



# SEGMENTEZ DES CLIENTS D'UN SITE E-COMMERCE

---

**DATA SCIENCE**

# MÉTHODOLOGIE

---

I. PRESENTATION

II. DESCRIPTION DES DONNEES

III. DATA CLEANING

IV. ANALYSE DES DONNEES

V. FEATURES ENGINEERING

VI. MODELISATION

VII. EVALUATION DU MODELE

VIII. CONCLUSION

# PRESENTATION

---

## Contexte

- Fournir aux équipes d'e-commerce de Olist une segmentation des clients .
- Pourvoir faire leurs campagnes de communication.

## Objectifs

- Faire la segmentation des clients
- Déterminer la fréquence nécessaire de mise à jour du modèle de segmentation

## Contraintes

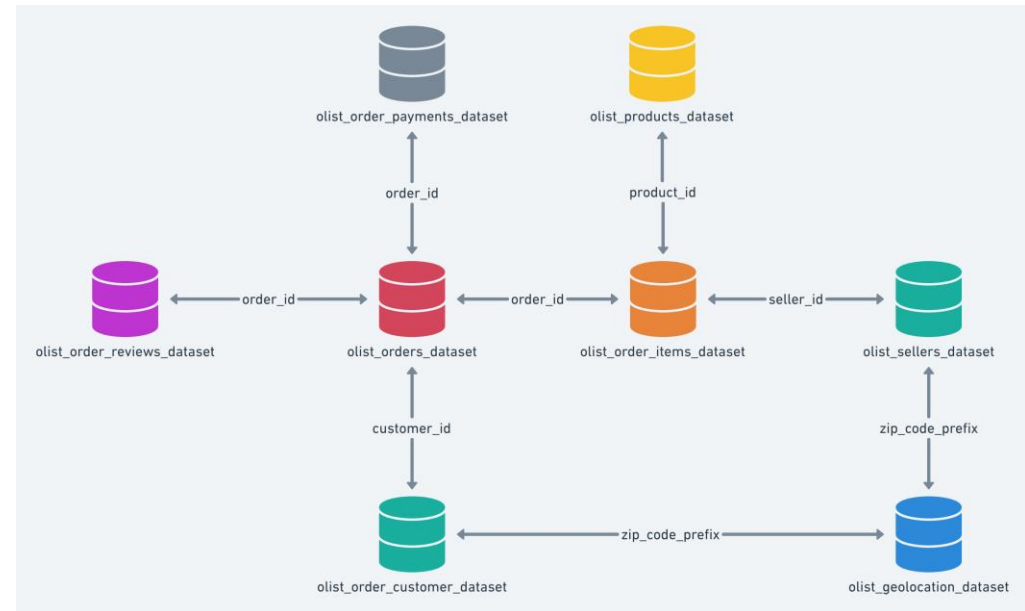
- La base de données est composée de plusieurs fichiers
- Les données sont non-supervisées et contiennent plusieurs valeurs manquantes et doublons

# DESCRIPTION DES DONNEES

## Base de données

Huit 09 fichiers dont :

- 01 fichier produit,
- 01 fichier catégorie produit,
- 01 fichier géolocalisation,
- 02 fichiers commande,
- 01 fichier acheteur
- 01 fichier paiement,
- 01 fichier commentaire,
- 01 fichier vendeur





# DESCRIPTION DES DONNEES

## Fusion en Dataframe unique

Elle consiste en deux (2)  
phases :

- Phase 1,
- Phase 2.

Une Dataframe centrée sur les  
commandes

Int64Index: 102652 entries, 0 to 102651

Data columns (total 17 columns):

#	Column	Non-Null	Count	Dtype
0	order_id	102652	non-null	object
1	customer_id	102652	non-null	object
2	customer_unique_id	102652	non-null	object
3	customer_state	102652	non-null	object
4	customer_city	102652	non-null	object
5	seller_id	102652	non-null	object
6	seller_city	102652	non-null	object
7	product_id	102652	non-null	object
8	product_size	102636	non-null	object
9	product_category	101633	non-null	object
10	order_approved_at	102638	non-null	object
11	order_estimated_delivery_date	102652	non-null	object
12	order_status	102652	non-null	object
13	order_delivered_customer_date	100409	non-null	object
14	price	102652	non-null	float64
15	payment_value	102651	non-null	float64
16	review_score	101855	non-null	float64

dtypes: float64(3), object(14)

# DATA CLEANING

---

Il a consisté à :

- ❖ Sélectionner les variables pertinentes
- ❖ Eliminer les valeurs manquantes,
- ❖ Supprimer des Variables
- ❖ Ajouter des Variables

	Données initiales	Données finales
Nbre de Lignes	102.652	99.401
Nbre de colonnes	17	17

# DATA CLEANING

---

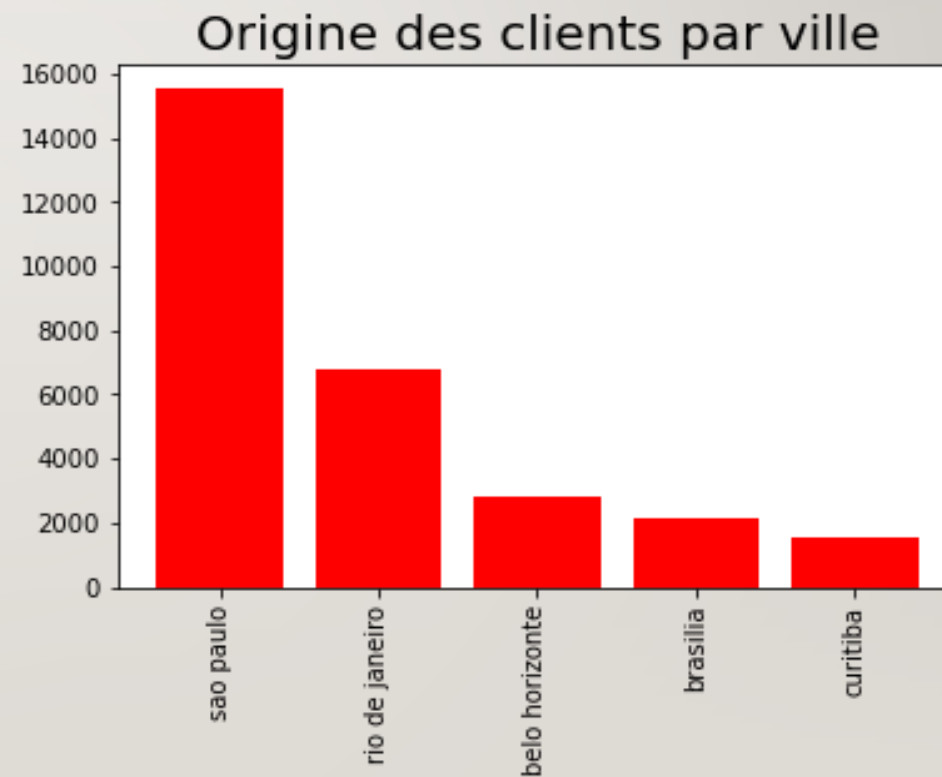
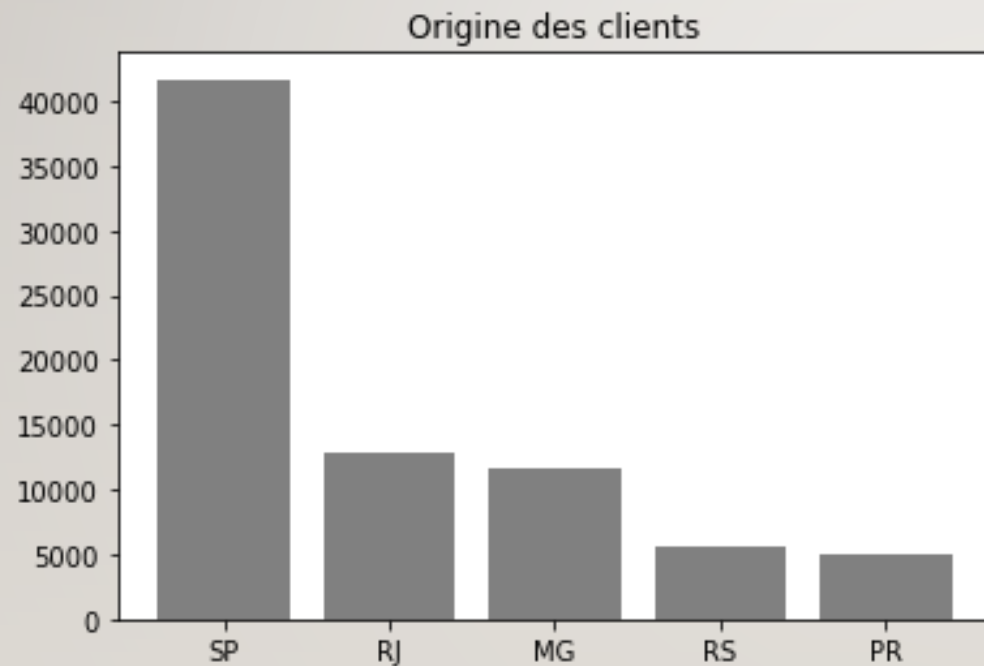
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 99401 entries, 0 to 102651
Data columns (total 17 columns):
#   Column                Non-Null Count  Dtype
---  -
0   order                  99401 non-null  object
1   customer_unique_id     99401 non-null  object
2   customer_state         99401 non-null  object
3   customer_city          99401 non-null  object
4   seller_id              99401 non-null  object
5   seller_city            99401 non-null  object
6   product_id             99401 non-null  object
7   product_size           99401 non-null  object
8   product_category       99401 non-null  object
9   order_status           99401 non-null  object
10  price                  99401 non-null  float64
11  payment_value          99401 non-null  float64
12  score                  99401 non-null  float64
13  order_approved_Date    99401 non-null  object
14  order_approved_hour    99401 non-null  object
15  Time(day)              99401 non-null  int32
16  Delay(day)             99401 non-null  int32
dtypes: float64(3), int32(2), object(12)
memory usage: 12.9+ MB
```

# ANALYSE DES DONNEES

## Analyse Exploratoire

---

### Analyse des variables

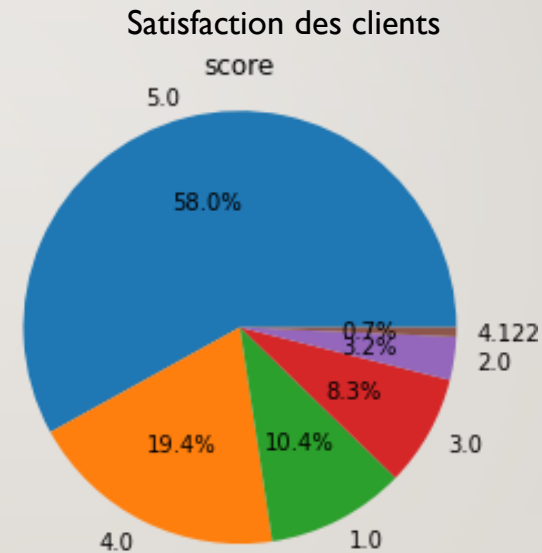




# ANALYSE DES DONNEES

## Analyse Exploratoire

### Analyse des variables



- ❑ Trois catégories de produits sont plus commandes sur les dix(10)
- ❑ Un taux de satisfaction meilleur dans l'ensemble

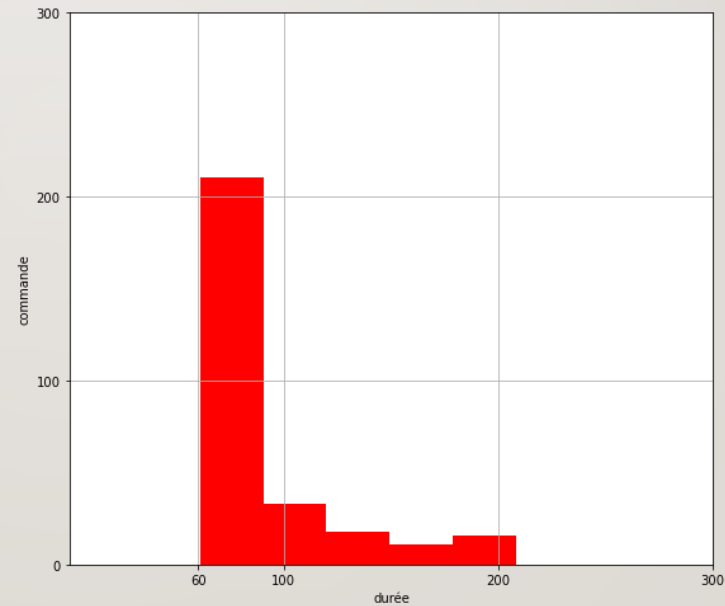
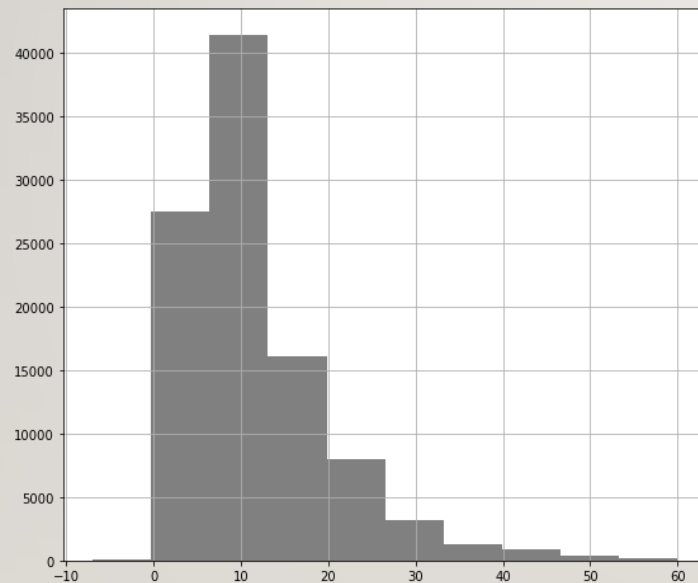
# ANALYSE DES DONNEES

## Analyse Exploratoire

---

### Analyse des variables

Nombre de commandes par intervalle de durée



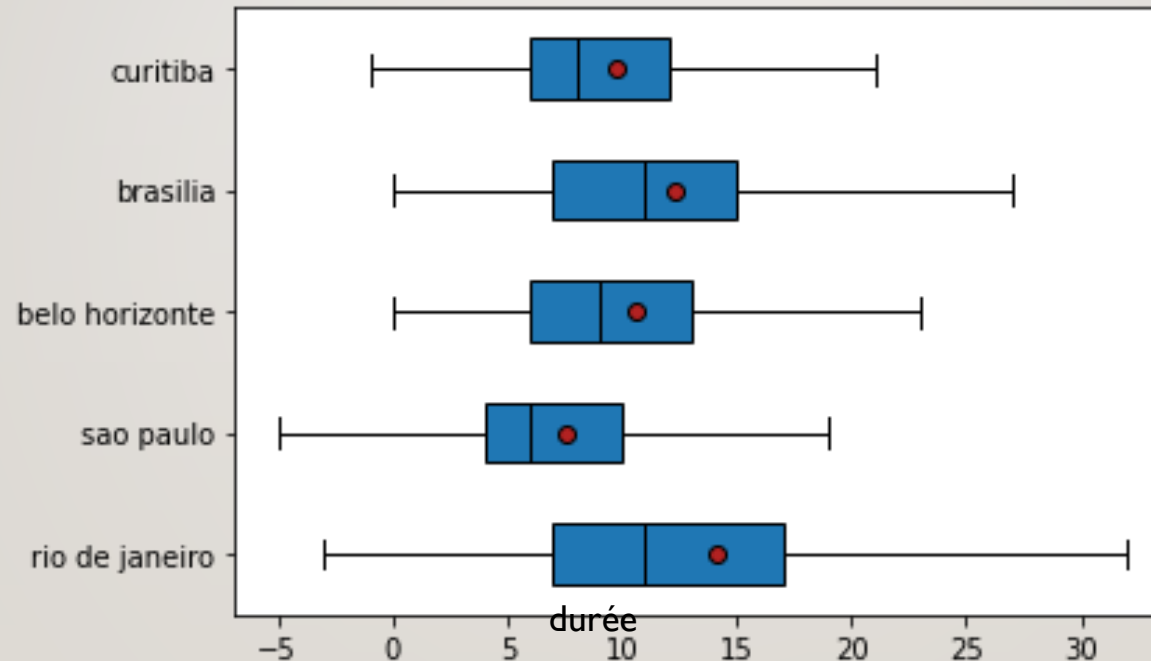
La majorité des commandes ont été livrées avant vingt (20) jours

# ANALYSE DES DONNEES

## Analyse Exploratoire

---

### Analyse Bivariée



La durée de livraison est relativement longue pour la ville de '**rio de Janeiro**'

# ANALYSE DES DONNEES

## Analyse Exploratoire

---

### Analyse des corrélations



Il y'a une forte corrélation entre les variables '**price**' et '**payment\_value**'



# ANALYSE DES DONNEES

## Analyse RFM

---

La table RFM

	customer_unique_id	last_order_date	order_count	total_spent
<b>87659</b>	f2aa8ff9b2dca8ccfc255b8bceda4e0c	2018-07-23	1	68.43
<b>34175</b>	5e5930a6046d05c80794ca72464762ec	2018-01-23	1	35.69
<b>85888</b>	ed9804f42ea58ef893439f6cd0355cf8	2018-08-22	1	112.77
<b>69188</b>	bf432226a944503e91c6ec2159c6663e	2018-02-01	1	87.38
<b>7090</b>	137979cbe1d895efb140ca1aad7e9915	2017-10-17	1	58.62
<b>53275</b>	93c19a2a1d633765971086fc540a1d36	2018-02-07	1	55.61
<b>70076</b>	c1b4427dc13af0fa51d44306cdef0e3e	2017-10-22	1	77.89

# ANALYSE DES DONNEES

## Analyse RFM

---

### Calcul de score

	customer_unique_id	last_order_date	order_count	total_spent	recency	frequency	monetary
8174	16537a831ca74e633b94cfa47164c420	2017-10-01	1	864.13	332	0.066667	0.054080
75043	cfb1b50c3f73e9a4da0186f0a7d959a0	2018-06-23	1	112.09	67	0.066667	0.007015
76351	d31233132950973bd0df88d4523ed465	2017-02-01	1	237.99	574	0.066667	0.014894
28808	4f916ba67d94ea608ff6915700595b49	2018-08-07	1	208.14	22	0.066667	0.013026
29879	529a8faa9d37064b655b398164e4fda7	2018-04-12	1	88.31	139	0.066667	0.005527
14370	27a86b0297f39efe6b94046f95a5f4bf	2018-01-22	1	44.75	219	0.066667	0.002801
52596	91dda3852e8a4757ed73fc91360cfa5f	2017-11-05	2	415.60	297	0.133333	0.026010
28972	50091850aa6ef7671af8c3e0e6762e1c	2017-10-18	1	158.52	315	0.066667	0.009921
39988	6e6fabd95676ab9f6fe71473894669ed	2017-12-11	1	144.12	261	0.066667	0.009020
32719	5a48e68daf6419abb65cd6fd1c03dda3	2017-06-10	1	97.10	445	0.066667	0.006077

# ANALYSE DES DONNEES

## Analyse RFM

---

### Catégorisation

	customer_unique_id	last_order_date	order_count	total_spent	recency	frequency	monetary	category
0	0000366f3b9a7992bf8c76cfd3221e2	2018-05-10	1	141.90	111	0.066667	0.008881	6
1	0000b849f77a49e4a4ce2b2a4ca5be3f	2018-05-07	1	27.19	114	0.066667	0.001702	6
2	0000f46a3911fa3c0805444483337064	2017-03-10	1	86.22	537	0.066667	0.005396	8
3	0000f6ccb0745a6a4b88665a16c9f078	2017-10-12	1	43.62	321	0.066667	0.002730	2
4	0004aac84e0df4da2b147fca70cf8255	2017-11-14	1	196.89	288	0.066667	0.012322	2

# FEATURES ENGINEERING

---

- ❖ Transformation de la dataframe
- ❖ Analyse en Composantes Principales (ACP)



# FEATURES ENGINEERING

## Transformation de la dataframe

---

	customer_unique_id	customer_state	order_count	total_spent	average_score	large	medium	small	Automotive_Industry	Beauty_Health_Well
0	871766c5855e863f6eccc05f988b23cb	RJ	1	72.19	5.0	0.0	0.0	1.0	0.0	
1	eb28e67c4c0b83846050ddfb8a35d051	SP	2	284.56	4.5	1.0	0.0	1.0	0.0	
2	3818d81c6709e39d06b2738a8d3a2474	MG	1	216.87	5.0	0.0	0.0	1.0	0.0	
3	af861d436cfc08b2c2ddefd0ba074622	SP	1	25.78	4.0	0.0	0.0	1.0	0.0	
4	64b576fb70d441e8f1b2d7d446e483c5	SP	1	218.04	5.0	0.0	0.0	1.0	0.0	

# FEATURES ENGINEERING

## Analyse en Composantes Principales (ACP)

---

	customer_unique_id	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
0	871766c5855e863f6eccc05f988b23cb	-0.453173	-0.475505	0.908050	-0.198068	1.982270	-1.692184	-0.218758	-0.570270
1	eb28e67c4c0b83846050ddfb8a35d051	0.563502	3.187697	2.066192	3.404257	1.274128	0.622491	-0.392839	-2.069185
2	3818d81c6709e39d06b2738a8d3a2474	0.318968	0.335463	-0.733735	-0.975728	0.005529	-0.745319	-0.194401	-0.192341
3	af861d436cfc08b2c2ddefd0ba074622	0.373827	-1.228459	1.049602	-0.339624	-0.995308	-0.327800	-0.064375	-0.065938
4	64b576fb70d441e8f1b2d7d446e483c5	-0.167254	0.000962	0.349539	-0.149792	-0.201059	1.940684	-0.518369	-2.227760

# MODELISATION

---

Nous allons tester trois algorithmes de clustering et ensuite choisir le mieux adapter nos données

❖ **K-means**

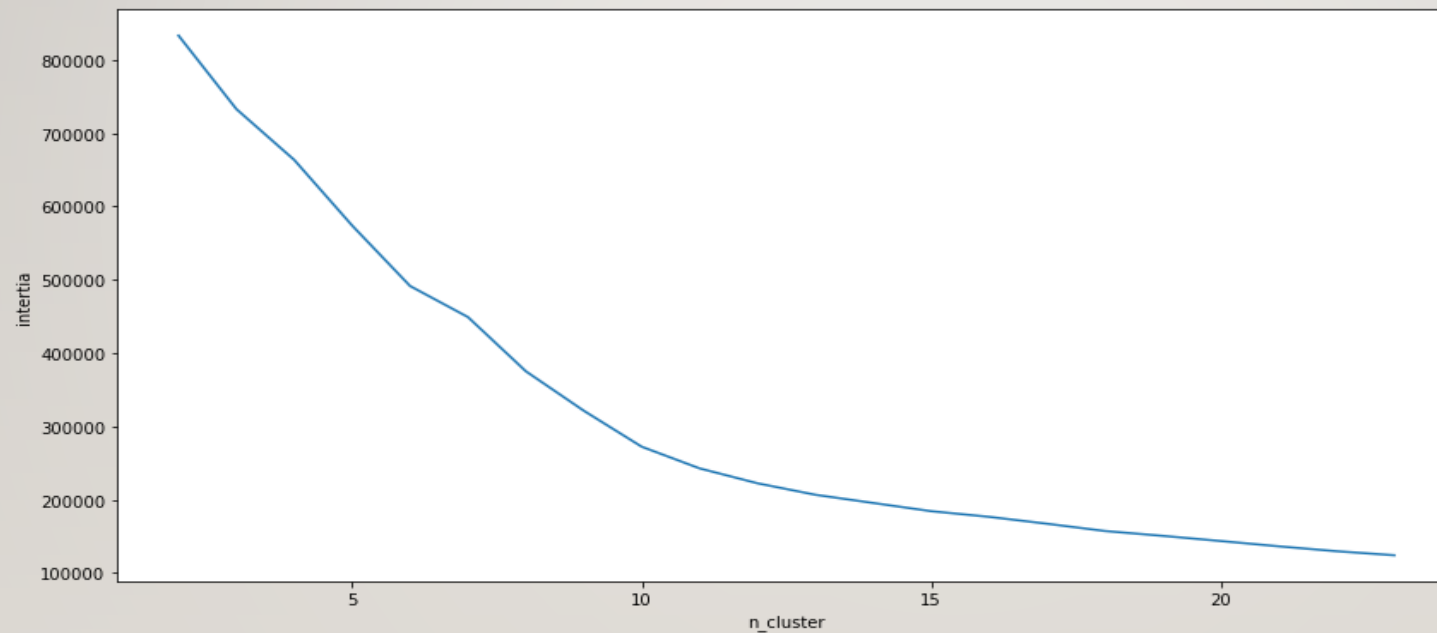
❖ **DBSCAN**

❖ **Agglomerative clustering**

# MODELISATION

## K-MEANS

La méthode du coude



Le coude est observé pour un nombre de cluster = 10



# MODELISATION

## DBSCAN

---

### Recherche des hyperparamètre

eps=0.1, min\_samples=2000, n\_clusters=7  
eps=0.1, min\_samples=4000, n\_clusters=3  
eps=0.1, min\_samples=6000, n\_clusters=1  
eps=0.1, min\_samples=8000, n\_clusters=0  
eps=0.1, min\_samples=10000, n\_clusters=0  
eps=0.3, min\_samples=2000, n\_clusters=8  
eps=0.3, min\_samples=4000, n\_clusters=5  
eps=0.3, min\_samples=6000, n\_clusters=3  
eps=0.3, min\_samples=8000, n\_clusters=3  
eps=0.3, min\_samples=10000, n\_clusters=3  
eps=0.6, min\_samples=2000, n\_clusters=9  
eps=0.6, min\_samples=4000, n\_clusters=7  
eps=0.6, min\_samples=6000, n\_clusters=3

eps=0.6, min\_samples=8000, n\_clusters=3  
eps=0.6, min\_samples=10000, n\_clusters=3  
eps=0.8, min\_samples=2000, n\_clusters=9  
eps=0.8, min\_samples=4000, n\_clusters=7  
eps=0.8, min\_samples=6000, n\_clusters=4  
eps=0.8, min\_samples=8000, n\_clusters=3  
eps=0.8, min\_samples=10000, n\_clusters=3  
eps=1, min\_samples=2000, n\_clusters=9  
eps=1, min\_samples=4000, n\_clusters=8  
eps=1, min\_samples=6000, n\_clusters=4  
eps=1, min\_samples=8000, n\_clusters=3  
eps=1, min\_samples=10000, n\_clusters=3

# MODELISATION

## Choix de l'Algorithme

---

### Qualité des clusters

#### K-means :

**clusters = 3** ; Score de silhouette moyen : **0.327**

**clusters = 10** ; Score de silhouette moyen : **0.652**

#### DBSCAN :

**clusters = 3** ; Score de silhouette moyen : **0.349**

**clusters = 9** ; Score de silhouette moyen : **0.646**

#### Agglomerative clustering :

**clusters = 3** ; Score de silhouette moyen : **0.003**

**clusters = 10** ; Score de silhouette moyen : **-0.10**

# MODELISATION

## Analyse des groupes de clusters

---

- ❖ Analyse des groupes de clustering par ANOVA
- ❖ Analyse des groupes avec les statistiques descriptives
- ❖ Analyse graphique des groupes du clustering

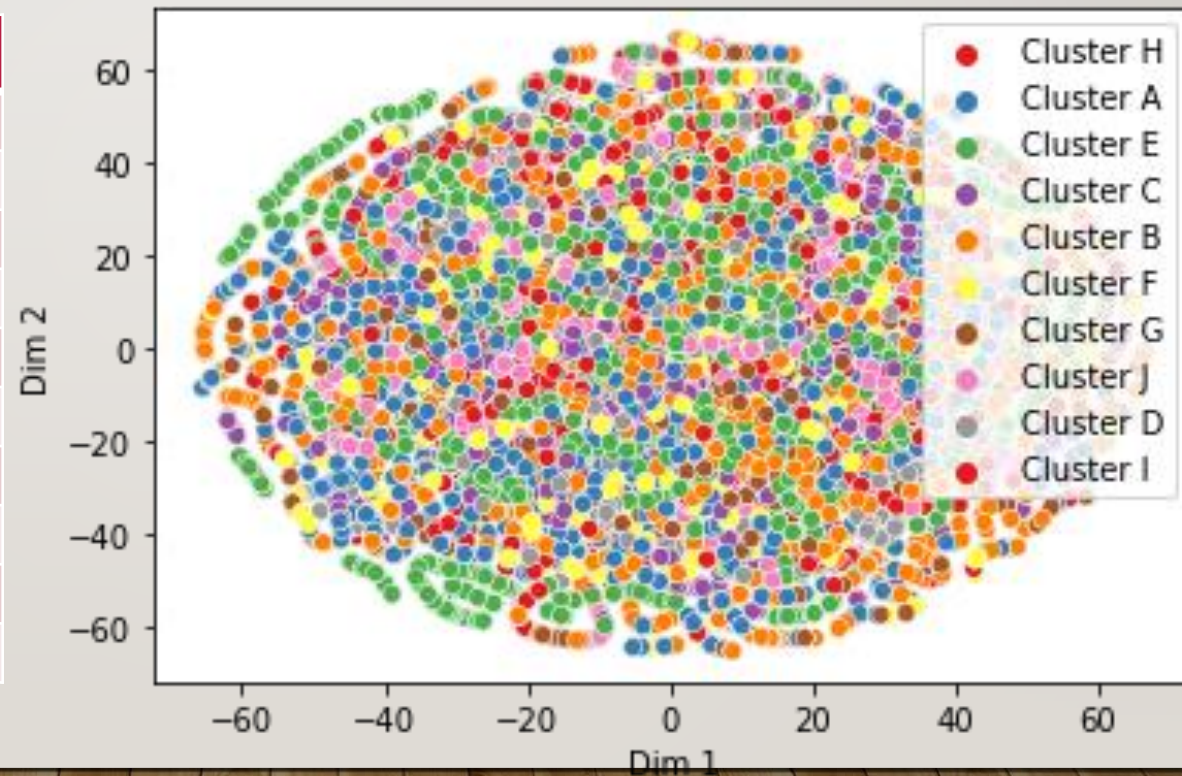


# MODELISATION

## Analyse des groupes de clusters

### Analyse statistique

Clusters	Nombre
Cluster E	18662
Cluster A	17426
Cluster B	15164
Cluster G	7375
Cluster F	6932
Cluster H	6931
Cluster C	6082
Cluster J	6002
Cluster D	5575
Cluster I	2295



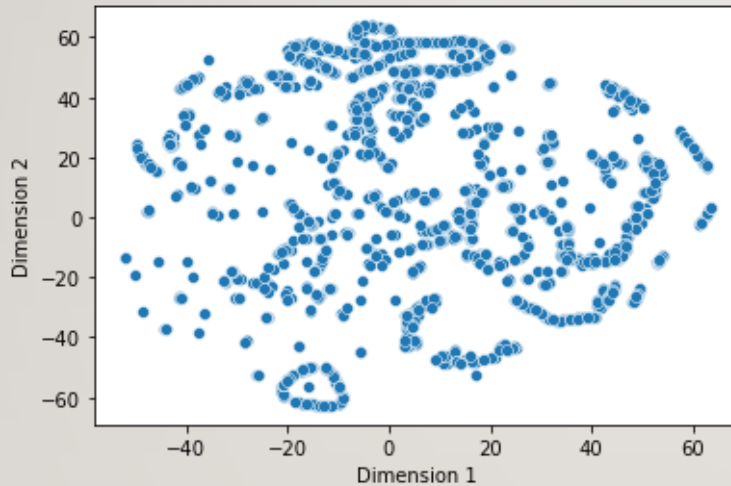


# MODELISATION

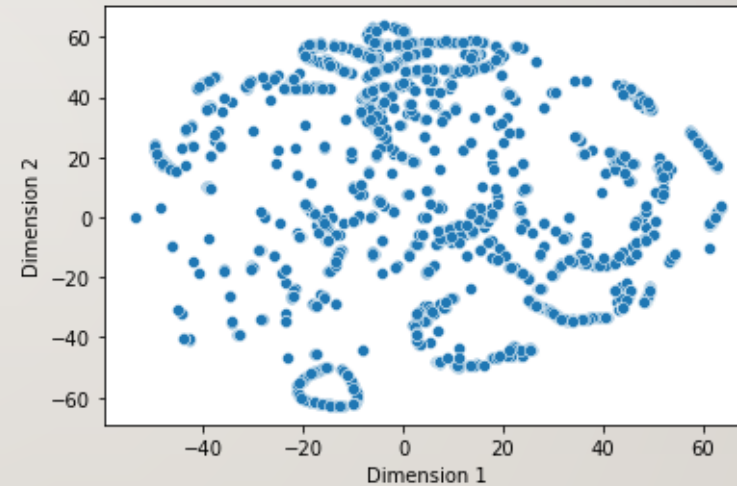
## Analyse des groupes de clusters

---

### Analyse Graphique



Groupe A



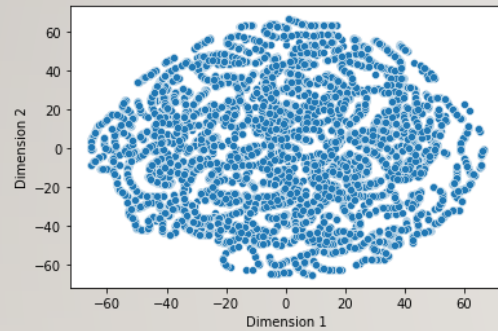
Groupe J

# MODELISATION

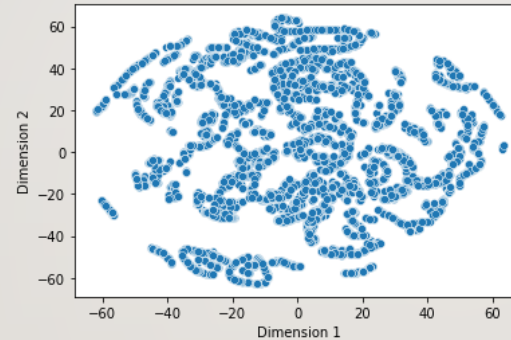
## Analyse des groupes de clusters

---

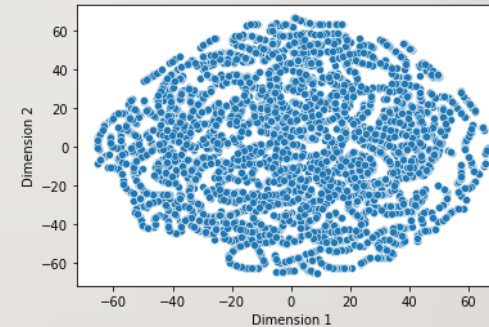
### Analyse Graphique



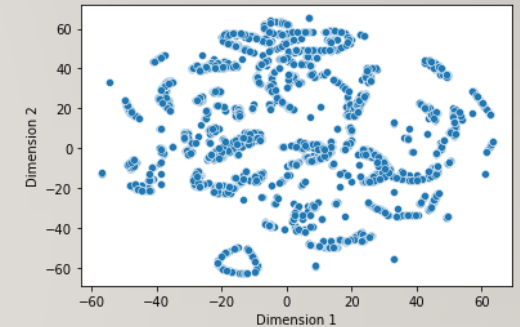
Groupe B



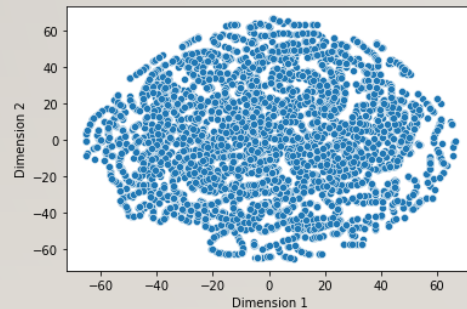
Groupe C



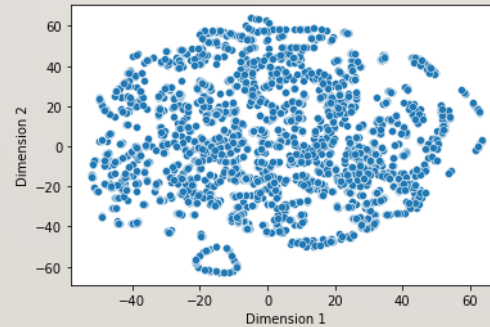
Groupe D



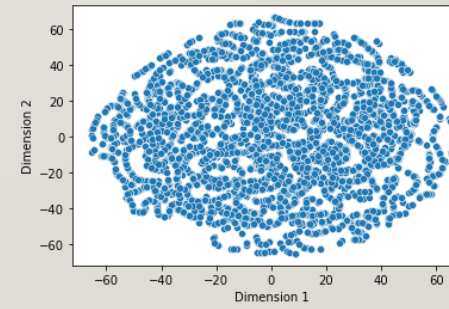
Groupe E



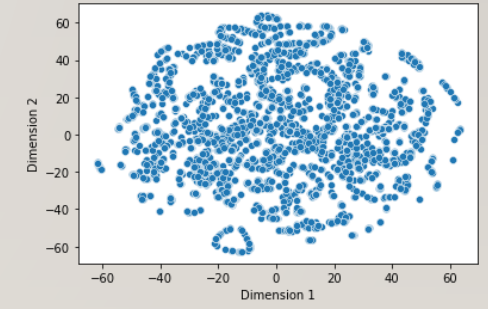
Groupe F



Groupe G



Groupe H



Groupe I

# MODELISATION

## Analyse des groupes de clusters

---

### Conclusion de l'analyse

- ❖ La majorité des clients de chacun des groupes sont issue de l'état SP et les groupes ont quasiment la même moyenne de satisfaction.
- ❖ Les groupes A et J ont des moyennes de commande et de dépenses plus élevées que celles des autres groupes qui ont moyenne de commande environ égale 1
- ❖ Les clients des groupes B, C, D et E commande uniquement que des produits de petite et moyenne taille ; le groupe G a une moyenne de commande en produit de grande taille largement supérieure
- ❖ Les clients des groupes A et J commande beaucoup plus les 3 catégories de produits les plus commandés que ceux des autres groupes qui en général ne commande que un (1) ou deux (2) de ces produits
- ❖ Les clients des groupes A et J sont les moins nombreux



# EVALUATION DU MODÈLE

---

L'évaluation consiste à déterminer le temps de mise à jour des clusters :

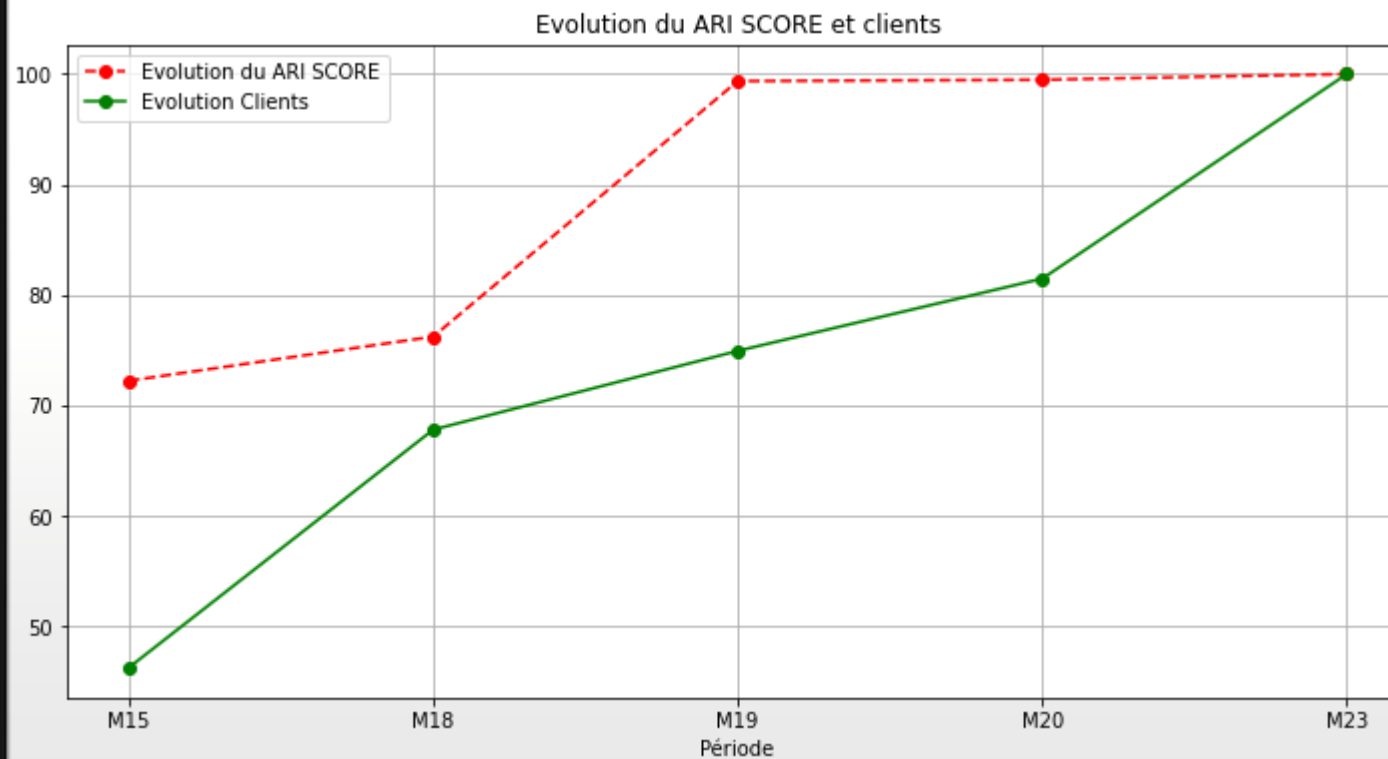
- ❖ CREATION DE DATAFRAME PERIODIQUE ;

- ❖ EVALUATION DE LA STABILITE;



# EVALUATION DU MODÈLE

---



# CONCLUSION

---

Une maintenance en vue de mettre à jour des clusters sera effectuée tous les quatre (4) mois

A large, solid blue arrow pointing downwards, connecting the first box to the second.


La majorité de ces clients est installée dans l'état SP

A large, solid blue arrow pointing downwards, connecting the second box to the third.

Deux groupes (A et J) se distinguent de façon significative des huit autres groupes existants dans la base de données des clients

# CONCLUSION

---



Périodicité de maintenance  
quatre (4) mois

Mener des campagnes  
marketing de fidélisations  
auprès des groupes A et J

Renforcer les campagnes  
d'incitation à l'achat auprès  
des autres groupes



MERCI

OPENCLASSROOM