INDIANA UNIVERSITY
BLOOMINGTON

H&M Personalised Fashion Recommendation

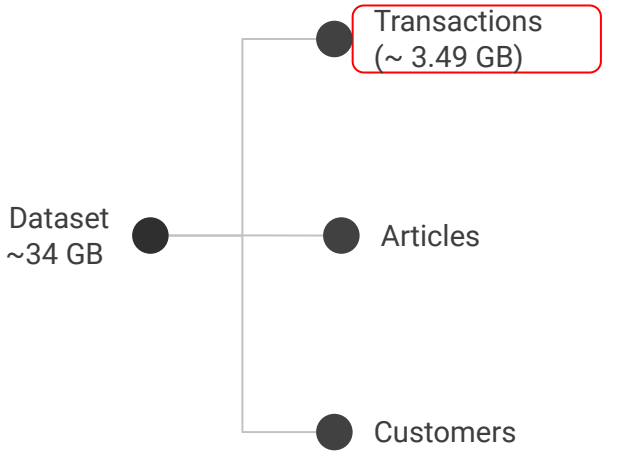
Motivation

H&M Group is a family of brands and businesses with 53 online markets and approximately 4,850 stores. Online store offers shoppers an extensive selection of products to browse through. But with too many choices, customers might not quickly find what interests them or what they are looking for, and ultimately, they might not make a purchase. To enhance the shopping experience, product recommendations are key. More importantly, helping customers make the right choices also has a positive implications for sustainability, as it reduces returns, and thereby minimizes emissions from transportation.

What are we trying to do?

Develop product recommendations based on data from previous transactions, as well as from customer and product meta data. From which we will provide personalised fashion recommendations to our customers tailored to their preferences.

Dataset



Purchases of each customer for each date,
5 columns 31M rows

Meta-data for each article available for purchase
25 columns 105k rows

Meta-data for each h&m customer
7 columns ? rows

Transactions

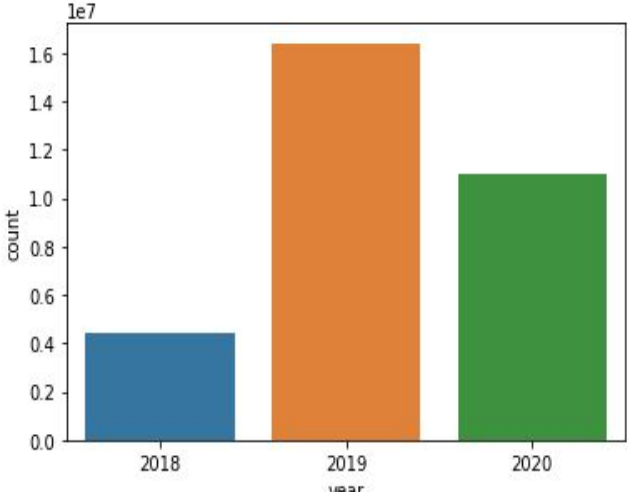
Preprocessing steps:

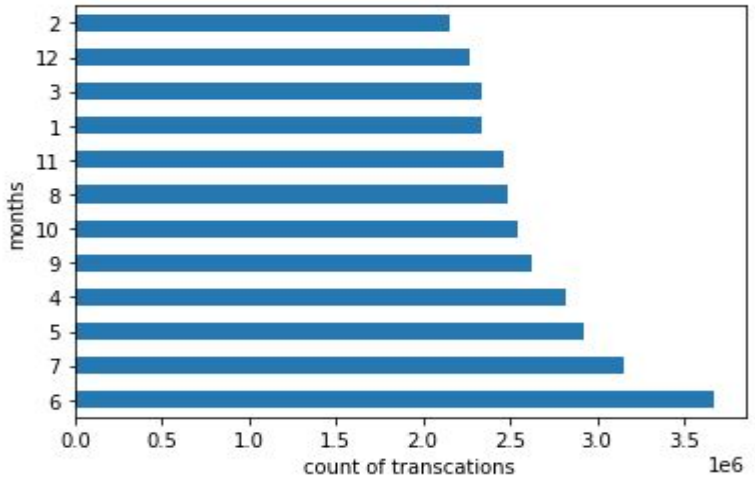
1. Convert t_dat column to datetime format
2. Convert customer_id from char to hex (reduces space)
3. Separate out year month and day of transaction into separate columns

[4]:

	t_dat	customer_id	article_id	price	year	month	day
0	2018-09-20	-6846340800584936	663713001	0.050831	2018	9	20
1	2018-09-20	-6846340800584936	541518023	0.030492	2018	9	20
2	2018-09-20	-8334631767138808638	505221004	0.015237	2018	9	20
3	2018-09-20	-8334631767138808638	685687003	0.016932	2018	9	20
4	2018-09-20	-8334631767138808638	685687004	0.016932	2018	9	20

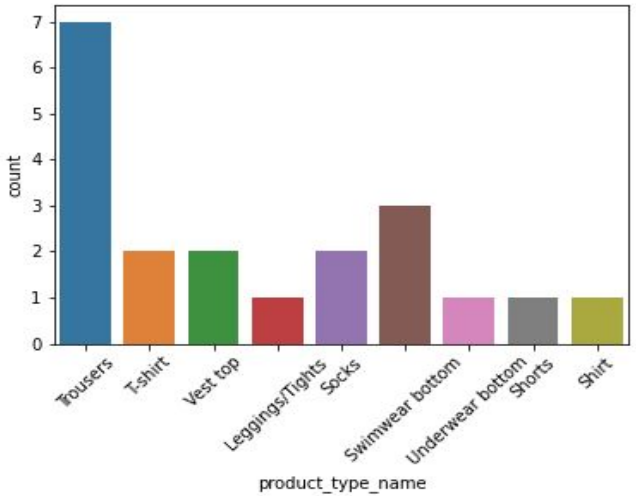
Distribution over years and months





Which clothes do people purchase the most?

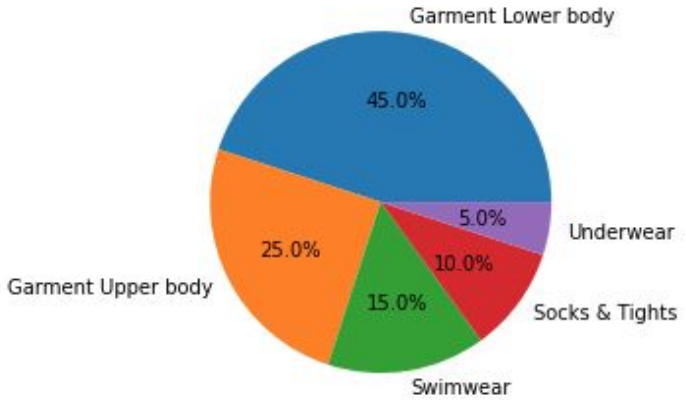
Surprisingly trousers are the most bought item.



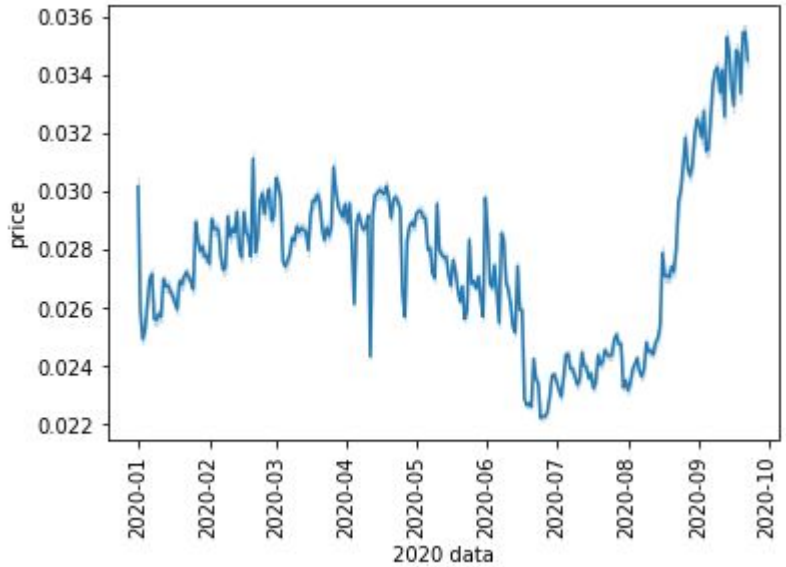
Let's look at the top purchases

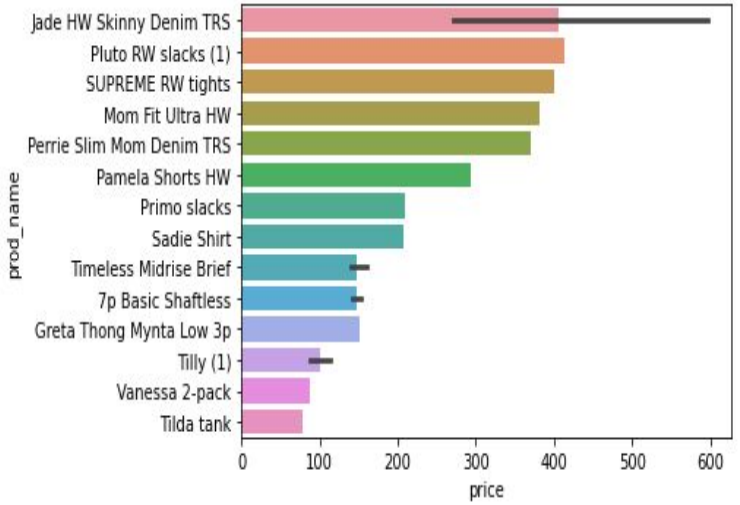
Which group bring in the most money?

Lower body garments dominate this category as observed before as well trousers were a major contributor in the top 20.



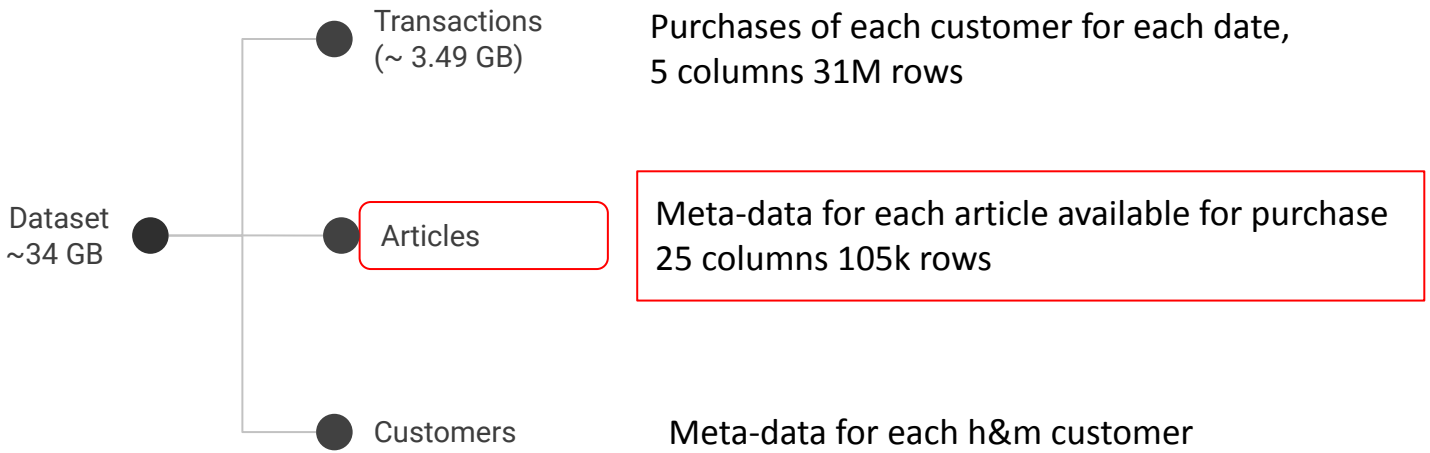
Let's talk a little more about sales!!





Somethings that people buy together!!

Dataset


Dataset
~34 GB
Transactions
(~ 3.49 GB)
Articles
Customers
Purchases of each customer for each date,
5 columns 31M rows
Meta-data for each article available for purchase
25 columns 105k rows
Meta-data for each h&m customer
7 columns ? rows

Articles
Data Description

- 105 k rows
- 25 columns
  - 11 numerical
  - 14 categorical

	article_id	product_code	prod_name	product_type_no	product_type_name	product_group_name	graphical_appearance_no	graphical_appearance_name	colour_group_code
0	108775015	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	9
1	108775044	108775	Strap top	253	Vest top	Garment Upper body	1010016	Solid	10
2	108775051	108775	Strap top (1)	253	Vest top	Garment Upper body	1010017	Stripe	11
3	110065001	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	9
4	110065002	110065	OP T-shirt (Idro)	306	Bra	Underwear	1010016	Solid	10

Articles

Columns

- 105 k rows
- 25 columns
  - 11 numerical
  (Are they actually numerical ?)
  - 14 categorical

```
[ 4  3  1  2  5  7  6 -1]
['Dark' 'Light' 'Dusty Light' 'Medium Dusty' 'Bright' 'Medium' 'Undefined'
 'Unknown']
```

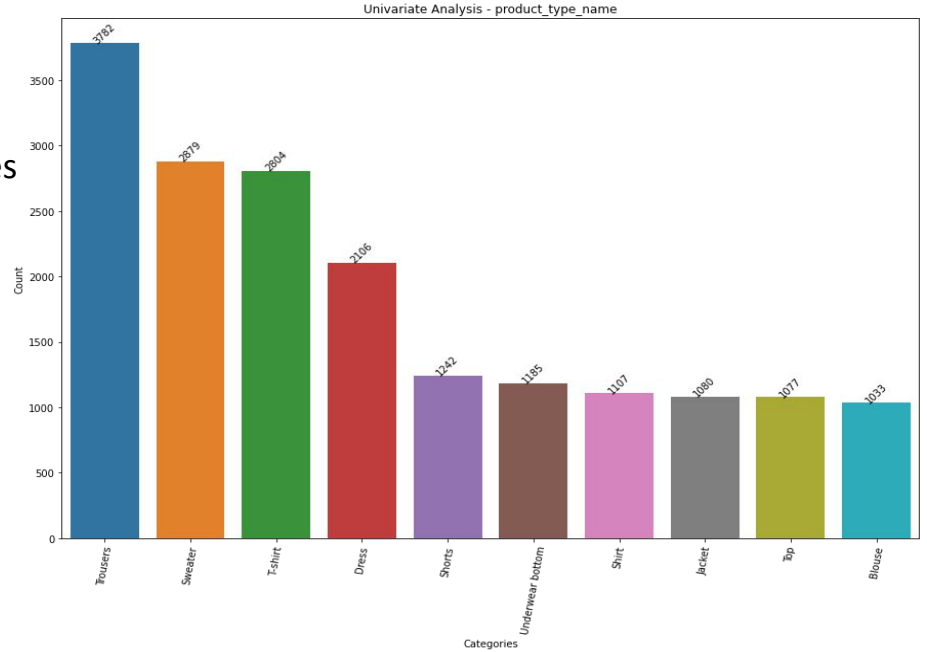
```
[ 1  4  3 26  2]
['Ladieswear' 'Baby/Children' 'Menswear' 'Sport' 'Divided']
```

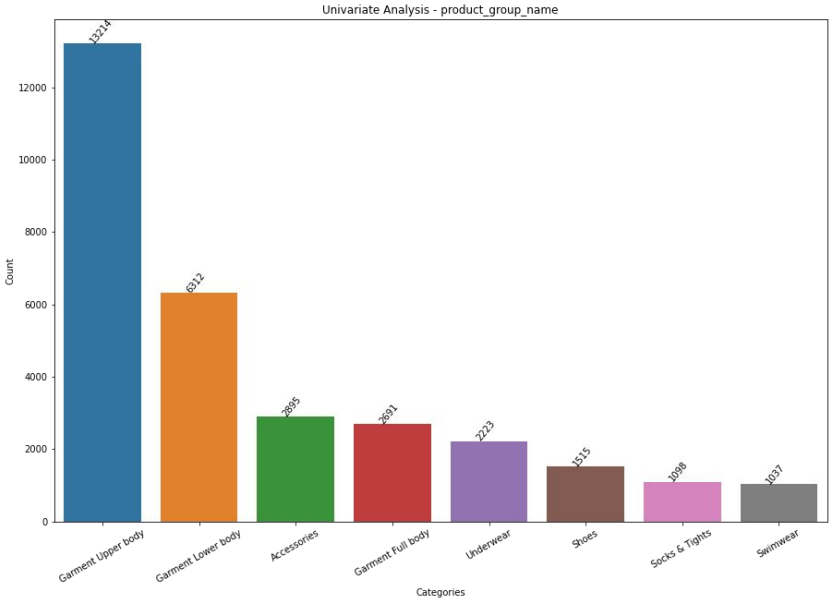
article_id
prod_name
product_type_no
product_code
product_type_name
product_group_name
graphical_appearance_no
graphical_appearance_name
colour_group_code
colour_group_name
perceived_colour_value_id
perceived_colour_value_name
perceived_colour_master_id
perceived_colour_master_name
department_no
department_name
index_code
index_name
index_group_no
index_group_name

section_no
section_name
garment_group_no
garment_group_name
detail_desc

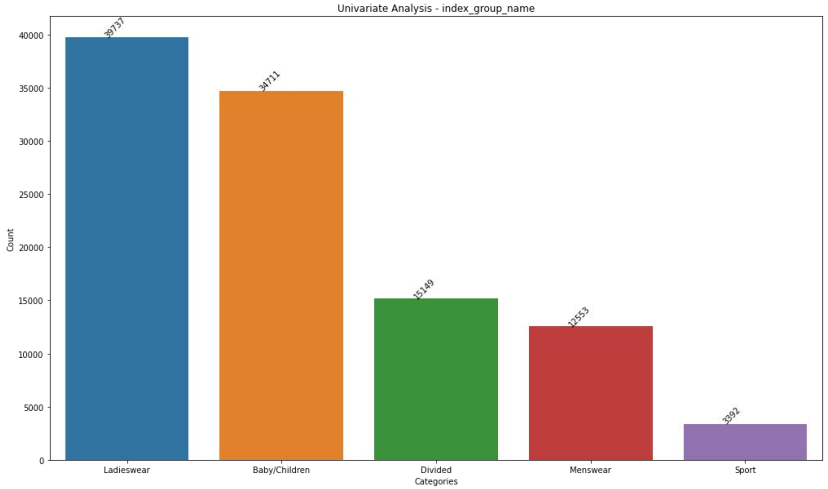
Univariate Analysis

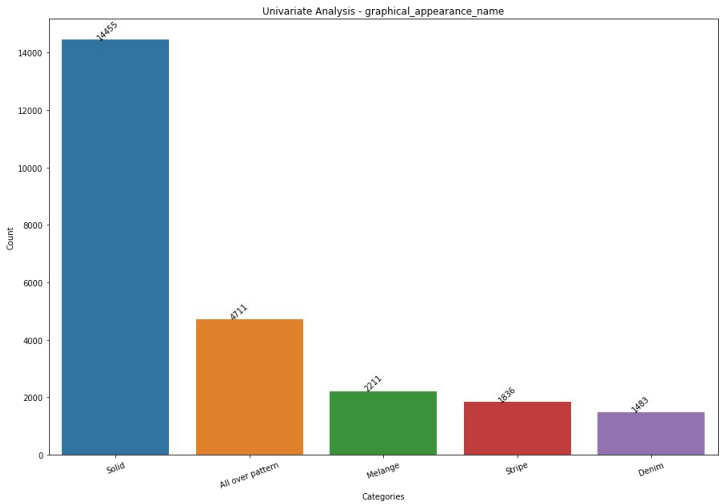
Trousers product type maximum number of articles and the mostly bought item.

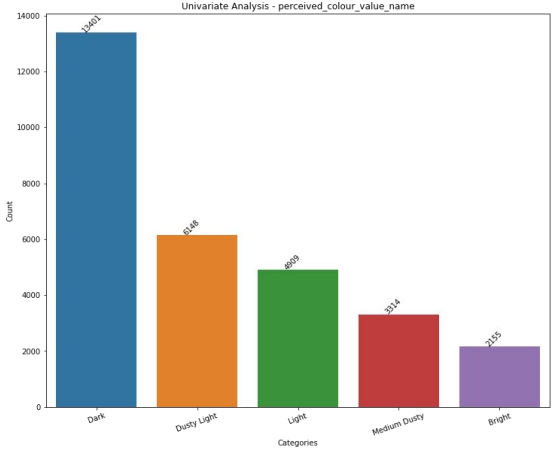




Upper body garments most number of articles followed by **lower body garments** but still lower body garments bring in the most money.







Categories: Index Name, Graphical Appearance and Colour Value

Some more Analysis using these three categories only:

Ladieswear article **:** 39737

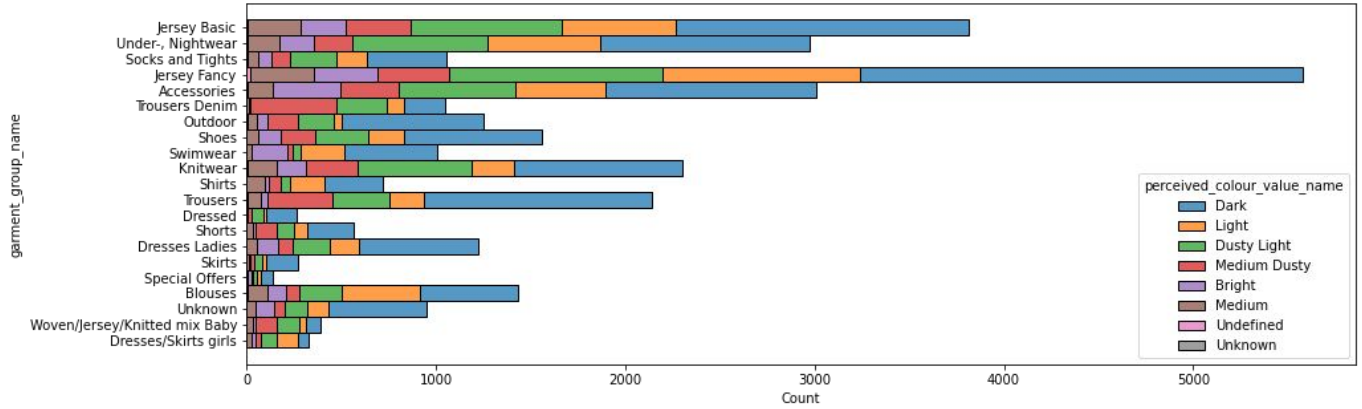
Ladieswear articles with **Solid** graphical appearance : 23426

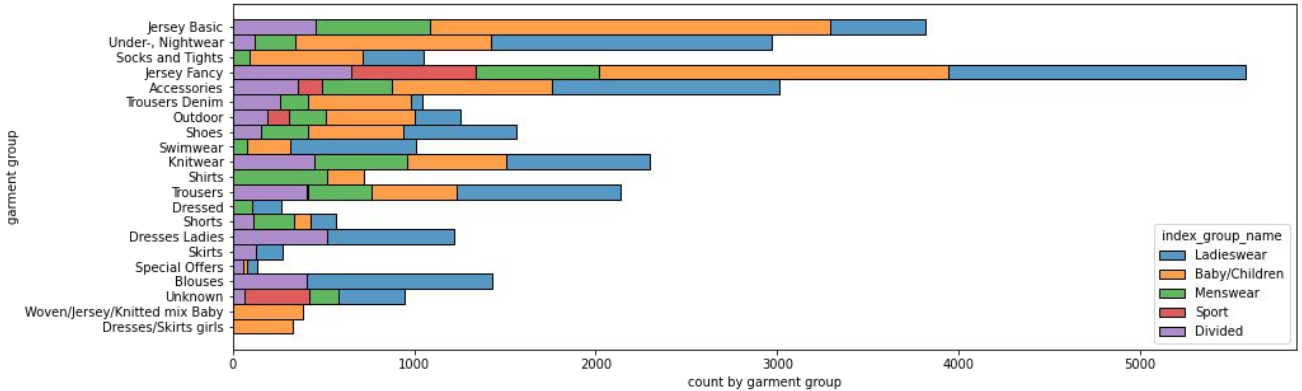
Ladieswear articles with **Solid** graphical appearance and **Dark** colour : 10436

Lower Body and **LadiesWear** articles : 5742

Lower Body and **Ladies Wear** articles with **Solid** graphical appearance : 2904

Bivariate Analysis



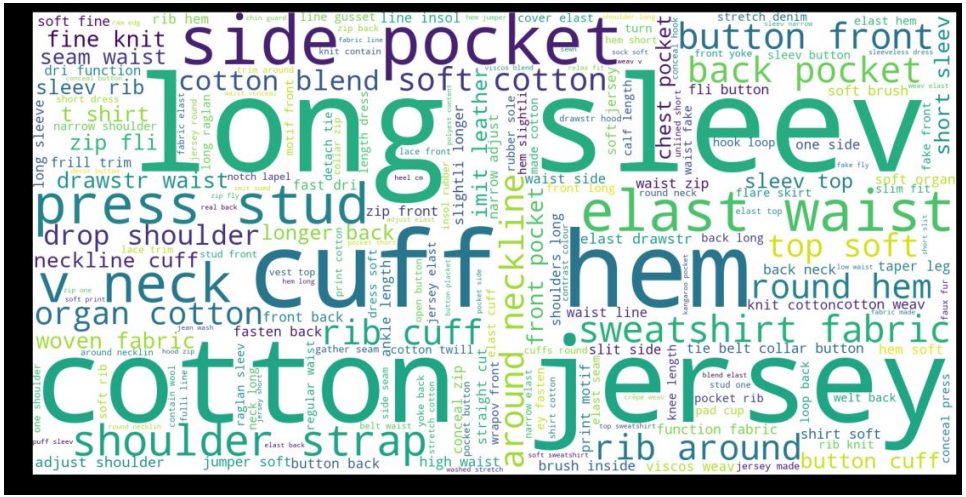


Textual data (description)

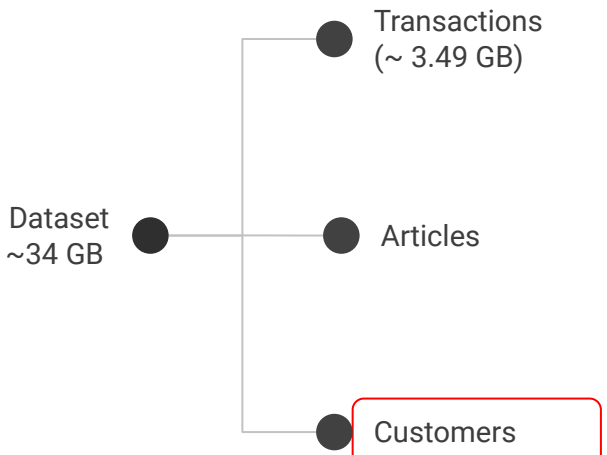
Pre-processing

- Tokenization
- Stop word removal
- Punctuation removal
- Special Character removal
- Stemming
- Lemmatization

cotton	10082
elast	8284
top	7494
soft	7135
jersey	6902
rib	6557
hem.	6231
pocket	6039
zip	5564
short	5486



Dataset

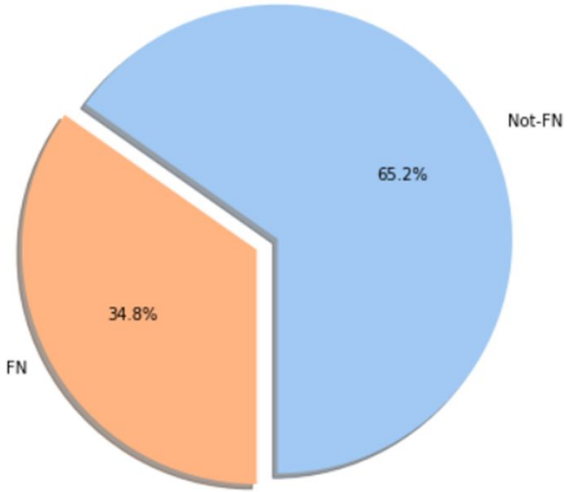


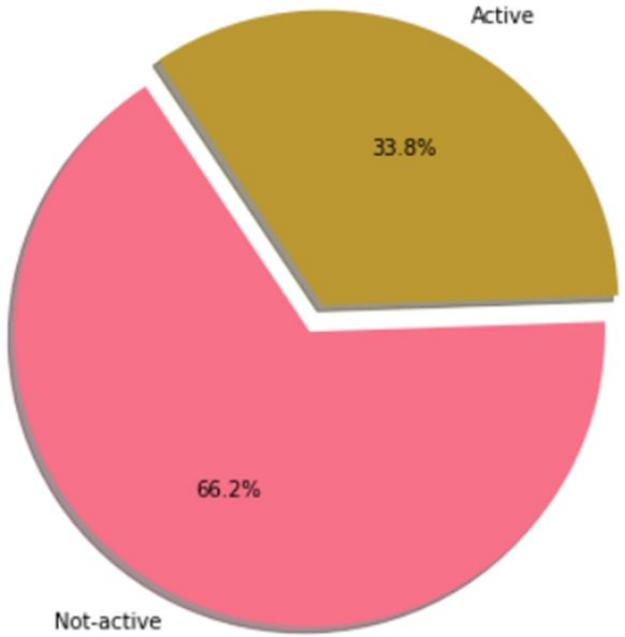
Purchases of each customer for each date,
5 columns 31M rows

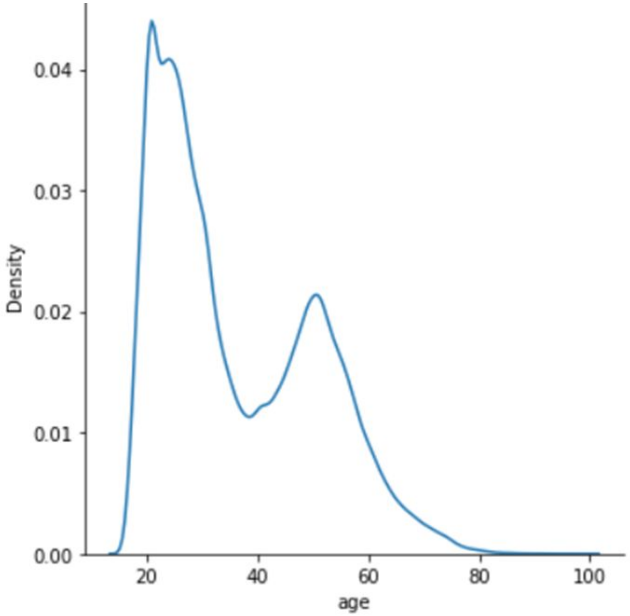
Meta-data for each article available for purchase
25 columns 105k rows

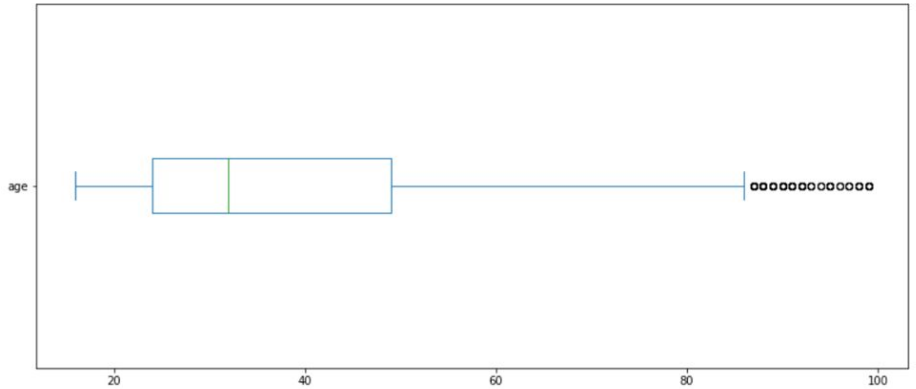
Meta-data for each h&m customer
7 columns ? rows

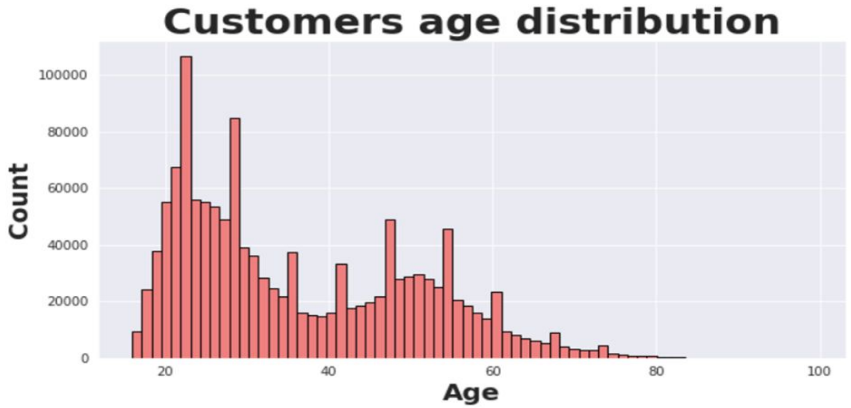
	customer_id	FN	Active	club_member_status	fashion_news_frequency	age	postal_code
0	00000dbacae5abe5e23885899a1fa44253a17956c6d1c3...	NaN	NaN	ACTIVE	NONE	49.0	52043ee2162cf5aa7ee79974281641c6f11a68d276429a...
1	0000423b00ade91418cceaf3b26c6af3dd342b51fd051e...	NaN	NaN	ACTIVE	NONE	25.0	2973abc54daa8a5f8ccfe9362140c63247c5eee03f1d93...
2	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	NaN	NaN	ACTIVE	NONE	24.0	64f17e6a330a85798e4998f62d0930d14db8db1c054af6...
3	00005ca1c9ed5f5146b52ac8639a40ca9d57aeff4d1bd2...	NaN	NaN	ACTIVE	NONE	54.0	5d36574f52495e81f019b680c843c443bd343d5ca5b1c2...
4	00006413d8573cd20ed7128e53b7b13819fe5cfc2d801f...	1.0	1.0	ACTIVE	Regularly	52.0	25fa5ddee9aac01b35208d01736e57942317d756b32ddd...

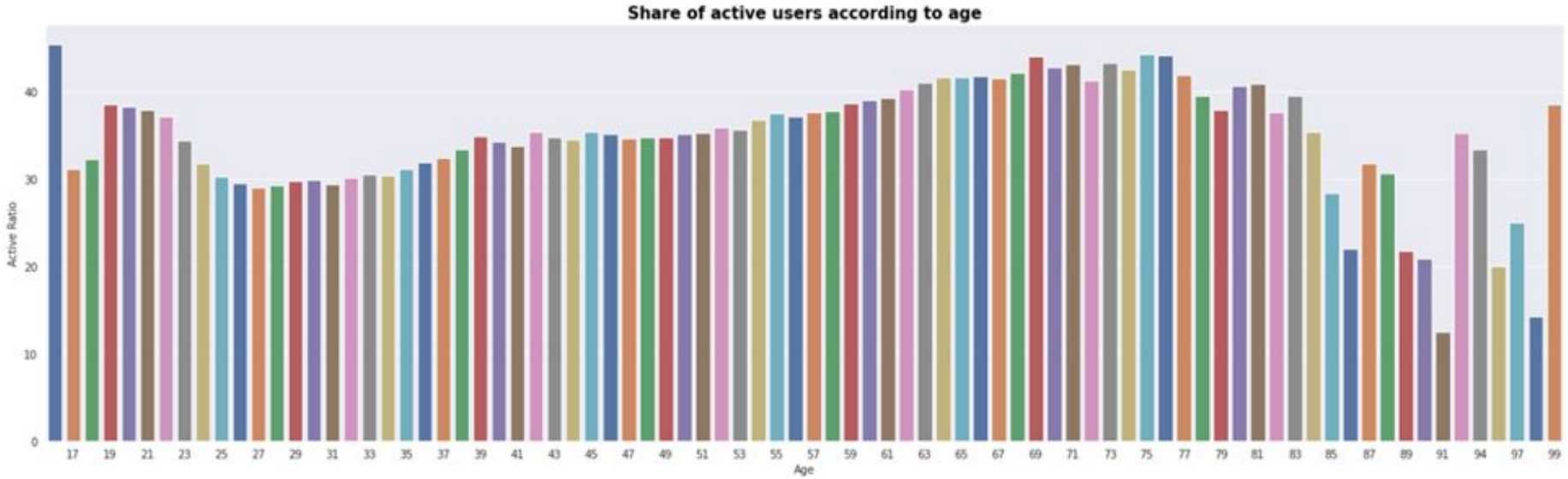

Not-FN
65.2%
34.8%
FN

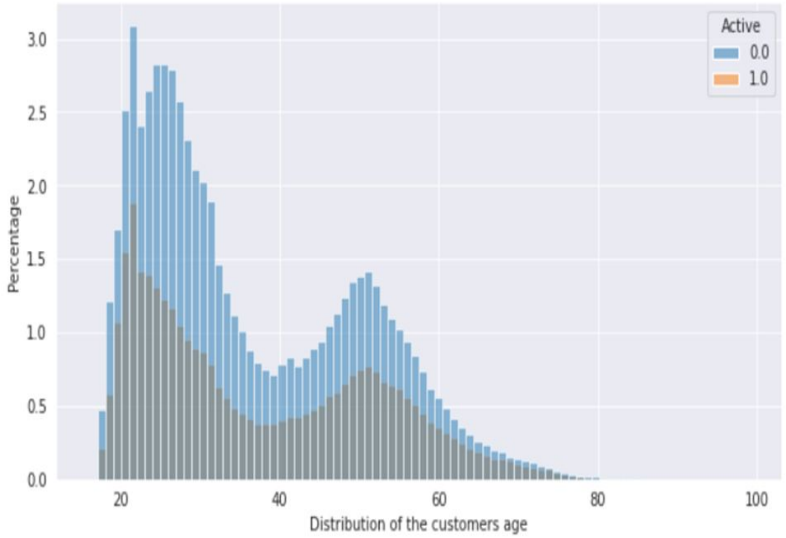

Active
33.8%
66.2%
Not-active

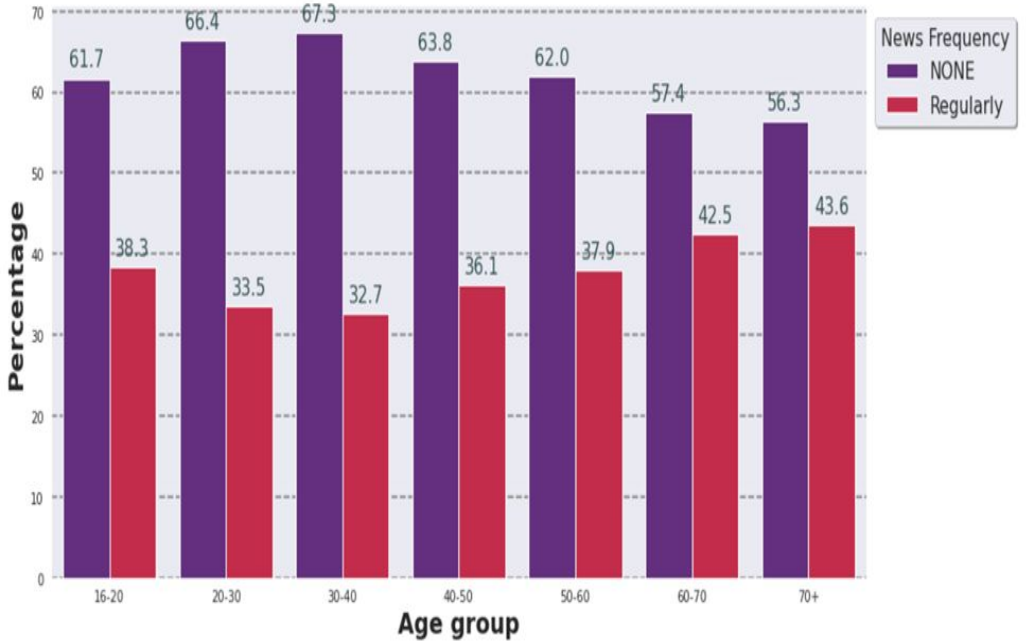

0.04
0.03
0.02
0.01
0.00
Density
20
40
60
80
100
age

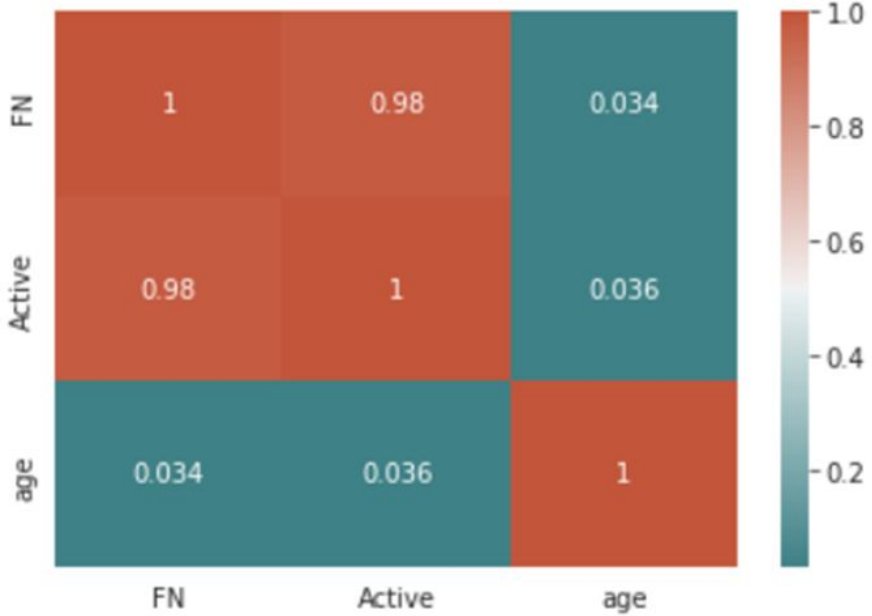

age
20
40
60
80
100

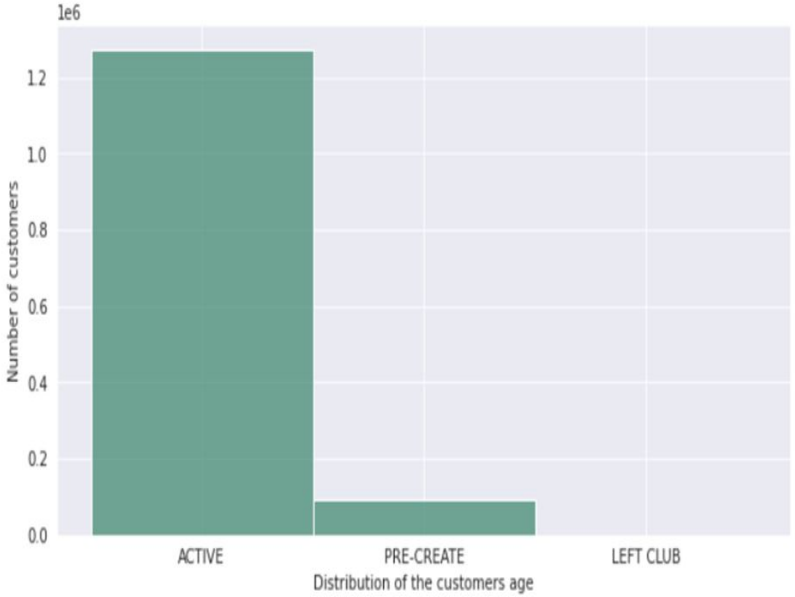

Customers age distribution
100000
80000
60000
40000
20000
0
Count
20
40
60
80
100
Age

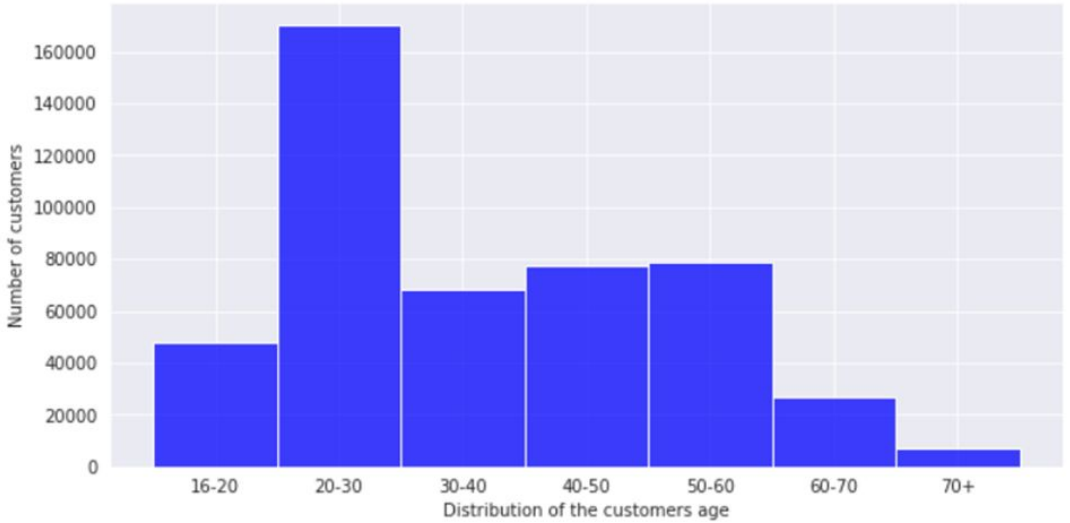

Share of active users according to age
Active Ratio
0
10
20
30
40
17
19
21
23
25
27
29
31
33
35
37
39
41
43
45
47
49
51
53
55
57
59
61
63
65
67
69
71
73
75
77
79
81
83
85
87
89
91
94
97
99
Age

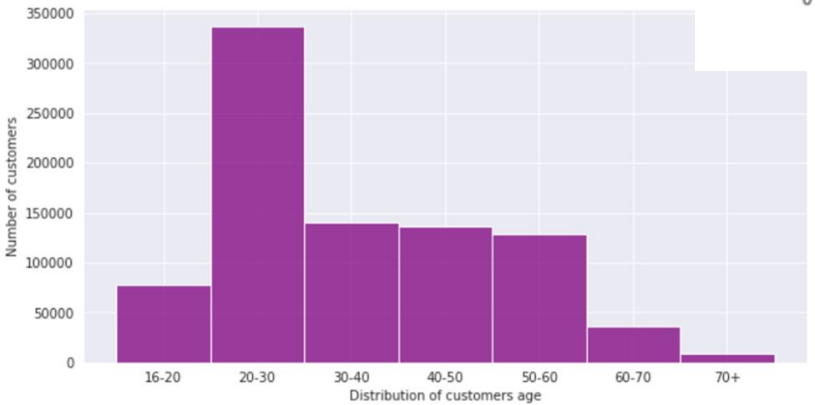

Active
0.0
1.0
Percentage
Distribution of the customers age

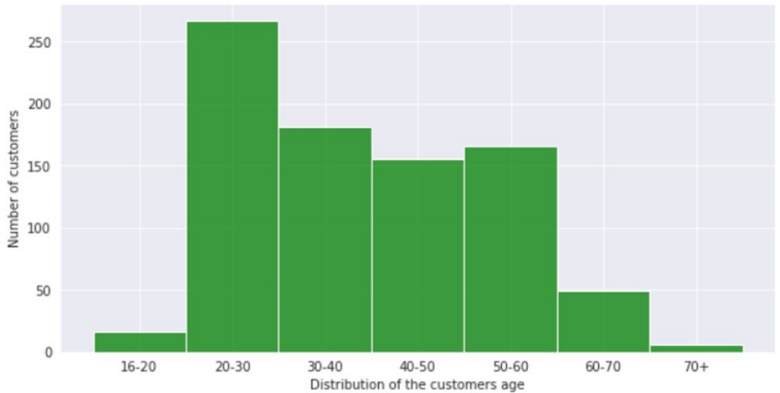

News Frequency
NONE
Regularly
Percentage
Age group
16-20
20-30
30-40
40-50
50-60
60-70
70+
61.7
38.3
66.4
33.5
67.3
32.7
63.8
36.1
62.0
37.9
57.4
42.5
56.3
43.6


FN
Active
age
1
0.98
0.034
0.98
1
0.036
0.034
0.036
1
FN
Active
age
1.0
0.8
0.6
0.4
0.2


1e6
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Number of customers
ACTIVE
PRE-CREATE
LEFT CLUB
Distribution of the customers age


160000
140000
120000
100000
80000
60000
40000
20000
0
Number of customers
16-20
20-30
30-40
40-50
50-60
60-70
70+
Distribution of the customers age


350000
300000
250000
200000
150000
100000
50000
0
Number of customers
16-20
20-30
30-40
40-50
50-60
60-70
70+
Distribution of customers age


250
200
150
100
50
0
Number of customers
16-20
20-30
30-40
40-50
50-60
60-70
70+
Distribution of the customers age

So how do we recommend??

Data Analysis

Extract data from our database if exists for this customer id.

Recommendation based past history

Recommend items most frequently purchased by the customer.

Recommendation based on similar purchases

Recommend products based on similar products purchased by other customers

Recommend our most purchased items

Recommend our most in demand products.

What's in the plan next?

- Run classification algorithms to classify product categories given different product attributes.
- Run Neural Networks to predict the prices of articles using only their images.
- Train model to generate labels for articles using their images.
- Develop more complex rules using past years data from 2019 & 2018 and verifying those rules.