PROJET5
Segmentez des clients d'un site e-commerce

Apprendre des modèles de clustering

AL SAMMAN Wassim
Data Scientist Apprenti

PAPOUTSIS Panayotis
Data Scientist -Mentor

Introduction et problématique

- Améliorer les campagnes de communication de l'entreprise Olist
- En tant que consultant chez Olist:
  - Fournir aux équipes e-commerces une segmentation des clients
  - Fournir à l'équipe marketing une description actionnable de la segmentation et proposer un contrat de maintenance

La focntion client()

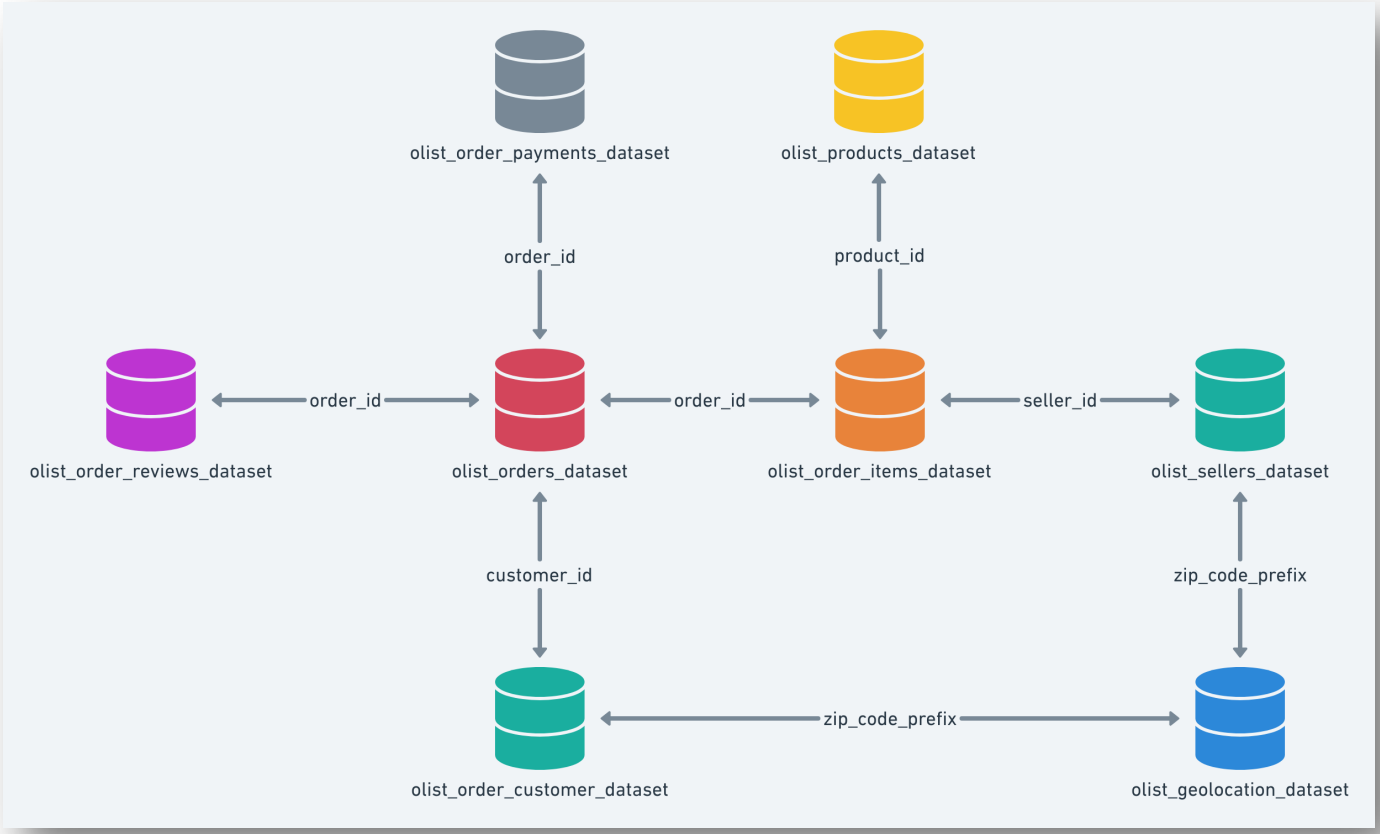
- Le code complet pour ce projet

```
def client(Choice='RFM',algo='KMeans',t0=150,nt=200,n_clustersAgglomerative=5,min_samplesDBSCAN=2,epsDBSCAN=3,
           OptimalK=5,seed=6,TestNewFeatures=False,showfliers=False):
```

- La variable Choice:
  - 'Percentage', 'RFM', 'Elblow', 'SilhouetteScore', 'RFMBox', 'InfoCluster', 'RadarChart', et 'Periode'
- La variable algo:
  - 'KMeans','DBSCAN', ou 'Agglomerative'
- Les autres variables

Cleanning effectué

- Le schéma des données
- Merge selon le lien donné sur le schéma
- Supprimer les doublons. Pas toujours comme ‘product_id’.
- Prendre en compte les commandes ‘delivered’



Feature engineering et pourcentage des clients

- Le pourcentage des clients:

```
client(Choice='Percentage')
```

```
Pourcentage des clients ayant acheté plus d'une commande 2.98%
```

- Feature engineering:
  - Calculer les variables R,F, et M

```
RFM=client(Choice='RFM')
```

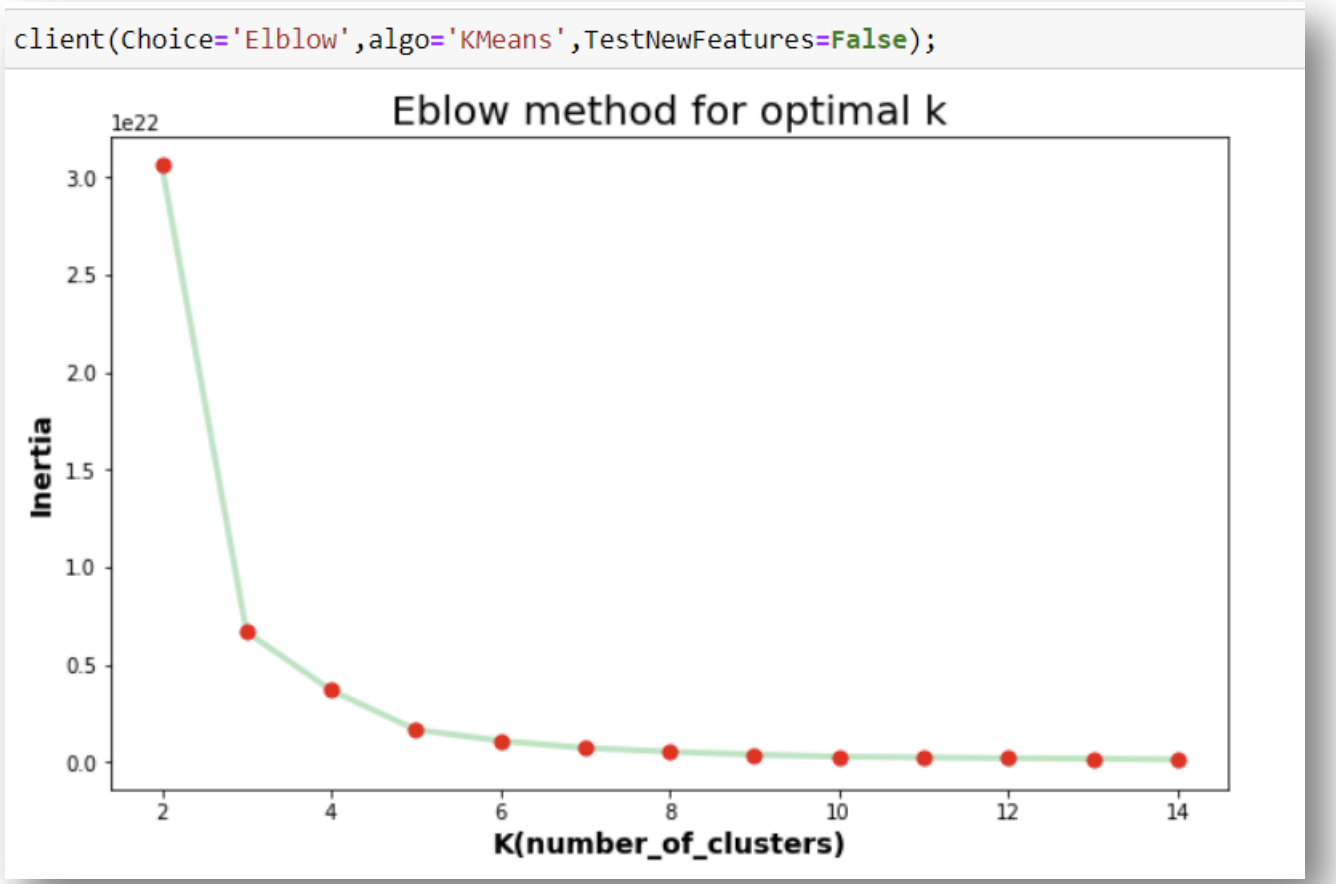
```
RFM
```

customer_unique_id	R	RFloat	F	M
2f64e403852e6893ae37485d5fcacdaf	2016-10-03 16:56:50	2.016100e+13	1	39.09
61db744d2f835035a5625b59350c6b63	2016-10-03 21:13:36	2.016100e+13	1	53.73
8d3a54507421dbd2ce0a1d58046826e0	2016-10-03 22:06:03	2.016100e+13	1	133.46

KMeans clustering

- Elbow méthode

```
client(Choice='Elblow',algo='KMeans',TestNewFeatures=False);
```



KMeans clustering

- SilhouetteScore

```
client(Choice='SilhouetteScore',algo='KMeans',OptimalK=5,TestNewFeatures=False)
```

```
SilhouetteScore est:64.38 %
```

- InfoCluster

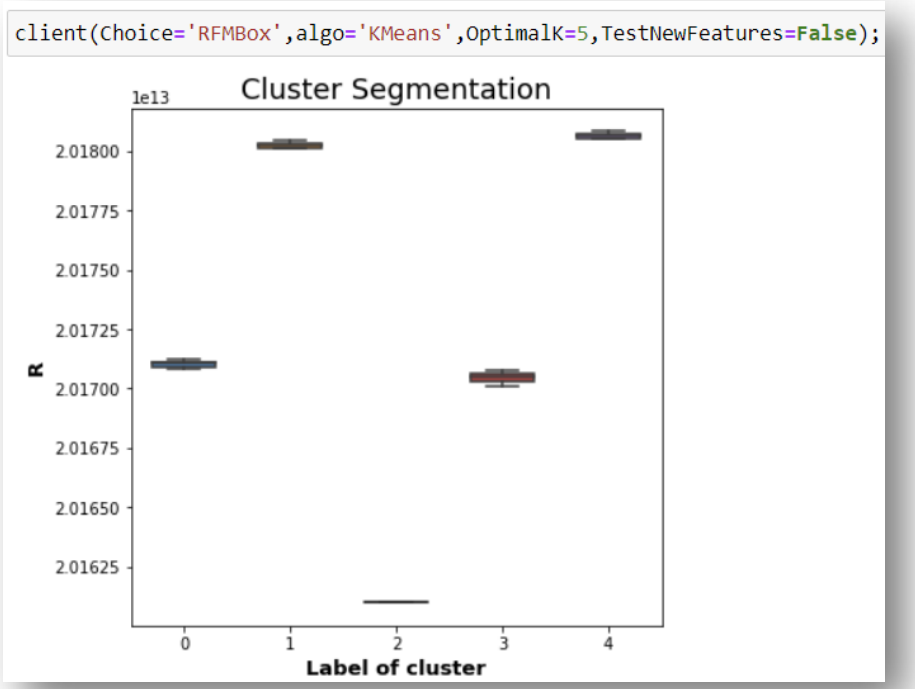
```
InfoCluster=client(Choice='InfoCluster',algo='KMeans',OptimalK=5, TestNewFeatures=False)
InfoCluster
```

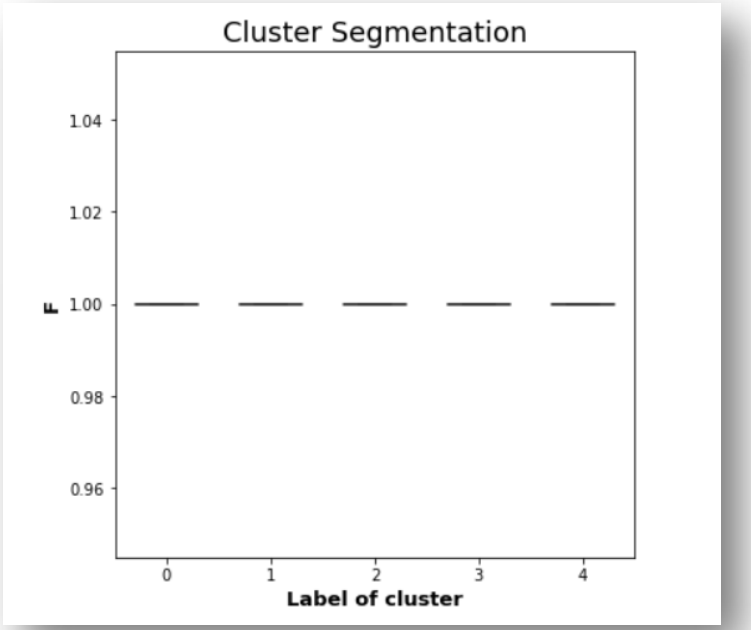
Labels	NumberOfClient	F	M	MeanDeltaDate
0	24436	1.03	161.99	0 days 00:09:00.892535603
1	26403	1.04	159.73	0 days 00:06:32.270764685
2	250	1.01	176.59	0 days 00:40:36.676000
3	16758	1.03	159.36	0 days 00:17:49.815073397
4	24907	1.04	166.79	0 days 00:06:58.432810053

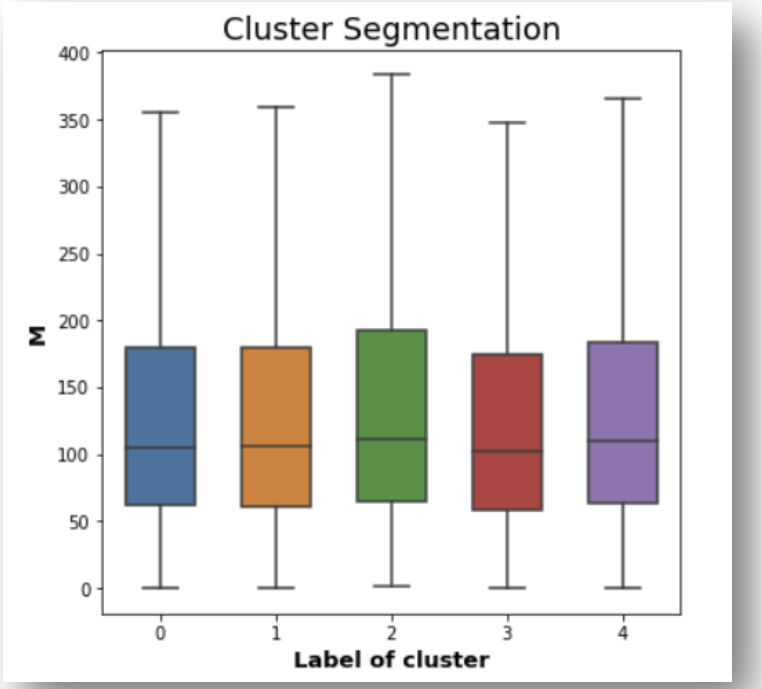
KMeans clustering

- RFMBox

```
client(Choice='RFMBox',algo='KMeans',OptimalK=5,TestNewFeatures=False);
```



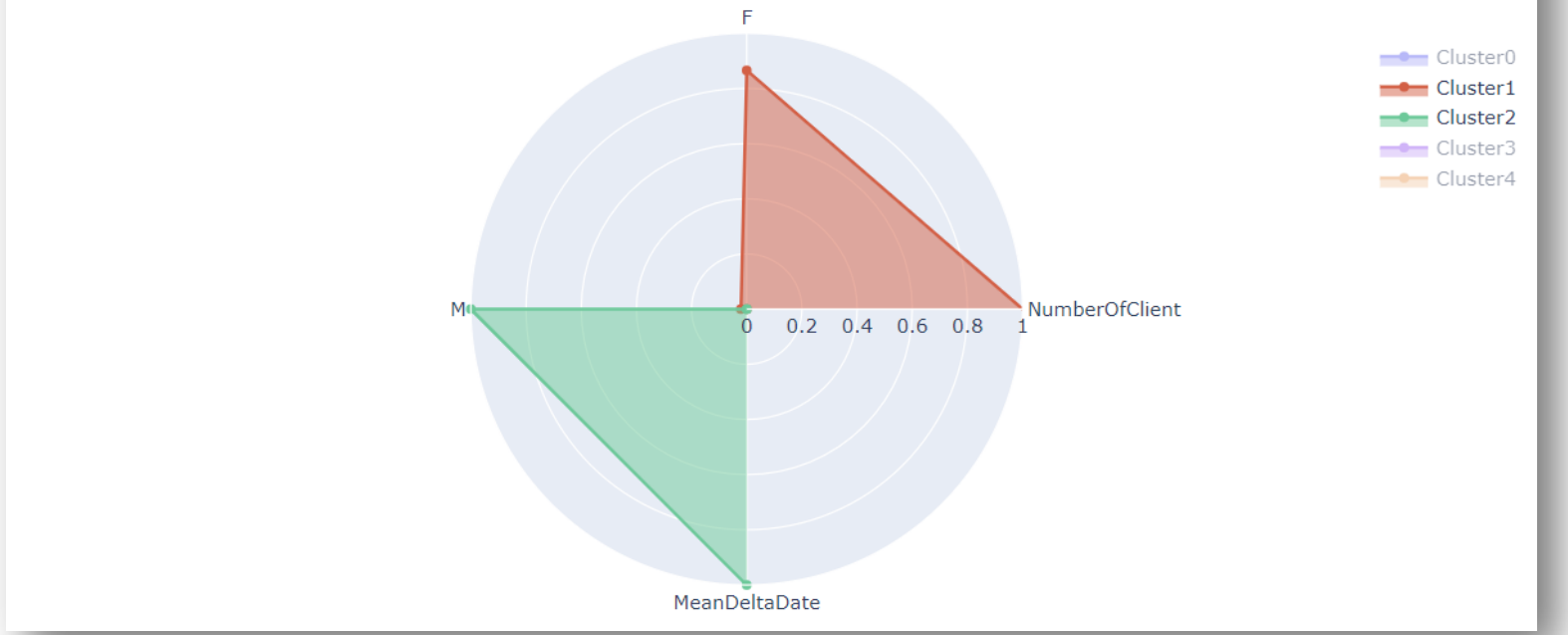




KMeans clustering

- RadarChart

```
client(Choice='RadarChart',algo='KMeans',OptimalK=5,TestNewFeatures=False)
```



KMeans clustering

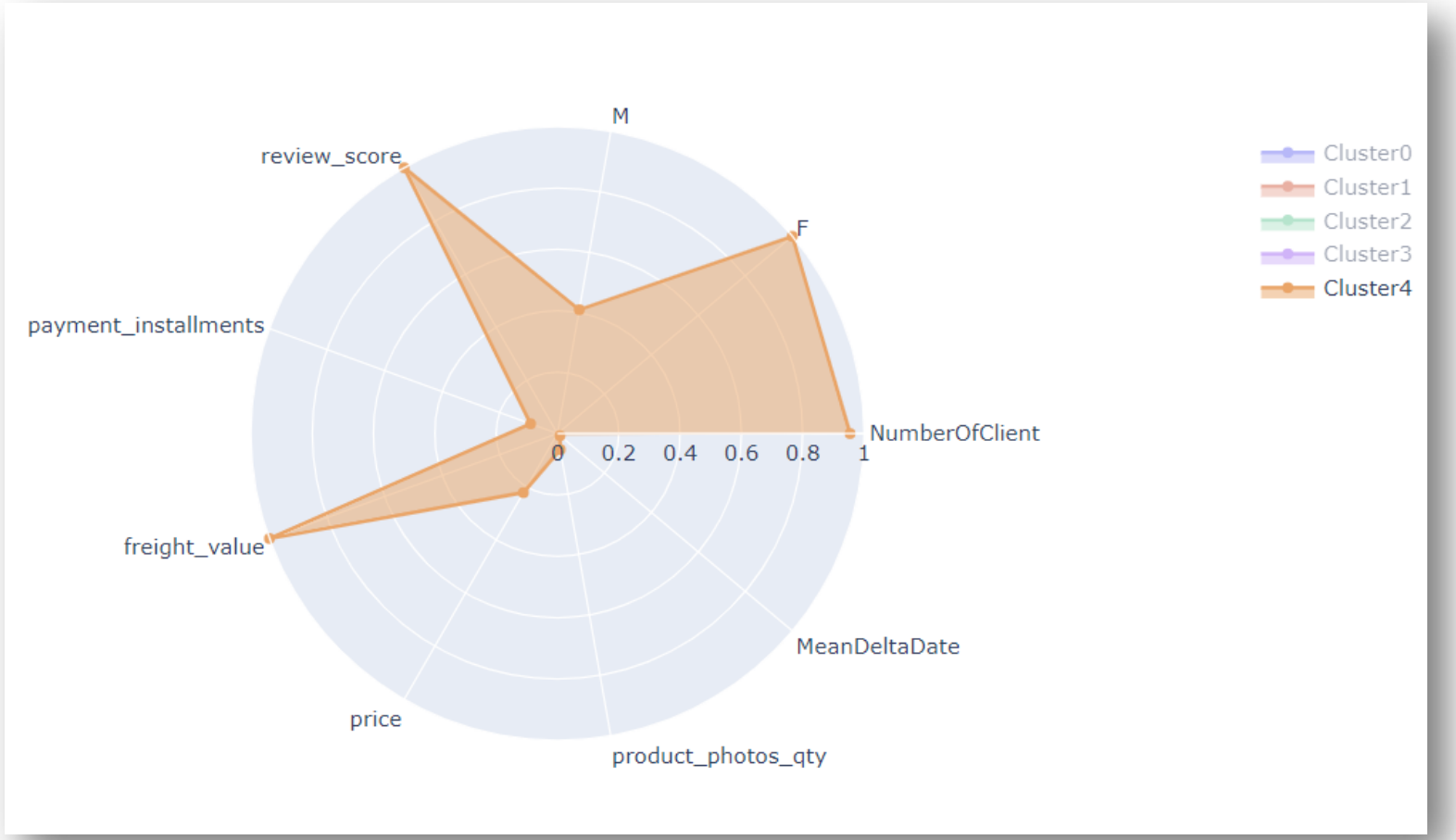
- Avec des nouvelles features:
  - OptimalK a la meme valeur
  - SilhouetteScore n'est pas différente
  - RFMBox n'est pas différent
  - InfoCluster avec les nouvelles features

```
InfoCluster=client(Choice='InfoCluster',algo='KMeans',OptimalK=5,TestNewFeatures=True)
InfoCluster
```

Labels	NumberOfClient	F	M	review_score	payment_type	payment_installments	freight_value	price	product_photos_qty	MeanDeltaDate
0	24052	1.03	162.14	4.15	[credit_card, boleto, voucher, debit_card]	3.02	19.66	125.36	2.28	0 days 00:09:09.528105770
1	25958	1.04	160.02	4.00	[credit_card, boleto, debit_card, voucher]	2.73	19.97	123.42	2.24	0 days 00:06:38.995492719
2	248	1.01	177.63	3.97	[boleto, credit_card,	3.68	20.13	141.40	2.57	0 days

KMeans clustering

- RadarChart avec les nouvelles features



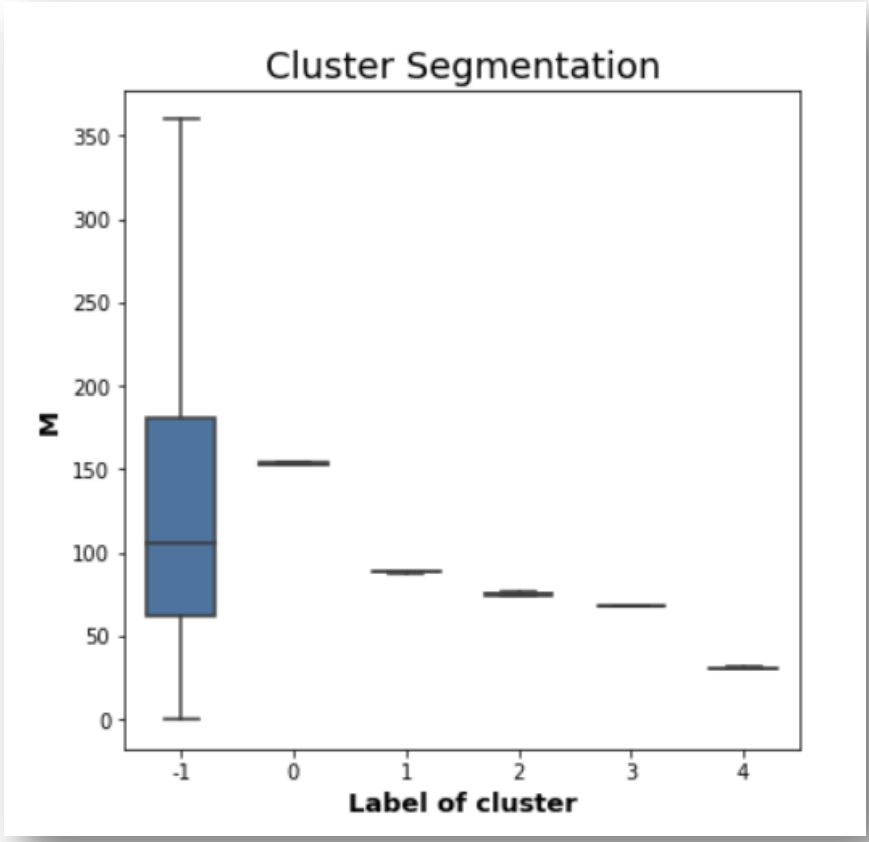
DBSCAN clustering

- SilhouetteScore

```
client(Choice='SilhouetteScore',algo='DBSCAN',min_samplesDBSCAN=2,epsDBSCAN=3,TestNewFeatures=False)
```

```
SilhouetteScore est:-97.5 %
```

- RFMBox



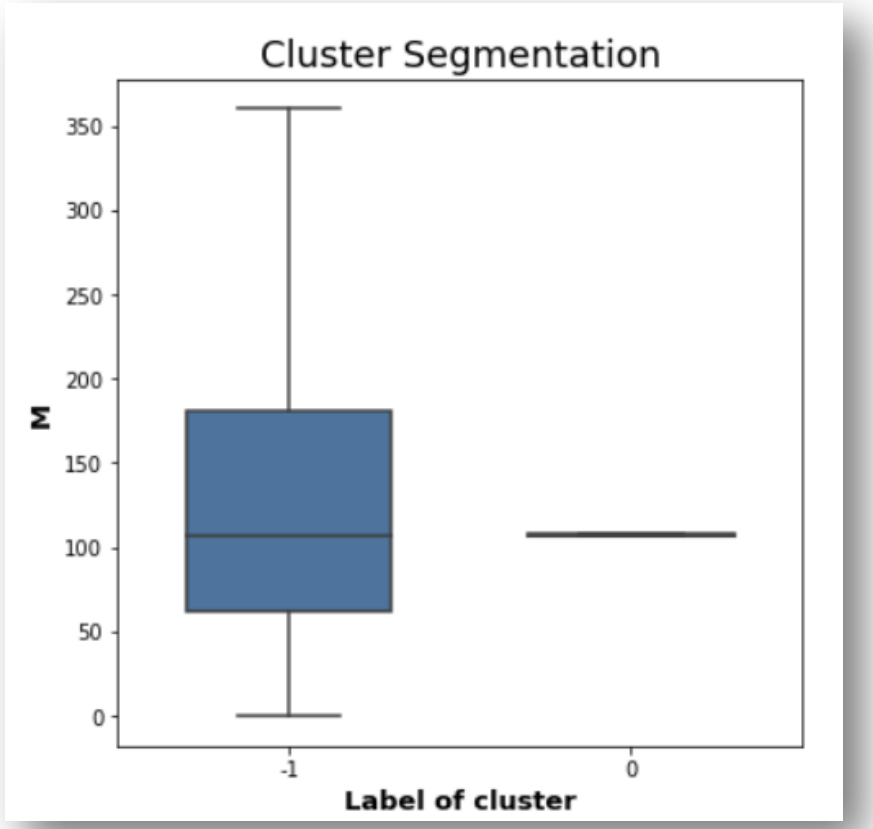
DBSCAN clustering

- SilhouetteScore avec des nouvelles features

```
client(Choice='SilhouetteScore',algo='DBSCAN',min_samplesDBSCAN=2,epsDBSCAN=3,TestNewFeatures=True)
```

```
SilhouetteScore est:-11.82 %
```

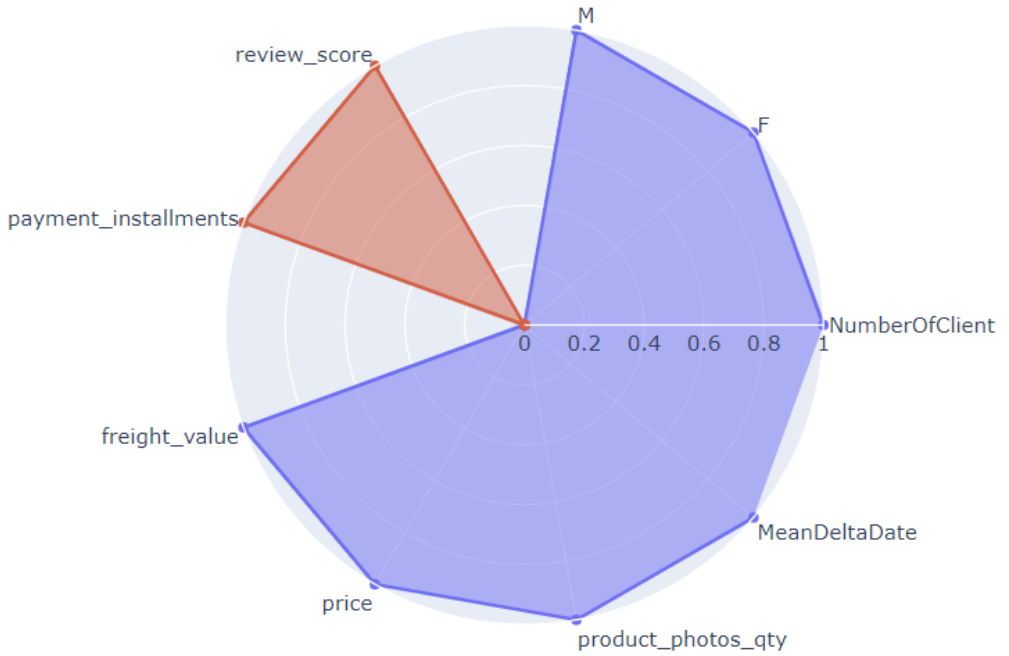
- RFMBox avec des nouvelles features



DBSCAN clustering

- RadarChart avec les nouvelles features

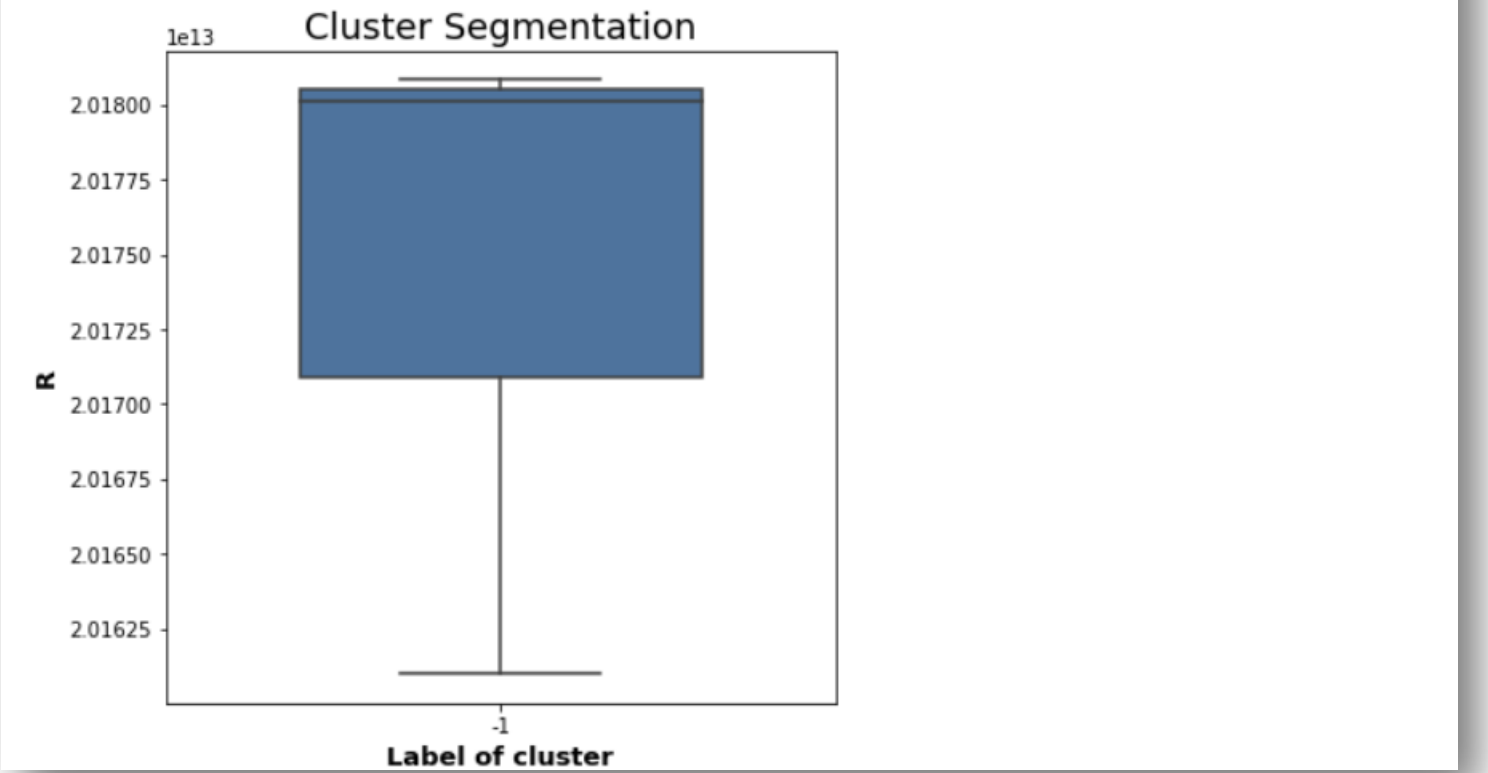
```
client(Choice='RadarChart',algo='DBSCAN',min_samplesDBSCAN=2,epsDBSCAN=3,TestNewFeatures=True)
```



DBSCAN clustering

- RFMBox en changeant le paramètre min_samples

```
client(Choice='RFMBox',algo='DBSCAN',min_samplesDBSCAN=5,epsDBSCAN=3,TestNewFeatures=False);
```



- Remarque : on peut aussi avoir InfoCluster.

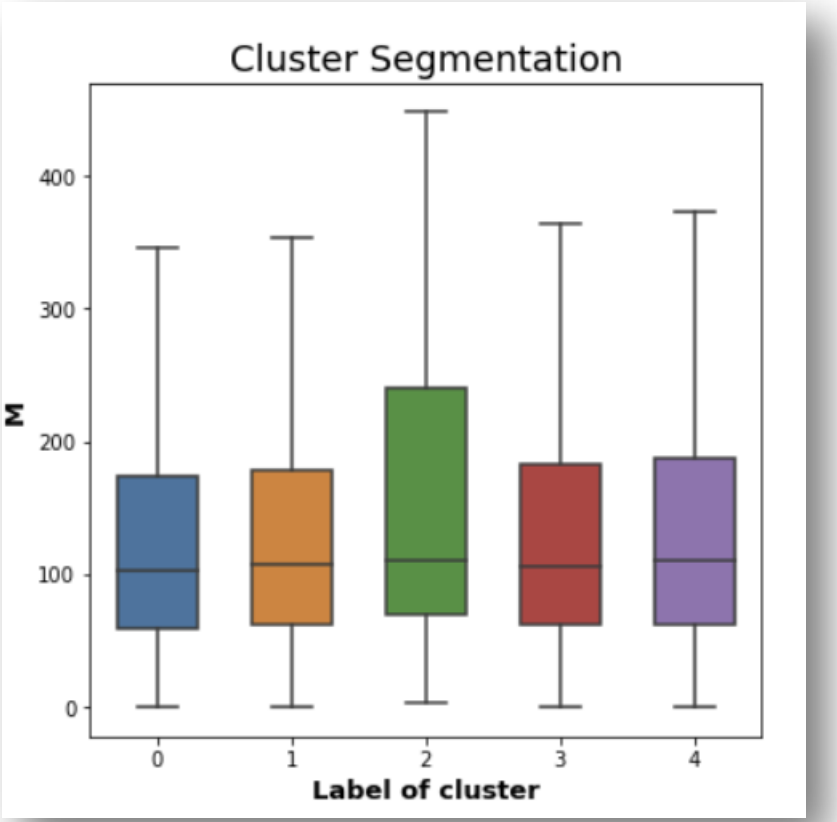
Agglomerative clustering

- SilhouetteScore

```
client(Choice='SilhouetteScore',algo='Agglomerative',n_clustersAgglomerative=5,TestNewFeatures=False)
```

```
SilhouetteScore est:61.68 %
```

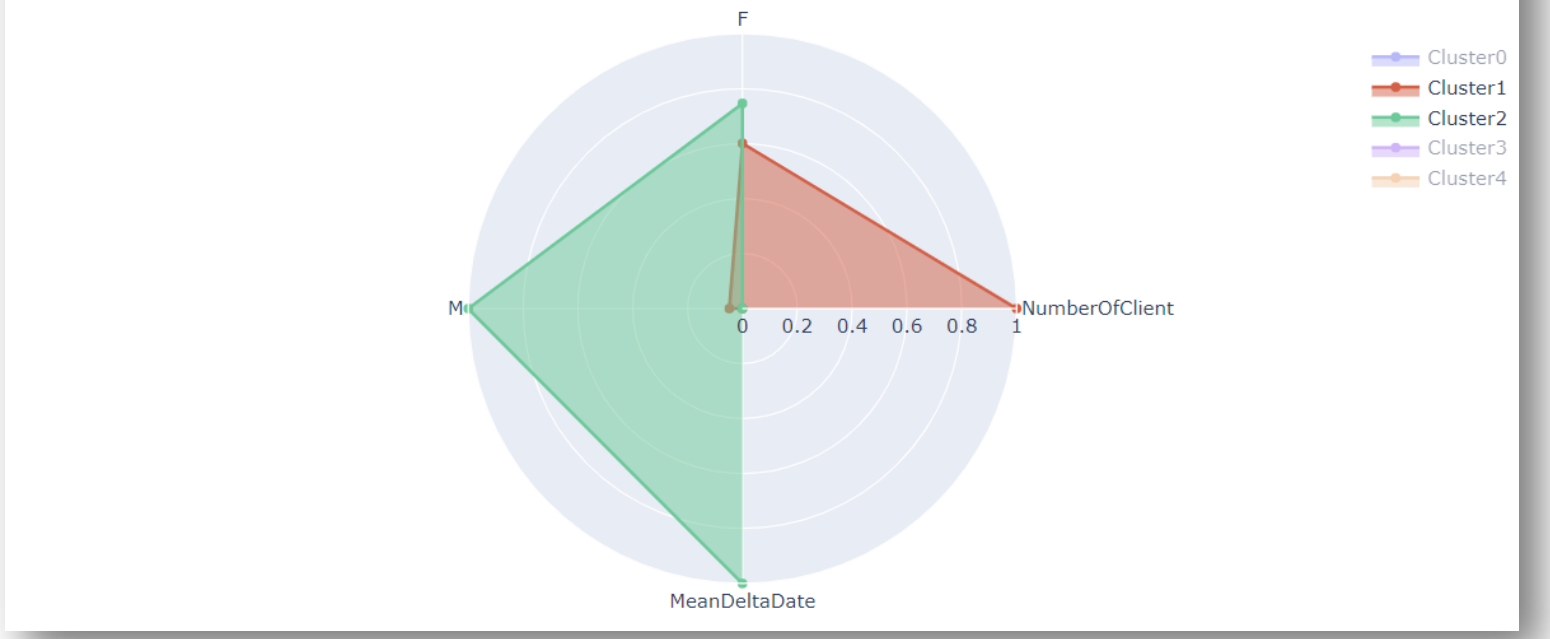
- RFMBox



Agglomerative clustering

- RadarChart

```
client(Choice='RadarChart',algo='Agglomerative',n_clustersAgglomerative=5,TestNewFeatures=False)
```



- Remarques:
  - On peut avoir les résultats avec les nouvelles features.
  - On peut aussi avoir InfoCluster.

Le choix final du modèle

- DBSCAN le plus mauvais
  - La grande quantité des données
- Agglomerative est bien mais n’est pas le meilleur
  - La grande quantité des données
  - Très sensible aux valeurs aberrantes
  - Sensible à l'ordre de données
- Kmeans est le meilleur
  - Pas de problème pour la qunatité des données
  - Pas sensible aux valeurs aberrantes

Contrat de maintenance

- Le but du contrat de maintenance
- La méthodologie suivie
- Le résultat pour le modèle Kmeans
- L'effet de nouvellles features

```
ARIScore=client(Choice='Periode',algo='KMeans',t0=50,nt=100)
```

```
ARIScore
```

	ARIScore	>0.8
50	0.0000	False
150	0.1374	False
250	0.0316	False
350	0.3737	False
450	0.4261	False
550	0.6993	False
650	0.8252	True

```
ARIScore=client(Choice='Periode',algo='KMeans',t0=600,nt=15)
```

```
ARIScore
```

	ARIScore	>0.8
600	0.6437	False
615	0.8376	True
630	0.8304	True
645	0.8261	True
660	0.8228	True
675	1.0000	True
690	1.0000	True

Contrat de maintenance

- Le résultat pour le modèle DBSCAN
- Pas d'effet avec le changement du paramètre

```
ARIScore=client(Choice='Periode',algo='DBSCAN',t0=50,nt=100)
```

```
ARIScore
```

	ARIScore	>0.8
50	1.0	True
150	1.0	True
250	1.0	True
350	1.0	True
450	1.0	True
550	1.0	True
650	1.0	True

	ARIScore	>0.8
10	1.0	True
50	1.0	True
90	1.0	True
130	1.0	True
170	1.0	True
210	1.0	True
250	1.0	True
290	1.0	True
330	1.0	True
370	1.0	True
410	1.0	True
450	1.0	True
490	1.0	True
530	1.0	True
570	1.0	True
610	1.0	True
650	1.0	True
690	1.0	True

Contrat de maintenance

- Le résultat pour le modèle Agglomerative

```
ARIScore=client(Choice='Periode',algo='Agglomerative',t0=50,nt=100)
```

```
ARIScore
```

	ARIScore	>0.8
50	0.0000	False
150	0.1302	False
250	0.0299	False
350	-0.0166	False
450	0.5378	False
550	0.8857	True
650	0.6954	False

	ARIScore	>0.8
500	0.8362	True
515	0.8534	True
530	0.8680	True
545	0.8811	True
560	0.8931	True
575	0.9040	True
590	0.9159	True
605	0.9218	True
620	0.8521	True
635	0.7427	False
650	0.6954	False
665	0.6702	False
680	1.0000	True

Conclusion générale

- Avoir d’autres résultats avec la fonction client()
- A l’aise avec des nouvelles applications
- Toujours motivé