P5 Customer segmentation e-commerce

Objectifs

understanding different types of clients

provide actionable information for the marketing team

analysing cluster stability

Actions

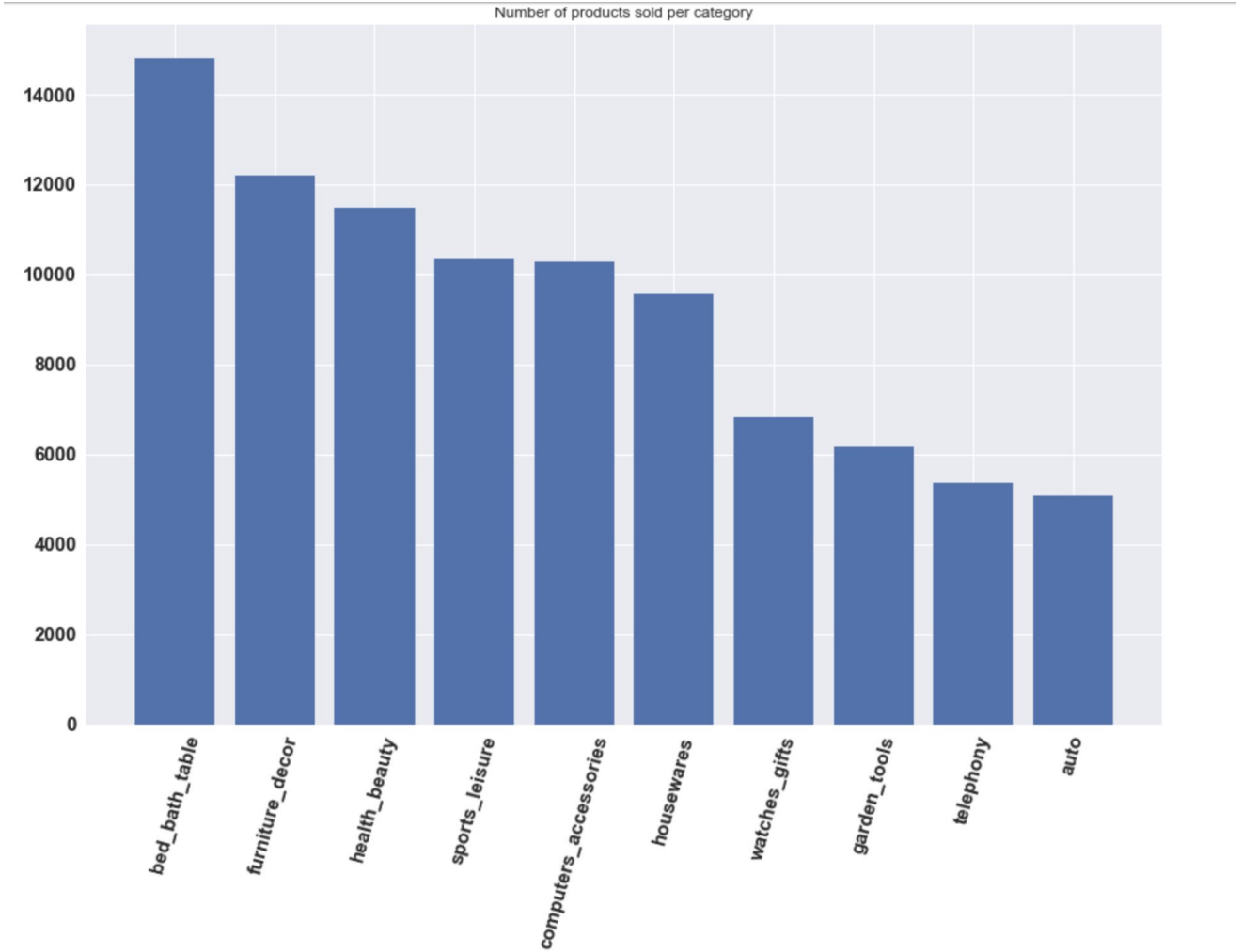
1. Cleaning and feature engineering
2. RFM
3. Kmeans
4. Analysis of cluster stability

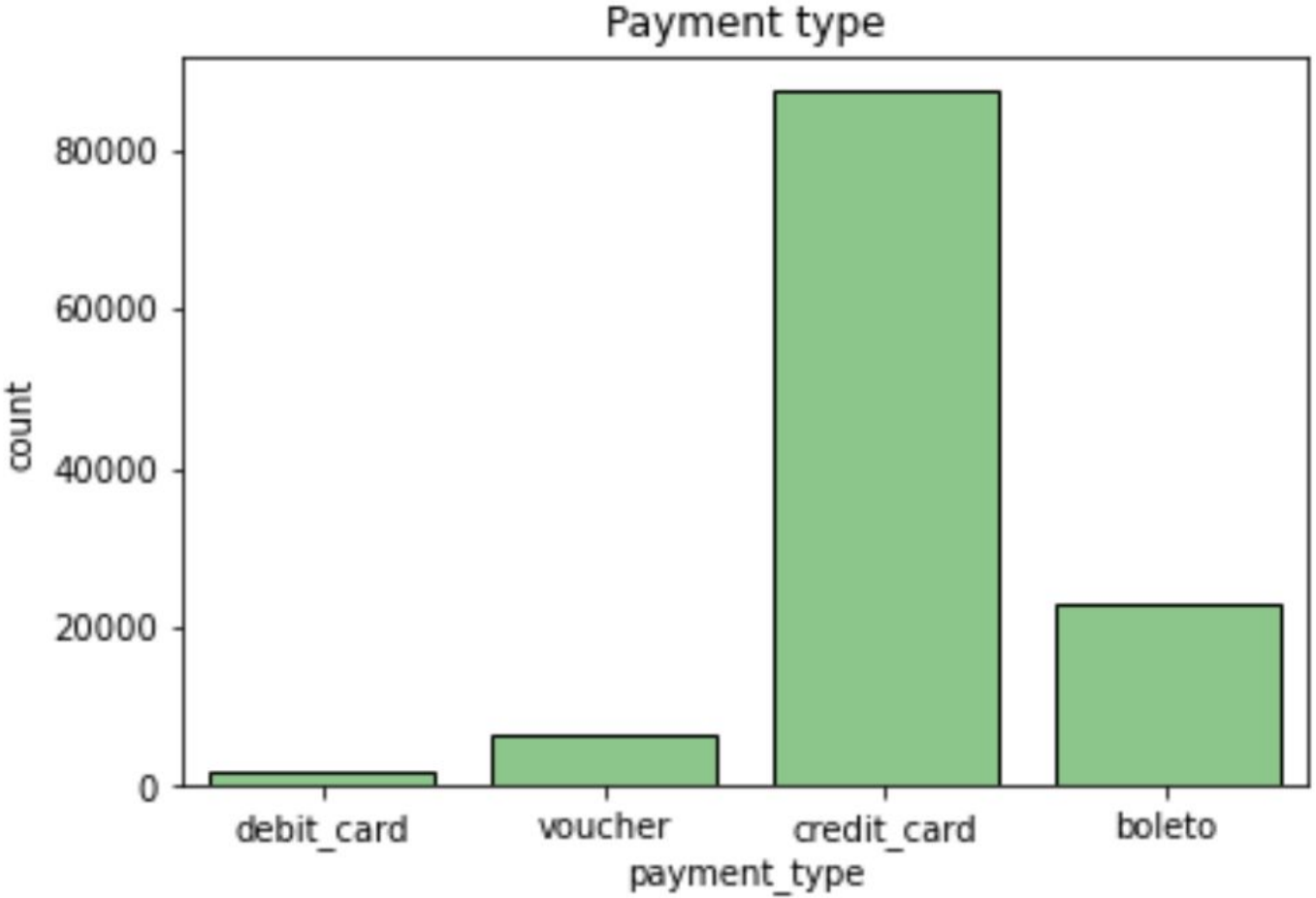
Cleaning and feature engineering

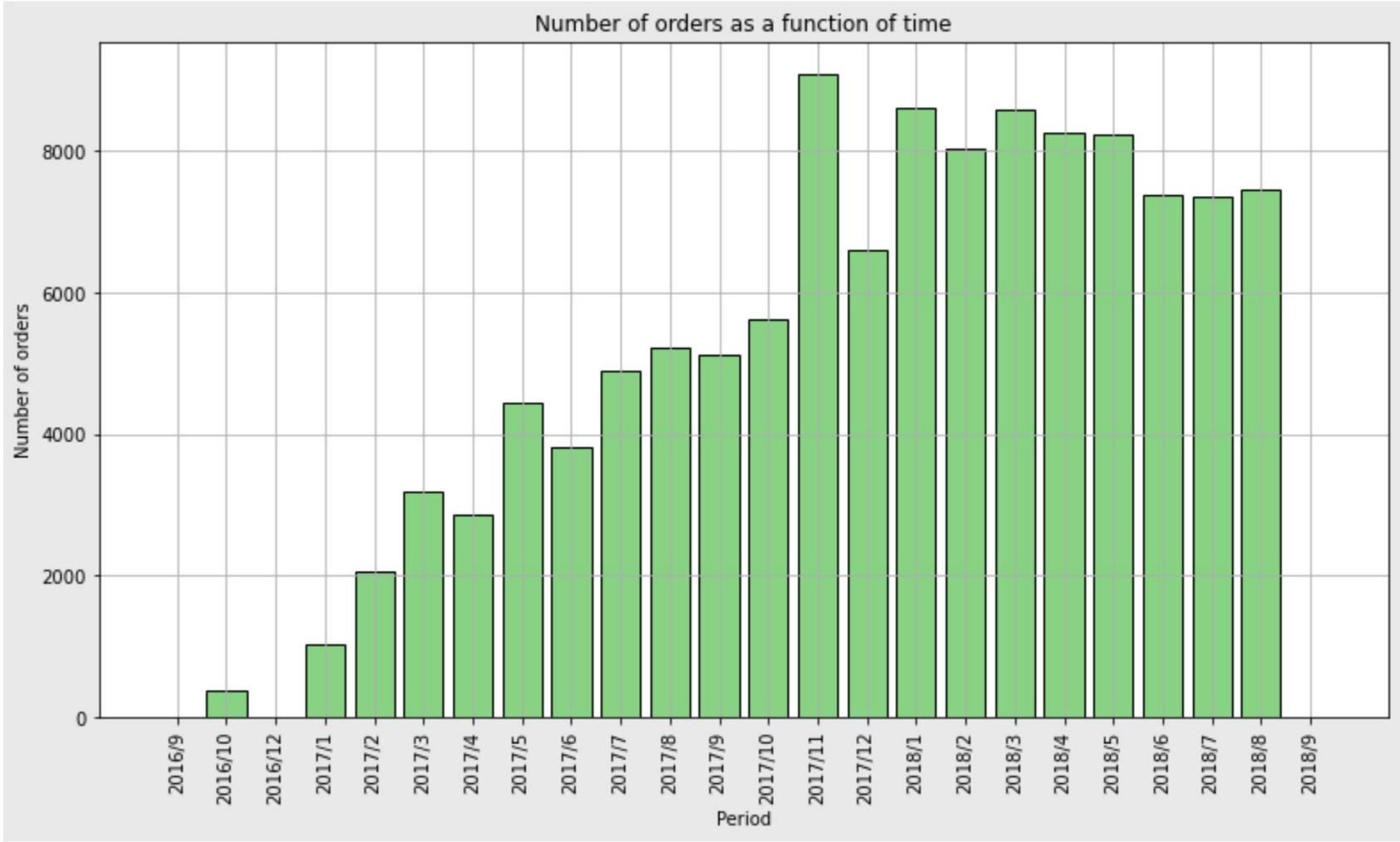
customer_unique_id	money	days	times	monetary	recency	frequency	summa	payment	state	ncategories	pcategory	review_score
0a0a92112bd4c708ca5fde585afaa872	109312.64	338	8	4	2	4	10	credit_card	RJ	1	fixed_telephony	1.0
698e1cf81d01a3d389d96145f7fa6df8	45256.00	375	20	4	1	4	9	credit_card	GO	1	auto	1.0
c402f431464c72e27330a67f7b94d4fb	44048.00	192	20	4	3	4	11	boleto	SP	1	computers_accessories	1.0
4007669dec559734d6f53e029e360987	36489.24	282	6	4	2	4	10	boleto	MG	1	agro_industry_and_commerce	1.0
ef8d54b3797ea4db1d63f0ced6a906e9	30186.00	136	10	4	3	4	11	boleto	RJ	1	drinks	5.0
...	...	...	...	...	...	...	...	...	...	...	...	...
6f5b9d1cdccc4d28f0483a612edecacf	11.63	365	1	1	1	1	3	credit_card	SP	1	baby	5.0
2878e5b88167faab17d4fb83a986d38b	11.63	308	1	1	2	1	4	credit_card	SP	1	baby	5.0
b33336f46234b24a613ad9064d13106d	10.89	73	1	1	4	1	6	credit_card	SP	1	auto	3.0
bd06ce0e06ad77a7f681f1a4960a3cc6	10.07	354	1	1	1	1	3	credit_card	SP	1	stationery	5.0
317cfc692e3f86c45c95697c61c853a6	9.59	8	2	1	4	2	7	credit_card	SP	1	health_beauty	5.0

95419 rows × 12 columns

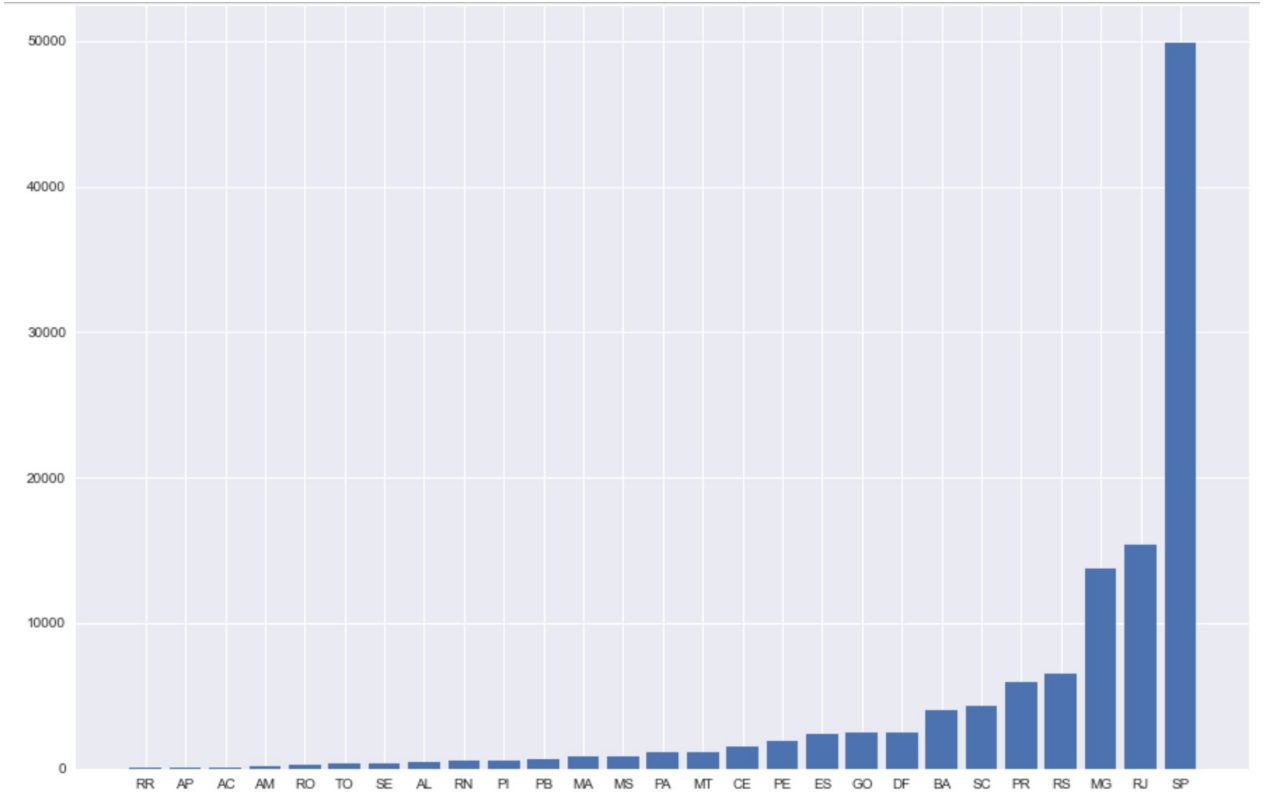
Favorite category



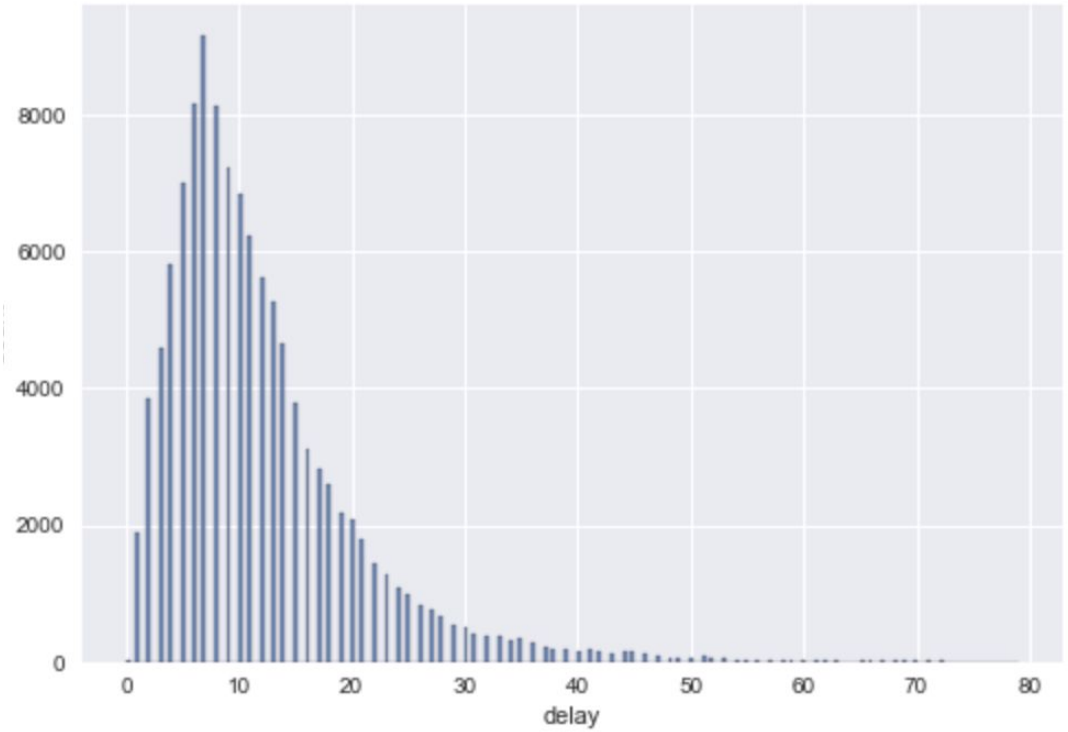

Payment type
80000
60000
40000
20000
0
count
debit_card
voucher
credit_card
boleto
payment_type


Number of orders as a function of time
Number of orders
0
2000
4000
6000
8000
2016/9
2016/10
2016/12
2017/1
2017/2
2017/3
2017/4
2017/5
2017/6
2017/7
2017/8
2017/9
2017/10
2017/11
2017/12
2018/1
2018/2
2018/3
2018/4
2018/5
2018/6
2018/7
2018/8
2018/9
Period

Customer by state

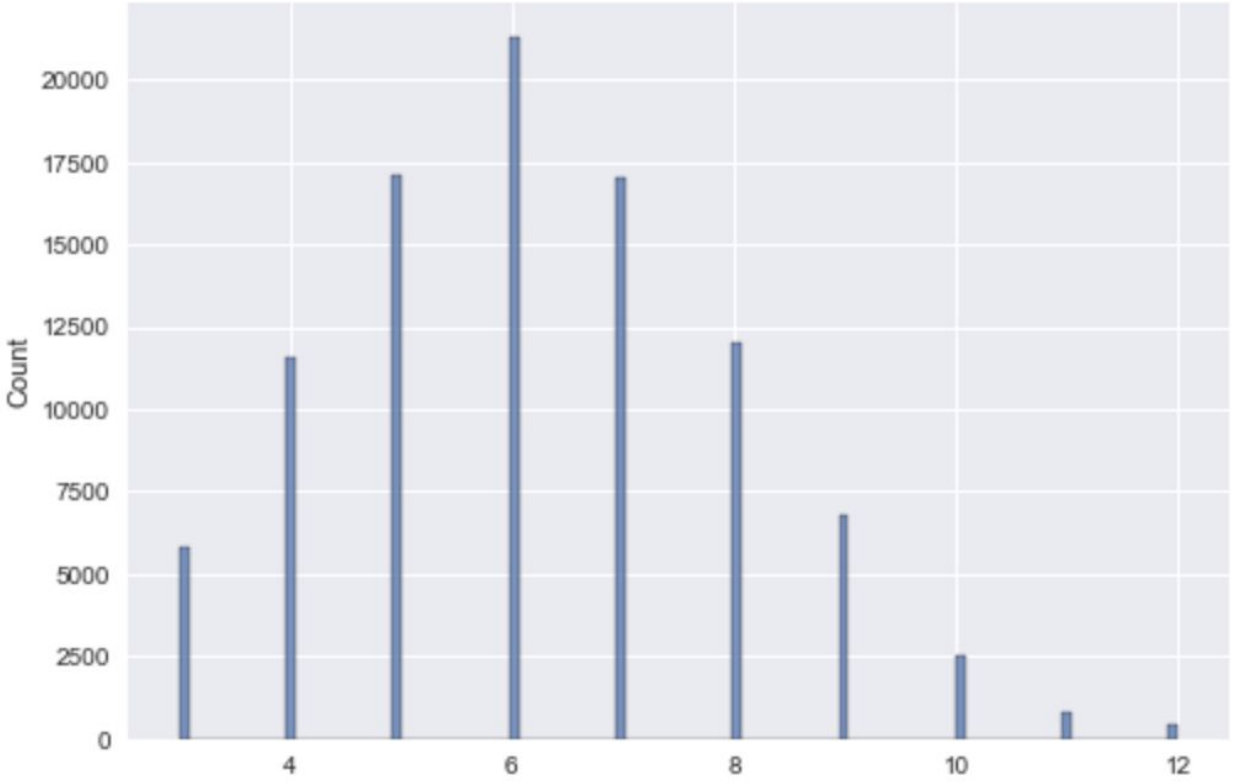


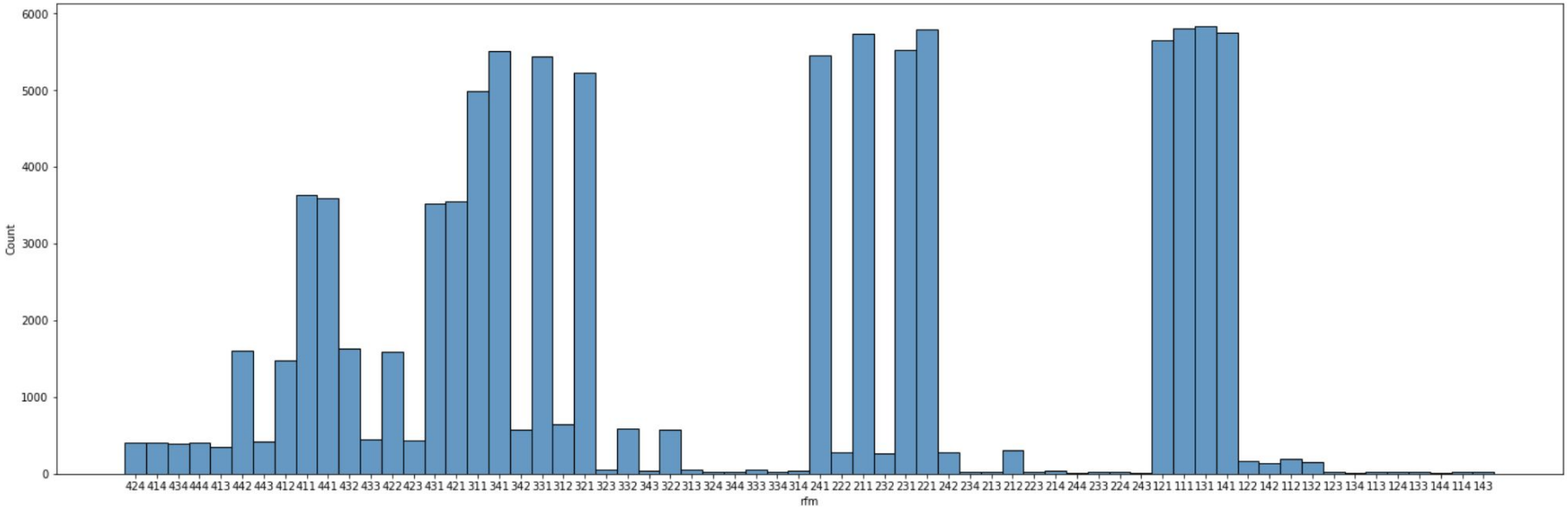
Delivery time



RFM

RFM sum


Count
20000
17500
15000
12500
10000
7500
5000
2500
0
4
6
8
10
12


Count
6000
5000
4000
3000
2000
1000
0
424 414 434 444 413 442 443 412 411 441 432 433 422 423 431 421 311 341 342 331 312 321 323 332 343 322 313 324 344 333 334 314 241 222 211 232 231 221 242 234 213 212 223 214 244 233 224 243 121 111 131 141 122 142 112 132 123 134 113 124 133 144 114 143
rfm

RFM groups

```
#1.Rich and recent
filter_VIP = df_trim[df_trim.rfm.isin(['444','434','424','414','344','334','324','314'])].index
df_trim.loc[filter_VIP,'class_RFM'] = 'VIP'
```

```
#2.Rich and not recent
filter_VIP_forgotten = df_trim[df_trim.rfm.isin(['244','234','224','214','343','333','323','313'])].index
df_trim.loc[filter_VIP_forgotten,'class_RFM'] = 'VIP not recent'
```

```
#3.Lost VIP
filter_VIP_lost = df_trim[df_trim.rfm.isin(['144','134','124','114','143','133','123','113'])].index
df_trim.loc[filter_VIP_lost,'class_RFM'] = 'VIP lost'
```

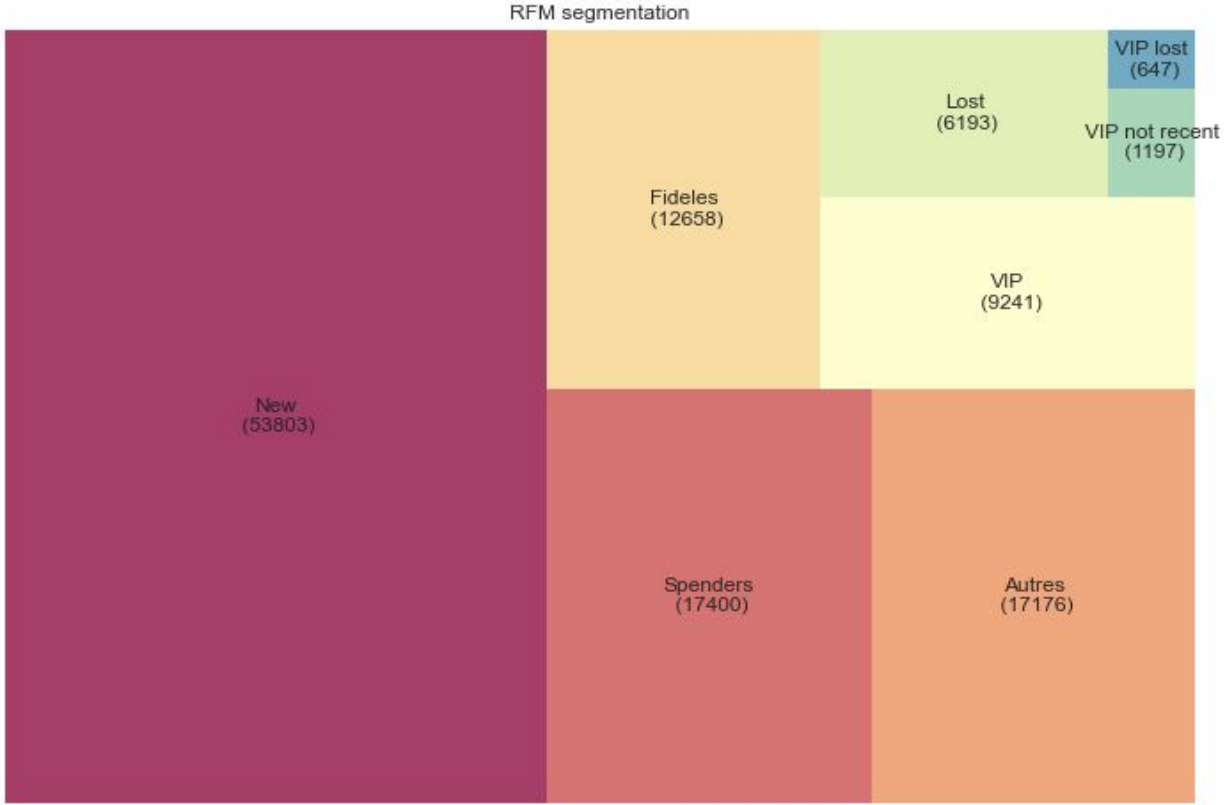
```
#4.New not VIP
filter_new = (df_trim.recency.isin([4,3])) & (df_trim.class_RFM.isna())
filter_new = df_trim[filter_new].index
df_trim.loc[filter_new,'class_RFM'] = 'New'
```

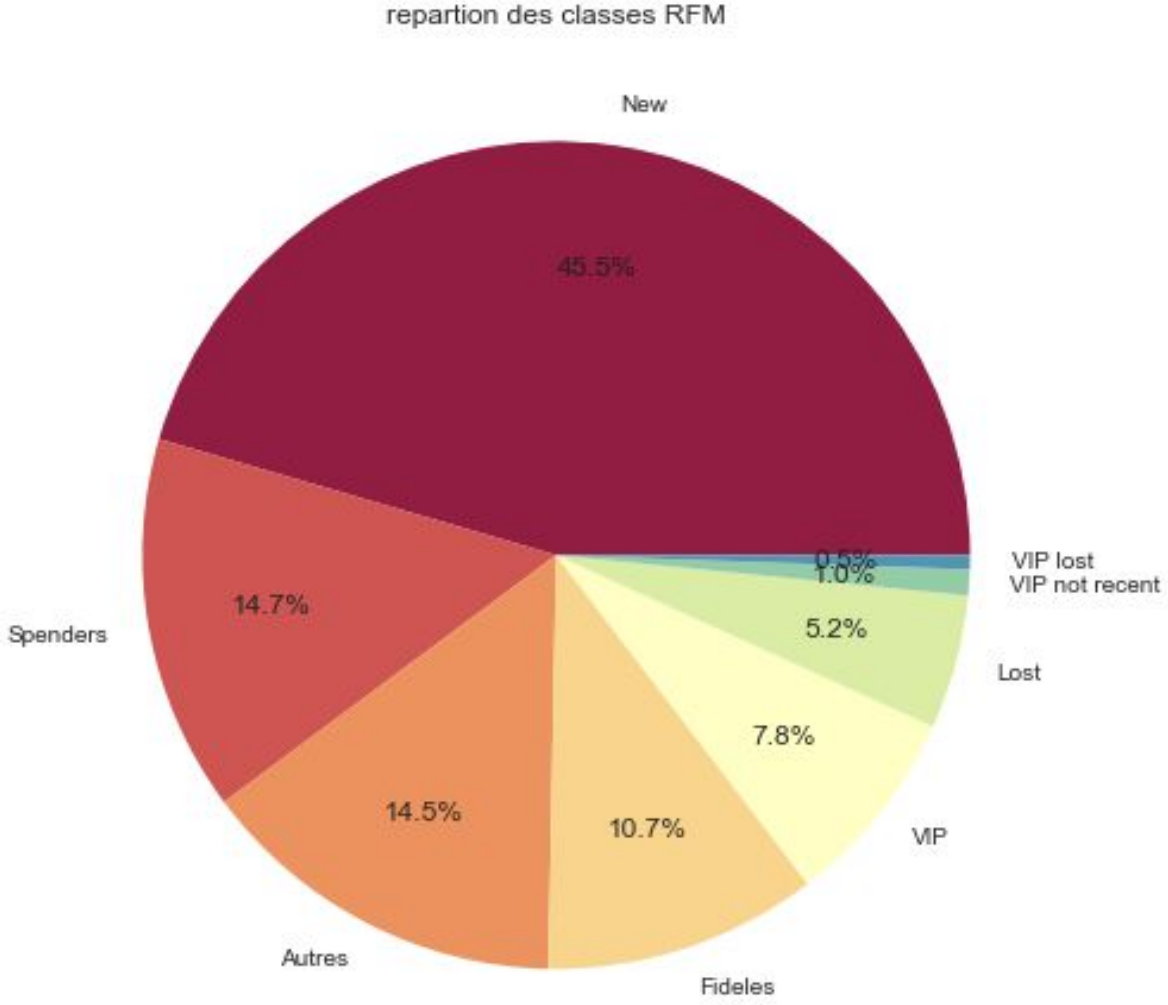
```
#5.More than once not new or VIP
filter_fidel = (df_trim.frequency!=1) & (df_trim.class_RFM.isna())
filter_fidel = df_trim[filter_fidel].index
df_trim.loc[filter_fidel,'class_RFM'] = 'Fideles'
```

```
#6.Spenders
filter_spenders = (df_trim.monetary.isin([4,3])) & (df_trim.class_RFM.isna())
filter_spenders = df_trim[filter_spenders].index
df_trim.loc[filter_spenders,'class_RFM'] = 'Spenders'
```

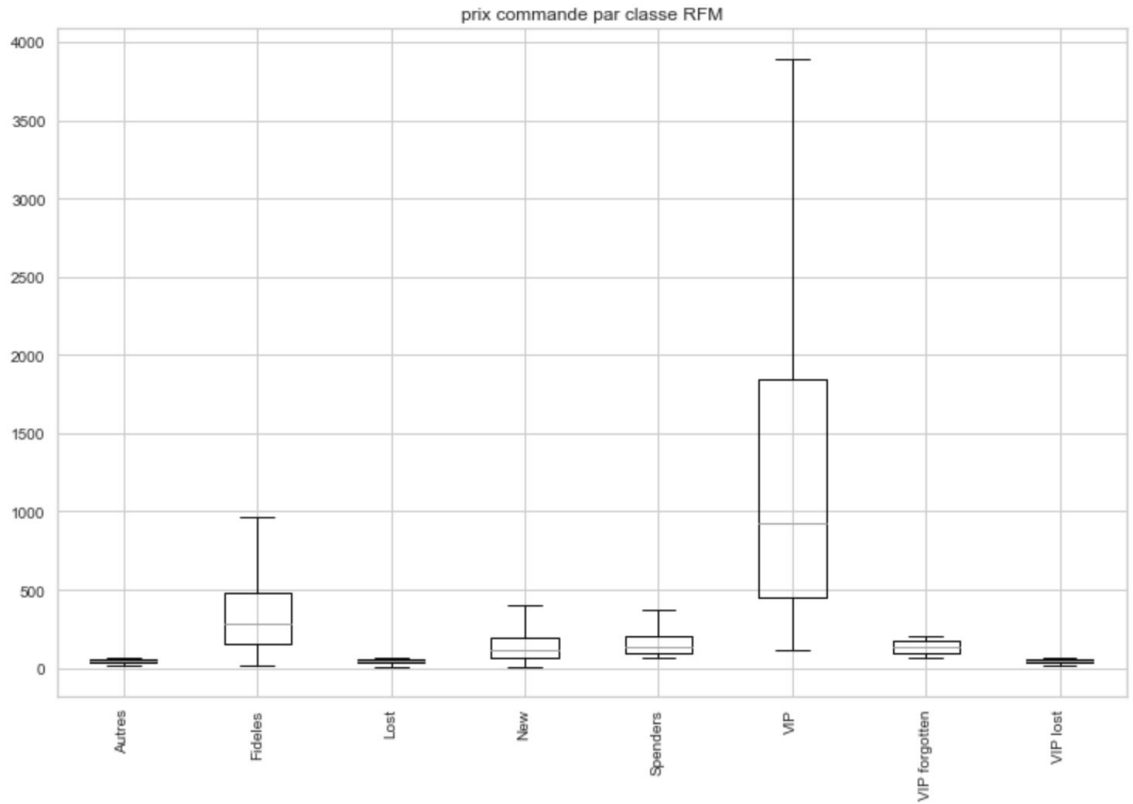
```
#7.No money and long gone and only once
filter_lost = df_trim[df_trim.rfm.isin(['111', '112'])].index
df_trim.loc[filter_lost,'class_RFM'] = 'Lost'
```

```
#8.Clients autres
df_trim.class_RFM.fillna(value='Autres', inplace=True)
```

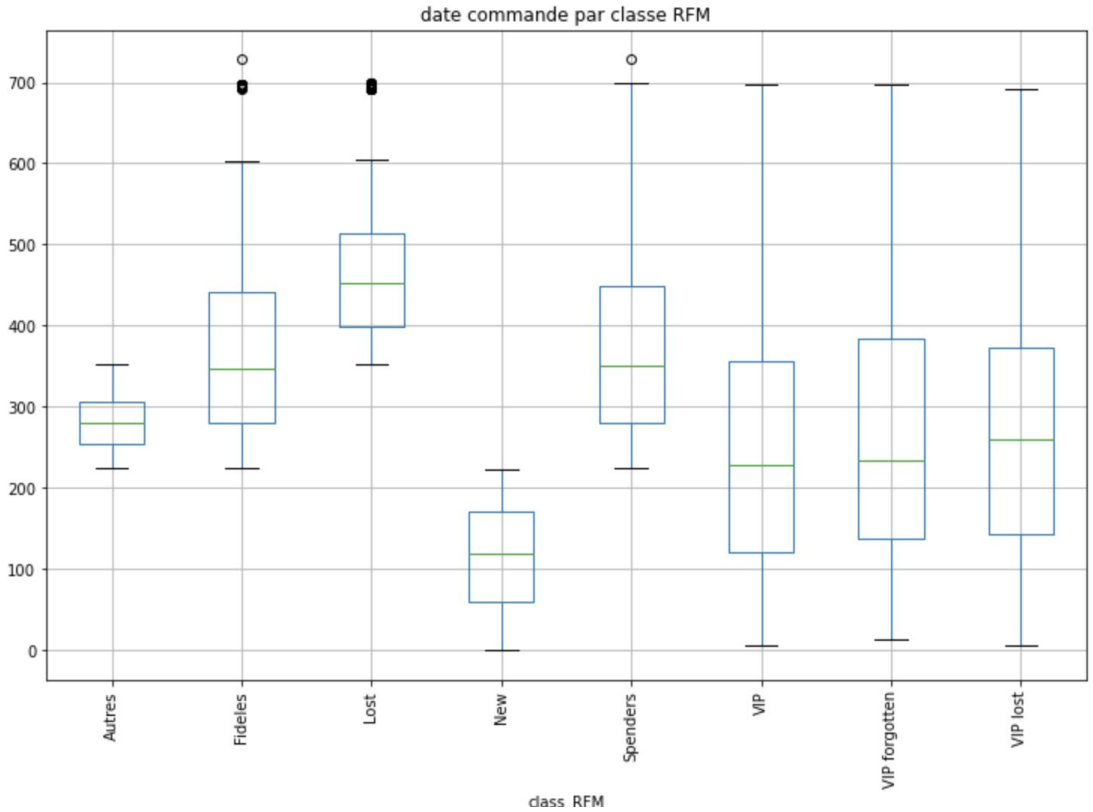

RFM segmentation
New
(53803)
Fideles
(12658)
Lost
(6193)
VIP lost
(647)
VIP not recent
(1197)
VIP
(9241)
Spenders
(17400)
Autres
(17176)


repartion des classes RFM
New
45.5%
VIP lost
0.5%
VIP not recent
1.0%
Lost
5.2%
VIP
7.8%
Fideles
10.7%
Autres
14.5%
Spenders
14.7%

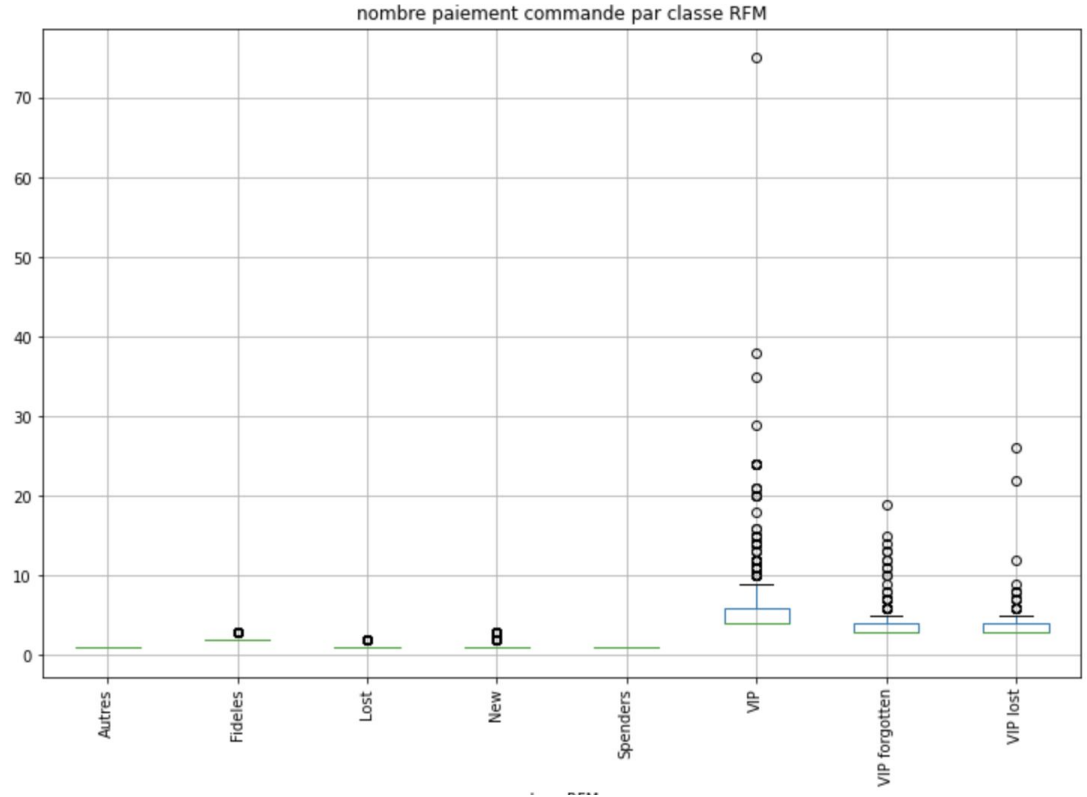
Monetary


prix commande par classe RFM
4000
3500
3000
2500
2000
1500
1000
500
0
Autres
Fideles
Lost
New
Spenders
VIP
VIP forgotten
VIP lost

Recency

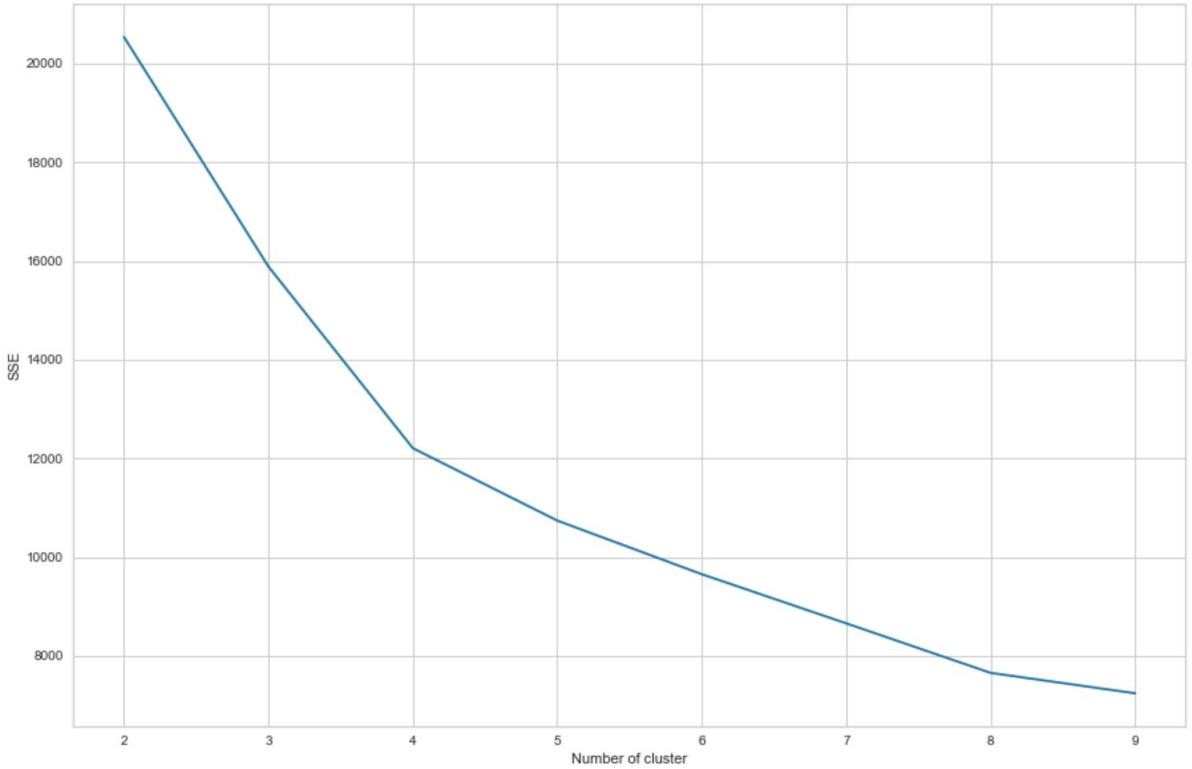

date commande par classe RFM
700
600
500
400
300
200
100
0
Autres
Fideles
Lost
New
Spenders
VIP
VIP forgotten
VIP lost
class_RFM

Frequency

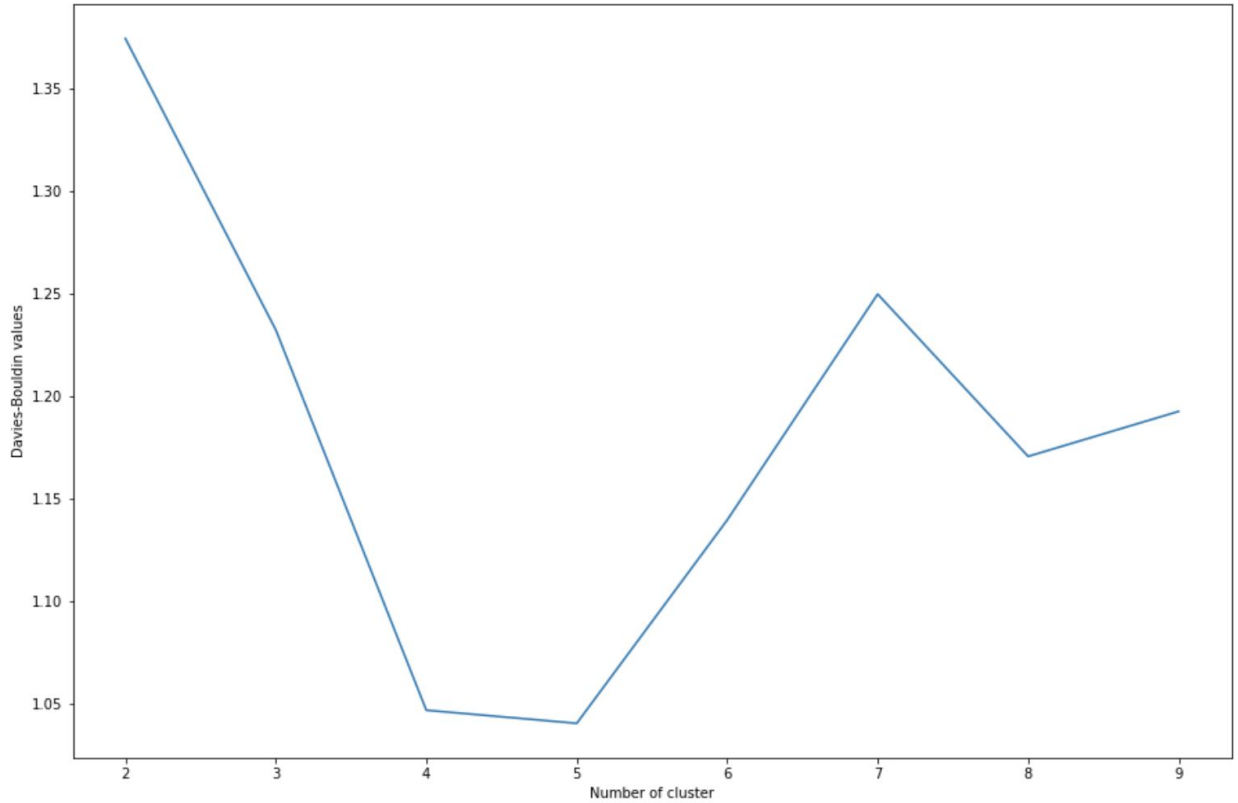

nombre paiement commande par classe RFM
70
60
50
40
30
20
10
0
Autres
Fideles
Lost
New
Spenders
VIP
VIP forgotten
VIP lost

Kmeans

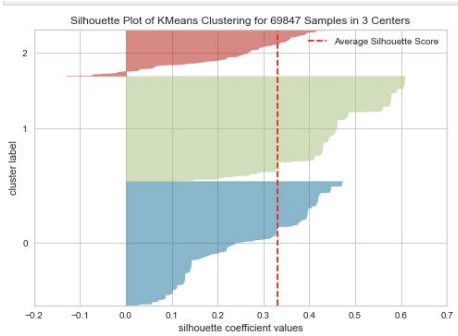
Inertia

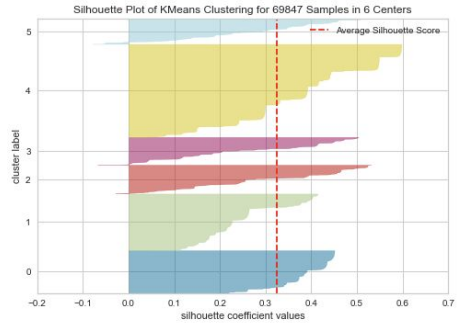

20000
18000
16000
14000
12000
10000
8000
SSE
2
3
4
5
6
7
8
9
Number of cluster

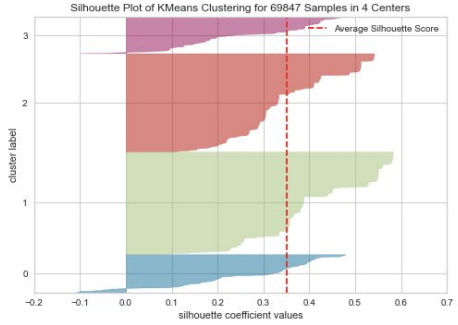
David Boudin values

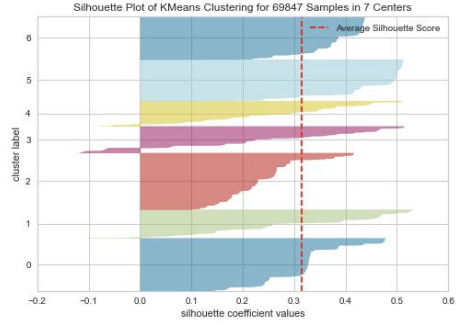


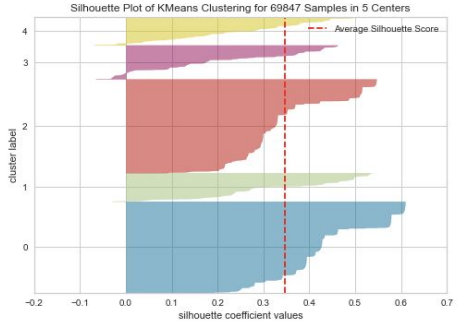
Siluette plot

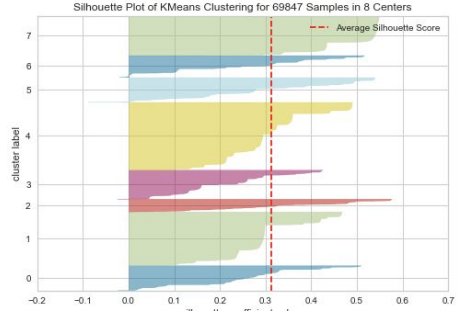






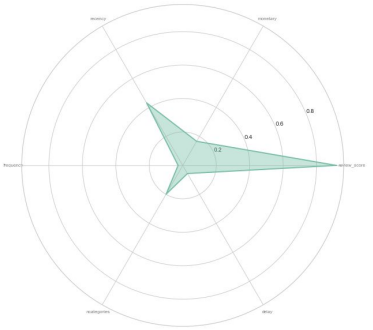






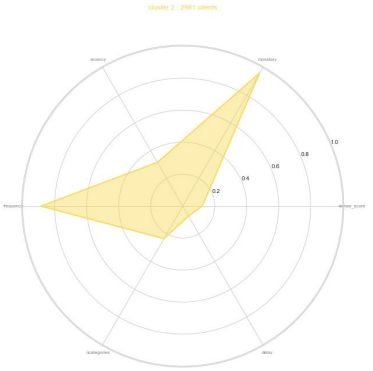
Clusters

1



2

3



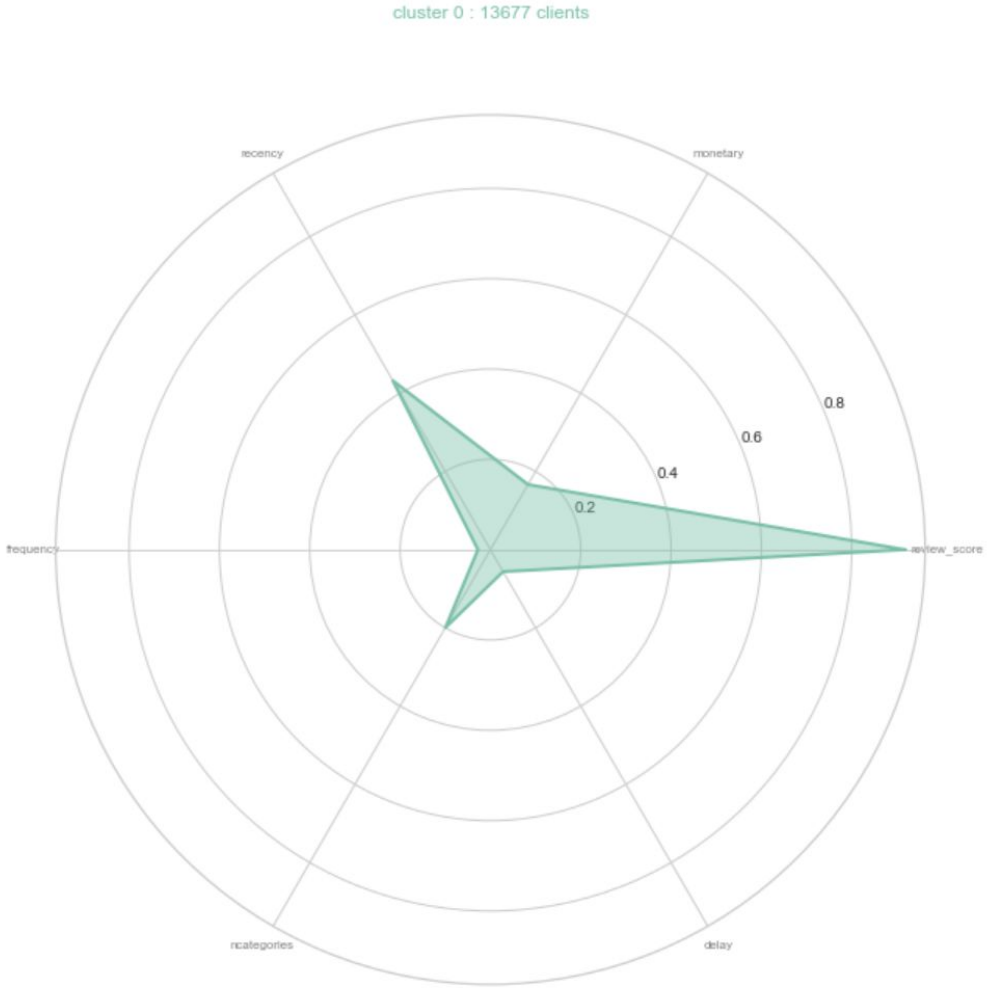
4

```
cluster 0 contenant 13677
review_score    0.918047
monetary        0.166484
recency         0.432161
frequency       0.027613
ncategories     0.199093
delay           0.056154
dtype: float64
cluster 1 contenant 10914
review_score    0.894436
monetary        0.835257
recency         0.000000
frequency       0.105553
ncategories     0.203408
delay           0.055535
dtype: float64
cluster 2 contenant 2981
review_score    0.121916
monetary        0.961758
recency         0.316449
frequency       0.885944
ncategories     0.234552
delay           0.072223
dtype: float64
cluster 3 contenant 4507
review_score    0.174081
monetary        0.174469
recency         0.297167
frequency       0.037053
ncategories     0.197382
delay           0.092766
dtype: float64
```

CLUSTER 1

New customer, who

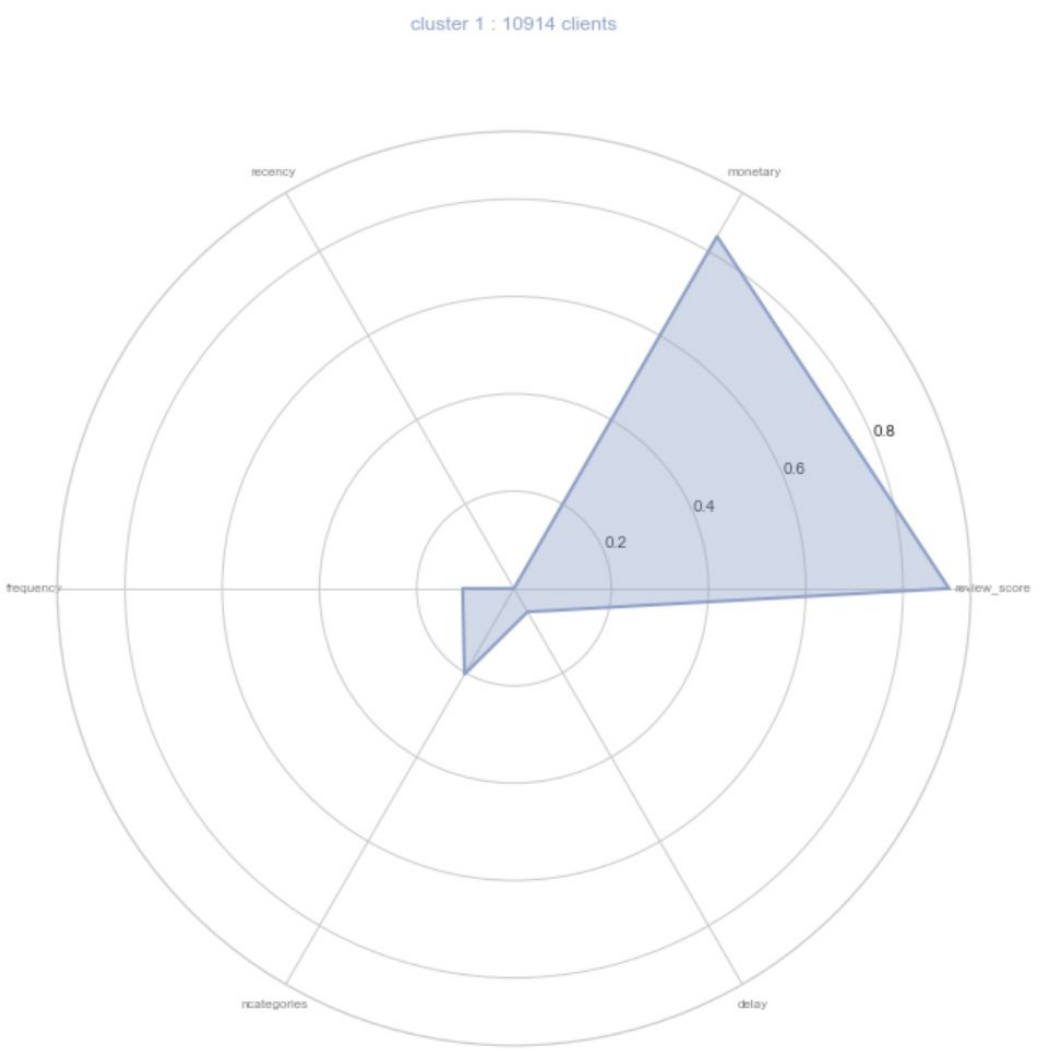
- liked our products
- have not spent a lot yet



CLUSTER 2

Client we cannot afford-to-lose:

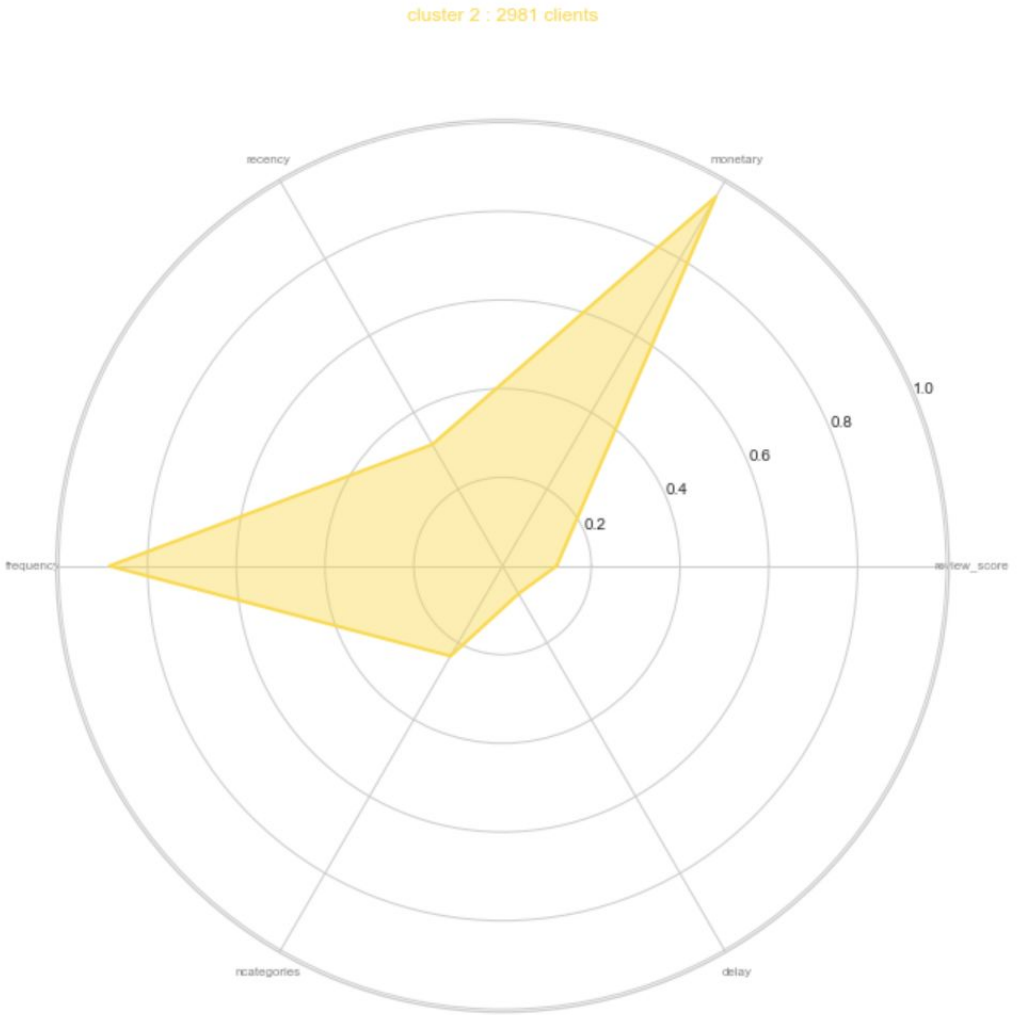
-spent a lot of money
-liked our products
-have not bought recently



CLUSTER 3

Demanding client

-spent money
-bought often
-bad review score
-delay 2nd longest

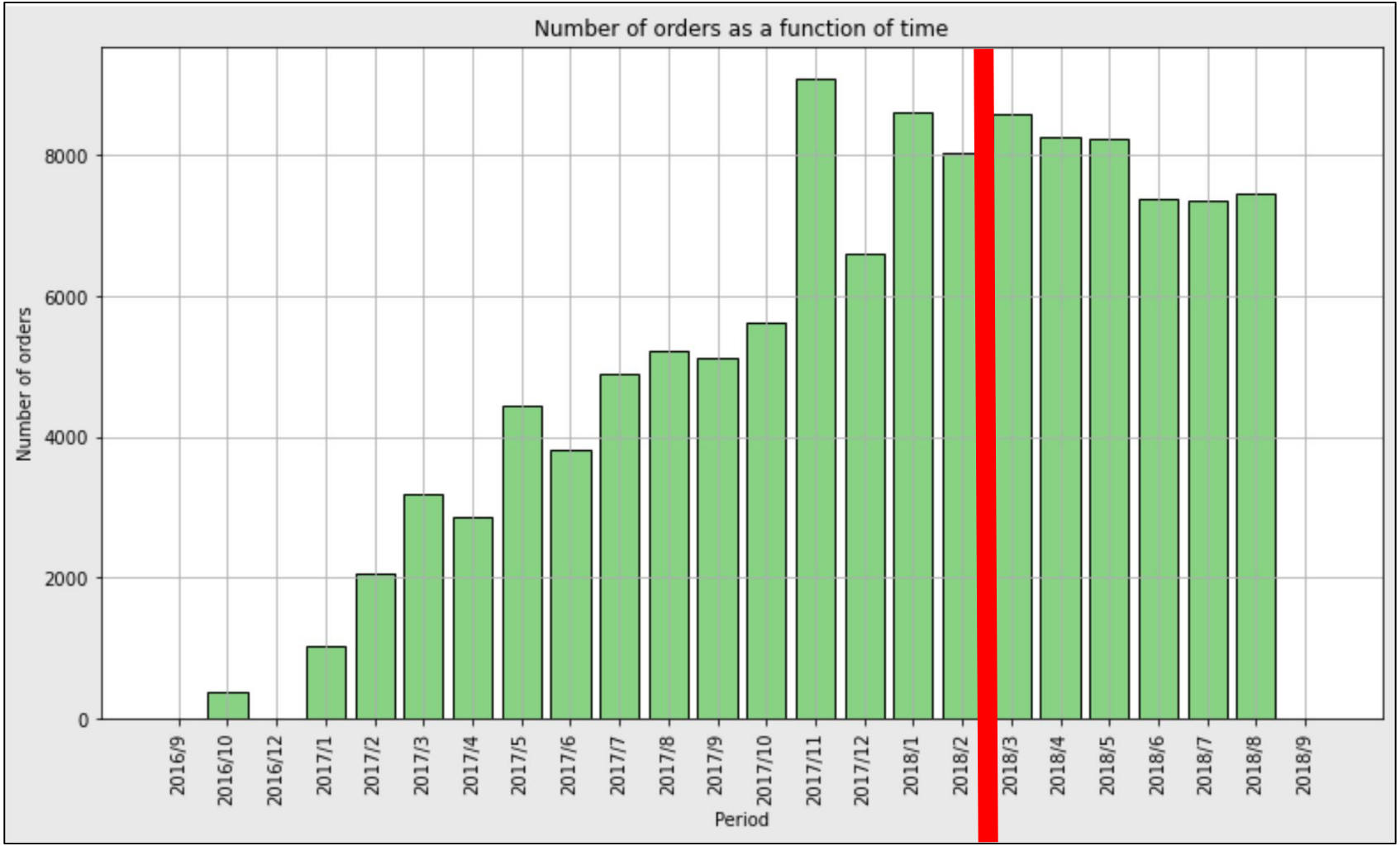


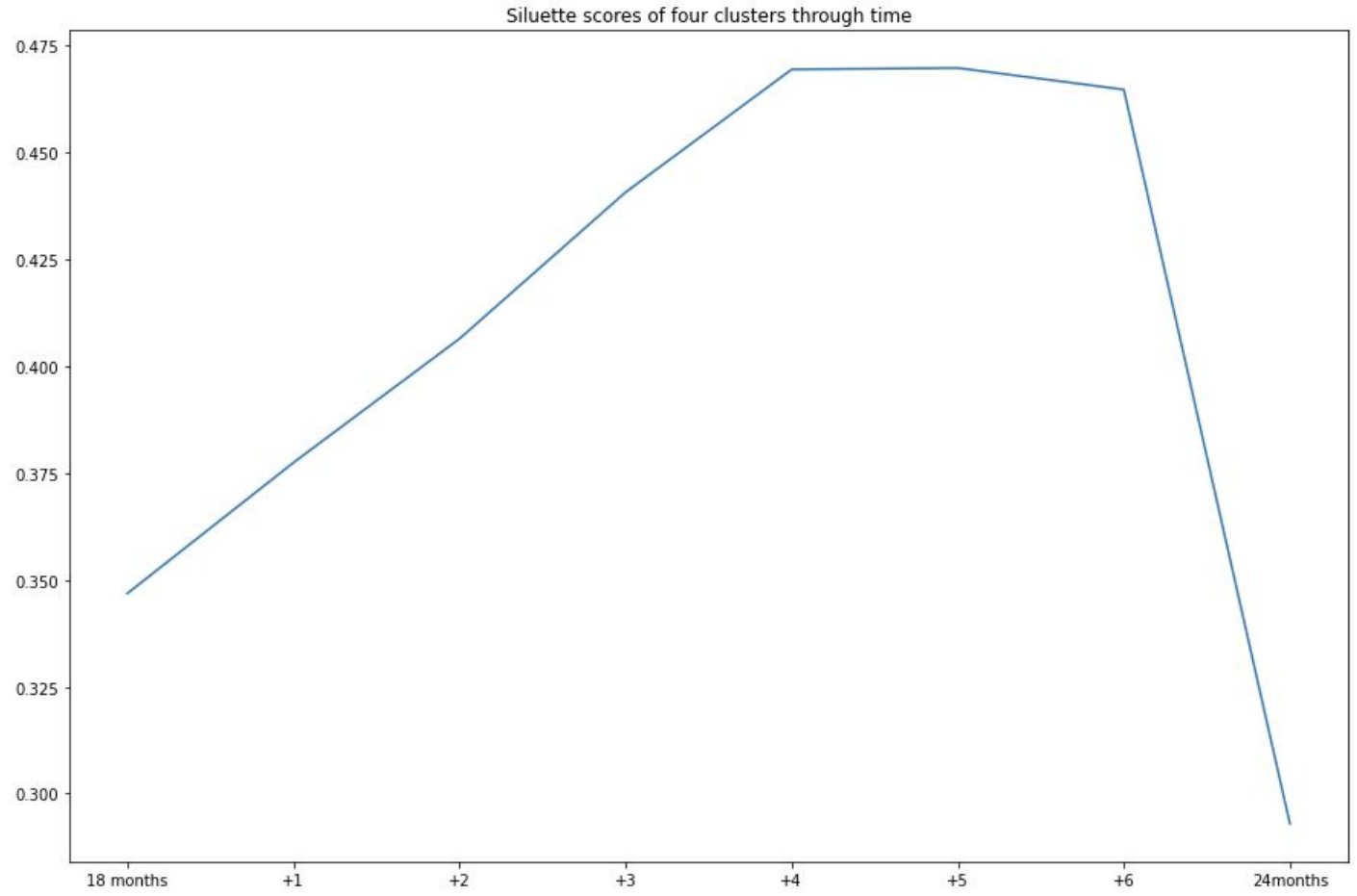
CLUSTER 4

Can’t wait so long?

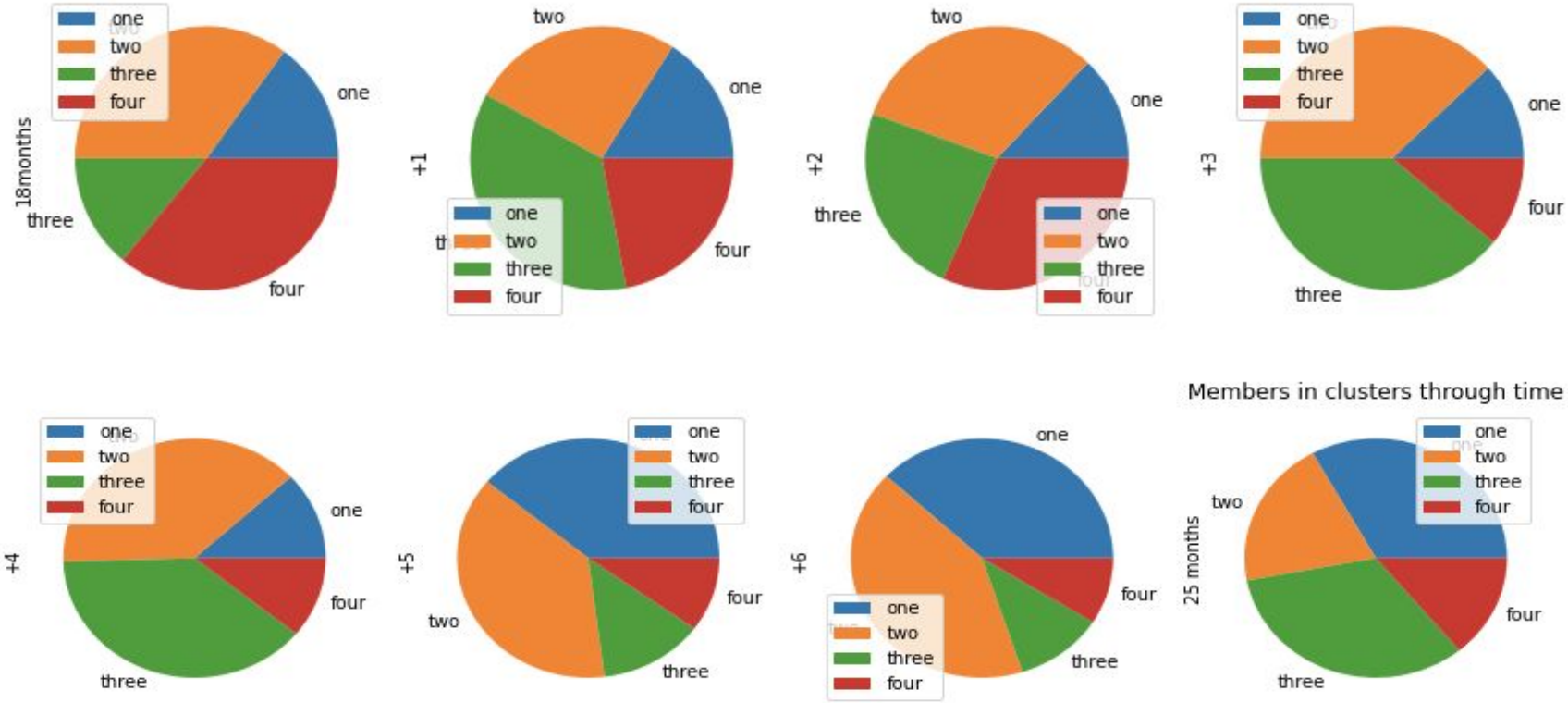
-longest delivery time
-poor review score
-recent but have not spent too much or often

Cluster stability


Number of orders as a function of time
Number of orders
0
2000
4000
6000
8000
2016/9
2016/10
2016/12
2017/1
2017/2
2017/3
2017/4
2017/5
2017/6
2017/7
2017/8
2017/9
2017/10
2017/11
2017/12
2018/1
2018/2
2018/3
2018/4
2018/5
2018/6
2018/7
2018/8
2018/9
Period


Siluette scores of four clusters through time
0.475
0.450
0.425
0.400
0.375
0.350
0.325
0.300
18 months
+1
+2
+3
+4
+5
+6
24months

Cluster members through time



Conclusion

Increasing number of customers, only few return
Opportunities for marketing

Lots of happy customers
Delivery time needs to be shortened