



Olist

Projet 5 - Raphaël Koudache--Louvet



# Contexte

## BtoBtoC :

- Marché brésilien fragmenté
- Intégration
- Olist n'est pas Amazon

## Proposition de valeur:

- Image Boutique Officielle
- Logistique optimisée (économies d'échelle, délais)
- Campagnes Marketing Exclusives
- Outils de compétitivité et de pricing
- Annonces Optimisées (référencement, présentation)

The screenshot displays the Olist website interface. At the top, there's a yellow header with the 'mercado livre' logo and a search bar. Below this, the Olist logo is prominently featured. The main content area shows a product listing for 'Kit Rack Suspensão Tv Com 4 Nichos Premium Organizador Livros'. The product image shows a modern white TV stand with shelves. To the right of the image, the product details are listed: 'Novo | 2323 vendidos', '4.5 estrelas', '43 opiniões', and 'MAIS VENDIDO'. The price is 'R\$ 68<sup>97</sup>' with a financing option 'em 12x de R\$ 16<sup>98</sup> pagando com Mercado Crédito'. Below the product image, there's a section for 'Anúncios de Olist' showing three similar products with their prices and ratings. At the bottom, there's a 'Características principais' section with a table listing 'Marca' and 'Vendas Pp' as 'Vendas Pp' and 'Cor' as 'Branco'.

Marca	Vendas Pp
Vendas Pp	

Cor	Branco
Branco	



# Mission

1 - Segmentation Clients (Apprentissage Machine non supervisé)

2 – Maintenance (Simulation)

N.B : PEP8



# Plan

I – Analyse Exploratoire des données et manipulations des données

II – Segmentation des clients (Clustering)

III – Contrat de Maintenance et PEP8



# I – Analyse Exploratoire des données et manipulations des données

**LEV E**  
LAVARELERO

TABELA DE PREÇOS

DE 0 À 1 HORA  
**R\$ 6,00**

POR HORA ADICIONAL DO PRAZO + R\$1,00

TEMOS VAGAS  
PARA MENSALISTAS

DE SEGUNDA A SEXTA DAS 07H AS 19H  
SÁBADO, DOMINGO E FÉRIADO FECHADO

www.lev.com.br



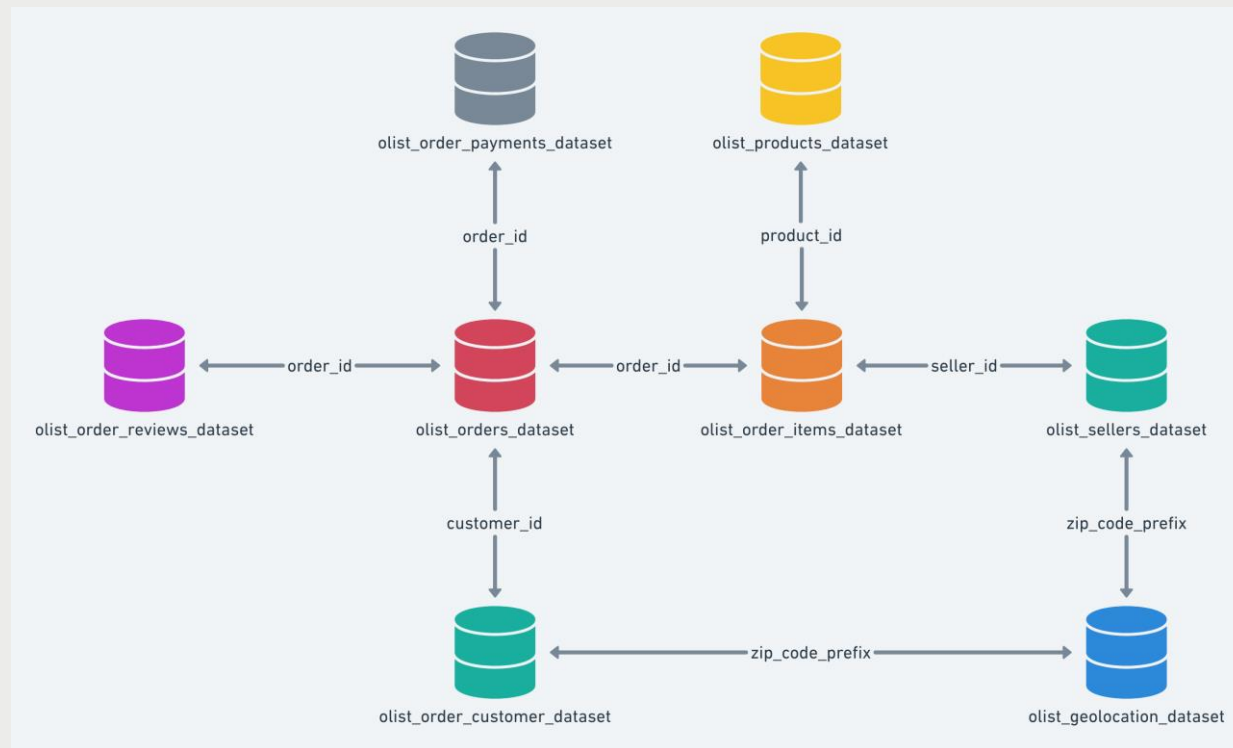
# I - Exploration

8 Tables

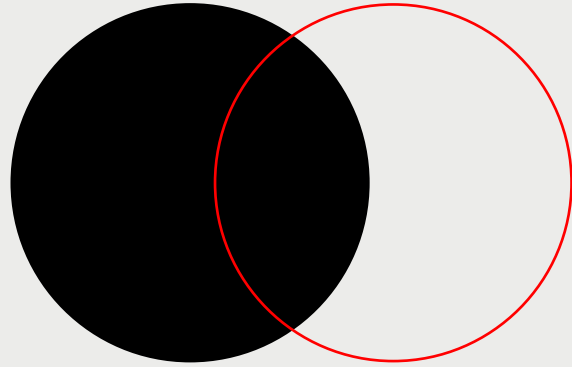
Brésil

2016-2018

Etonnement :  
97 % = 1 seule  
commande

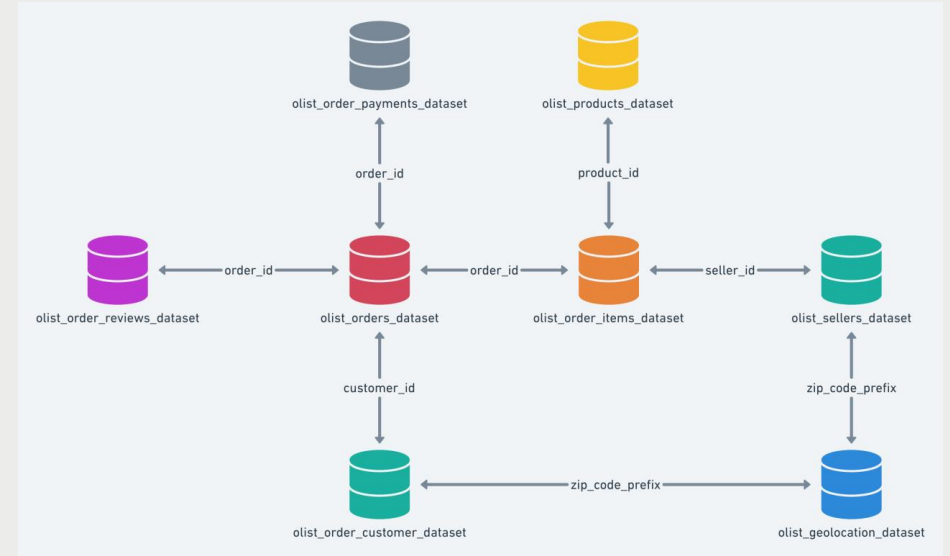


# I - Jointure



Left Joins sur les commandes

Attention : agrégations pour éviter doublons



Clefs

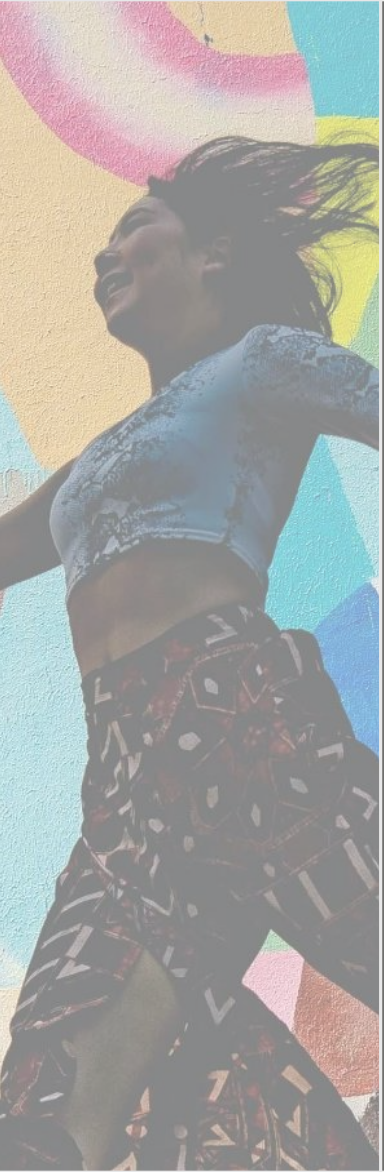
# I - Sélection des features

~~Variables qualitatives :~~

OneHotEncode VS Distances  
Catégorie de produits VS 1 seule commande

RFM + Proposition de valeur d'Olist :

Sélection par itération  
Segmentation comportementale





# I - Sélection des features

Progressivement RFM va devenir RMS TBE:

**R**écense

**M**ontant

**S**atisfaction

**T**iming (avance ou retard)

**B**asket (Panier)

**E**vénement



# I - Nettoyage

Si manques dans features  
sélectionnées drops

Commandes non délivrées ne  
sont pas prises en compte

Missing Values	
order_approved_at	0.160899
order_delivered_carrier_date	1.793023
order_delivered_customer_date	2.981668
payment_value_sum_per_order	0.001006
payment_type_mode_per_order	0.001006
payment_installments_max_per_order	0.001006
payment_sequential_max_per_order	0.001006
last_review_answer_timestamp	0.772317
review_score_mean_per_order	0.772317
review_id	0.772317
review_score	0.772317
review_comment_title	88.385073
review_comment_message	58.895786
review_creation_date	0.772317
review_answer_timestamp	0.772317
order_item_id_max	0.779357
product_id	0.779357
seller_id	0.779357
shipping_limit_date	0.779357
price	0.779357
freight_value	0.779357
seller_zip_code_prefix	0.779357
seller_city	0.779357
seller_state	0.779357
product_category_name	2.205328
product_name_lenght	2.205328
product_description_lenght	2.205328
product_photos_qty	2.205328
product_weight_g	0.795447
product_length_cm	0.795447
product_height_cm	0.795447
product_width_cm	0.795447
geolocation_zip_code_prefix_x	0.279563
customer_geolocation_lat	0.279563
customer_geolocation_lng	0.279563
geolocation_city_x	0.279563
geolocation_state_x	0.279563
geolocation_zip_code_prefix_y	0.998582
seller_geolocation_lat	0.998582
seller_geolocation_lng	0.998582
geolocation_city_y	0.998582
geolocation_state_y	0.998582

Missing Values	
review_comment_title	88.220586
review_comment_message	59.420552



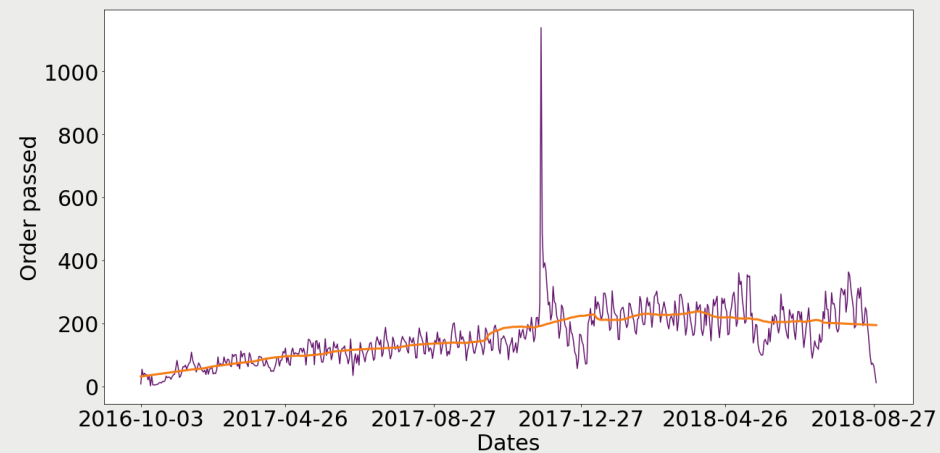
# I - Feature Engineerings

Récence

Avance ou retard  
sur la livraison

Evènements

Number of sells per day against moving average 90 days



# I - Agrégation

```
.groupby() ; .agg({})
```

Customer\_Unique\_Id

1 sample = 1 individu

90 000 samples

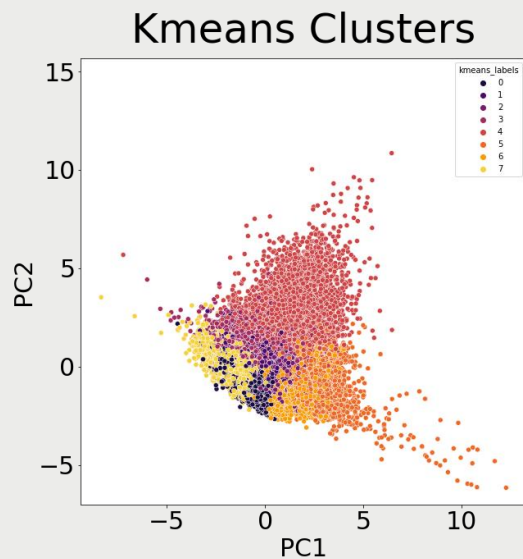




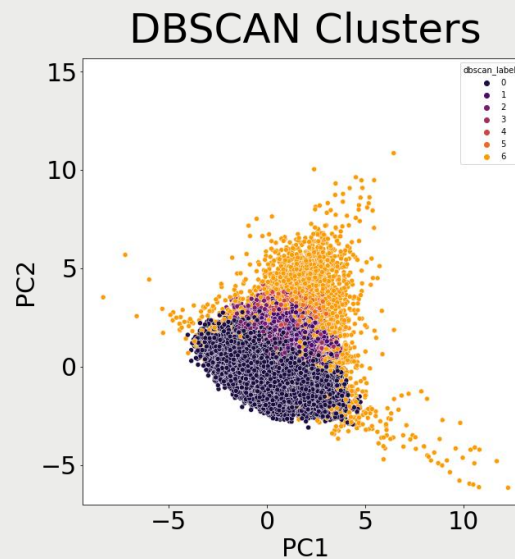
## II – Segmentation des clients (Clustering)



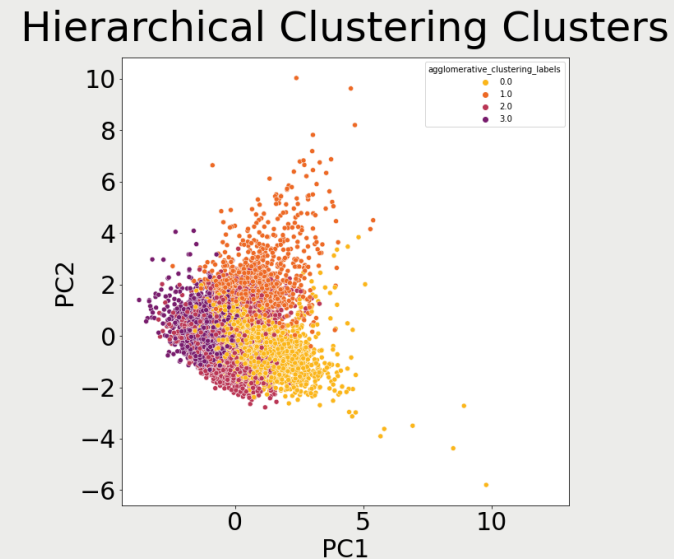
## II – Anachronie : choix final de l'algorithme



(Sélectionné)



(non non linéaires)



(trop couteux en temps)

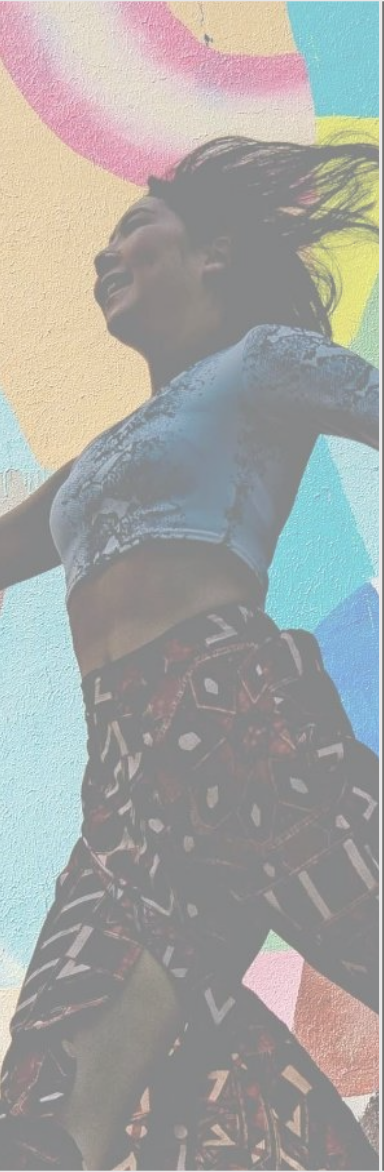


## II – Choix d'une métrique

Inertie, Silhouette, fit-time focus sur les changement brusques de tendances

Focus sur l'interprétabilité des clusters, les scores mathématiques sont utiles relativement à eux-mêmes moins absolument

« Pouvoir de discrimination » Centroids

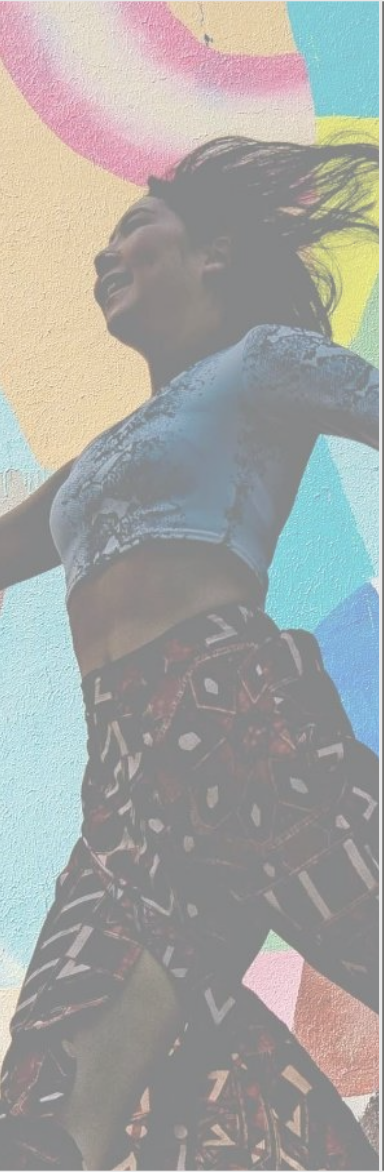


## II – Transformations des distributions

Logarithme Népérien [étalée à droite]

Fonction Réciproque  $1/x$  (avec inversion au préalable)  
[étalée à gauche]

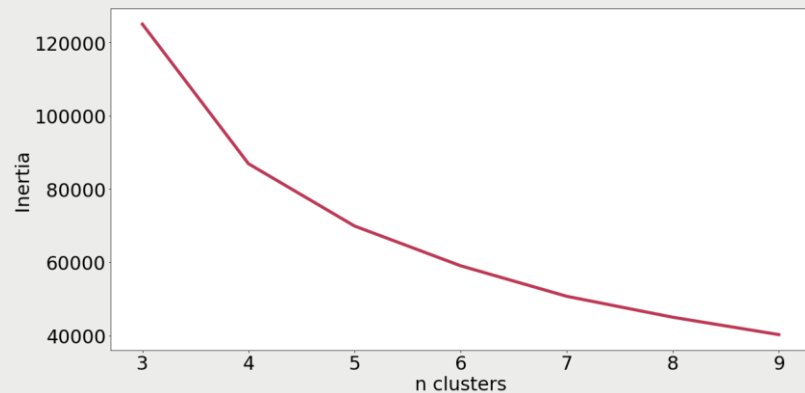
StandardScaler() centrage et réduction des données



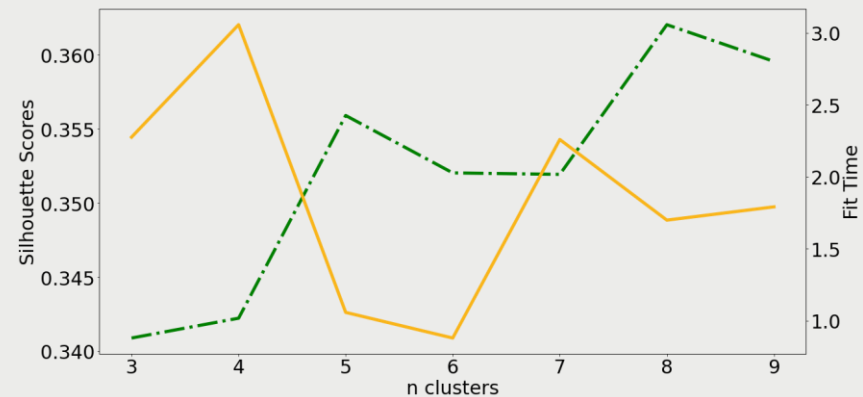


## II – RFM (4 Means)

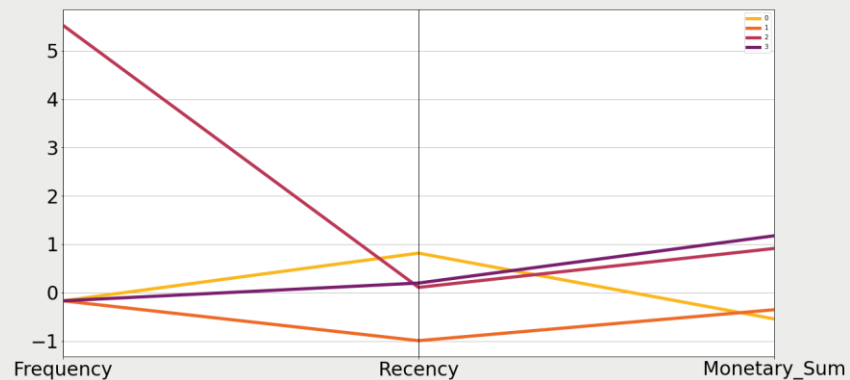
RFM Elbow Plot



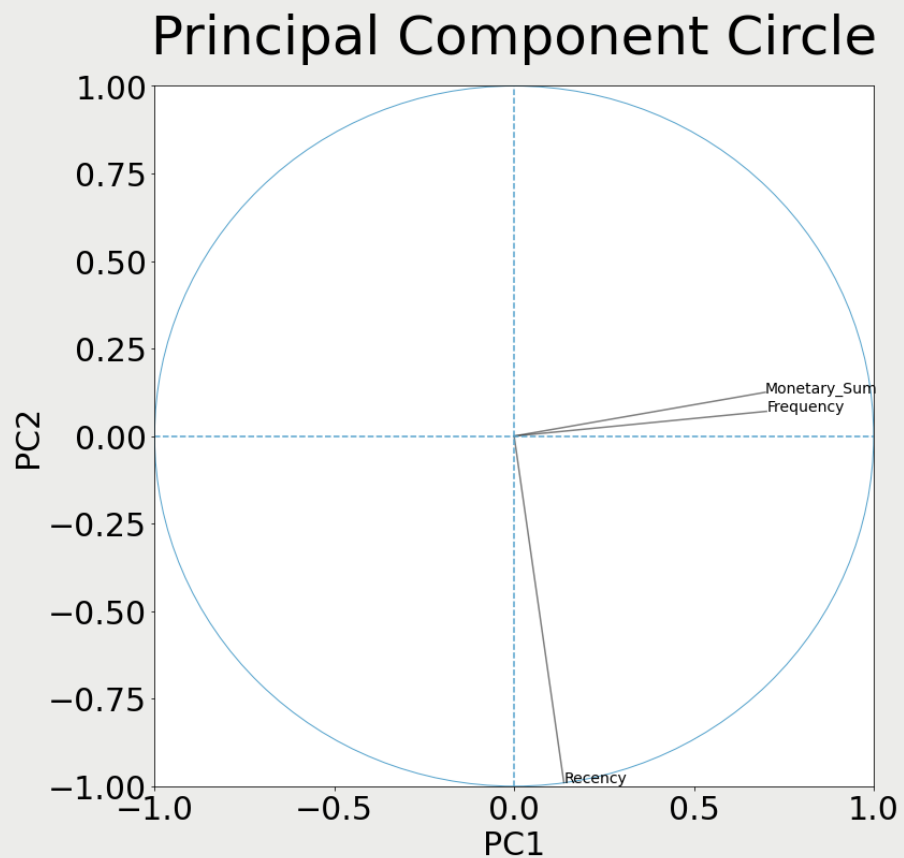
Kmeans Silhouette & Fit Time



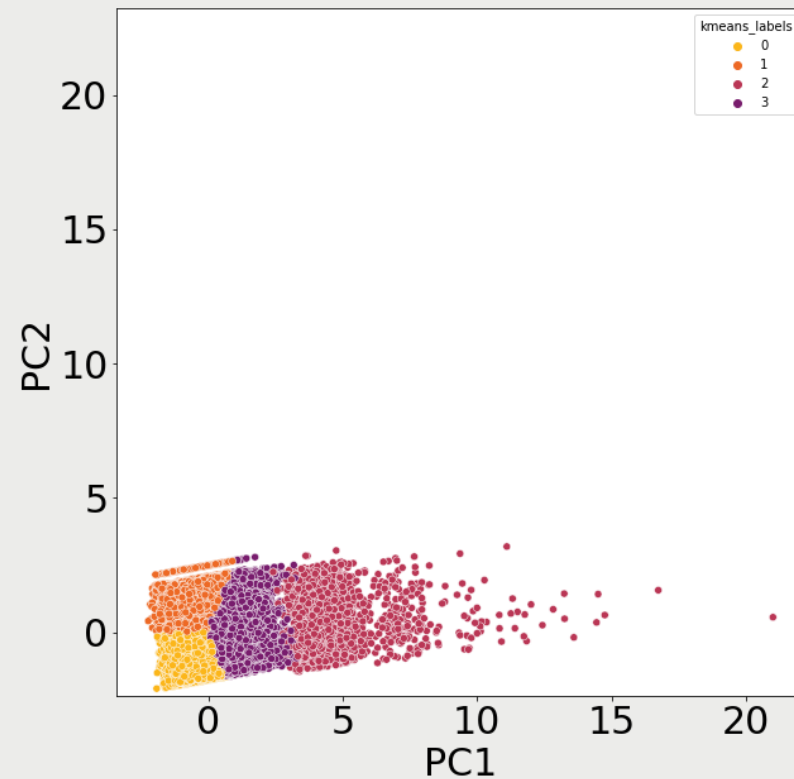
Centroids Parallel Plot



## II – RFM (4 Means – PCA 38%)



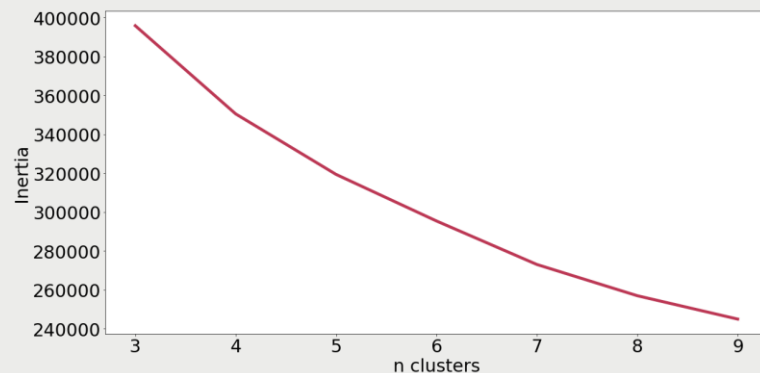
### Kmeans Clusters



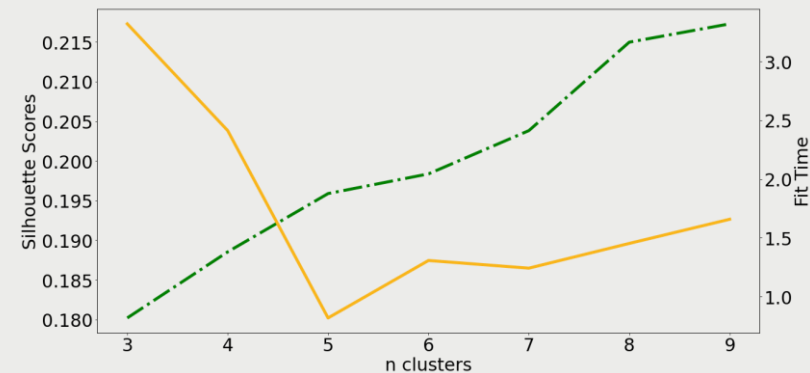


## II – RMS TBE (8 Means)

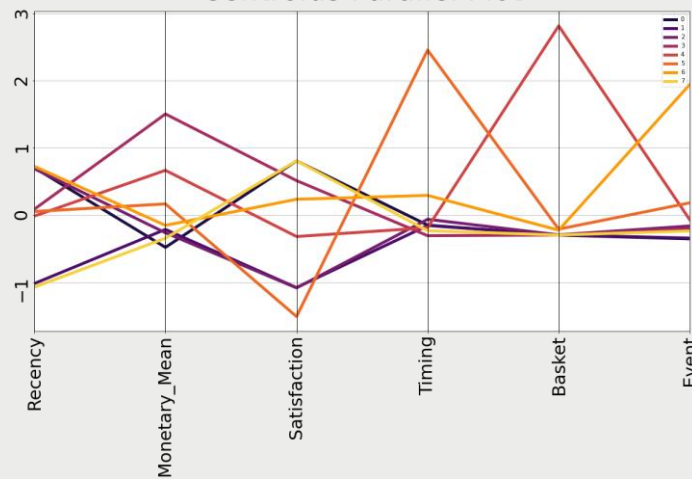
RFM Elbow Plot



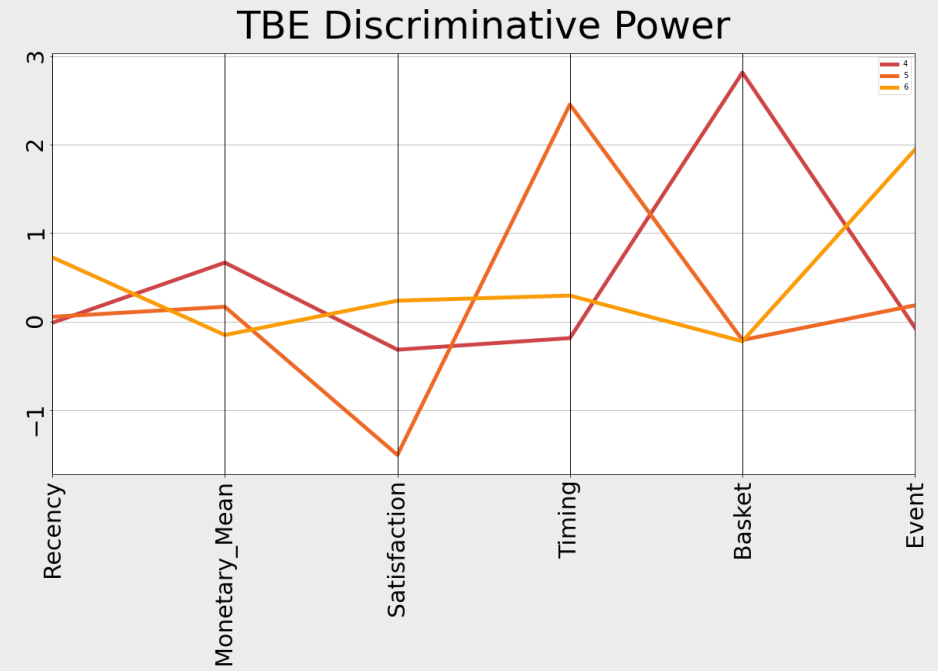
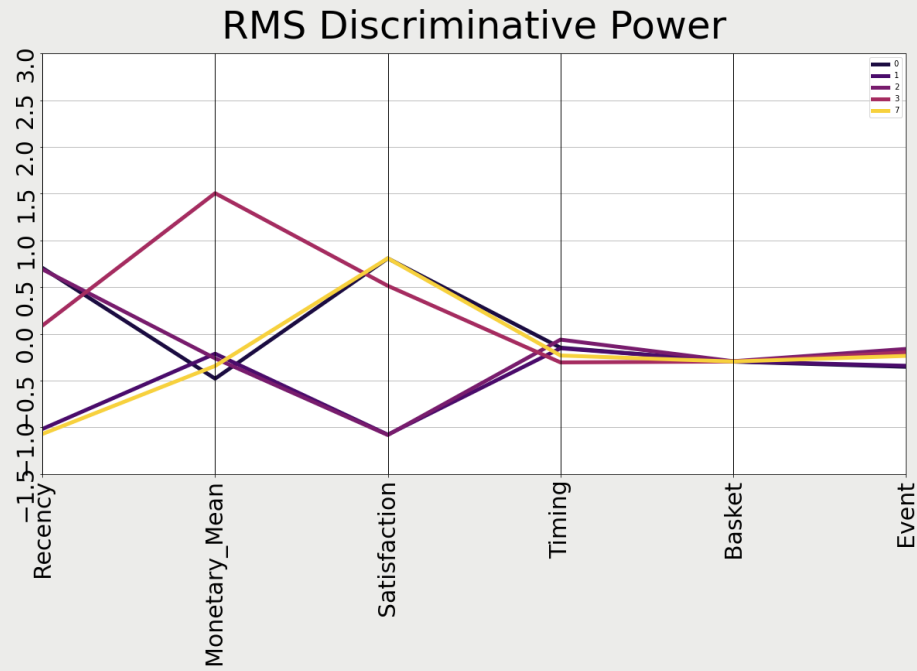
Kmeans Silhouette & Fit Time



Centroids Parallel Plot



## II – RMS TBE (8 Means) – « Pouvoir de discrimination »





Content  
Récent



20%

Triste  
Récent



14%

Content  
Lointain



18%

Triste  
Lointain



13%

Article  
Coûteux



11%

Retard  
Livraison



5%

Articles  
Panier



9%

Evène-  
ment



10%



### III – Contrat de Maintenance et PEP8

**LEV E**  
LAVARETOS

TABELA DE PREÇOS

DE 0 À 1 HORA  
**R\$ 6,00**

POR HORA ADICIONAL DO PÁRQUEO R\$ 1,00

TEMOS VAGAS  
PARA MENSALISTAS

DE SEGUNDA A SEXTA DAS 07H ÀS 19H  
SÁBADO, DOMINGO E FÉRIADO FECHADO

www.lev.com.br







### III – Méthode

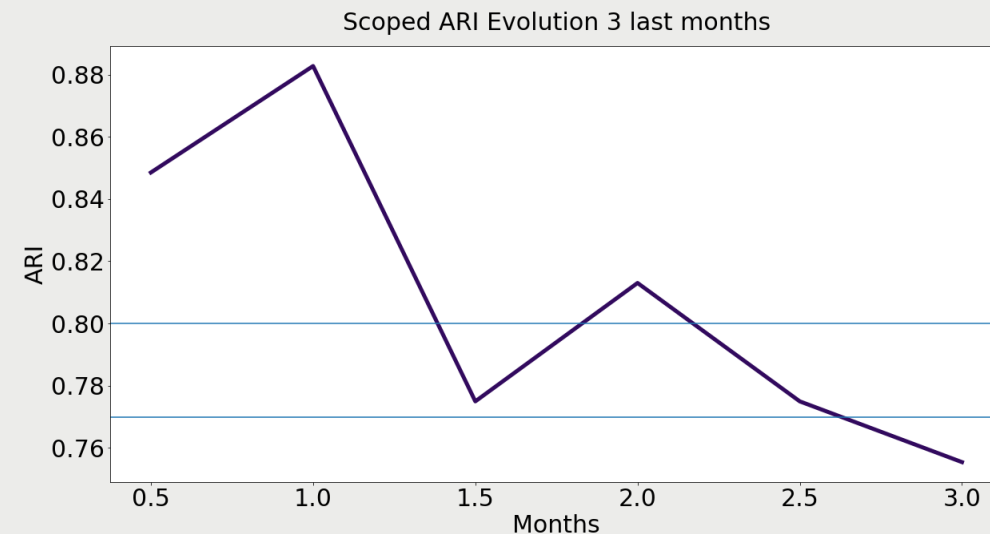
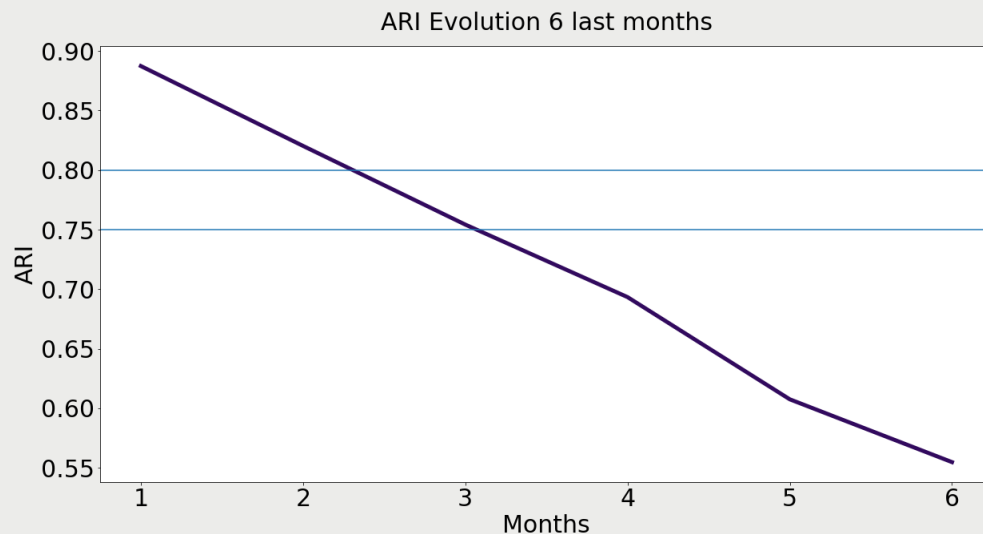
Comment les segments auraient évolué si la mission avait été donnée il y a 6 mois ?

Il y a 6 mois un modèle aurait été livré et utilisé sur les nouveaux clients jusqu'à la fin de nos données

On confronte ce modèle (standardisation incluse) à d'autres comme s'ils étaient livrés en prenant en compte les nouveaux clients (les nouveaux clients servent au .fit ici et pas dans le modèle précédent)

On se sert de l'ARI pour savoir si les clusters mois après mois sont similaires ou non

# III – Maintenance



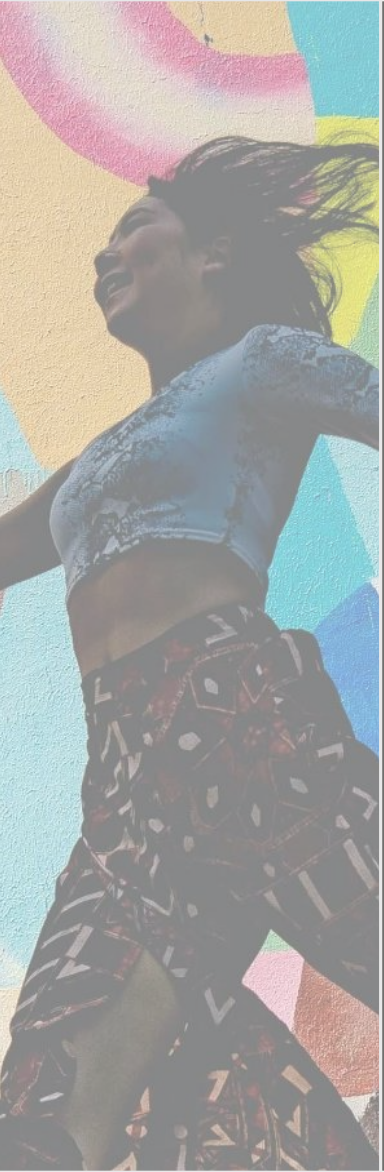
Le modèle pourra être mis à jour entre le 2  
et 3-ème mois pour rester efficace




### III – PEP8 Forme des livrables

La convention PEP8 est respectée sauf pour les « ; » à la fin des cellules afin de cacher les sorties Matplotlib et Seaborn.

”Some other good reasons to ignore a particular guideline:  
When applying the guideline would make the code less readable, even for someone who is used to reading code that follows this PEP.”





A woman with long dark hair, wearing a light blue patterned crop top and a dark skirt with a colorful geometric pattern, is captured in a dynamic pose. She is holding two black shopping bags, one in each hand, and her arms are extended. The background is a vibrant, abstract composition of various colored geometric shapes and patterns, including a large circular mirror in the upper right corner. The overall scene conveys a sense of joy and fashion.

**Merci pour votre attention !**