OLIST

Solution de vente sur les market places en ligne

24082021 –
Catherine LE

« Pendant que vous êtes au repos, la compétition est en marche. changez la donne. Vendez sur internet avec olist store »

Compétences évaluées

- Mettre en place le modèle d'apprentissage non supervisé adapté au problème métier

- Transformer les variables pertinentes d'un modèle d'apprentissage non supervisé

- Adapter les hyper-paramètres d'un algorithme non supervisé afin de l'améliorer

- Évaluer les performances d’un modèle d'apprentissage non supervisé

Problématique

- Une **segmentation** selon les différents types d’utilisateurs à utiliser au quotidien pour les campagnes de communication.
- Comprendre les différents types d’utilisateurs grâce à leur **comportement** et à leurs **données personnelles**.

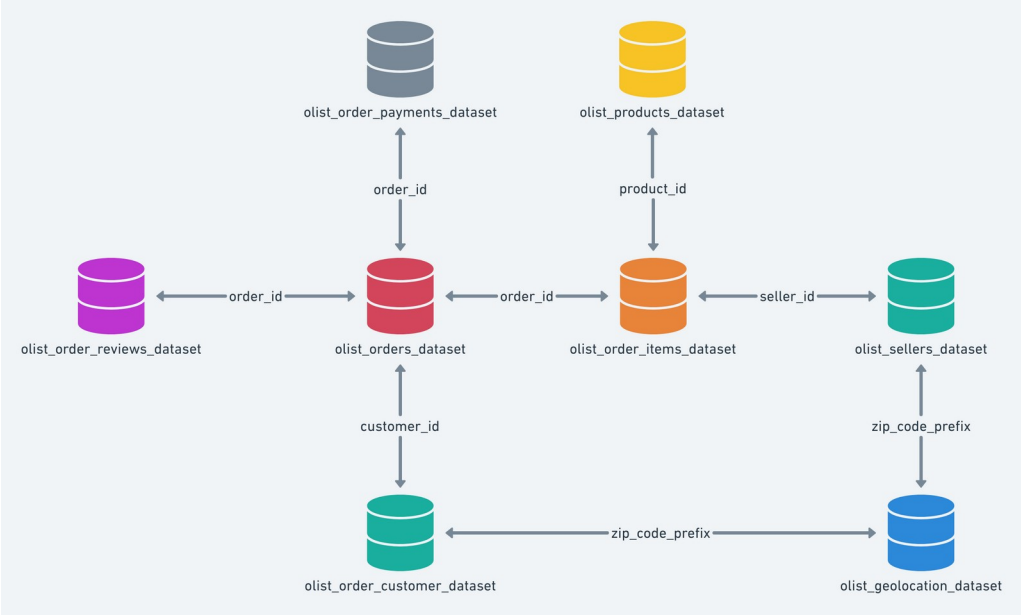
À fournir :

- une description actionnable de la segmentation
- une proposition de contrat de maintenance
- un code respectant la convention PEP8

Pistes de recherche

- Méthodes non supervisées sur les données normalisées
  - DBSCAN – détection des outliers pour le clustering
  - K-means comme baseline
  - K-means & AgglomerativeClustering
  - Log Regression avec régularisation l1 pour l’interprétation
- ACP en 3 composantes principales pour la visualisation des clusters
- Visualisation finale des clusters sur les données initiales non normalisées

Les données



raw_data (114312, 44)

Fusion des tables sur plusieurs clés :

- order_id
- product_id
- product_category_name
- customer_id
- seller_id
- geolocation_zip_code_prefix
- geolocation_city
- geolocation_state

Nettoyer les erreurs de formatage pour :

- customer_city
- geolocation_city
- seller_city

Features engineering

- Suppression des 11 variables non utiles
- Création de 7 nouvelles variables :

+ most_payment_type

+ order_count

+ payment_total

+ payment_mean

+ hod (heure de la commande)

+part_day

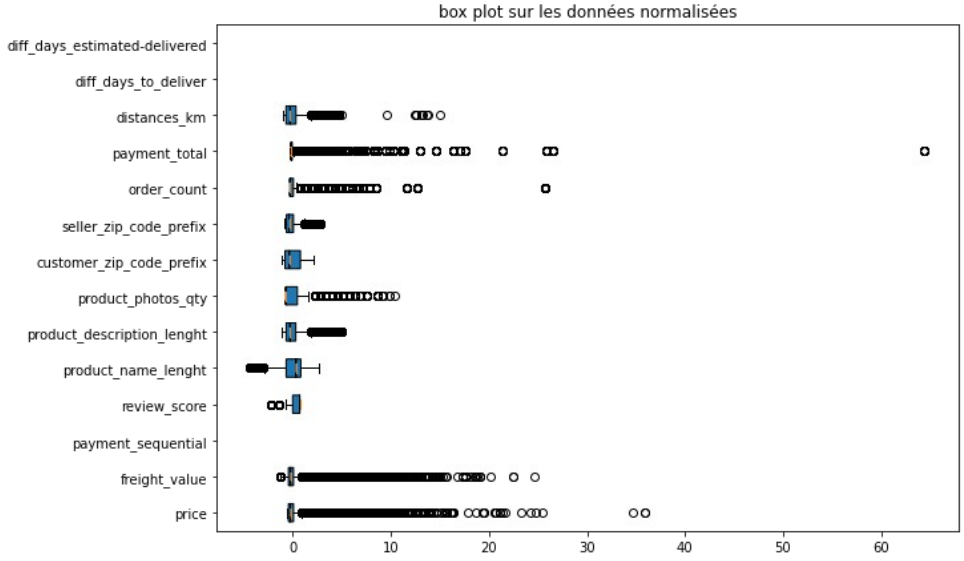
+distances_km

+diff_days_to_deliver

+diff_days_estimated_delivered

- Analyse de corrélation et suppression de 14 variables redondantes

Cleaning

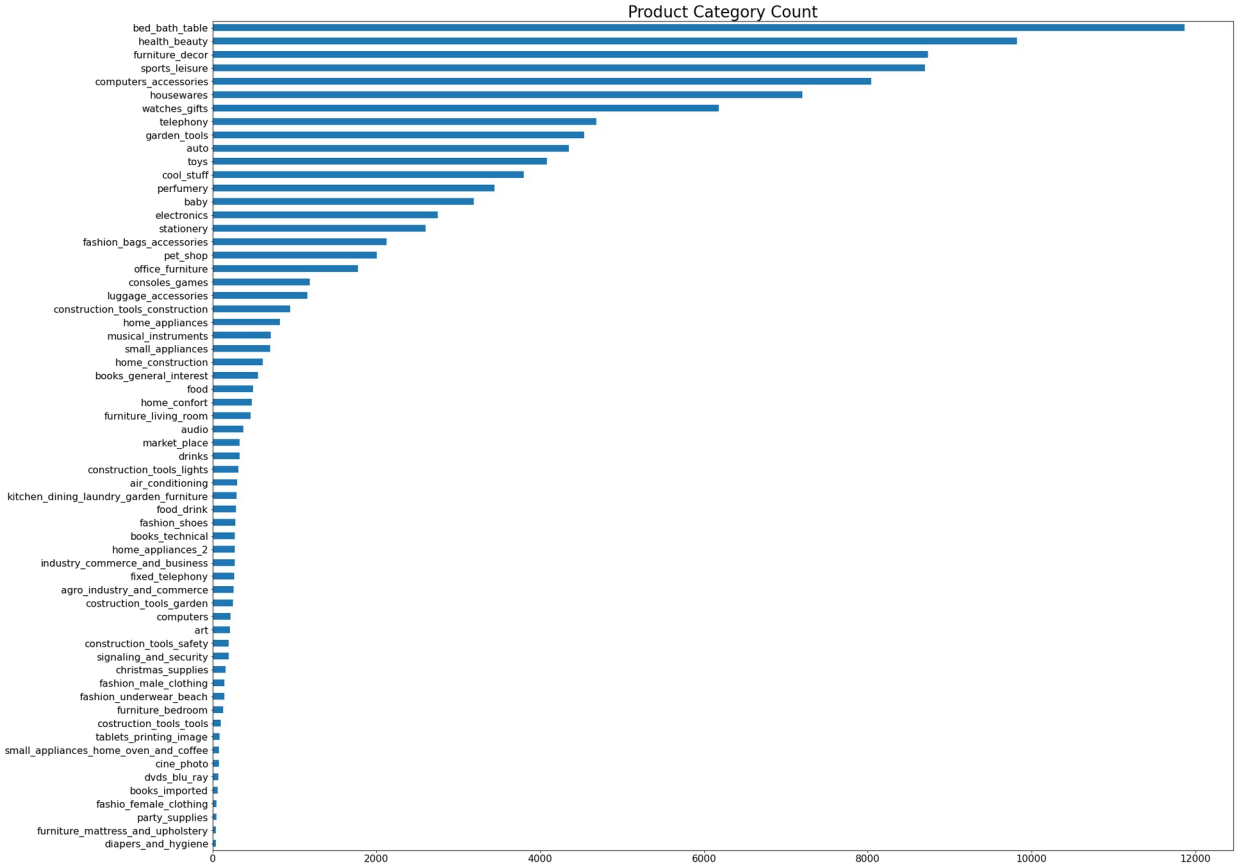


Présence de valeurs aberrantes ?

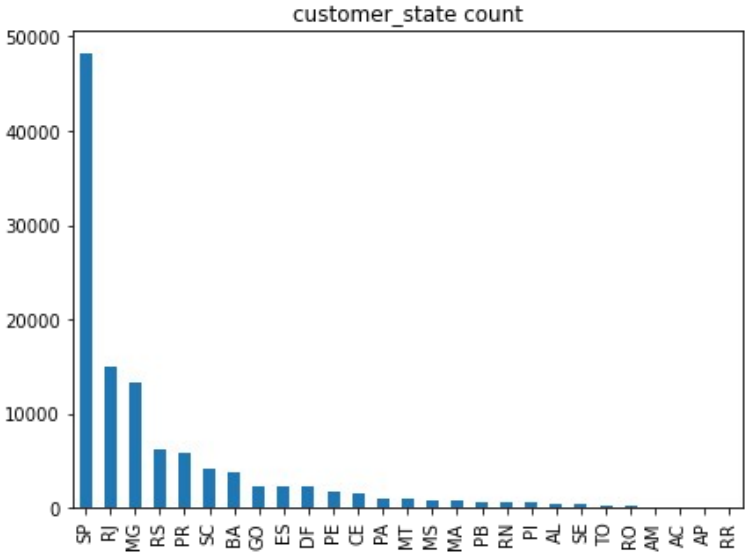
- Suppression des **doublons** sur table ‘geolocation’ sur les colonnes ["geolocation_city","geolocation_zip_code_prefix","geolocation_state"]
- Suppression des **doublons** sur les colonnes ["order_id","product_id", "customer_unique_id"]
- Présences d’**outliers** mais ce sont des **valeurs atypiques** donc elles sont conservées.

- Imputation des NaN avec KNN Imputer
- Normalisation des X
- Encoding des variables catégoriques

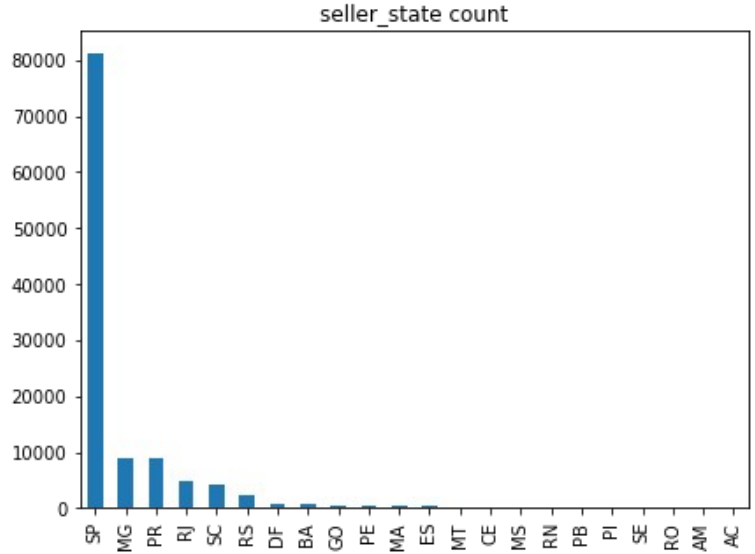
Exploration



Exploration

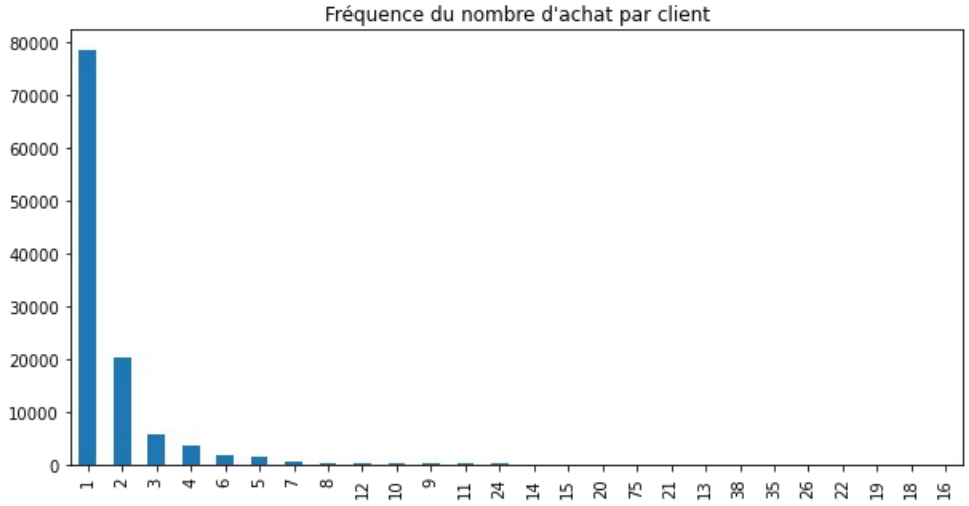


Les pays majoritaires sont :
- SP, Sao Paulo,
- MG, Minas Gerais
- RJ, Rio de Janeiro



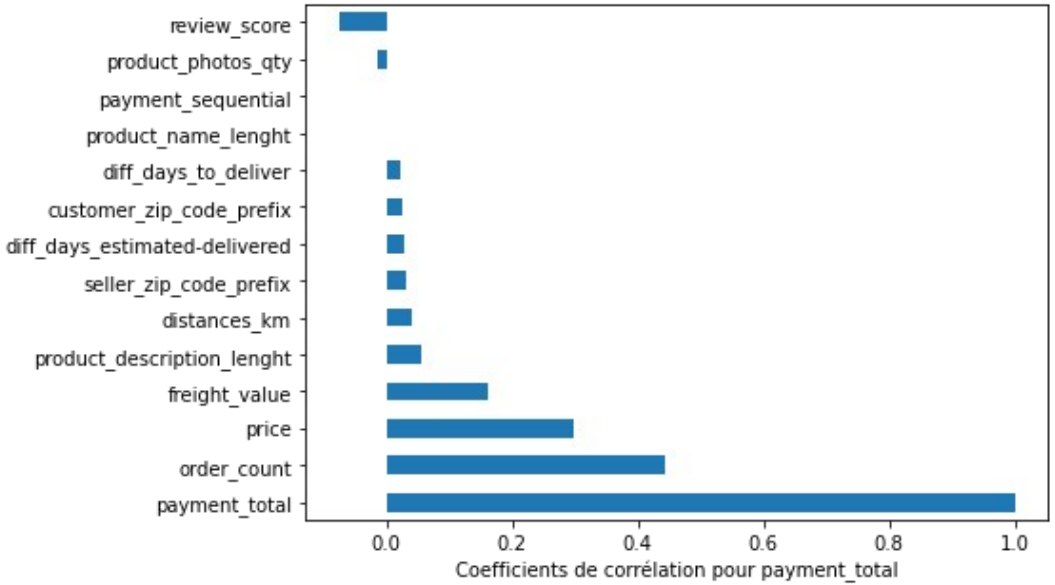
Les pays majoritaires sont :
- SP, Sao Paulo,
- MG, Minas Gerais
- PR, Paraná
- RJ, Rio de Janeiro

Exploration



Les nombres de commande supérieurs à 1 sont peut-être sous représentés. Il est alors possible d’enrichir la base de données.

Exploration

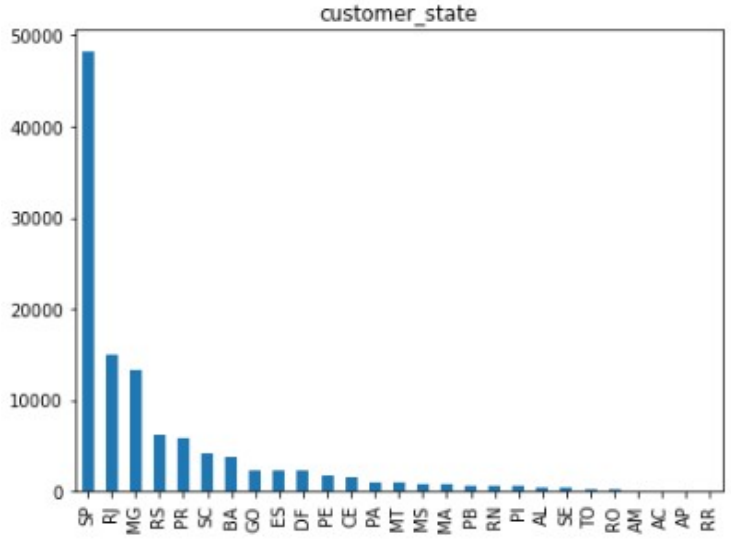


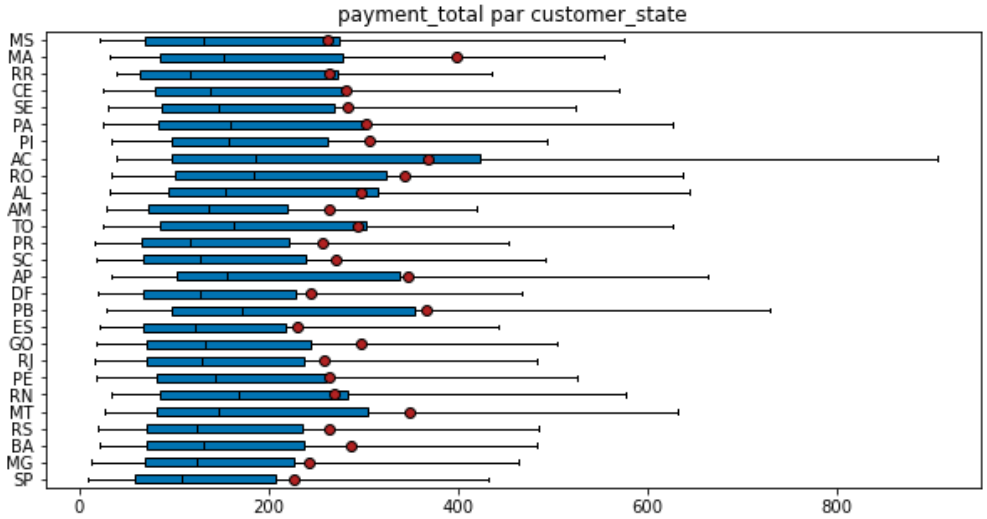
Les variables les plus corrélées à payment_total sont:

- order_count
- price

Ces variables sont susceptibles d'expliquer la formation des clusters.

Exploration



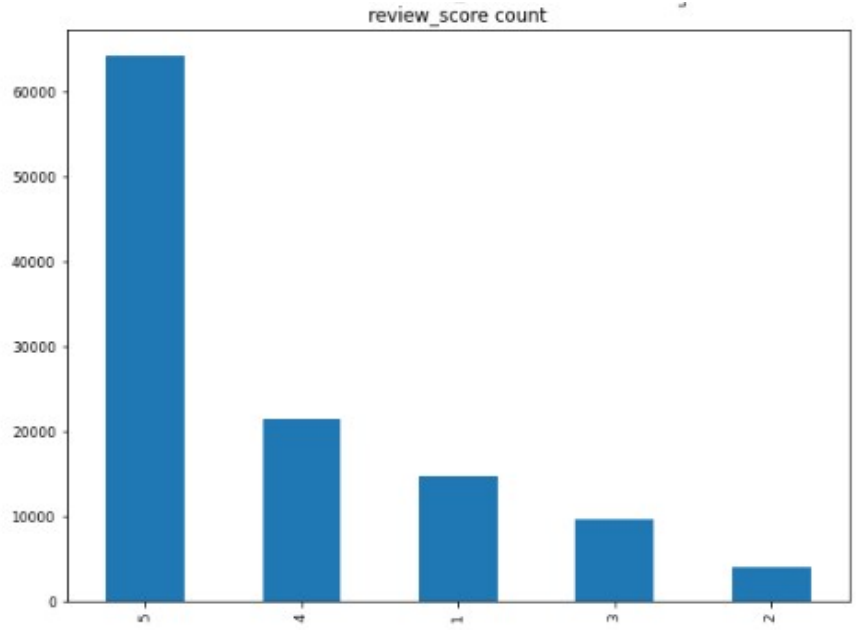


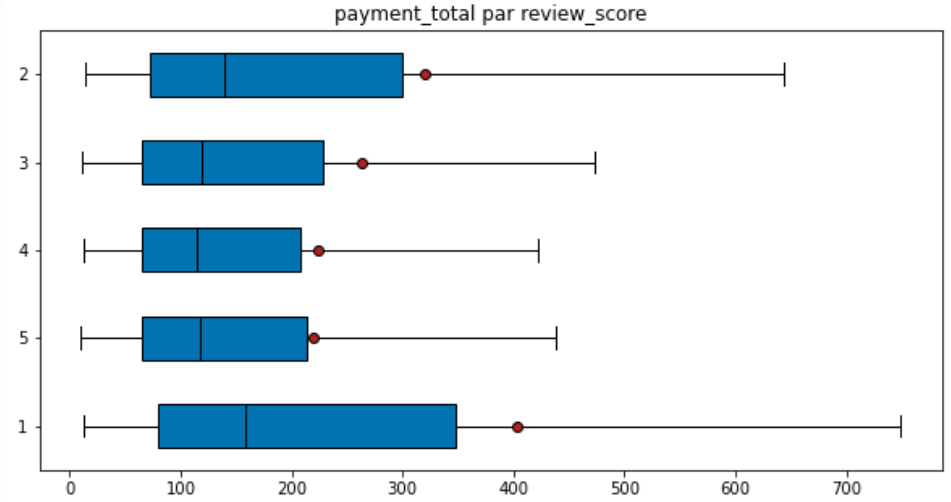
Les pays majoritaires sont :

- SP, Sao Paulo,
- MG, Minas Gerais
- RJ, Rio de Janeiro

Malgré tout les autres pays dépensent autant voir plus que SP, MG, RJ sur leurs payment_total

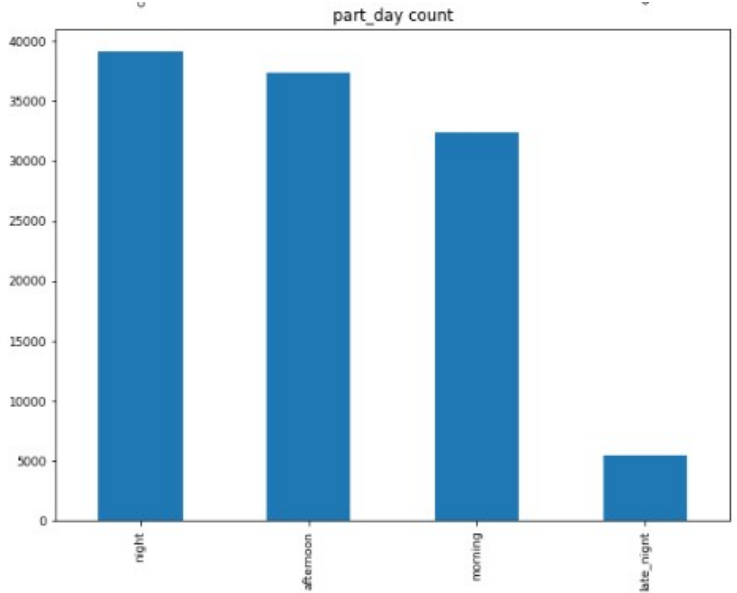
Exploration

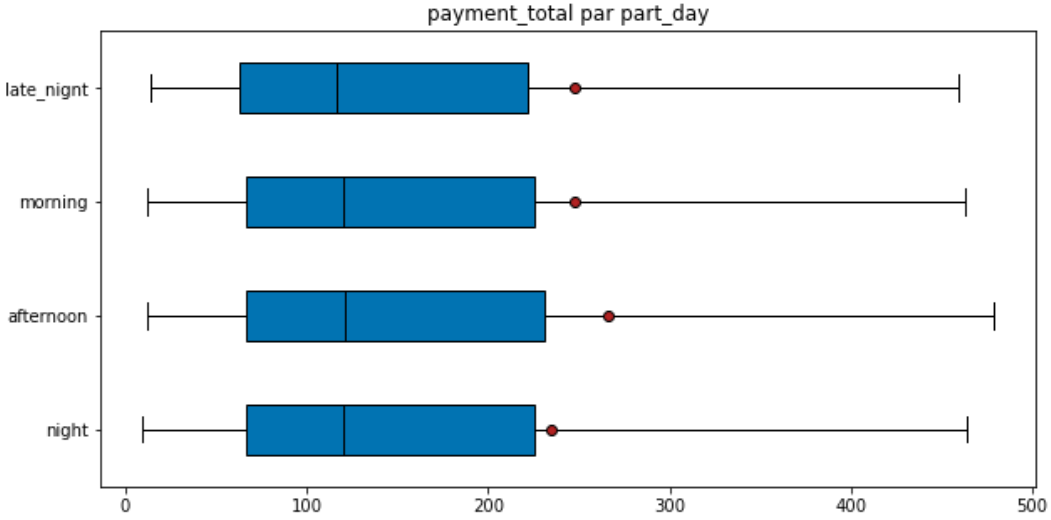




Plus la valeur du panier total est élevée plus la note est faible et vice versa.

Exploration

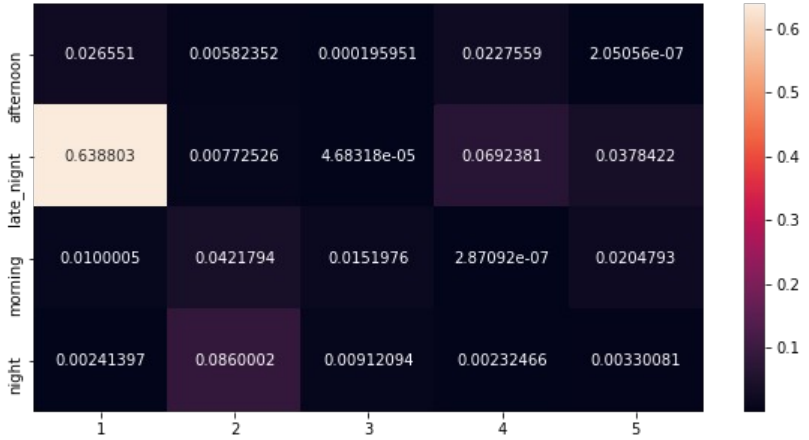




Peu de transactions se font tard la nuit, mais ils représentent une part équivalentes aux autres parties de la journée sur le payment_total des clients.

Exploration

part_day et customer_state



review_score et part_day

Les variables dépendantes sont :
review_score et part_day
review_score et customer_state
part_day et customer_state

Ces variables sont susceptibles d'expliquer la formation des clusters.

Classification non supervisée sur les données à grande dimension

- **DBSCAN pour supprimer les outliers**
- **K-means pour la baseline**
  - n_clusters avec la méthode du coude
  - apprentissage et prédiction
  - plot avec PCA en 3 composantes principales
  - vérification du nombres d'individus par cluster
  - évaluation avec avec Silhouette Score, Calinski Harabasz Score et Davies Bouldin Score
- **Phase 1 : K-means**
  - apprentissage et prédiction
  - récupération des centroïdes
  - vérification du nombres d'individus par cluster
- **Phase 2 : AggolmerativeClustering**
  - n_clusters avec la méthode du dendogramme et du coude
  - apprentissage et prédiction
  - plot avec PCA en 3 composantes principales
  - vérification du nombres d'individus par cluster
  - évaluation avec Silhouette Score, Calinski Harabasz Score et Davies Bouldin Score
- **Régression logistique régularisée (supervisé) en l1 pour l'importance des variables**

Améliorations & modèle final

Les étapes d'amélioration :

- Méthodologie en 2 phases avec **K-means** et **AgglomerativeClustering** pour les données à grande dimension.

- **DBSCAN** avec eps = 1.5 pour supprimer les outliers et rendre les clusters plus homogènes.

- Suppression des variables corrélées restantes telles que payment_mean.

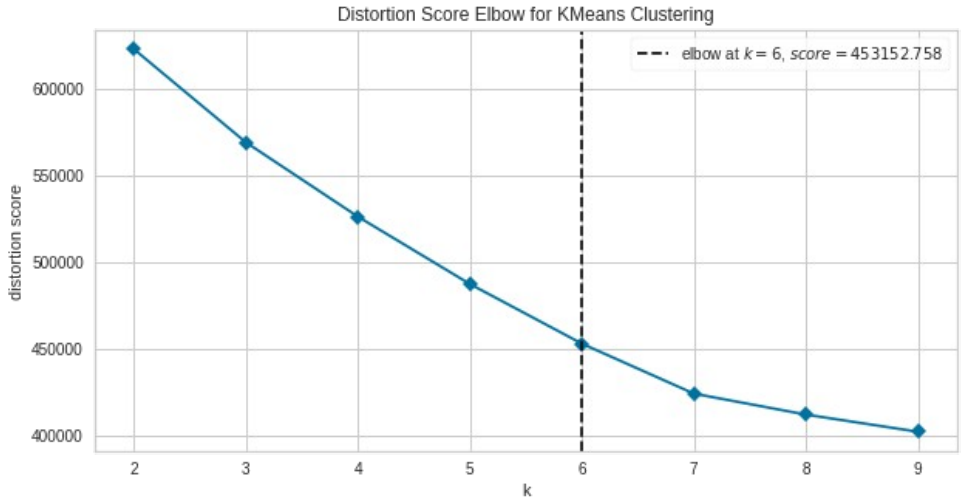
- Suppression des variables non utiles telles que product_weight_g, product_length_cm, (...), order_status_approved ect...

- **n_clusters** de 200 à 1000 pour le k-means.

- **n_clusters** de 12 à 6 pour l'AgglomerativeClustering.

Modélisation

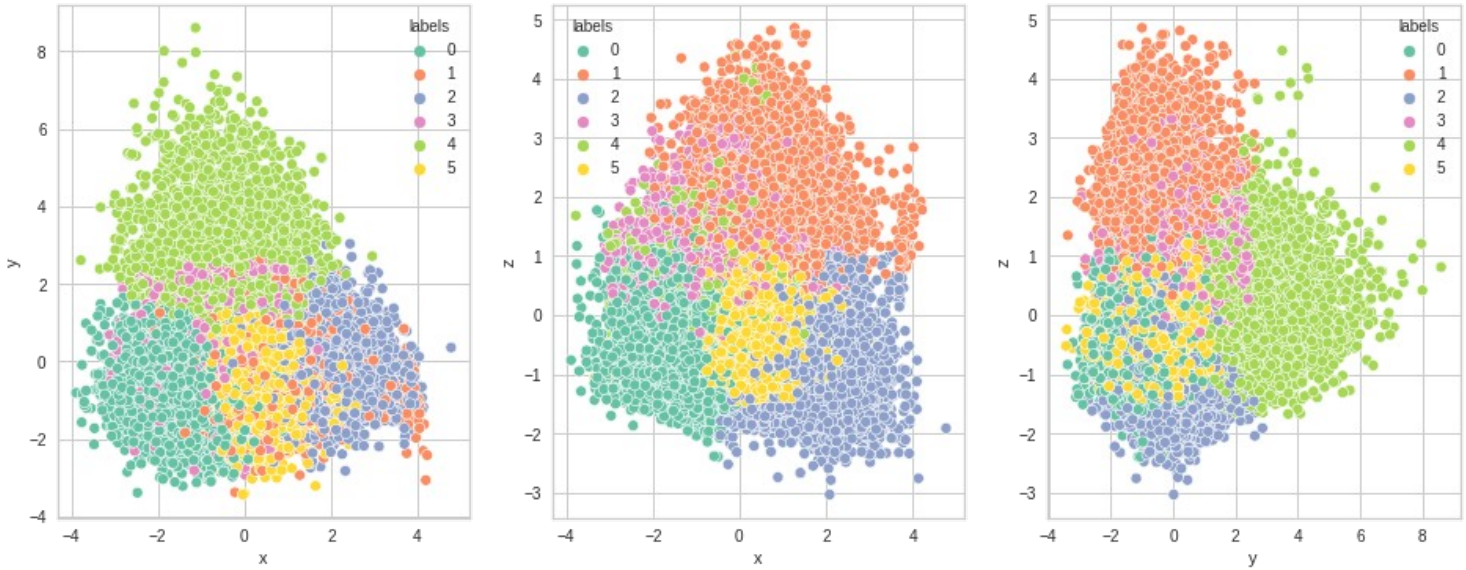
Baseline : K-means



n_clusters = 6

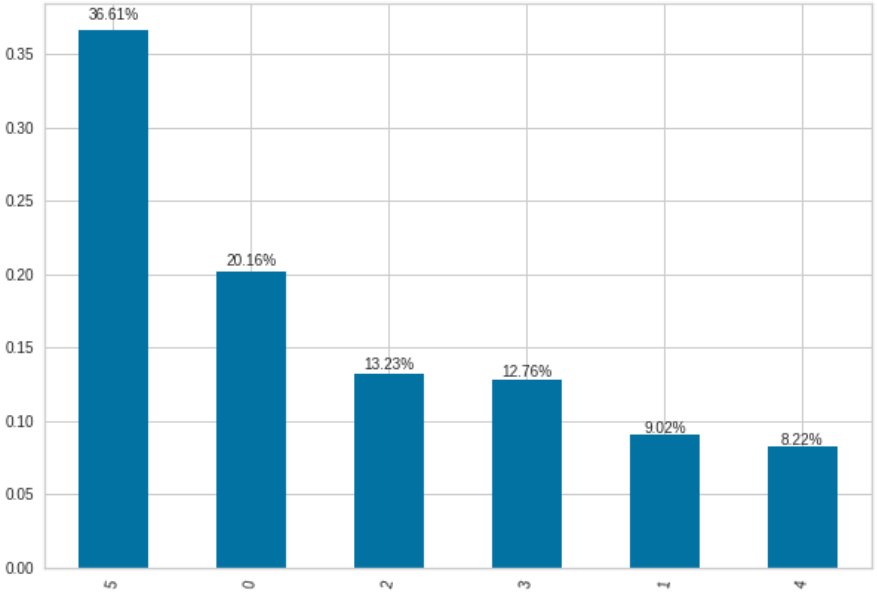
Modélisation

Baseline: K-means



Modélisation

Baseline: K-means



silhouette_score = 0.136
Score en progression de 0,06 à 0.136

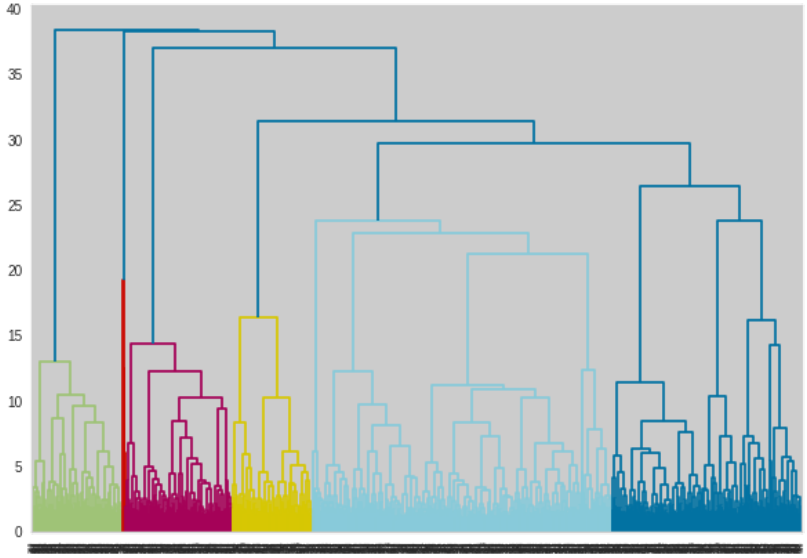
Bon signe pour l'obtention de clusters plus homogènes par la suite.

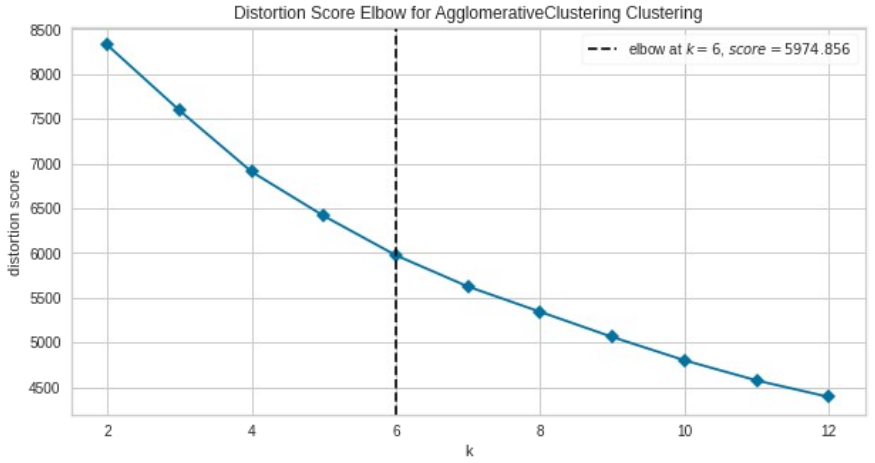
Scores pour l'itération 1:

Silhouette Score: 0.136
Calinski Harabasz Score: 8884.851
Davies Bouldin Score: 1.883

Modélisation

K-means & AgglomerativeClustering

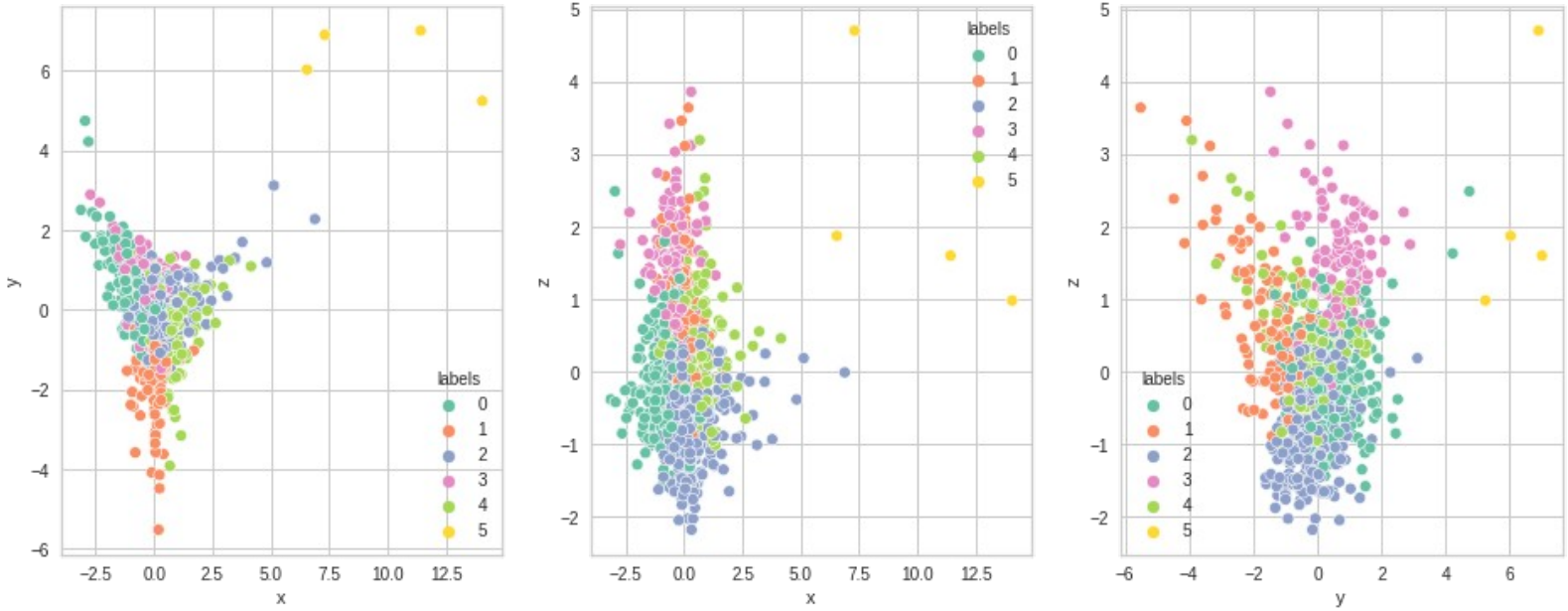




n_clusters = 6

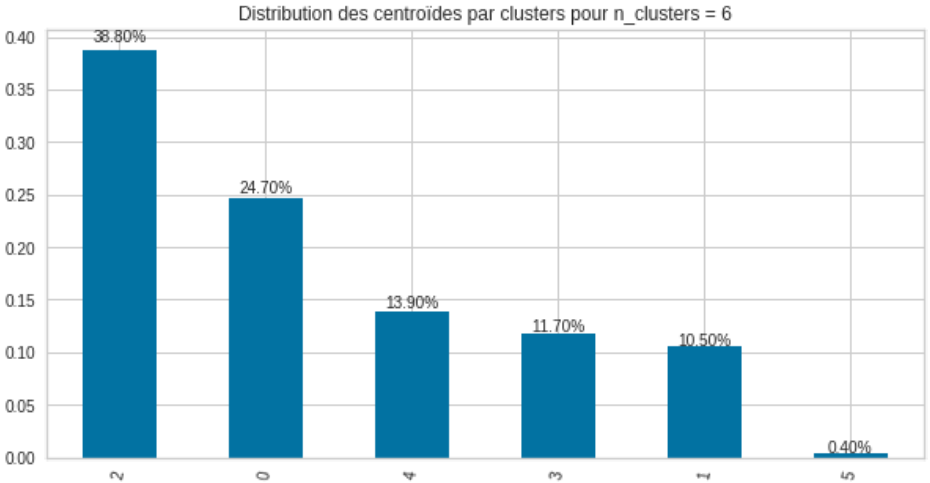
Modélisation

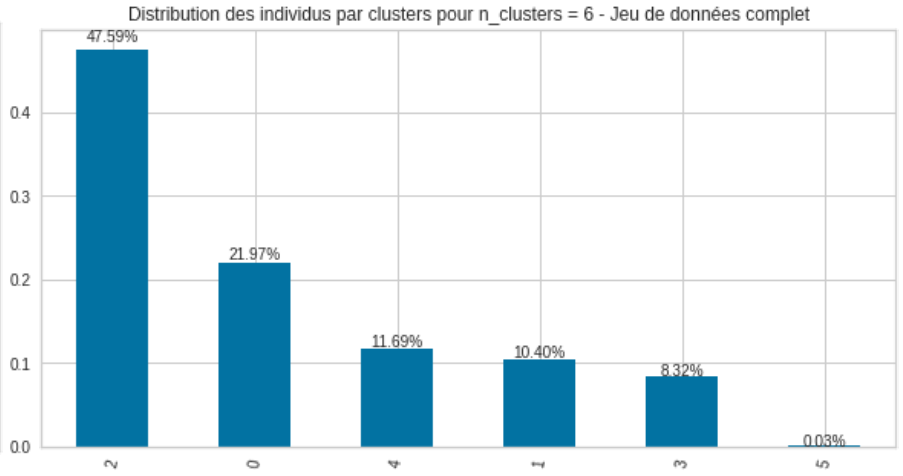
K-means & AgglomerativeClustering



Modélisation

K-means & AgglomerativeClustering





n_clusters = 6 >> n_clusters = 5

Scores pour l'itération 1:

Silhouette Score: 0.127
Calinski Harabasz Score: 102.709
Davies Bouldin Score: 1.917

Comparaison des Résultats

K-means & AgglomerativeClustering

K-means

Scores pour l'itération 1:

Silhouette Score: 0.136
Calinski Harabasz Score: 8884.851
Davies Bouldin Score: 1.883

K-means & AgglomerativeClustering

Scores pour l'itération 1:

Silhouette Score: 0.127
Calinski Harabasz Score: 102.709
Davies Bouldin Score: 1.917

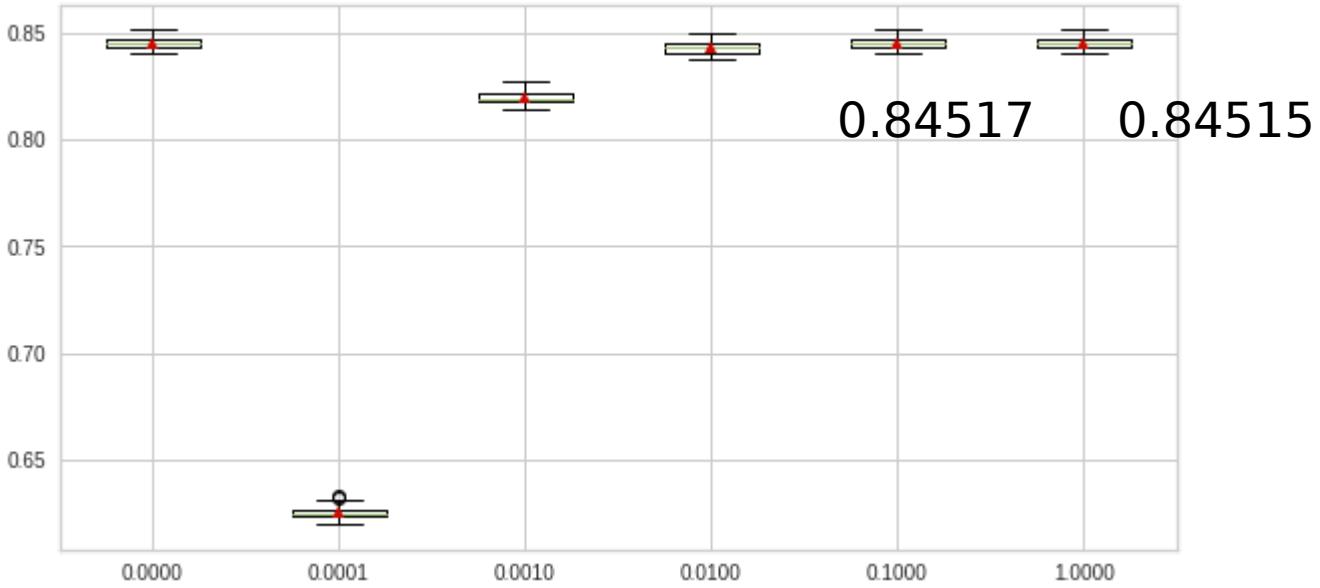
2 résultats proches versus grande amélioration du Calinski Harabasz Score.
Donc choix du modèle 2.

Le **coefficient silhouette** est donc compris entre -1 et 1, plus proche le coefficient est proche de 1, plus l'assignation de x à son cluster est satisfaisante

Plus les ratios **C.Harabasz** et **D.Bouldin** sont petits, plus les groupes sont séparés et regroupés.

Interprétation

Régression logistique régularisée par l1



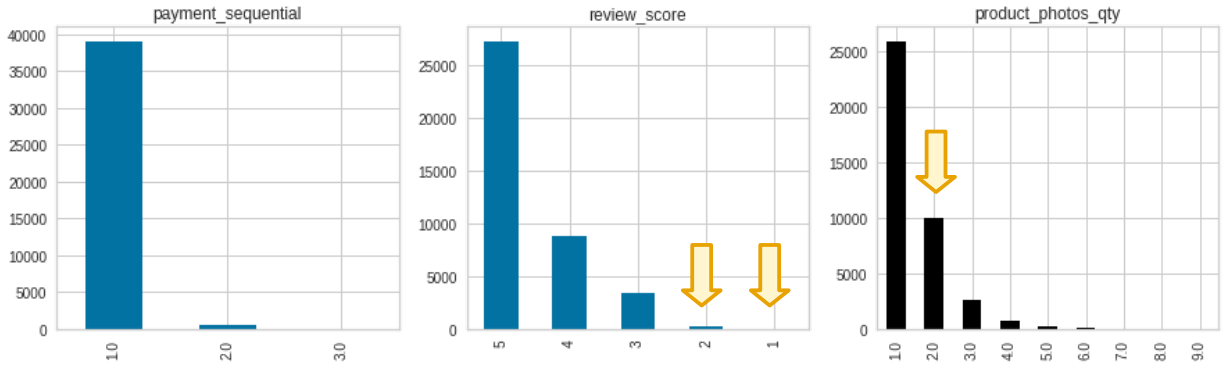
LogisticRegression(C=0.1, multi_class='multinomial', penalty='l1', solver='saga')

model.score = 0.85 (Précision ou mean accuracy)

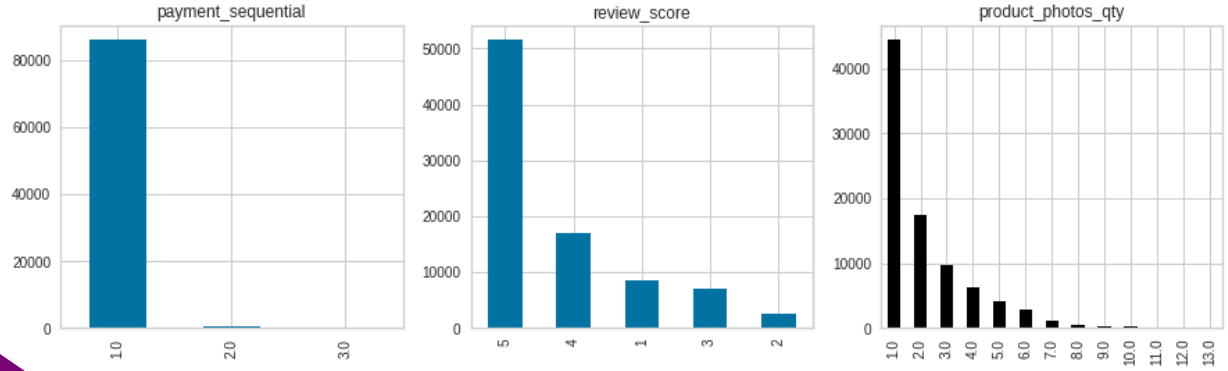
Interprétation: Cluster 2

Régression logistique régularisée par l1

Cluster 2



Échantillon N



review_score : 1.27
payment_sequential : 1.61
product_photos_qty : -1.38

Clients qui notent plutôt bien.

Leurs produits sont présentés avec 1 ou 2 photos.

Interprétation Cluster 0

Régression logistique régularisée par l1

Cluster 2

Échantillon N

most_payment_type_credit_card : 1.06
product_photos_qty : 1.21
review_score : 1.36

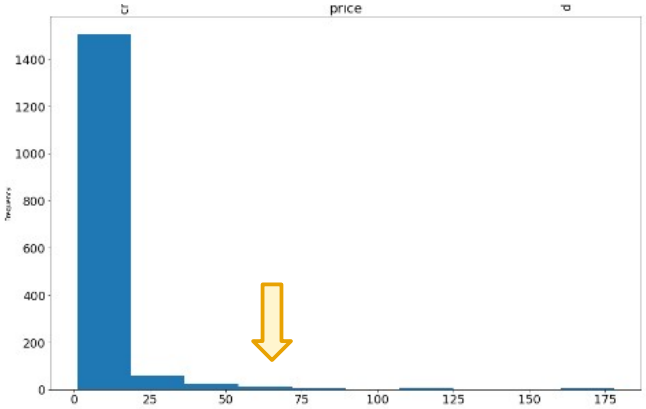
Ces clients notent plutôt bien.

Ils aiment acheter des produits avec plusieurs photos

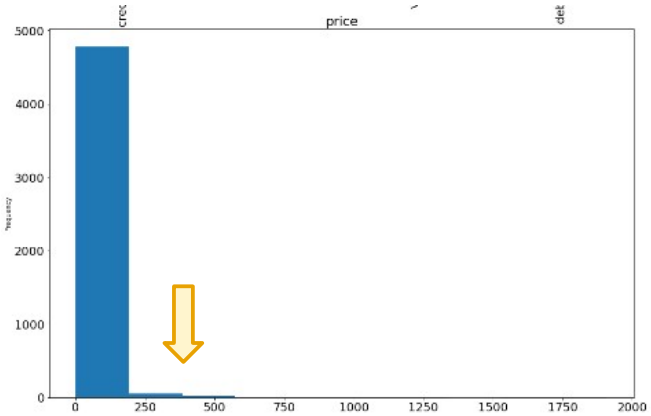
Interprétation Cluster 4

Régression logistique régularisée par l1

Cluster 4



Échantillon N

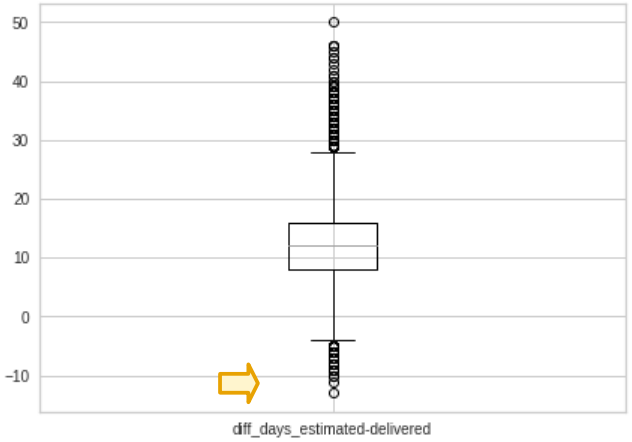


price : 0.23
part_day_afternoon : 0.36
diff_days_estimated-delivered : 0.36
most_payment_type_boleto : 0.40
order_count : 0.81
review_score : -3.14

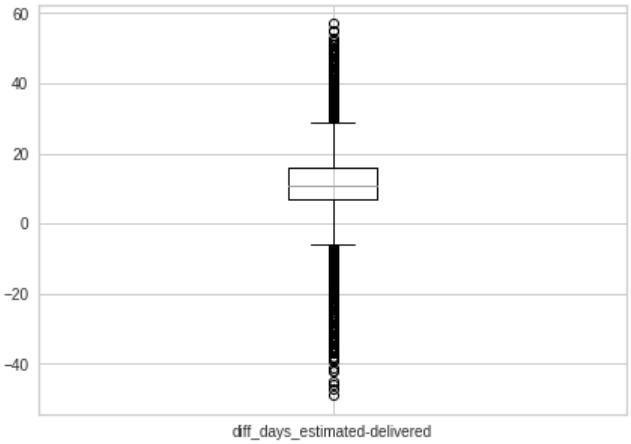
Ces clients achètent des produits inférieurs à 100 réal environ. Et en espèce de préférence.

Interprétation Cluster 4

Régression logistique régularisée par l1



Cluster 4



Échantillon N

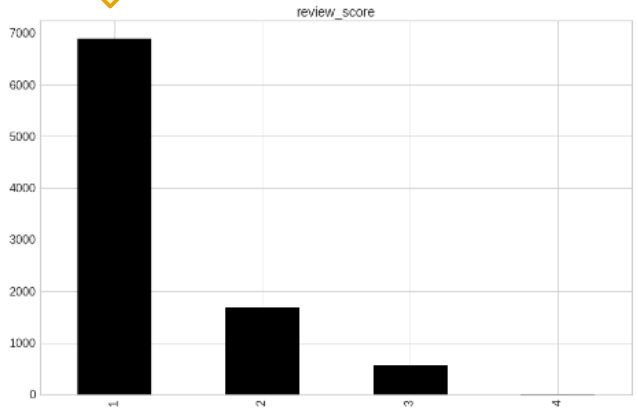
price :	0.23
part_day_afternoon :	0.36
diff_days_estimated-delivered :	0.36
most_payment_type_boleto :	0.40
order_count :	0.81
review_score :	-3.14

Ces clients ne connaissent pas de retard de livraison de plus de 10 jours environ sur la date de livraison estimée.

Interprétation Cluster 4

Régression logistique régularisée par l1

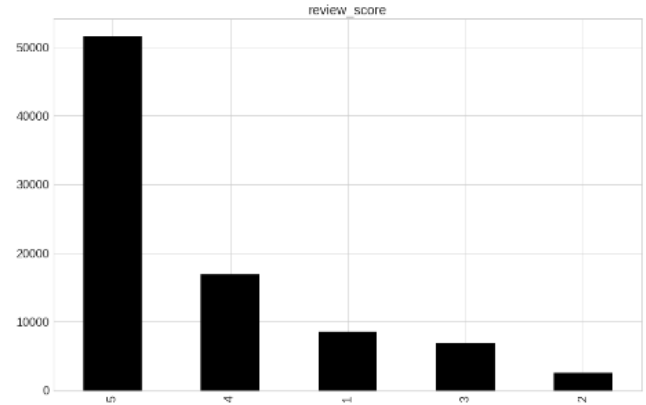
Cluster 4



price :	0.23
part_day_afternoon :	0.36
diff_days_estimated-delivered :	0.36
most_payment_type_boleto :	0.40
order_count :	0.81
review_score :	-3.14

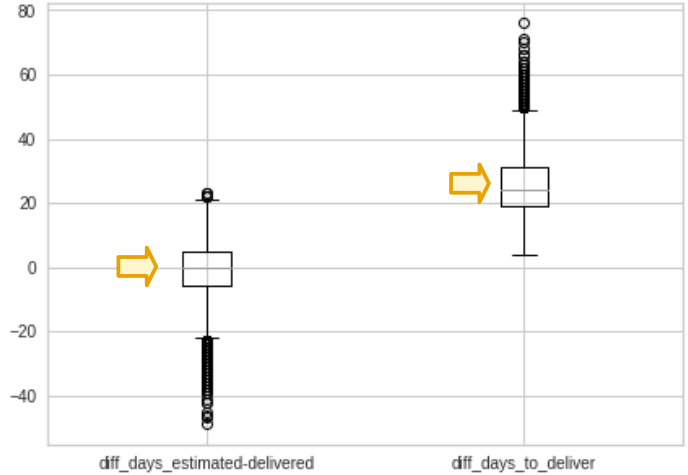
Ces clients notent sévèrement.

Échantillon N

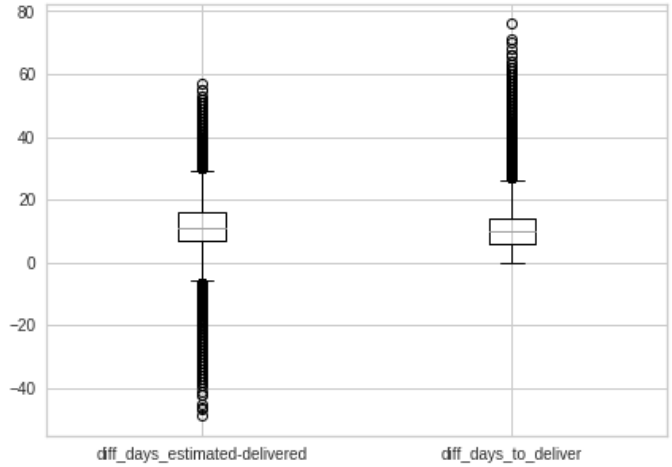


Interprétation Cluster 3

Régression logistique régularisée par l1



Cluster 3



Échantillon N

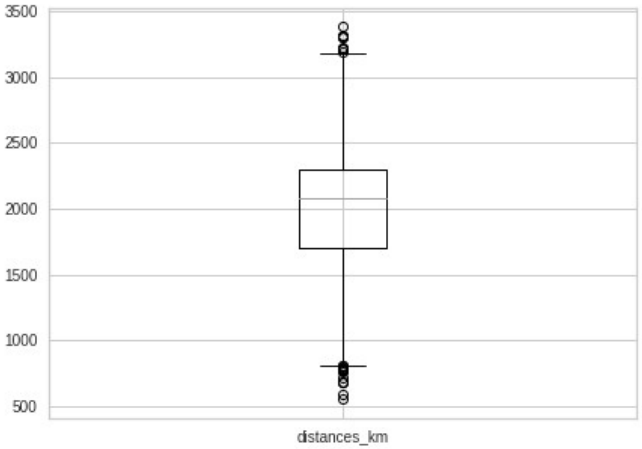
diff_days_to_deliver : 2.73
diff_days_estimated-delivered : -1.64
product_photos_qty : -1.64

*10 jours est la valeur médiane du nombre de jours en avance sur la date estimée de livraison pour la population N, alors qu'elle est de **0 jours** pour le cluster 3.*

*10 jours est le temps d'attente en médiane sur l'échantillon N alors que le cluster 3 attend en médiane **25 jours** pour recevoir la commande.*

Interprétation Cluster 5

Régression logistique régularisée par l1



Cluster 5

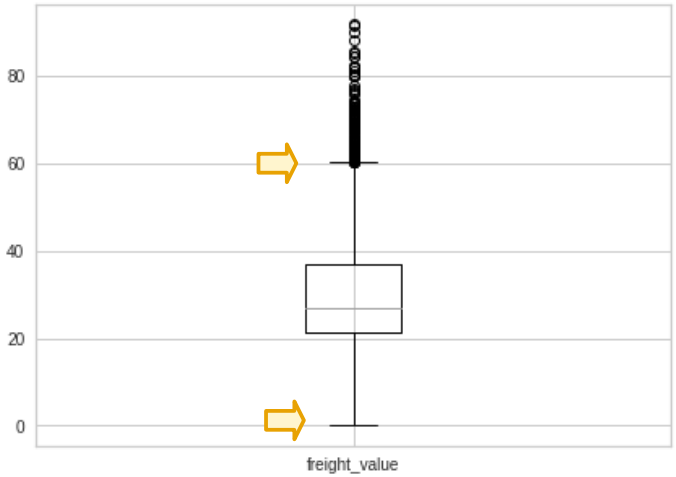
Échantillon N

freight_value : 0.85
part_day_morning : 0.87
distances_km : 4.59

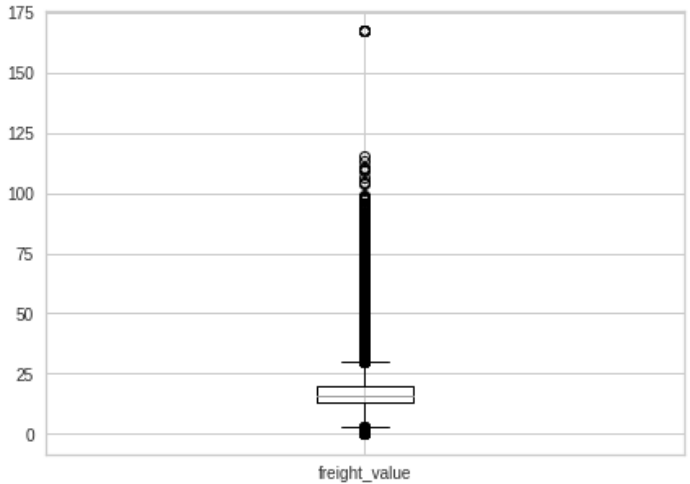
Sur la population N la distance en km entre le customer et le seller est au alentour de 400 km. Alors que pour le cluster 5 cette distance est au alentour de 2000 km. Donc ce sont des clients transfrontaliers.

Interprétation Cluster 5

Régression logistique régularisée par l1



Cluster 5



Échantillon N

freight_value : 0.85
part_day_morning : 0.87
distances_km : 4.59

L'écart min-max des frais de transport pour le cluster 5 est plus important que pour la population N.

Contrat de maintenance : pré-étude

```
youngest = data.index.max()
youngest
```

```
Timestamp('2018-08-29 15:00:37')
```

```
oldest = data.index.min()
oldest
```

```
Timestamp('2016-09-05 00:15:34')
```

	order_count_day	variation_day
2016-09-05	1	NaN
2016-10-02	1	0.00
2016-10-03	6	5.00
2016-10-04	53	7.83
2016-10-05	28	-0.47

Nombre d'achat par jour et sa variation journalière

```
In [396]: date_to_review = df_day[(df_day["variation_day"] > 10) | (df_day["variation_da
          date_to_review
```

Out[396]:

	order_count_day	variation_day
2017-01-05	29	13.5

Lorsque la variation dépasse 10%, on relance le clustering.

Donc le 2017-01-05 le clustering peut être relancé.

En deux années, le dépassement s'est produit une seule fois.

Contrat de maintenance

Création des jeux de données :
- Table avec « order_purchase_timestamp » en index

- Découpage de la table initiale en partant de la date la plus récente **moins *x* mois** avec *x* = 1, 2, 3, 4, 6, et 12

Exemple avec date_fin – 1 mois :

```
#t= 2018-08-29
#dataset  t-1
data_set1 = data.loc['2016-09-05':'2018-07-29']
print(data_set1.shape)
data_set1.head()
```

```
(80351, 22)
```

order_purchase_timestamp	price	freight_value	payment_sequential	review_score	product_name_lenght
2018-03-19 18:40:33	0.11	-5.33e-01	-0.09	-2.24	1.01
2017-12-03 17:28:57	-0.06	9.42e-03	-0.09	0.69	-1.09
2018-03-12 11:56:58	0.02	-6.84e-01	-0.09	0.69	-0.09
2017-04-13 02:11:17	-0.46	1.08e-01	-0.09	-0.05	-0.29
2017-03-05 19:25:45	1.14	-4.49e-01	-0.09	-0.05	0.31

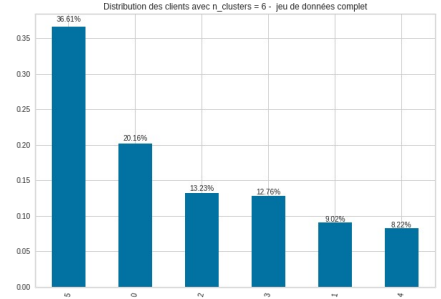
```
youngest = data.index.max()
youngest
```

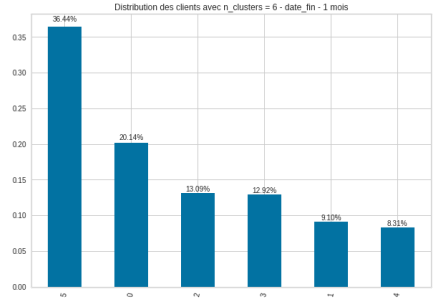
```
Timestamp('2018-08-29 15:00:37')
```

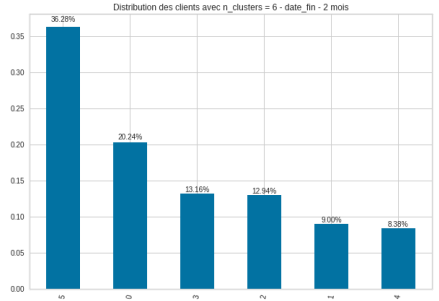
```
oldest = data.index.min()
oldest
```

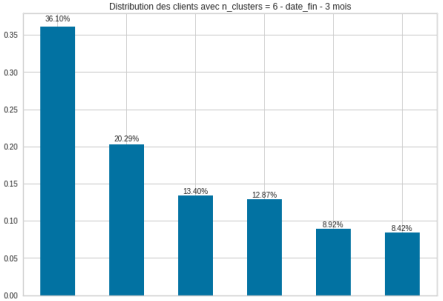
```
Timestamp('2016-09-05 00:15:34')
```

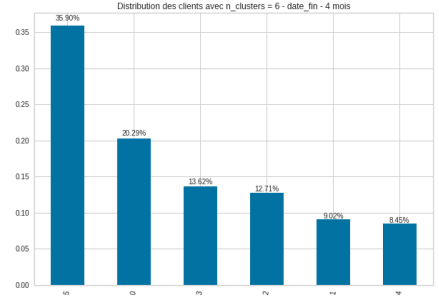
Contrat de maintenance

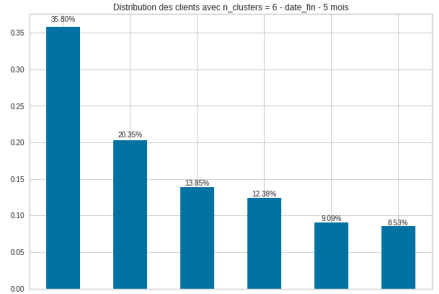


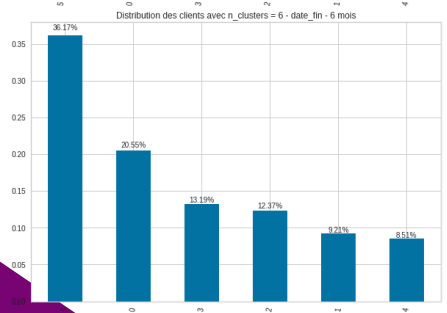


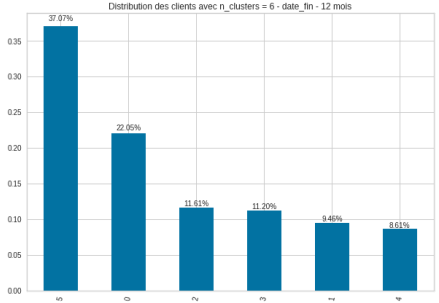






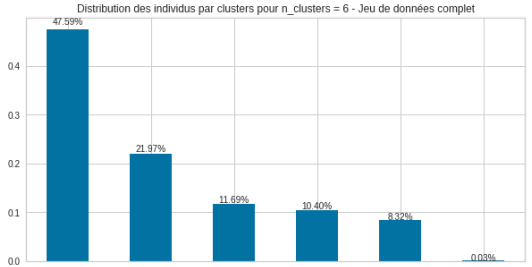


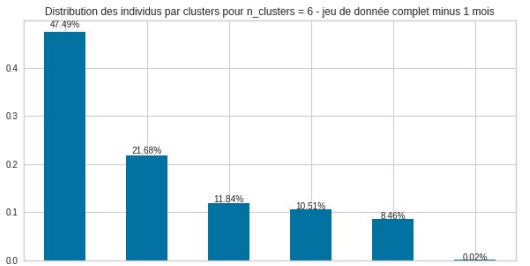


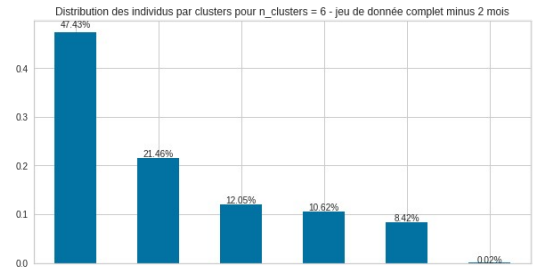


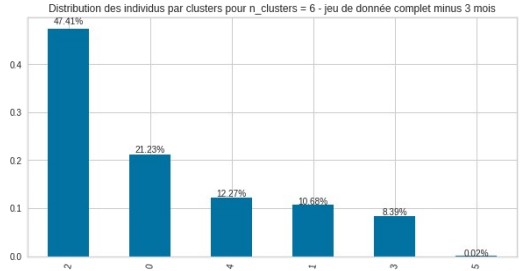
Baseline K-means : les changements sont perceptibles au bout de 12 mois pour les 2 premiers clusters.

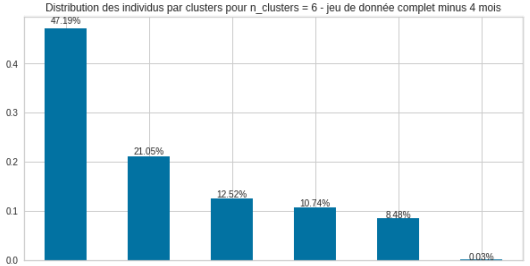
Contrat de maintenance

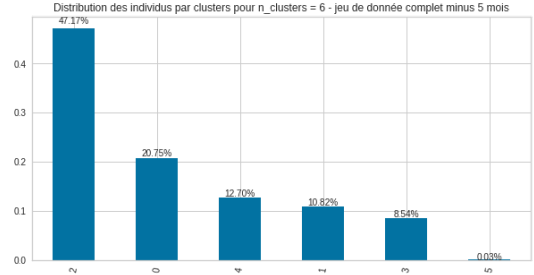


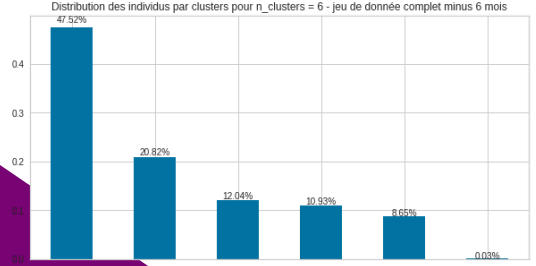


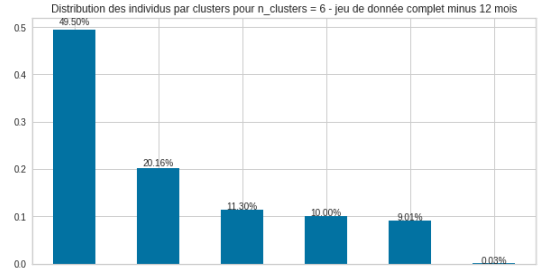












K-means et AgglomerativeClustering : les changements sont perceptibles au bout de 12 mois pour les 2 premiers cluster.

Contrat de maintenance : conclusion

Le contrat de maintenance se repose sur trois études :

1- La pré-étude basée sur le nombre d’achats par jour avec un déclenchement au-delà d’un seuil de 10 %.

2- Distribution des clients suite au K-means clustering

3- Distribution des clients suite suite au K-means et à l’AgglomerativeClustering

De ces 3 études, il est préconisé de faire un renouvellement de l’étude un fois par an.

PEP8

```python
def fancy_dendrogram(*args, **kwargs):

    '''
    *arg: varying number of positional arguments
    **kwargs: named arguments
    '''

    max_d = kwargs.pop('max_d', None)
    if max_d and 'color_threshold' not in kwargs:
        kwargs['color_threshold'] = max_d
    annotate_above = kwargs.pop('annotate_above', 0)

    ddata = dendrogram(*args, **kwargs)

    if not kwargs.get('no_plot', False):
        plt.title('Hierarchical Clustering Dendrogram (truncated)')
        plt.xlabel('sample index or (cluster size)')
        plt.ylabel('distance')
        for i, d, c in zip(ddata['icoord'], ddata['dcoord'], ddata['color_list']):
            x = 0.5 * sum(i[1:3])
            y = d[1]
            if y > annotate_above:
                plt.plot(x, y, 'o', c=c)
                plt.annotate("%.3g" % y, (x, y), xytext=(0, -5),
                             textcoords='offset points',
                             va='top', ha='center')
        if max_d:
            plt.axhline(y=max_d, c='k')
    return ddata
```

```python
from scipy.cluster.hierarchy import dendrogram, linkage

# set cut-off to 50
max_d = 50  # max_d as in max_distance
Z = linkage(centers, 'ward')

fancy_dendrogram(
    Z,
    truncate_mode='lastp',
    p=12,
    leaf_rotation=90.,
    leaf_font_size=12.,
    show_contracted=True,
    annotate_above=10,
    max_d=max_d,  # plot a horizontal cut-off line
)
plt.show()
```

Quelques exemples :

- Convention de nommage des fonctions, des variables
- Mise en page (indentation, sauts de lignes, doctring)
- Opérateurs =,+, ect..
- Commentaires en anglais