

## Segmentation d'une clientèle

Basée sur un relevé de factures

Francois BANGUI

# Parcours Datascientist : projet 5

## Annexes

Annexe 1 : fichiers du projet

Annexe 2 : organisation et processus de l'étude

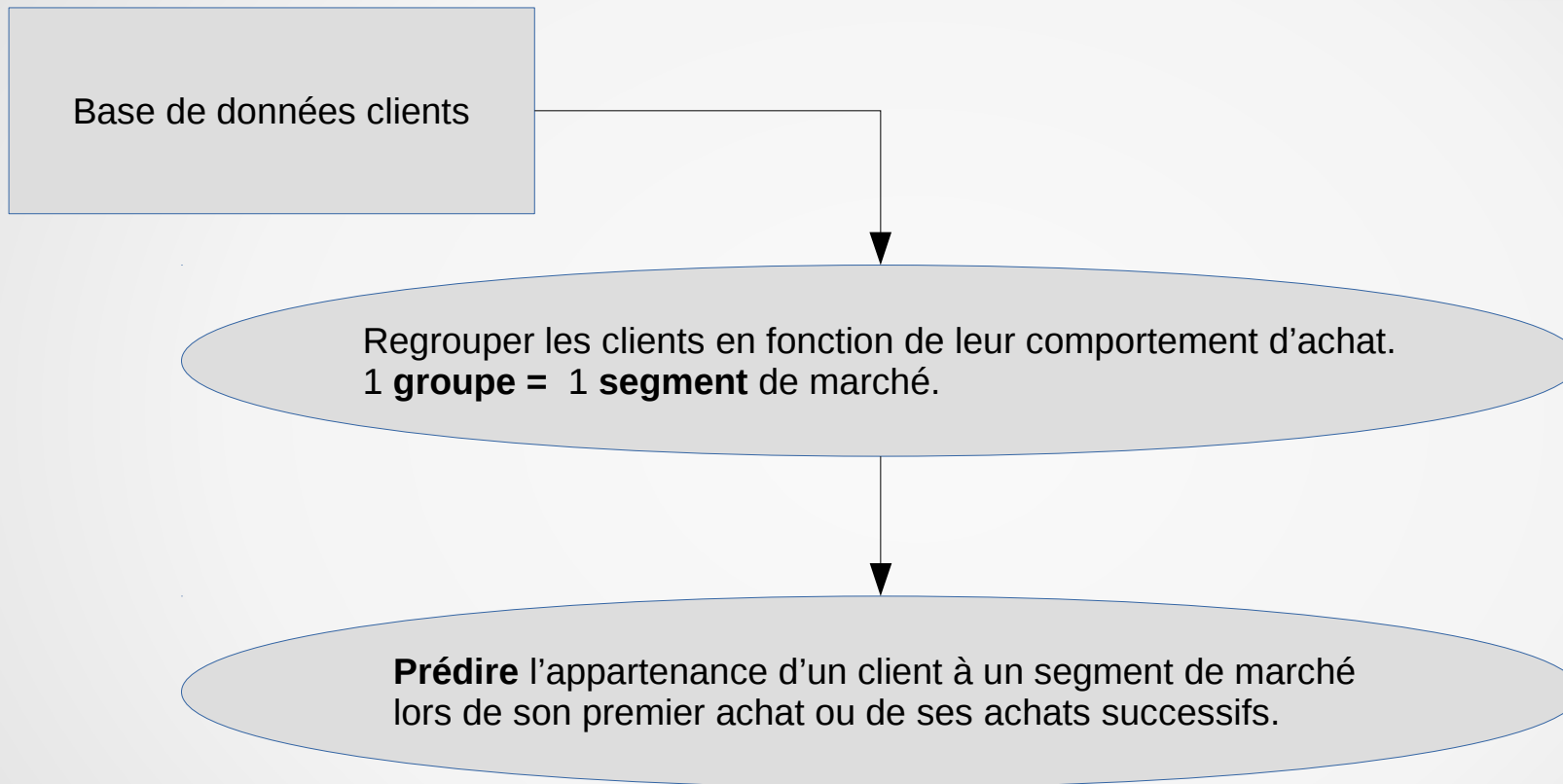
Annexe 3 : Variables issues du score RFM

Annexe 4 : variables issues du traitement NLP

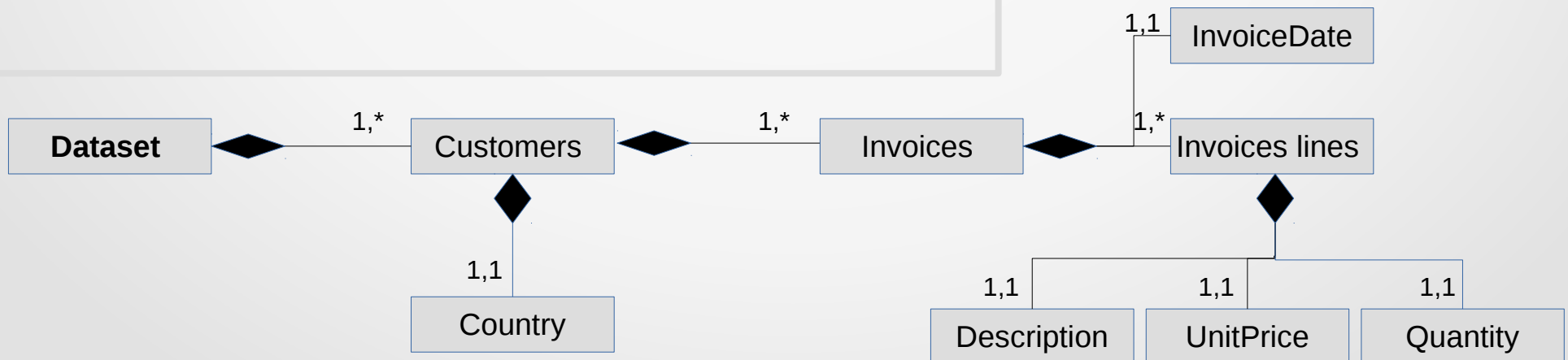
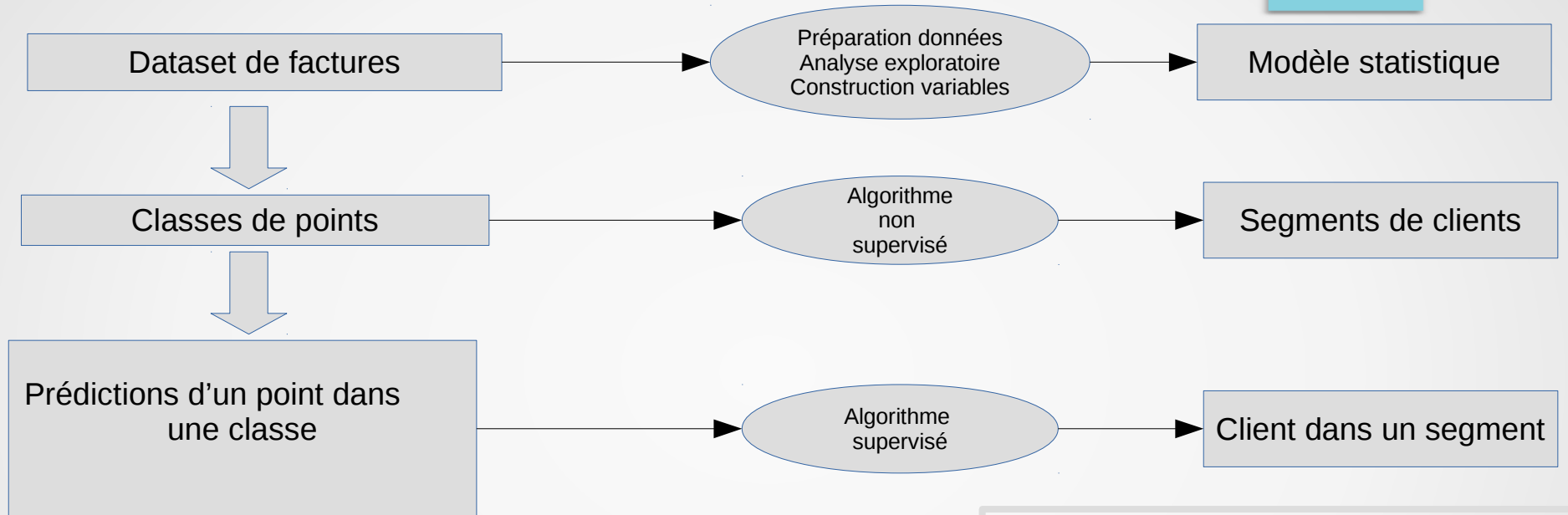
Annexe 5 : variables issues de la date de facturation

Annexe 6 : API WEB

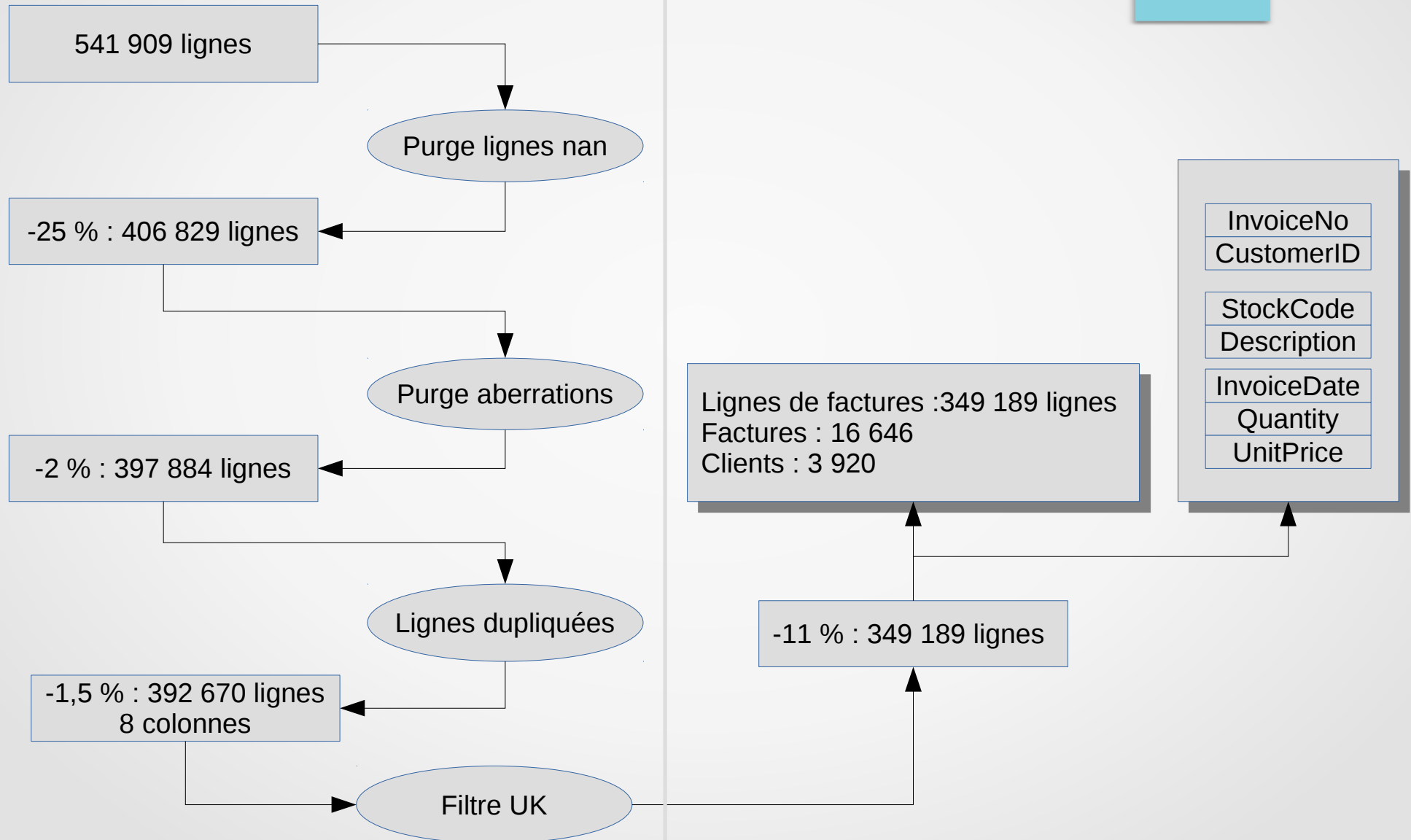
# Formulation du problème : mission



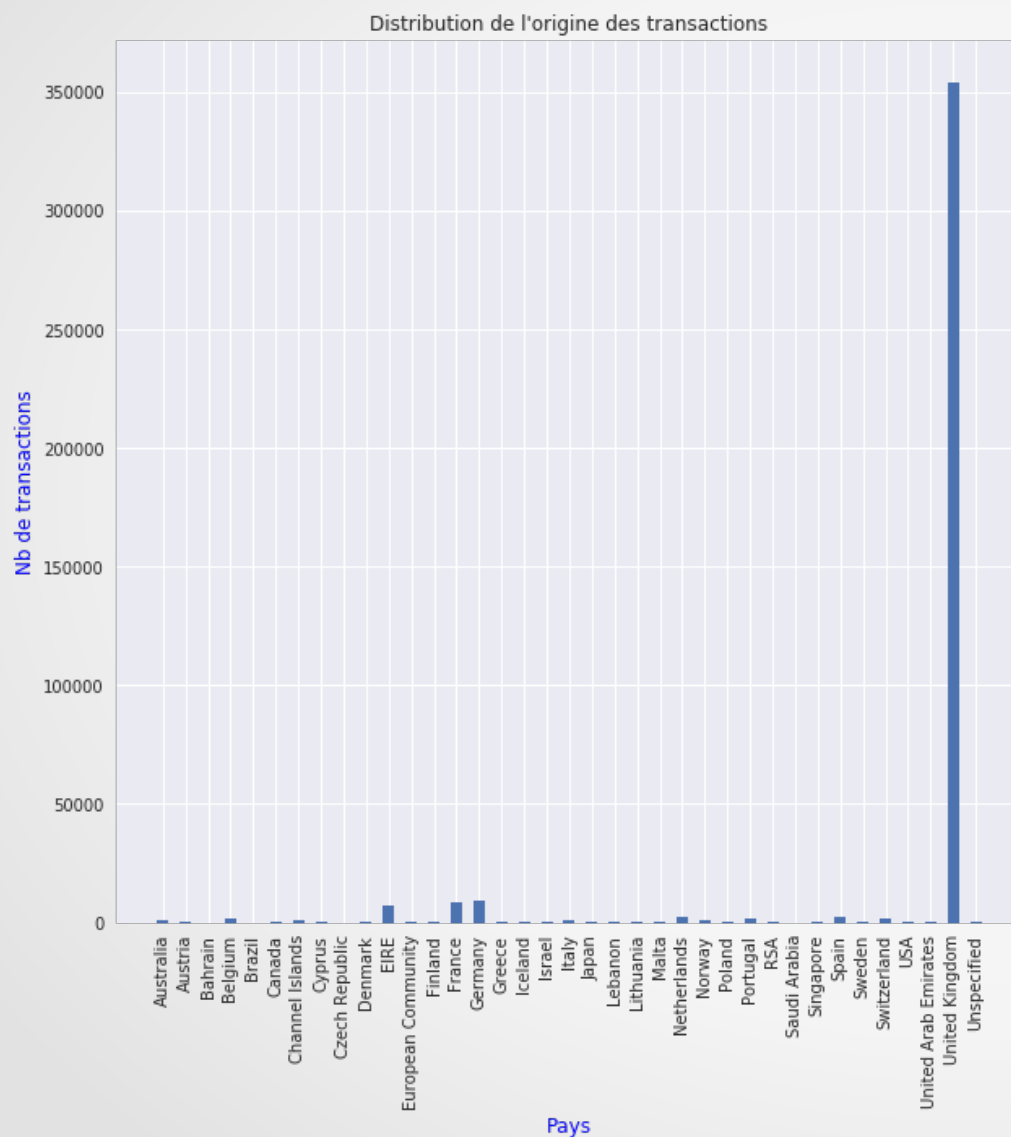
# Processus global du projet



# Préparation des données



# Exploration des données



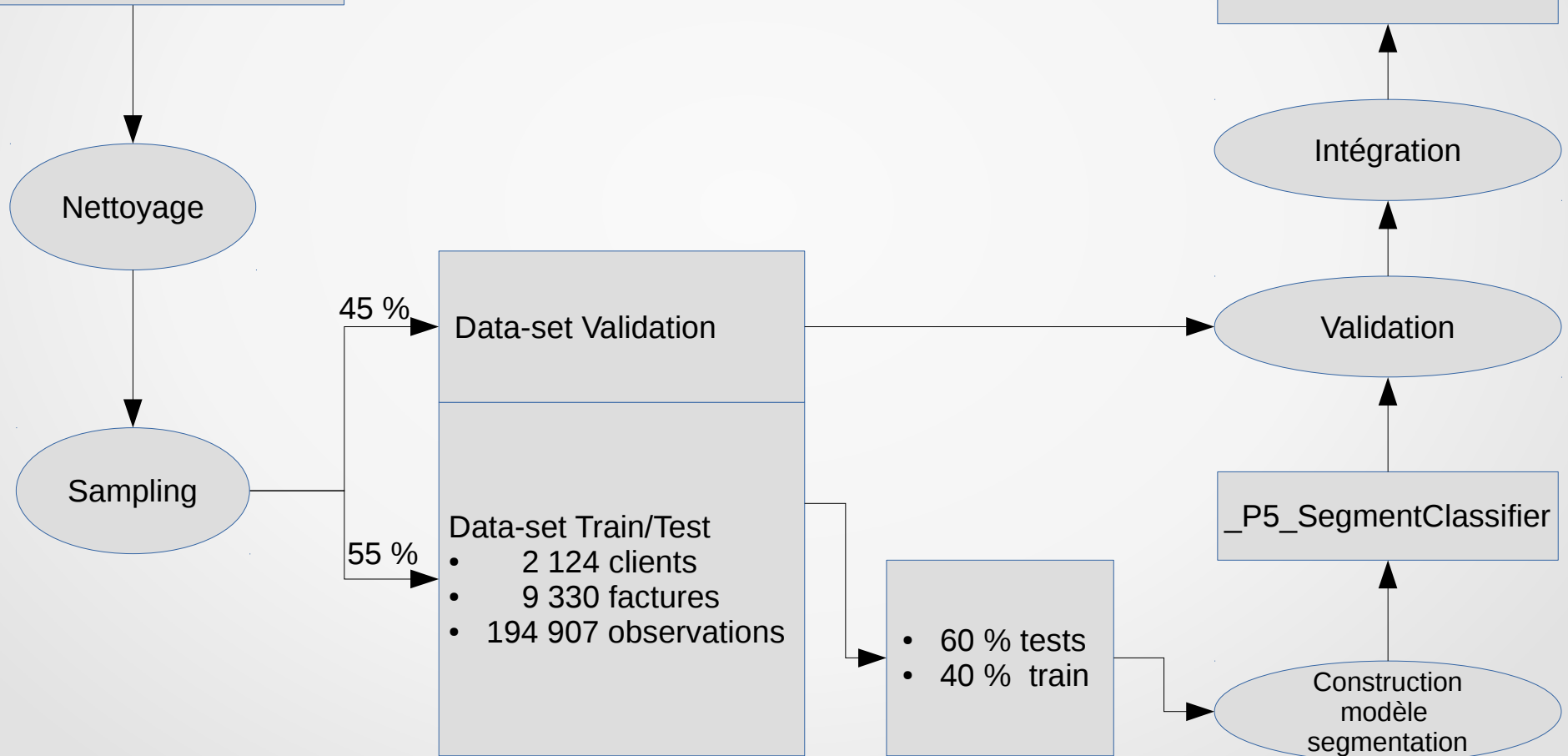
Transactions dominées par UK

Filtrage  
Nettoyage

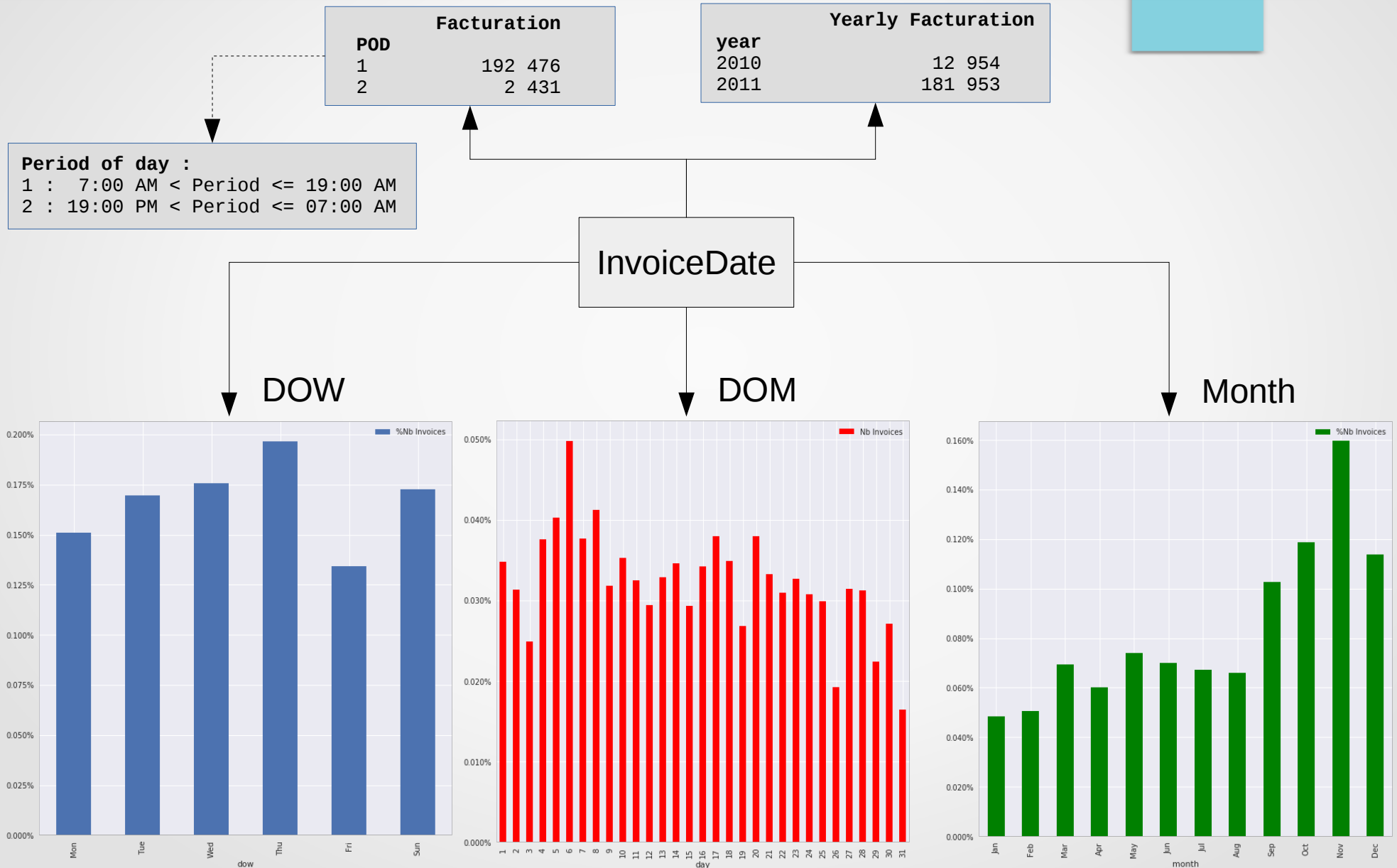
Data-set UK

# Sampling

- 3 920 clients
- 16 646 factures
- 349 189 lignes de facture

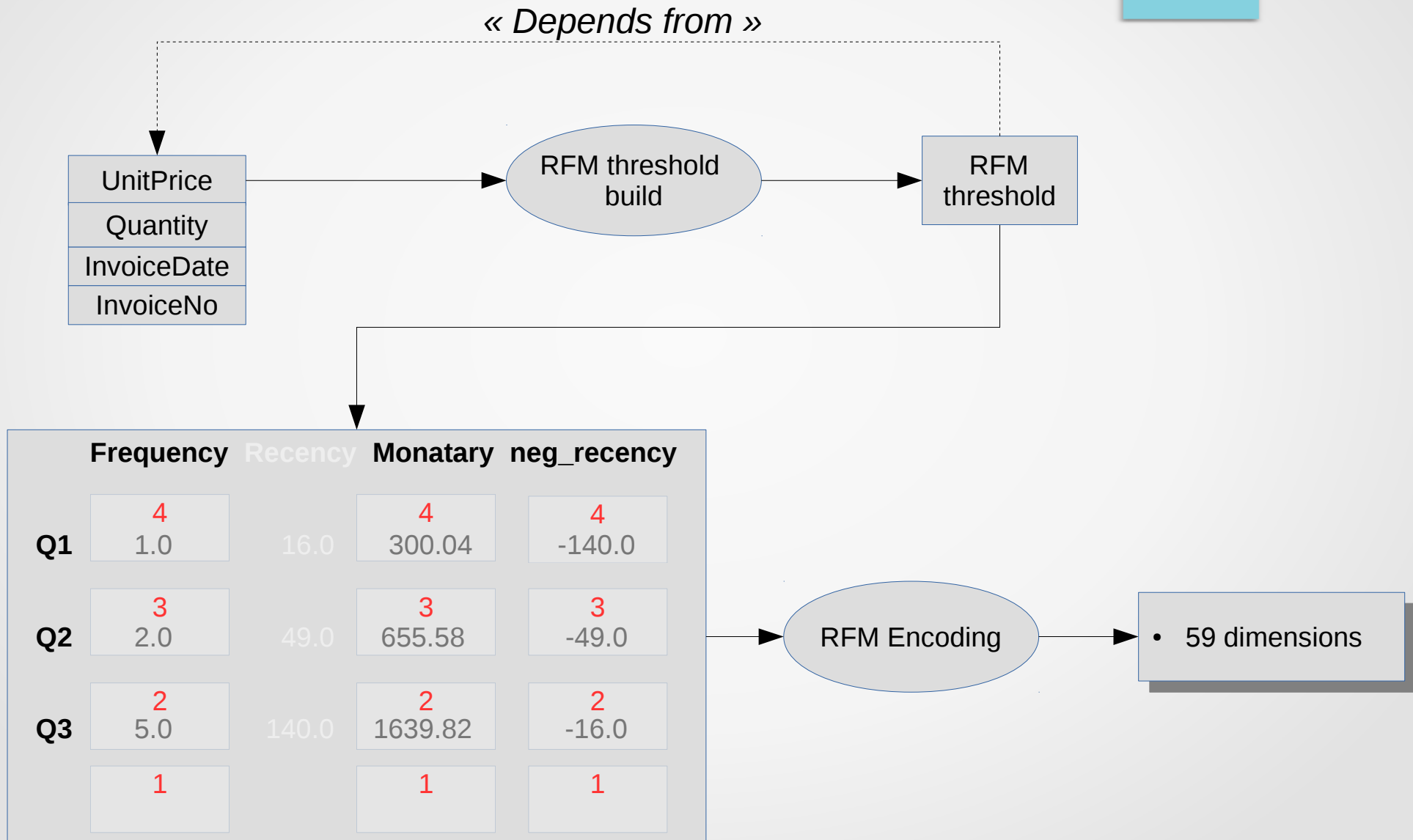


# Analyse exploratoire : Activité = F(temps)

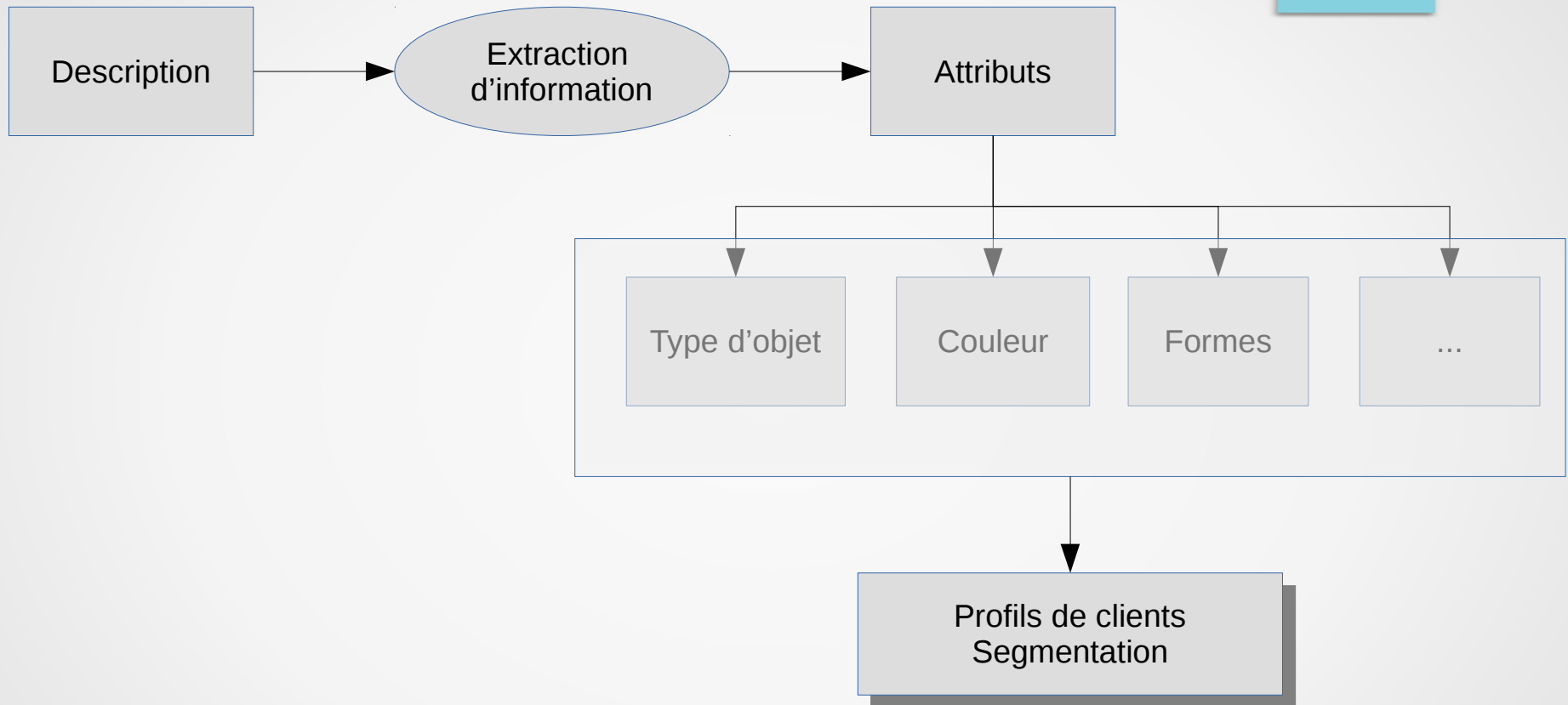




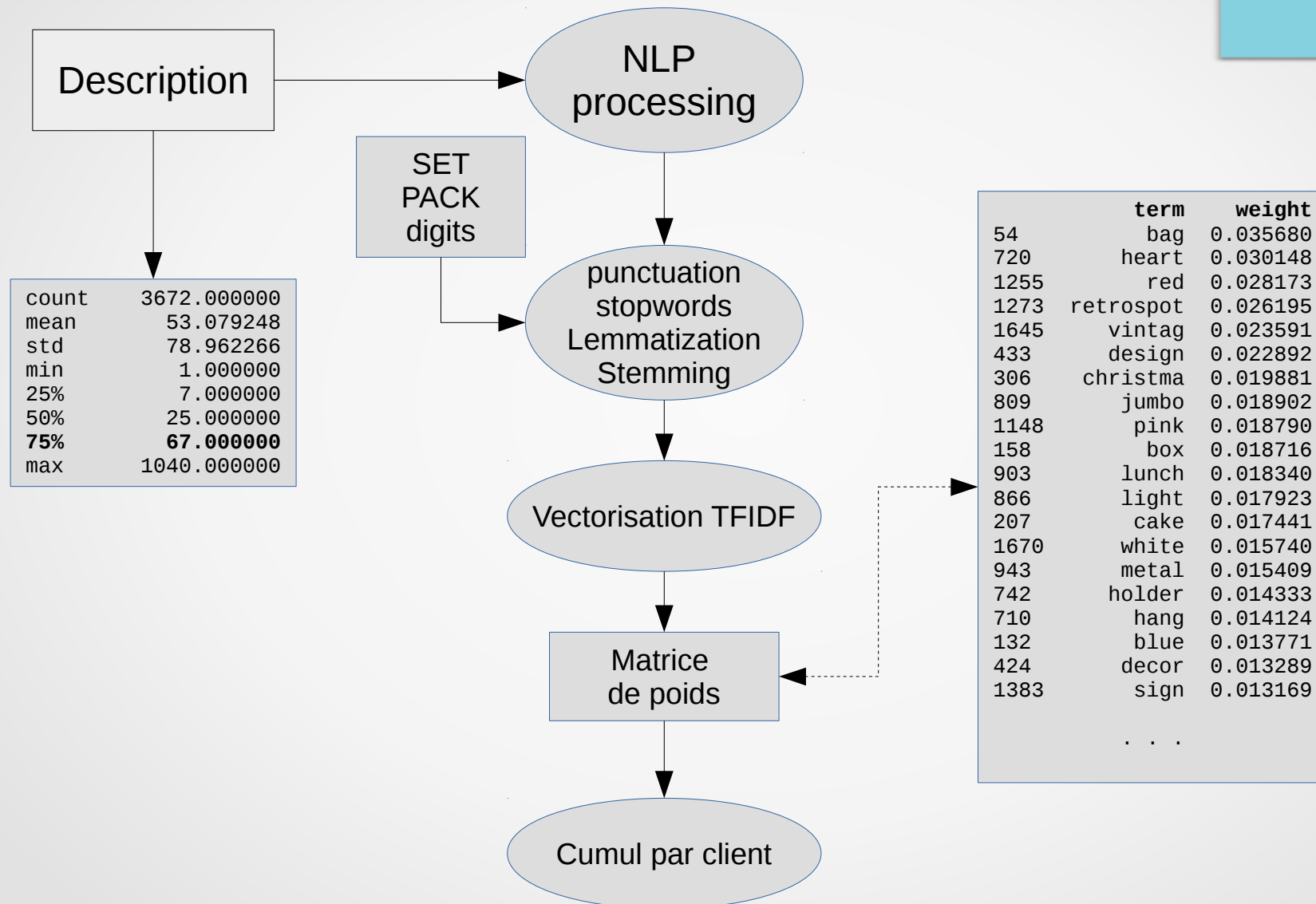
# RFM : Traitement & durée de vie



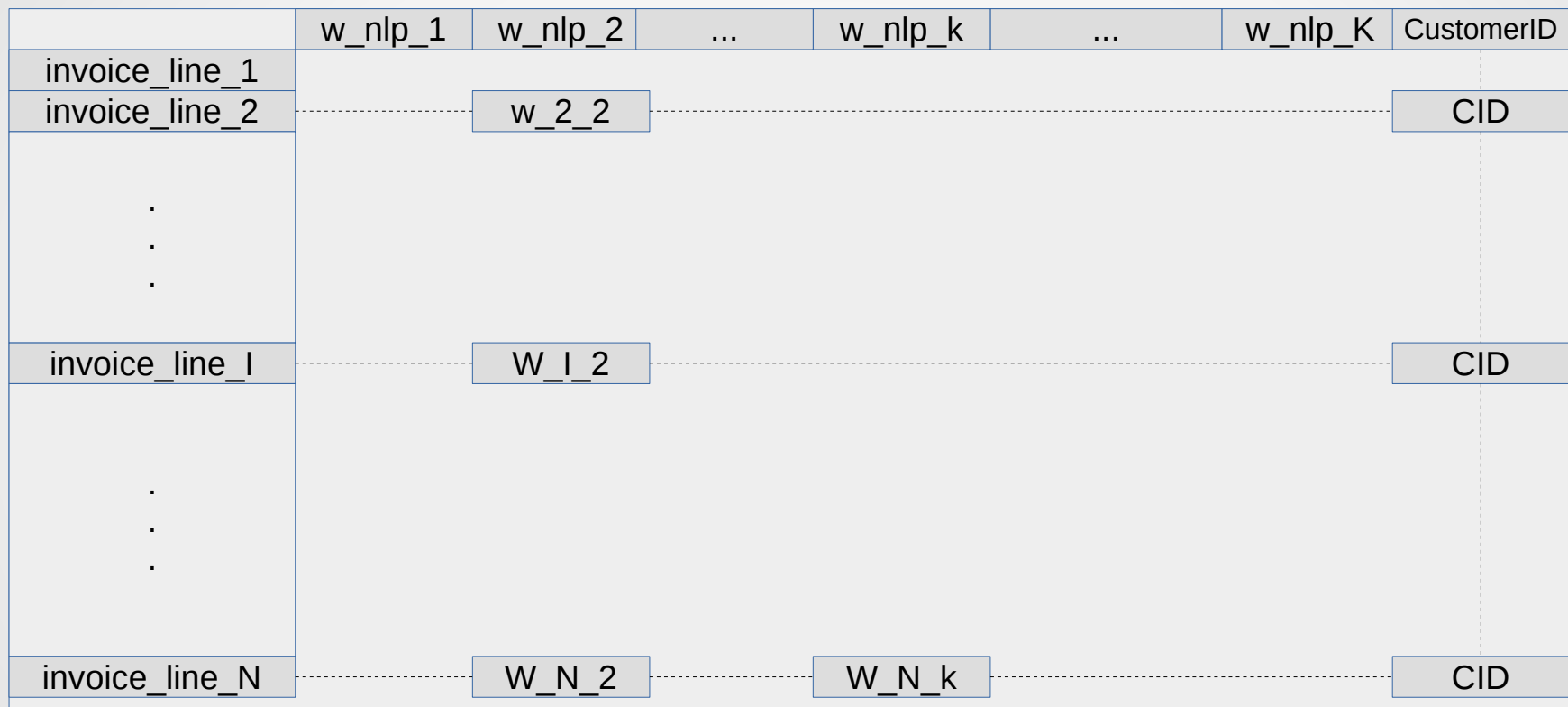
# Analyse exploratoire : Description



# Analyse exploratoire : items



# Traitement NLP: cumul des poids

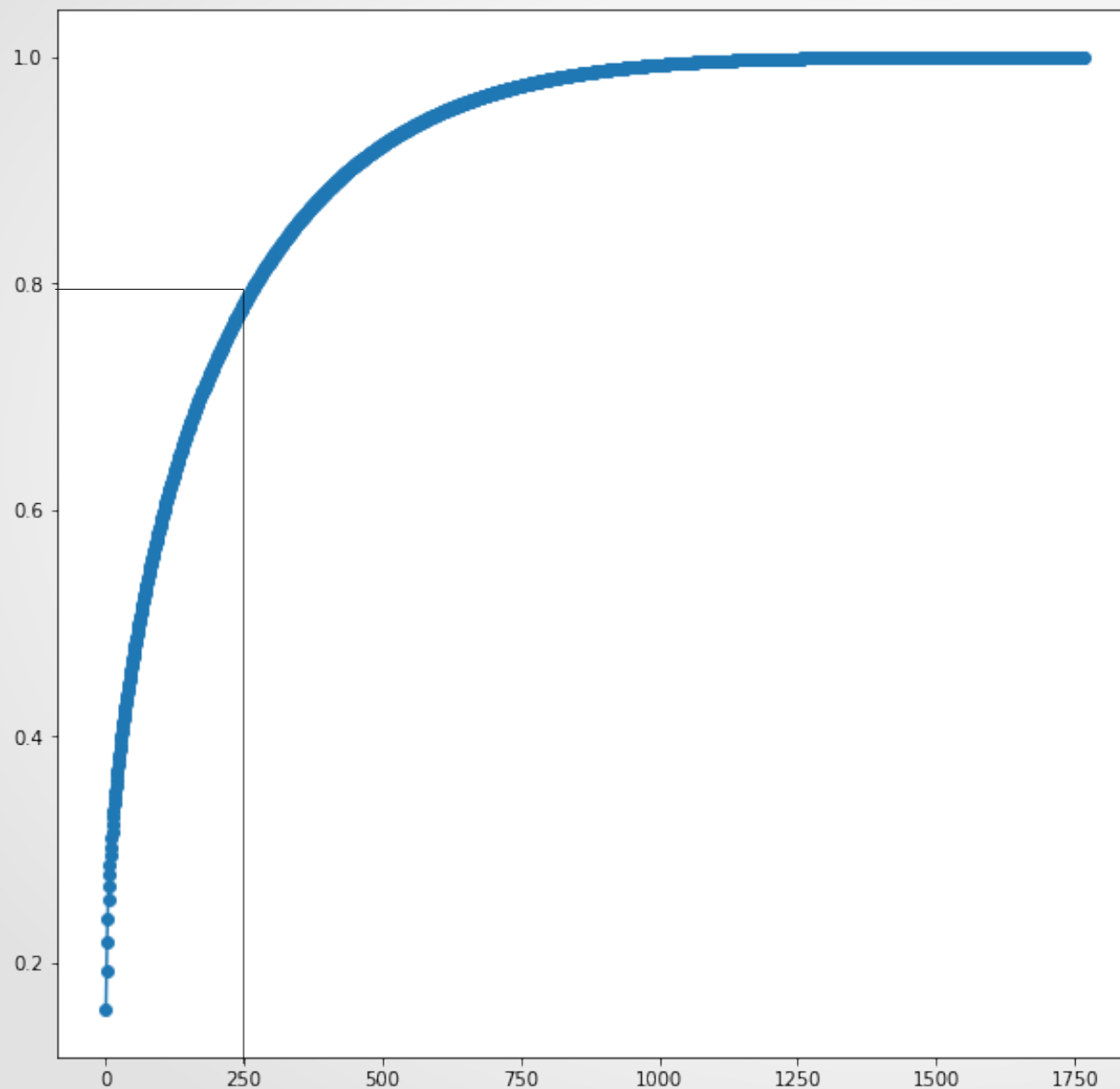


CustomerID

$\text{Sum}(w_{j_2})$

$\text{Sum}(w_{j_k})$

# Variables NLP : réduction de dimension

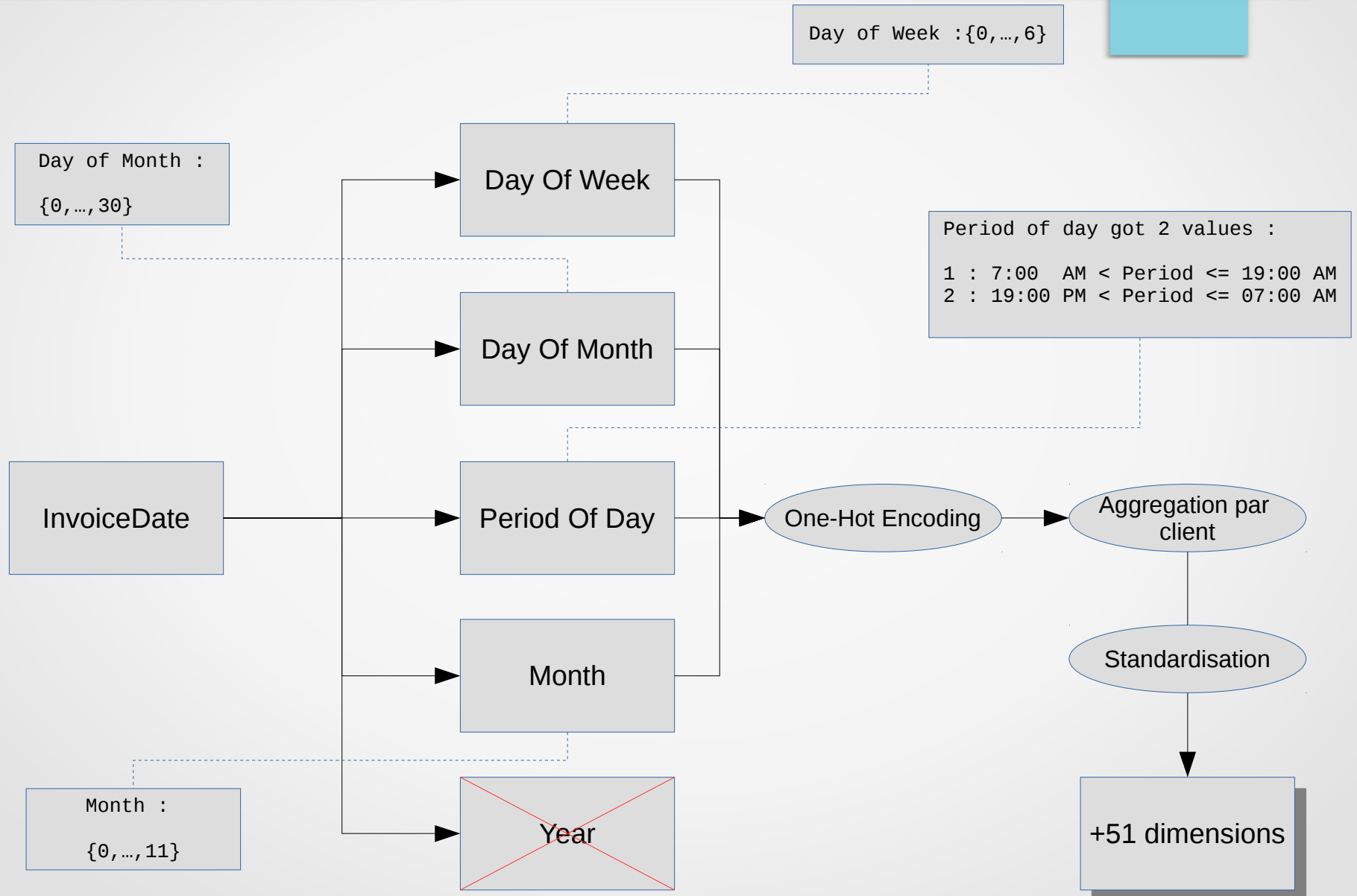


Traitement NLP : 1766 dimensions

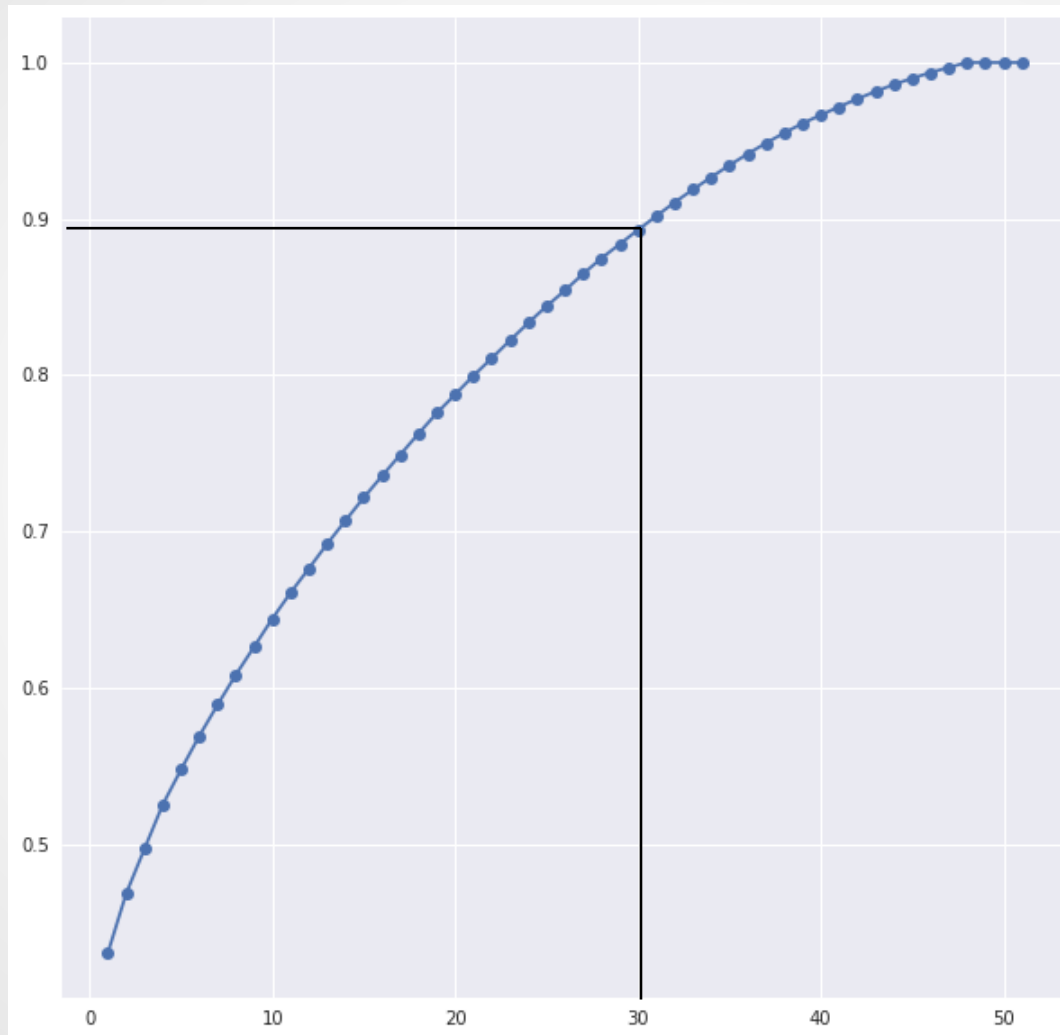
Hypothèse :  
linéarité du phénomène

Nb dimensions :  
250  $\Rightarrow$  80 % variance expliquée

# Time : Nouvelles dimensions (2010,2011)



# Time : réduction de dimension



Nb de dimensions réduites: 30

# Parcours Data scientist : projet 5

## Clustering

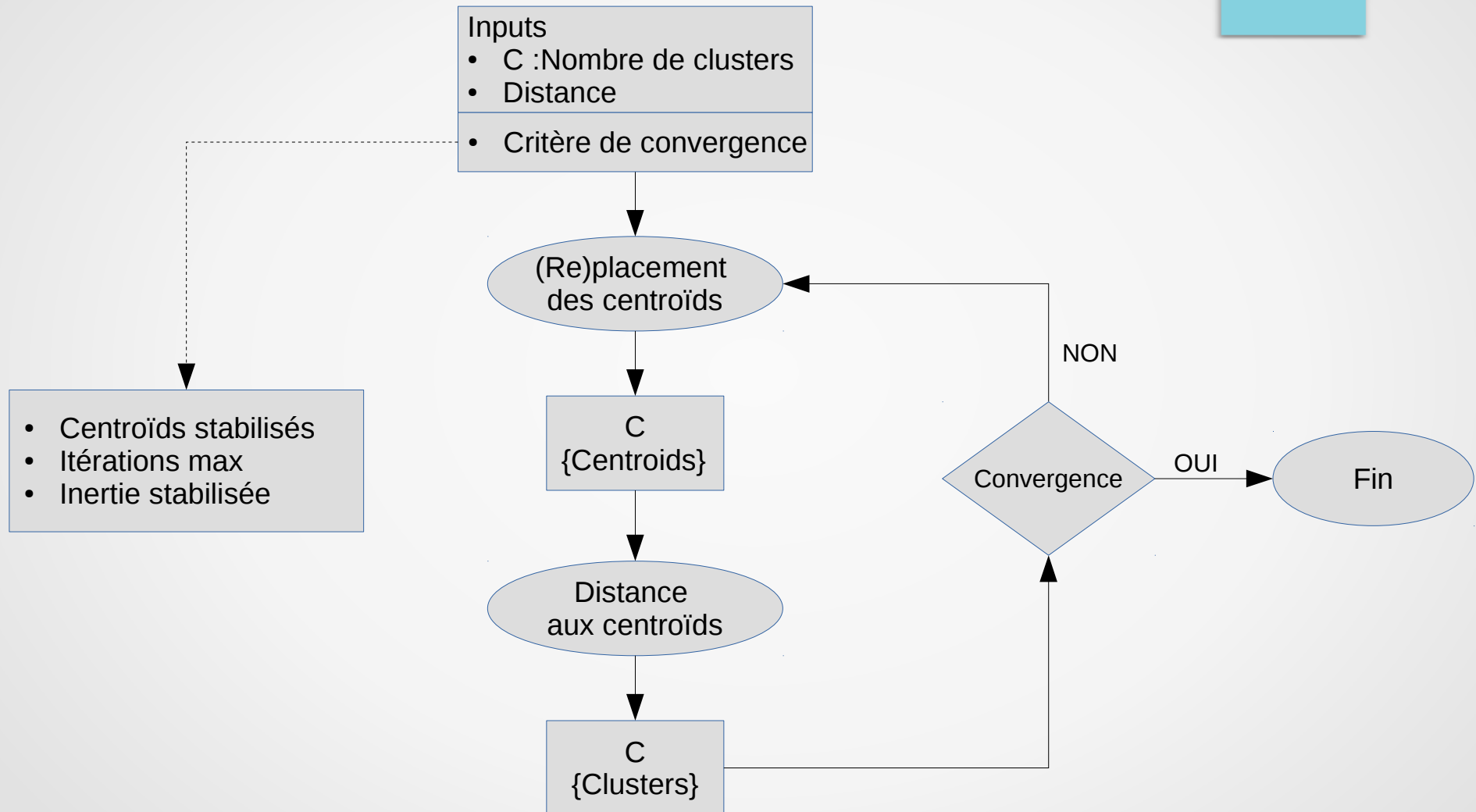
- Kmeans
- GMM

Data model :

- Complexity : 2 124
- Dimensions : 339

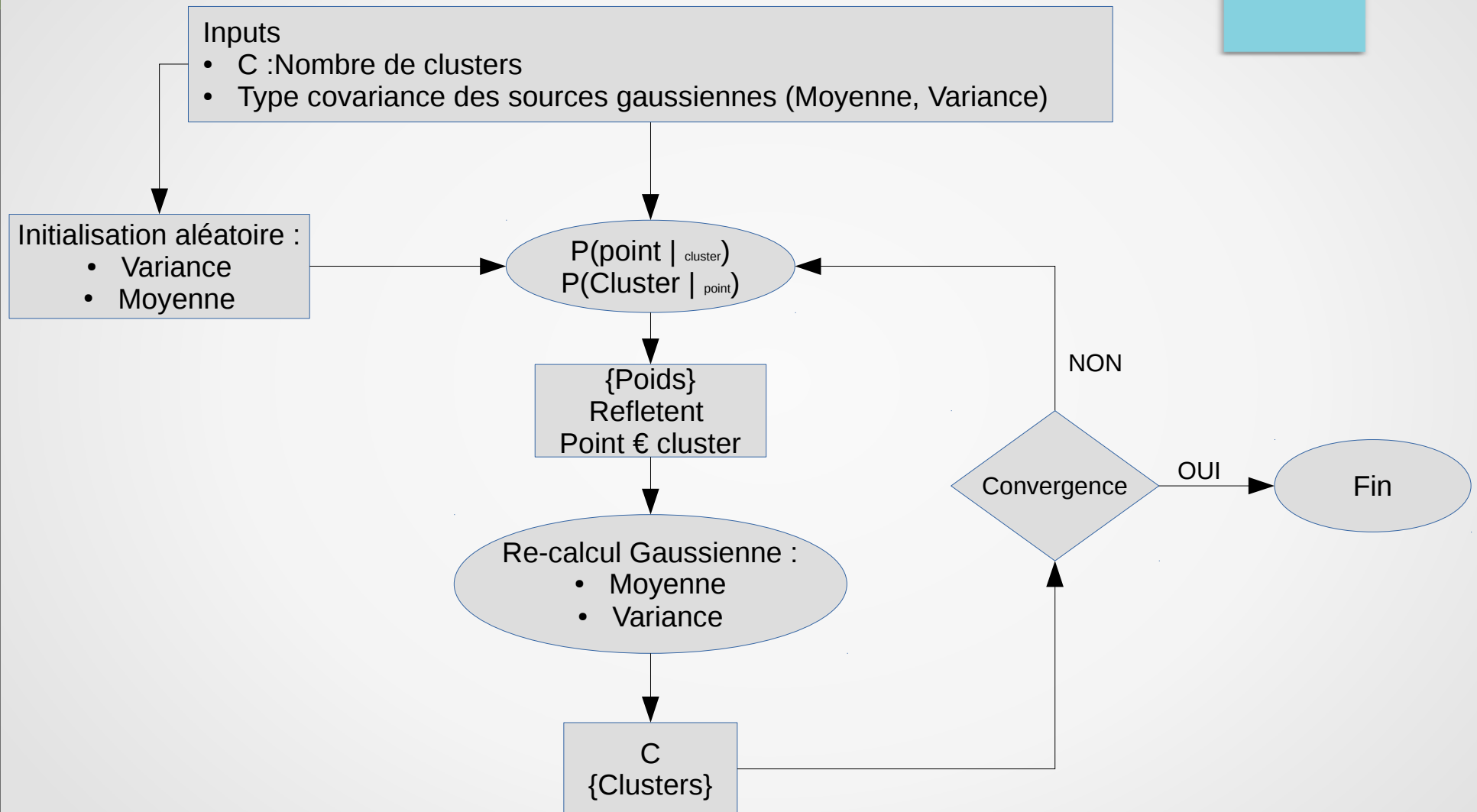


# Kmeans : description

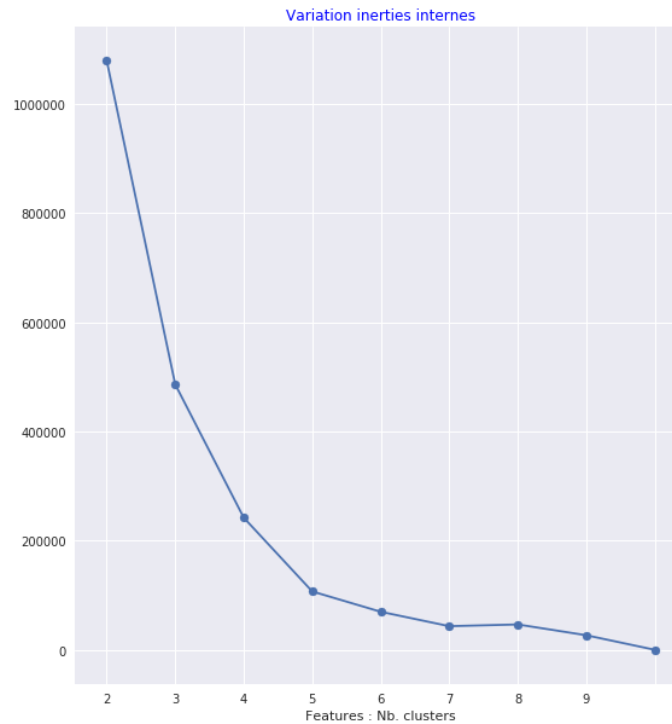
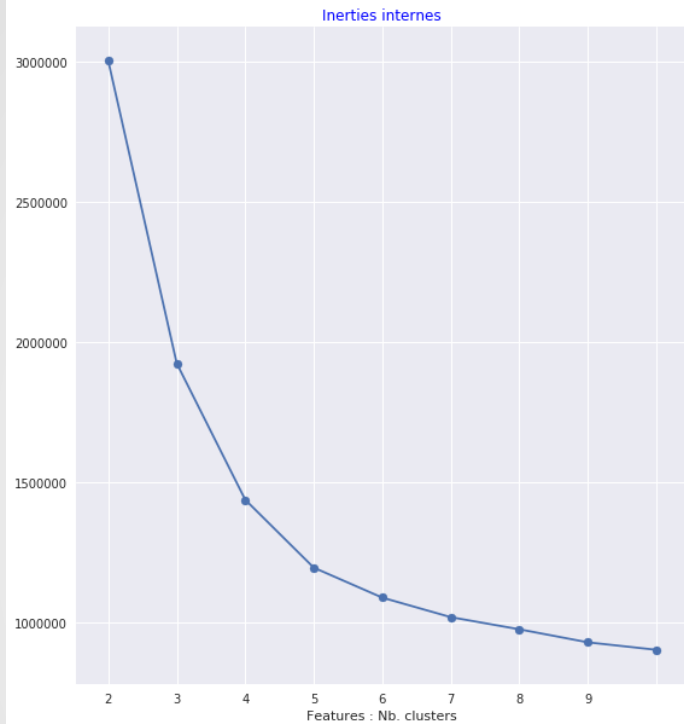


Adapté aux nuages sphériques  
Outliers → biaisent les centroïds

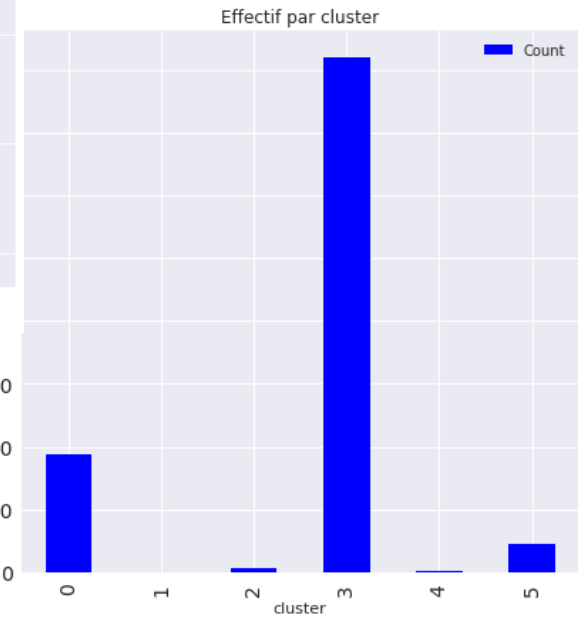
# GMM: description



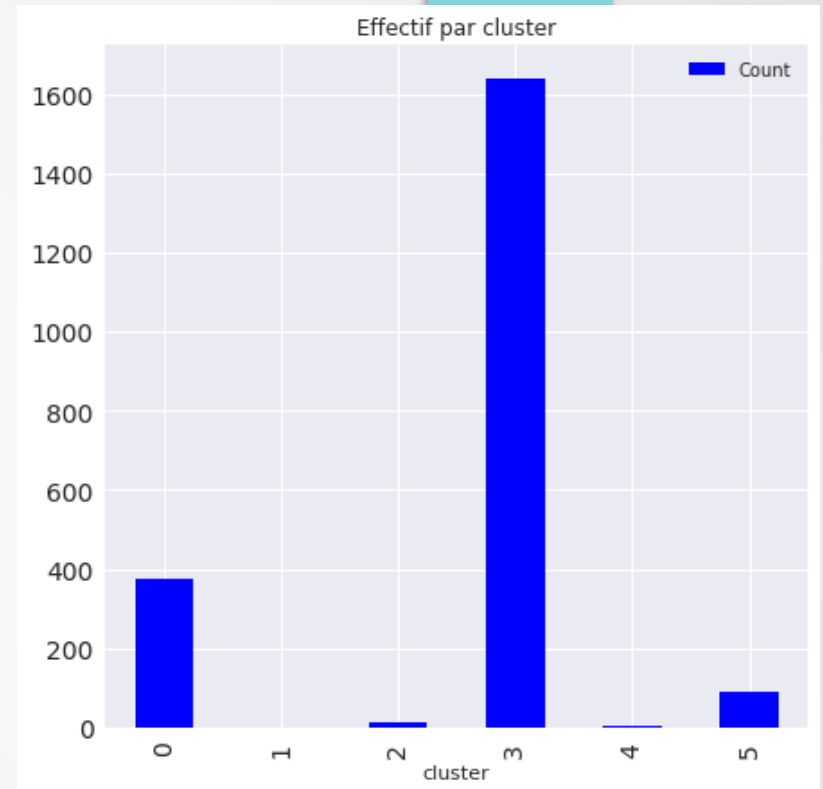
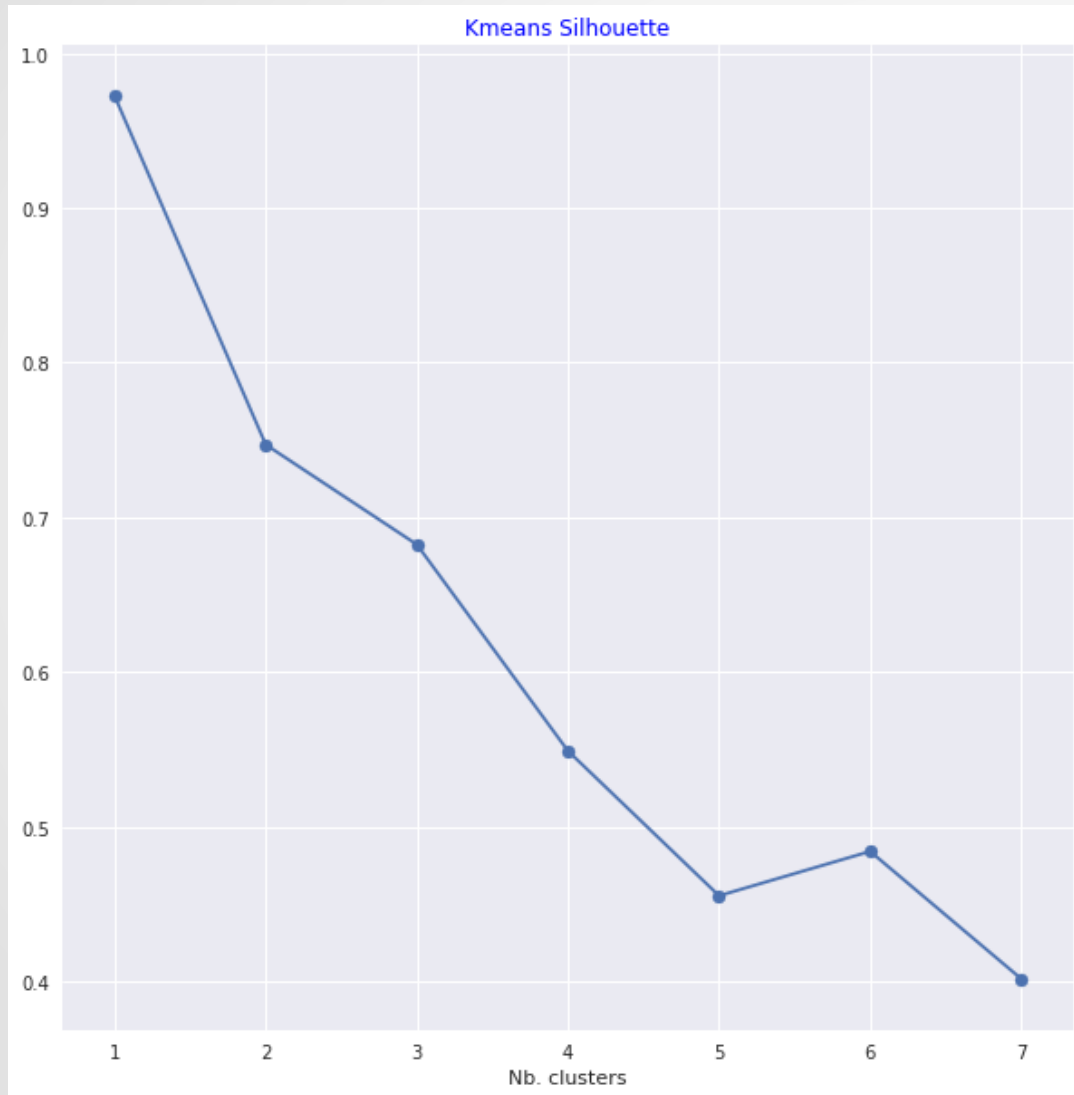
# Kmeans : clustering vs intra-cluster inertias



cluster	Count
0	377
1	1
2	11
3	1641
4	2
5	92



# Kmeans : clustering vs Silhouette



cluster	Count
0	377
1	1
2	11
3	1641
4	2
5	92

# GMM clustering

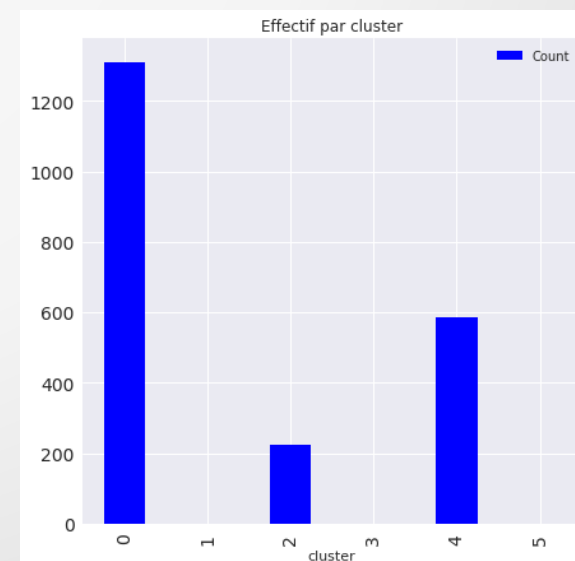


Nb optimal de clusters

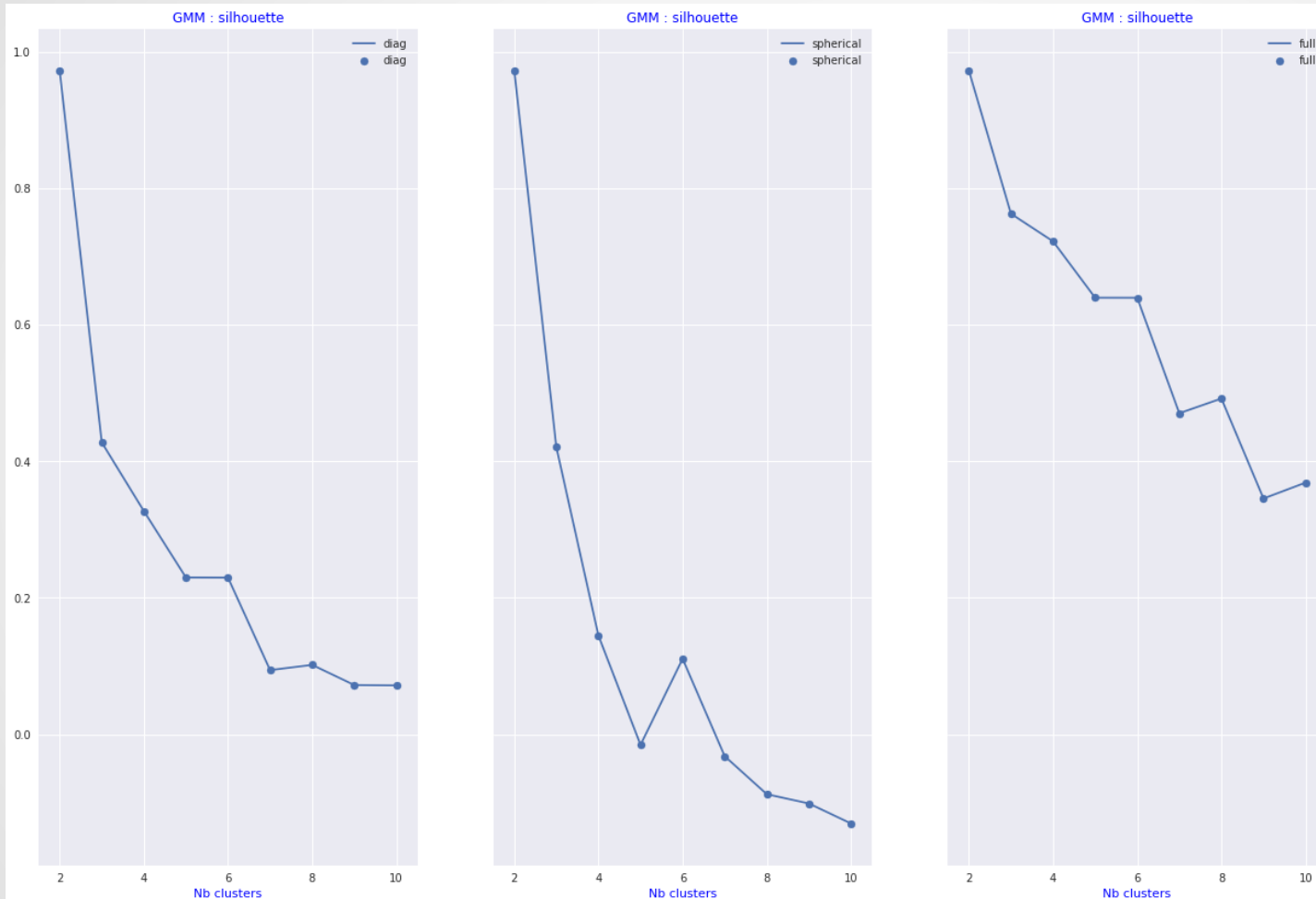
Type de covariance :

- Sphérique : 3 clusters
- Diagonale : 3 clusters

cluster	Count
0	1312
1	1
2	224
3	1
4	585
5	1



# GMM : silhouette vs covariance type



Nb optimal de clusters :

- Diagonale : 3 clusters
- Sphérique : 3 clusters
- Full : 3 clusters

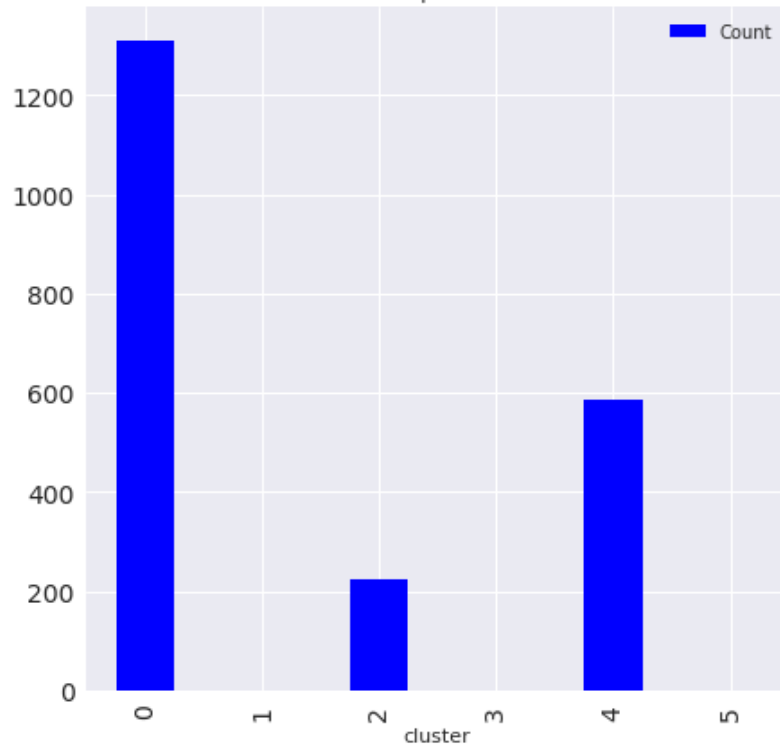
# Clustering : choix de l'algo & hyp. Param.

GMM

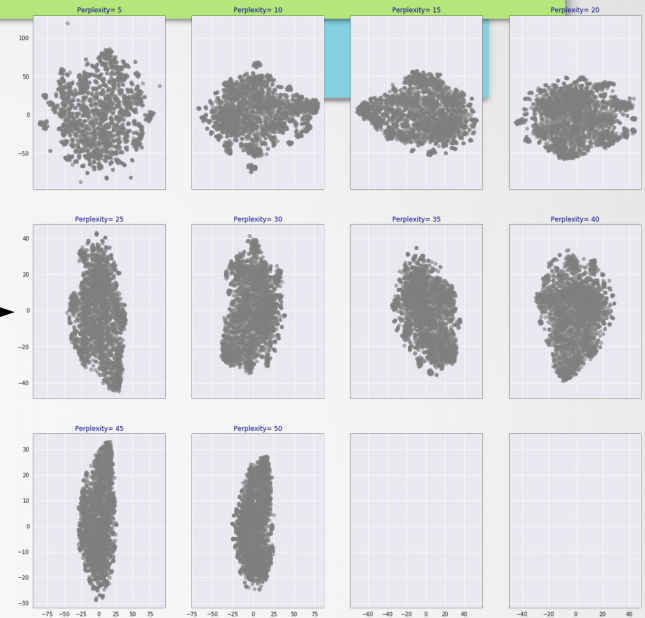
Type de covariance : Diagonale

Nombre de clusters : 6  $\rightarrow$  3 : 0, 2, 4

Effectif par cluster



T-SNE 2D



## Parcours DataScientist : projet 5

# Analyse des segments de marché



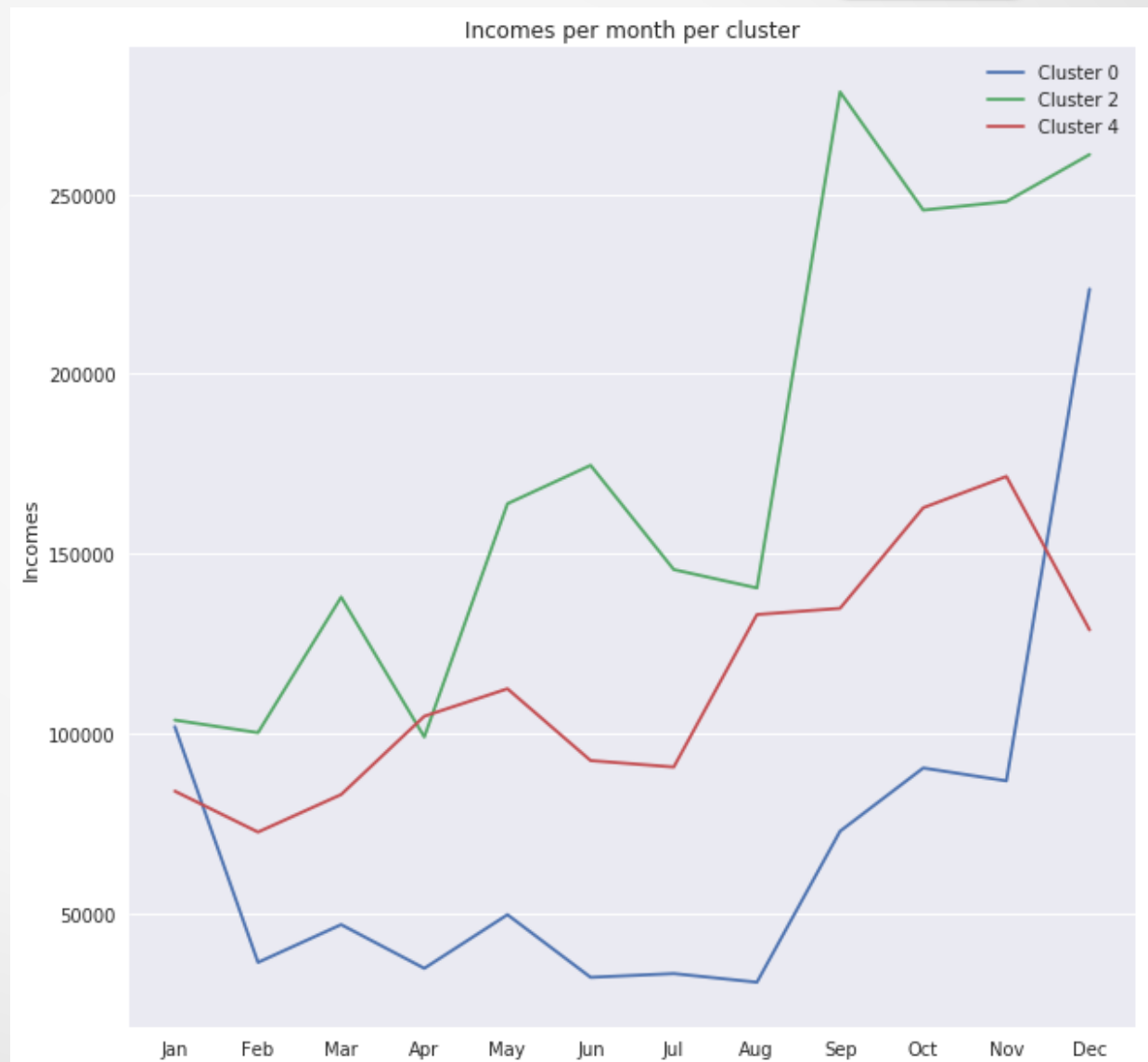
# Market segments : Incomes

cluster	Count	%
0	1312	62
2	224	11
4	585	27

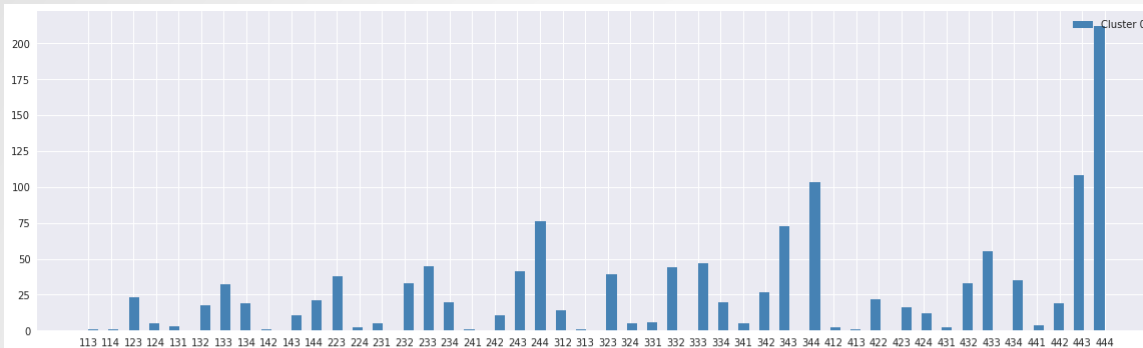
Segments 2 and 4 :

Feb → Jun : increases buy  
Jul → Aug : decrease buy  
Sep → Dec : high buy activity

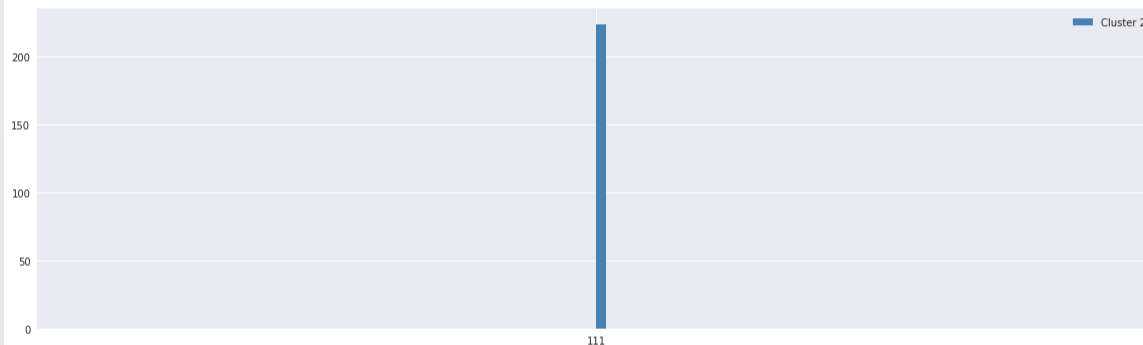
Segment 0 : Q3 increases buy



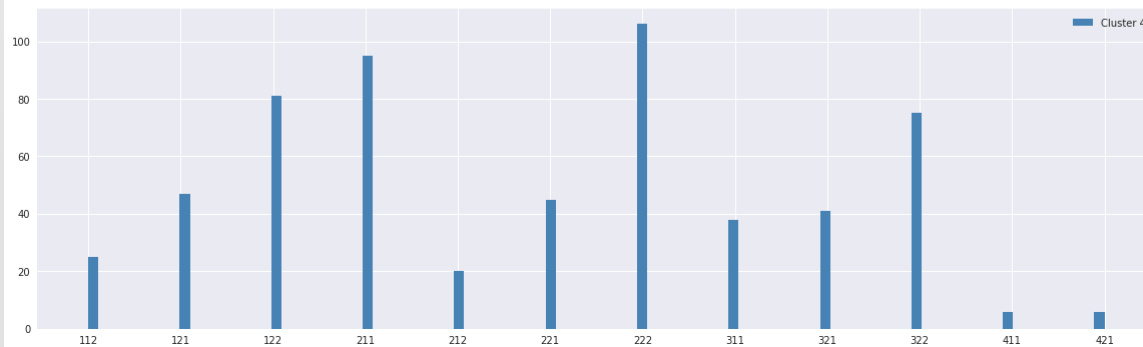
# Markets segments: RFM distribution



Cluster 0 :  
Week moneraty value  
Low fidelity



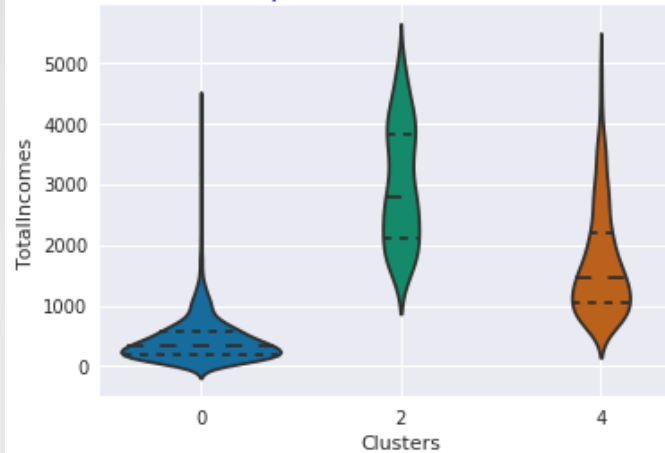
Cluster 2 :  
High moneraty value  
Buy often  
High fidelity



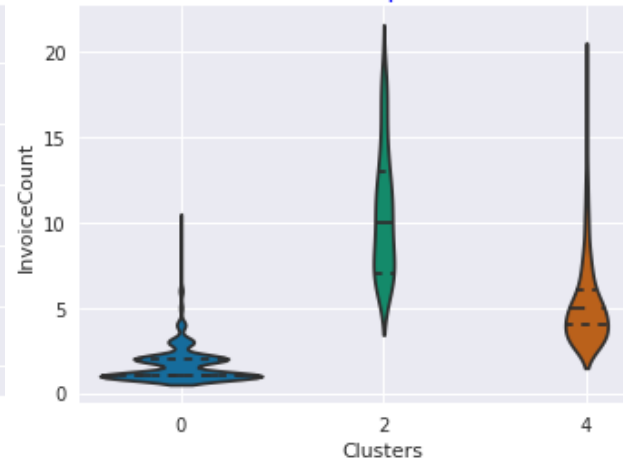
Cluster 4 :  
Middle moneraty value  
Buy often  
Middle fidelity

# Markets segments: behaviours (1)

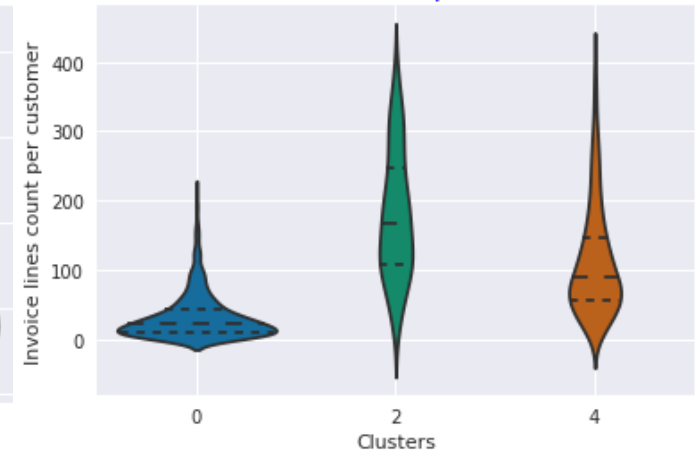
Incomes per customer for each cluster



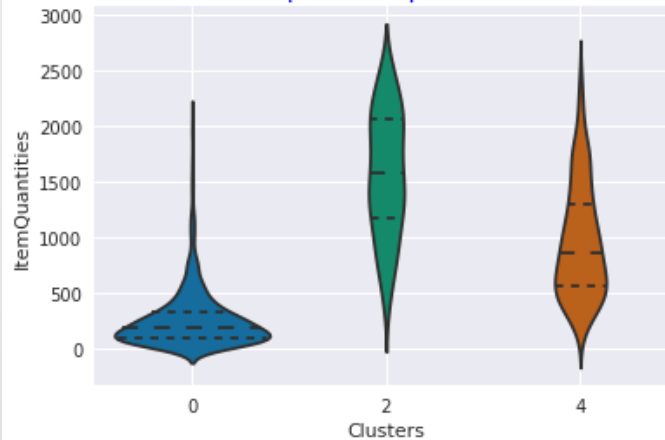
Invoice count per customer



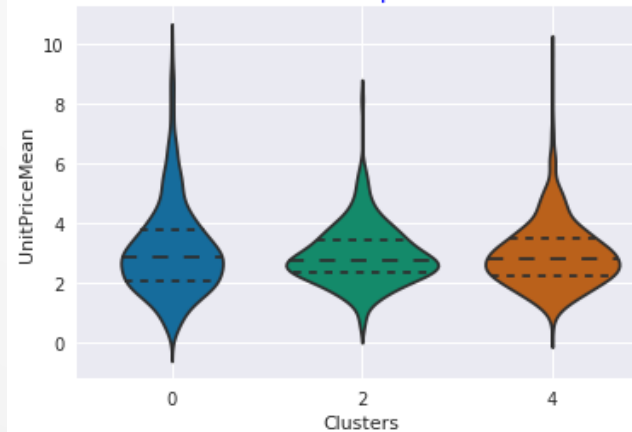
Invoice lines count per customer



Items quantities per customer



Mean Unit Price per customer



Clusters	Unique items	% items
0	3187	91 %
2	3098	88 %
4	3134	90 %

## Cluster 0 :

- Few invoices
- Few items per invoice

## Cluster 2 :

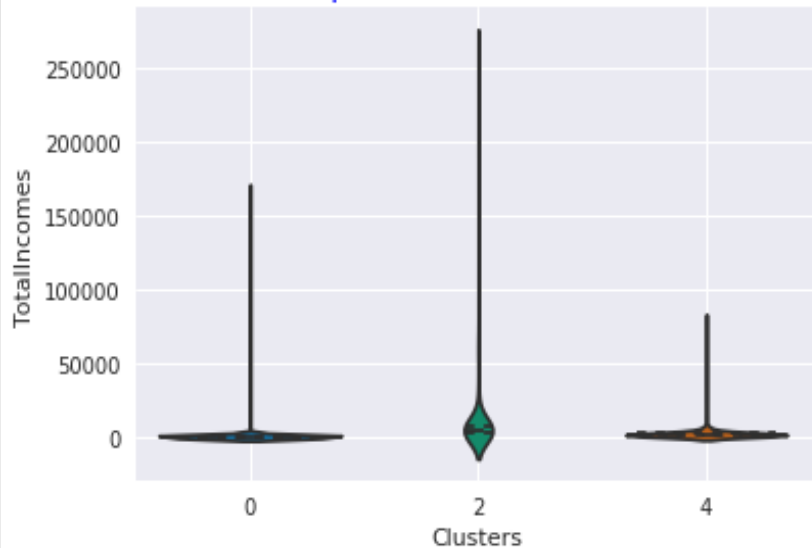
- Lot of invoices
- Lot of items per invoice

## Cluster 4 :

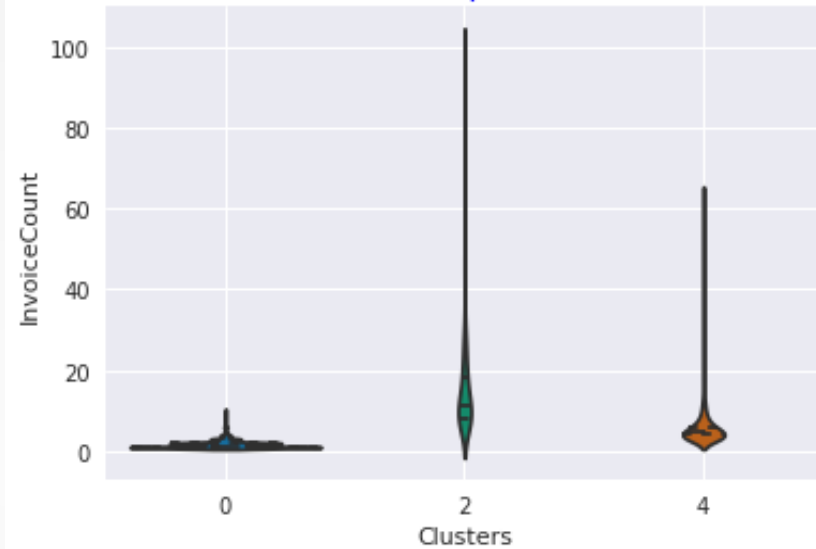
- Few invoices
- Few items per invoice

# Markets segments: behaviours (2)

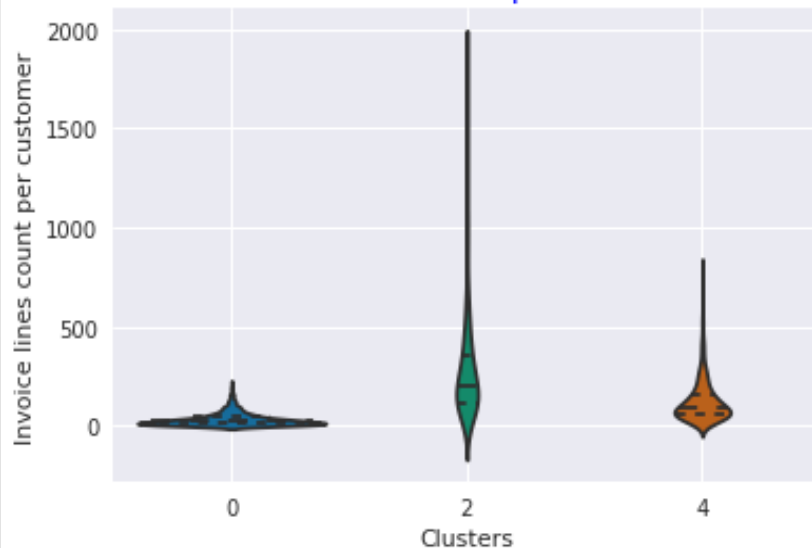
Incomes per customer for each cluster



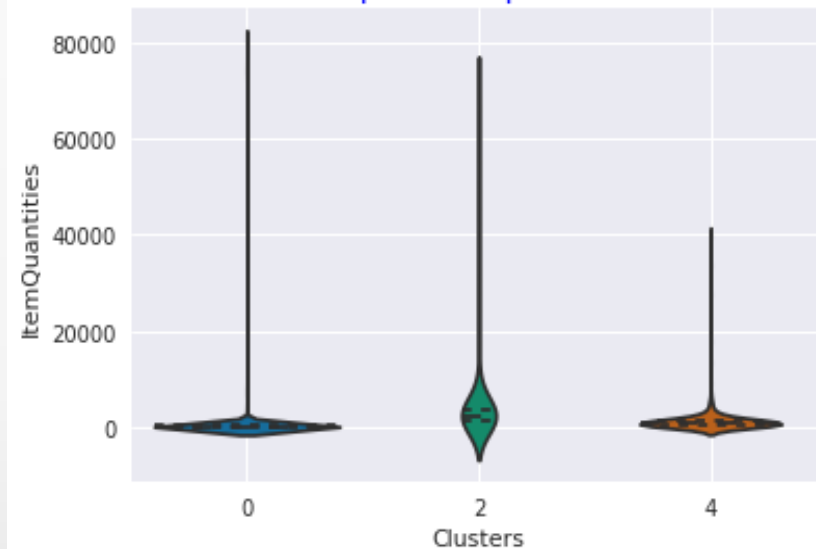
Invoice count per customer



Invoice lines count per customer



Items quantities per customer



# Parcours DataScientist : projet 5

## Prediction d'appartenance

- Random Forest
- SVC

# Random Forests: précision

Nombre de répétitions : 10  
Variation d'estimateurs: 1 à 20

**18 Estimateurs**

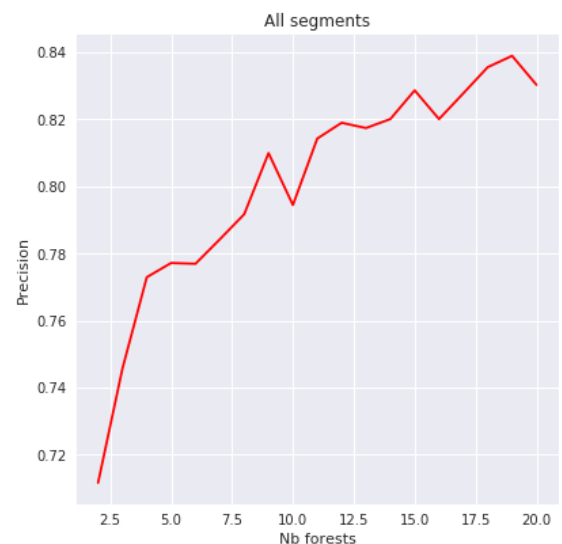
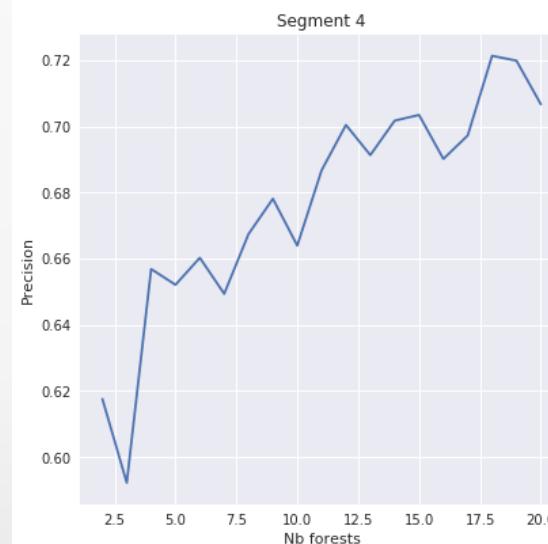
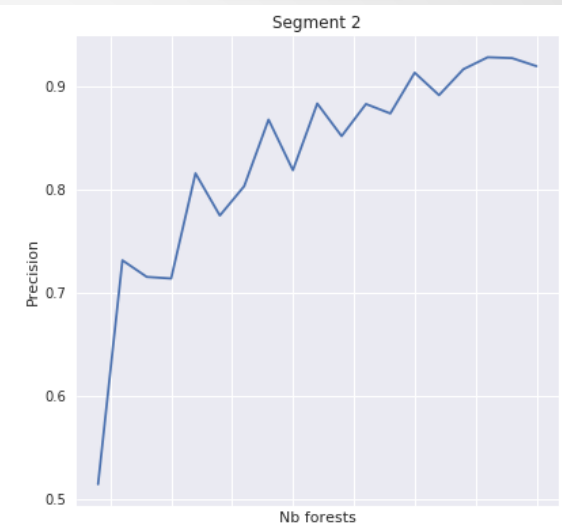
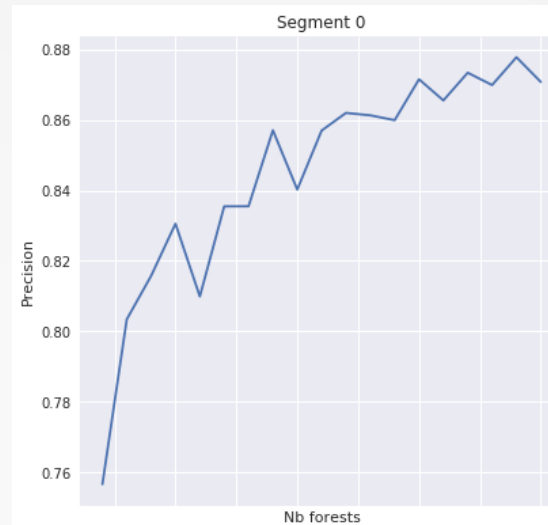
Répétitions : 10

Global accuracy = 0.84

Segment : 0 / : 0.87

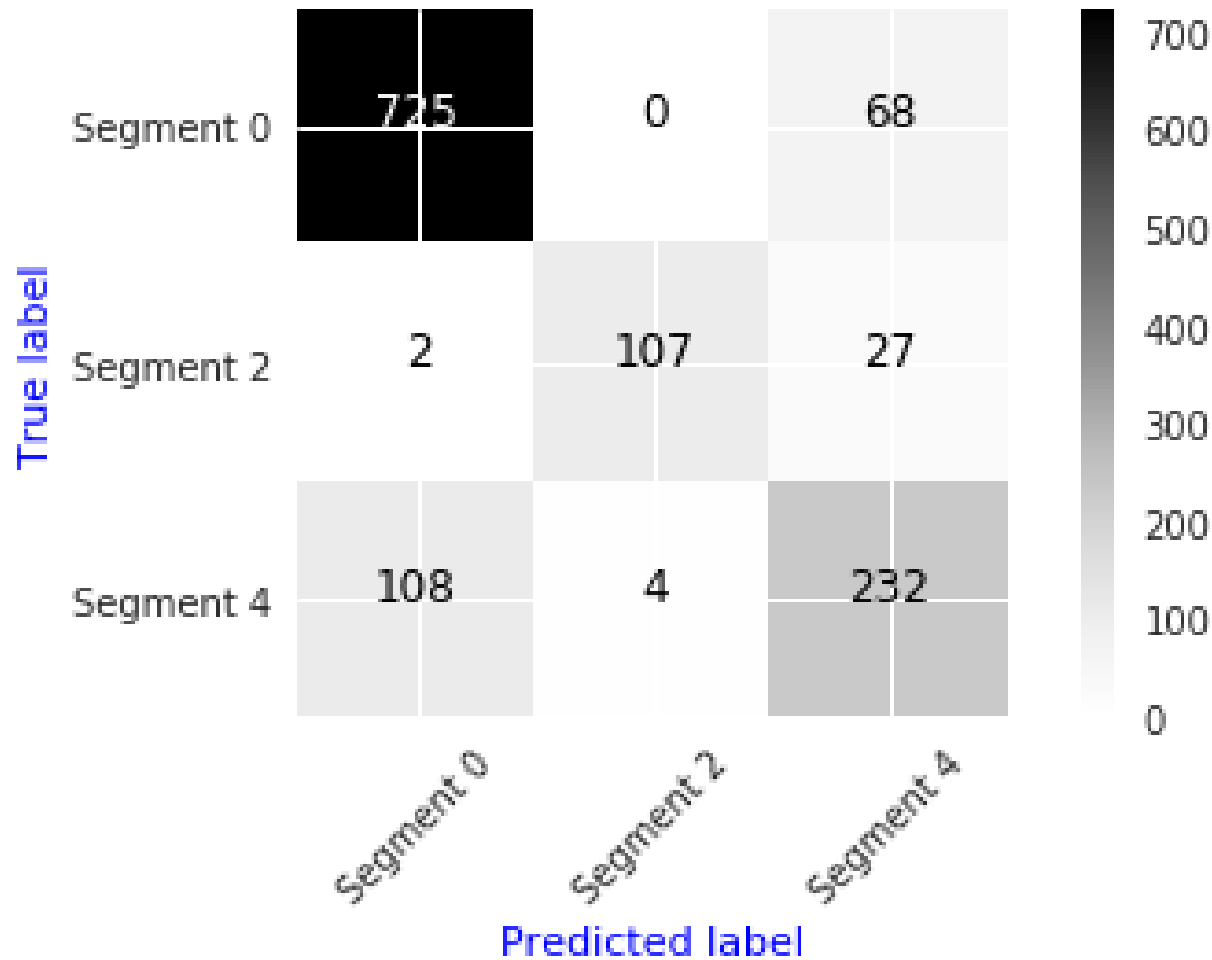
Segment : 2 / : 0.93

Segment : 4 / : 0.72



# Random Forests: matrice de confusion

Confusion matrix for RF predictions



	Segment 0	Segment 2	Segment 4
Recall	87 %	96 %	70 %
Précision	91 %	79 %	67 %
Spécificité	84 %	97 %	88 %
F-mesure	89 %	86 %	69 %

# Estimateur: SVC

Noyaux :

- Linéaire
- RBF
- Poly
- Sigmoid

SVC : One vs Rest

**kernel= linear**

**Accuracy / segment : {0: 0.88, 2: 0.73, 4:0.70}**

**Global accuracy : 0.82**

**kernel= rbf**

Accuracy / segment : {0: 0.83, 2: 0.73, 4 : 0.61}

Global accuracy : 0.77

**kernel= poly**

Accuracy / segment : {0: 0.70, 2: 0.66, 4 : 0.54}

Global accuracy : 0.68

**kernel= sigmoid**

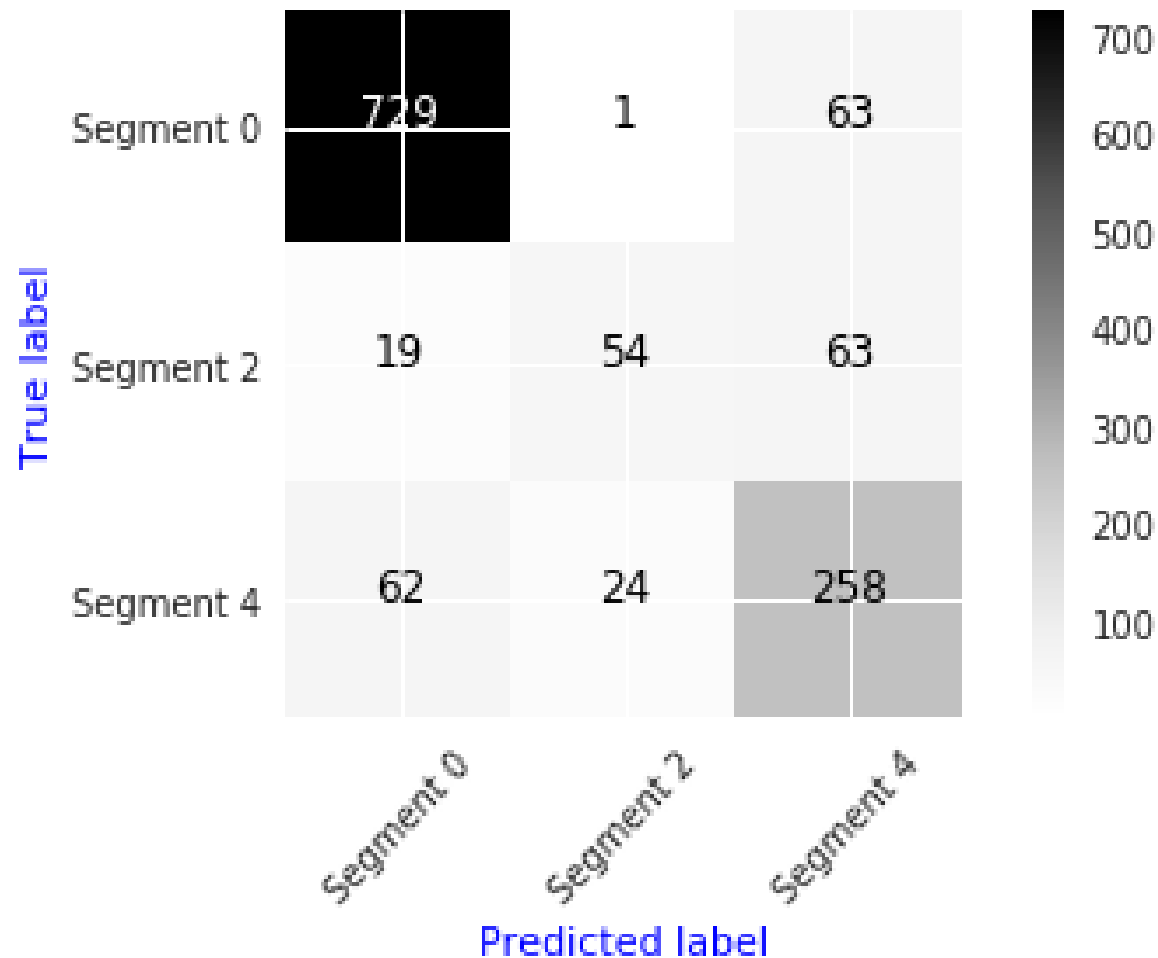
Accuracy / segment : {0: 0.79, 2: 0.27, 4:0.61}

Global accuracy : 0.66

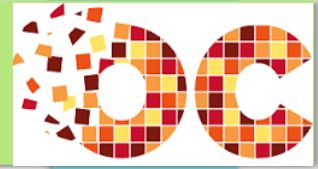


# SVC: matrice de confusion

Confusion matrix for SVC predictions, 60% tests

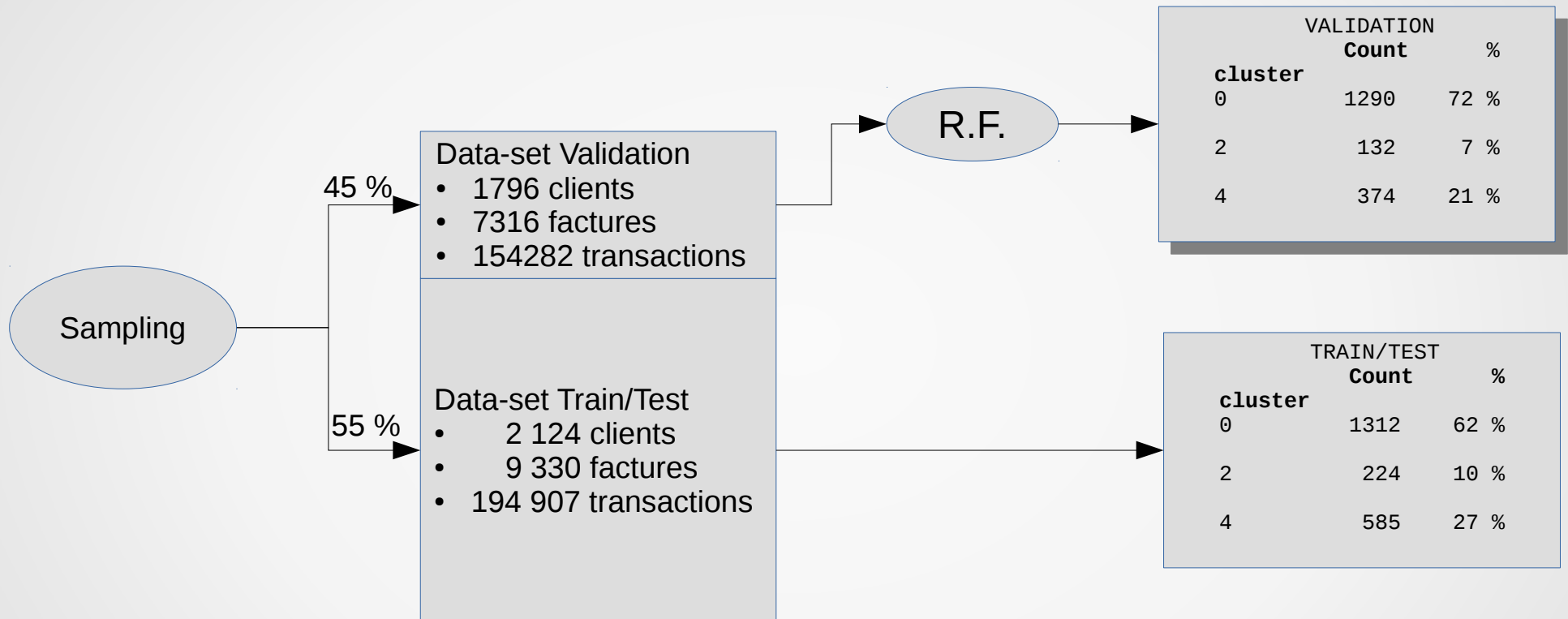


	Segment 0	Segment 2	Segment 4
Recall	90 %	68 %	75 %
Précision	92 %	39 %	67 %
Spécificité	86 %	93 %	86 %
F-mesure	91 %	50 %	71 %



## Validation

# Validation : data-set prediction



# Validation : Segment 2 vs RFM = 111

CustomerID	Frequency	Recency	Monatary	neg_recency	R_score	F_score	M_score	RFM
12839	14	2	5591.42	-2	1	1	1	111
...								

get\_market\_segment()

RFM threshold matrix

	Frequency	Recency	Monatary	neg_recency
Q1	1.0	16.0	296.170	-140.0
Q2	2.0	49.0	654.025	-49.0
Q3	5.0	140.0	1552.130	-16.0

0: [13013]

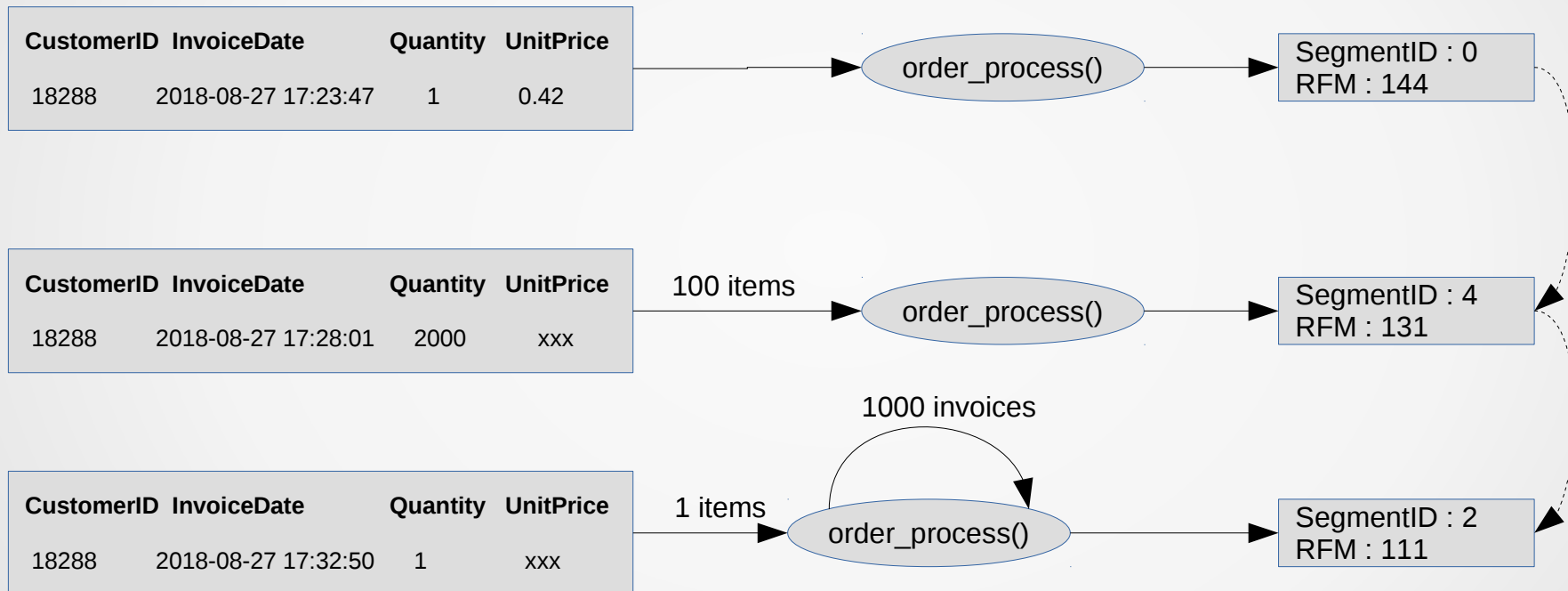
2: [12839, 12856, 12901, 12921, 12971, 13050, 13069, 13081]

4: [12935, 13012, 13090, 13126]

Segment 0 : 1  
Segment 2 : 8  
Segment 4 : 4

Segment 2  
Précision :  $8/13 = 60\%$

# Validation : Market segment path



# Conclusions

- Problème linéaire
- Traitement NLP : nécessite d'importantes ressources
- Améliorations / évolutions:
  - Re-calcul de la matrice de seuils RFM
  - Intégration en environnement BIG DATA
    - Traitement d'une facture > 100 transactions
    - Traitement NLP sur nb. descriptions > 4000
    - Traitement simultané de factures

# Parcours Datascientist : projet 5

## Annexes

Annexe 1 : fichiers du projet

Annexe 2 : organisation et processus de l'étude

Annexe 3 : Variables issues du score RFM

Annexe 4 : variables issues du traitement NLP

Annexe 5 : variables issues de la date de facturation

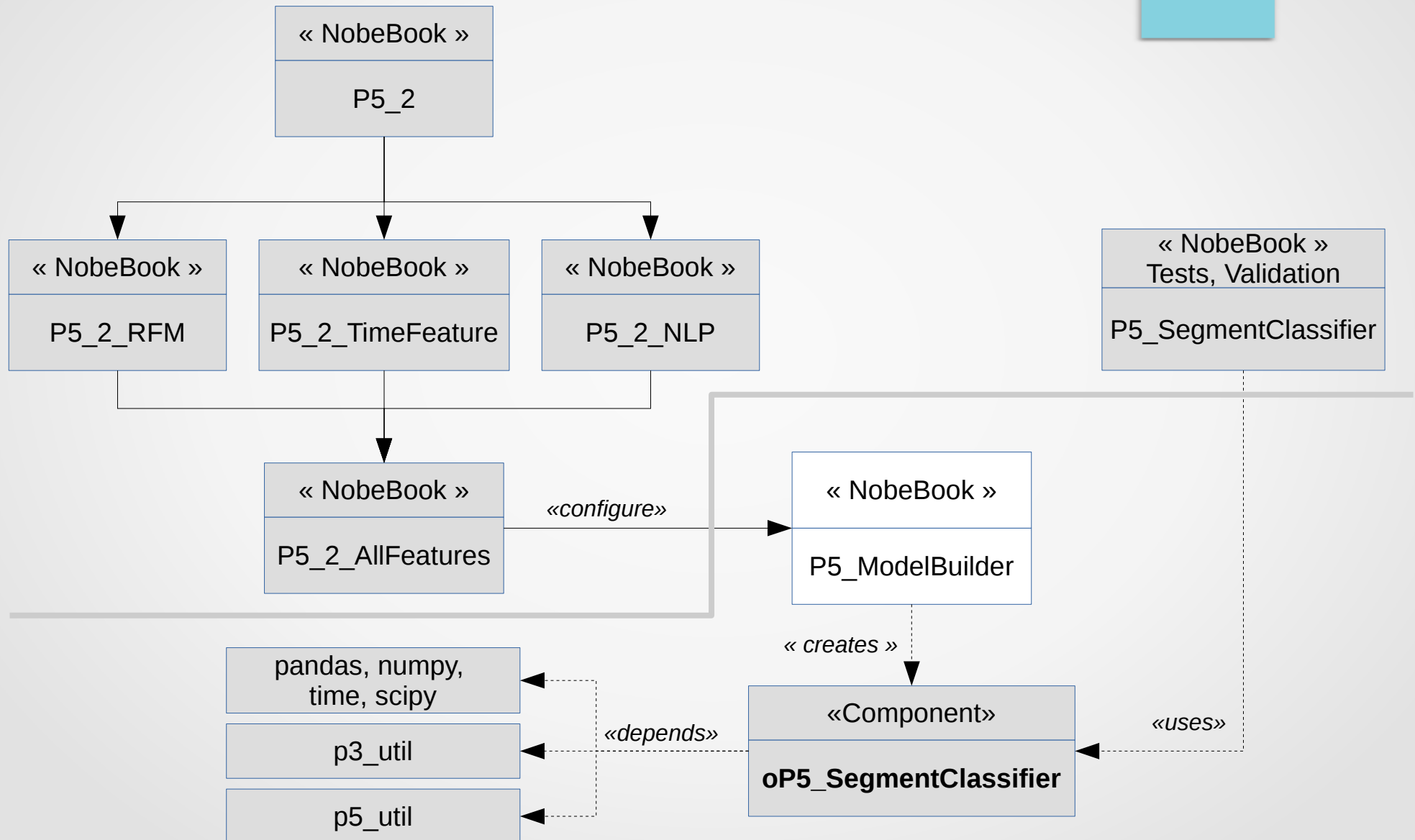
Annexe 6 : API WEB

# Annexe 1 : fichiers du projet

- **Fichiers source python :**
  - p3\_util\_plot.py : utilitaires d'affichage issus du projet P3
  - p3\_util.ppy : utilitaires du projet P3
  - p5\_util\_plot.py : utilitaires d'affichage issus du projet P5 (projet courant)
  - p5\_util.py : utilitaires issus du projet P5 (projet courant)
  - P5\_ModelBuilder.py : générateur de modèle de prédiction
  - P5\_SegmentClassifier.py : implémentation du modèle de prédiction
- **Notebooks de l'analyse exploratoire :**
  - P5\_2.ipynb : nettoyage / exploration
  - P5\_2\_RFM.ipynb : analyse des features dérivées du score RFM
  - P5\_2\_timeFeature.ipynb : analyse des features dérivées de la date de facturation
  - P5\_2\_NLP.ipynb : analyse des features dérivées de la description traitées en NLP
- **Notebook des approches de modélisation :**
  - P5\_2\_AllFeature.ipynb : algorithmes de M.L. non supervisés et supervisés.
- **Notebook de test / validation**
  - P5\_SegmentClassifier.ipynb
- **Rapport sous forme de présentation pdf:**
  - Openclassrooms\_ParcoursDatascientist\_P5.pdf
- **Points d'entrée de l'API :**
  - Pour récupérer une liste de vols :
    - <https://francois-bangui-oc-p4.herokuapp.com/predictor/?>\*
  - Pour récupérer l'évaluation du retard d'un vol à partir de son identifiant :
    - [https://francois-bangui-oc-p4.herokuapp.com/predictor/?flight\\_id=<ID>](https://francois-bangui-oc-p4.herokuapp.com/predictor/?flight_id=<ID>)
- [Pour récupérer un el](#)
- <http://localhost:5000/?customerID=12822>



# Annexe 2 : artefacts et processus d'étude

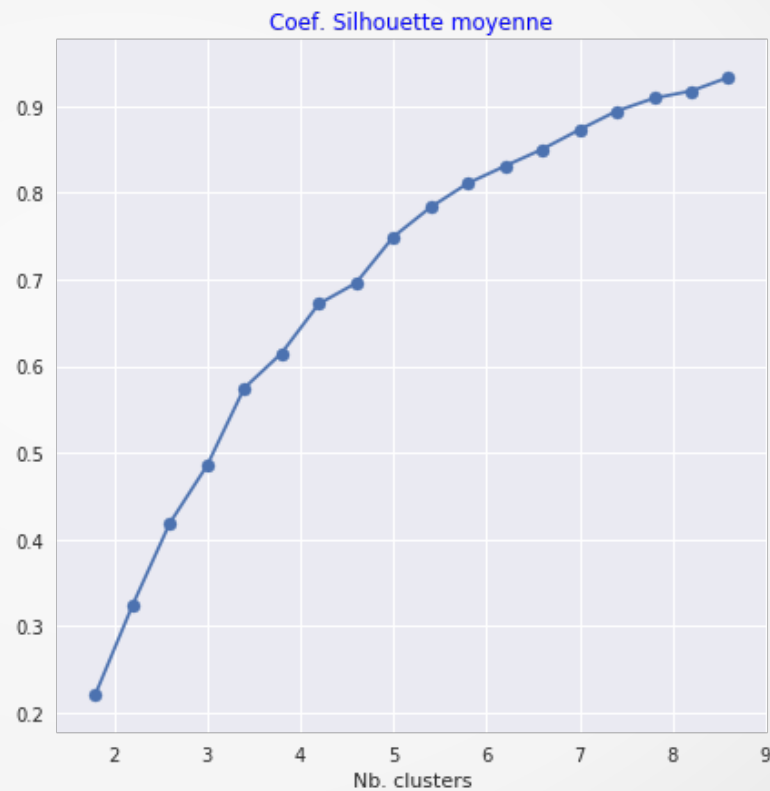
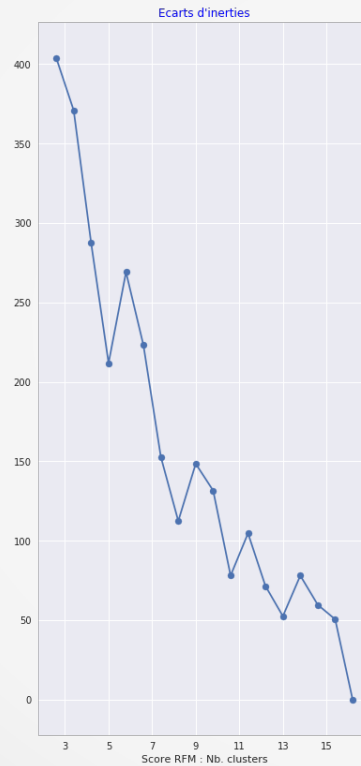
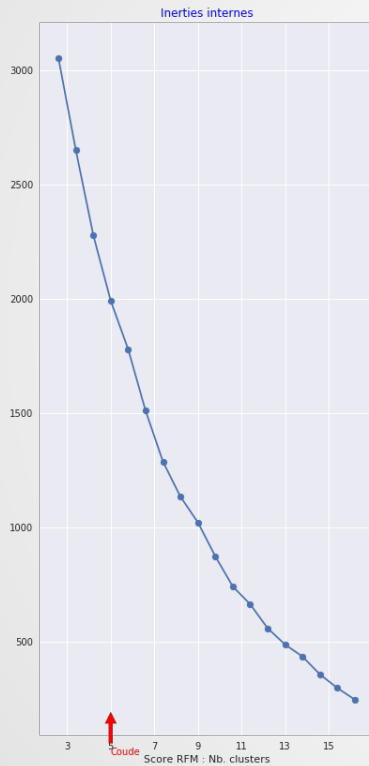


## Annexe 3 : Étude RFM

# RFM : Kmeans Clustering

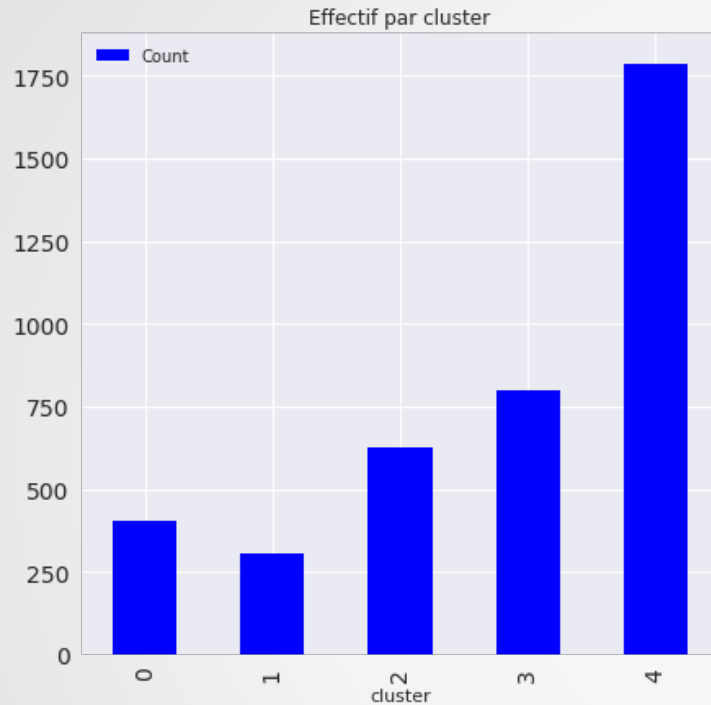
Encodage RFM

- 2373 clients
- 58 dimensions



Nb optimal de clusters : 5

# Clustering RFM : Kmeans effectifs par cluster



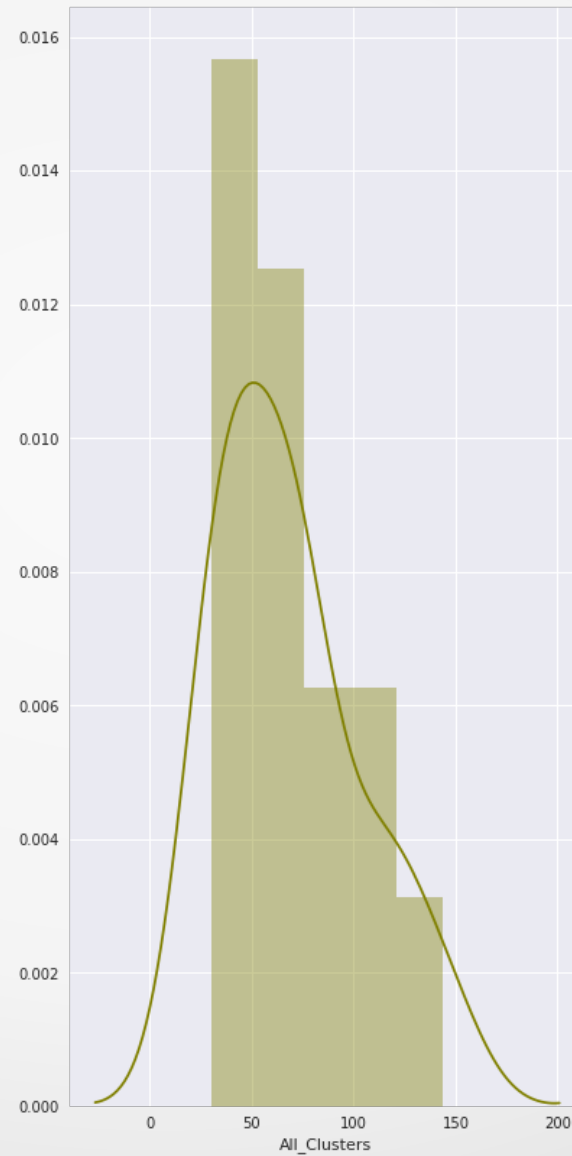
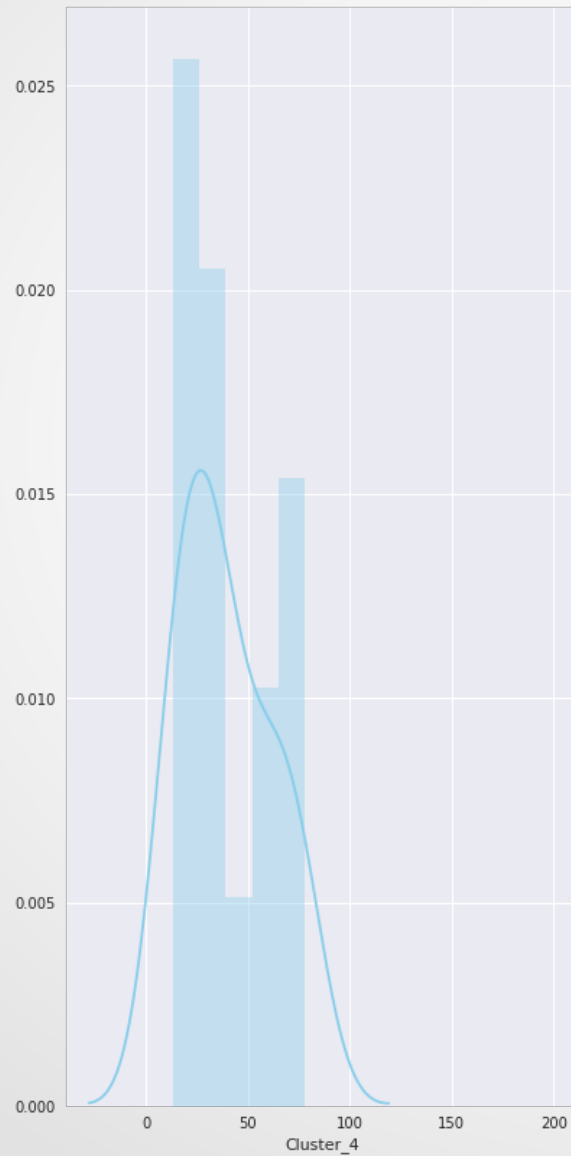
- Clusters : > 300
- Exclusion mutuelle

Clusters distincts :

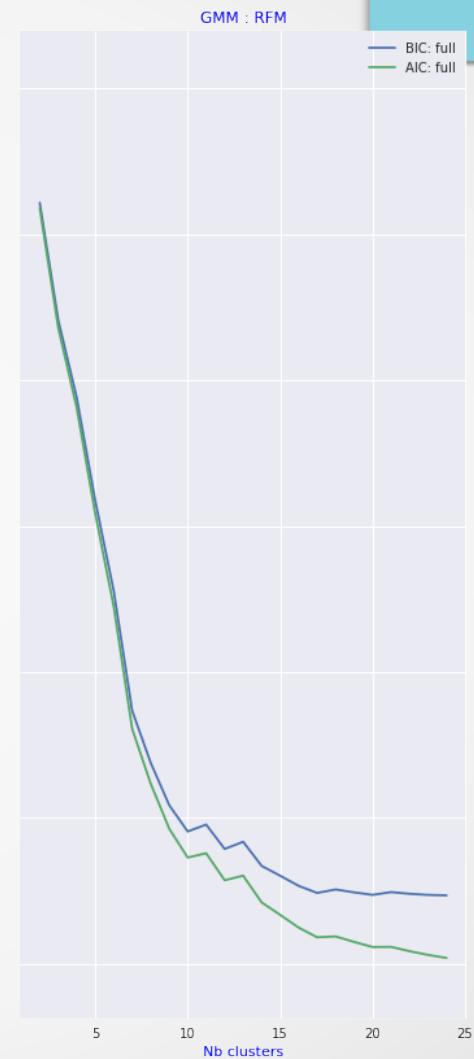
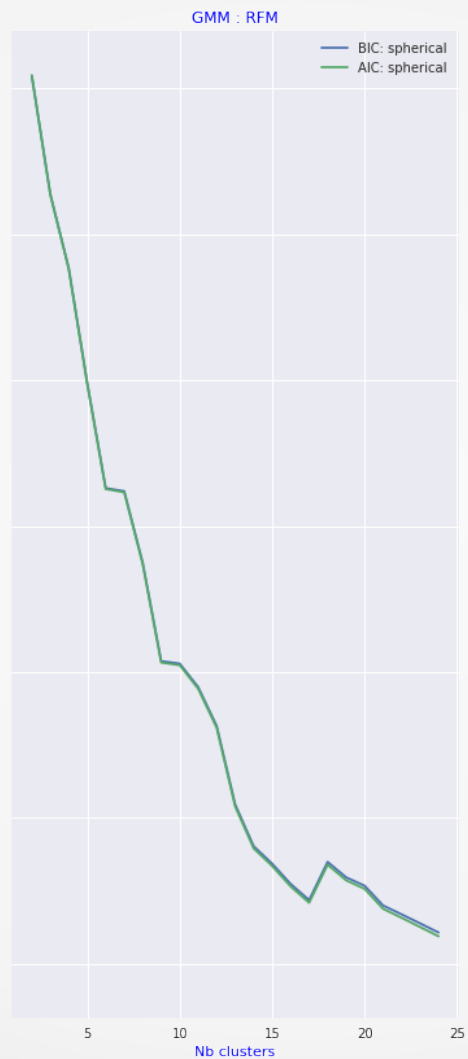
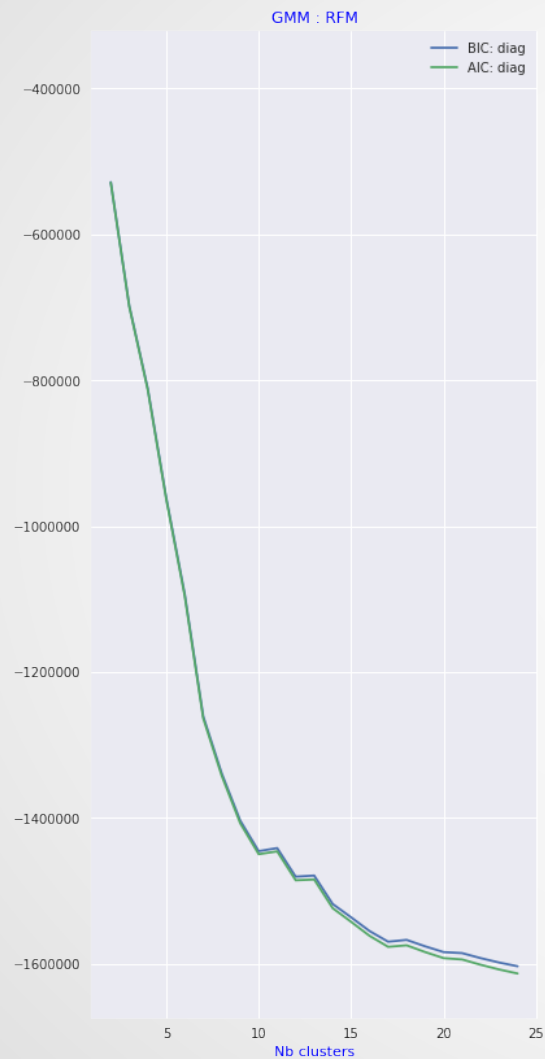
- 111, 411, 422, 433, 444

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4	All_segments
RFM						
111	402	0	0	0	0	402
112	0	0	0	0	78	78
113	0	0	0	0	16	16
114	0	0	0	0	1	1
121	0	0	0	0	75	75
122	0	0	0	0	115	115
123	0	0	0	0	38	38
124	0	0	0	0	13	13
131	0	0	0	0	16	16
132	0	0	0	0	40	40
133	0	0	0	0	61	61
134	0	0	0	0	19	19
141	0	0	0	0	4	4
142	0	0	0	0	12	12
143	0	0	0	0	19	19
144	0	0	0	0	55	55
411	0	305	0	0	0	305
412	0	0	0	0	144	144
413	0	0	0	0	34	34
421	0	0	0	0	115	115
422	0	0	0	365	0	365
423	0	0	0	0	161	161
424	0	0	0	0	80	80
431	0	0	0	0	33	33
432	0	0	0	0	161	161
433	0	0	0	433	0	433
434	0	0	0	0	186	186
441	0	0	0	0	30	30
442	0	0	0	0	65	65
443	0	0	0	0	218	218
444	0	0	626	0	0	626

# Clustering RFM : distributions

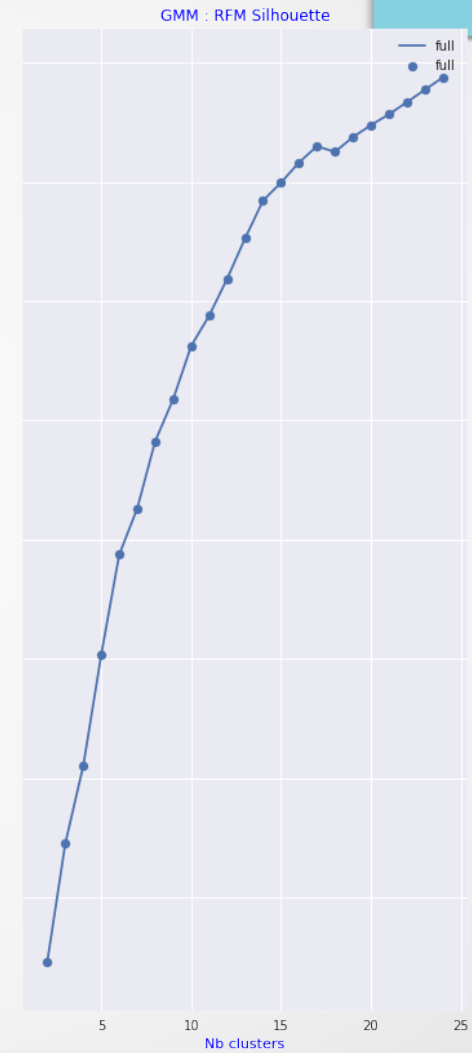
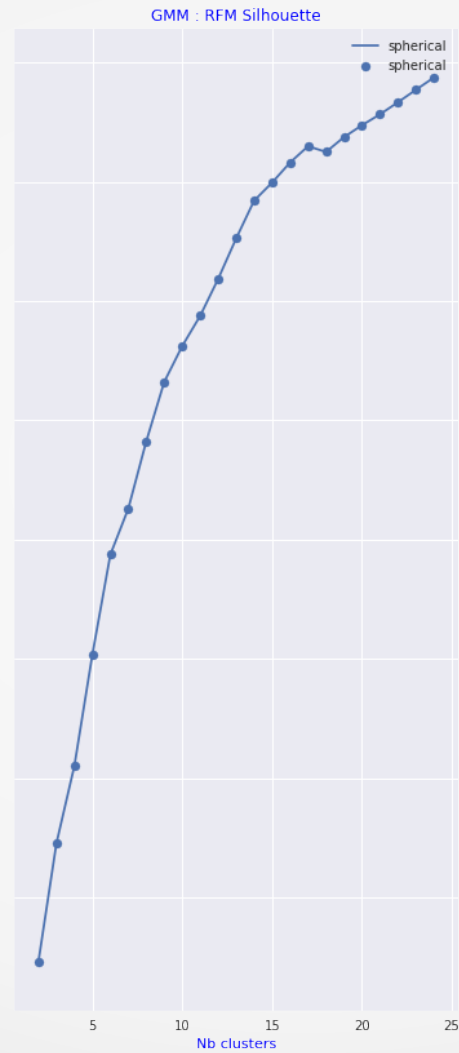
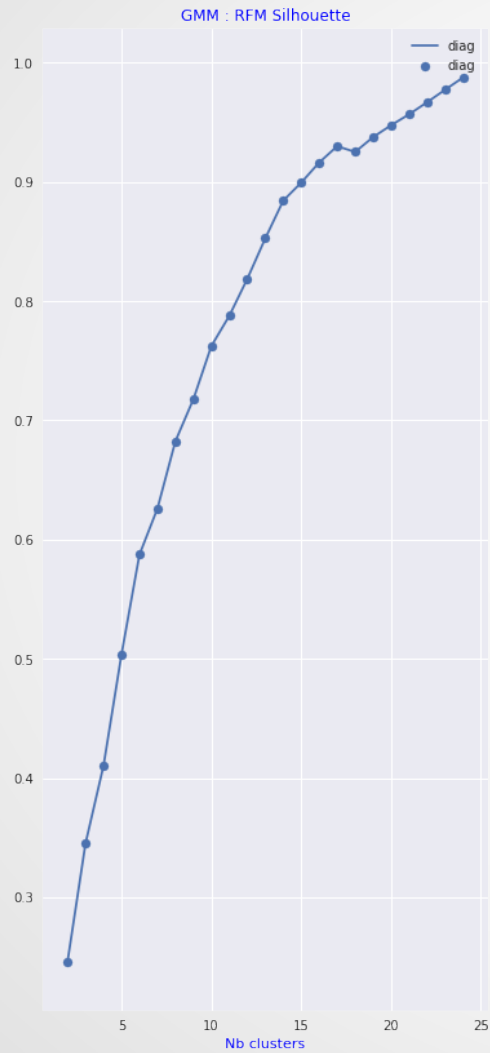


# Clustering RFM : GMM models

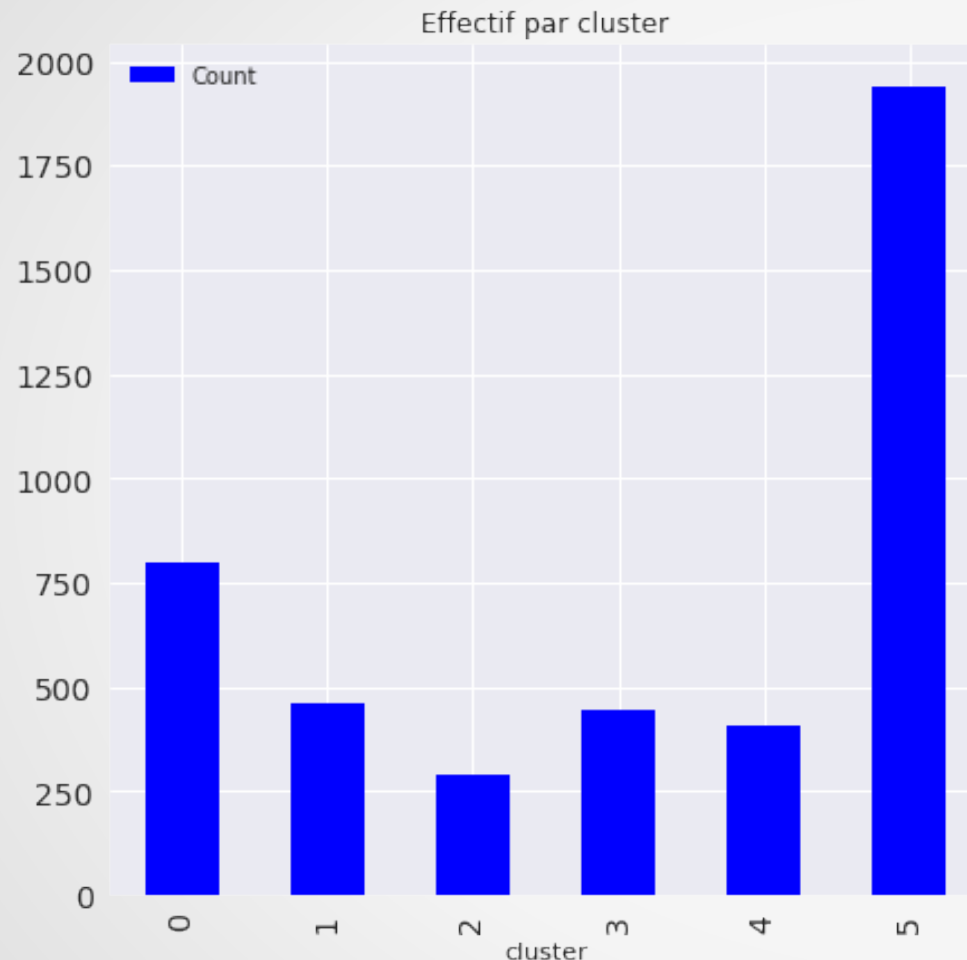


Nb clusters: 6

# Clustering RFM : GMM silhouette



# Clustering RFM : Mapping cluster/RFM (1)

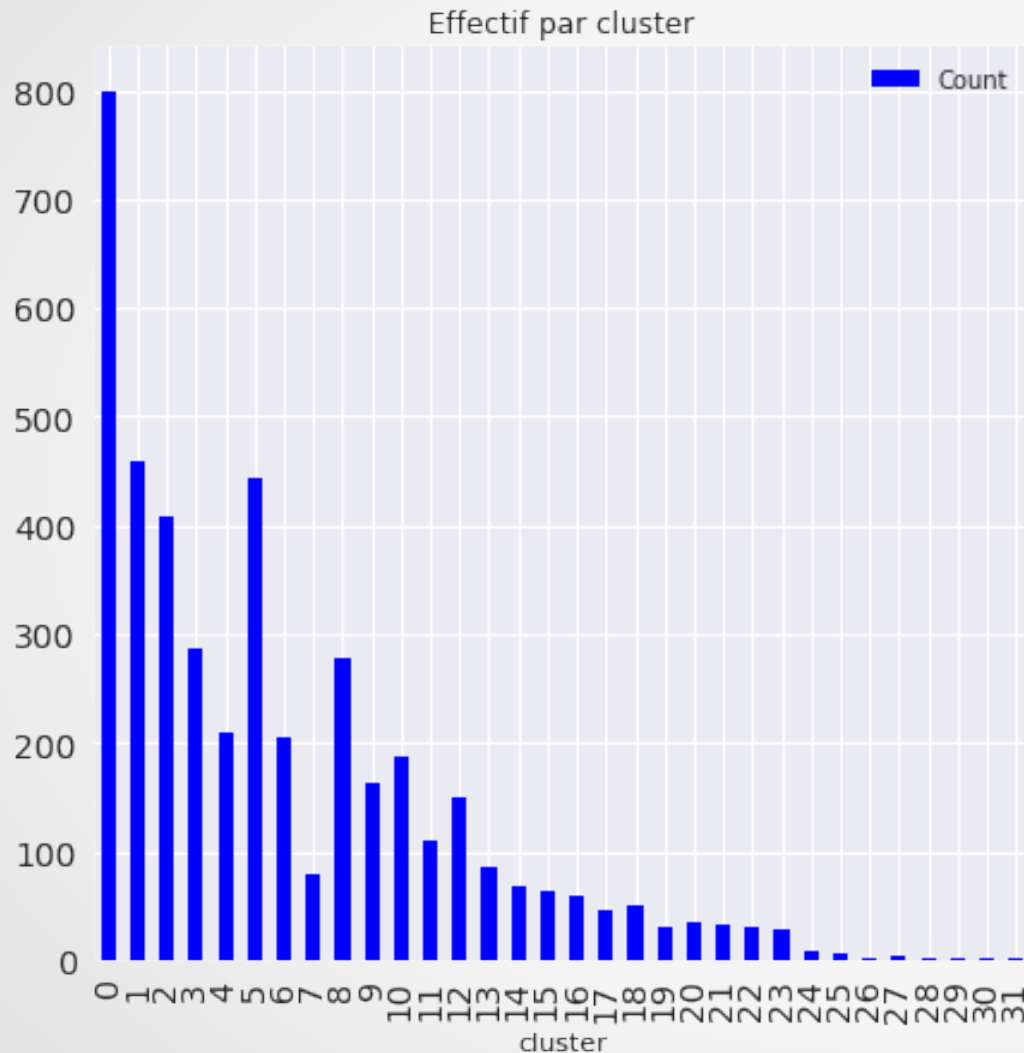


```
Cluster : 0 RFM = ['444']  
Cluster : 1 RFM = ['443']  
Cluster : 2 RFM = ['433']  
Cluster : 3 RFM = ['111']  
Cluster : 4 RFM = ['422']  
Cluster : 5 RFM = ['441' '421' '411'  
'442' '132' '122' '144' '431' '133'  
'432' '121' '423'  
'131' '434' '112' '142' '143' '412'  
'123' '134' '424' '124' '413' '141'  
'113' '114' '414']
```

5 clusters distincts



# Clustering RFM : Mapping cluster/RFM (2)



```
Cluster : 0 RFM = ['444']
Cluster : 1 RFM = ['443']
Cluster : 2 RFM = ['422']
Cluster : 3 RFM = ['433']
Cluster : 4 RFM = ['432']
Cluster : 5 RFM = ['111']
Cluster : 6 RFM = ['421']
Cluster : 7 RFM = ['412']
Cluster : 8 RFM = ['411']
Cluster : 9 RFM = ['122']
Cluster : 10 RFM = ['423']
Cluster : 11 RFM = ['442']
Cluster : 12 RFM = ['434']
Cluster : 13 RFM = ['121']
Cluster : 14 RFM = ['133']
Cluster : 15 RFM = ['112']
Cluster : 16 RFM = ['144']
Cluster : 17 RFM = ['132']
Cluster : 18 RFM = ['123']
Cluster : 19 RFM = ['441']
Cluster : 20 RFM = ['431']
Cluster : 21 RFM = ['134']
Cluster : 22 RFM = ['424']
Cluster : 23 RFM = ['143']
Cluster : 24 RFM = ['124']
Cluster : 25 RFM = ['131']
Cluster : 26 RFM = ['413']
Cluster : 27 RFM = ['142']
Cluster : 28 RFM = ['113']
Cluster : 29 RFM = ['141']
Cluster : 30 RFM = ['114']
Cluster : 31 RFM = ['414']
```

Covariance sphérique  
32 clusters : RFM séparés

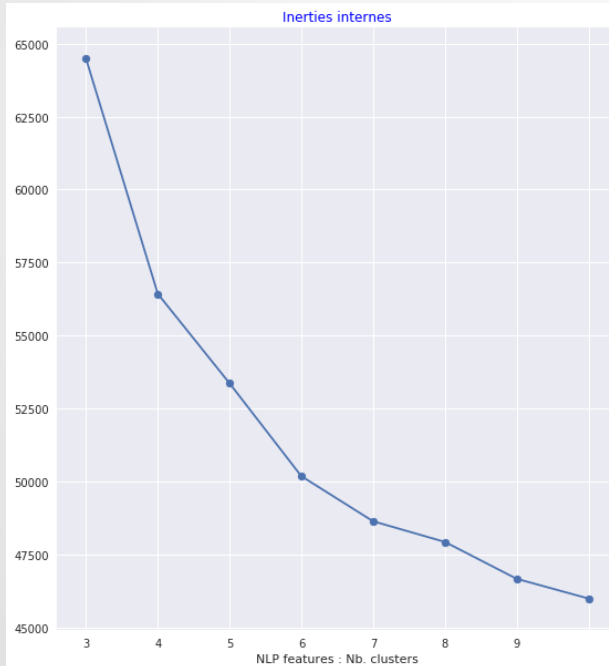
## Annexe 3 : Étude NLP

# NLP: Kmeans clusters

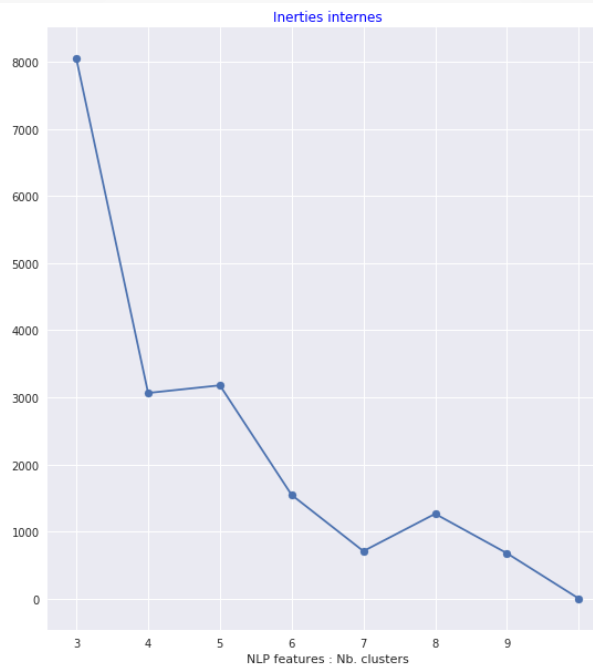
NLP

- 2373 clients
- 250 dimensions

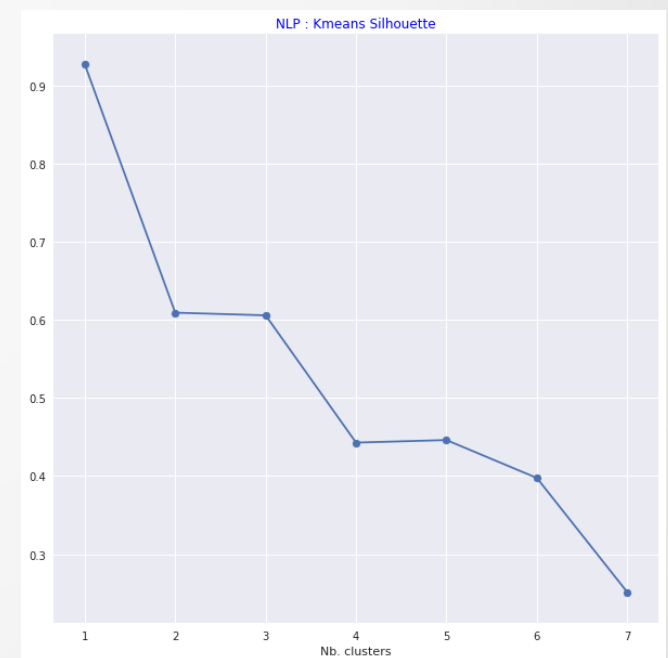
Inerties interne



Inerties interne :  
taux décroissance

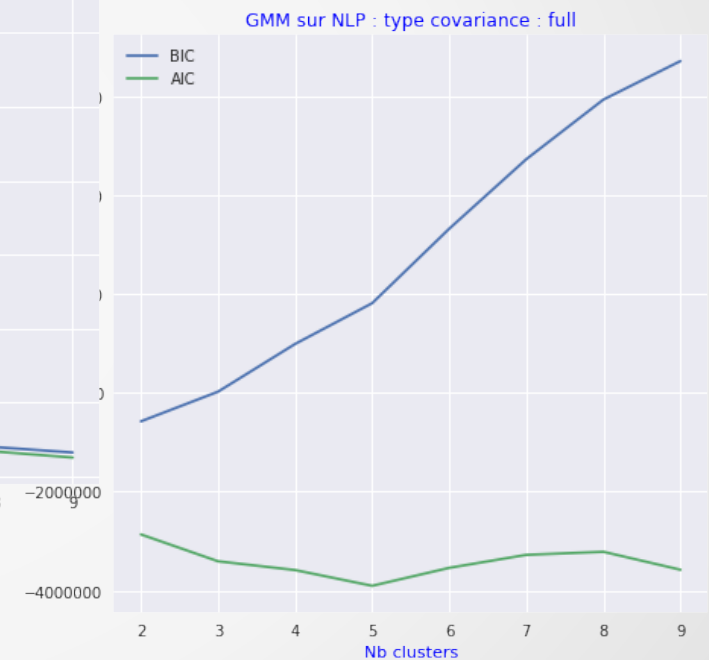
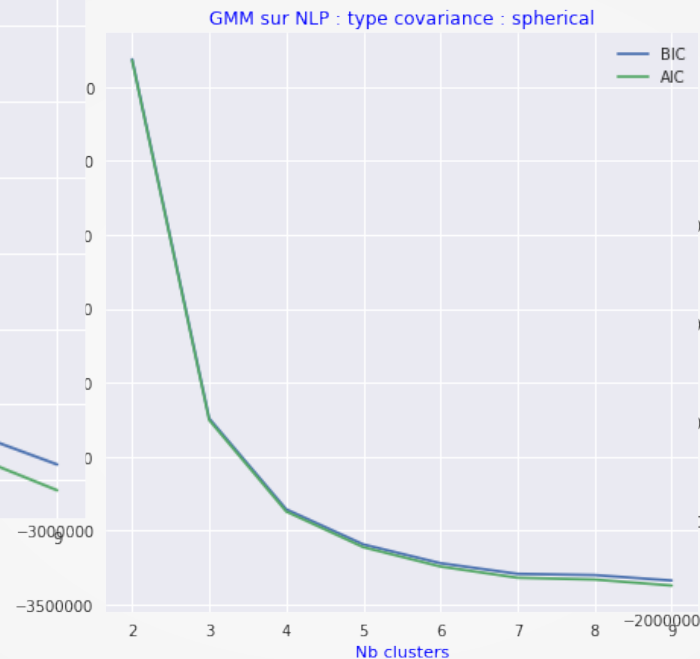
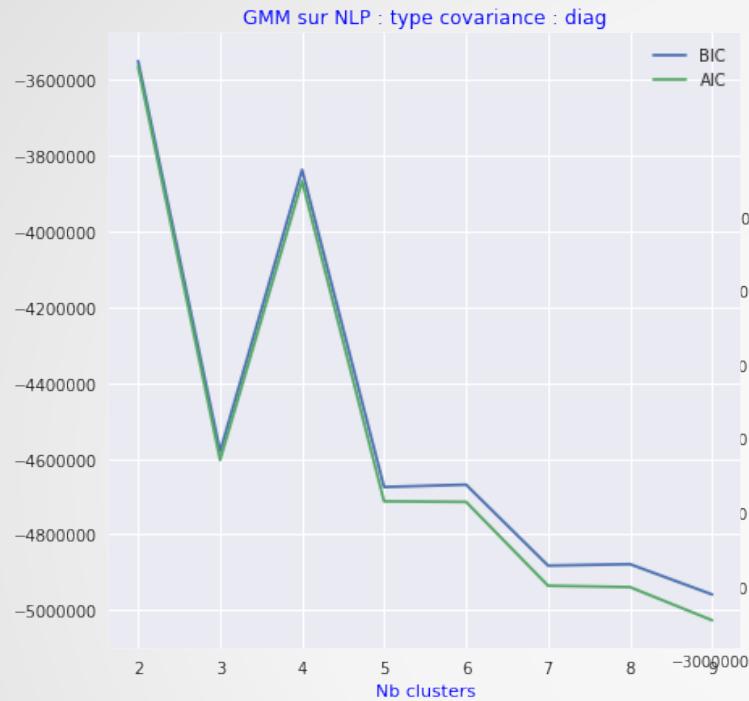


Coefficient de silhouette



Nb optimal de clusters : 3

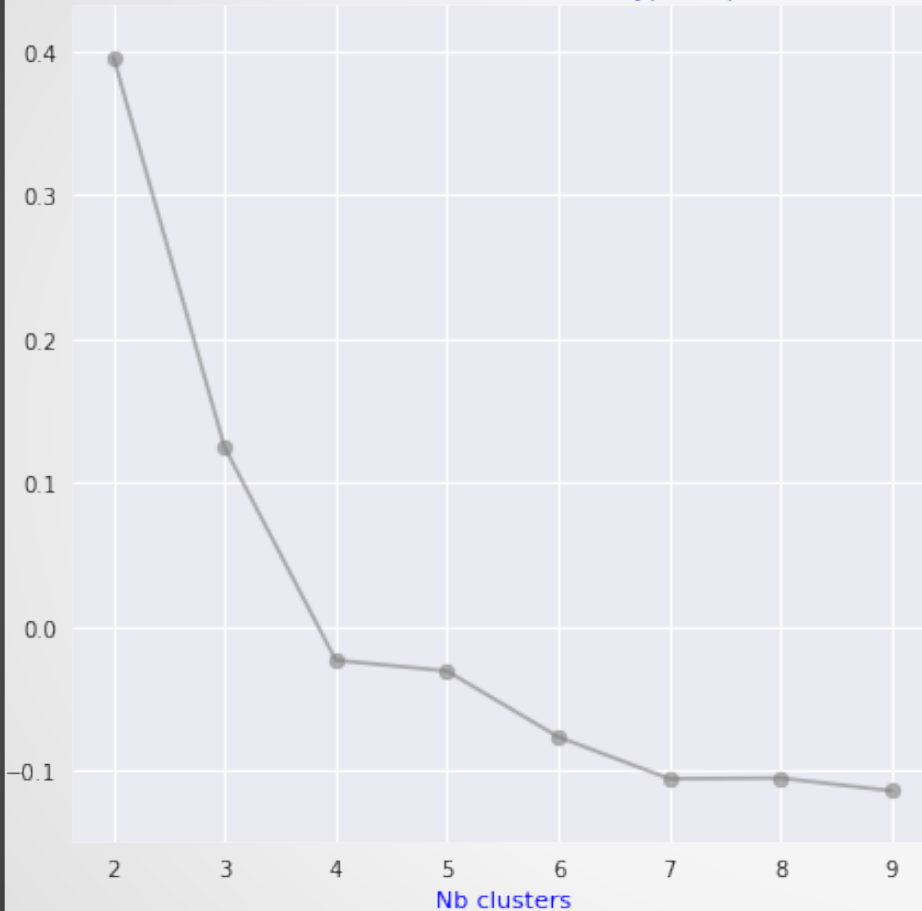
# NLP: GMM clusters et type de covariance



Nb optimal de clusters : 3  
Co-variance des axes : sphérique

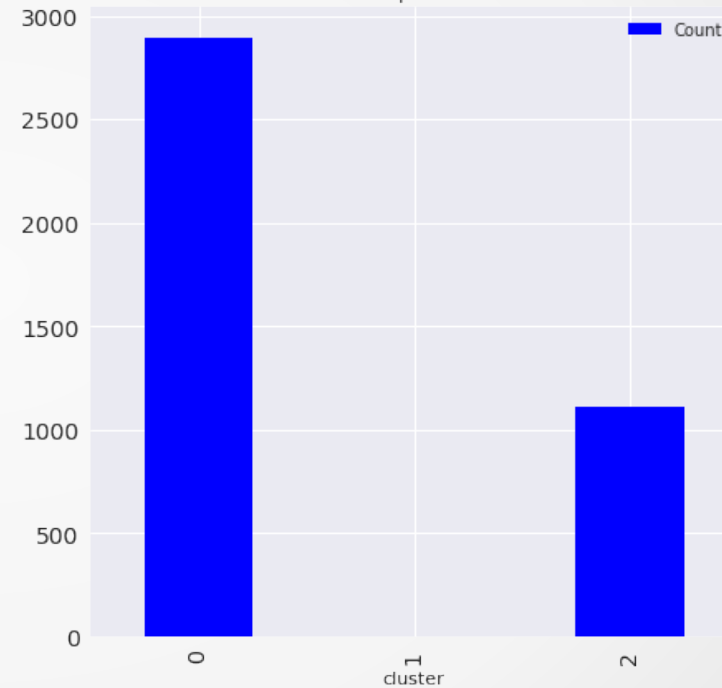
# NLP: GMM clusters et silhouette

GMM : silhouette with covariance type= spherical



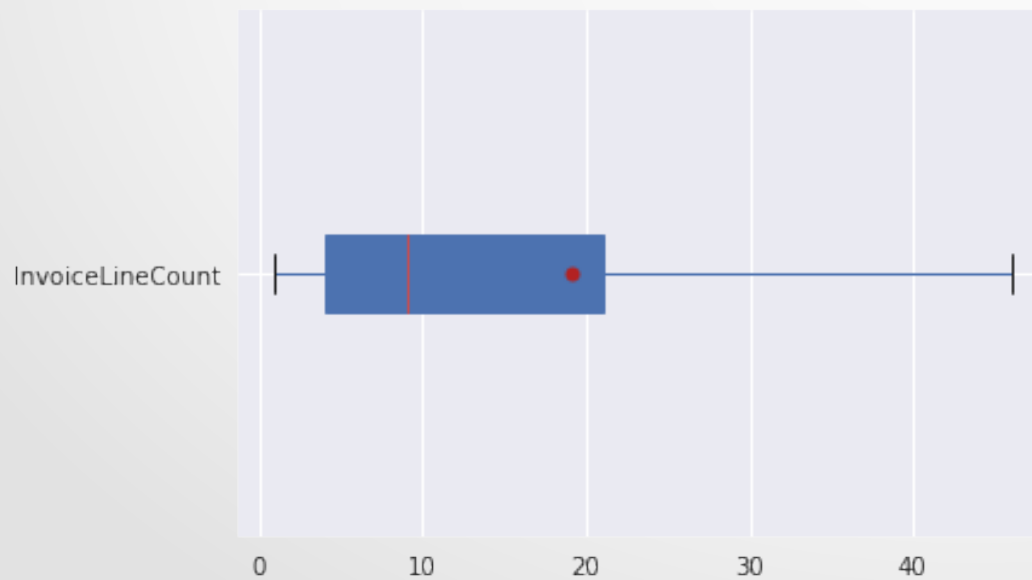
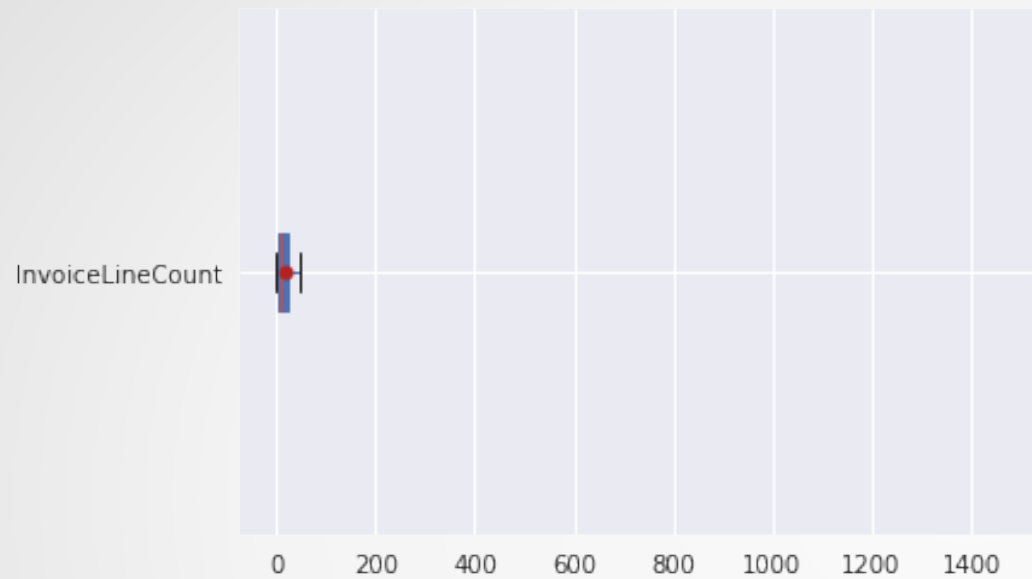
Nb optimal de clusters : 3  
Convergence : True

Effectif par cluster



cluster	Count
0	2896
1	1
2	1113

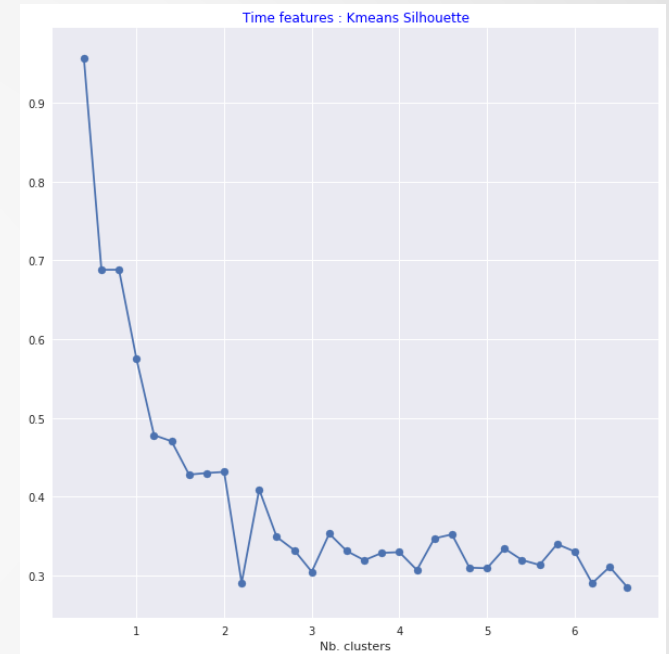
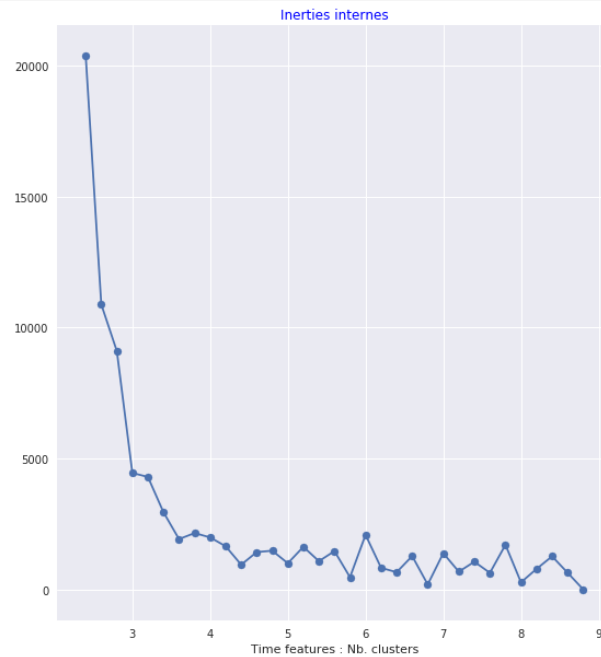
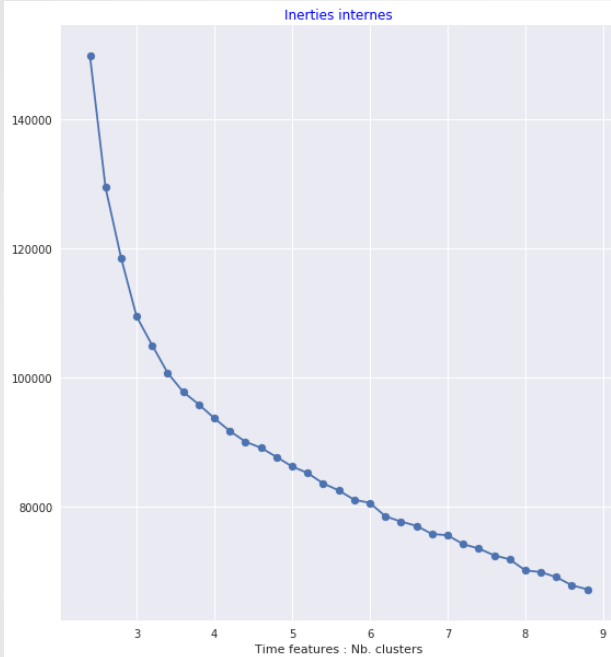
# NLP: Cluster 1



Cluster : 1  
Customer ID : 17841  
Transactions : 1484

## Annexe 4 : Étude time

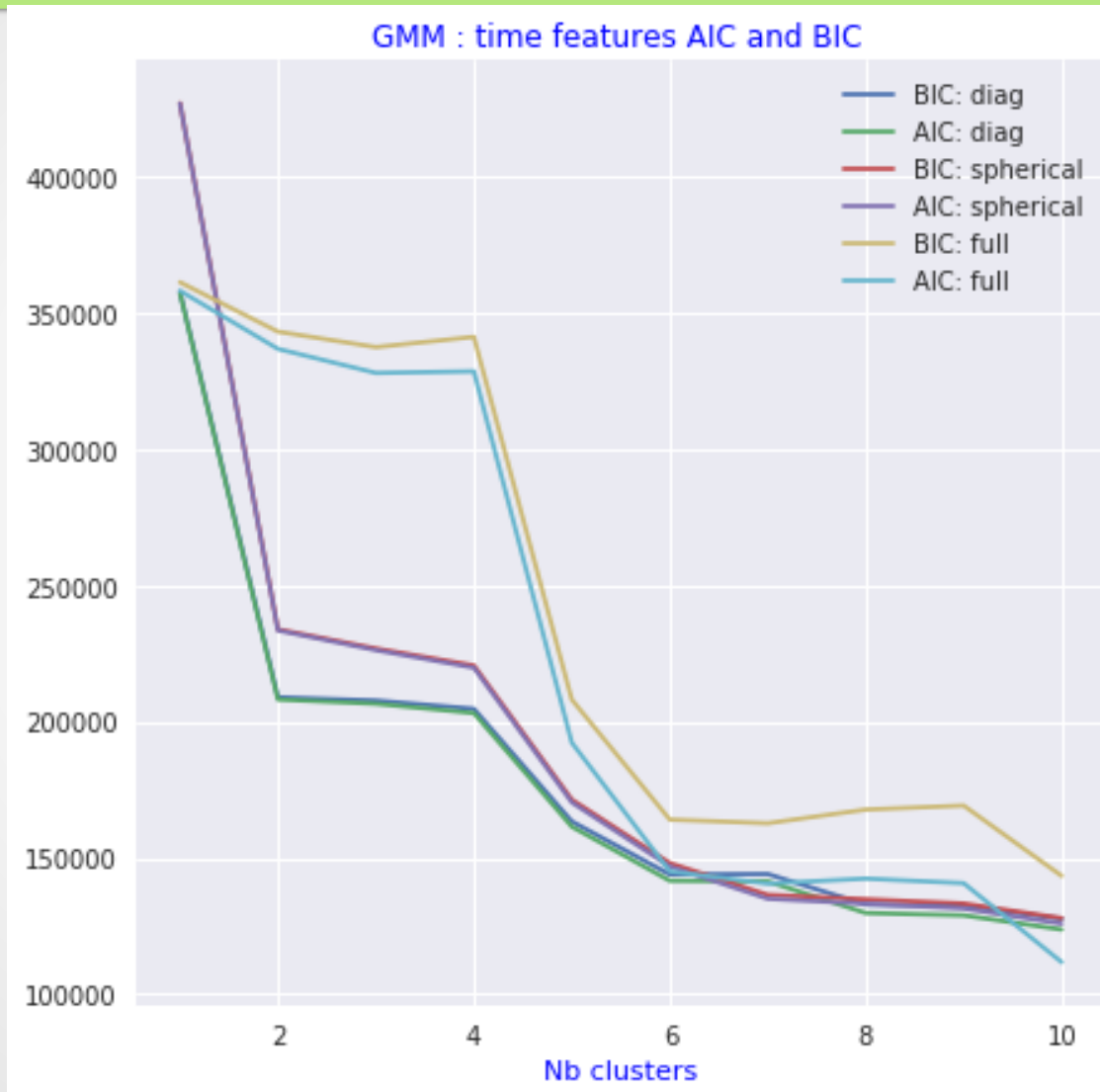
# Time : Kmeans clustering



Nb optimal de clusters : 3

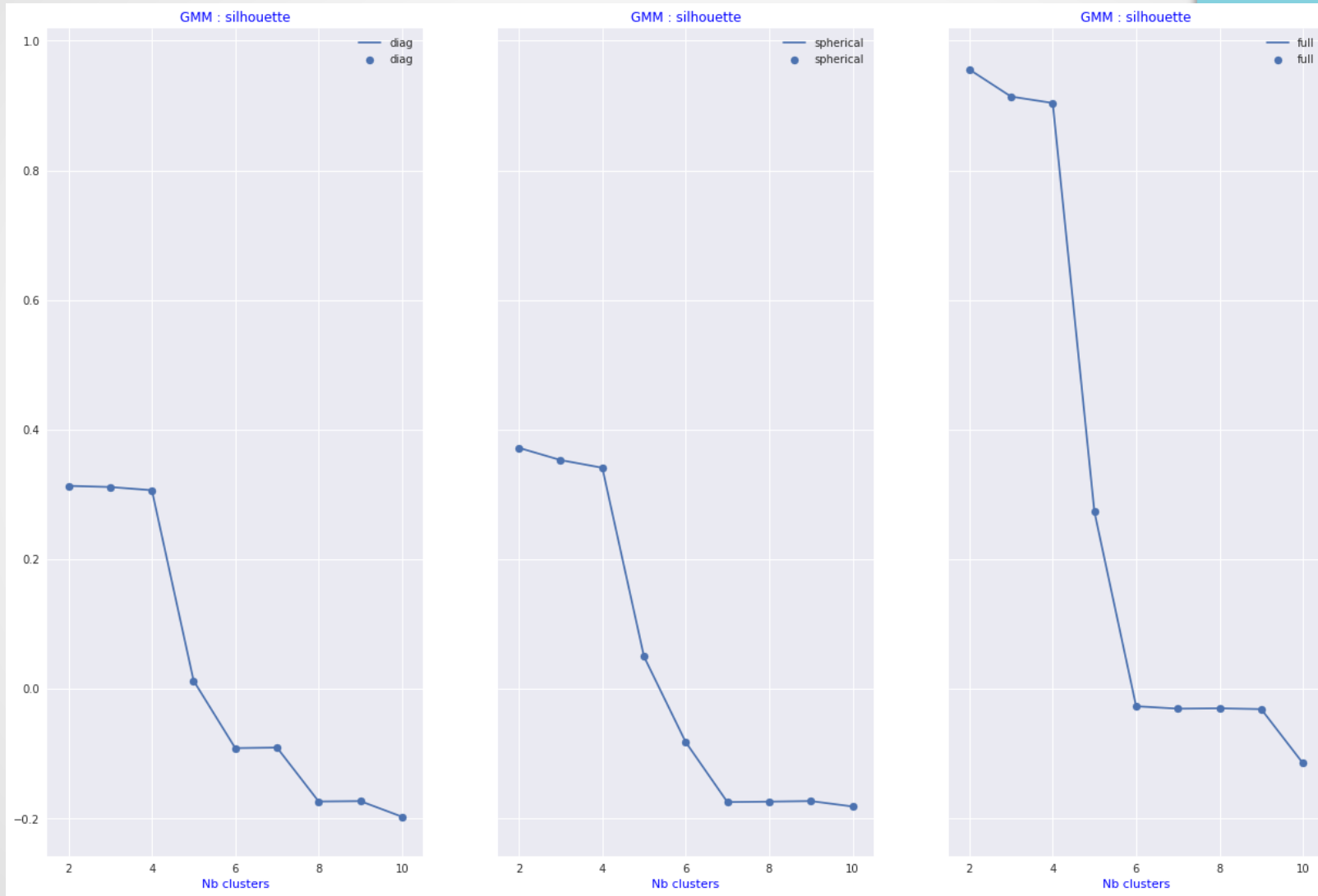


# Time : GMM clustering vs AIC and BIC



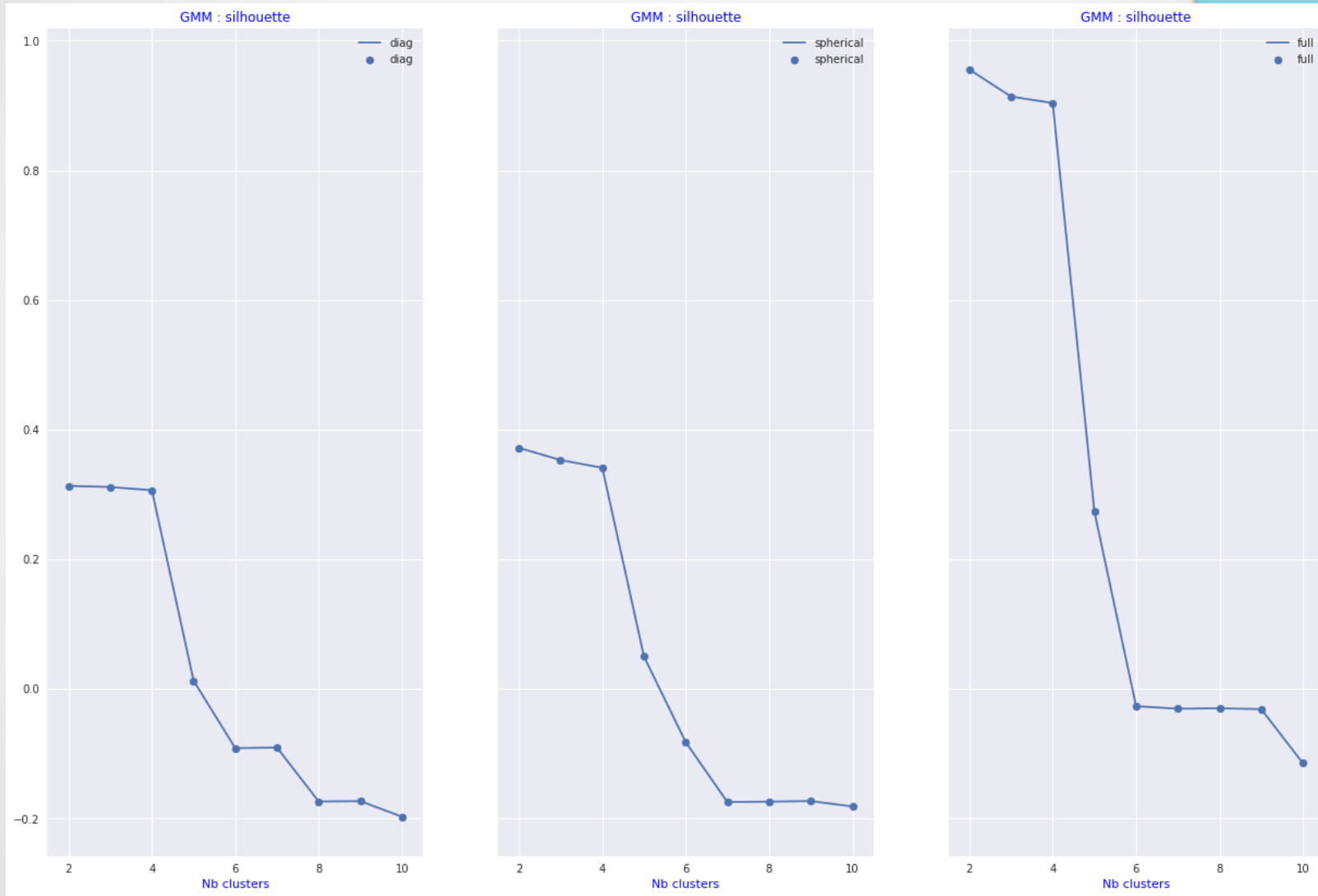
Nb optimal de clusters : 6

# Time : GMM clustering vs Silhouette



Nb optimal de clusters : 4  
Type covariance : Full

# Time : GMM clustering vs Silhouette



Nb optimal de clusters : 4  
Type covariance : Full

# Annexe 6 : API WEB

- Informations du data-set :

- [http://localhost:5000/?\\*](http://localhost:5000/?*)

- ```
{ "_results": [ { "customer_count": "3921", "invoice_count": "16661", "invl_count": "349216" } ] }
```

- Informations client

- <http://localhost:5000/?customerID=12822>

- ```
{ "_results":  
[ { "customerID":12822 ,"marketID":2 ,"invoice_count":16  
,"item_count":13756 ,"invl_count":72 ,"mean_unit_price":2.28  
,"incomes":31308.58 ,"old_date":2011-09-13 13:46:00  
,"new_date":2018-08-29 12:14:08 ,"RFM":144 } ] }
```

- Achat en ligne et prédiction d'un client inexistant

- <http://localhost:5000/?order&customerID=0&stockCode=22812&quantity=3&orderDate=NONE>

- ```
{ "_results": [ { "customerID": "18288", "marketID": "0" } ] }
```

- Achat en ligne et prédiction d'un client existant (dataset validation)

- <http://localhost:5000/?order&customerID=0&stockCode=22812&quantity=1&orderDate=NONE>