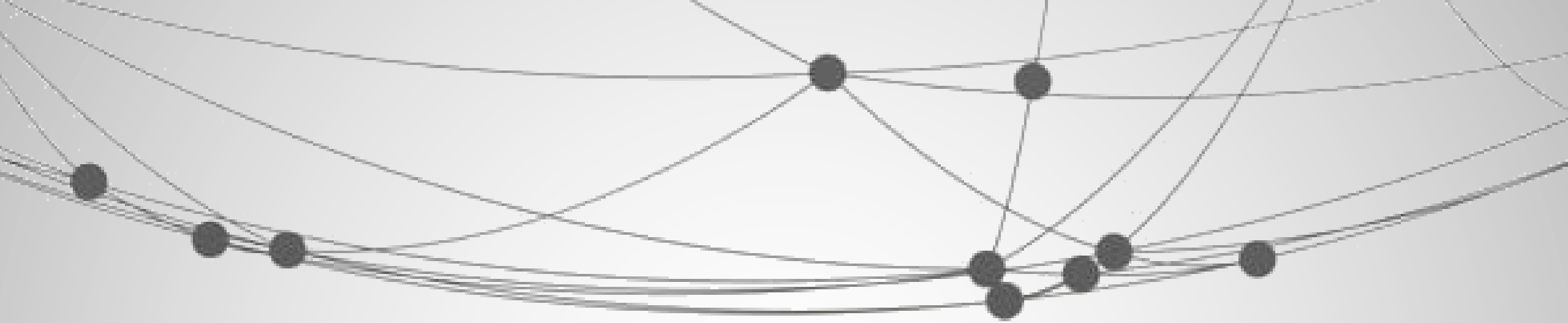
AI Engineer PATH

Project 5

Segment the customer base of an e-commerce Website

Axel Favreul

Contents



01

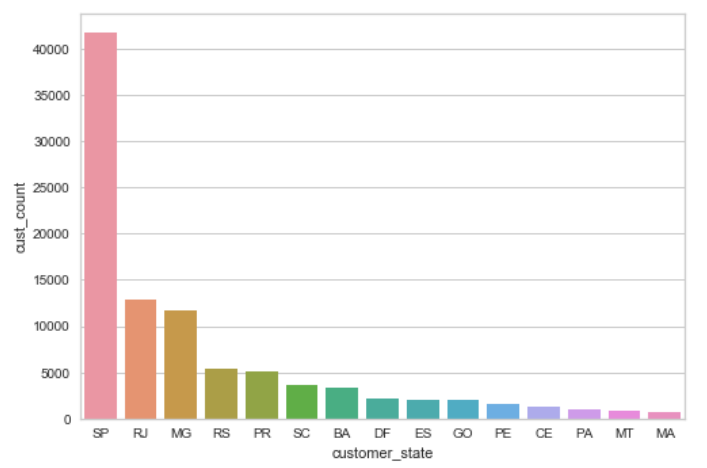
Exploratory Analysis

Dataset Overview

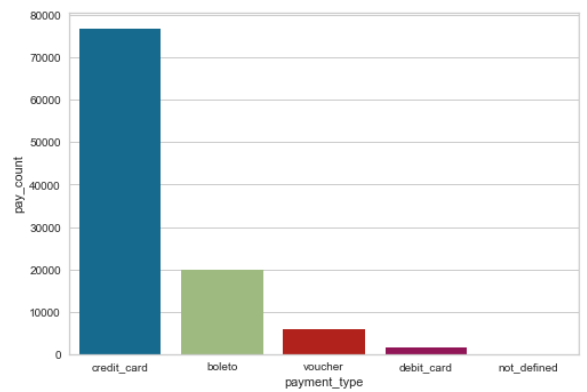
Olis wants to develop a customer segmentation that its marketing teams can use on a routine basis for their communication campaigns.

5 differents datasets containing :

Information about customer, seller, product, order and payment.

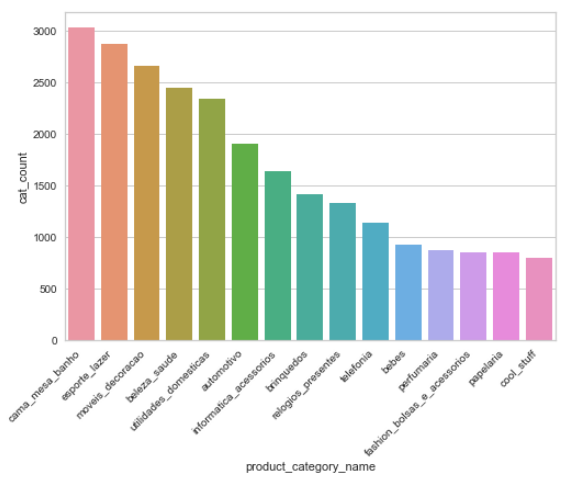


Customer state distribution



Customer payment distribution

	order_status	status_count
3	delivered	96478
6	shipped	1107
1	canceled	625
7	unavailable	609
4	invoiced	314

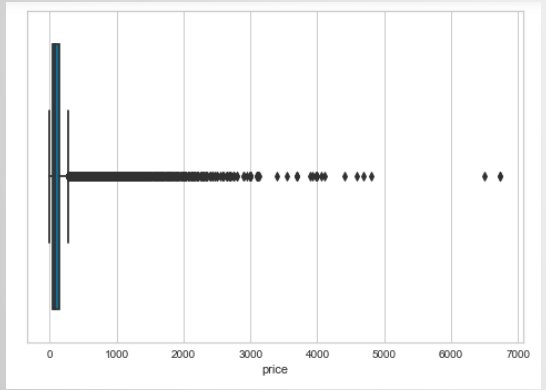


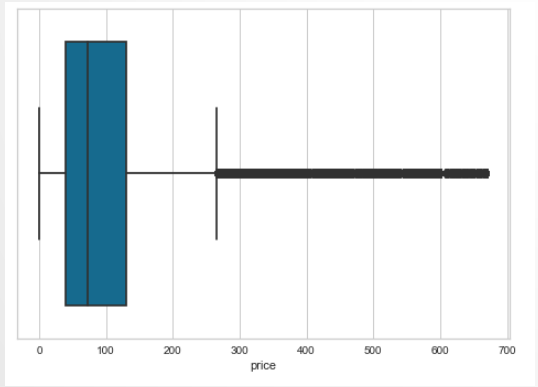
Dataset Overview

Outlier

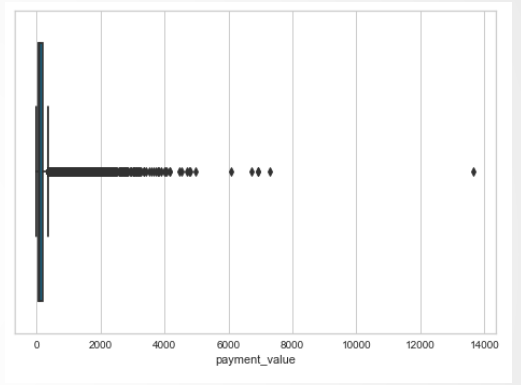
Method: z score outside 3

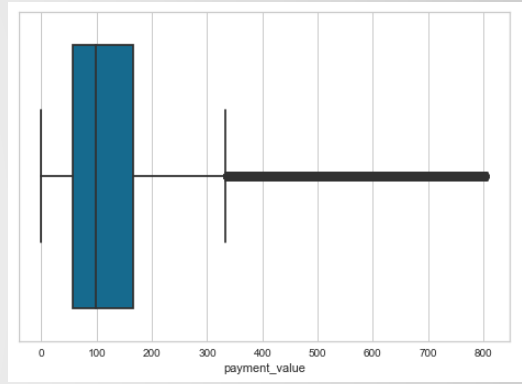
Price



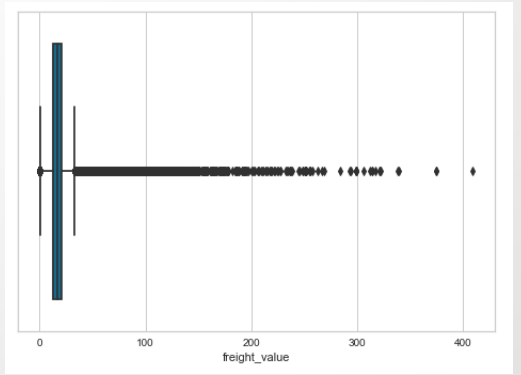


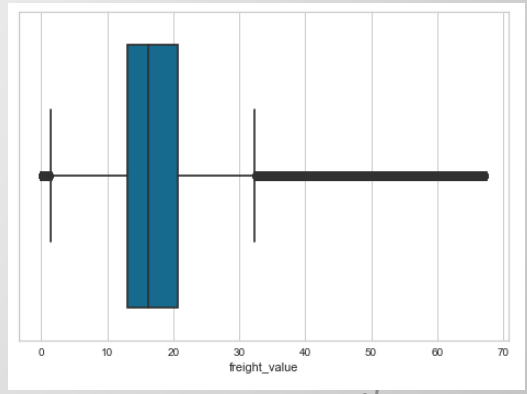
Payment





Freight value





Dataset Overview

Missing numerical value

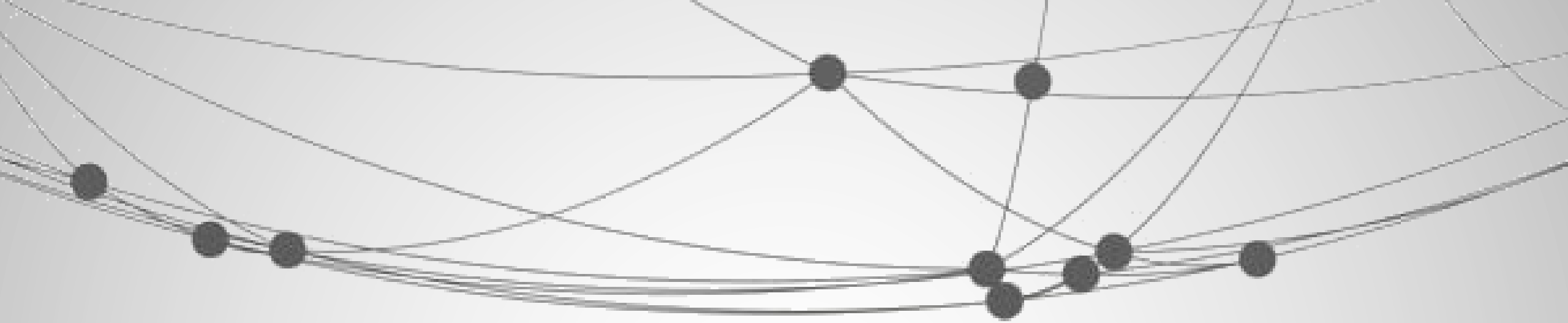
	Miss_value	Miss_value_100
payment_value	2739	2.314948
price	2089	1.765581

Method: drop due to the small number of missing value

Method chosen : Recency Frequency Monetary value

Recency: Latest order
Frequency: Number of order
Monetary value: Sum of payment value

customer_unique_id	recency	frequency	expense
0000366f3b9a7992bf8c76cfdf3221e2	160	1	141.90
0000b849f77a49e4a4ce2b2a4ca5be3f	163	1	27.19
0000f46a3911fa3c0805444483337064	585	1	86.22
0000f6ccb0745a6a4b88665a16c9f078	369	1	43.62
0004aac84e0df4da2b147fca70cf8255	336	1	196.89



02

Model testing

Overview

+ Scaler

Scaler : Standard

+ Model tested:

Kmeans

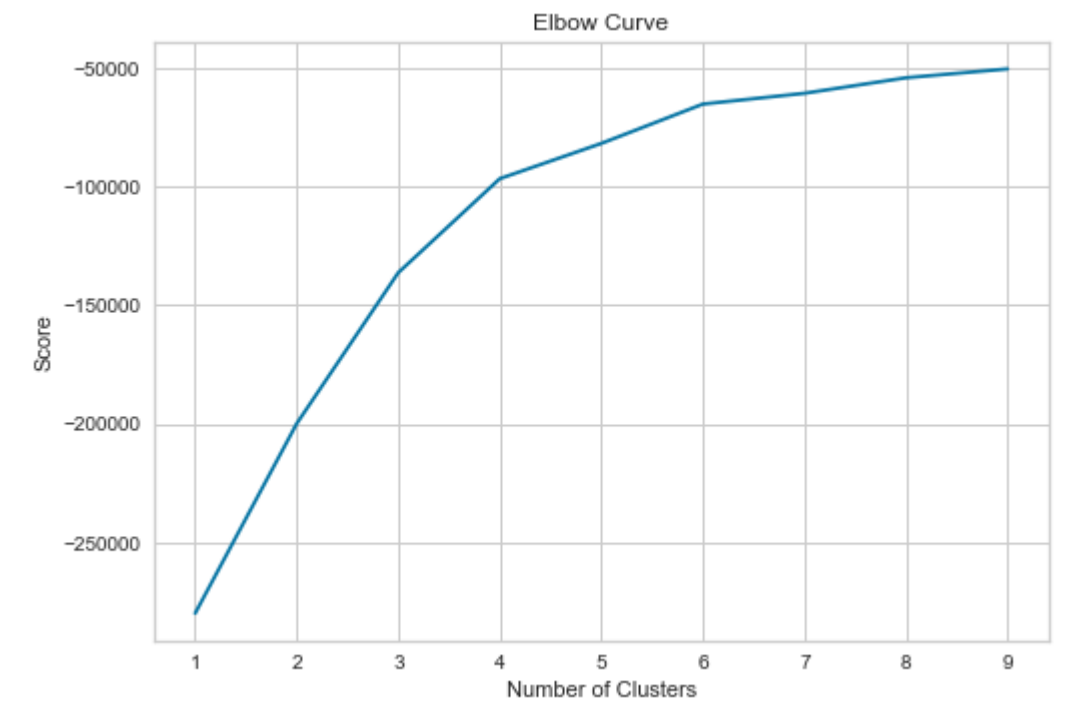
Gaussian

+ Method

- Choose number of cluster
- Perform clusterization
- Visualization
- Comparison (silhouette score)
- Stability

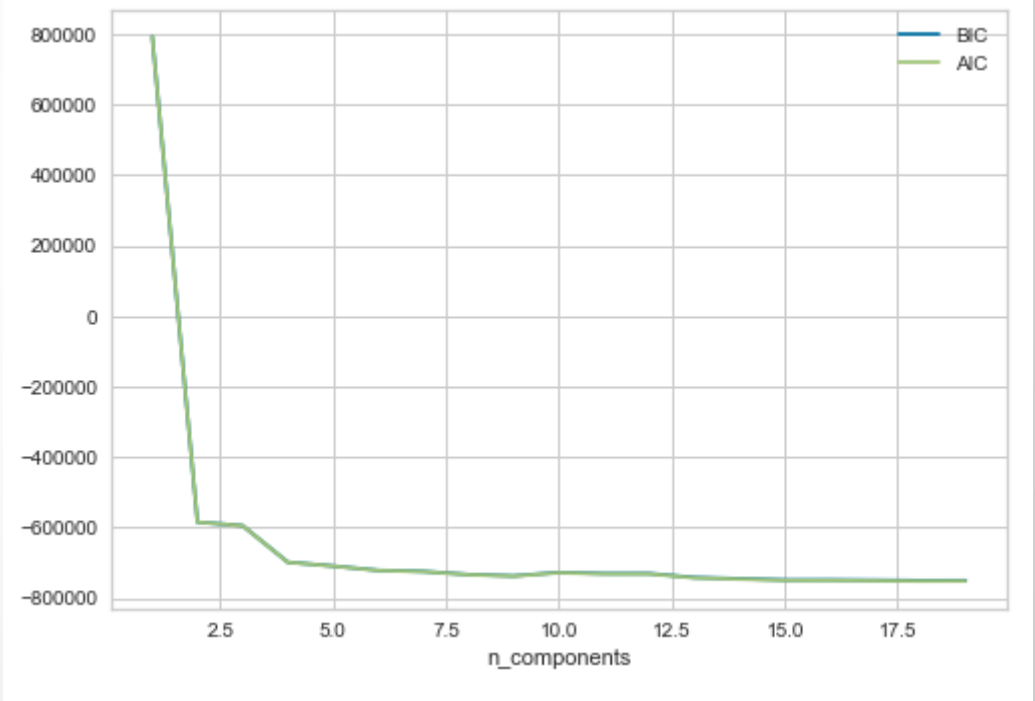
Number of cluster

+ Kmeans



Elbow curve

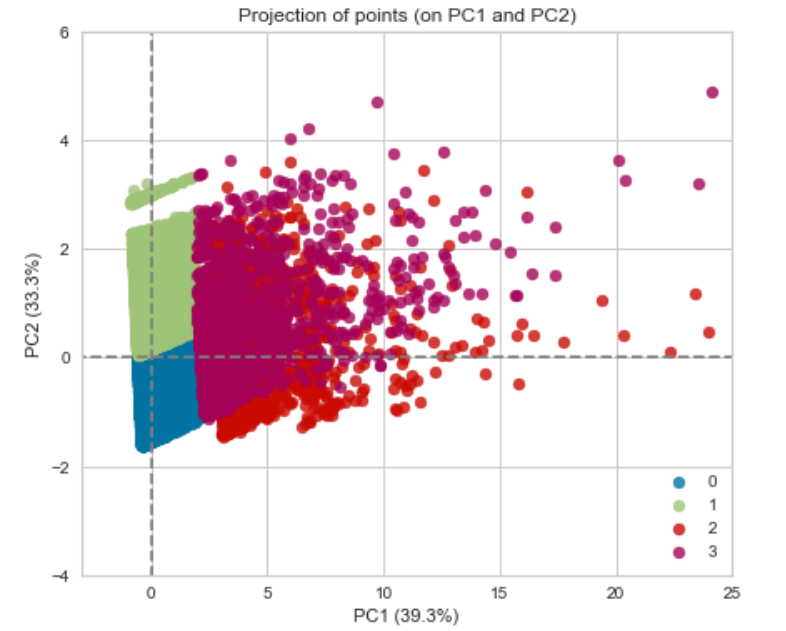
+ GMM



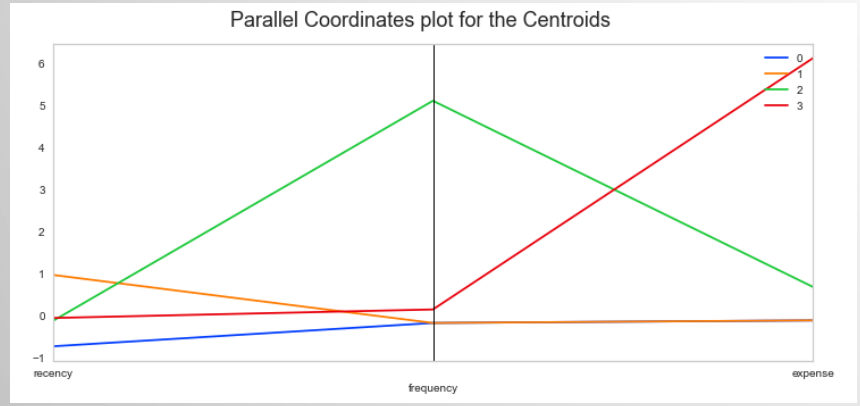
Akaike information criterion

Visualization

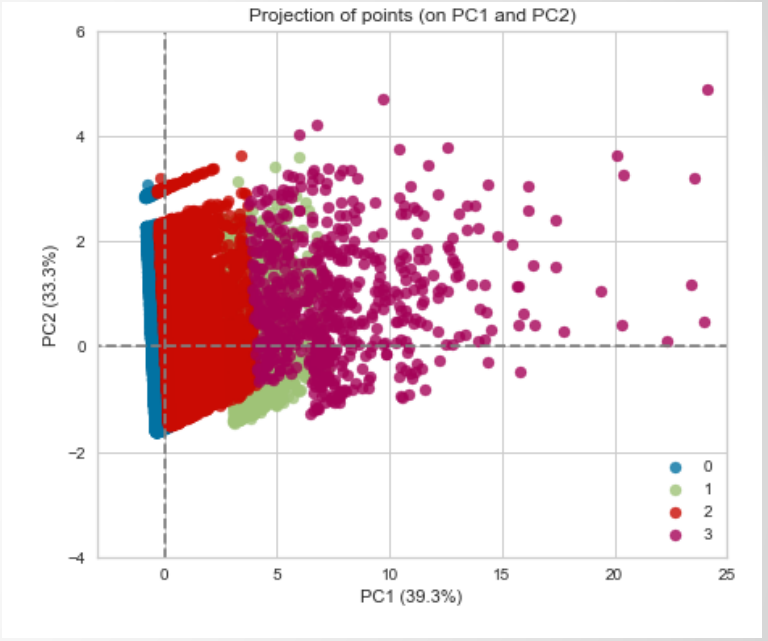
+ Kmeans

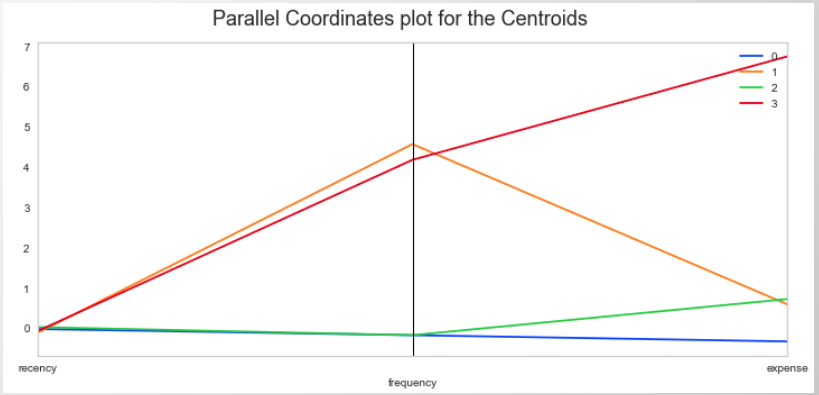


	PC1	PC2
recency	-0.107919	0.992037
frequency	0.705436	0.030391
expense	0.700509	0.122226



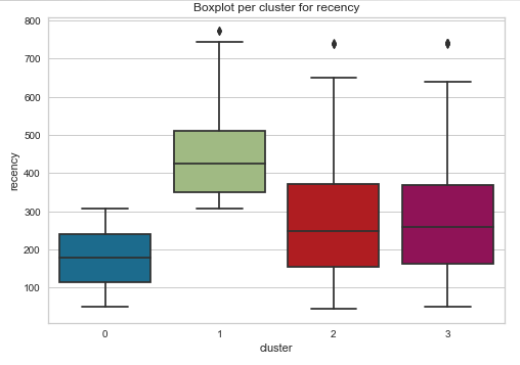
+ GMM

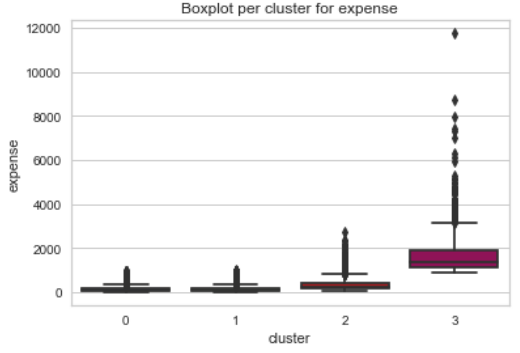


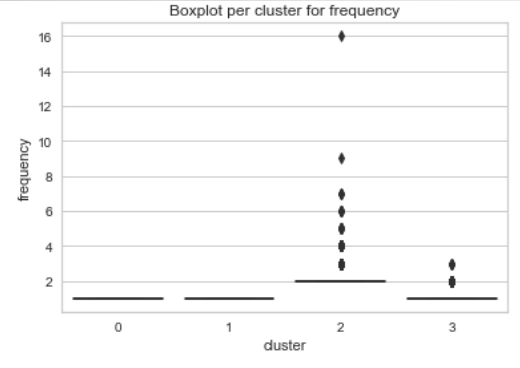


RFM Exploration

+ Kmeans

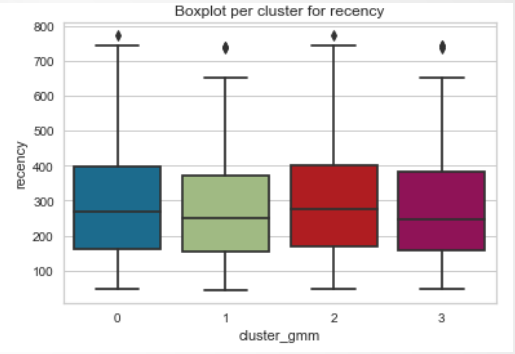


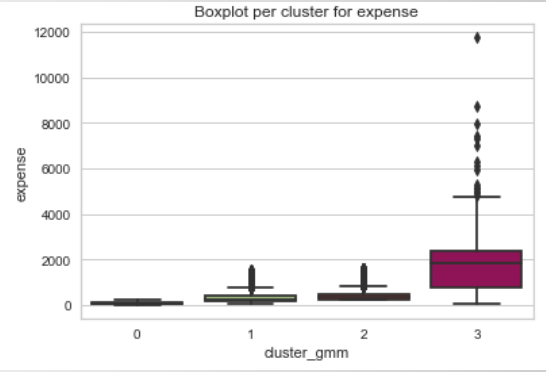


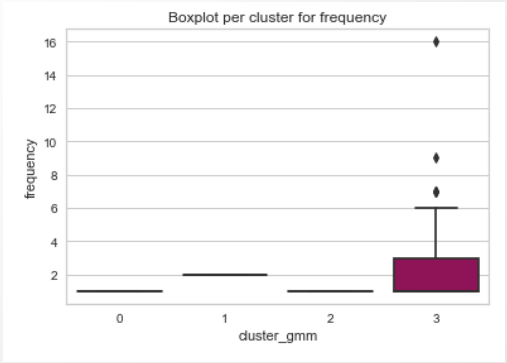


0	51498
1	37920
2	2770
3	1136

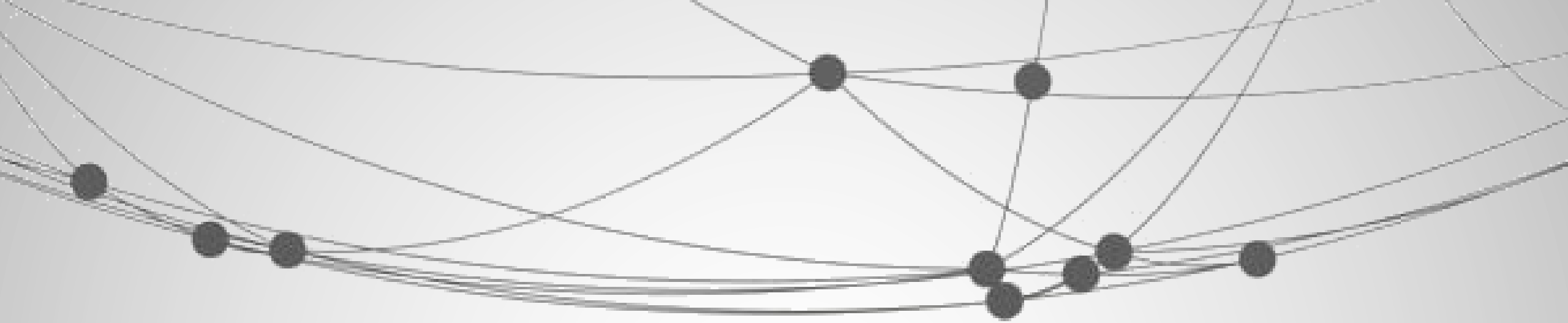
+ GMM







0	73095
2	17025
1	2552
3	652

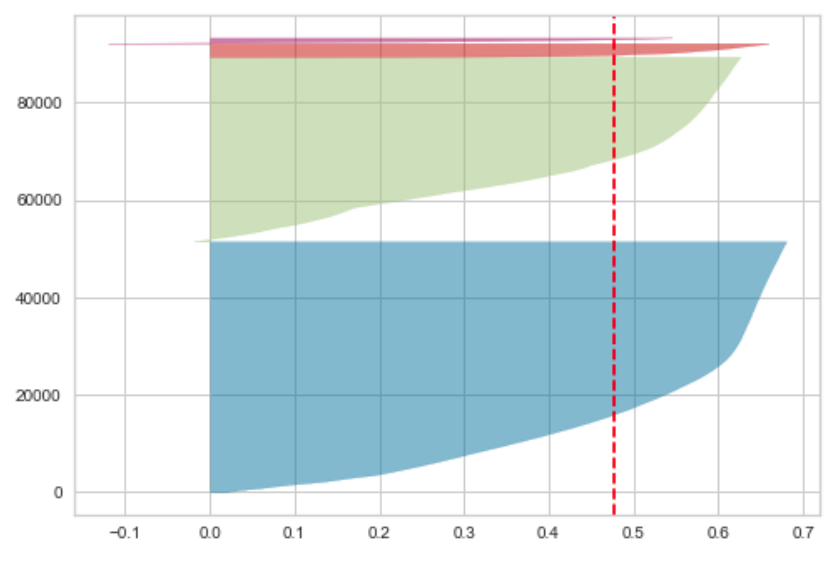


03

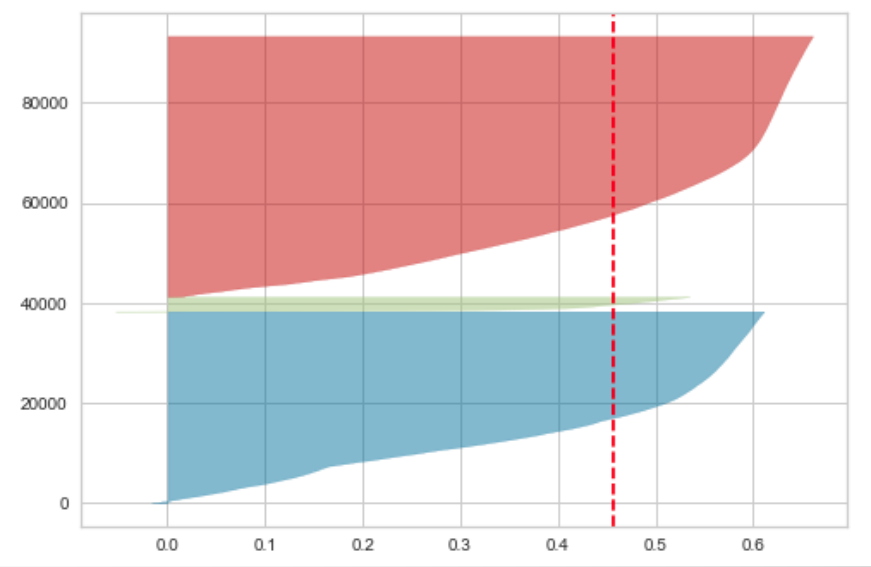
Kmeans Exploration

Silhouette score

+ 4 clusters



+ 3 clusters

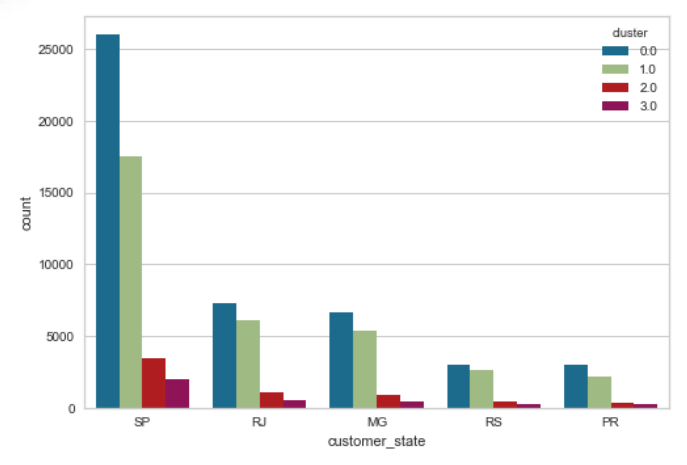


Silhouette score Kmeans: 0,477

Silhouette score GMM: 0,32

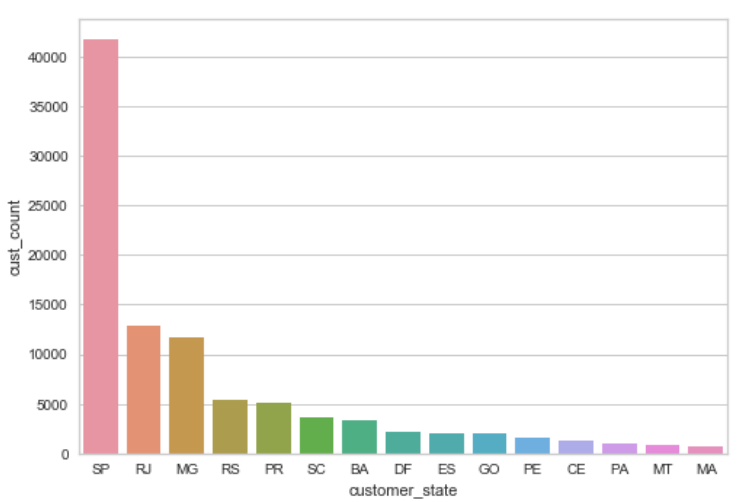
Exploration

+ Kmeans distribution

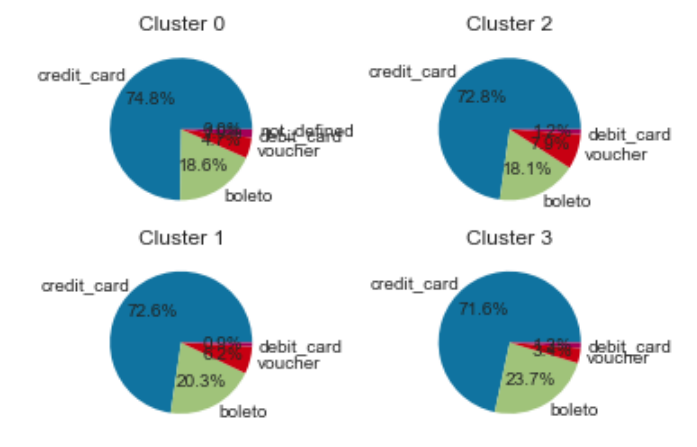


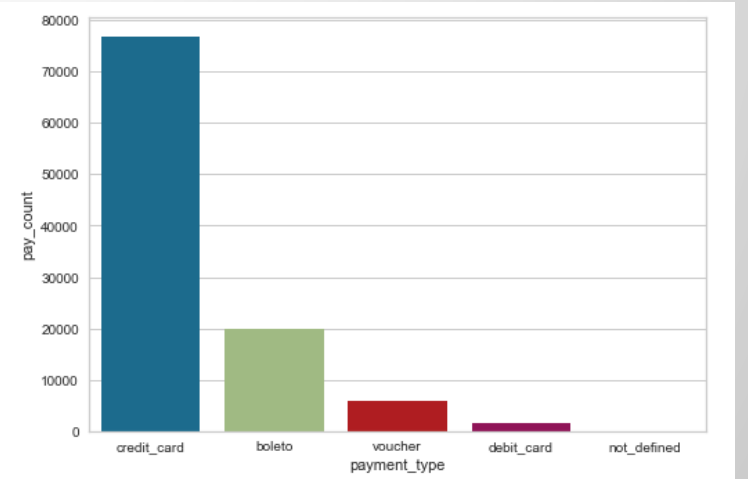
State distribution

+ Global distribution



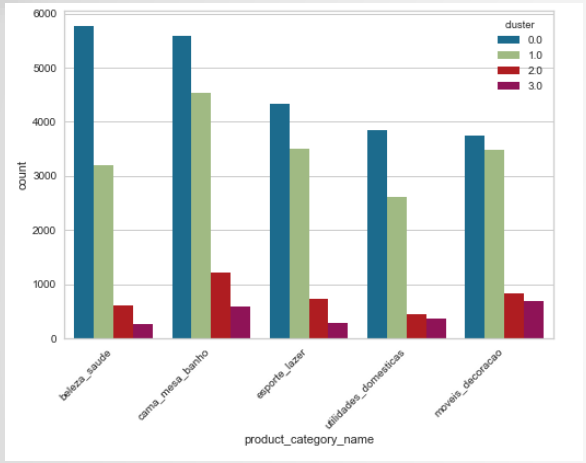
Payment methods



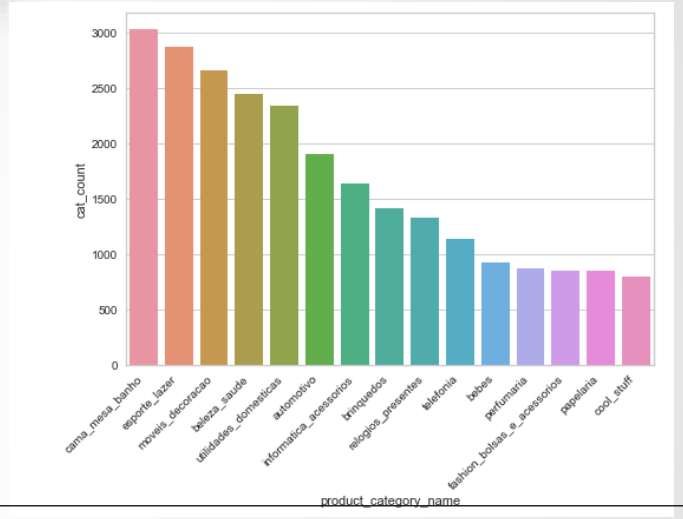


Exploration

+ Kmeans distribution

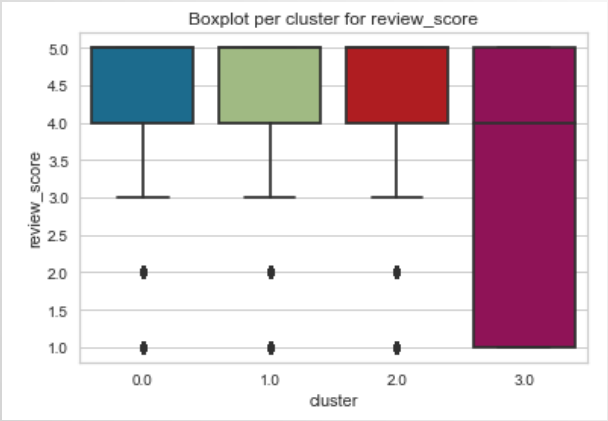


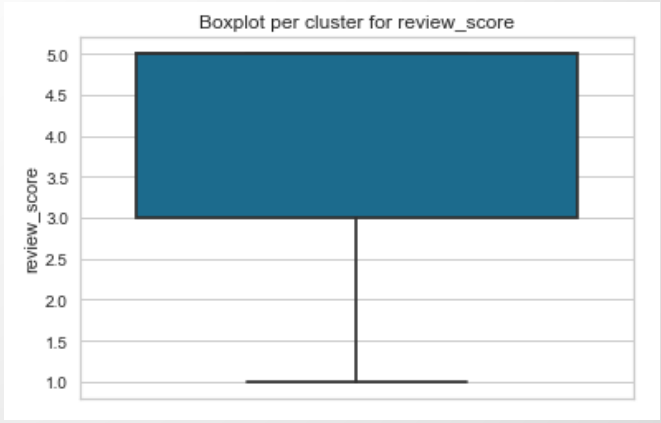
+ Global distribution



Product distribution

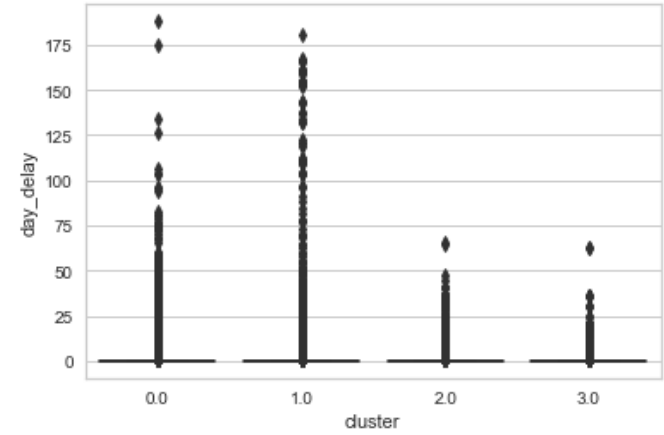
Review score





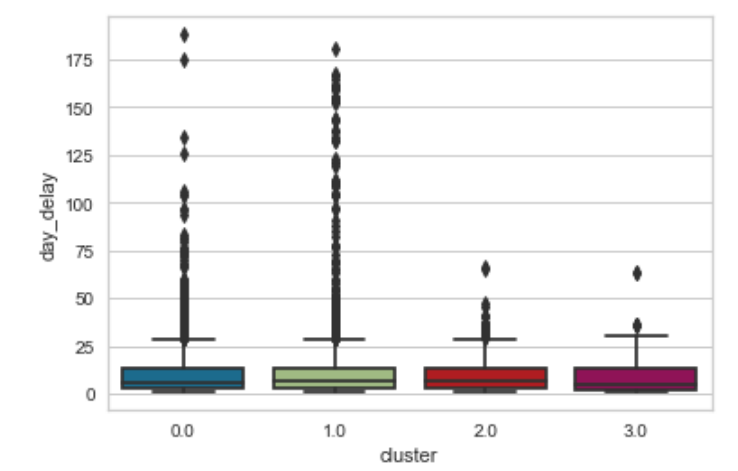
Exploration

+ Kmeans delay distribution



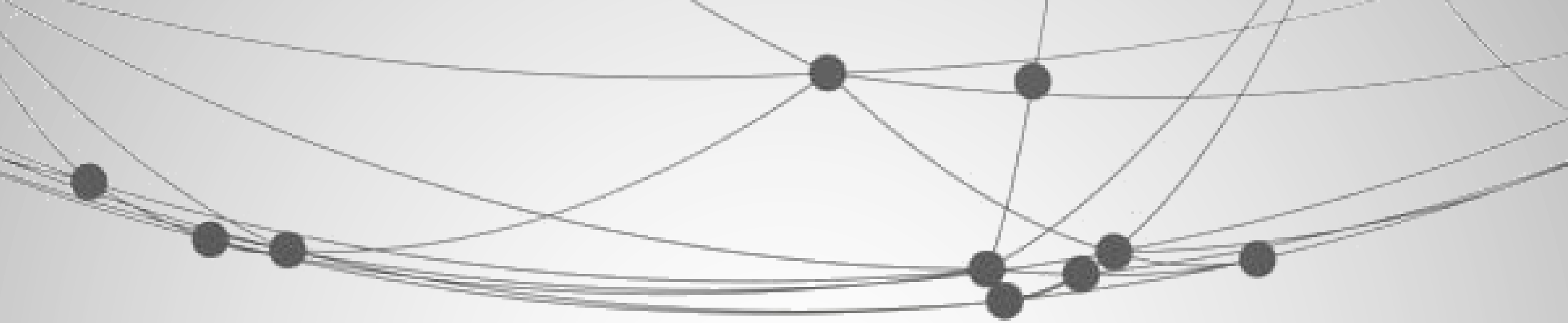
With no delay

cluster	day_delay
0.0	0.716591
1.0	0.659056
2.0	0.539414
3.0	0.417678



Only when delay occurs

cluster	day_delay
0.0	9.921989
1.0	11.951169
2.0	10.123487
3.0	8.382609



04

Stability

Stability

Tested on a subdataset of 90%

Cluster distribution :

1	46343
2	34152
3	2463
0	1034

Train and Distribution on the 90% dataset

0	46356
1	34143
2	2463
3	1030

Predict and Distribution on the full dataset (filter with same index)

13 data were misplaced within the clusters which is 0,02 %

05

Conclusion

Conclusion

Cluster 0 : former /occasional

- Low recency
- Low frequency
- Low monetary value
- 54% of the customer base
- Facing delay

Cluster 1: recent/ occasional

- High recency
- Low frequency
- Low monetary value
- 41,5% of the customer base
- Facing delay

Cluster 2: loyal client

- Average recency
- High frequency
- Low monetary value
- 3% of the customer base

Cluster 3: occasional high expense client

- Average recency
- Low frequency
- Very high monetary value
- 1,5% of the customer base
- Where bad reviews are located
- More housing product orders in this cluster

QUESTION ?

OpenClassRooms: Project 3

A x e l F a v r e u l