PROJET 5
Segmentez des clients d'un site e-commerce

Problématique et axes de recherche

Olist souhaite que vous fournissiez à ses équipes d'e-commerce une segmentation des clients qu'elles pourront utiliser au quotidien pour leurs campagnes de communication.

Votre objectif est de comprendre les différents types d'utilisateurs grâce à leur comportement et à leurs données personnelles.

Vous devrez fournir à l'équipe marketing une description actionnable de votre segmentation et de sa logique sous-jacente pour une utilisation optimale, ainsi qu'une proposition de contrat de maintenance basée sur une analyse de la stabilité des segments au cours du temps.

Enfin, votre client, Olist, a spécifié sa demande ainsi :

- La segmentation proposée doit être exploitable et facile d'utilisation pour l'équipe marketing.
- Vous évaluerez la fréquence à laquelle la segmentation doit être mise à jour, afin de pouvoir effectuer un devis de contrat de maintenance.
- Le code fourni doit respecter la convention PEP8, pour être utilisable par Olist.

Problématique et axes de recherche

- segmentation de type RFM
- essai de différents algorithme de clustérisations :
  - KMEANS
  - DBSCAN
  - Hiérarchique
- test de stabilité des clusters
- maintenance et rafraichissement algorithme

Analyse exploratoire des données

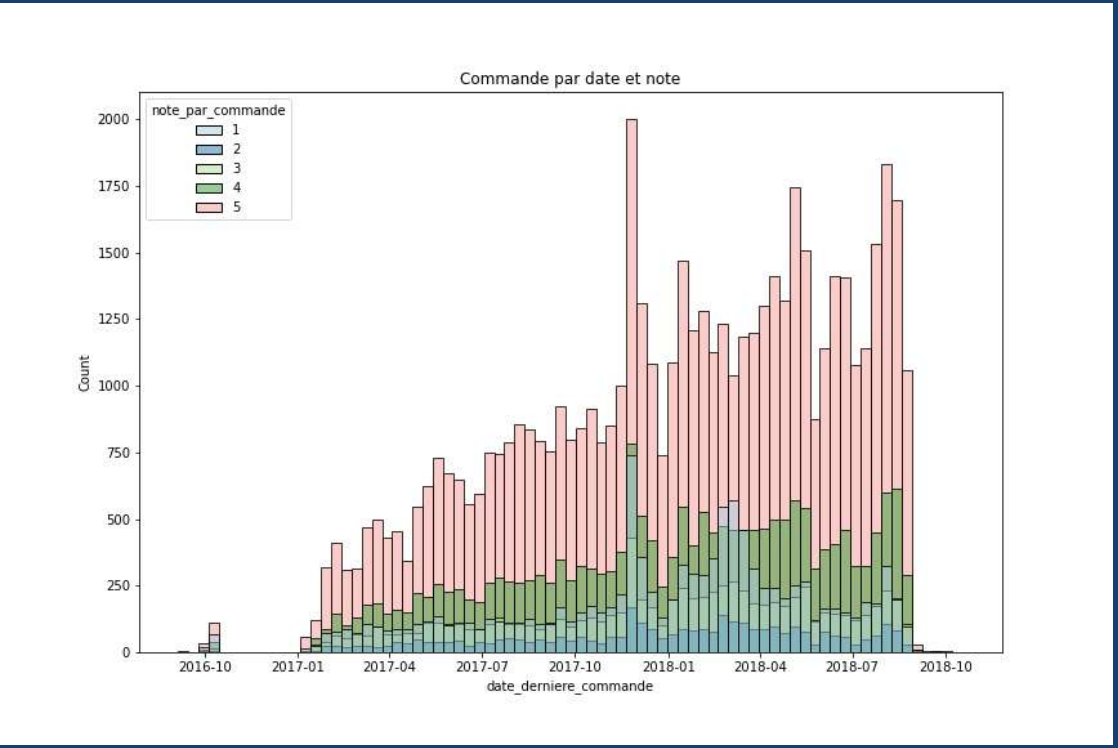
Les données

Pour cette mission, Olist vous fournit une base de données anonymisée comportant des informations sur l'historique de commandes, les produits achetés, les commentaires de satisfaction, et la localisation des clients depuis janvier 2017.

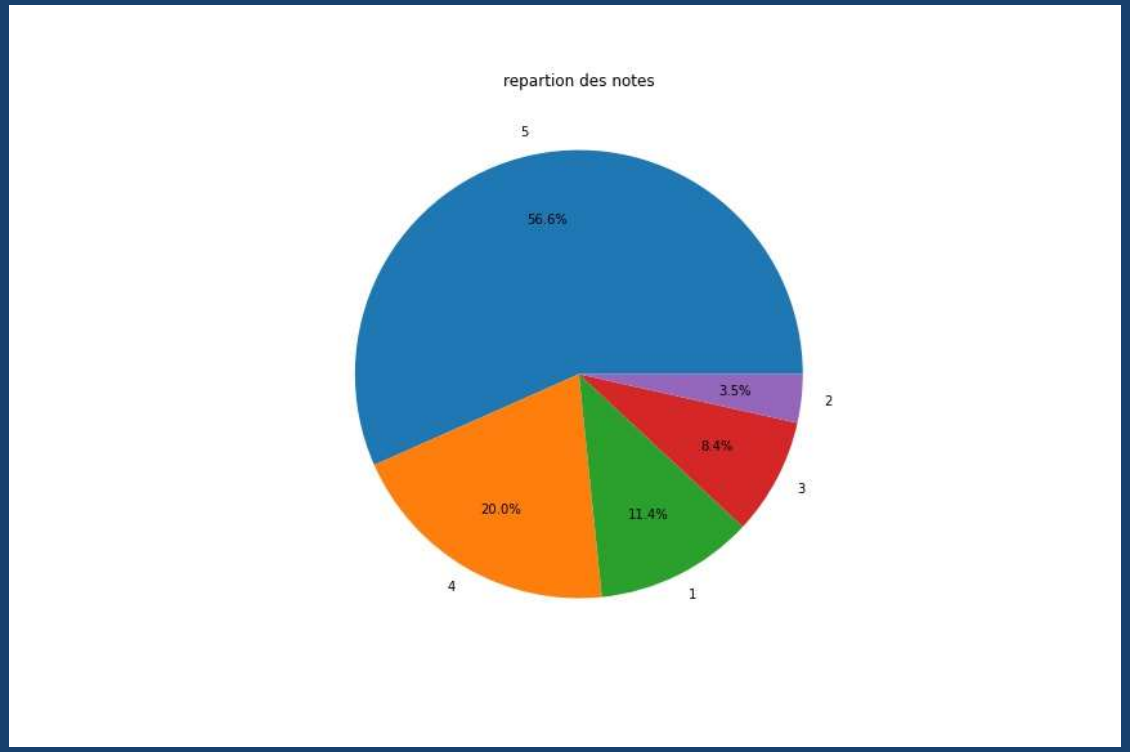
9 fichiers csv :

- Client
- Localisation
- Produit par commande
- Paiement par commande
- Note par commande
- Commande
- Produits
- Vendeurs
- Traduction produits

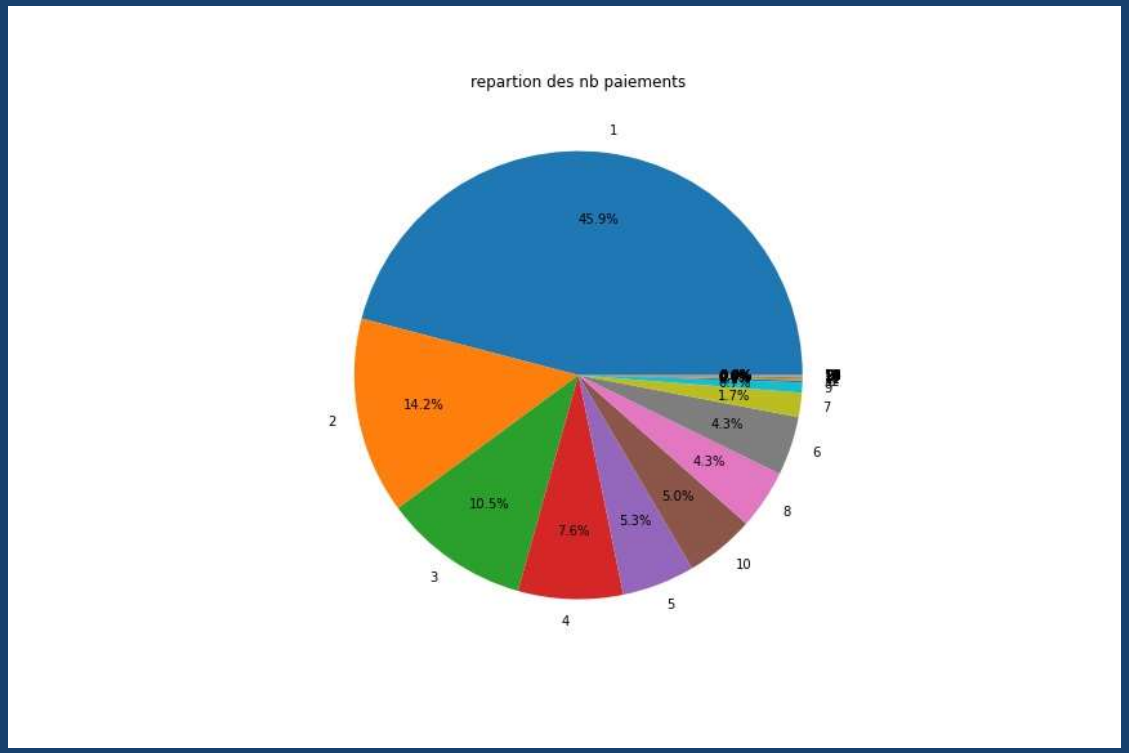
Analyse exploratoire des données



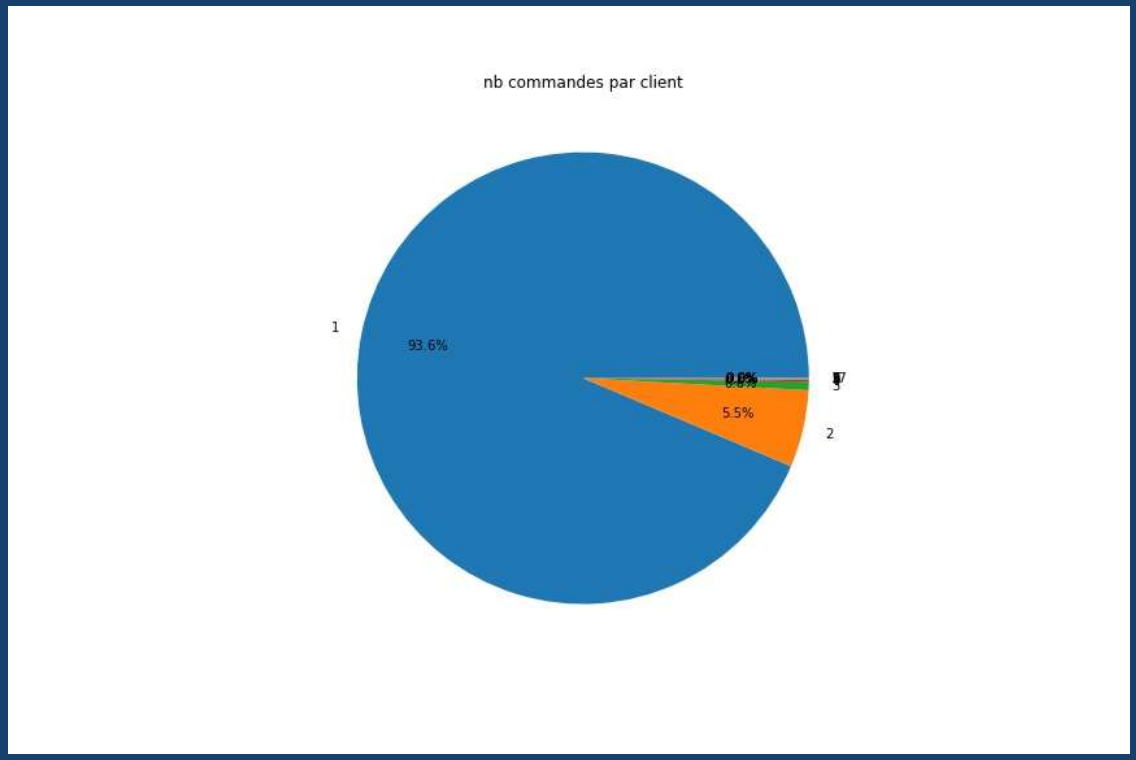
Analyse exploratoire des données



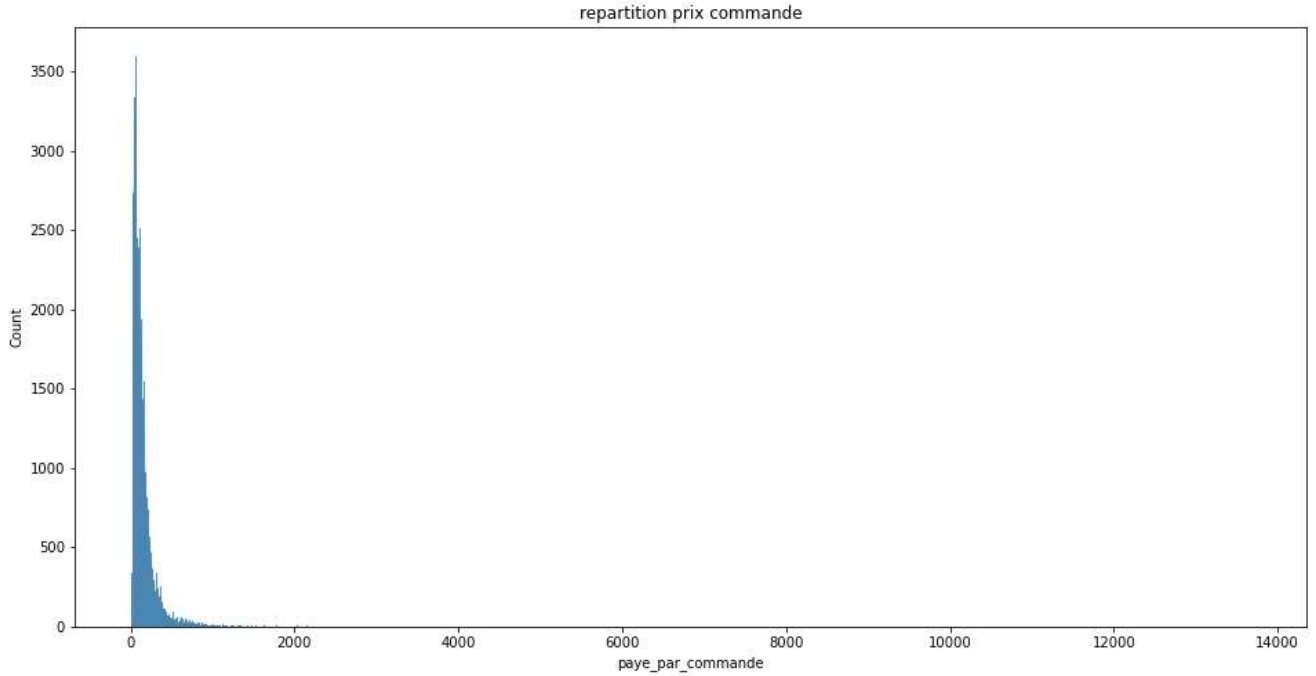
Analyse exploratoire des données



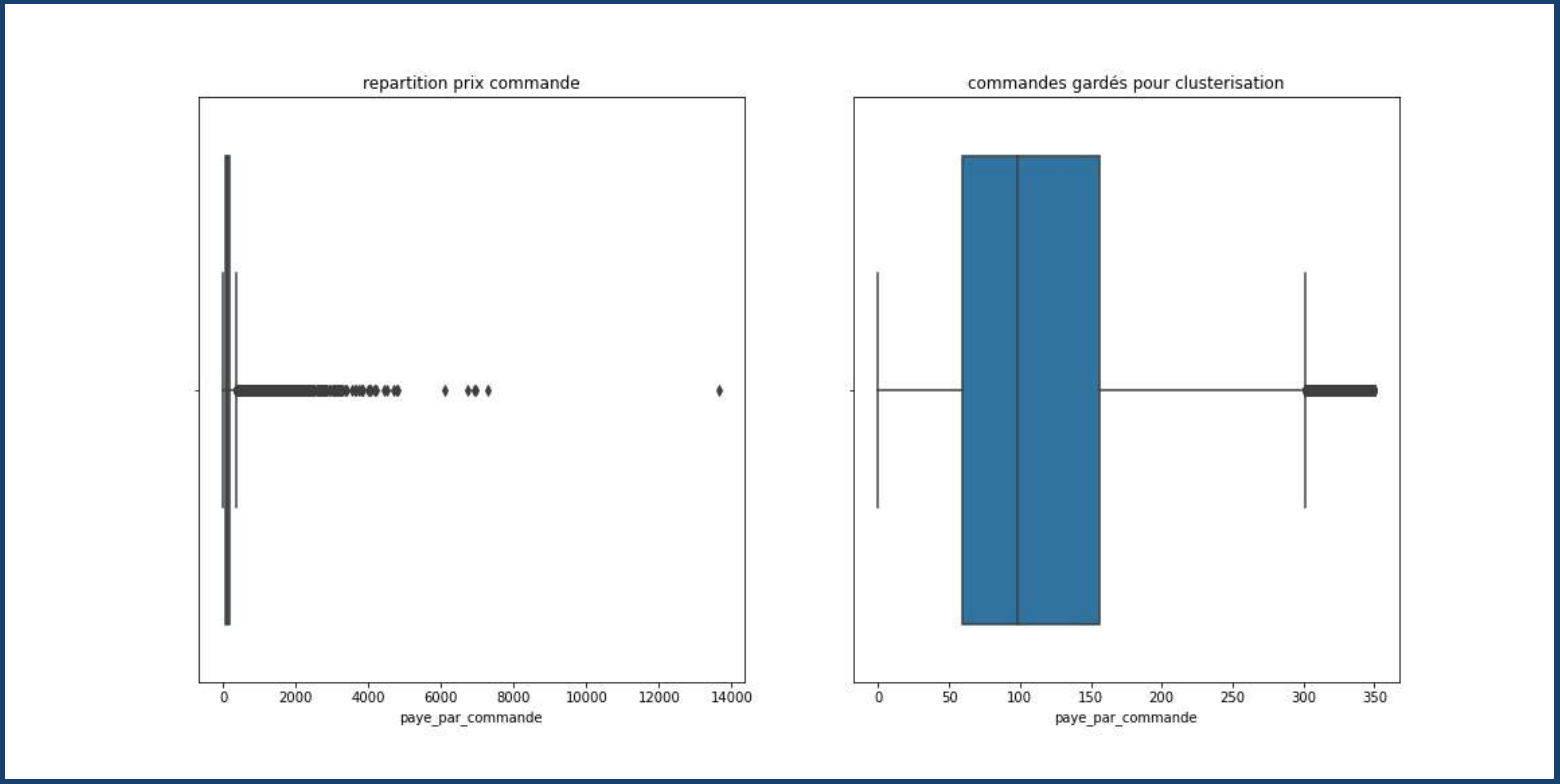
Analyse exploratoire des données



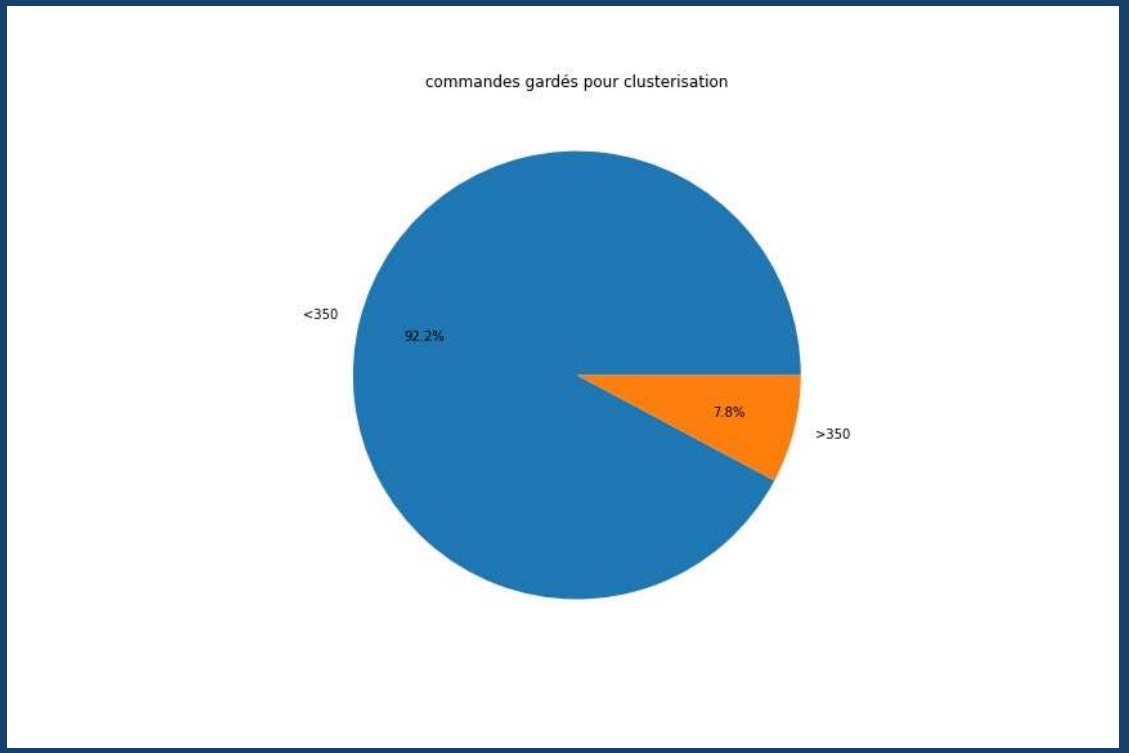
Analyse exploratoire des données



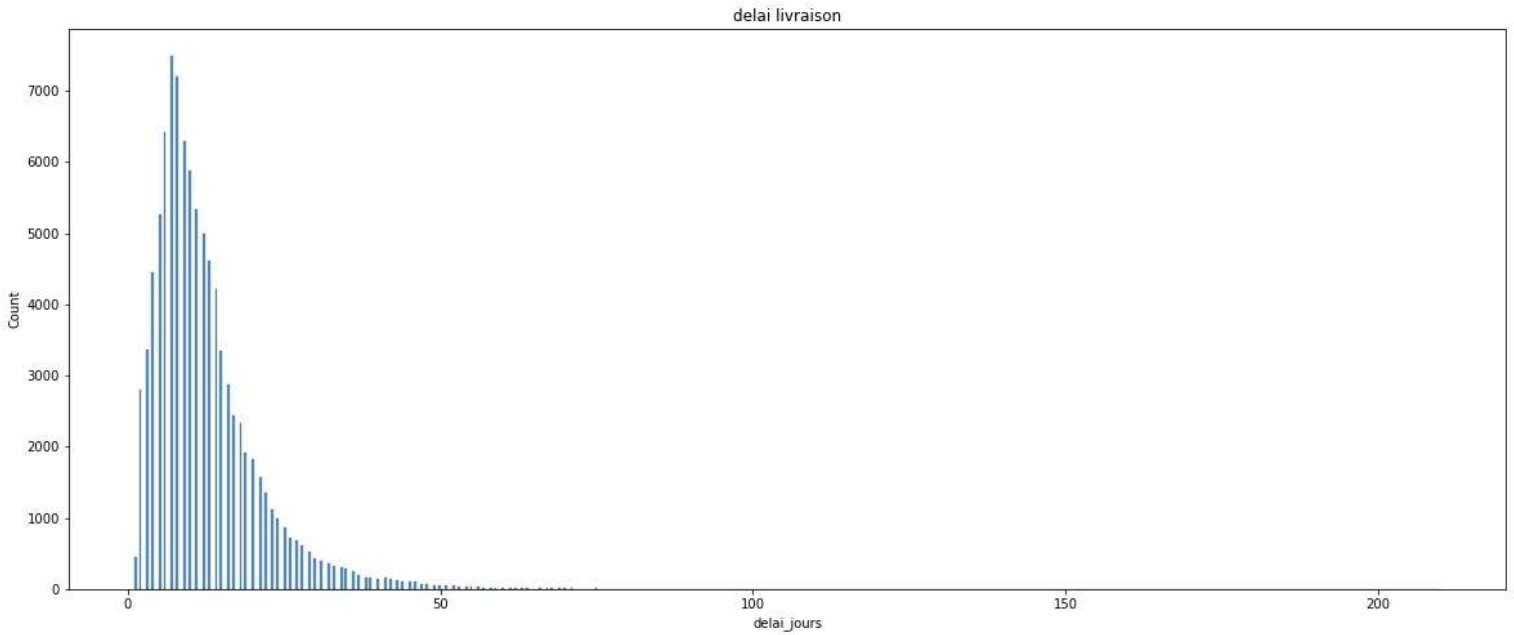
Analyse exploratoire des données



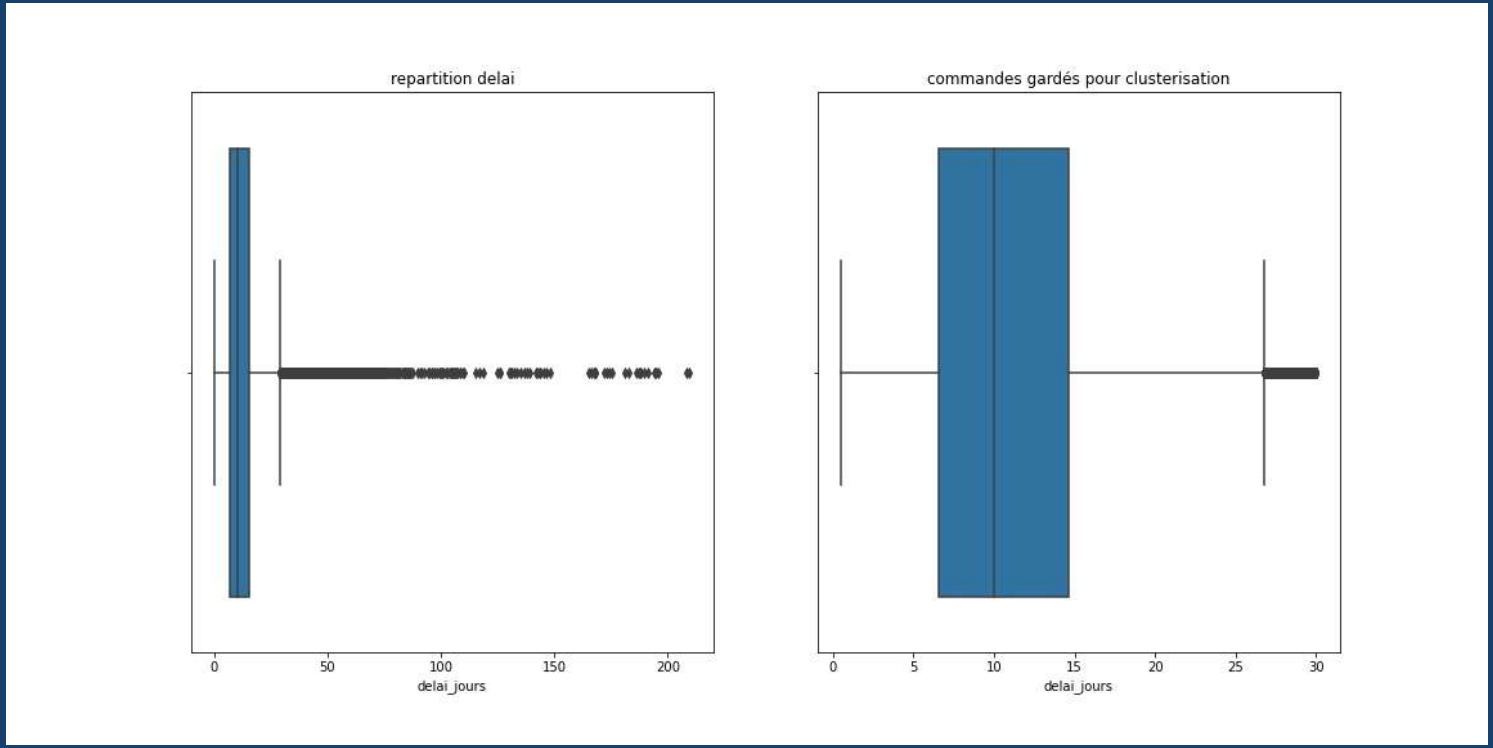
Analyse exploratoire des données



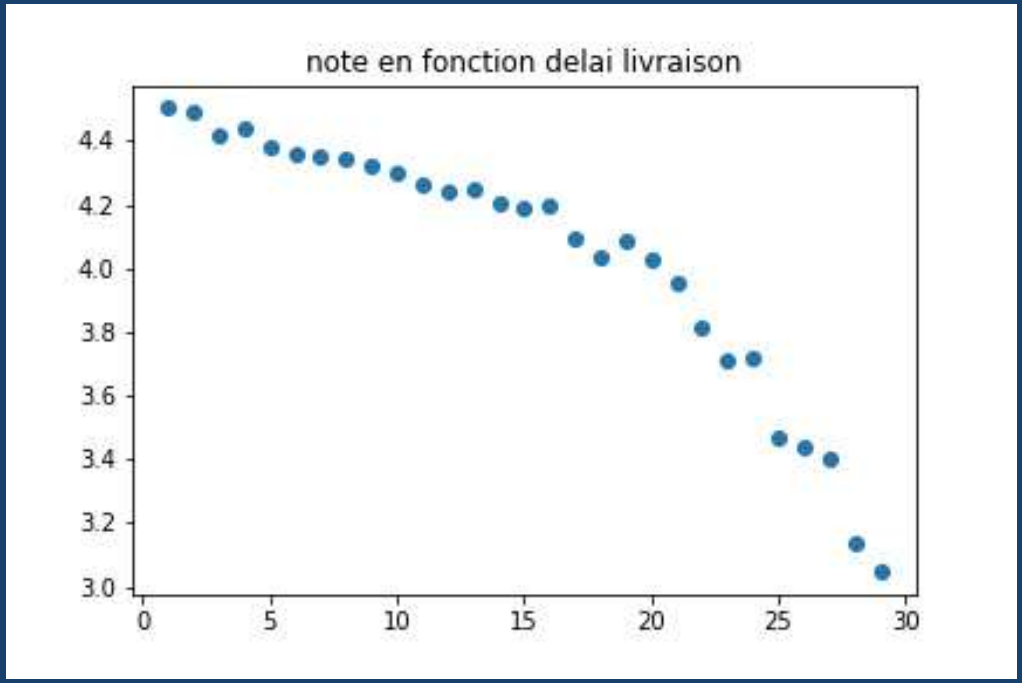
Analyse exploratoire des données

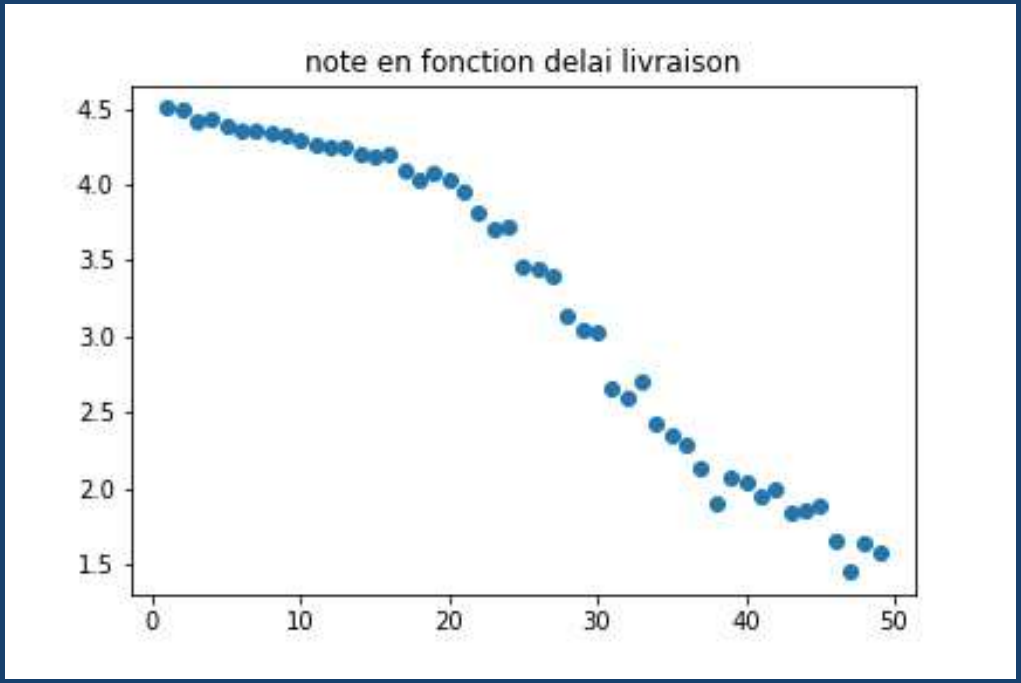


Analyse exploratoire des données



Analyse exploratoire des données





Segmentation RFM

Découpage en quartile (1,2,3,4) :

- Montant

- Récence (date commande)

Découpage à partir MinMax (1,2,3,4) :

- Fréquence (nb_commande)

Score RFM en ajoutant les trois

String RFM pour garder trace des trois

	paye_par_commande	nb_commande	date_commande	RFM_R	RFM_F	RFM_M	str_RFM	class_RFM
0	146.87	1	45.0	3	1	3	313	Depensiers
21	102.03	1	50.0	3	1	3	313	Depensiers
26	94.63	1	77.0	2	1	3	213	Depensiers
30	47.59	1	46.0	3	1	1	311	Recents à relancer
37	218.89	2	16.0	4	2	4	424	VIP
...	...	...	...	...	...	...	...	...
99420	64.42	1	156.0	1	1	2	112	Non interesses perdus
99424	102.03	1	46.0	3	1	3	313	Depensiers
99428	130.85	1	95.0	2	1	3	213	Depensiers
99430	102.03	1	37.0	3	1	3	313	Depensiers
99433	84.32	1	107.0	2	1	2	212	Presque perdus

12044 rows × 8 columns

Segmentation RFM

Regroupement à partir du str_RFM :

‘4x4’ : VIP

‘2x4’ : VIP presque perdus

‘1x4’ : VIP Perdus

‘111,112’ : Non intéressés, perdus

‘411,412’ : Nouveaux

RFM_F!=1 : Fidèles

RFM_M==3ou4 : Dépensiers

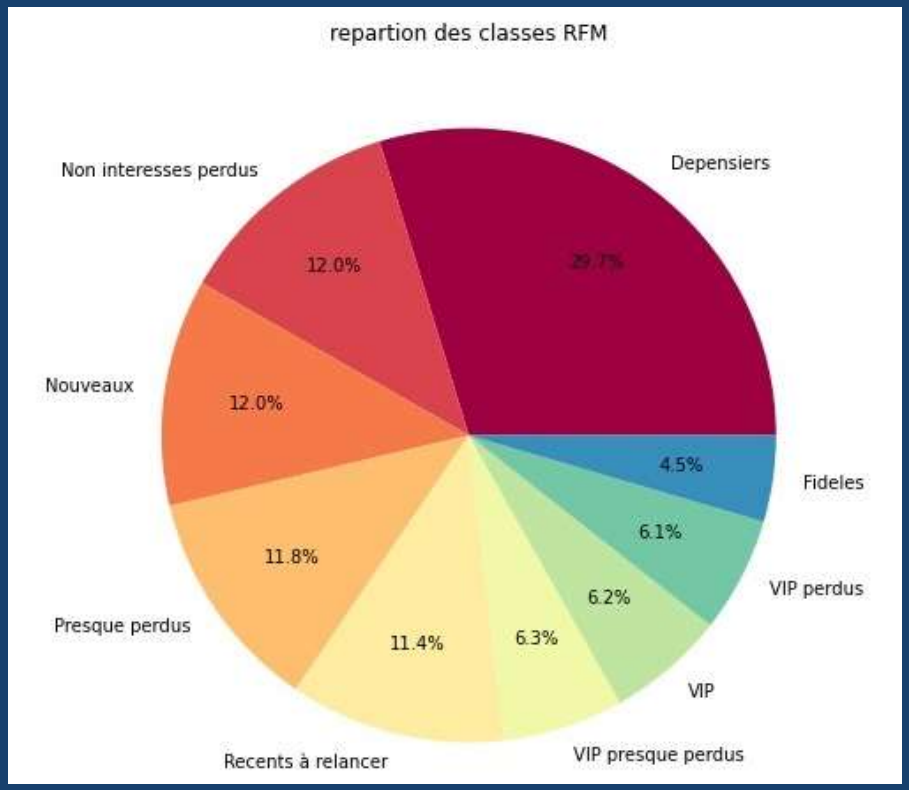
‘311, 312’ : Récents à relancer

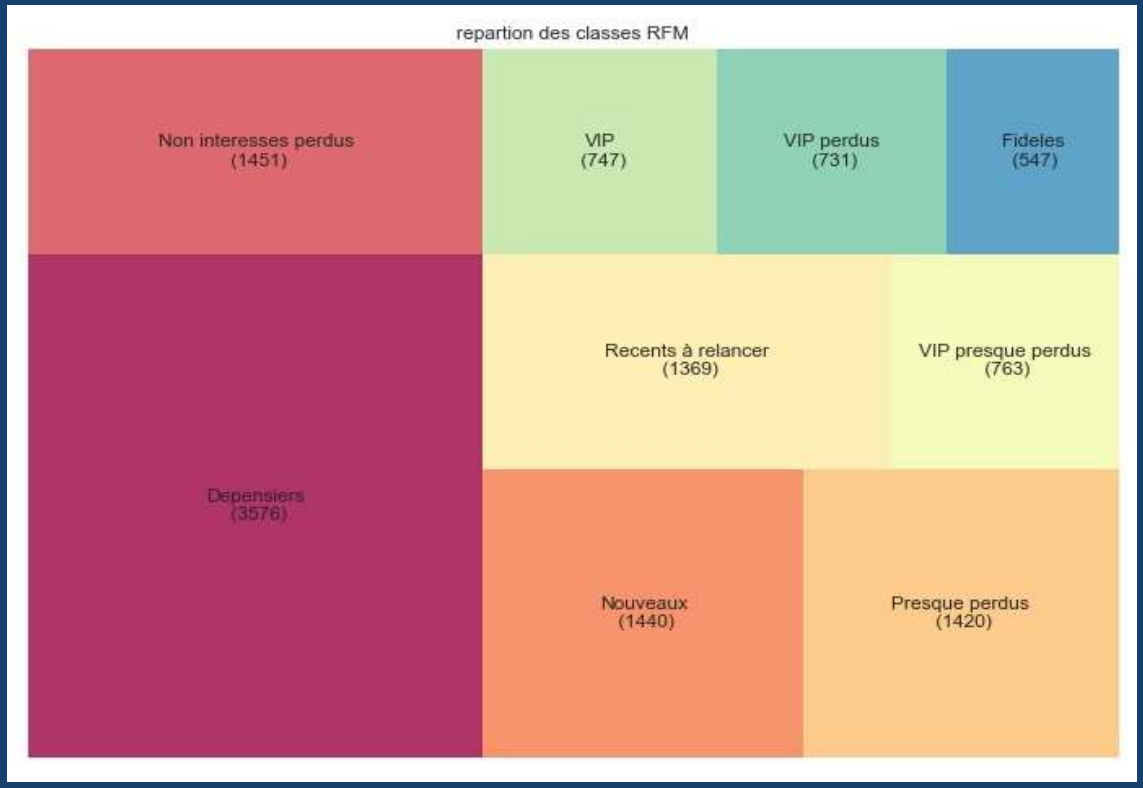
‘211, 212’ : Presque perdus

	paye_par_commande	nb_commande	date_commande	RFM_R	RFM_F	RFM_M	str_RFM	class_RFM
0	146.87	1	45.0	3	1	3	313	Depensiers
21	102.03	1	50.0	3	1	3	313	Depensiers
26	94.63	1	77.0	2	1	3	213	Depensiers
30	47.59	1	46.0	3	1	1	311	Recents à relancer
37	218.89	2	16.0	4	2	4	424	VIP
...	...	...	...	...	...	...	...	...
99420	64.42	1	156.0	1	1	2	112	Non interesses perdus
99424	102.03	1	46.0	3	1	3	313	Depensiers
99428	130.85	1	95.0	2	1	3	213	Depensiers
99430	102.03	1	37.0	3	1	3	313	Depensiers
99433	84.32	1	107.0	2	1	2	212	Presque perdus

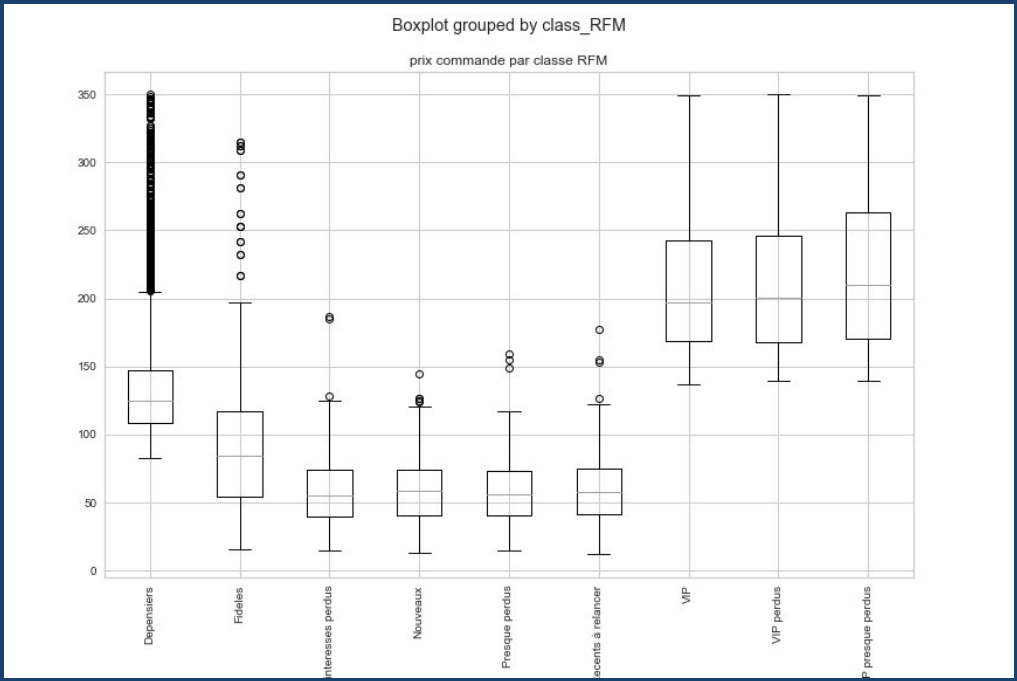
12044 rows × 8 columns

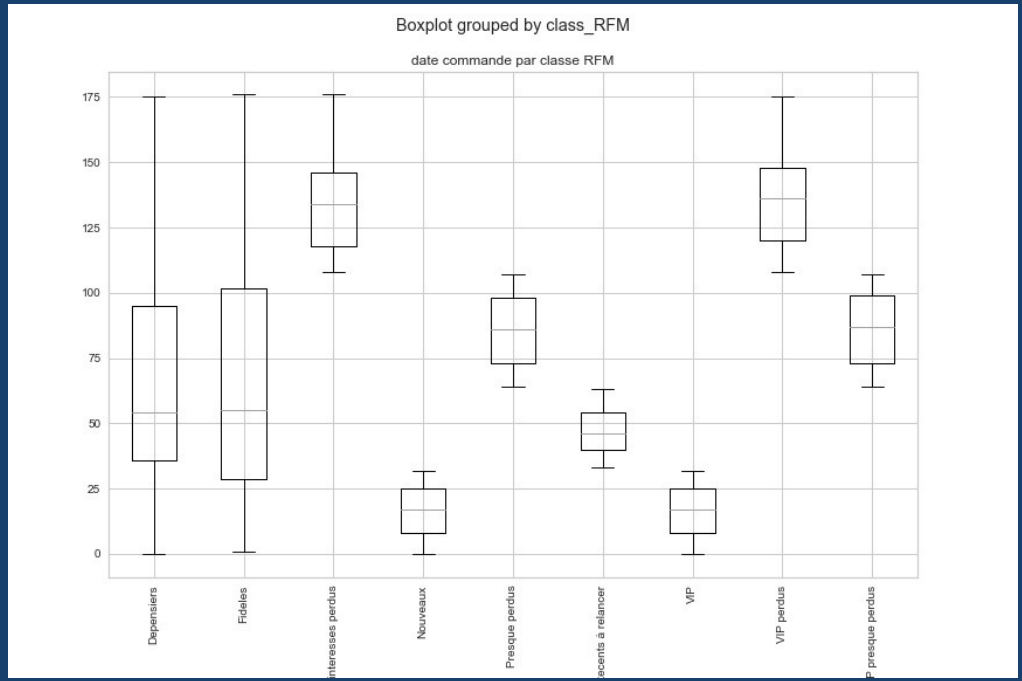
Segmentation RFM





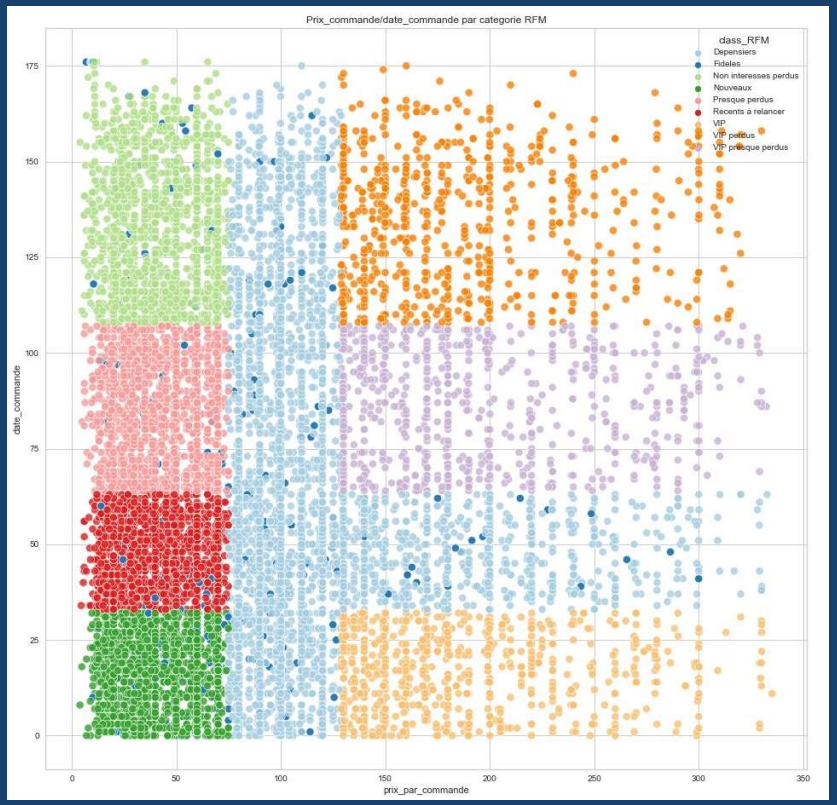
Segmentation RFM





Segmentation RFM

	class_RFM	Total	main_categ_count_en	main_categ_price_en	prix_par_commande	nb_paye_par_commande	nb_commande	date_commande	delai_jours	note_par_commande
0	Depensiers	3576	bed_bath_table	bed_bath_table	117.637671	3.499161	1.000000	64.498322	11.410466	4.242450
1	Fideles	547	bed_bath_table	bed_bath_table	71.861298	3.259598	2.180987	68.444241	11.525498	4.124314
2	Non interesses perdus	1451	furniture_decor	furniture_decor	40.609649	1.946244	1.000000	133.782908	11.038487	4.246726
3	Nouveaux	1440	housewares	housewares	41.955299	2.213889	1.000000	16.759722	10.851810	4.243056
4	Presque perdus	1420	telephony	telephony	40.575268	2.138028	1.000000	85.730282	11.679872	4.219718
5	Recents à relancer	1369	telephony	telephony	41.583798	2.246896	1.000000	47.499635	10.675067	4.248356
6	VIP	747	bed_bath_table	bed_bath_table	187.304444	4.475234	1.066934	16.544846	11.574210	4.255689
7	VIP perdus	731	furniture_decor	furniture_decor	187.116772	3.763338	1.030096	134.800274	12.078524	4.240766
8	VIP presque perdus	763	perfumery	perfumery	196.067313	4.128440	1.020970	86.073394	12.277457	4.201835

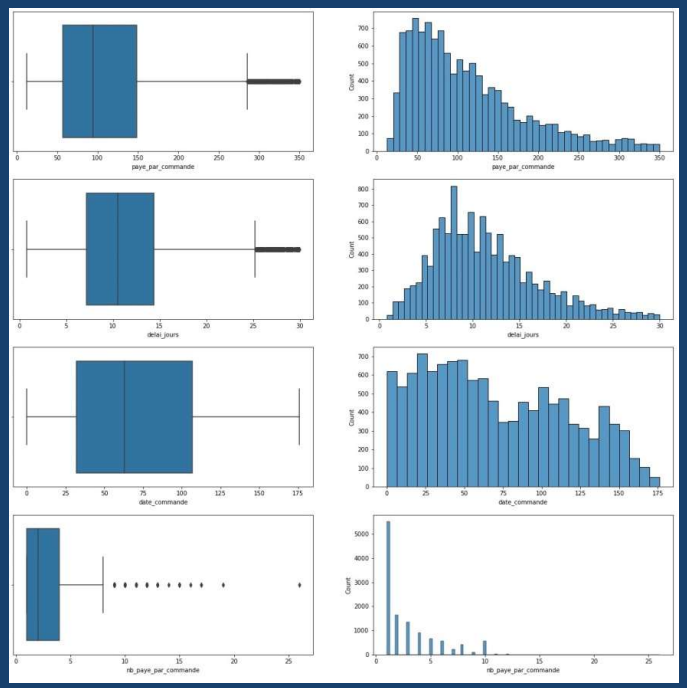


Feature Engineering

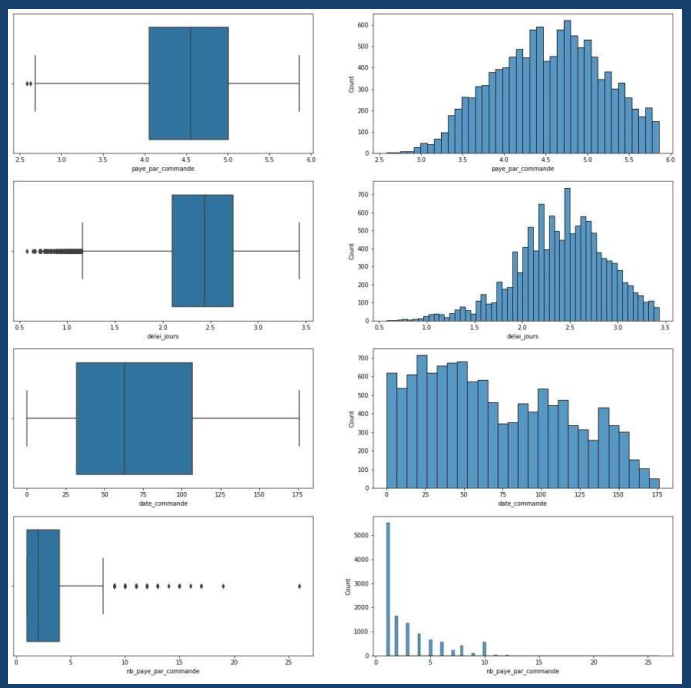
- quatre Variable par client :
  - Prix payé
  - Delai
  - Date
  - Nombre de paiement

Feature Engineering

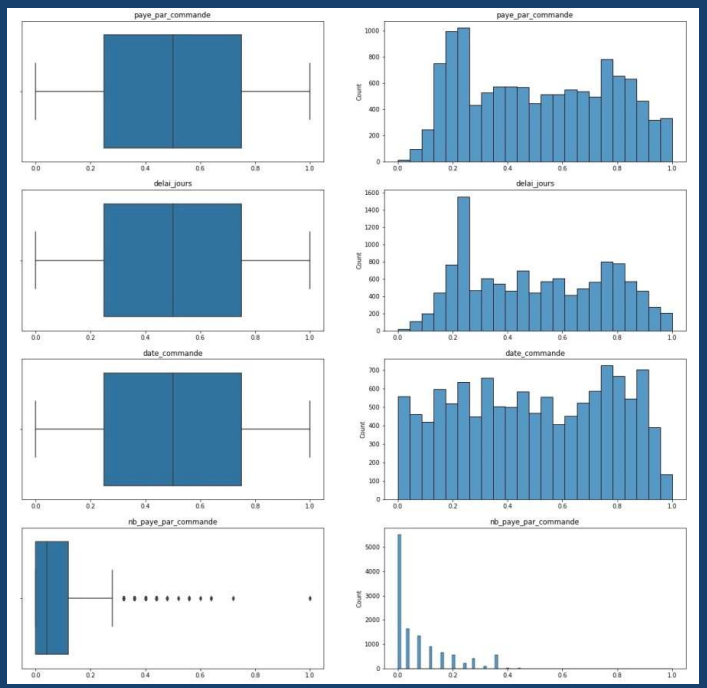
données départ



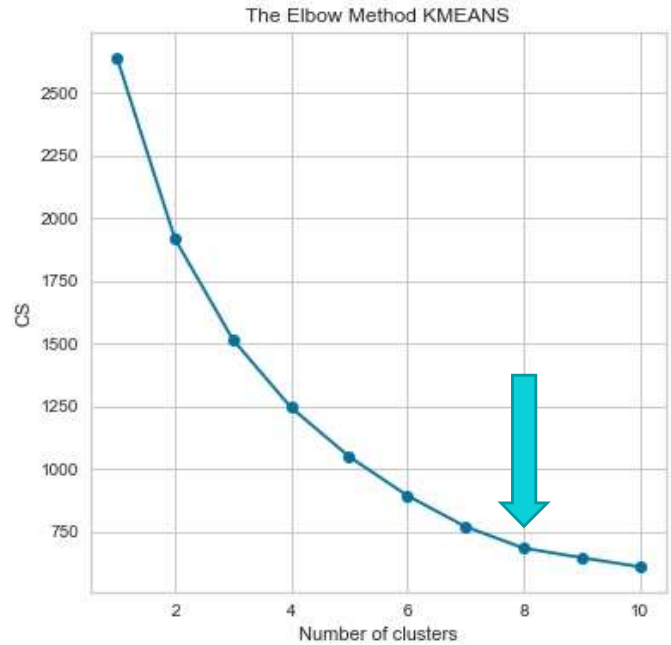
passage au log

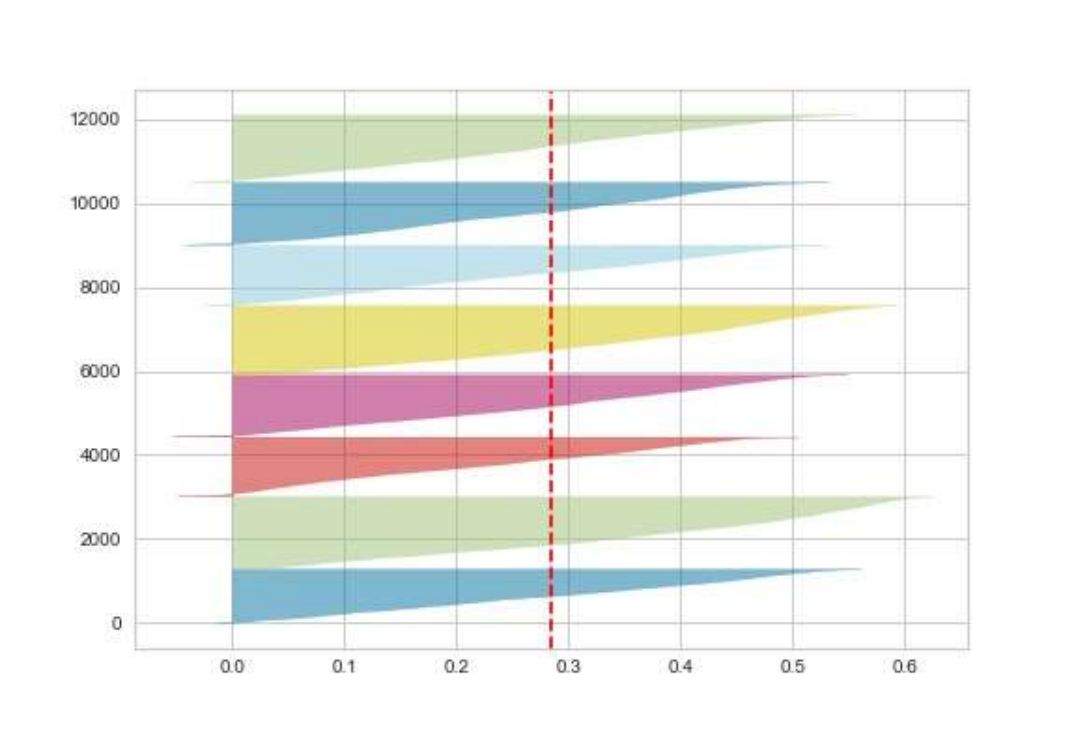


QuantileTransformer + MinMax

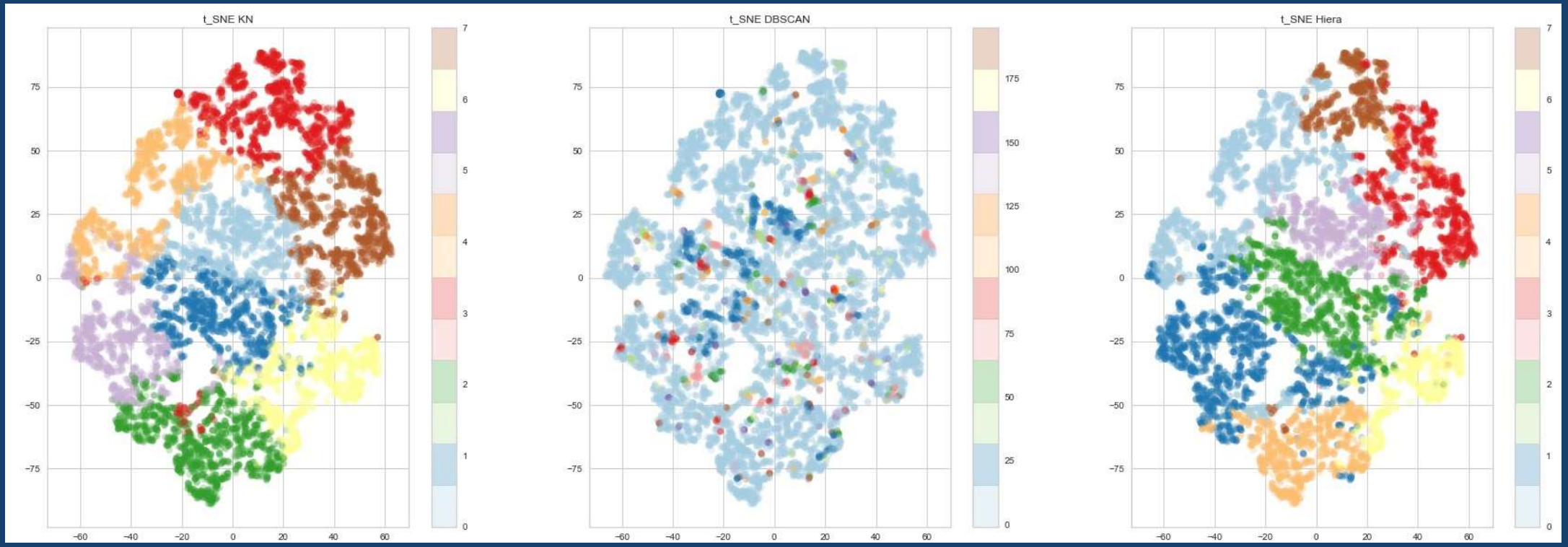


CLUSTERISATION

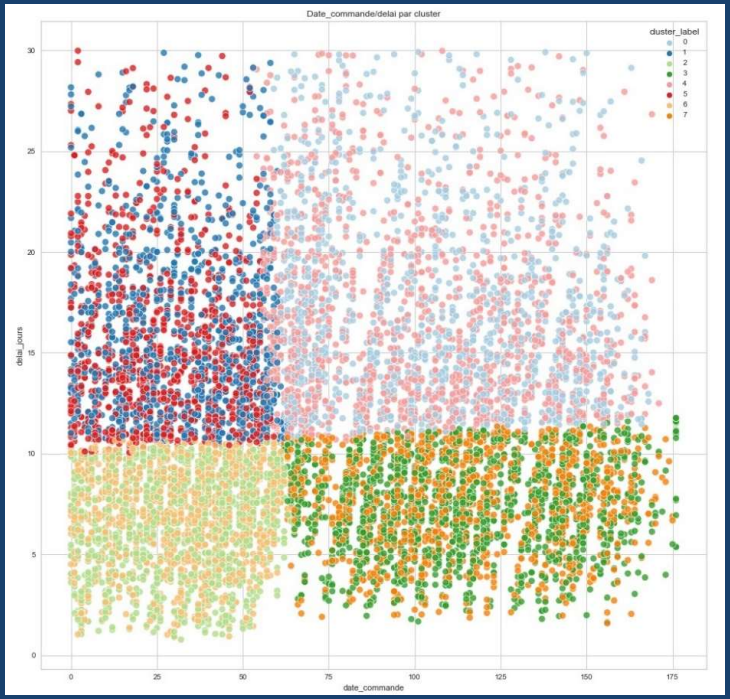


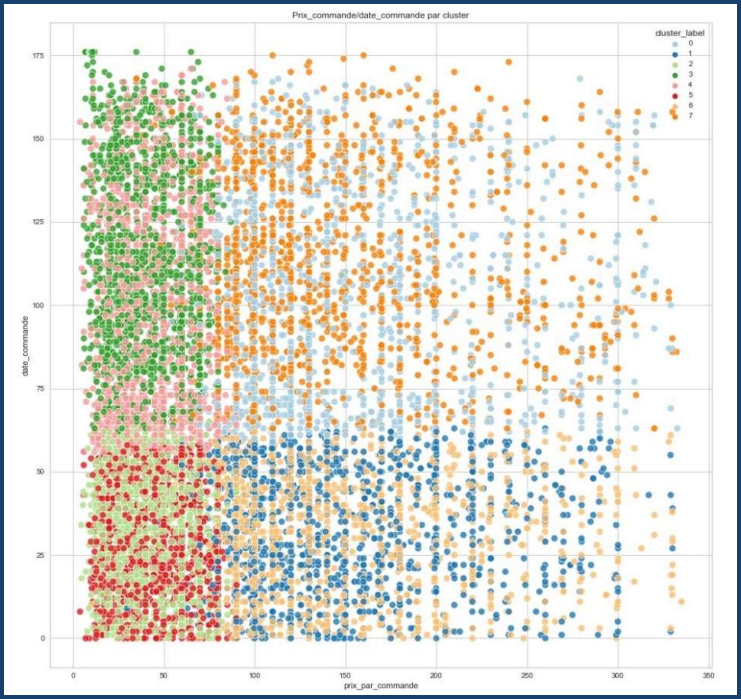


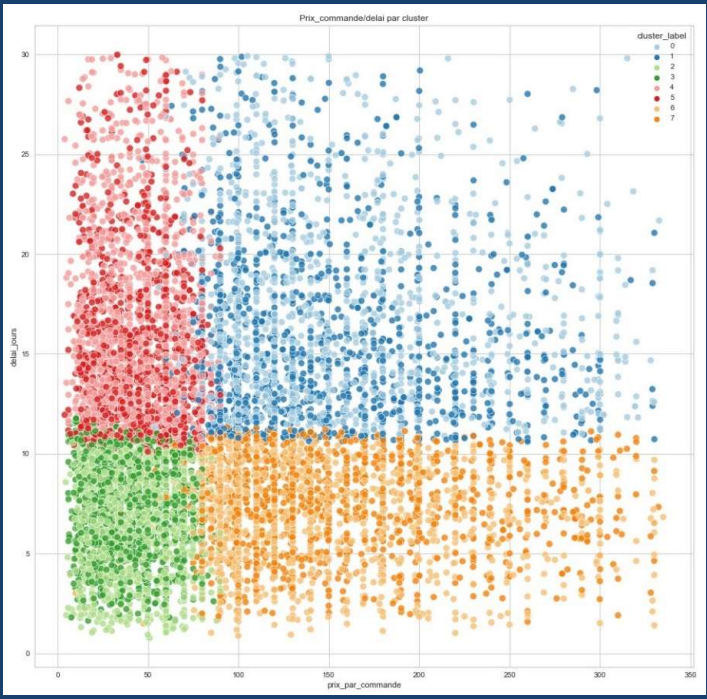
CLUSTERISATION



CLUSTERISATION

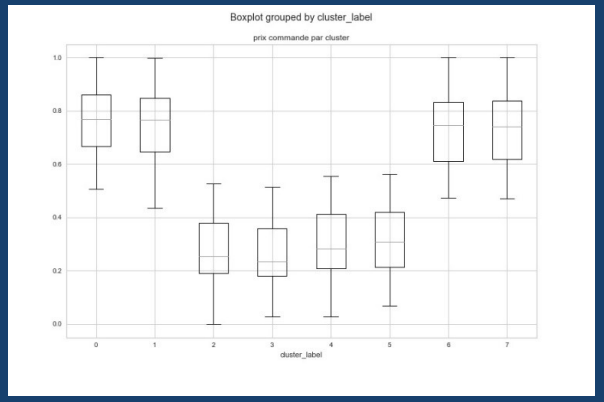


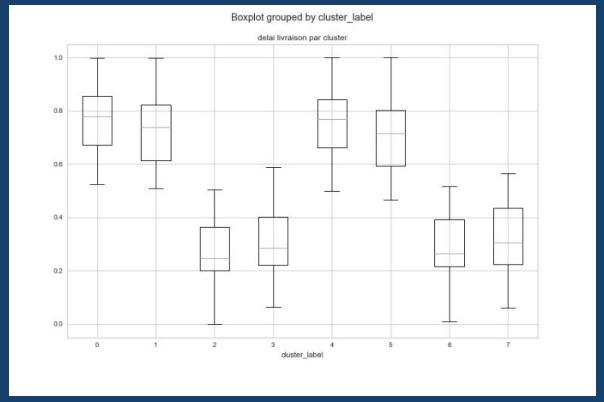


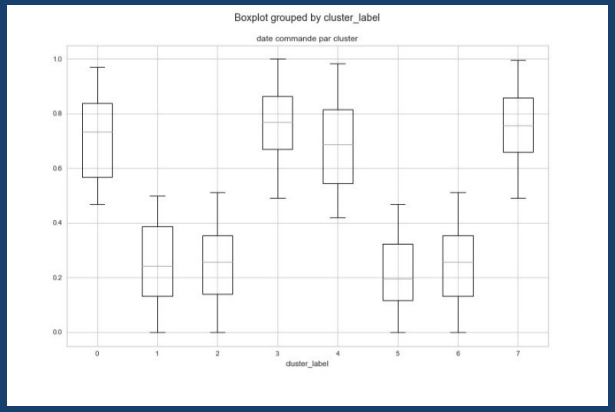


CLUSTERISATION

	cluster_label	Total	main_categ_count_en	main_categ_price_en	class_RFM	prix_par_commande	nb_paye_par_commande	nb_commande	date_commande	delai_jours	note_par_commande
0	0	1440	bed_bath_table	bed_bath_table	Depensiers	152.765764	3.676389	1.043056	104.860417	16.871560	4.077778
1	1	1389	bed_bath_table	bed_bath_table	Depensiers	145.718891	4.081353	1.093593	32.121670	15.517230	4.110871
2	2	1642	housewares	housewares	Recents à relancer	43.383356	2.114495	1.045067	31.606577	6.705551	4.354446
3	3	1729	furniture_decor	furniture_decor	Non interesses perdus	40.130902	1.905147	1.082128	114.266628	7.432379	4.357432
4	4	1579	furniture_decor	furniture_decor	Presque perdus	44.156080	2.110196	1.049398	99.398353	16.461273	4.105130
5	5	1290	bed_bath_table	bed_bath_table	Nouveaux	44.839620	2.328682	1.085271	26.921705	14.895065	4.110078
6	6	1500	bed_bath_table	bed_bath_table	Depensiers	144.062793	4.207333	1.050667	31.365333	7.074563	4.348000
7	7	1475	bed_bath_table	bed_bath_table	Depensiers	144.743132	3.585085	1.042034	112.451525	7.637923	4.352542







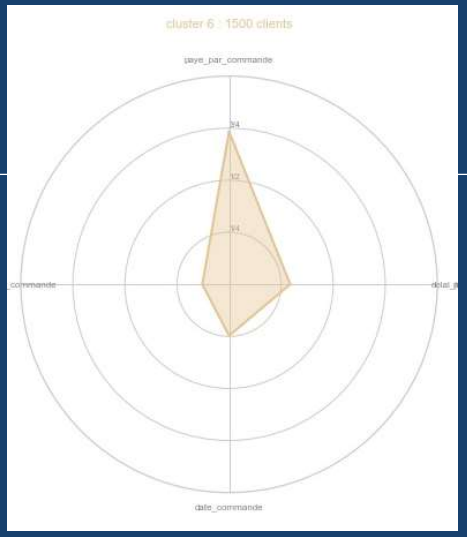
CLUSTERISATION

6 : Dépensier, récent, délai court, bonne note
- Champions bonne note, à récompenser

1 : Dépensier, récent, délai moyen, note moyen
- Champions note moyenne, à fidéliser

7 : Dépensier, vieille commande, délai court, bonne note
- Client risque de perdre, à relancer récompense si retour ?

0 : Dépensier, vieille commande , note pas bonne, délai pas bon
- Client ancien pas satisfait, mérite une attention



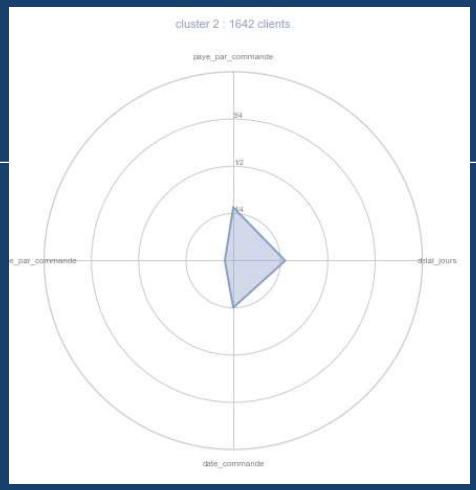
CLUSTERISATION

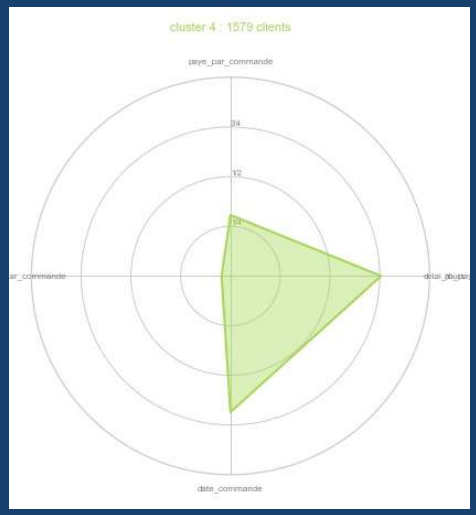
2 : petit montant, récent, délai court, bonne note
- Nouveaux clients et prometteurs

5 : petit montant, récent, délai pas bon, note basse
- Clients récents mais pas satisfait

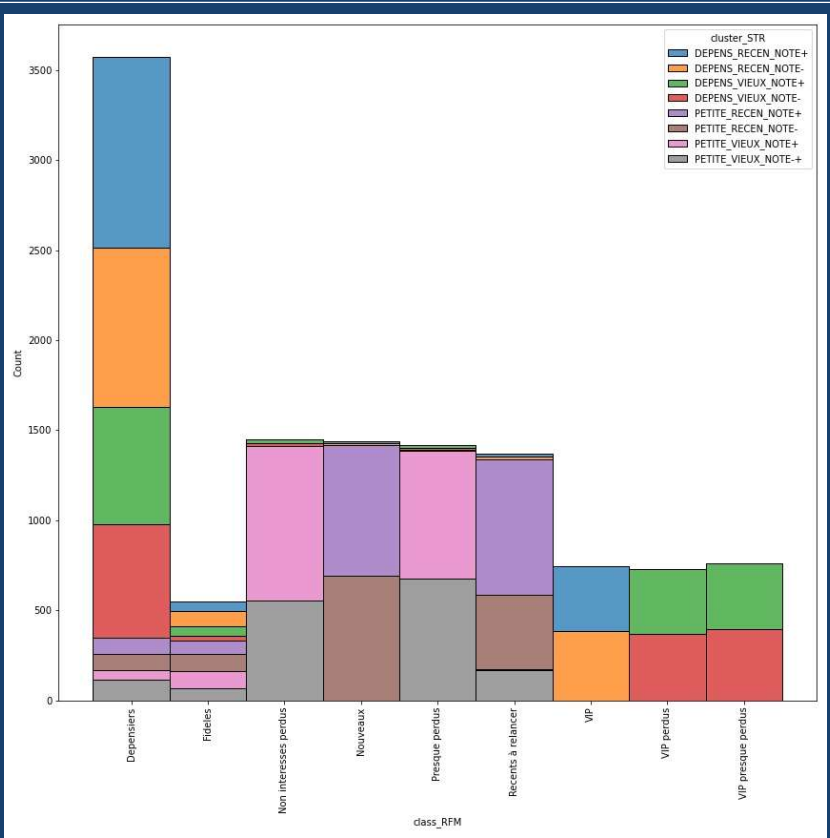
3 : petit montant, vieille commande, délai court, bonne note
- Client sur le point de disparaitre

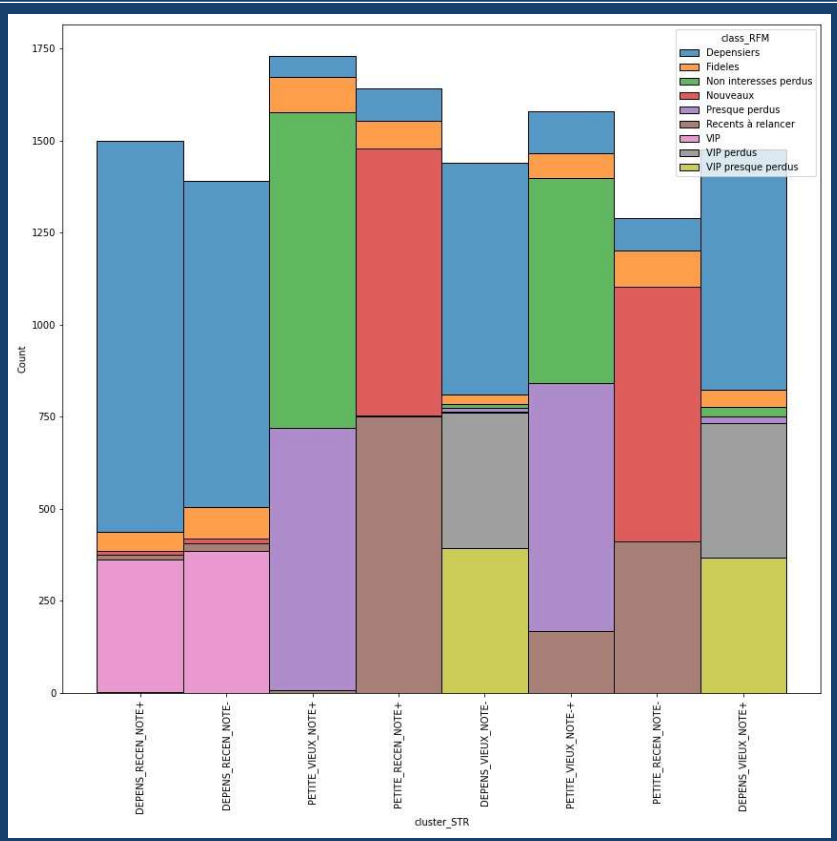
4 : petit montant, vieille commande, délai pas bon, note basse
- Client perdu





CLUSTERISATION





Stabilité CLUSTER

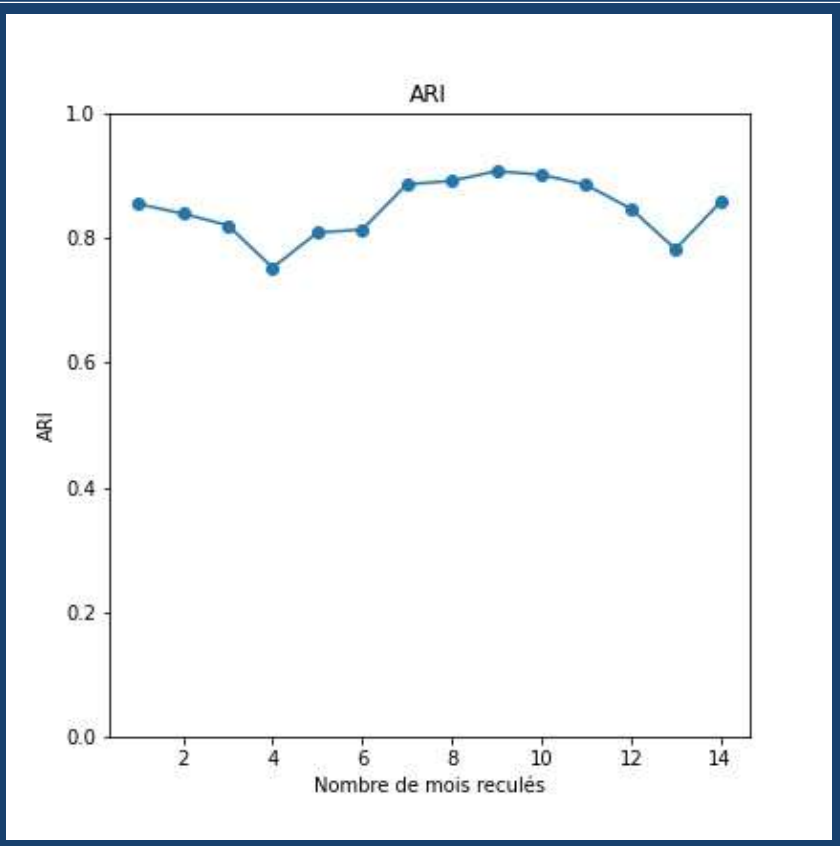
- Entrainement de 51 clusters KMEANS
- calcul de l'indice de Rand ajusté (ARI) pour chaque couple de cluster
- moyenne de tous les indices :
- résultat : 0,9885

```
print('ARI moyen : ', np.array(stab_clust).mean())
```

```
ARI moyen :  0.9885122291684986
```

Maintenance CLUSTER

- Entrainement cluster sur un semestre
- prédiction pour le mois suivant = jeu validation
- on décale les données entrainement 1 mois
- entrainement d'un nouveau modèle
- prédiction sur le jeu de validation
- calcul du score de Rand ajusté



MERCI

Questions et Réponses