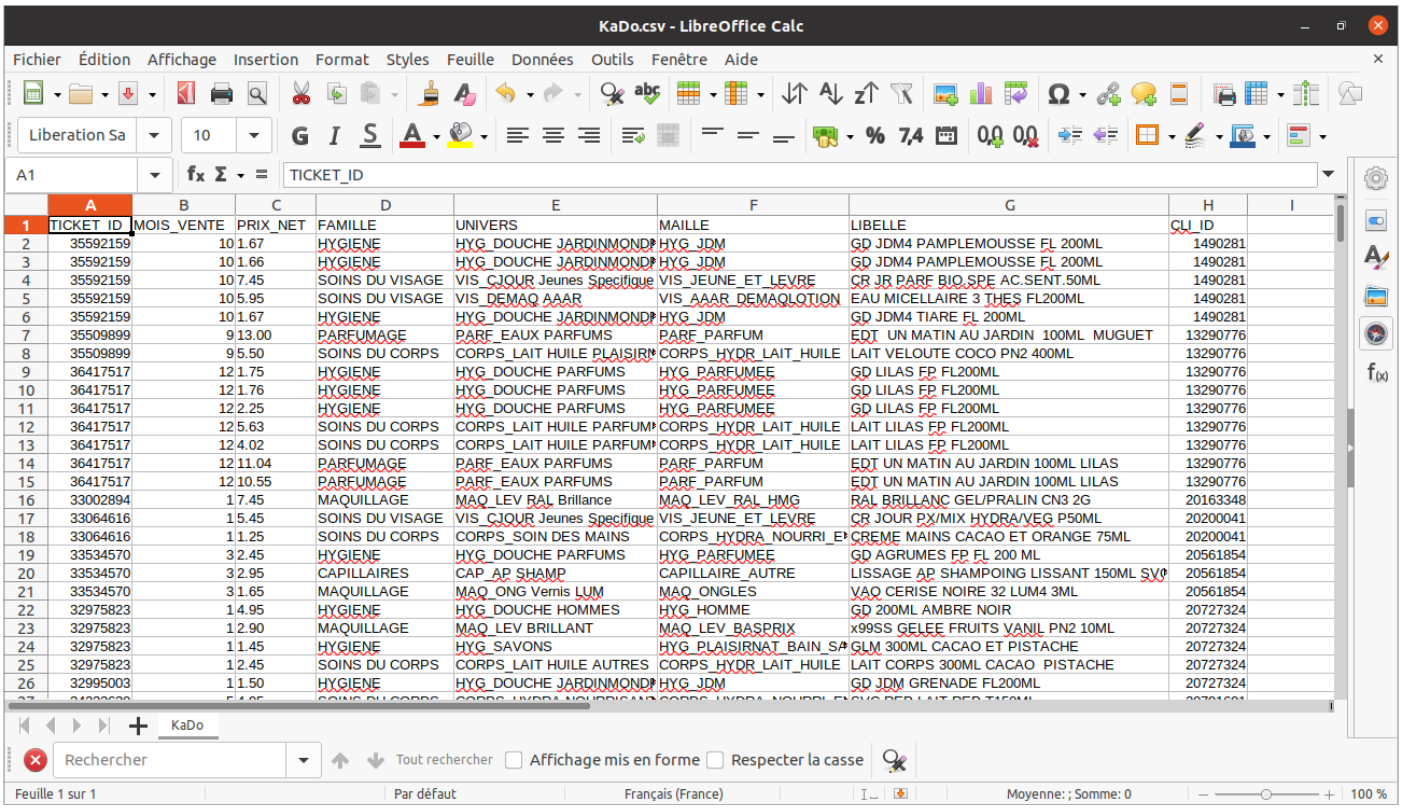
A company granted you access to KaDo: a database containing millions of bought items, divided in 3 categories. For instance, a bottle of red wine belongs to Famille: alcohol, Univers: wine, Maille: red wine.

Dataset :



TICKET_ID	MOIS_VENTE	PRIX_NET	FAMILLE	UNIVERS	MAILLE	LIBELLE	CLI_ID
35592159	10	1.67	HYGIENE	HYG_DOUCHE JARDINMONDI	HYG_JDM	GD JDM4 PAMPLEMOUSSE FL 200ML	1490281
35592159	10	1.66	HYGIENE	HYG_DOUCHE JARDINMONDI	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
35592159	10	1.67	HYGIENE	HYG_DOUCHE JARDINMONDI	HYG_JDM	GD JDM4 TIARE FL 200ML	1490281
35509899	9	13.00	PARFUMAGE	PARF_EAUX PARFUMS	PARF_PARFUM	EDT UN MATIN AU JARDIN 100ML MUGUET	13290776
35509899	9	5.50	SOINS DU CORPS	CORPS_LAIT HUILE PLAISIRN	CORPS_HYDR_LAIT_HUILE	LAIT VELOUTE COCO PN2 400ML	13290776
36417517	12	1.75	HYGIENE	HYG_DOUCHE PARFUMS	HYG_PARFUMEE	GD LILAS FP FL200ML	13290776
36417517	12	1.76	HYGIENE	HYG_DOUCHE PARFUMS	HYG_PARFUMEE	GD LILAS FP FL200ML	13290776
36417517	12	2.25	HYGIENE	HYG_DOUCHE PARFUMS	HYG_PARFUMEE	GD LILAS FP FL200ML	13290776
36417517	12	5.63	SOINS DU CORPS	CORPS_LAIT HUILE PARFUM	CORPS_HYDR_LAIT_HUILE	LAIT LILAS FP FL200ML	13290776
36417517	12	4.02	SOINS DU CORPS	CORPS_LAIT HUILE PARFUM	CORPS_HYDR_LAIT_HUILE	LAIT LILAS FP FL200ML	13290776
36417517	12	11.04	PARFUMAGE	PARF_EAUX PARFUMS	PARF_PARFUM	EDT UN MATIN AU JARDIN 100ML LILAS	13290776
36417517	12	10.55	PARFUMAGE	PARF_EAUX PARFUMS	PARF_PARFUM	EDT UN MATIN AU JARDIN 100ML LILAS	13290776
33002894	1	7.45	MAQUILLAGE	MAQ_LEV RAL Brillance	MAQ_LEV_RAL_HMG	RAL BRILLANC GEL/PRALIN CN3 2G	20163348
33064616	1	5.45	SOINS DU VISAGE	VIS_CJOUR Jeunes Specifique	VIS_JEUNE_ET_LEVRE	CR JOUR PX/MIX HYDRA/VEG P50ML	20200041
33064616	1	1.25	SOINS DU CORPS	CORPS_SOIN DES MAINS	CORPS_HYDRA_NOURRI_E	CREME MAINS CACAO ET ORANGE 75ML	20200041
33534570	3	2.45	HYGIENE	HYG_DOUCHE PARFUMS	HYG_PARFUMEE	GD AGRUMES FP FL 200 ML	20561854
33534570	3	2.95	CAPILLAIRES	CAP_AP SHAMP	CAPILLAIRE_AUTRE	LISSAGE AP SHAMPOING LISSANT 150ML SV	20561854
33534570	3	1.65	MAQUILLAGE	MAQ_ONG Vernis LUM	MAQ_ONGLES	VAO CERISE NOIRE 32 LUM4 3ML	20561854
32975823	1	4.95	HYGIENE	HYG_DOUCHE HOMMES	HYG_HOMME	GD 200ML AMBRE NOIR	20727324
32975823	1	2.90	MAQUILLAGE	MAQ_LEV BRILLANT	MAQ_LEV_BASPRIX	x99SS GELEE FRUITS VANIL PN2 10ML	20727324
32975823	1	1.45	HYGIENE	HYG_SAVONS	HYG_PLAISIRNAT_BAIN_SA	GLM 300ML CACAO ET PISTACHE	20727324
32975823	1	2.45	SOINS DU CORPS	CORPS_LAIT HUILE AUTRES	CORPS_HYDR_LAIT_HUILE	LAIT CORPS 300ML CACAO PISTACHE	20727324
32995003	1	1.50	HYGIENE	HYG_DOUCHE JARDINMONDI	HYG_JDM	GD JDM GRENADE FL200ML	20727324

The dataset contains more than 700,000 rows, illustrating all customer transactions in the company.

In order to provide a quick solution to show statistics on the global information, the development team proposes the PowerBi.

PowerBi, allows us from a CSV file, to have several dynamic and interactive graphs in order to have all the information necessary for the management of data with important quantity

PowerBi :

T-DAT-901

853,51K
Nombre de clients

7,25M
Nombre de produits vendus

2,73M
Nombre de commandes

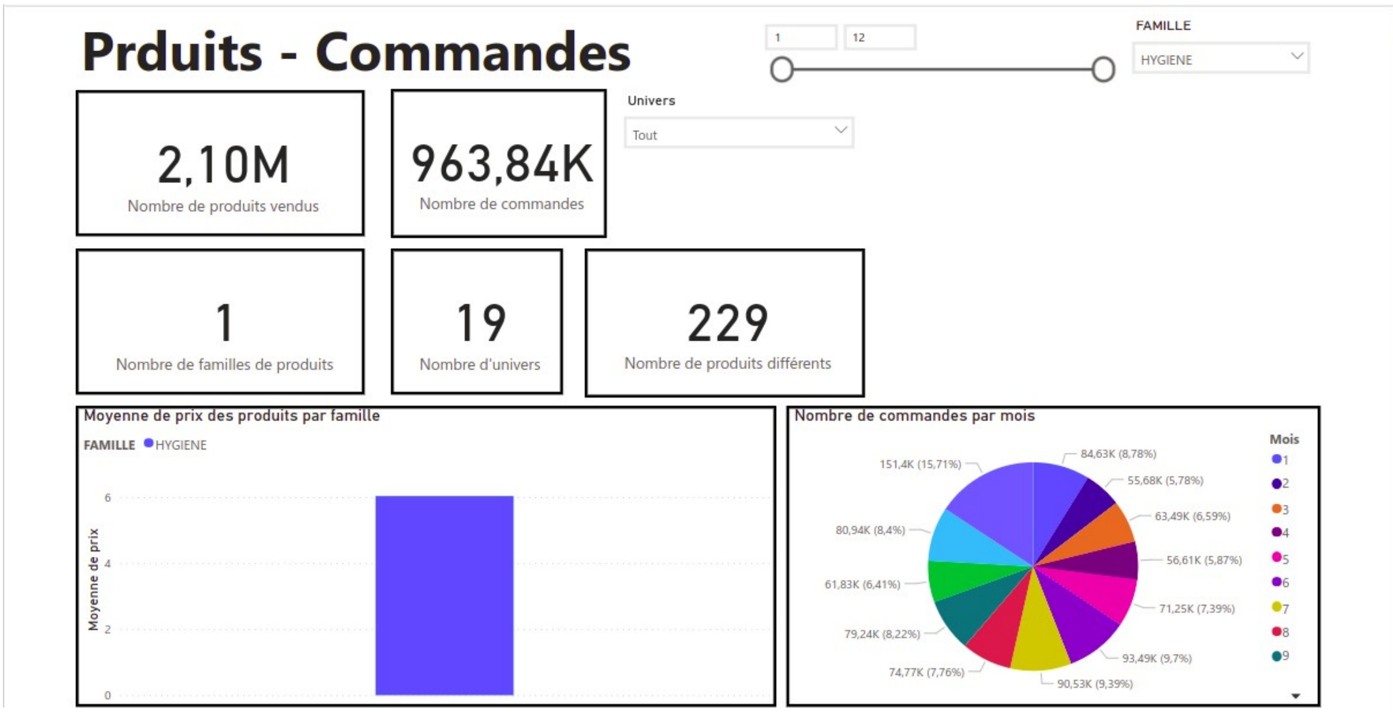
9
Nombre de familles de produits

105
Nombre d'univers

1484
Nombre de produits différents

5,97
Moyenne du prix d'un produit vendu


Clients
1
12
FAMILLE
Tout
853,51K
Nombre de clients
Nombre de clients distincts par mois
148,01K (6,42%)
257,42K (11,17%)
166,73K (7,24%)
228,08K (9,9%)
169,14K (7,34%)
170,45K (7,4%)
213,54K (9,27%)
176,44K (7,66%)
211,94K (9,2%)
179,31K (7,78%)
183,95K (7,98%)
198,77K (8,63%)
Mois
12
7
6
1
9
11
5
8
10
3
Nombre de produits vendus par familles
0,15M (2,12%)
0,63M (8,67%)
0,88M (12,13%)
2,1M (28,92%)
1,49M (20,62%)
1,69M (23,39%)
Familles
HYGIENE
MAQUILLAGE
SOINS DU VISAGE
SOINS DU CORPS
PARFUMAGE
Moyenne de prix d'achat d'un produit par mois
Moyenne de prix d'un produit
Mois
Prix d'un panier moyen par mois
Moyenne de prix d'une commande
Mois

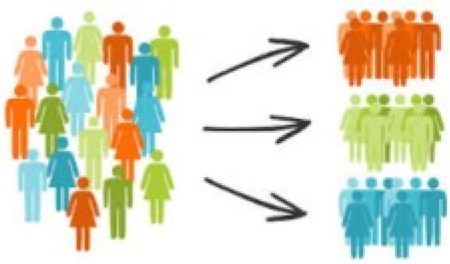

Prduits - Commandes
1
12
FAMILLE
HYGIENE
Univers
Tout
2,10M
Nombre de produits vendus
963,84K
Nombre de commandes
1
Nombre de familles de produits
19
Nombre d'univers
229
Nombre de produits différents
Moyenne de prix des produits par famille
FAMILLE HYGIENE
Moyenne de prix
Nombre de commandes par mois
151,4K (15,71%)
84,63K (8,78%)
55,68K (5,78%)
63,49K (6,59%)
80,94K (8,4%)
56,61K (5,87%)
61,83K (6,41%)
71,25K (7,39%)
79,24K (8,22%)
93,49K (9,7%)
74,77K (7,76%)
90,53K (9,39%)
Mois
1
2
3
4
5
6
7
8
9

Recommender System :

There are many ways to build recommender systems for ratings-based data, such as movies and songs. The problem with rating-based models is that they cannot be easily normalized for data with unscaled target values, such as purchase or frequency data. For example, scores typically range from 0 to 5 or 0 to 10 for songs and movies.

The goal is to create a system to recommend products to customers using purchase data with Python and the Turicreate machine learning module. These steps include:

- Data transformation and normalization
- Training of the models
- Evaluation of model performance
- Selection of the optimal model



When a customer taps for the first time on the "order" page, we can recommend the first 10 items to add to their cart, for example, hygiene products, vintages, etc.

The tool will also be able to search a list of recommendations based on a given user, for example:

- Entry: customer ID
- Return: ranked list of items (product IDs) that the user is most likely to want to put in their (empty) "shopping cart".

I. Import modules

- pandas for data manipulation
- turicreate for model selection and evaluation
- sklearn for splitting data into training and test sets.

II. Load data

III. Prepare the data

Our goal here is to break down each item list in the product column into rows and count the number of products purchased by a user.

III. 1 Create data with user, item and target fields

- This table will serve as input for our later modeling.
- In this case, our user is customerId, productId, and purchase_count.

III.2 Create dummy data

- Dummy to mark whether a customer has purchased this item or not.
- If someone buys an item, then purchase_dummy is marked as 1.
- Normalizing the number of purchases, for example for each user, would not work because customers may have different purchase frequencies and different tastes. However, we can normalize the items by purchase frequency for all users

III.3. Normalize item values across users

To do this, we normalize the purchase frequency of each item across users by first creating a user-item matrix as follows

In this step, we normalized their purchase history from 0 to 1 (1 being the highest number of purchases for an item and 0 being the zero number of purchases for that item).

IV. Separation of Training and Test Sets

- Splitting the data into training and test sets is an important part
- We use an 80:20 ratio for our training and test set size.
- The training part will be used to develop a predictive model, while the other part will be used to evaluate the performance of the model.

Now that we have three data sets with purchase accounts, dummy purchases, and scaled purchase accounts, we would like to split them for modeling.

V. Define the models using the Turicreate library

Before running a more complicated approach like collaborative filtering, we need to run a base model to compare and evaluate the models. We will use the popularity model.

A common approach to predicting purchased items is collaborative filtering. Let's first define our variables to use in the models:

Turicreate made it super easy for us to call a modeling technique, so let's define our function for all models as follows:

VI. Popularity model as baseline

- The popularity model takes the most popular items for recommendation. These items are products with the highest number of sales among customers.
- The training data is used for model selection

VII.1. Similarity in cosine

The similarity is the cosine of the angle between the 2 vectors of the item vectors of A and B
The closer the vectors are, the smaller the angle and the larger the cosine.

It allows to quantify the similarity between 2 entities

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

VII.2. Pearson similarity

- The similarity is the Pearson coefficient between the two vectors.

The Pearson correlation coefficient, is an alternative method to normalize the count of common neighbors. This method compares the number of common neighbors to the expected value in a network where the vertices are randomly connected. This value is strictly between -1 and 1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

VIII. Model evaluation

To evaluate recommendation engines, we can use the concept of RMSE and precision-recall. Several evaluation criteria:

RMSE (Root Mean Squared Errors)

- Measures the error of predicted values
- The lower the RMSE value, the better the recommendations.

Recall

- What is the percentage of products a user buys that are actually recommended?
- If a customer buys 5 products and the recommendation decides to show 3, the recall is 0.6.

Accuracy

- If the customer was recommended 5 products and bought 4, the precision is 0.8.

Why are recall and precision both important?

- Let's take the case where we recommend all products, so our customers will surely cover the items they liked and purchased. In this case, we have a 100% recall rate!
- We need to consider accuracy. If we recommend 300 items but the user only likes and buys 3 of them, the precision is 0.1%! This very low precision indicates that the model is not excellent, despite its excellent recall.
- So our goal should be to optimize both recall and precision (to be as close to 1 as possible).

Popularity Model | Cosine Similarity | Pearson Similarity

Purchase counts

Popularity Model			Cosine Similarity			Pearson Similarity		
cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall
1	0.0007199424046076297	0.00029346223730672953	1	0.06335493160547198	0.03506211424106219	1	0.06357091432685413	0.035144393373017405
2	0.003167746580273574	0.0031647468202543752	2	0.06263498920086356	0.07230791509969706	2	0.0624910007199426	0.07222392181915947
3	0.0034317254619630373	0.005225953352303244	3	0.05097192224622066	0.08726852084863941	3	0.051091912646988306	0.08727520602811081
4	0.0029157667386609147	0.005793716226204202	4	0.04321454283657309	0.09728549302056154	4	0.043178545716342755	0.0968562746107815
5	0.006133909287257029	0.01590833513184247	5	0.03832973362131041	0.10678978981967435	5	0.03832973362131065	0.10616781100655155
6	0.00643148548116152	0.020268614880891146	6	0.034641228701704024	0.1143935243780528	6	0.034569234461243395	0.11430058895574437
7	0.0059035277177825855	0.021857490368254646	7	0.032150570811478006	0.12289104457643715	7	0.03175974493469068	0.12174051064512559
8	0.0055795536357091365	0.023666719655885578	8	0.030192584593232753	0.1315870411077297	8	0.030039596832253546	0.13105989027280296
9	0.005391568674506037	0.02583266066631922	9	0.02850971922246227	0.13891659760035277	9	0.028413726901847833	0.13911332731652126
10	0.005363570914326852	0.028737828252912132	10	0.027048236141108645	0.1465656410215347	10	0.027041036717062584	0.14659112729432097

Purchase dummy

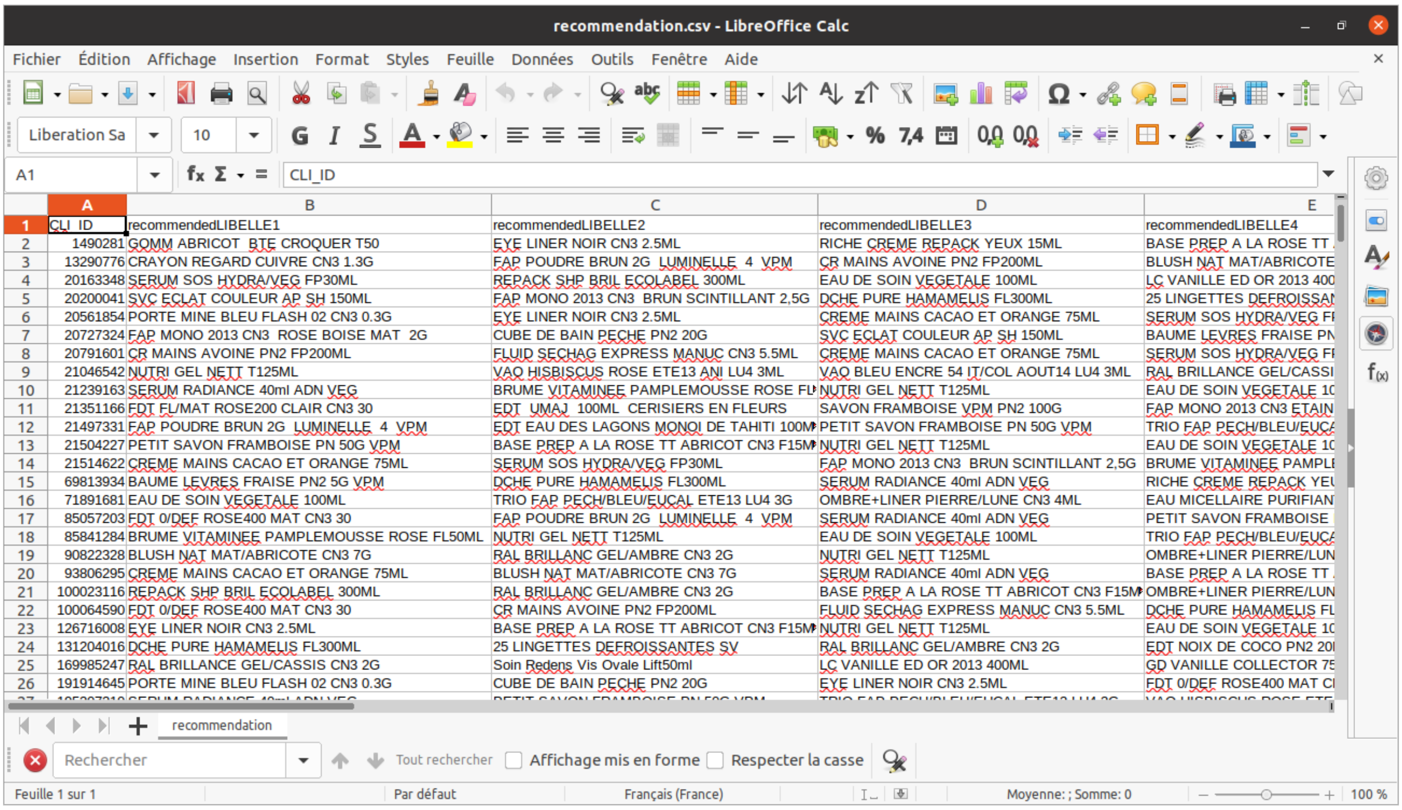
Popularity Model			Cosine Similarity			Pearson Similarity		
cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall
1	0.05430644350262212	0.030314648895390185	1	0.0549529487824151	0.03061267414672341	1	0.054881114862438046	0.03059300533530103
2	0.054521945262552975	0.06031738893404705	2	0.05448602830256436	0.06034552388603806	2	0.054737447022484356	0.06055703487263677
3	0.04575820702535739	0.07481887179538214	3	0.045758207025357586	0.07498768562309908	3	0.04575820702535712	0.07499666486309599
4	0.03837727174771911	0.08234835646340455	4	0.037784641907909054	0.0815998700808042	4	0.03778464190790865	0.08163578704079291
5	0.03409237842109061	0.0905290880664835	5	0.03356080741326061	0.09043569918636968	5	0.03364700811723305	0.09059373381031925
6	0.03139142302995476	0.09898045781970626	6	0.03131958910997786	0.09997398456210141	6	0.03131958910997744	0.09999963953352173
7	0.029534003099120503	0.108245485128516	7	0.029585313041961357	0.10916704060613812	7	0.029605837019097744	0.10925854333753758
8	0.027835643991092555	0.11626777864484733	8	0.02803318727102923	0.11750801162579602	8	0.028140938150995002	0.11805506113304967
9	0.026482771831525472	0.12401884389627199	9	0.026913775351387562	0.12632150625478447	9	0.026953683084708142	0.1265609526547087
10	0.025623159255800585	0.1343114499883889	10	0.025759643703756923	0.13419136827664874	10	0.025716543351770624	0.13383989516818834

Scaled counts

Popularity Model			Cosine Similarity			Pearson Similarity		
cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall	cutoff	mean_precision	mean_recall
1	0.022683084899546416	0.011808641269330261	1	0.022755094692878064	0.01187224992010684	1	0.023475192626197055	0.012186231083083403
2	0.03229639230935397	0.037299229209449225	2	0.03175631885936494	0.03654672686913105	2	0.03208036292935828	0.03676344205668188
3	0.029524015266076113	0.04928467481446728	3	0.029692038117184297	0.04931707922146633	3	0.02993207076162342	0.04983143488812296
4	0.028605890401094437	0.06270341369040609	4	0.028731907539425724	0.06293024453940188	4	0.028983941816086736	0.06361982403649855
5	0.02963923093540727	0.08210212724012415	5	0.02978325052207093	0.0823035546342493	5	0.029783250522070984	0.082266120971843
6	0.028767912436091326	0.09633992744649825	6	0.028959938551643184	0.09685699776806231	6	0.028827920597201143	0.09634521387973909
7	0.026818504459463634	0.10393507431392464	7	0.0269728111594606	0.1042201988051407	7	0.026911088479461782	0.10388903948175872
8	0.024960394613667346	0.10991651636147132	8	0.02511341542449774	0.11034237427814785	8	0.024996399510333337	0.10985690825476863
9	0.023707224182488874	0.11704257121921632	9	0.02384324268100469	0.11738050289221046	9	0.023771232887672755	0.11705134384086605
10	0.022726290775545558	0.12475205087728285	10	0.022877511341542388	0.12512255581140802	10	0.02281270252754377	0.12477339403969312

IX. Final output

A csv file and then send it to the web application POWERBI to visualize its data.
We can also have the results for only one custommer with his ID.



	A	B	C	D	E
1	CLI_ID	recommendedLIBELLE1	recommendedLIBELLE2	recommendedLIBELLE3	recommendedLIBELLE4
2	1490281	GOMM ABRICOT BTE CROQUER T50	EYE LINER NOIR CN3 2.5ML	RICHE CREME REPACK YEUX 15ML	BASE PREP A LA ROSE TT
3	13290776	CRAYON REGARD CUIVRE CN3 1.3G	FAP POUDRE BRUN 2G LUMINELLE 4 VPM	CR MAINS AVOINE PN2 FP200ML	BLUSH NAT MAT/ABRICOTE
4	20163348	SERUM SOS HYDRA/VEG FP30ML	REPACK SHP BRIL ECOLABEL 300ML	EAU DE SOIN VEGETALE 100ML	LC VANILLE ED OR 2013 400
5	20200041	SVC ECLAT COULEUR AP SH 150ML	FAP MONO 2013 CN3 BRUN SCINTILLANT 2,5G	DCHE PURE HAMAMELIS FL300ML	25 LINGETTES DEFROISSAN
6	20561854	PORTE MINE BLEU FLASH 02 CN3 0.3G	EYE LINER NOIR CN3 2.5ML	CREME MAINS CACAO ET ORANGE 75ML	SERUM SOS HYDRA/VEG FI
7	20727324	FAP MONO 2013 CN3 ROSE BOISE MAT 2G	CUBE DE BAIN PECHE PN2 20G	SVC ECLAT COULEUR AP SH 150ML	BAUME LEVRES FRAISE PN
8	20791601	CR MAINS AVOINE PN2 FP200ML	FLUID SECHAG EXPRESS MANUC CN3 5.5ML	CREME MAINS CACAO ET ORANGE 75ML	SERUM SOS HYDRA/VEG FI
9	21046542	NUTRI GEL NETT T125ML	VAO HISBISCUS ROSE ETE13 ANI LU4 3ML	VAO BLEU ENCRE 54 IT/COL AOUT14 LU4 3ML	RAL BRILLANCE GEL/CASSI
10	21239163	SERUM RADIANCE 40ml ADN VEG	BRUME VITAMINEE PAMPLEMOUSSE ROSE FL	NUTRI GEL NETT T125ML	EAU DE SOIN VEGETALE 10
11	21351166	FDT FL/MAT ROSE200 CLAIR CN3 30	EDT UMAJ 100ML CERISIERS EN FLEURS	SAVON FRAMBOISE VPM PN2 100G	FAP MONO 2013 CN3 ETAIN
12	21497331	FAP POUDRE BRUN 2G LUMINELLE 4 VPM	EDT EAU DES LAGONS MONOI DE TAHITI 100M	PETIT SAVON FRAMBOISE PN 50G VPM	TRIO FAP PECH/BLEU/EUCA
13	21504227	PETIT SAVON FRAMBOISE PN 50G VPM	BASE PREP A LA ROSE TT ABRICOT CN3 F15M	NUTRI GEL NETT T125ML	EAU DE SOIN VEGETALE 10
14	21514622	CREME MAINS CACAO ET ORANGE 75ML	SERUM SOS HYDRA/VEG FP30ML	FAP MONO 2013 CN3 BRUN SCINTILLANT 2,5G	BRUME VITAMINEE PAMPLE
15	69813934	BAUME LEVRES FRAISE PN2 5G VPM	DCHE PURE HAMAMELIS FL300ML	SERUM RADIANCE 40ml ADN VEG	RICHE CREME REPACK YEU
16	71891681	EAU DE SOIN VEGETALE 100ML	TRIO FAP PECH/BLEU/EUCAL ETE13 LU4 3G	OMBRE+LINER PIERRE/LUNE CN3 4ML	EAU MICELLAIRE PURIFIAN
17	85057203	FDT 0/DEF ROSE400 MAT CN3 30	FAP POUDRE BRUN 2G LUMINELLE 4 VPM	SERUM RADIANCE 40ml ADN VEG	PETIT SAVON FRAMBOISE
18	85841284	BRUME VITAMINEE PAMPLEMOUSSE ROSE FL50ML	NUTRI GEL NETT T125ML	EAU DE SOIN VEGETALE 100ML	TRIO FAP PECH/BLEU/EUCA
19	90822328	BLUSH NAT MAT/ABRICOTE CN3 7G	RAL BRILLANC GEL/AMBRE CN3 2G	NUTRI GEL NETT T125ML	OMBRE+LINER PIERRE/LUN
20	93806295	CREME MAINS CACAO ET ORANGE 75ML	BLUSH NAT MAT/ABRICOTE CN3 7G	SERUM RADIANCE 40ml ADN VEG	BASE PREP A LA ROSE TT
21	100023116	REPACK SHP BRIL ECOLABEL 300ML	RAL BRILLANC GEL/AMBRE CN3 2G	BASE PREP A LA ROSE TT ABRICOT CN3 F15M	OMBRE+LINER PIERRE/LUN
22	100064590	FDT 0/DEF ROSE400 MAT CN3 30	CR MAINS AVOINE PN2 FP200ML	FLUID SECHAG EXPRESS MANUC CN3 5.5ML	DCHE PURE HAMAMELIS FL
23	126716008	EYE LINER NOIR CN3 2.5ML	BASE PREP A LA ROSE TT ABRICOT CN3 F15M	NUTRI GEL NETT T125ML	EAU DE SOIN VEGETALE 10
24	131204016	DCHE PURE HAMAMELIS FL300ML	25 LINGETTES DEFROISSANTES SV	RAL BRILLANC GEL/AMBRE CN3 2G	EDT NOIX DE COCO PN2 20
25	169985247	RAL BRILLANCE GEL/CASSIS CN3 2G	Soin Redens Vis Ovale Lift50ml	LC VANILLE ED OR 2013 400ML	GD VANILLE COLLECTOR 75
26	191914645	PORTE MINE BLEU FLASH 02 CN3 0.3G	CUBE DE BAIN PECHE PN2 20G	EYE LINER NOIR CN3 2.5ML	FDT 0/DEF ROSE400 MAT CI