Projet P4 – Segmentez des clients d'un site e-commerce

Synthèse et nettoyage jeu de données

geolocalisation_data

olist_geolocation_dataset
(1000163 x 5)

- geolocation_zip_code_prefix
- geolocation_lat
- geolocation_lng
- X geolocation_city
- X geolocation_state

- Group by 'mean' on 'zip_code' (19015 x 3)

clients_data

olist_customers_dataset
(99441 x 5)

- customer_id
- customer_unique_id
- customer_zip_code_prefix
- customer_city
- customer_state

- Merge sur variable 'zip_code_prefix'

order_data

olist_orders_dataset
(99441 x 8)

- order_id
- customer_id
- order_status
- order_purchase_timestamp
- X order_approved_at
- X order_delivered_carrier_date
- order_delivered_customer_date
- X order_estimated_delivery_date

customers_df

(99441 x 7)

- customer_id
- customer_unique_id
- zip_code_prefix
- customer_city
- customer_state
- geolocation_lat
- geolocation_lng

- Valeurs manquantes ''order_delivered customer_date' remplis avec 'order_estimated_delivery_date'
- Filtre sur la catégorie 'delivered' de la variable 'order_status' (96478 x 5)

- Merge sur variable 'customer_id'

order_items

olist_order_items_dataset
(112650 x 7)

- order_id
- order_item_id
- product_id
- X seller_id
- X shipping_limit_date
- price
- freight_value

customers_df

(96478 x 11)

- customer_id
- customer_unique_id
- zip_code_prefix
- customer_city
- customer_state
- geolocation_lat
- geolocation_lng
- order_id
- order_status
- order_purchase_timestamp
- order_delivered_customer_date

- Merge sur variable 'order_id'

order_payment_data

olist_order_payments_dataset
(103886 x 5)

- order_id
- payment_sequential
- payment_type
- payment_installments
- payment_value

- Remplacement valeurs 0 de 'payment_installements' avec 1
- Suppression lignes doublons avec même produit et differentes payment_type

customers_df

(110197 x 15)

- customer_id
- customer_unique_id
- zip_code_prefix
- customer_city
- customer_state
- geolocation_lat
- geolocation_lng
- order_id
- order_status
- order_purchase_timestamp
- order_delivered_customer_date
- order_item_id
- product_id
- price
- freight_value

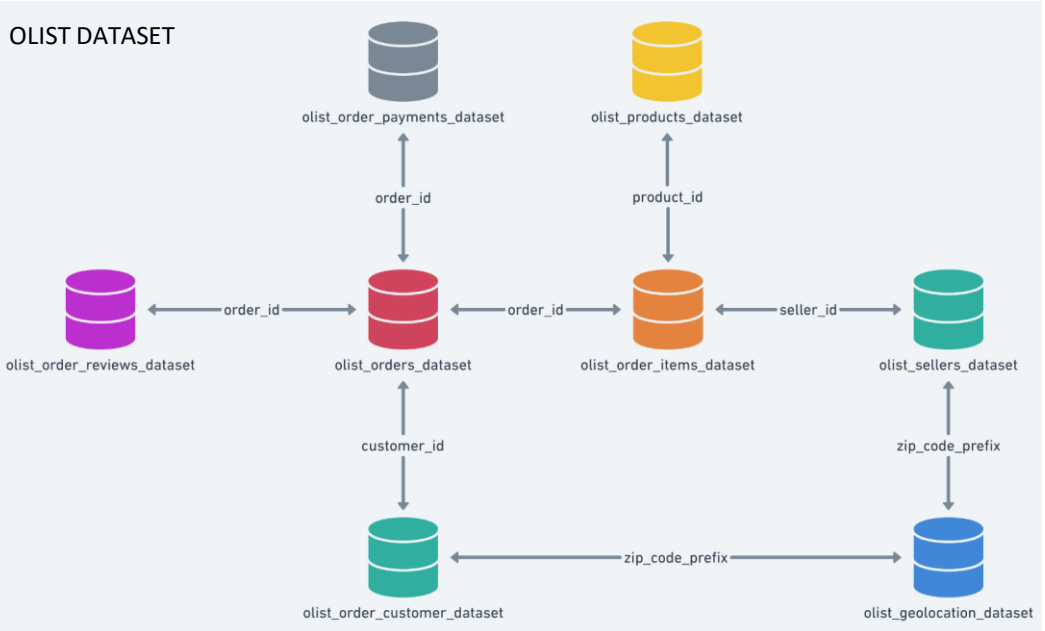
- Merge sur variable 'order_id'

customers_df

(110197 x 19)

- customer_id
- customer_unique_id
- zip_code_prefix
- customer_city
- customer_state
- geolocation_lat
- geolocation_lng
- order_id
- order_status
- order_purchase_timestamp
- order_delivered_customer_date
- order_item_id
- product_id
- price
- freight_value
- payment_sequential
- payment_type
- payment_installments
- payment_value

OLIST DATASET



Synthèse et nettoyage jeu de données

products_data

products_category_transl_data (71 x 2)

product_category_name
product_category_name_english

- Remplacement catégories 'pc_gamer' et 'portateis_cozinha_e_preparadores_de_alimentos' (70 x 2)

products_data

olist_products_dataset (32951 x 7)

product_id
product_category_name
X product_name_lenght
X product_description_lenght
product_photos_qty
X product_weight_g
X product_length_cm
X product_height_cm
X product_width_cm

- Suppression 7 variables
- Suppression lignes avec NaN pour 'product_category_name' (32341 x 3)

order_reviews_data

olist_order_reviews_dataset (99224 x 7)

X review_id
order_id
review_score
X review_comment_title
X review_comment_message
X review_creation_date
X review_answer_timestamp

- Suppression 5 variables (99224 x 2)

customers_df

(110197 x 19)

customer_id
customer_unique_id
zip_code_prefix
customer_city
customer_state
geolocation_lat
geolocation_lng
order_id
order_status
order_purchase_timestamp
order_delivered_customer_date
order_item_id
product_id
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value

- Merge sur variable'order_id'

customers_df

(110197 x 20)

customer_id
customer_unique_id
zip_code_prefix
customer_city
customer_state
geolocation_lat
geolocation_lng
order_id
order_status
order_purchase_timestamp
order_delivered_customer_date
order_item_id
product_id
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value
review_score

- Merge sur variable 'product_id'

- Merge sur variable 'product_category_name'

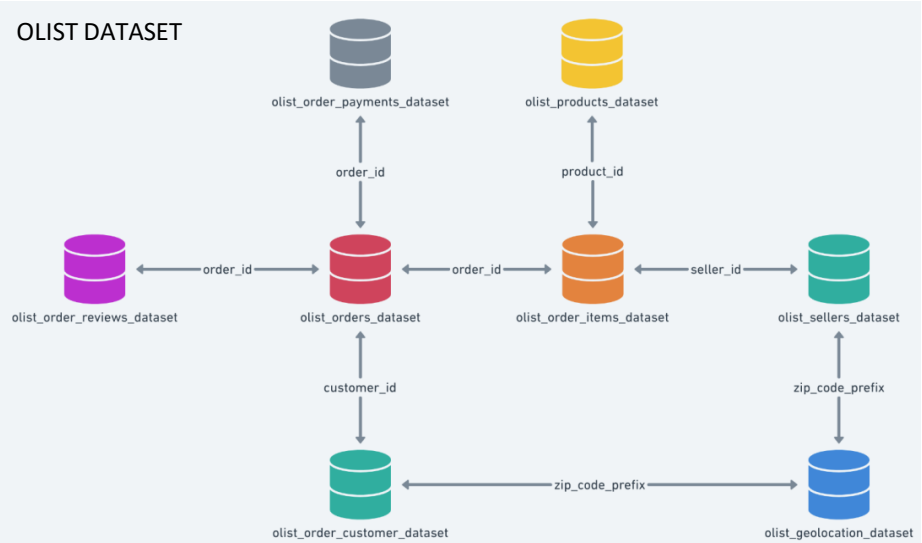
- Suppression variable 'product_category_name' (32341 x 2)

customers_df

(110197 x 22)

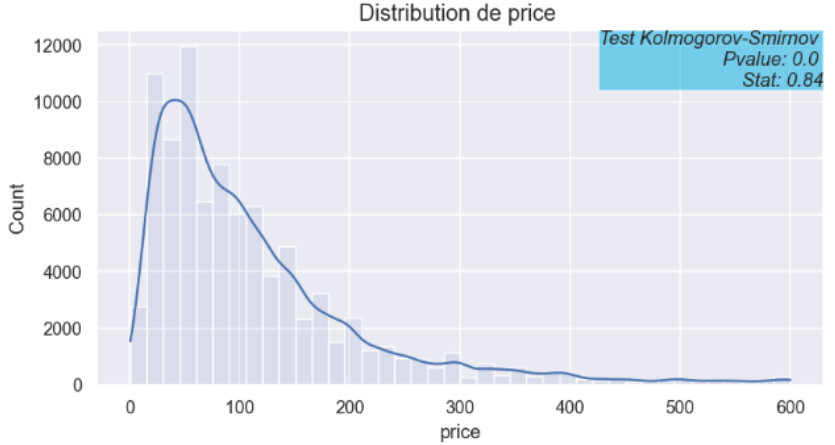
customer_id
customer_unique_id
zip_code_prefix
customer_city
customer_state
geolocation_lat
geolocation_lng
order_id
order_status
order_purchase_timestamp
order_delivered_customer_date
order_item_id
product_id
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value
review_score
product_photos_qty
product_category_name_english

OLIST DATASET

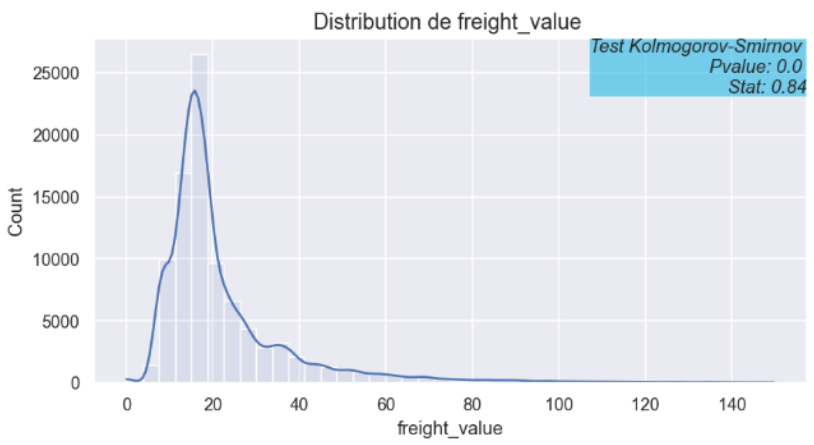



```
Data columns (total 22 columns):
 #   Column                         Non-Null Count   Dtype
---  ------                         --------------   -----
 0   customer_id                    110197 non-null  object
 1   customer_unique_id             110197 non-null  object
 2   zip_code_prefix                110197 non-null  int64
 3   customer_city                  110197 non-null  object
 4   customer_state                 110197 non-null  object
 5   geolocation_lat                109909 non-null  float64
 6   geolocation_lng                109909 non-null  float64
 7   order_id                       110197 non-null  object
 8   order_status                   110197 non-null  object
 9   order_purchase_timestamp       110197 non-null  object
 10  order_delivered_customer_date  110197 non-null  object
 11  order_item_id                  110197 non-null  int64
 12  product_id                     110197 non-null  object
 13  price                          110197 non-null  float64
 14  freight_value                  110197 non-null  float64
 15  payment_sequential             110194 non-null  float64
 16  payment_type                   110194 non-null  object
 17  payment_installments           110194 non-null  float64
 18  payment_value                  110194 non-null  float64
 19  review_score                   109370 non-null  float64
 20  product_photos_qty             108660 non-null  float64
 21  product_category_name_english  108660 non-null  object
dtypes: float64(9), int64(2), object(11)
```

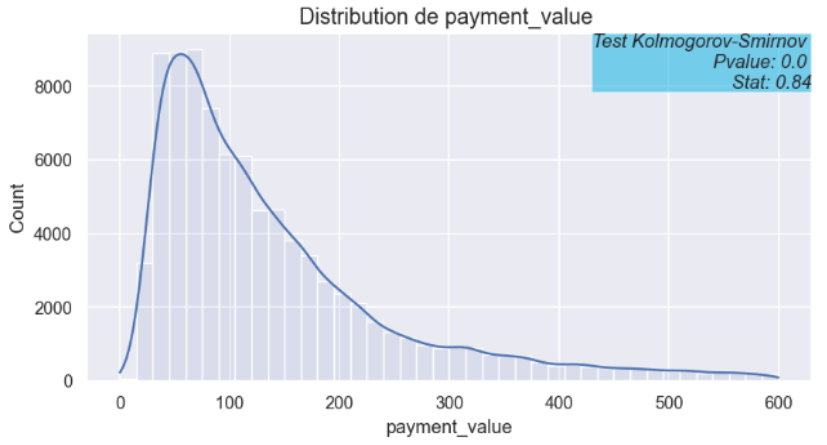
Exploration jeu de données – Analyse univariée variables quantitatives



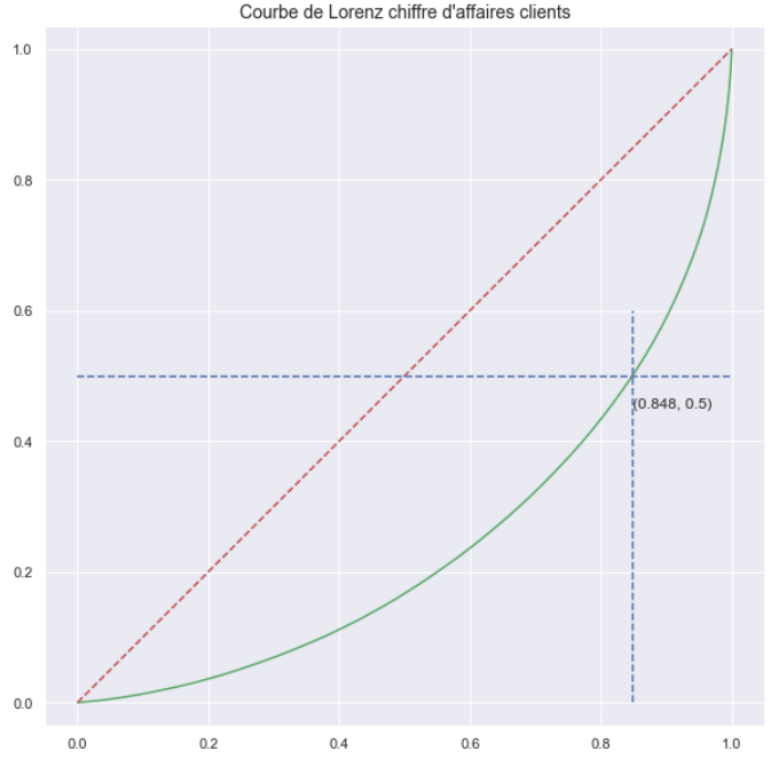
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	93358	0	0.850000	13440.000000	13439.100000	141.600000	89.730000	[-74.1, 357.3]	9.746869	265.212219



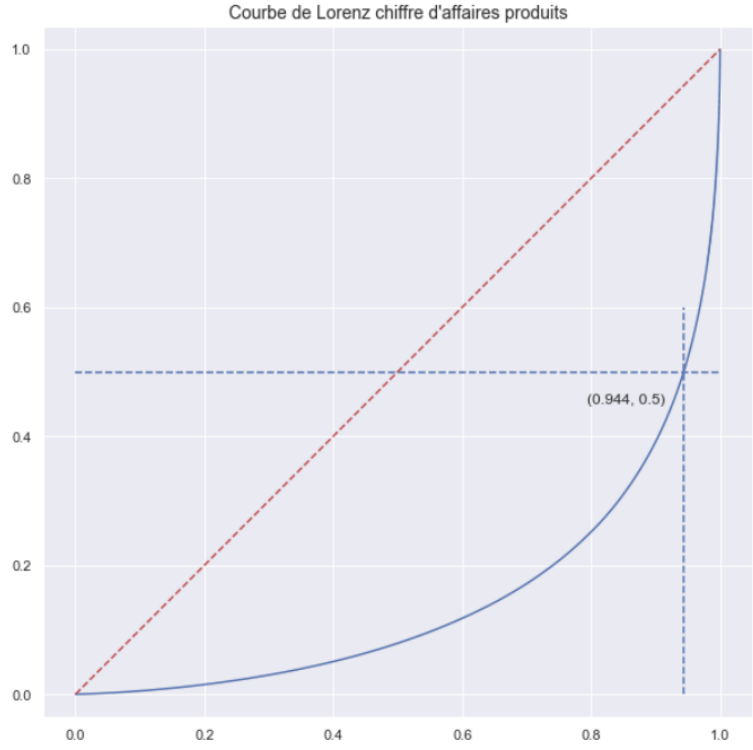
	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	93358	0	0.000000	1795.000000	1795.000000	23.500000	17.600000	[0.7, 46.3]	11.849747	519.711834



	count	missing values	min	max	range	mean	median	stdev	skewness	kurtosis
0	93358	0	9.590000	109312.600000	109312.600000	211.800000	112.950000	[-430.4, 854.0]	70.336527	9785.107388



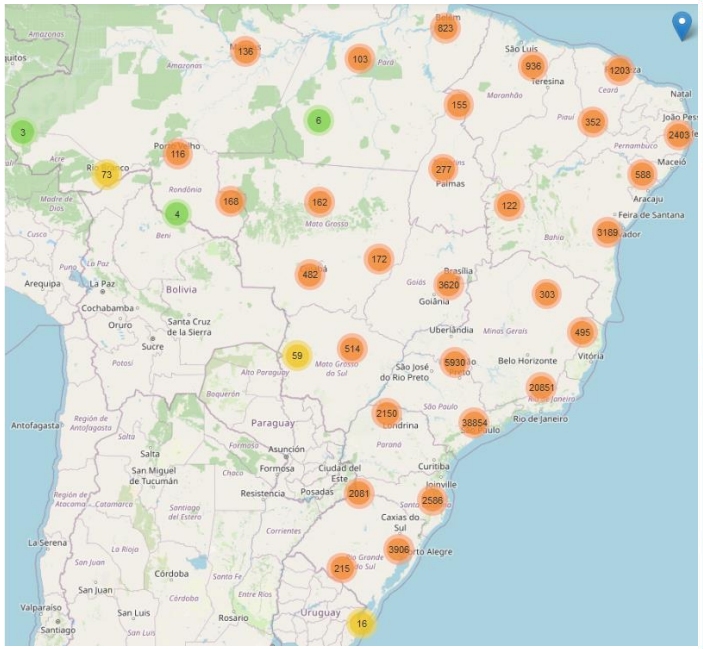
Percentage clients premium avec 50% chiffre d'affaires: 15.2 %

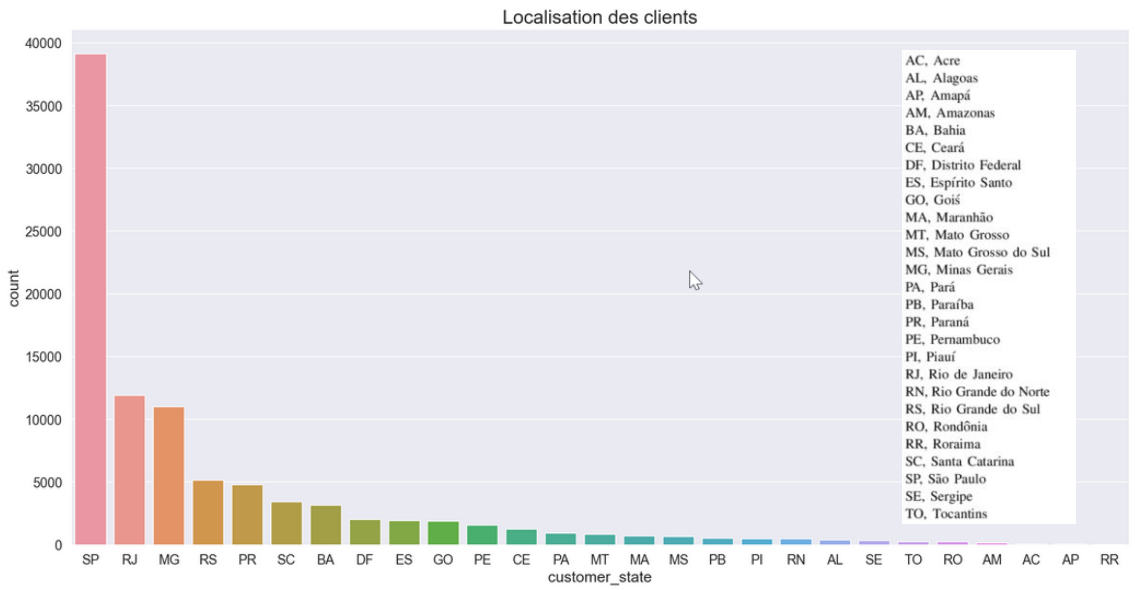


Percentage produits premium avec 50% chiffre d'affaires: 5.6 %

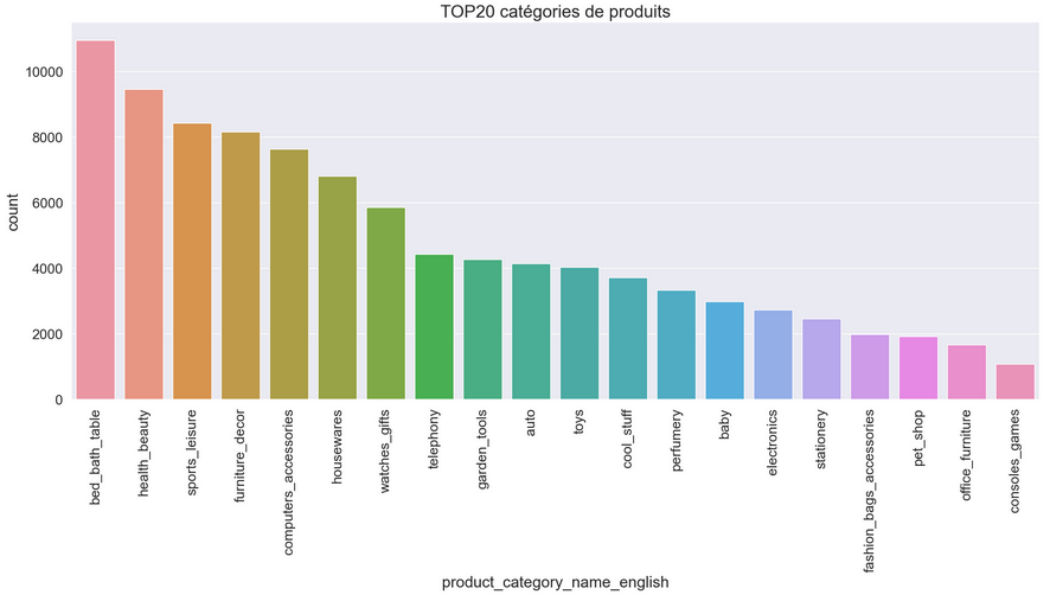
Exploration jeu de données – Analyse univariée et bivariée des variables qualitatives

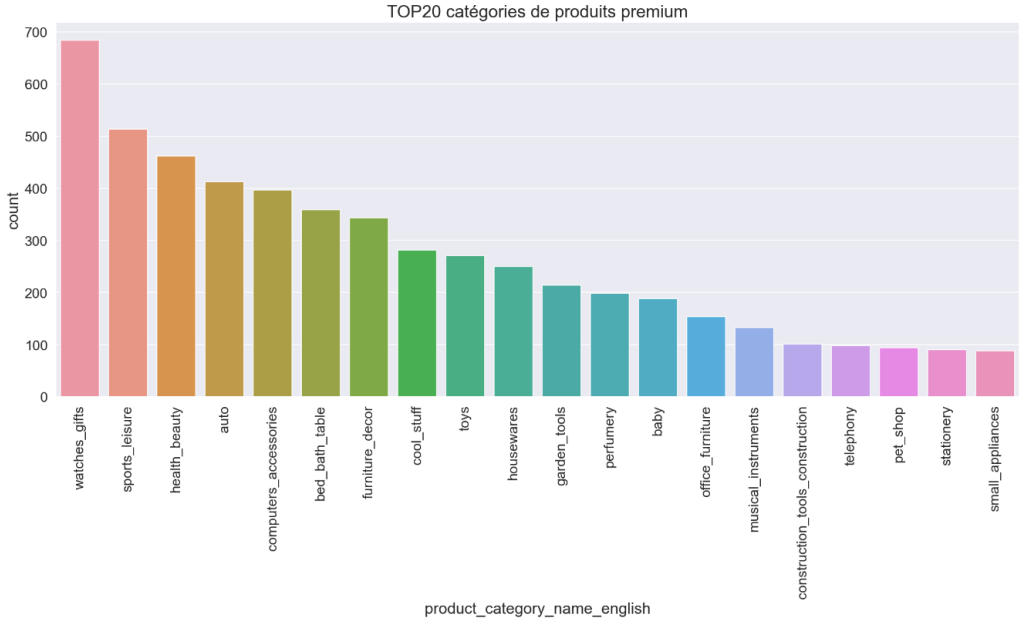
LOCALISATION CLIENTS





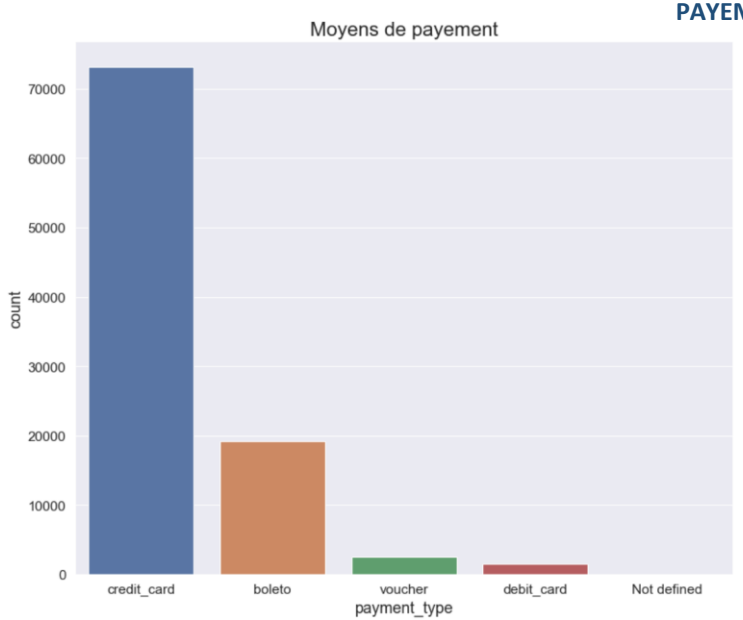
PRODUITS

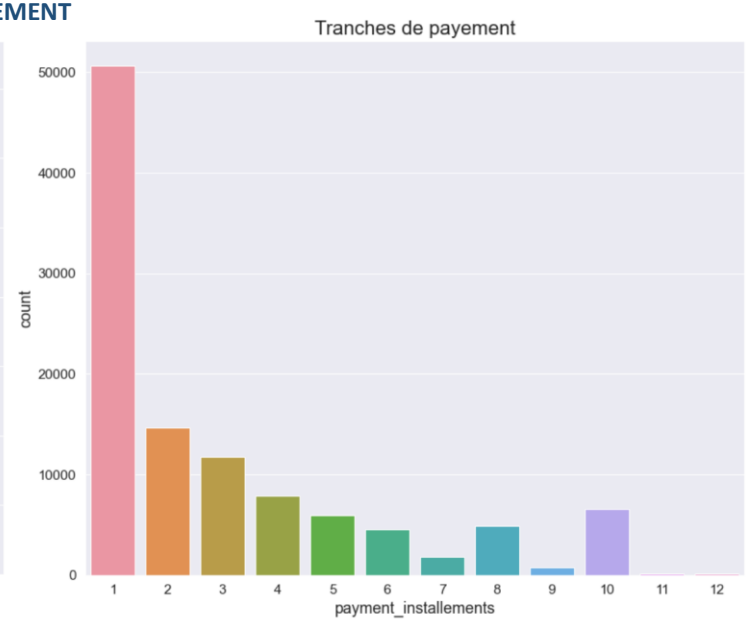




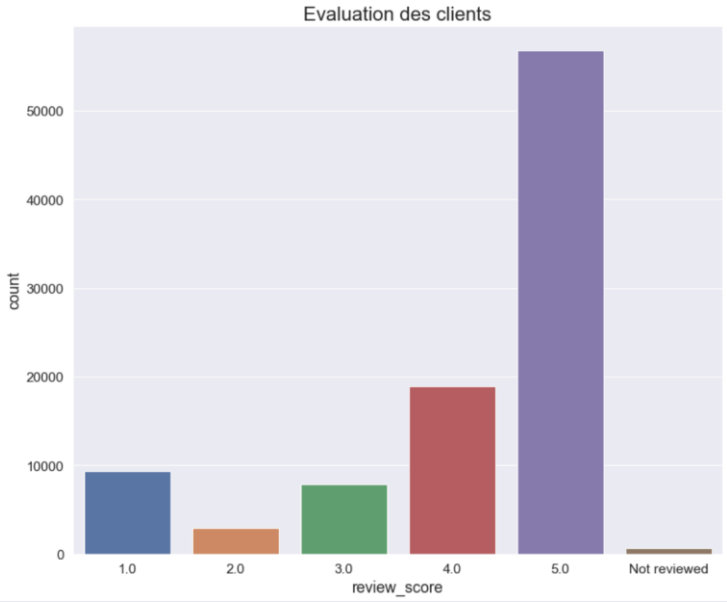
Exploration jeu de données – Analyse univariée et bivariée des variables qualitatives

PAYEMENT

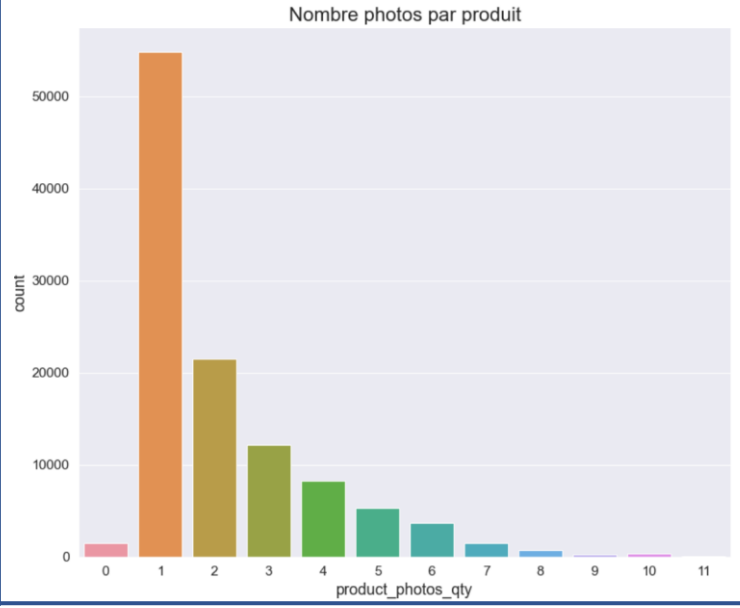




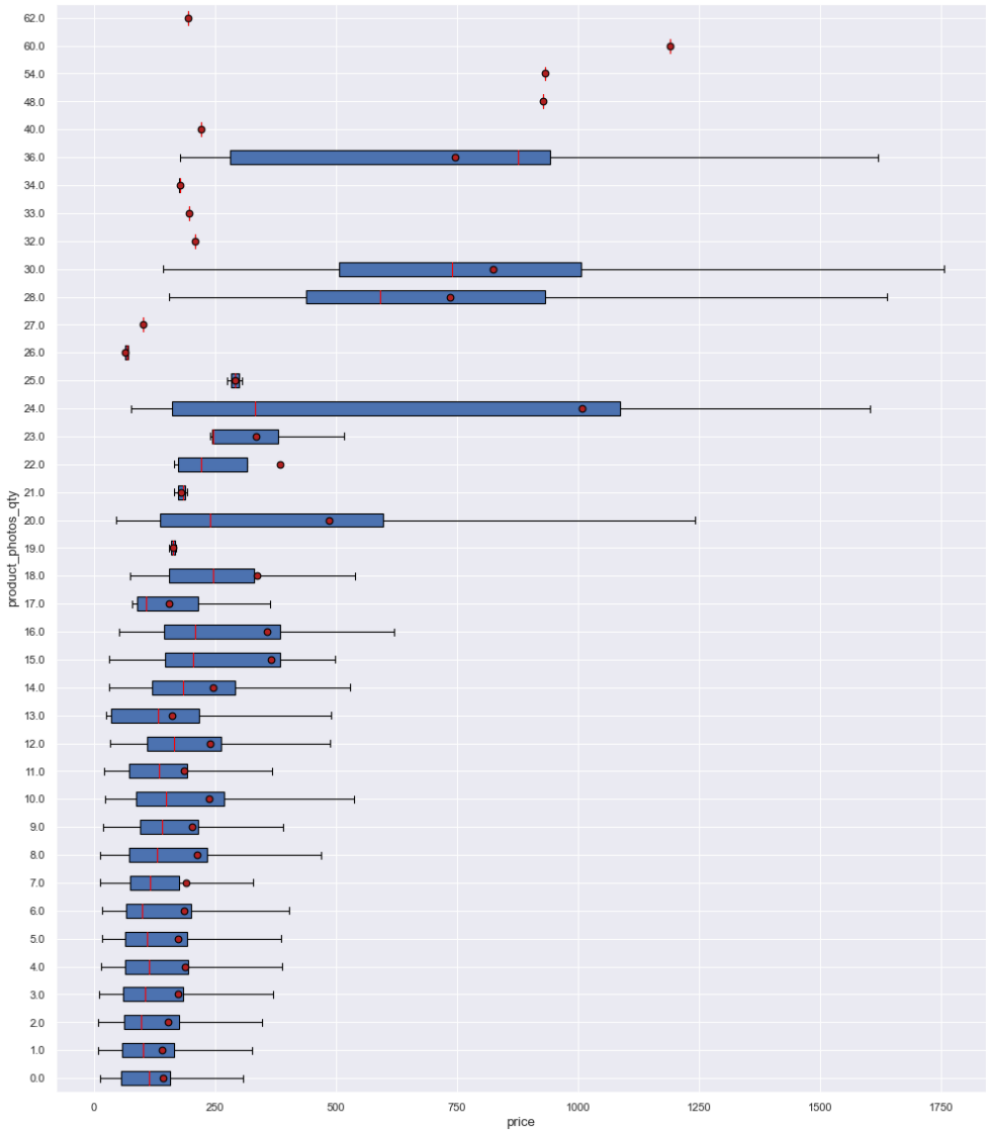
EVALUATION



NOMBRE PHOTOS PAR PRODUIT



CORRELATION ENTRE NOMBRE PHOTOS ET PRIX



Test Kruskal Wallis

	Source	ddof1	H	p-unc
Kruskal	product_photos_qty	38	1278.360829	1.302386e-243

Pre-processing et feature engineering

customers_df

(110197 x 22)

customer_id
customer_unique_id
zip_code_prefix
customer_city
customer_state
geolocation_lat
geolocation_lng
order_id
order_status
order_purchase_timestamp
order_delivered_customer_date
order_item_id
product_id
price
freight_value
payment_sequential
payment_type
payment_installments
payment_value
review_score
product_photos_qty
product_category_name_english

- Remplissage valeurs manquantes product_photos_qty avec 0
- Remplissage manquantes payment_sequential avec 0
- Remplissage manquantes payment_installments avec 1
- Remplissage manquantes payment_value avec summe de 'price' et 'freight_value'
- Création variable délai_livraison = order_delivered_customer_date - order_purchase_timestamp
- Synthèse des categories de produit

customers_df

(110197 x 23)

X customer_id
customer_unique_id
X zip_code_prefix
X customer_city
customer_state
X geolocation_lat
X geolocation_lng
X order_id
X order_status
order_purchase_timestamp
X order_delivered_customer_date
order_item_id
product_id
price
freight_value
X payment_sequential
X payment_type
X payment_installments
payment_value
review_score
product_photos_qty
product_category_name_english
delai_livraison

- Group by orders
- Agrégations par colonnes
- Suppression 11 variables
- Rename variables
- Création catégorie 'Multiple' pour les catégories de produit
- Création variable 'nb_produits_premium'
- Création variable 'frais/valeur_commande'
- Création variable 'nb_type_payement'
- Création variable 'nb_tranches'

data_orders_df

(96478 x 16)

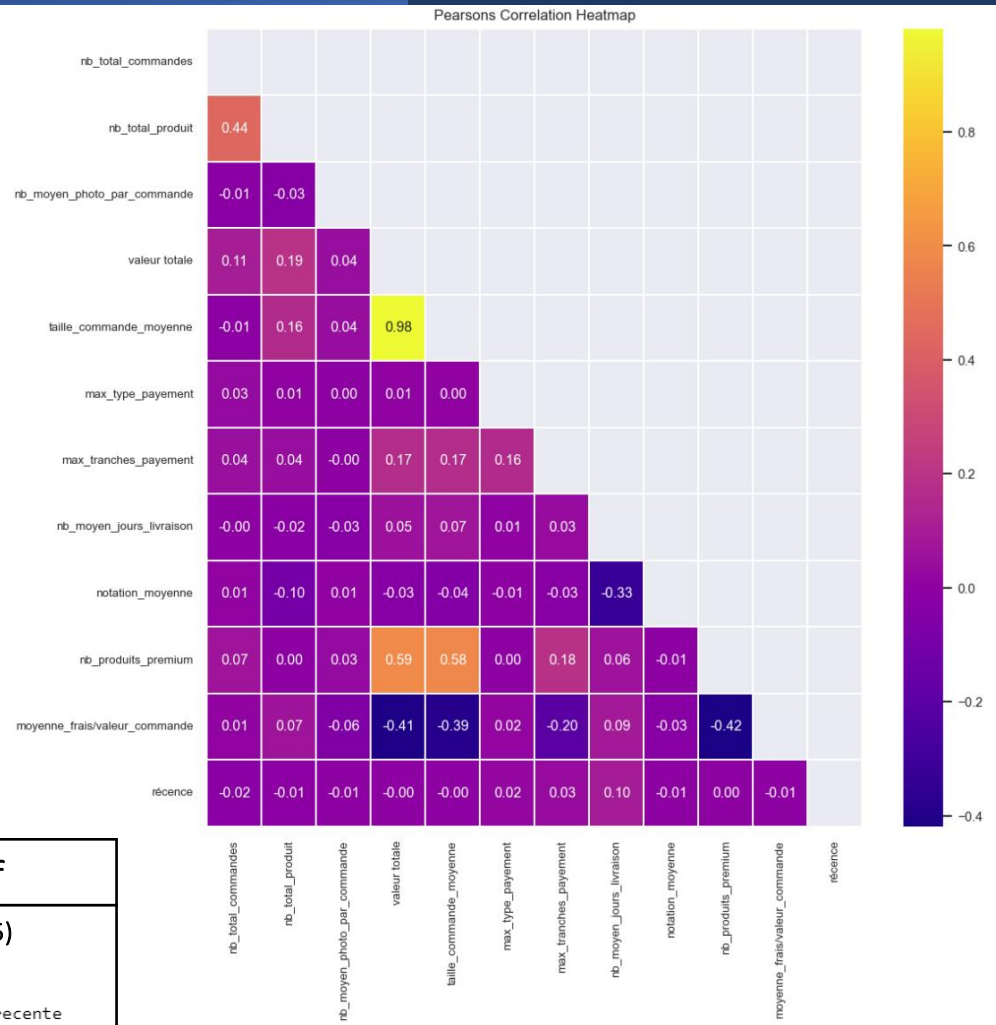
X customer_unique_id
nb_items
date_commande
id_produit
nb_photo_produit
prix_totale_commande
X valeur_totale_frais_transport
nb_type_payement
nb_tranches
valeur_payement
nb_jours_délai_livraison
notation
customer_state
product_category
nb_produits_premium
frais/valeur_commande

- Group by customers
- Agrégations par colonnes
- Suppression variables
- Rename variables
- Création variable récence = max_date - date_commande_plus_recente
- Création variable 'taille commande moyenne'
- Synthèse des categories de produit
- Synthèse variable 'customer_state'

data_df

(93358 x 15)

nb_total_commandes
nb_total_produit
date_commande_plus_recente
nb_moyen_photo_par_commande
valeur totale
taille_commande_moyenne
max_type_payement
max_tranches_payement
nb_moyen_jours_livraison
notation_moyenne
nb_produits_premium
moyenne_frais/valeur_commande
récence
product_category
customer_state

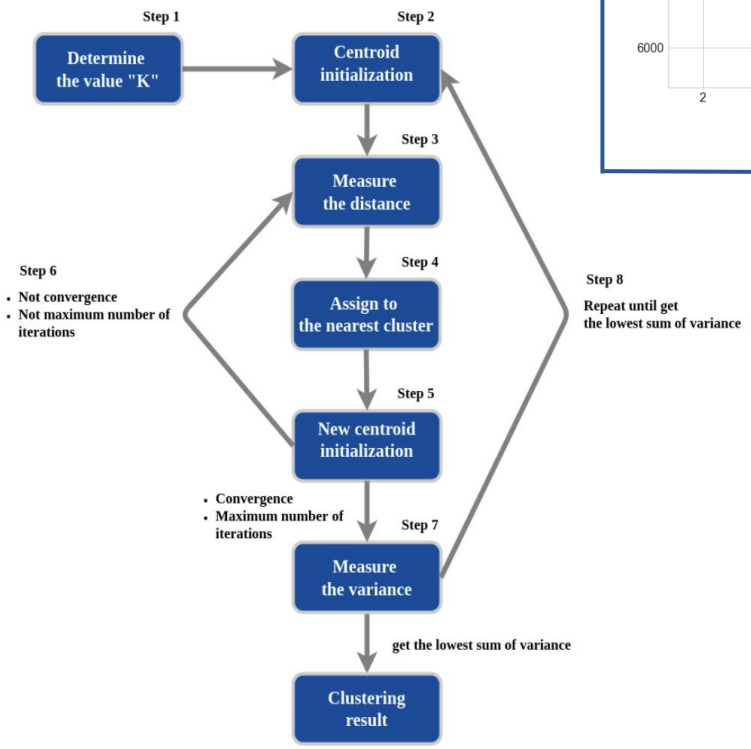



```
0   nb_total_commandes              93358 non-null  int64
1   nb_total_produit                93358 non-null  int64
2   date_commande_plus_recente      93358 non-null  datetime64[ns]
3   nb_moyen_photo_par_commande     93358 non-null  float64
4   valeur totale                   93358 non-null  float64
5   taille_commande_moyenne         93358 non-null  float64
6   max_type_payement               93358 non-null  int32
7   max_tranches_payement           93358 non-null  int32
8   nb_moyen_jours_livraison        93358 non-null  float64
9   notation_moyenne                93358 non-null  float64
10  nb_produits_premium             93358 non-null  int64
11  moyenne_frais/valeur_commande   93358 non-null  float64
12  récence                         93358 non-null  int64
13  product_category                93358 non-null  object
14  customer_state                  93358 non-null  object
```

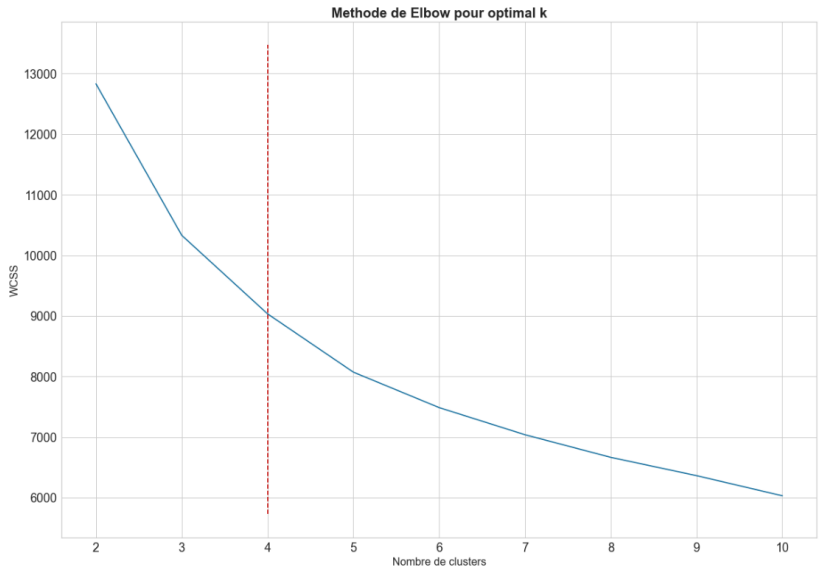
Normalisation du jeu de données

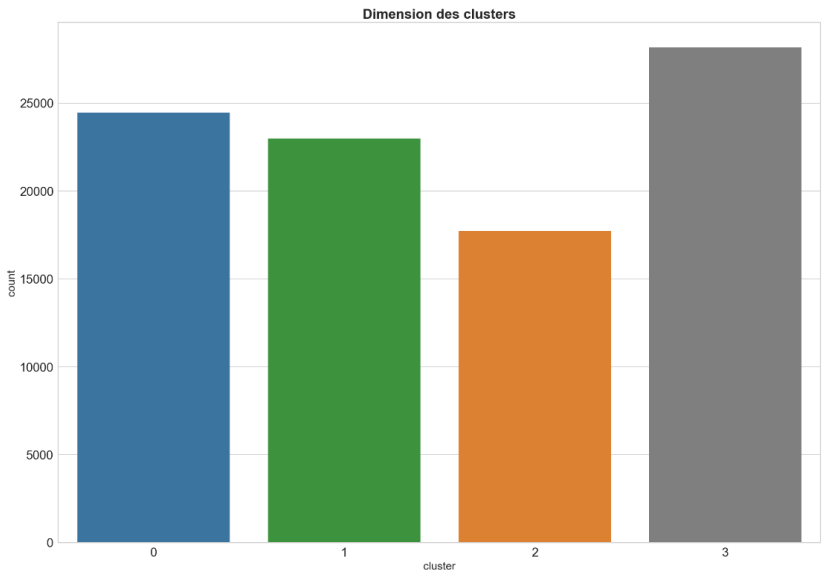
```
0   nb_total_commandes              93358 non-null  int64
1   nb_total_produit                93358 non-null  int64
2   nb_moyen_photo_par_commande     93358 non-null  float64
3   valeur totale                   93358 non-null  float64
4   taille_commande_moyenne         93358 non-null  float64
5   nb_moyen_jours_livraison        93358 non-null  float64
6   notation_moyenne                93358 non-null  float64
7   nb_produits_premium             93358 non-null  int64
8   moyenne_frais/valeur_commande   93358 non-null  float64
9   récence                         93358 non-null  int64
```

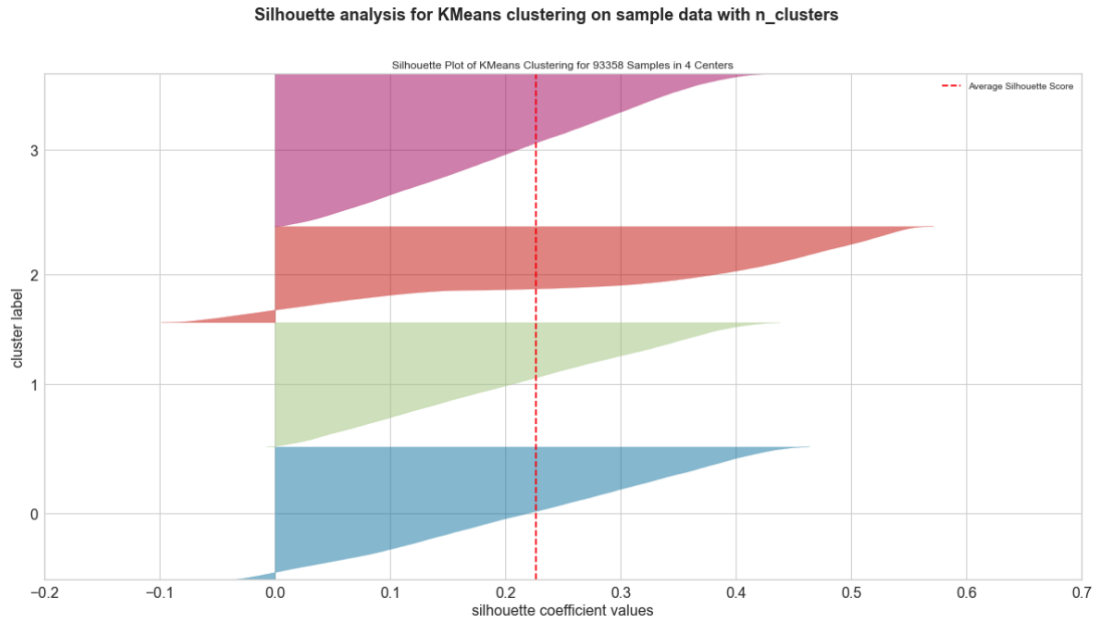
Algorithme Kmeans



Définition nombre de cluster

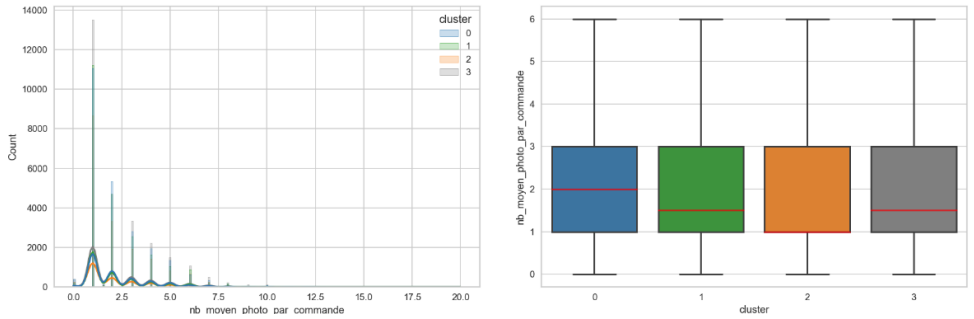




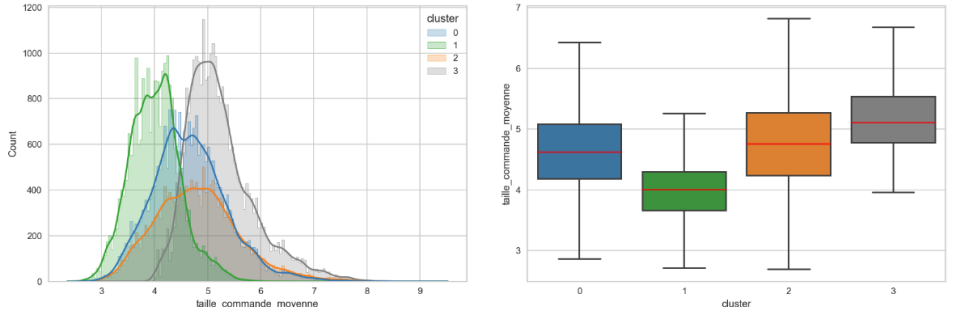


Algorithme K-means: analyse statistique des clusters

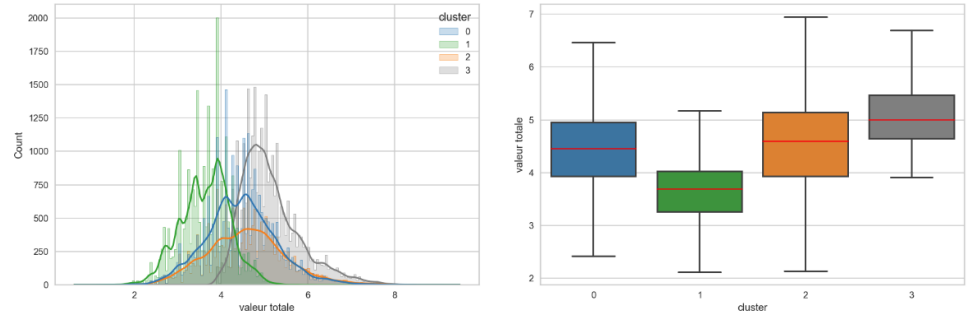
Nb. moyen photo par commande



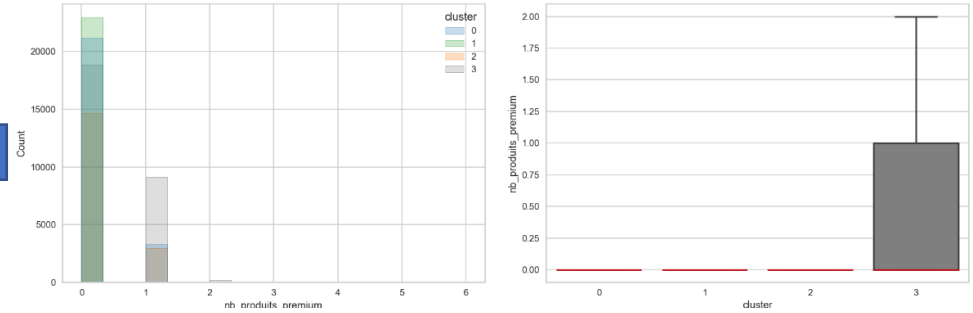
Taille commande moyenne



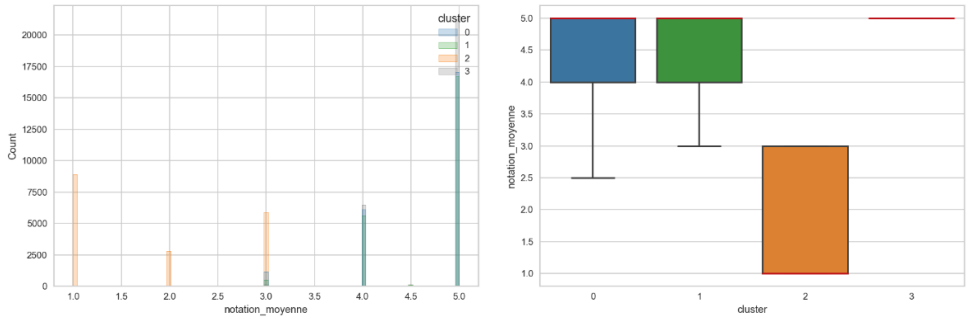
Valeur totale



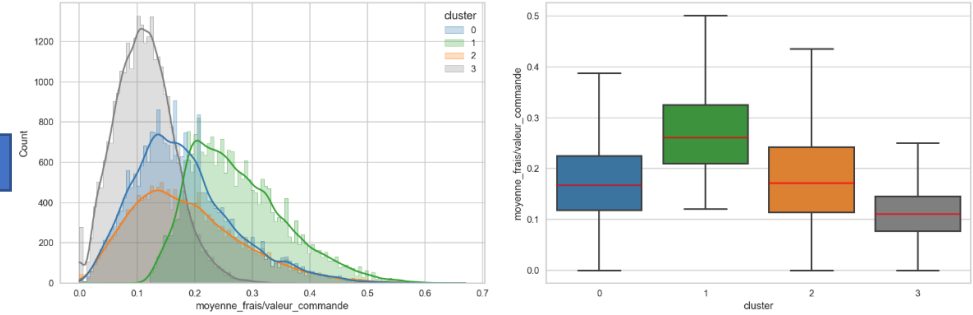
Nb. produits premium



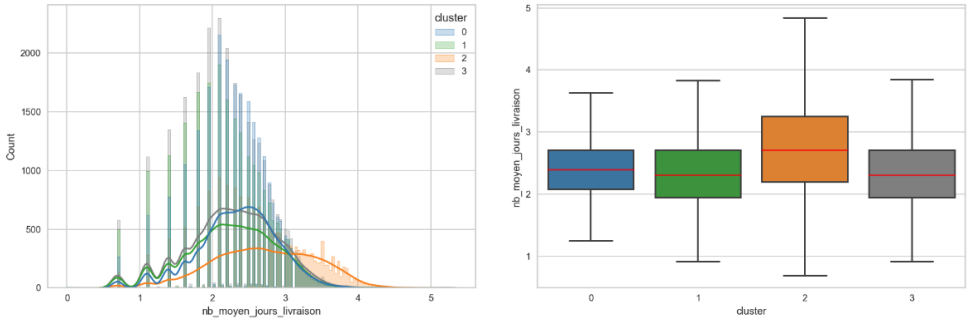
Notation moyenne



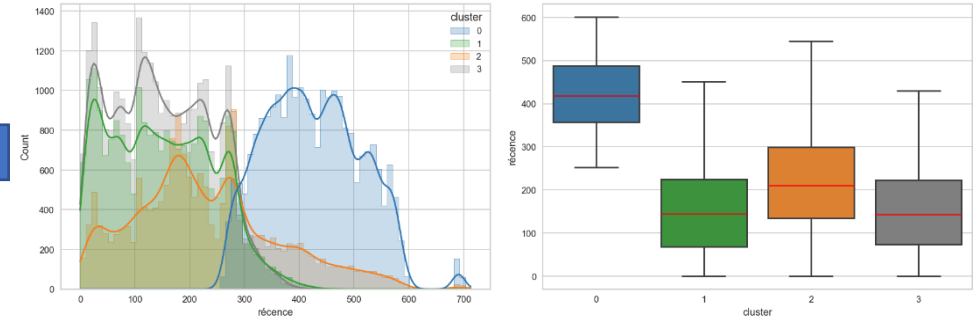
Moyenne frais/valeur commande



Nb. moyen jours de livraison



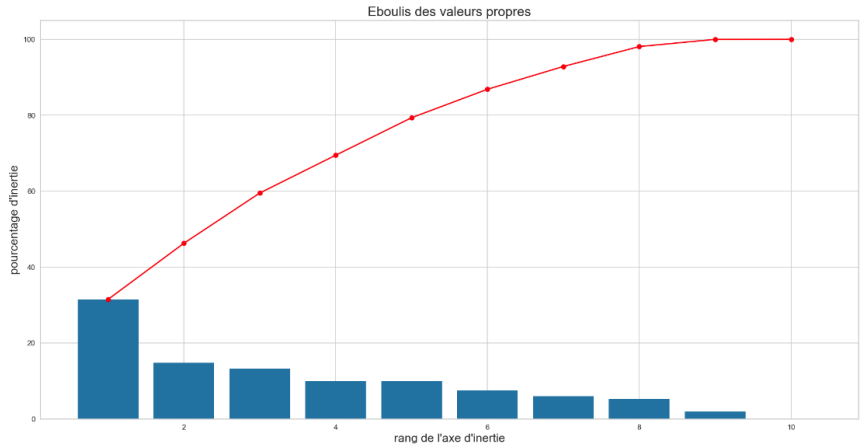
Récence



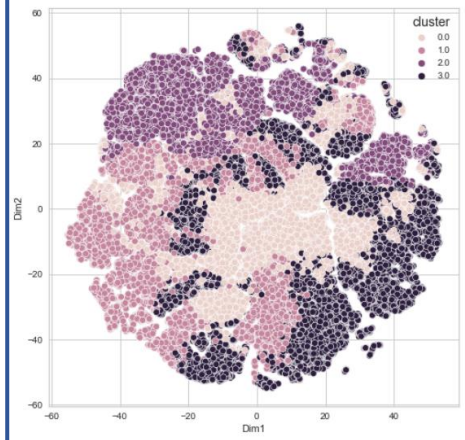
Algorithme K-means: caractérisation et visualisation des clusters

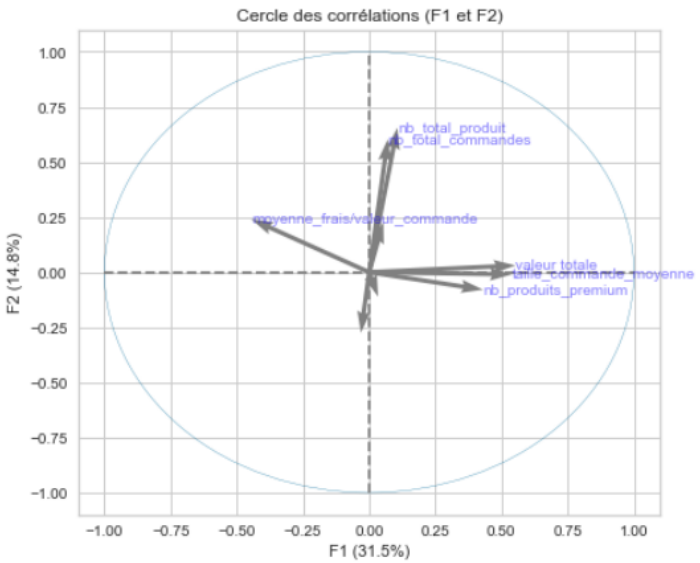
Visualisation des clusters

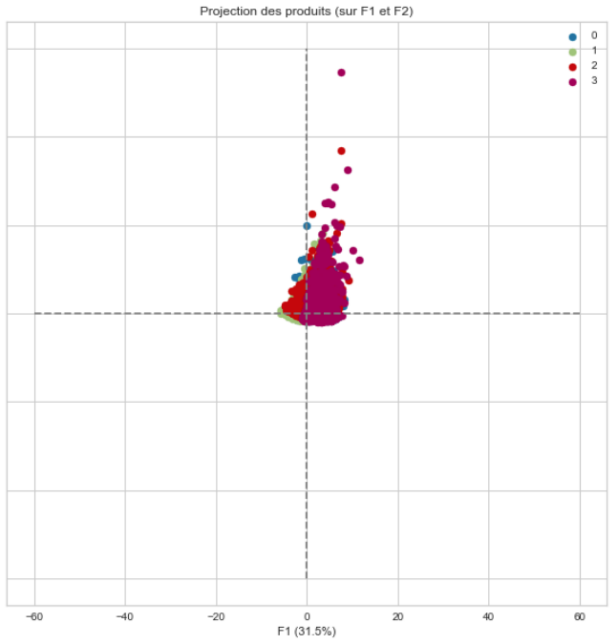
PCA



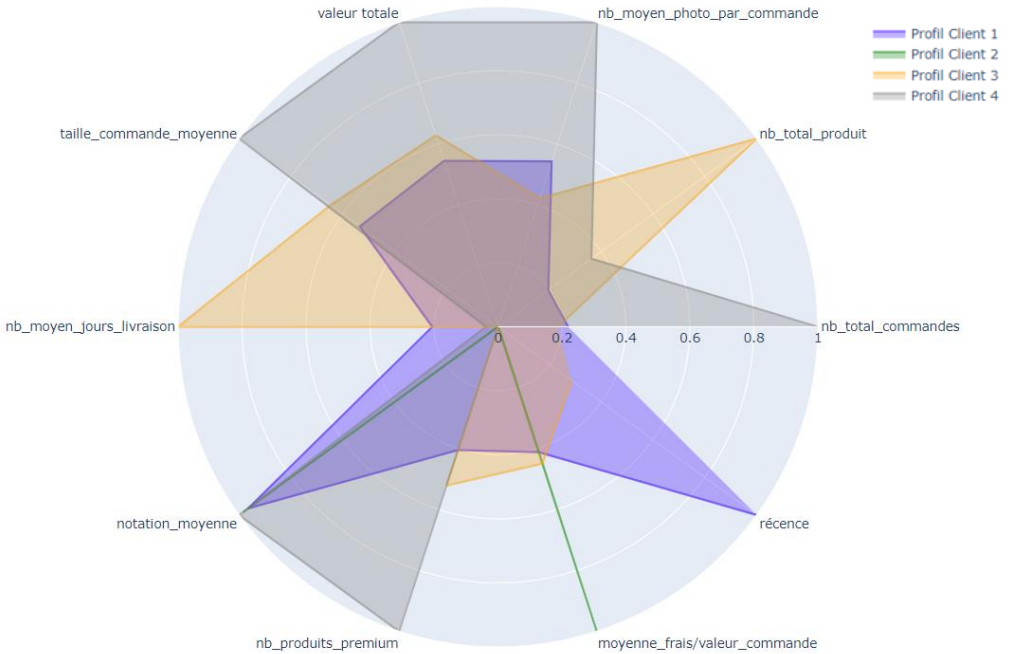
t-SNE







Caractérisation profils clients



Profil Client 1:

- Anciens clients
- Valeurs totales et tailles de commandes moyennes
- Ils achètent parfois produits premium
- Pas trop éloignées
- Clients satisfaits

Profil Client 2:

- Clients récents
- Valeurs totales et tailles de commandes petites
- Plutôt éloignées
- Clients satisfaits

Profil Client 3:

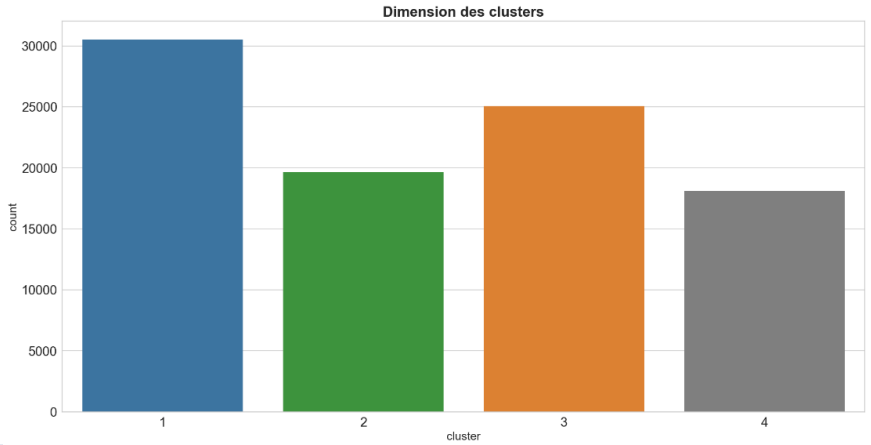
- Clients relativement récents
- Valeurs totales et tailles de commandes moyennes
- Pas trop éloignées
- Clients moins satisfaits

Profil Client 4:

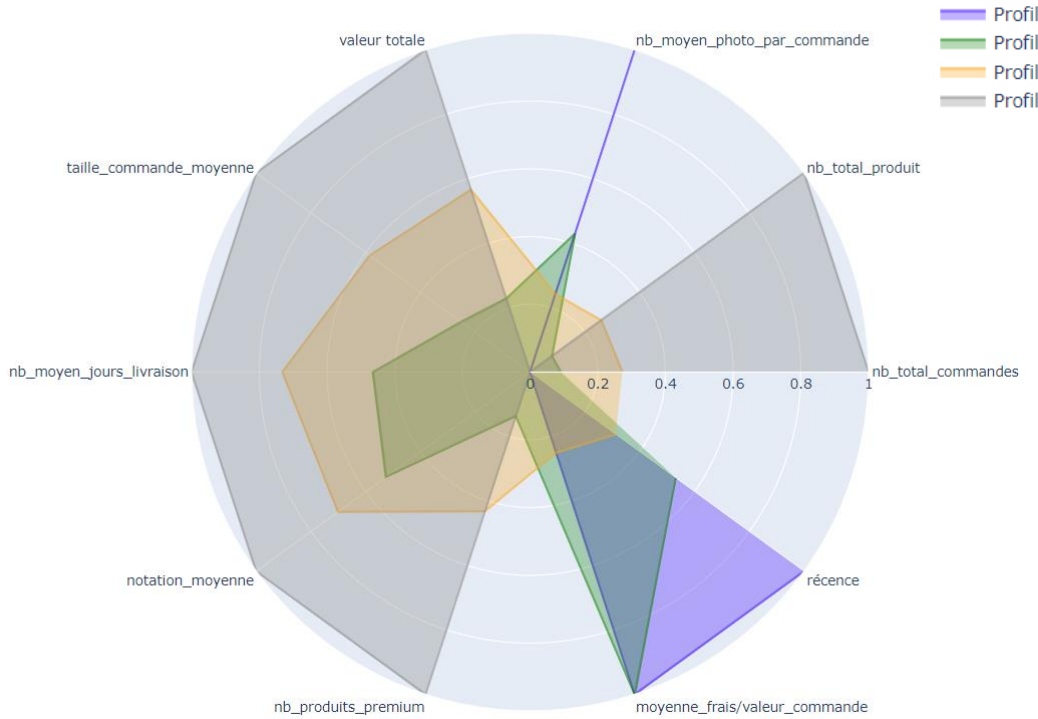
- Clients récents
- Valeurs totales et tailles de commandes élevées
- Ils achètent aussi des produits premium
- Pas trop éloignés
- Clients satisfaits.

Critères de regroupement des variables

Variable	Critère	Valeurs	Variable finale
nb_total_commandes	binning	[1,2,3,4,>4] -> [1,2,3,4,5]	nb_total_commandes_quant
nb_total_produit	binning	[1,2,3,4,>4] -> [1,2,3,4,5]	nb_total_produit_quant
nb_moyen_photo_par_ commande	binning	[0,1],]1,2],]2,3],]3,4],]4,>4] -> -> [1,2,3,4,5]	nb_moyen_photo_ par_commande_quant
notation_moyenne	binning	[0,1],]1,2],]2,3],]3,4],]4,5] -> -> [1,2,3,4,5]	notation_moyenne_quant
nb_produits_premium	binning	[1,2,3,4,>4] -> [1,2,3,4,5]	nb_produits_premium_quant
nb_moyen_jours_livraison	quantiles	[1,2,3,4,5]	nb_moyen_jours_livraison_quant
valeur_totale	quantiles	[1,2,3,4,5]	valeur_totale_quant
taille_commande_moyenne	quantiles	[1,2,3,4,5]	taille_commande_moyenne_ quant
moyenne_frais/ valeur_commande	quantiles	[1,2,3,4,5]	moyenne_frais/ valeur_commande_quant
récence	quantiles	[5,4,3,2,1]	récence_quant
score= $\sum$ Variables finales	quantiles	[1,2,3,4]	cluster



Caractérisation profils clients



Profil Client 1:

- Clients récents
- Valeurs totales et tailles de commandes moyennes/basses
- Plutôt éloignés
- Clients moins satisfaits

Profil Client 2:

- Clients plutôt récents
- Valeurs totales et tailles de commandes moyennes/basses
- Ils achètent parfois des produits premium
- Plutôt éloignées
- Clients moyennement satisfaits

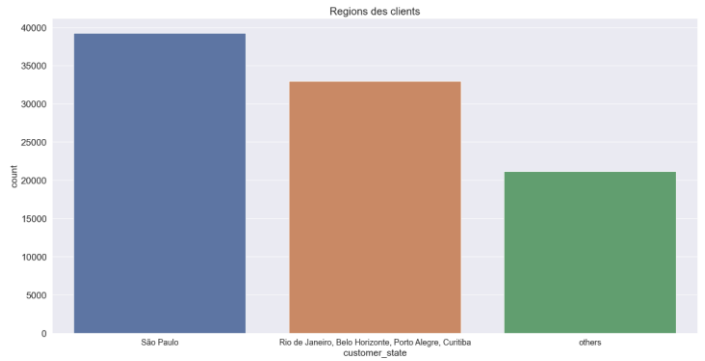
Profil Client 3:

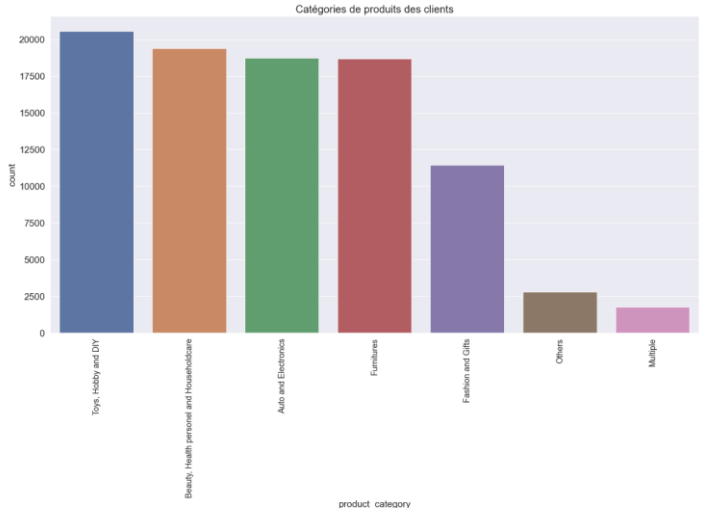
- Clients plutôt anciens
- Valeurs totales et tailles de commandes moyennes/élevées
- Pas trop éloignées
- Clients satisfaits

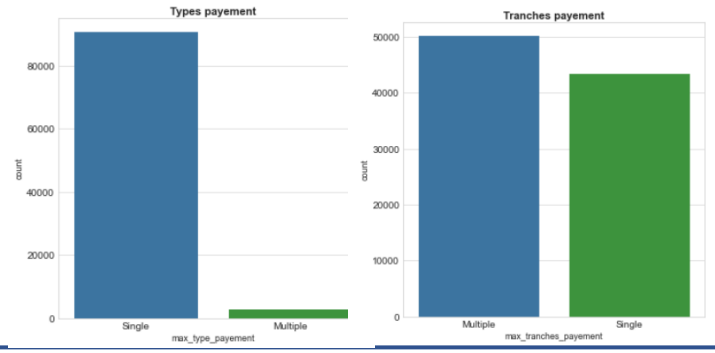
Profil Client 4:

- Clients anciens
- Valeurs totales et tailles de commandes élevées
- Ils achètent aussi des produits premium
- Pas trop éloignés
- Clients satisfaits.

Jeu de données







Algorithme K-modes

Step 1 Determine the value 'K'

Step 2 Modes initialization

Hamming distance

$$d(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j)$$

where

$$\delta(x_j,y_j)=\begin{cases}0, & x_j=y_j\\ 1, & x_j\neq y_j.\end{cases}$$

Step 3 Mesure the distance from modes

Step 4 Assign to the nearest mode

Step 5 New modes initialization

Max number of iterations

Step 6 Not Max number of iterations

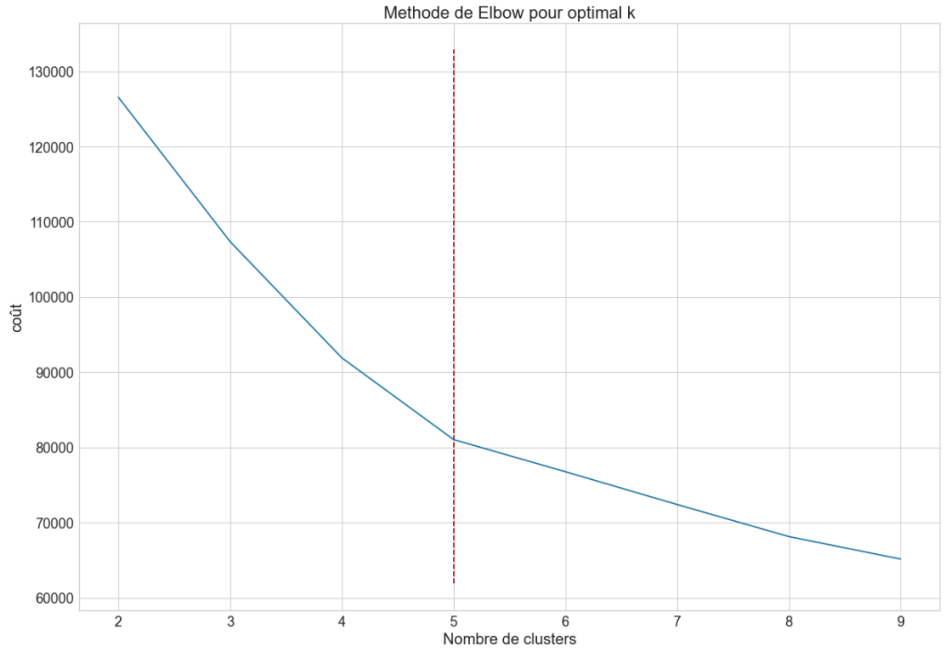
Step 7 Evaluate cost

Step 8 Repeat until get the lowest cost

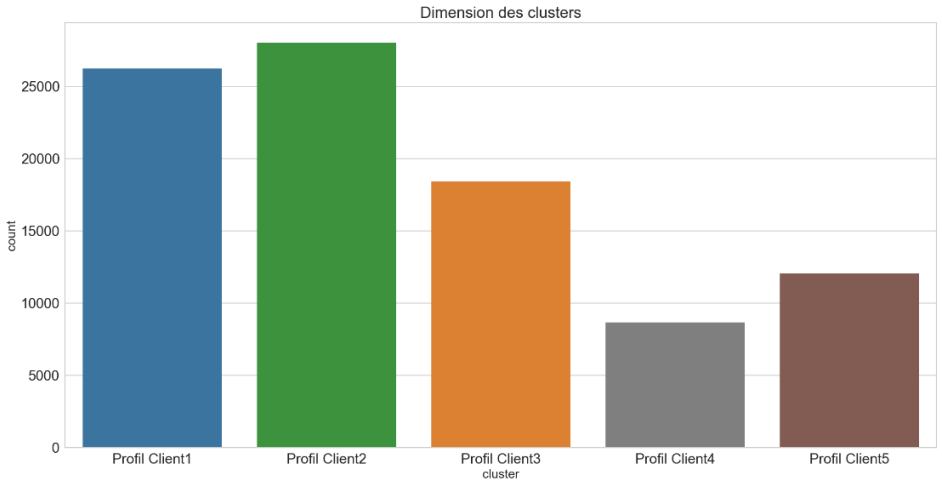
Get the lowest cost

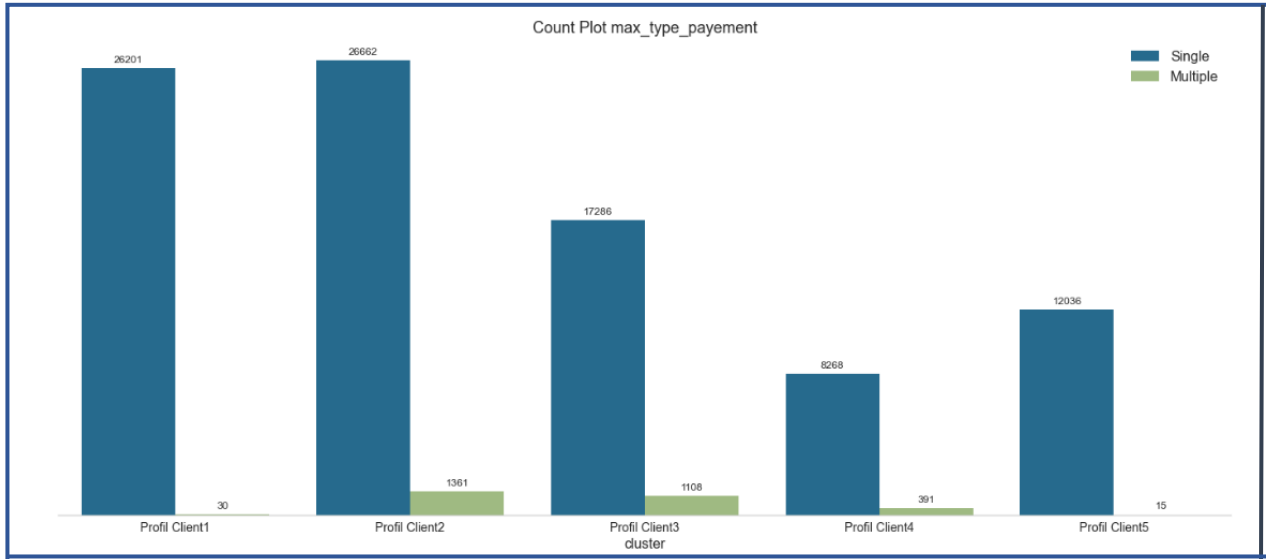
Clustering result

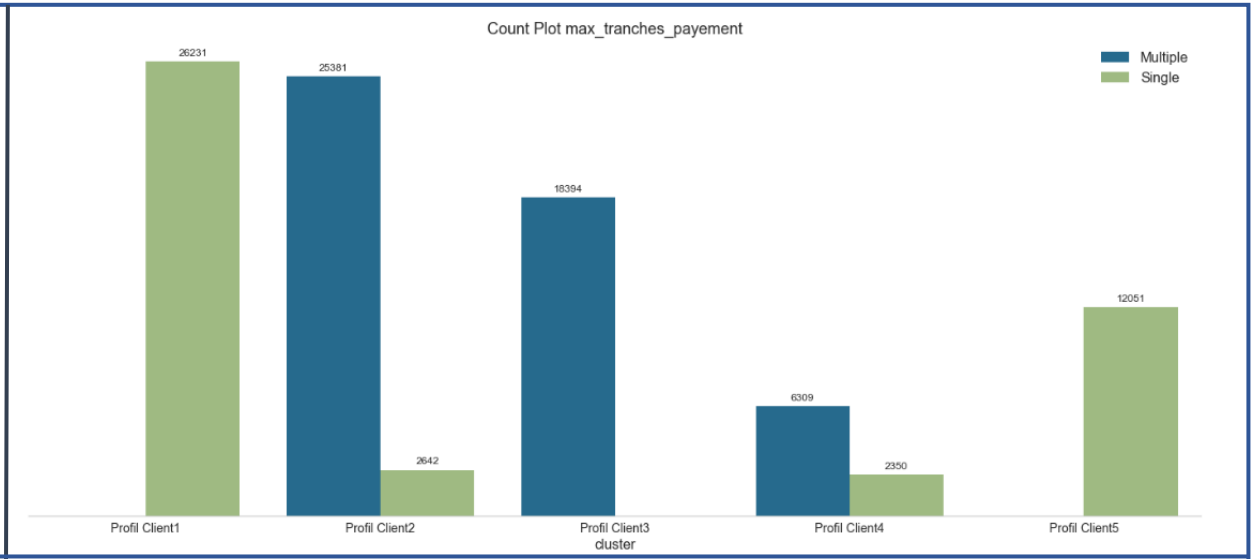
Définition nombre de cluster

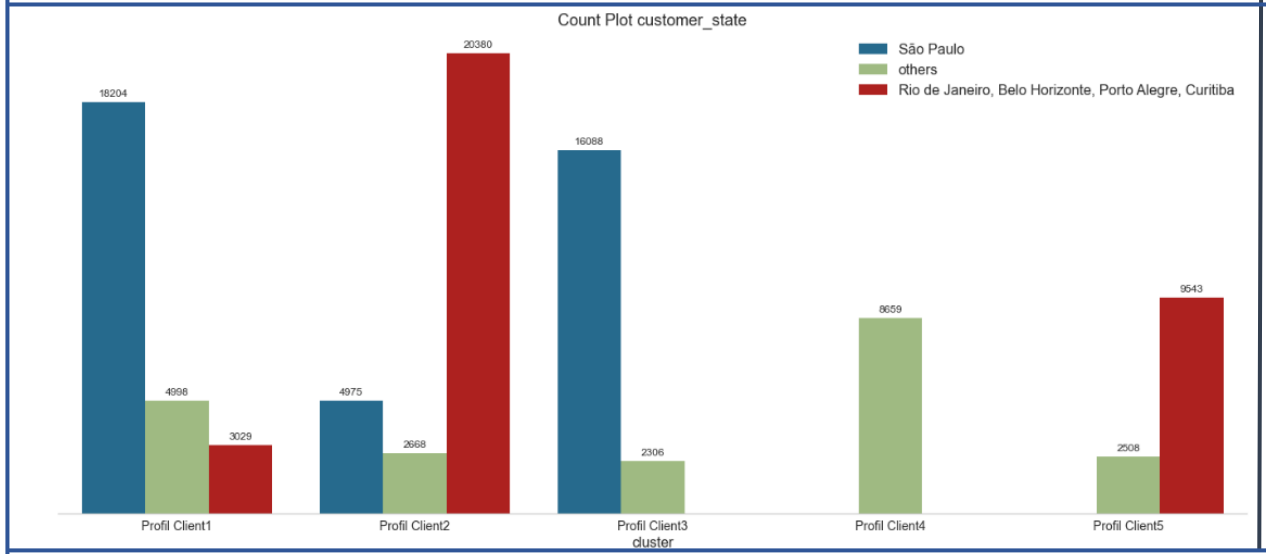


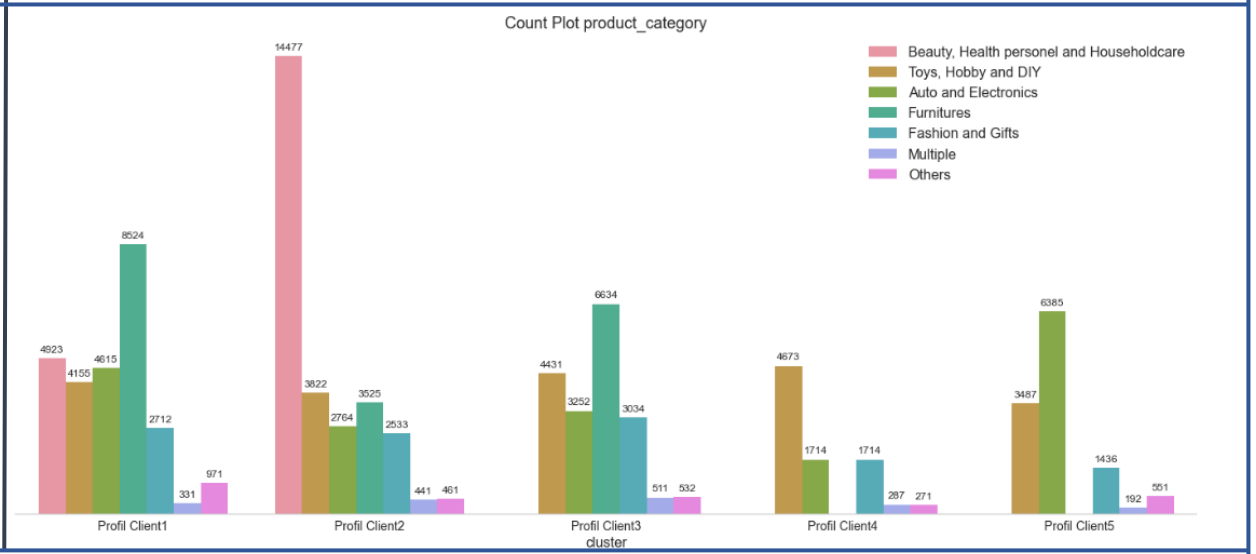
Clustering result











Profil Client 1:
- Clients principalement de la région de Sao Paulo
- Ils payent en une seule tranche avec un seul moyen de payement
- Ils achètent surtout fournitures pour la maison

Profil Client 2:
- Clients principalement des régions des grandes villes
- Ils payent en une plusieurs tranches avec un seul moyen de payement
- Ils achètent surtout produits pour la beauté, santé et soins ménagers

Profil Client 3:
- Clients principalement de la région de Sao Paulo
- Ils payent en plusieurs tranches, avec un seul moyen de payement
- Ils achètent surtout fournitures pour la maison et produits pour hobby et bricolage

Profil Client 4:
- Clients qui n'habitent pas dans les région des grands villes
- ls payent principalement en plusieurs tranches, avec un seul moyen de payement
- Ils achètent surtout produits pour hobby et bricolage.

Profil Client 5:
- Clients principalement régions des grandes villes
- Ils payent en une seule tranche avec un seul moyen de payement
- Ils achètent surtout auto et produits hi-tech et produits pour hobby et bricolage.

Algorithme K-means - stabilité temporelle du clustering

Intervalle des commandes dans le jeu de données: 23 mois

Période initiale			Evolution ARI dans le mois suivants										
Nr. mois	Nr. clients	Début-fin analyse	Oct 2017	Nov 2017	Dec 2017	Jan 2018	Fév 2018	Mar 2018	Apr 2018	Mai 2018	Juin 2018	Juil 2018	Aout 2018
12	23647	Sep 2017	0,97	0,93	0,4	0,36	0,4	0,38	0,35	0,34	0,34	0,34	0,34
13	27635	0ct 2017		0,96	0,93	0,38	0,33	0,33	0,33	0,32	0,31	0,31	0,31
14	32129	Nov 2017			0,96	0,93	0,33	0,33	0,33	0,32	0,32	0,32	0,32
15	40264	Dec 2017				0,96	0,91	0,88	0,33	0,31	0,33	0,32	0,33
16	45461	Jan 2018					0,96	0,93	0,9	0,34	0,33	0,33	0,33
17	52268	Fév 2018						0,96	0,91	0,9	0,35	0,35	0,35
18	58696	Mar 2018							0,94	0,94	0,9	0,37	0,37
19	65295	Apr 2018								0,94	0,95	0,93	0,37
20	72644	Mai 2018									0,96	0,92	0,87
21	78265	Juin 2018										0,96	0,91
22	83397	Juil 2018											0,96

Période de maintenance pour le programme de segmentation: 4 mois