Advanced Data analysis

Website sales analysis

Is it possible to get insights about customers to help the company to improve the sales strategy?

Fabien Calderini

Content

1) Quick words about the company

Payot Libraire, first shop in Lausanne, opened in 1877

>3 M
books sold in 2023

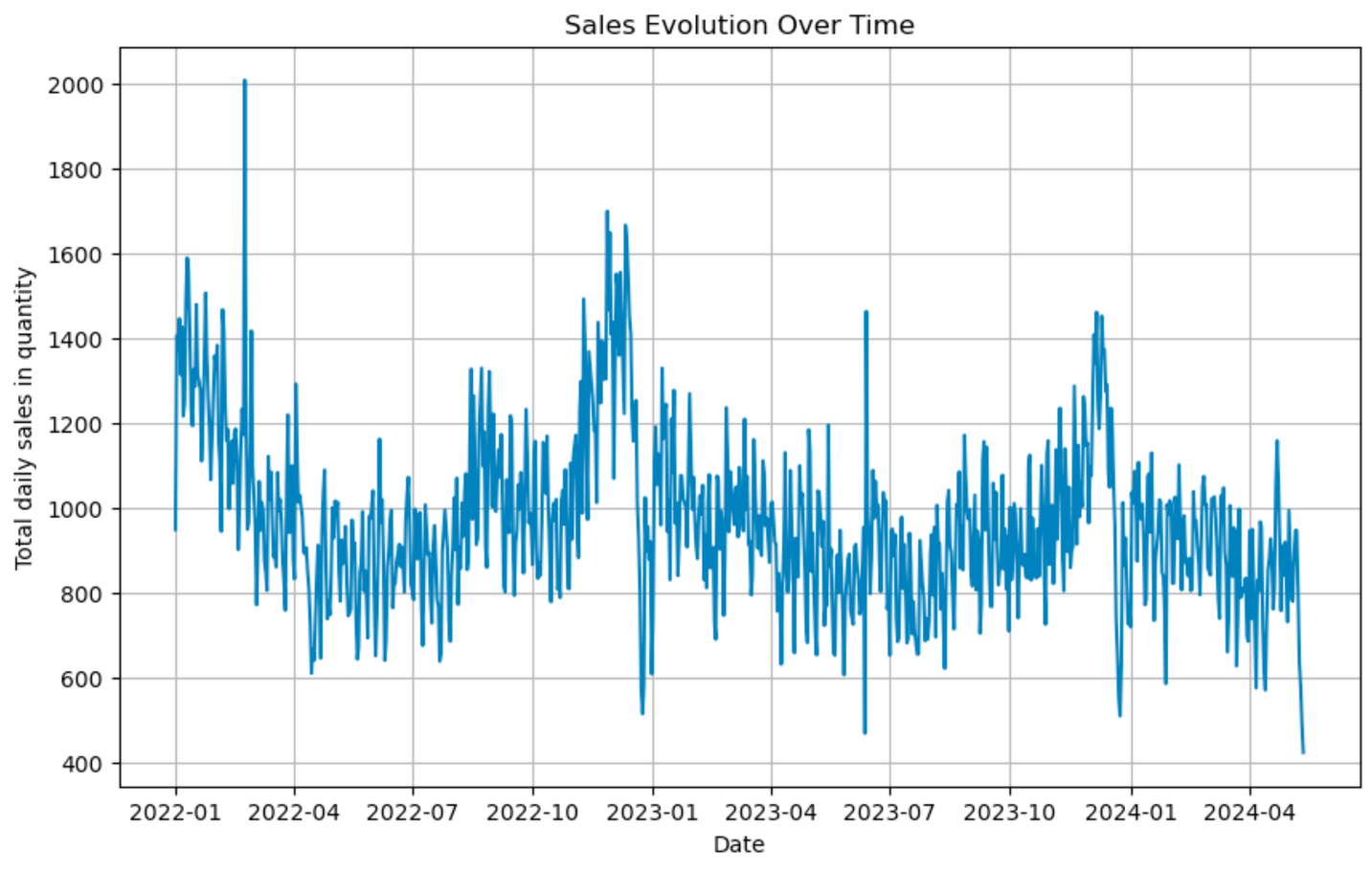
14
Shops

180 K
website orders in 2023

>10 M
Books available on website

1) The problem

With around 1000 books ordered per day and 10 millions books available, how to get insight about customers to improve our service ?



1) Datasets pre-treatment

Some data needed to be deleted or anonymized for privacy concerns

- Delete private informations such as street adresses

- Anonymization for the client email adress in the first dataframe and client number in the second dataframe to be able to make statistic per client while keeping everything private

1) Datasets

First dataset: Book sold on website between 2022 and 2024-03

	ean13	titre	auteur	editeur	quantity	code_rayon	unit_normal_pricettc	client_bill_email	order_number	order_id	creation_date	client_bill_comp_name	Format	Language	Nom rayon	Famille
0	9782848931463	Et il ne restera que poussière	Cornwell Patricia	Editions des Deux Terres	1	NaN	9.0	P 1	2705930	4634620	2024-01-01	NaN	ePub	Français	NaN	Unknown
1	9782253174332	Cadavre X	Cornwell Patricia	Le Livre de Poche	1	NaN	9.0	P 1	2705930	4634620	2024-01-01	NaN	ePub	Français	NaN	Unknown
2	9782290249949	La part des anges	Combes Bruno	J'ai Lu	1	NaN	13.4	P 2	2705931	4634621	2024-01-01	NaN	NaN	Français	LIT. FRANCOPHONE POCHE	LITTERATURE
3	9782221116081	San-Antonio	San-Antonio	Robert Laffont	1	NaN	44.3	P 3	2705932	4634622	2024-01-01	NaN	NaN	Français	POLICIER GF	LITTERATURE
4	9791034763160	Pépin et Olivia	Jourdy Camille	Editions Dupuis	1	NaN	30.2	P 4	2705933	4634623	2024-01-01	NaN	NaN	Français	BD JEUNESSE	BD
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
928979	9782212571592	Au risque d'être soi	Brousse Myriam	Eyrolles	1	NaN	23.9	P 5012	2525201	4404138	2022-12-31	NaN	NaN	Français	SANTE GENERALITES	LOISIRS PRATIQUE
928980	9782501135665	Votre corps a une mémoire	Brousse Myriam	Marabout	1	NaN	11.7	P 5012	2525201	4404138	2022-12-31	NaN	NaN	Français	SANTE GENERALITES	LOISIRS PRATIQUE
928981	9781783788217	Dangers of Smoking in Bed	Enriquez Mariana	Granta Publications	1	NaN	16.5	P 128520	2525202	4404139	2022-12-31	NaN	NaN	Anglais	FICTION POCKET	LANGUES EN VO
928982	9782070323517	Les faits et les mythes	Beauvoir Simone de	Editions Gallimard	1	NaN	18.5	P 128520	2525202	4404139	2022-12-31	NaN	NaN	Français	SOCIO SOCIALISATION	SCIENCES HUMAINES
928983	9780141181875	Ada or Ardor	Nabokov Vladimir	Penguin Books	1	NaN	19.9	P 128520	2525202	4404139	2022-12-31	NaN	NaN	Anglais	FICTION POCKET	LANGUES EN VO

814148 rows × 16 columns

client_bill_email have been modified for privacy reasons

Second dataset: Orders on website between 2022 and 2024-03

	order_number	total_pricettc	creation_date	delivery	quantity	totalttc_ligne	client_inst	store	payment_mode	order_source	city	Ligne synthèse	client_number
0	2331079	46.5	2022-01-01	dm_economy	1.0	46.5	0	CTCP OLF	pm_bill	mobile	Estavayer le Lac	1	733001
1	2331080	60.6	2022-01-01	dm_economy	1.0	28.8	0	CTCP OLF	pm_bill	www	Gorgier	0	698205
2	2331080	60.6	2022-01-01	dm_economy	1.0	31.8	0	CTCP OLF	pm_bill	www	Gorgier	1	698205
3	2331081	13.1	2022-01-01	dm_economy	1.0	13.1	0	CTCP OLF	pm_bill	mobile	Neuchâtel	1	362003
4	2331082	38.4	2022-01-01	dm_priority	1.0	29.4	0	CTCP OLF	pm_bill	mobile	Chigny	1	613851
...	...	...	...	...	...	...	...	...	...	...	...	...	...
893437	2749744	34.1	2024-03-25	dm_economy	1.0	34.1	0	CTCP OLF	pm_bill	mobile	Genève	1	356791
893438	2749745	21.3	2024-03-25	dm_shop	1.0	21.3	0	Lausanne Pépinet	pm_payInShop	www	Vallorbe	1	836651
893439	2749746	38.0	2024-03-25	dm_economy	1.0	19.0	0	Lausanne Pépinet	pm_creditcard	www	Le Bry	0	546323
893440	2749746	38.0	2024-03-25	dm_economy	1.0	10.0	0	Lausanne Pépinet	pm_creditcard	www	Le Bry	0	546323
893441	2749746	38.0	2024-03-25	dm_economy	1.0	9.0	0	Lausanne Pépinet	pm_creditcard	www	Le Bry	1	546323

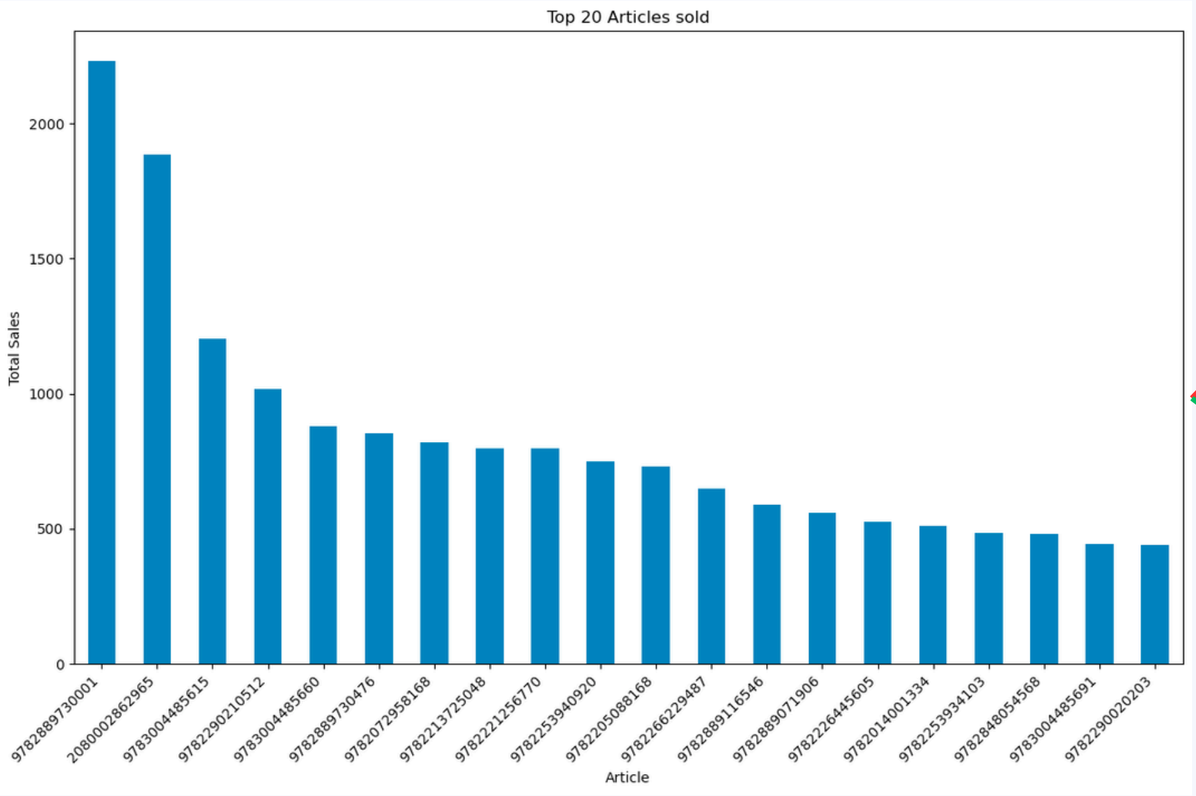
893442 rows × 13 columns

A table of correspondence for client number has been used for privacy reasons

2) EDA

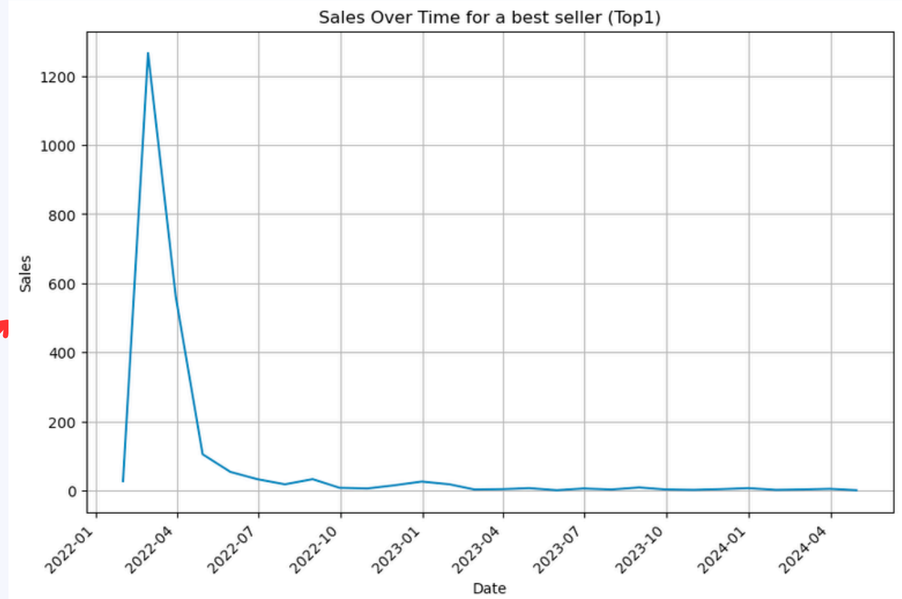
First dataset used: Book sold on website

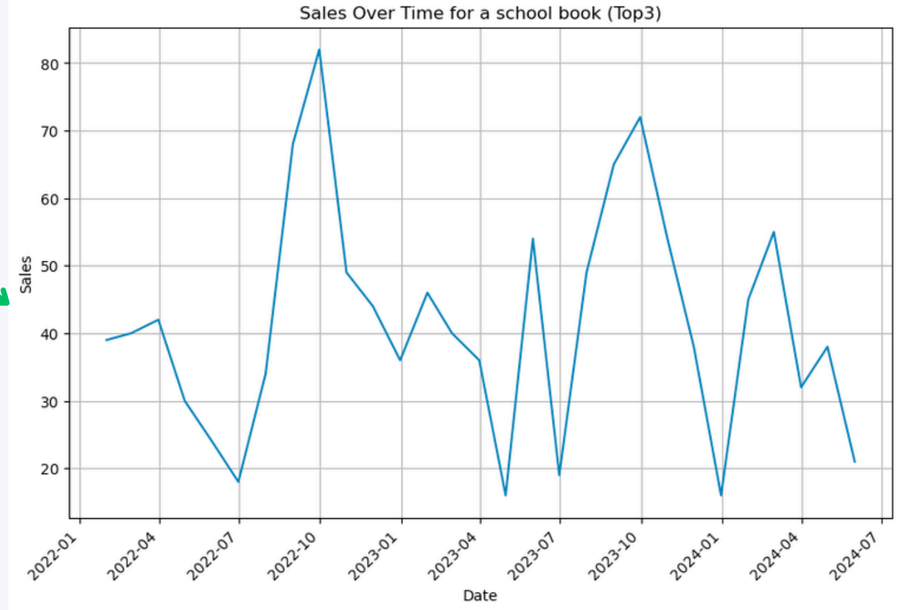
In the company, it is well known a small portion of books make the majority of the total sales, but the top 20 has different type of books:



Best seller

School book

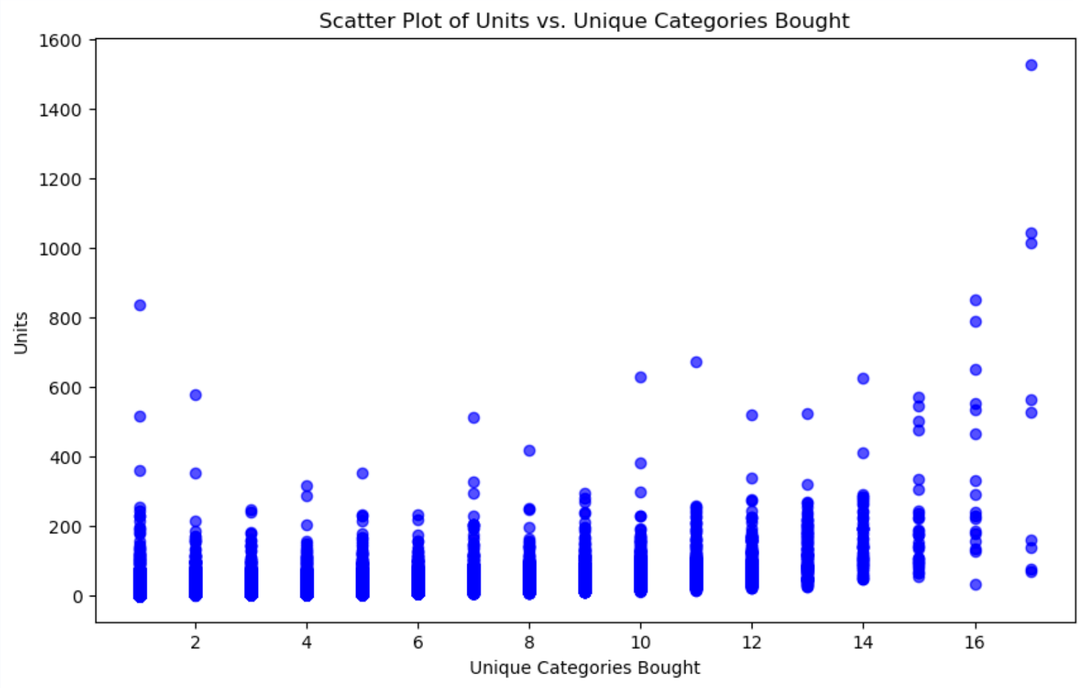


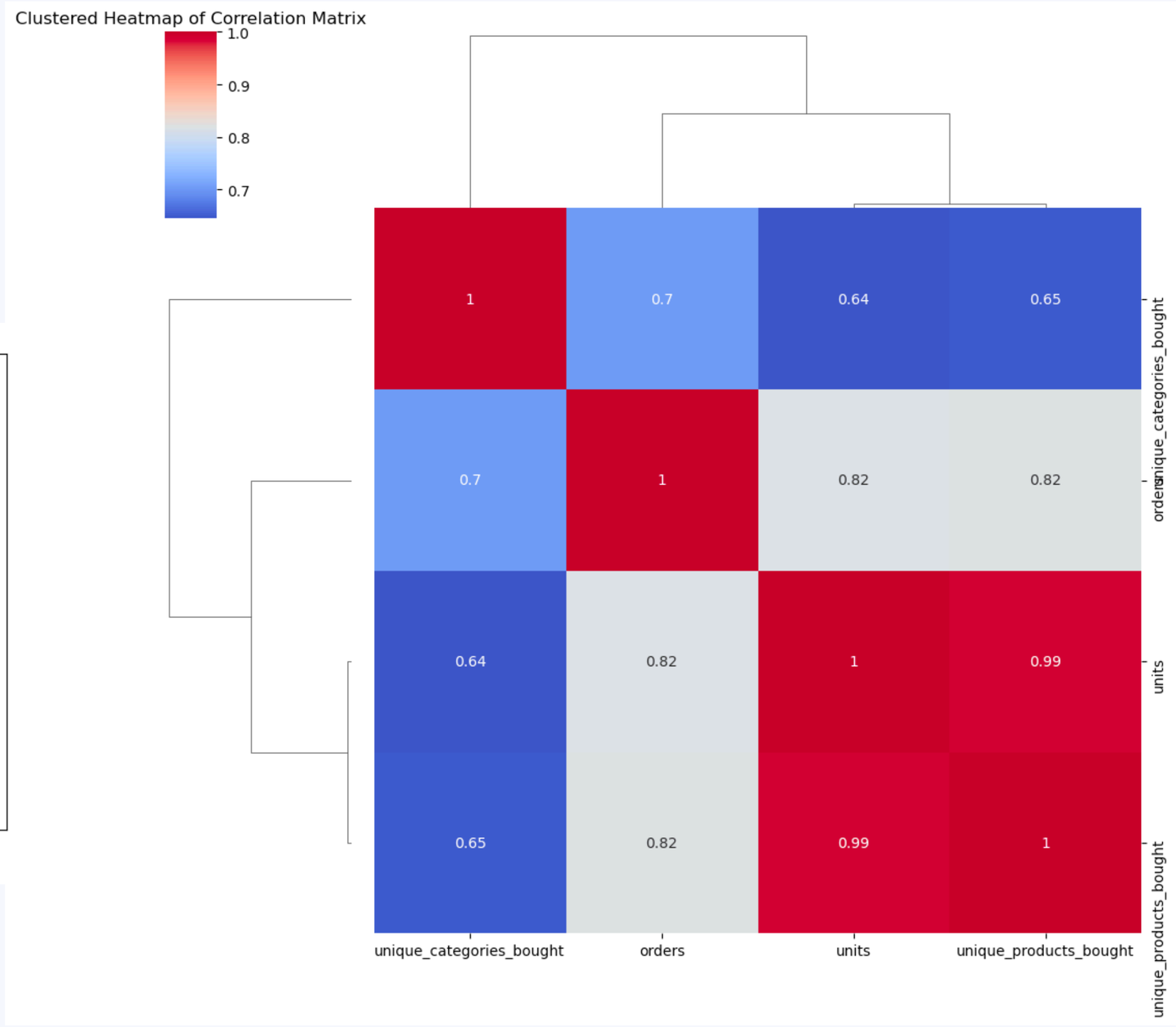


2) EDA

First dataset used: Book sold on website

Is there a link between the number of categories of books bought and the number of books bought?

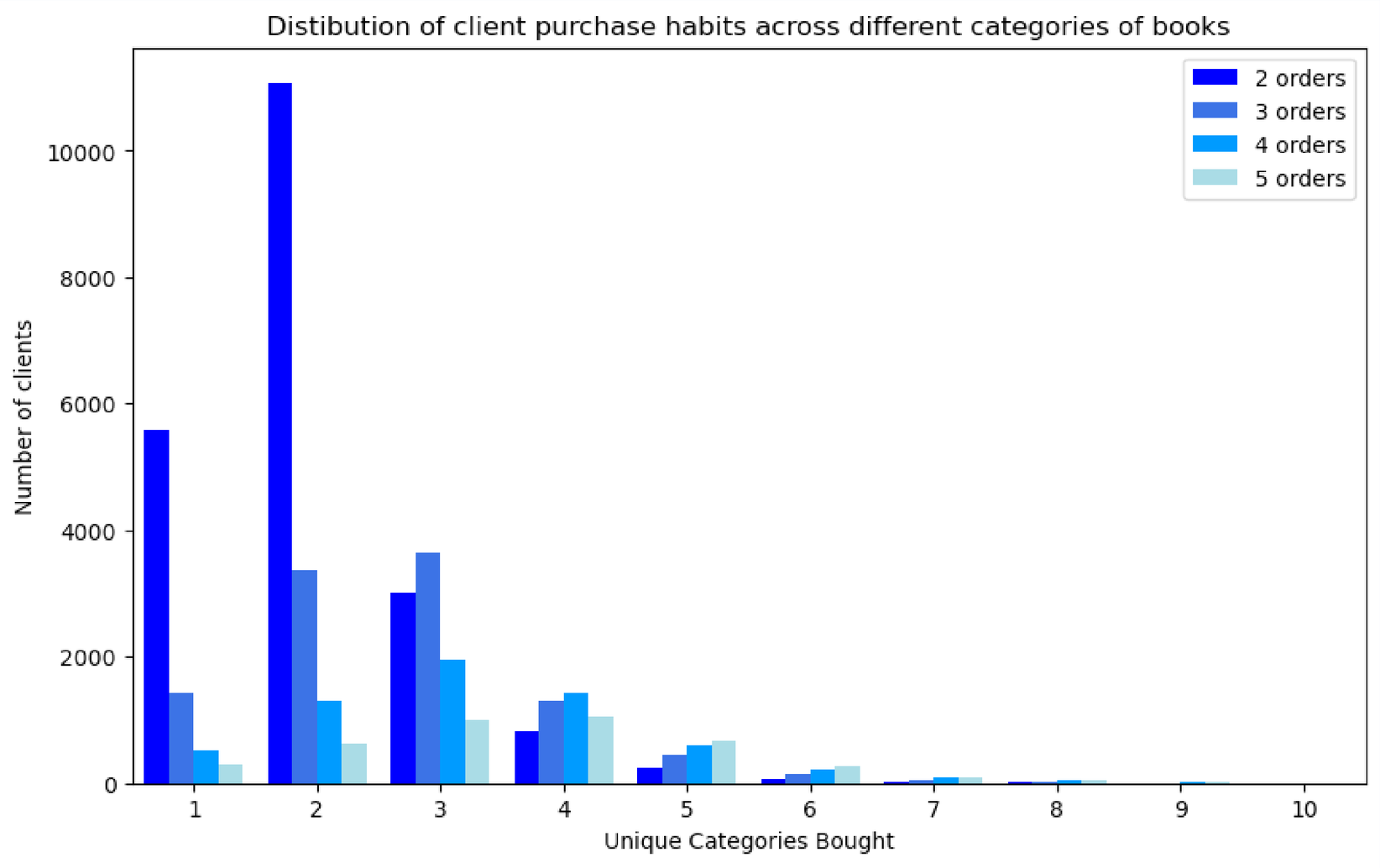




2) EDA

First dataset used: Book sold on website

How do the number of orders changes the number of books categories?



-> Looks like a Poisson distribution

-> Need to offer and recommend books of the same category, but not only as customers with higher orders have bought from more categories! Could change the strategy to make client come back

3) Clustering

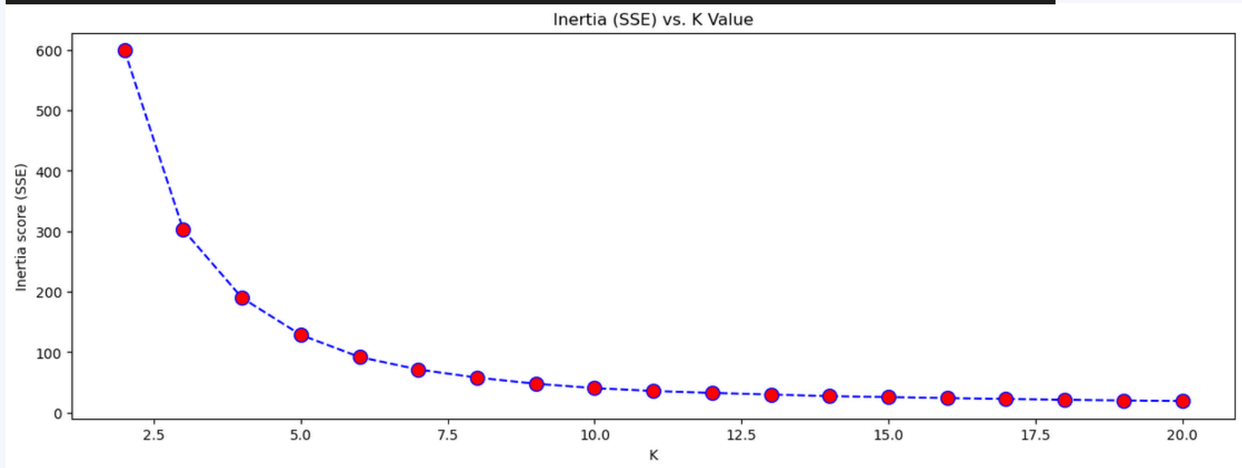
First dataset used: Book sold on website

Can we differentiate customers into groups to help for analysis and strategies improvement?

client_bill_email	units	orders	unique_products_bought	unique_categories_bought
P 1	19	10	19	1
P 10	1	1	1	1
P 100	14	3	14	5
P 1000	8	6	8	4
P 10000	1	1	1	1
...	...	...	...	...
P 99994	5	2	5	5
P 99995	2	1	2	1
P 99996	1	1	1	1
P 99997	1	1	1	1
P 99998	4	1	4	1

124819 rows × 4 columns

group by client and count the sum of units, number of orders and unique products and category bought

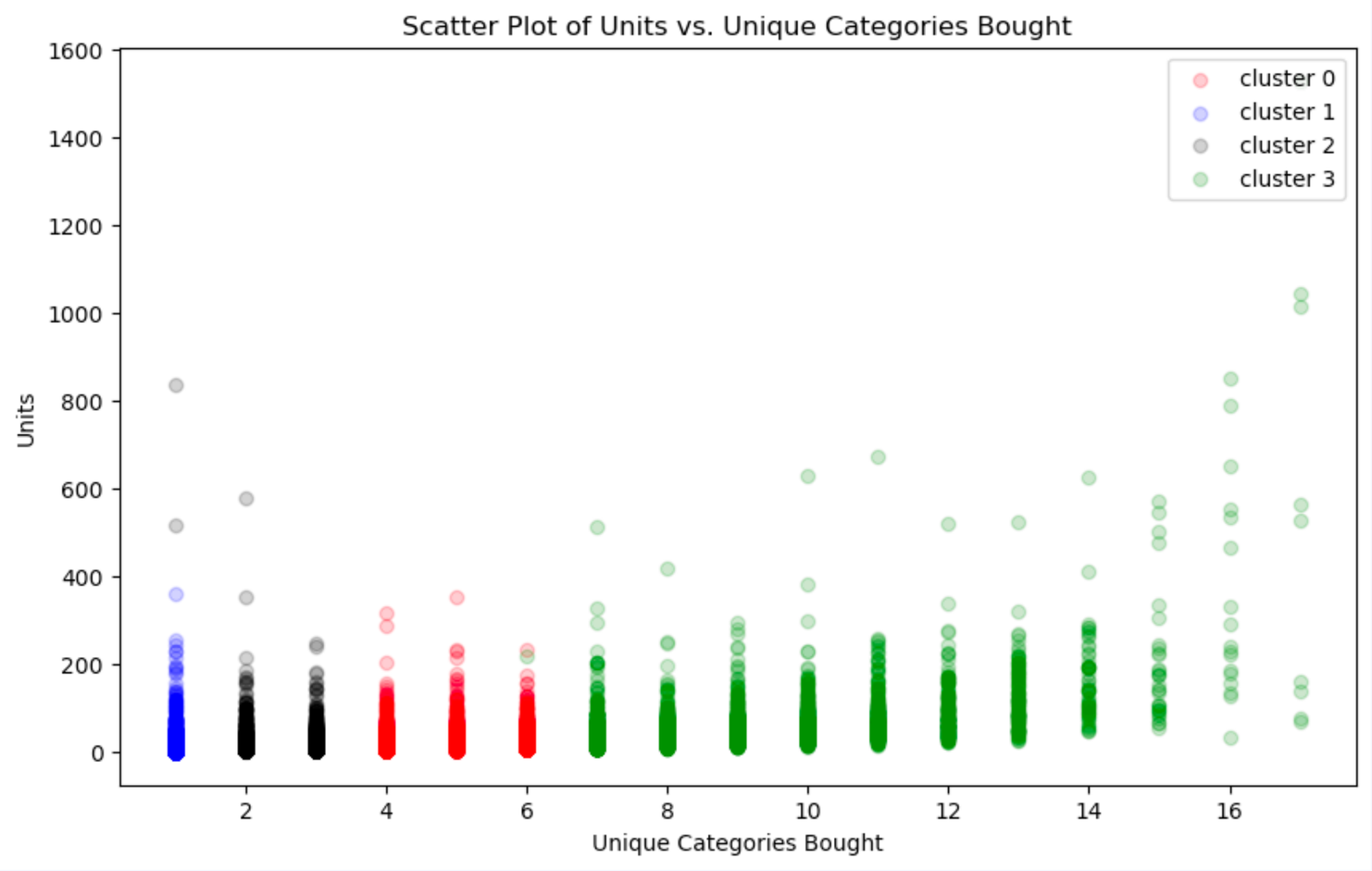


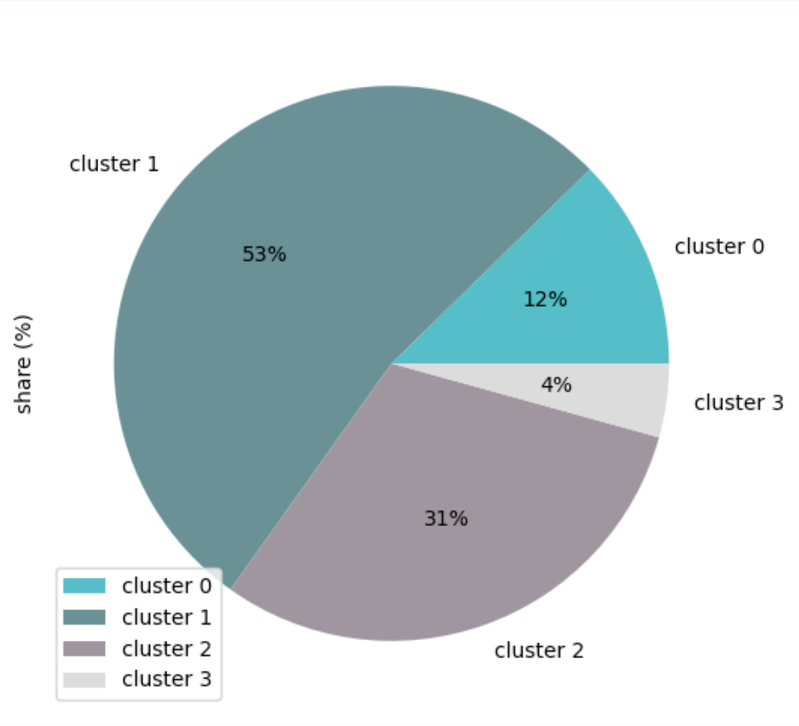
Number of cluster chosen : 4

3) Clustering

First dataset used: Book sold on website

Can we differentiate customers into 4 groups and help for analysis and strategies improvement?



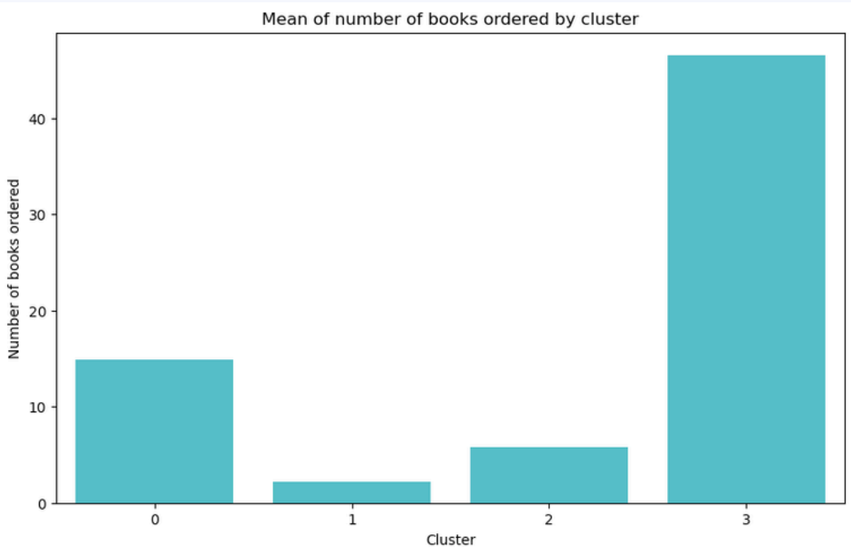


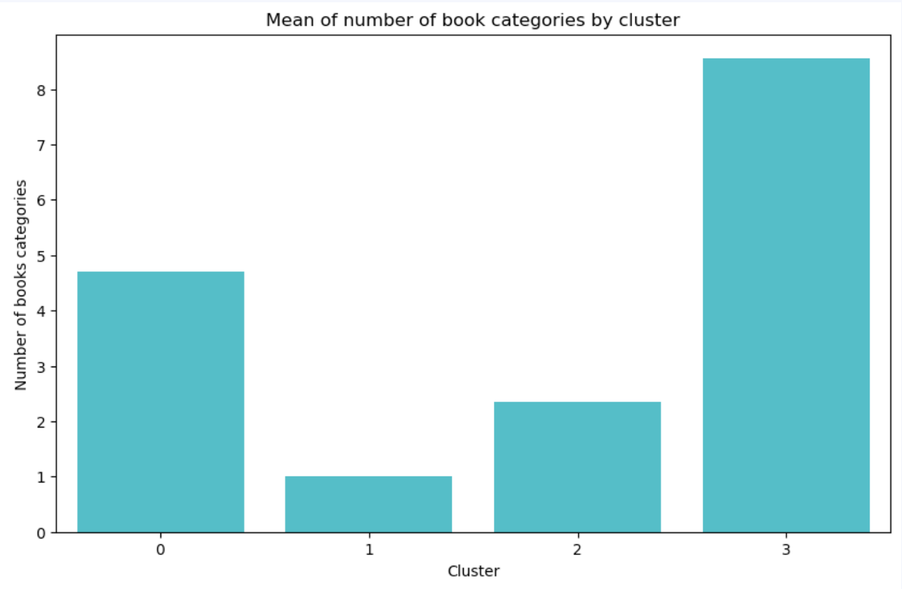
Here we can see a clear separation based on number of different categories bought

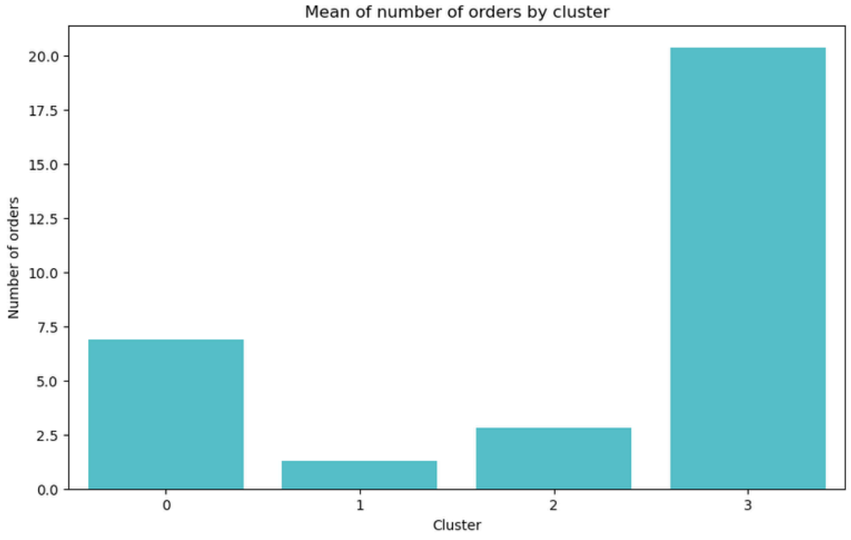
3) Clustering – profiles

First dataset used: Book sold on website

Can we differentiate clients into 4 groups and help for analysis and strategies improvement?







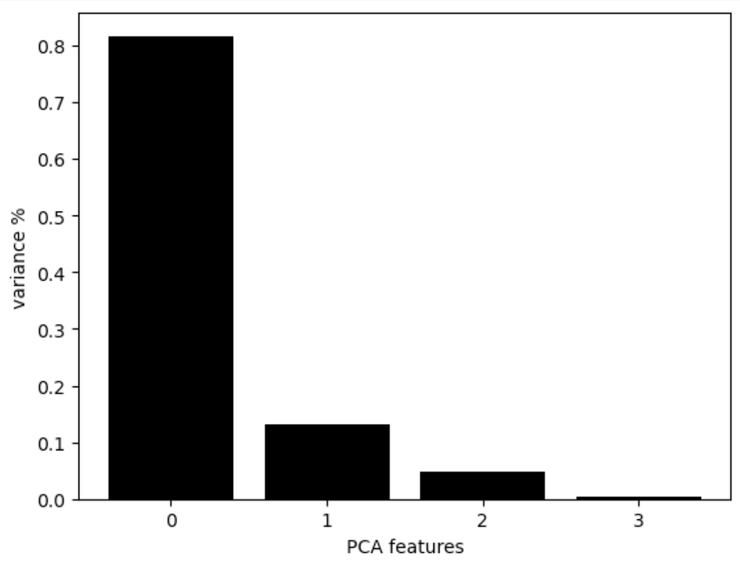
-> Since all variables are correlated, we see the same type of differences between clusters for each variable

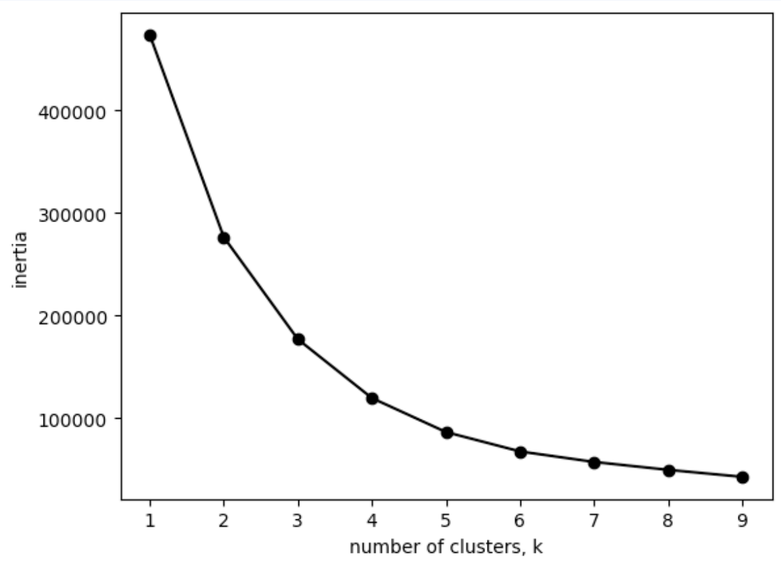
3) Clustering – PCA

First dataset used: Book sold on website

Another method to separate customers : this time we use PCA analysis with the weight of each categories for each customers

	units	orders	unique_products_bought	unique_categories_bought	ART ET SPECTACLE	BD	DIVERS	DROIT ET AFFAIRES	HISTOIRE & POLITIQUE	JEUNESSE	...	LOISIRS PRATIQUE	PAPETERIE	PARALIBRAIRIE	PSYCHO ET DEVLPT	SCIENCES HUMAINES	SCIENCES TECHNIQUES	SCOLAIRE	TOURISME	Unknown	VIE SPIRITUELLE
0	1	1	1	1	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	1.000000	0.000000
1	3	2	2	2	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.666667	0.000000	0.0	0.0	0.0	0.333333	0.000000
2	8	6	7	3	0.0	0.000000	0.0	0.0	0.000000	0.375000	...	0.0	0.0	0.0	0.125000	0.000000	0.0	0.0	0.0	0.000000	0.000000
3	8	1	8	1	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	1.0	0.0	0.000000	0.000000
4	1	1	1	1	0.0	1.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
124814	38	4	36	7	0.0	0.131579	0.0	0.0	0.131579	0.052632	...	0.0	0.0	0.0	0.000000	0.105263	0.0	0.0	0.0	0.026316	0.026316
124815	1	1	1	1	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000
124816	30	7	30	2	0.0	0.000000	0.0	0.0	0.000000	0.166667	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000
124817	2	2	2	1	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	1.0	0.000000	0.000000
124818	3	3	3	1	0.0	0.000000	0.0	0.0	0.000000	0.000000	...	0.0	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.000000	0.000000



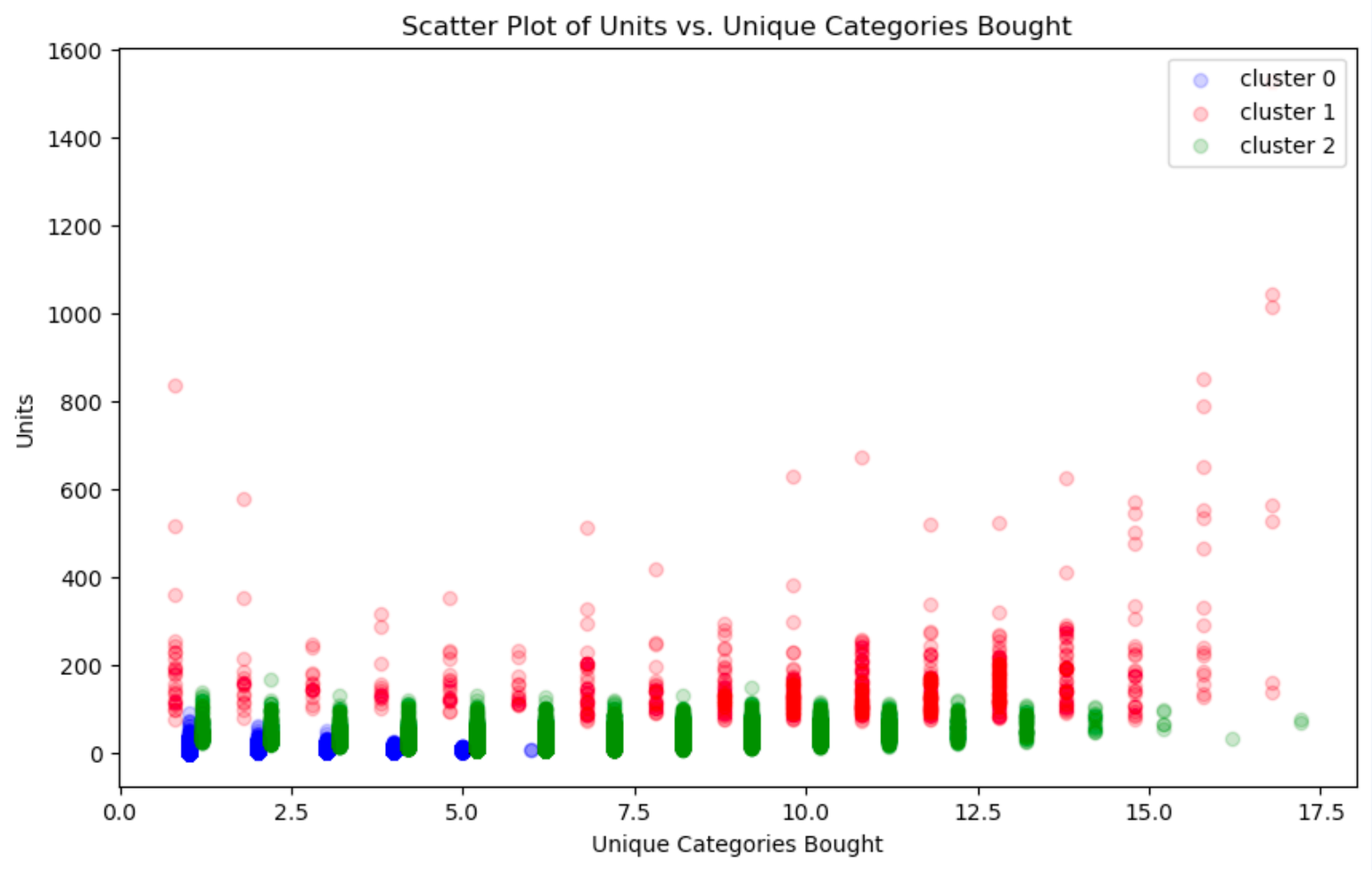


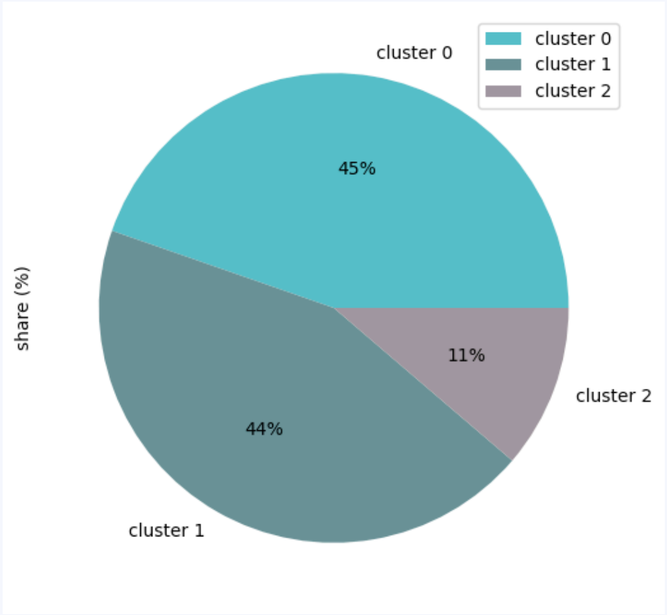
-> Number of features chosen : 3

3) Clustering – PCA

First dataset used: Book sold on website

Another method to separate customers : this time we use PCA analysis on weight of each categories for each customers





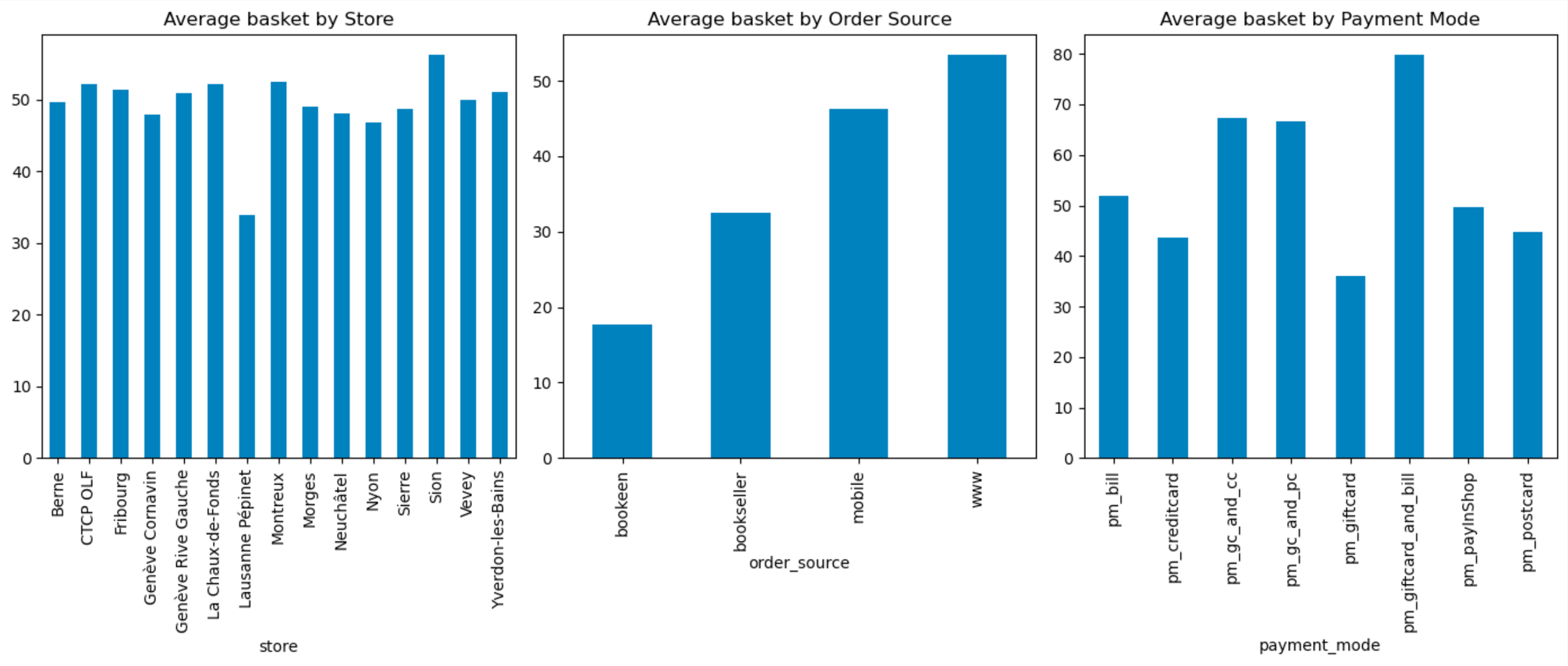
-> This time the separation is horizontal, we have high buyers in green and in red small buyers
But in quantity the first group in red account for 45% of clients

2) EDA on orders

Second dataset used: orders on website

This dataset allow us to see the difference between pick-up at stores or deliveries (=CTCP OLF)

The average basket per order :

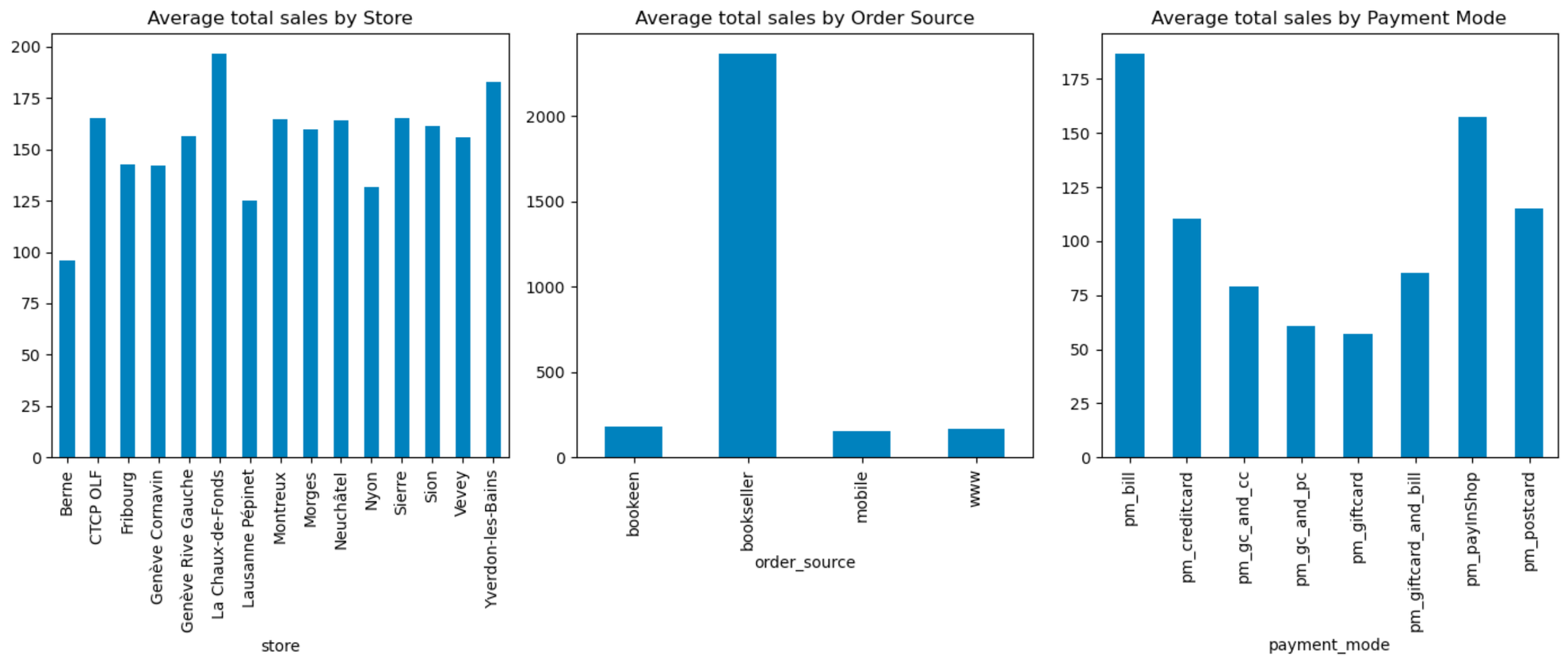


2) EDA on orders

Second dataset used: orders on website

This dataset allow us to see the difference between pick-up at stores or deliveries (=CTCP OLF)

The average total sales per client :



3) Randomforest on total sales per client

Second dataset used: orders on website

Shop and payment_mode in column to see the effect on total sales

	store_CTCP OLF	store_Fribourg	store_Genève Cornavin	store_Genève Rive Gauche
0	1	0	0	0
1	1	0	0	0
2	1	0	0	0
3	0	0	0	0
4	0	0	0	1
...	...	...	...	...
123385	0	0	0	0
123386	1	0	0	0
123387	0	0	0	0
123388	0	0	0	0
123389	1	0	0	0

...

payment_mode_pm_payInShop	payment_mode_pm_postcard	total_pricettc
0	0	87.80
0	0	689.95
0	1	23.00
0	0	3445.80
1	0	543.90
...	...	...
0	1	7.00
0	0	18.10
0	0	11.00
1	0	19.90
0	0	38.50

	Feature	Importance
24	total_pricettc	9.976927e-01
0	store_CTCP OLF	6.660983e-04
22	payment_mode_pm_payInShop	5.935029e-04
16	order_source_www	5.894252e-04
3	store_Genève Rive Gauche	2.680598e-04
15	order_source_mobile	1.771411e-04
11	store_Sion	8.934067e-06
4	store_La Chaux-de-Fonds	1.263610e-06
12	store_Vevey	1.108856e-06
1	store_Fribourg	7.413498e-07
13	store_Yverdon-les-Bains	4.679077e-07
17	payment_mode_pm_creditcard	2.926554e-07
5	store_Lausanne Pépinet	1.867992e-07
14	order_source_bookseller	4.749712e-08
23	payment_mode_pm_postcard	1.543798e-08
8	store_Neuchâtel	1.300818e-08
2	store_Genève Cornavin	1.036305e-08
7	store_Morges	1.223453e-09
10	store_Sierre	4.879280e-10
9	store_Nyon	1.978795e-10
6	store_Montreux	1.512698e-10
18	payment_mode_pm_gc_and_cc	6.686969e-13
21	payment_mode_pm_giftcard_and_bill	9.442506e-14
20	payment_mode_pm_giftcard	2.304396e-14
19	payment_mode_pm_gc_and_pc	5.683288e-21

4) Final words and improvements

Observations

- Majority of client only order once during the studied period
  - People who order more tend to order across more categories -> need to be taken into account for a recommendation system
- Lausanne shop has the lowest average basket and also the lowest average total sales per client (without taking into account Berne because it's a new shop since end of 2022).
- Kmeans, PCA and RandomForest used to try to divide clients into different type of customers
- Bookshop industry is hard because most of the book will be sold only 1 time, making it not easy to build a recommendation system. Only the best sellers are easy to recommend.

Limitations and improvements

- Deeper analysis with books, find a way to get all the resume of each book to build a recommendation system on this
- Need more informations on customers to have perhaps better segmentation of customers
- Association rules used in data mining could be tried to recommend books