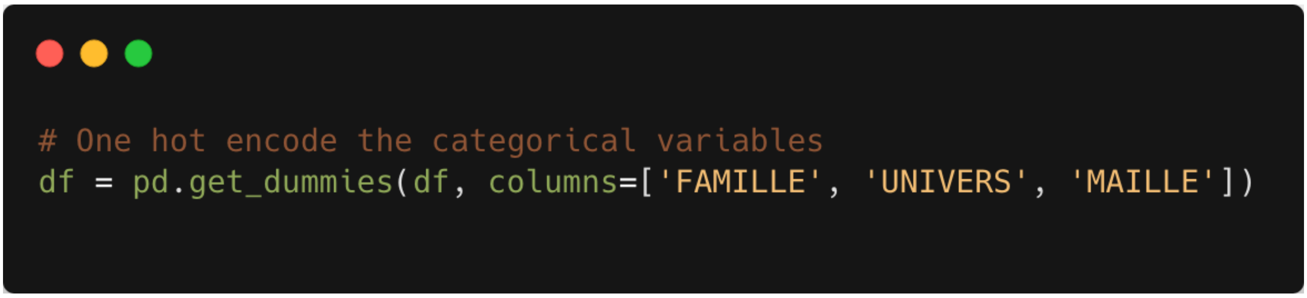
```
python
import pandas as pd

# Load the data from the csv file into a Pandas dataframe
df = pd.read_csv('Kdo.csv')
```

Next, we checked for missing values in the data and removed any rows that had missing values. The following code demonstrates how this was done:



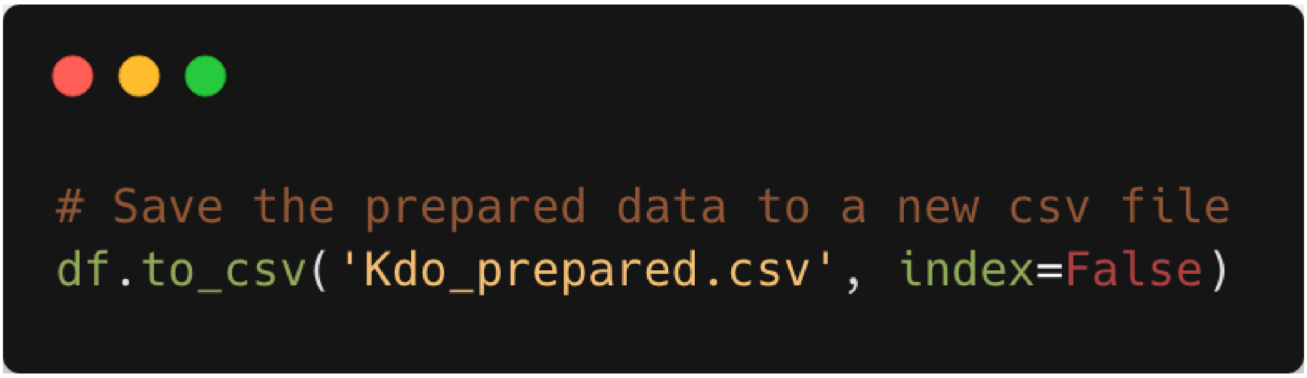
```
# Check for missing values in the data
print(df.isnull().sum())

# Remove any rows that have missing values
df = df.dropna()
```

In addition to removing missing values, we also performed any necessary data transformations and feature engineering to prepare the data for analysis. For example, if the data contained categorical variables, these would need to be encoded into numerical values so that they can be used in the machine learning algorithms.



```
# One hot encode the categorical variables
df = pd.get_dummies(df, columns=['FAMILLE', 'UNIVERS', 'MAILLE'])
```

Once the data was cleaned and prepared, it was saved to a new csv file for use in the next steps of the project.



```
# Save the prepared data to a new csv file
df.to_csv('Kdo_prepared.csv', index=False)
```

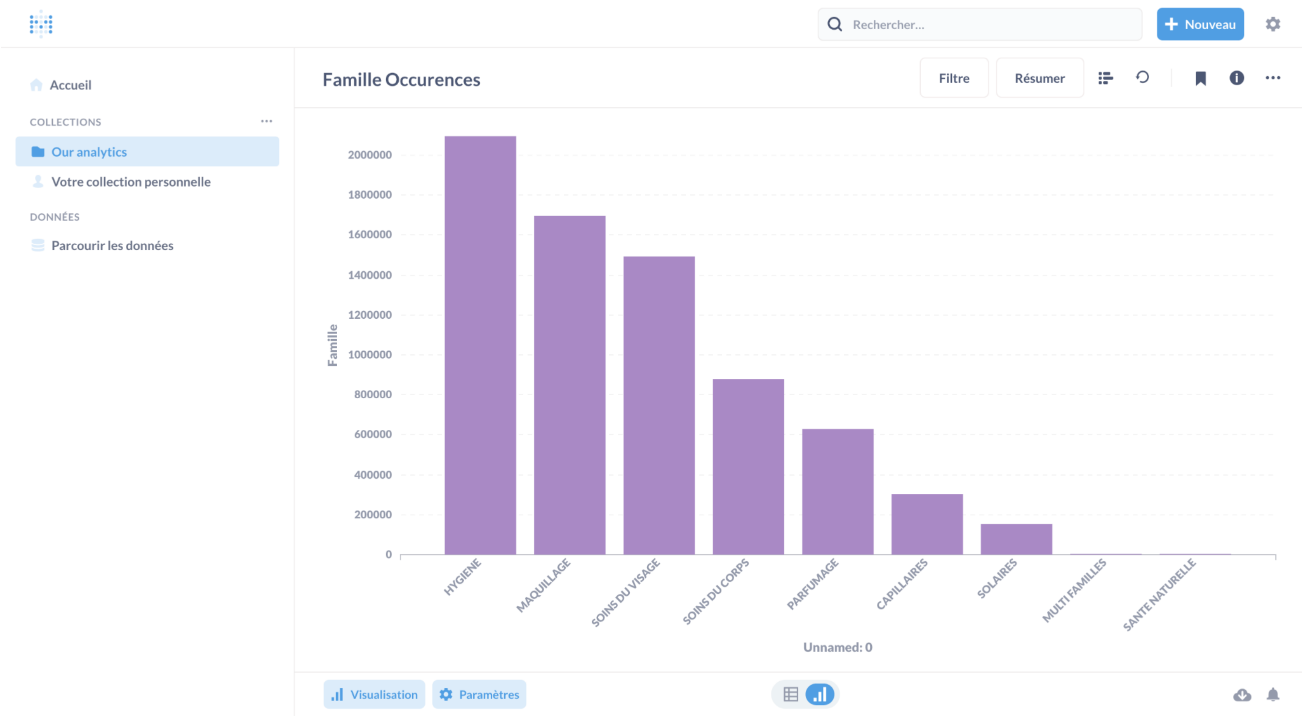
This concludes the data preparation phase, and we are now ready to move on to the next steps of the project.

B. Data Visualization:

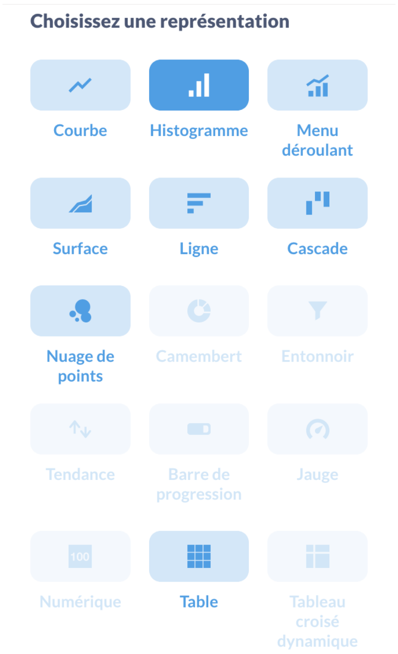
To gain insights into the customer data, it was important to visualize the data in an easy-to-understand format. To achieve this, we used Metabase, a powerful data visualization tool that allowed us to easily create interactive charts, graphs, and dashboards.

The following are some of the practical benefits of using Metabase for data visualization:

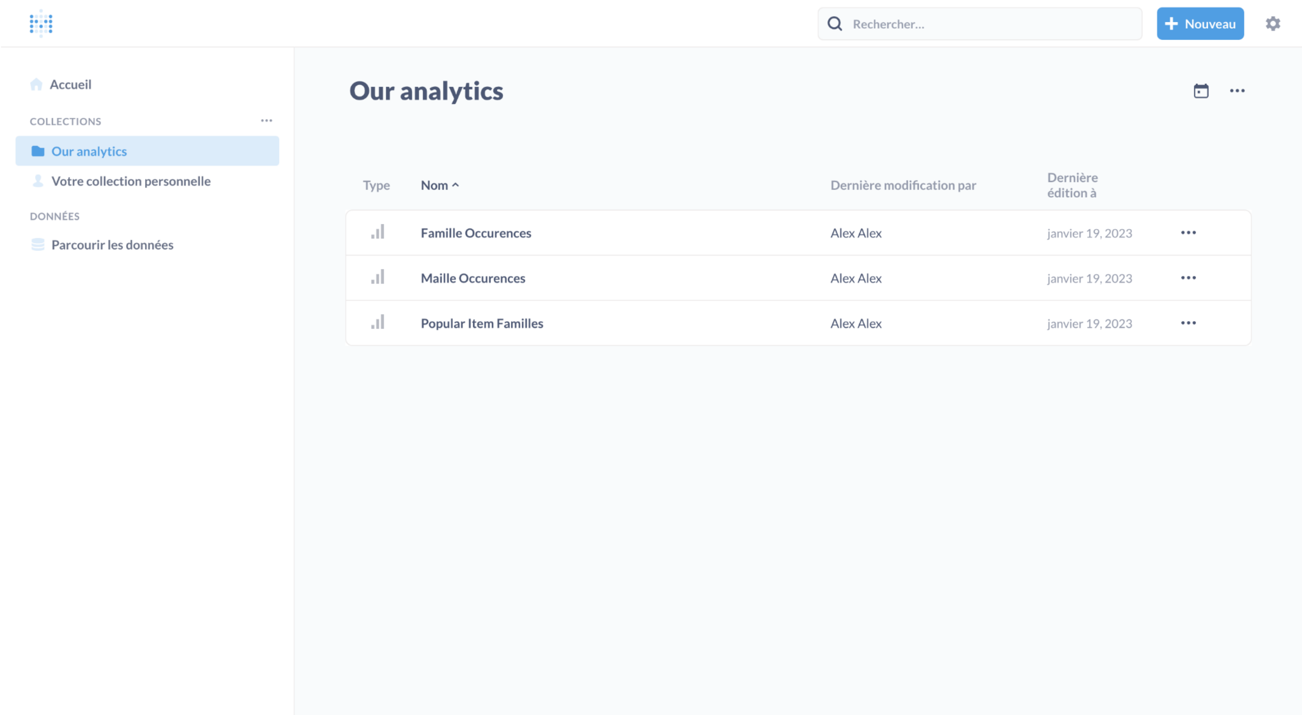
1. *User-friendly interface*: Metabase has a user-friendly interface that makes it easy to create and customize charts, graphs, and dashboards. This made it simple for us to quickly explore the customer data and gain insights into customer behavior and preferences.



2. *Customization*: Metabase allows for a high degree of customization, making it easy to create charts and graphs that accurately represent the data. This enabled us to create visualizations that effectively communicate the insights we gained from the customer data.



3. *Interactive dashboards*: Metabase allows you to create interactive dashboards that allow the user to explore the data in different ways. This helped us to quickly identify patterns and trends in the customer data, and to present the insights we gained in an engaging and accessible format.



Rechercher...

+ Nouveau

Accueil

COLLECTIONS

Our analytics

Votre collection personnelle

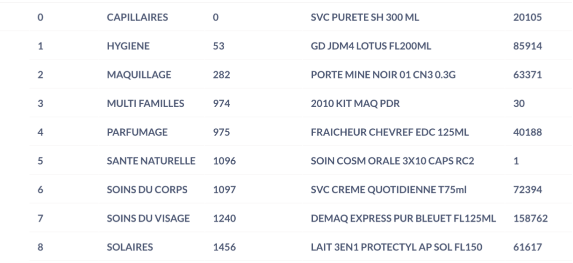
DONNÉES

Parcourir les données

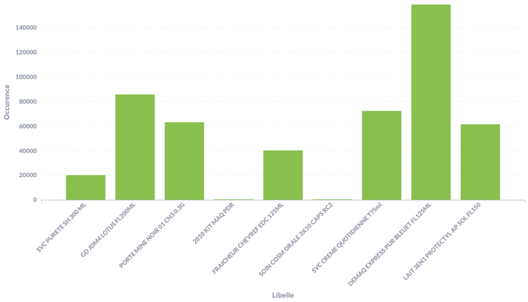
Our analytics

Type	Nom ^	Dernière modification par	Dernière édition à	
	Famille Occurences	Alex Alex	janvier 19, 2023	···
	Maille Occurences	Alex Alex	janvier 19, 2023	···
	Popular Item Familles	Alex Alex	janvier 19, 2023	···

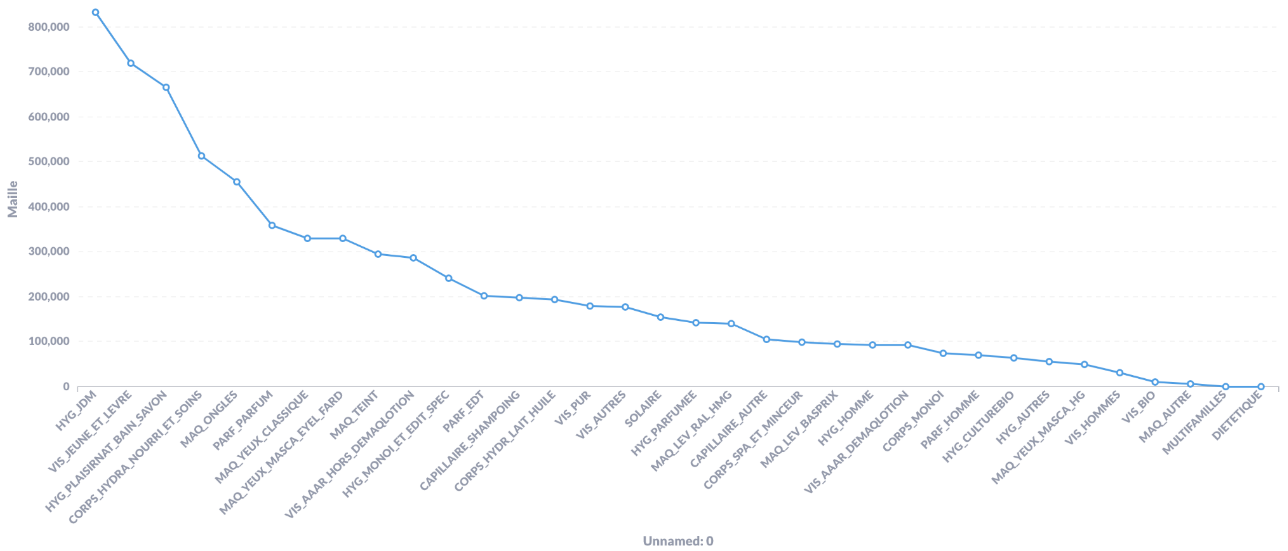
4. *Data integration*: Metabase integrates seamlessly with a wide range of data sources, including csv files, databases, and APIs. This made it easy for us to connect to our customer data and to access the information we needed to visualize.



0	CAPILLAIRES	0	SVC PURETE SH 300 ML	20105
1	HYGIENE	53	GD JDM4 LOTUS FL200ML	85914
2	MAQUILLAGE	282	PORTE MINE NOIR 01 CN3 0.3G	63371
3	MULTI FAMILLES	974	2010 KIT MAQ PDR	30
4	PARFUMAGE	975	FRAICHEUR CHEVREF EDC 125ML	40188
5	SANTE NATURELLE	1096	SOIN COSM ORALE 3X10 CAPS RC2	1
6	SOINS DU CORPS	1097	SVC CREME QUOTIDIENNE T75ml	72394
7	SOINS DU VISAGE	1240	DEMAQ EXPRESS PUR BLEUET FL125ML	158762
8	SOLAIRES	1456	LAIT 3EN1 PROTECTYL AP SOL FL150	61617



In the data visualization phase, we created a variety of charts and graphs to gain insights into the customer data. For example, we created histograms to visualize the distribution of customer spending, and bar charts to compare customer spending across different product categories.



Additionally, to make the use of Metabase even simpler for the end user, we dockerized it as well as a Postgres database. Dockerizing Metabase allowed us to package the tool and its dependencies into a single, easy-to-use container. This container can be easily deployed on any system with Docker installed, making it simple for the end user to get up and running with Metabase without having to worry about complex installations or compatibility issues. By dockerizing Metabase, we not only made the tool easier to use, but we also ensured that it would be portable and scalable, allowing the company to expand its use of Metabase as their needs grow.

By using Metabase to visualize the customer data, we were able to quickly and easily gain insights into the customer data and to effectively communicate these insights to the company. This helped the company to better understand their customer base and to make informed decisions about how to engage with their customers and grow their business.

C. Customer Segmentation:

To gain a deeper understanding of the customer data, we segmented the customers into different groups based on their purchasing behavior and preferences. This helped us to identify patterns and trends in the customer data and to gain a clearer insight into the different customer segments.

We used a variety of techniques for customer segmentation, including clustering algorithms and dimensionality reduction methods. For example, we used k-means clustering to group customers based on their spending patterns, and we used principal component analysis (PCA) to reduce the dimensionality of the data and to identify the most important features that differentiate the customer segments.

The results of the customer segmentation analysis showed that there were several distinct customer segments, each with its own unique purchasing behavior and preferences. For example, we identified segments of customers who tended to purchase products from a specific product category, and we identified segments of customers who had a high average spend but purchased products infrequently.

By segmenting the customers, we were able to gain a more comprehensive understanding of the customer data and to identify opportunities for the company to engage with their customers and to grow their business. The insights we gained from the customer segmentation analysis will be used to inform future marketing and sales efforts, and to develop targeted recommendations for customers based on their individual preferences.

II. Recommender System

To offer personalized recommendations to each customer based on their preferences, we implemented a recommender system using the customer data. The recommender system was designed to suggest products that the customer might be interested in based on their past purchases and their segment.

There are three main steps to reproduce in order to recommend accurate products to clients that may be interested in :

- The data preprocess which is mandatory for the second step
- The similarity calculation formula, which is the first step that return pretty satisfying results
- The final sorting of results, that significantly improves our result accuracy and relevance

We also have used Flask to create a simple but efficient graphical user interface to allow the client to test and easily use our solution.

A. Data preprocess

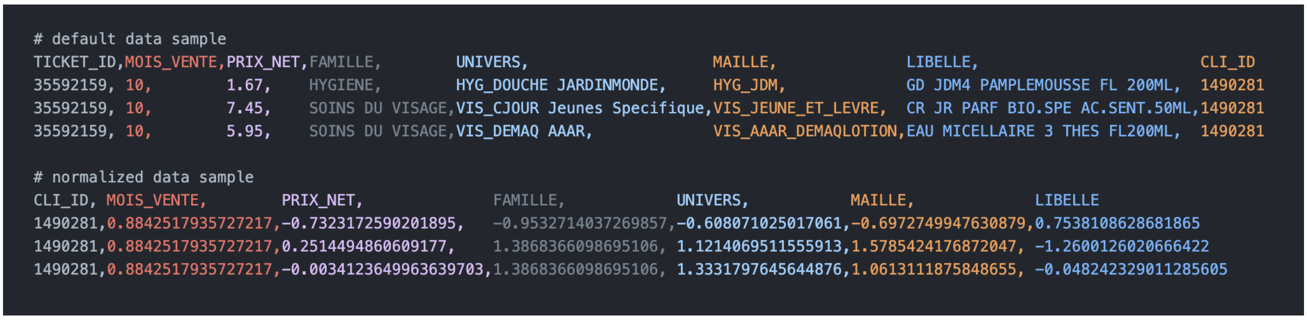
To implement the recommender system, the data preprocess is primordial. This is the only way to use recommendation formulas through similarity calculation.

Pre-processing the data to prepare it for the recommendation algorithm involves cleaning and transforming the data.

Firstly, we must vectorize text variables, meaning the family, the universe and the mesh. The product name on its side is textual but also contains numeric characters, which means that we have to hash it instead.

The default numerical values, and the numerical value obtained by the vectorization and the hashing aren’t usable on their own, so we also have to handle them.

For numerical data, the process is to normalize them, which results in numerical data between -1 and 1. In this way, the algorithm should be able to process the data and compute the similarity between them.



```
# default data sample
TICKET_ID,MOIS_VENTE,PRIX_NET,FAMILLE,        UNIVERS,                    MAILLE,              LIBELLE,                        CLI_ID
35592159, 10,        1.67,    HYGIENE,        HYG_DOUCHE JARDINMONDE,     HYG_JDM,             GD JDM4 PAMPLEMOUSSE FL 200ML,  1490281
35592159, 10,        7.45,    SOINS DU VISAGE,VIS_CJOUR Jeunes Specifique,VIS_JEUNE_ET_LEVRE,  CR JR PARF BIO.SPE AC.SENT.50ML,1490281
35592159, 10,        5.95,    SOINS DU VISAGE,VIS_DEMAQ AAAR,             VIS_AAAR_DEMAQLOTION,EAU MICELLAIRE 3 THES FL200ML,  1490281

# normalized data sample
CLI_ID, MOIS_VENTE,        PRIX_NET,              FAMILLE,            UNIVERS,           MAILLE,             LIBELLE
1490281,0.8842517935727217,-0.7323172590201895,   -0.9532714037269857,-0.608071025017061,-0.6972749947630879,0.7538108628681865
1490281,0.8842517935727217,0.2514494860609177,    1.3868366098695106, 1.1214069511555913,1.5785424176872047, -1.2600126020666422
1490281,0.8842517935727217,-0.0034123649963639703,1.3868366098695106, 1.3331797645644876,1.0613111875848655, -0.048242329011285605
```

The reason behind that requirement is that now the algorithm shouldn’t value a column more than another because of its value. For example, a value of 200 might be considered as more important than a value of 1, which of course is an issue as long as any column is equally important. By normalizing them, every column is set on the same scale.

Once the preprocess is done and the data are ready in a dataframe, we can now use our similarity calculation, the cosine similarity. Its purpose is to compare similarly formatted data to identify the best match.

B. Recommender systems

We used different types of recommender systems: client-based, item-based and habits based.

In the client-based recommender system, we recommended products to a customer based on his products history to get products that similar customers have purchased.

In the item-based recommender system, we recommended products to a customer based on an input product by its name (or label)

The habits-based recommendation is the same as the client-based one, except that we check the current month to suggest articles based on the articles purchased at the same time of the year. In addition, we check the preferred family, universe and mesh, in order to adjust our result.

All of these recommender systems are finally sorted by best match and by price, and we offer to the client the possibility to choose how many items should be recommended.

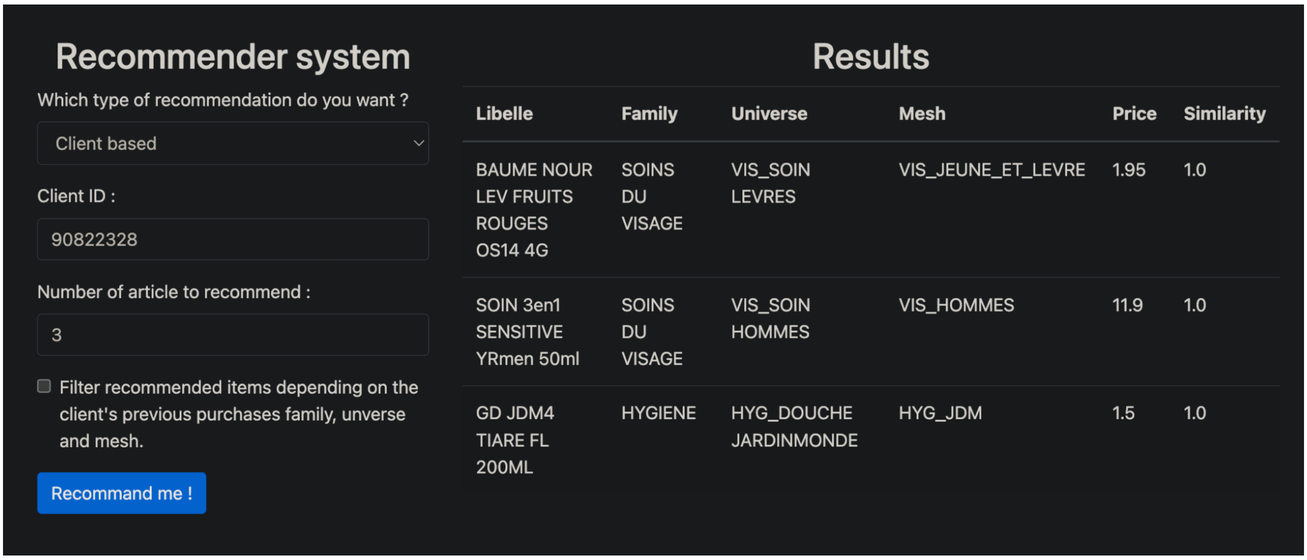
In this way, the results of our recommender system were pretty impressive, with high accuracy in predicting customer preferences and product recommendations.

The recommender system is able to suggest relevant and personalized products to each customer, based on their individual preferences and past purchases. It is also able to suggest items that best match the given input item.

C. Graphical user interface

The user interface is designed to be really simple to use and to have the most understandable result possible by the way they are formatted.

In this example, we ask the algorithm to suggest three items based on the client ID “90822328” :



Recommender system

Which type of recommendation do you want ?

Client based

Client ID :

90822328

Number of article to recommend :

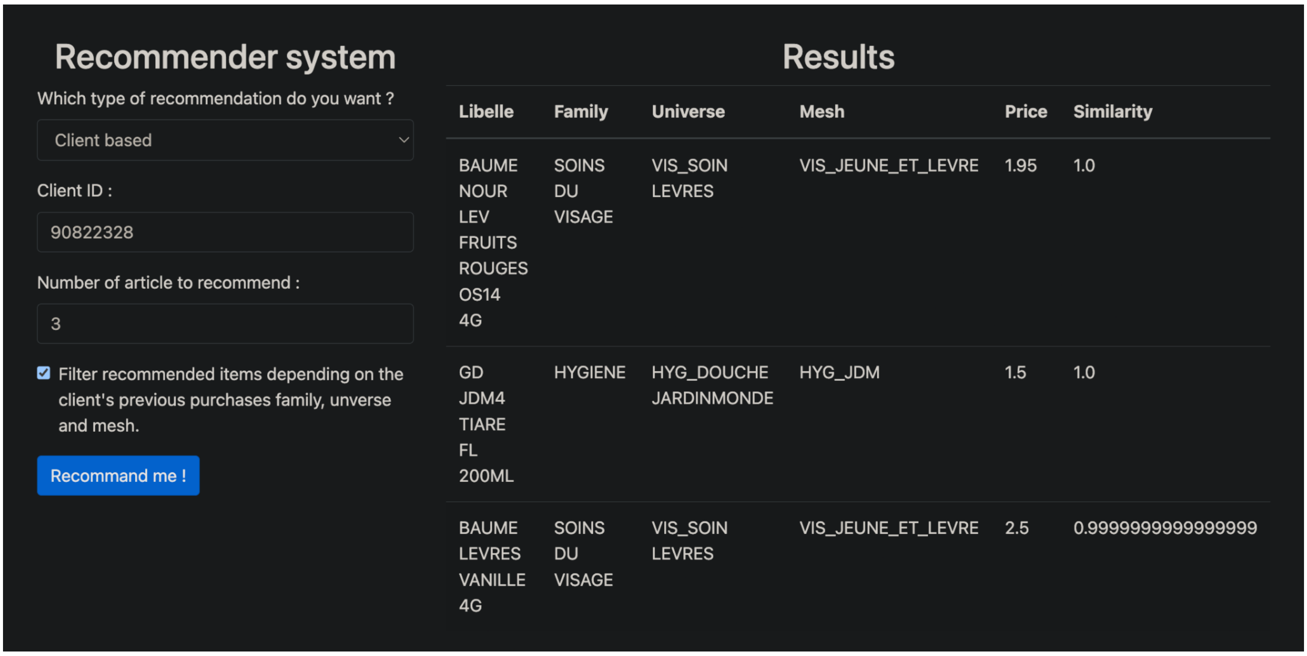
3

☐ Filter recommended items depending on the client's previous purchases family, unverse and mesh.

Results

Libelle	Family	Universe	Mesh	Price	Similarity
BAUME NOUR LEV FRUITS ROUGES OS14 4G	SOINS DU VISAGE	VIS_SOIN LEVRES	VIS_JEUNE_ET_LEVRE	1.95	1.0
SOIN 3en1 SENSITIVE YRmen 50ml	SOINS DU VISAGE	VIS_SOIN HOMMES	VIS_HOMMES	11.9	1.0
GD JDM4 TIARE FL 200ML	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.5	1.0

On this second example, we ask to recommend the same client and the same number of items, but this time we ask the program to check for item generally purchased on the same time of the year, and with the preferred family, universe and mesh of the given client :



Recommender system

Which type of recommendation do you want ?

Client based

Client ID :

90822328

Number of article to recommend :

3

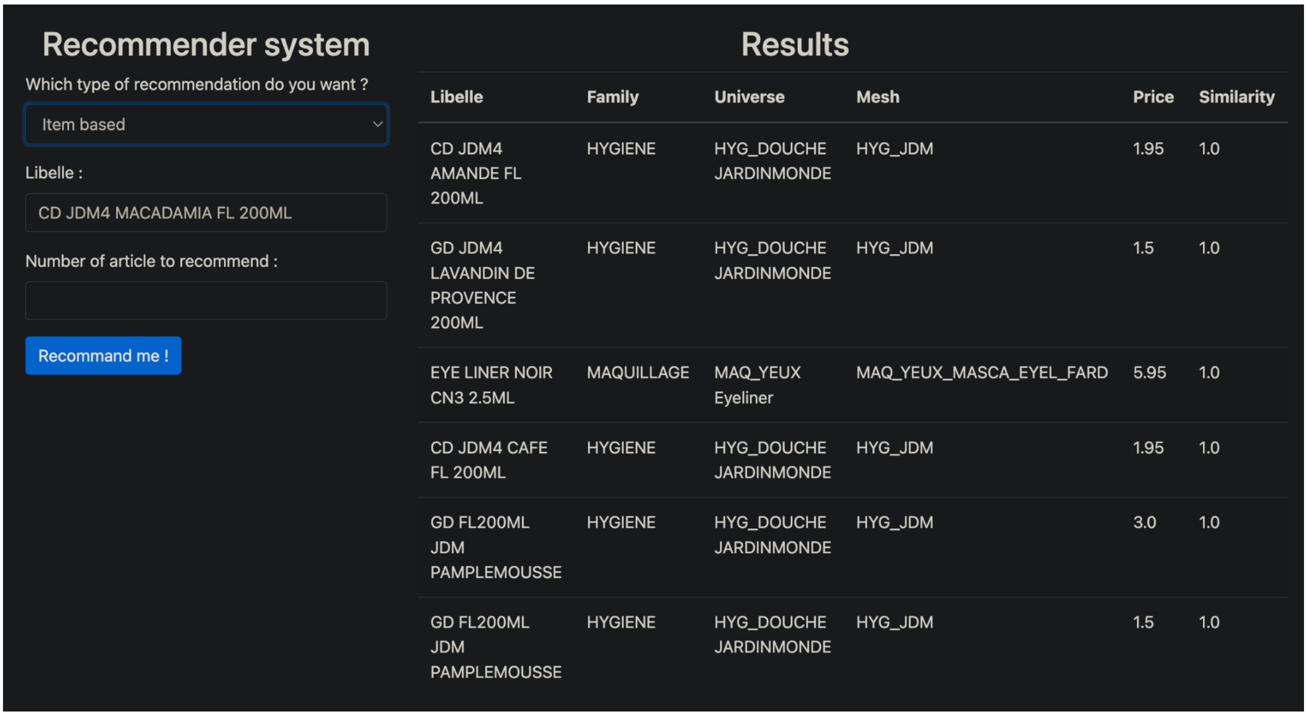
☑ Filter recommended items depending on the client's previous purchases family, unverse and mesh.

Results

Libelle	Family	Universe	Mesh	Price	Similarity
BAUME NOUR LEV FRUITS ROUGES OS14 4G	SOINS DU VISAGE	VIS_SOIN LEVRES	VIS_JEUNE_ET_LEVRE	1.95	1.0
GD JDM4 TIARE FL 200ML	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.5	1.0
BAUME LEVRES VANILLE 4G	SOINS DU VISAGE	VIS_SOIN LEVRES	VIS_JEUNE_ET_LEVRE	2.5	0.9999999999999999

In this second case, the result is also sorted by a combination of the best match and the lowest price.

In this last example, we recommend articles based on a item input, in this case “CD JDM4 MACADAMIA FL 200ML” :



Recommender system

Which type of recommendation do you want ?

Item based

Libelle :

CD JDM4 MACADAMIA FL 200ML

Number of article to recommend :

Results

Libelle	Family	Universe	Mesh	Price	Similarity
CD JDM4 AMANDE FL 200ML	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.95	1.0
GD JDM4 LAVANDIN DE PROVENCE 200ML	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.5	1.0
EYE LINER NOIR CN3 2.5ML	MAQUILLAGE	MAQ_YEUX Eyeliner	MAQ_YEUX_MASCA_EYEL_FARD	5.95	1.0
CD JDM4 CAFE FL 200ML	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.95	1.0
GD FL200ML JDM PAMPLEMOUSSE	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	3.0	1.0
GD FL200ML JDM PAMPLEMOUSSE	HYGIENE	HYG_DOUCHE JARDINMONDE	HYG_JDM	1.5	1.0

Note that if we hadn’t specified a number of articles to recommend, the solution would recommend a single article by default.

In conclusion of the recommendation system we made, the algorithm was a valuable addition to the customer data analysis and helped to provide insights into the customer preferences and purchasing behavior. The recommender system will be a valuable tool for the company in their future marketing and sales efforts, and will help to drive customer engagement and customer satisfaction.

Conclusion

In conclusion, this project has successfully analyzed the customer data of the company, providing valuable insights into their customer profiles through the implementation of a customer segmentation algorithm. The use of metabase as a data visualization tool has made it easier to visualize and interpret the data, providing a clear and concise representation of the customer segments. The implementation of a recommender system using the cosine similarity algorithm has also proven to be effective in offering personalized gifts to each customer based on their preferences.

This project has demonstrated the importance of using data-driven approaches in understanding customer behavior and making informed business decisions. With the information gathered and the recommendations provided, the company can make targeted and effective marketing strategies, resulting in increased customer satisfaction and ultimately, a boost in revenue.

In the future, further refinement and improvement can be made to the recommender system, incorporating more data and advanced algorithms such as Pairwise distance or ACP based reinforcement learning to enhance the accuracy of the recommendations. By utilizing the power of data, the company can continue to stay ahead of the competition and remain relevant in a rapidly changing market.