Dataset от компании H&M

На изображении 1. приведен пример структуры датасета.

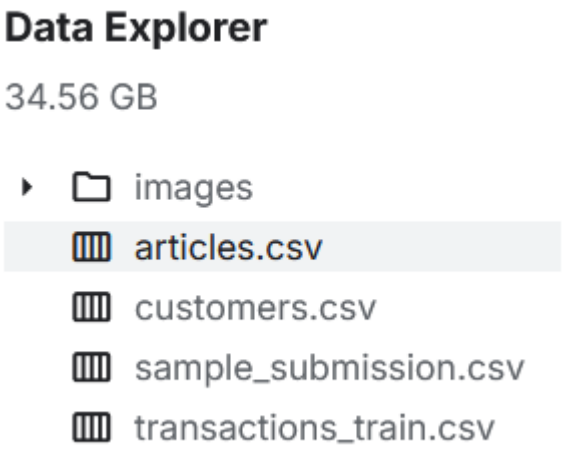


Рисунок 1 – Структура датасета

Images - папка изображений, соответствующая каждому идентификатору вещи (article_id); изображения помещаются в подпапки, начинающиеся с первых трех цифр (article_id). Нужно уточнить, что в данном датасете не все значения article_id имеют соответствующее изображение;

articles.csv - подробные метаданные для каждого товара, доступного для покупки;

customers.csv - метаданные для каждого покупателя в наборе данных;

sample_submission.csv - образец файла подачи в правильном формате (это использовалось в соревновании для правильного составления данных на выход. В нашем проекте он не нужен)

transactions_train.csv - обучающие данные, состоящие из покупок каждого клиента на каждую дату, а также дополнительную информацию. Повторяющиеся строки соответствуют нескольким покупкам одного и того же товара.

Разберем каждый файл более подробно. Пример данных из папки images на рисунке 2:

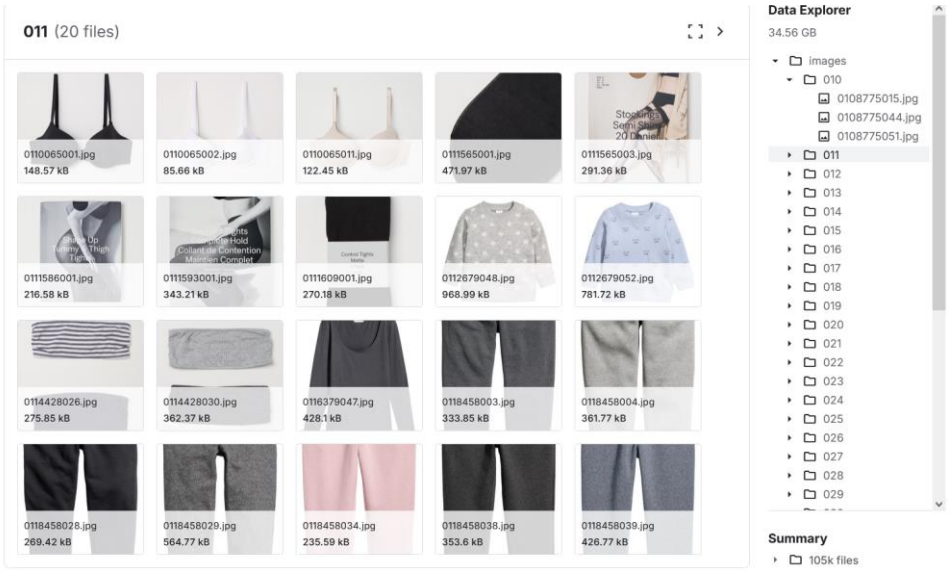


Рисунок 2 – Пример фото из папки images

Таблица articles.csv содержит в себе следующие 25 столбцов, рассмотрим их подробнее:

article_id – уникальный id;

product_code – уникальный код каждого продукта;

prod_name – имя продукта;

product_type – код категории одежды;

product_type_name – имя категории одежды;

graphical_appearance_no – номер графического изображения;

graphical_appearance_name – имя графического изображения на одежде;

colour_group_code – код группы цвета одежды;

colour_group_name – имя цвета одежды;

perceived_colour_value_id – id дополнительного цвета на одежде;

perceived_colour_value_name – название дополнительно цвета;

perceived_colour_master_id – id главного цвета;

perceived_colour_master_name – имя главного цвета;

department_no – уникальный номер департамента;

department_name – имя департамента;

index_code – уникальный идентификатор индекса;

index_name – имя индекса (по сути это тоже категории товаров);

index_group_no – номер группы;

index_group_name – имя группы индексов;

section_no – уникальный идентификатор каждого раздела;

section_name – название раздела;

garment_group_no – уникальный идентификатор предмета одежды;

garment_group_name – название предмета одежды;

detail_desc – полное описание товара.

Произведем запрос head() к таблице:

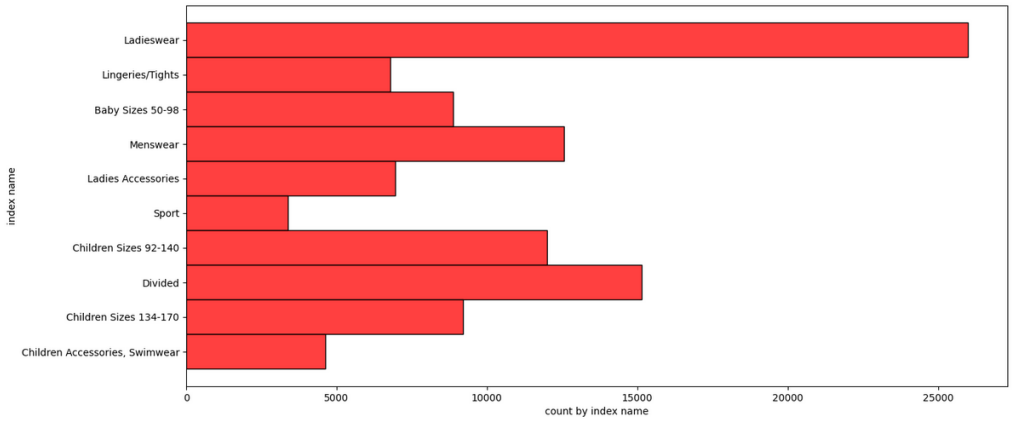
[3]:

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

5 rows × 25 columns

Посмотрим распределение по индексу товаров:

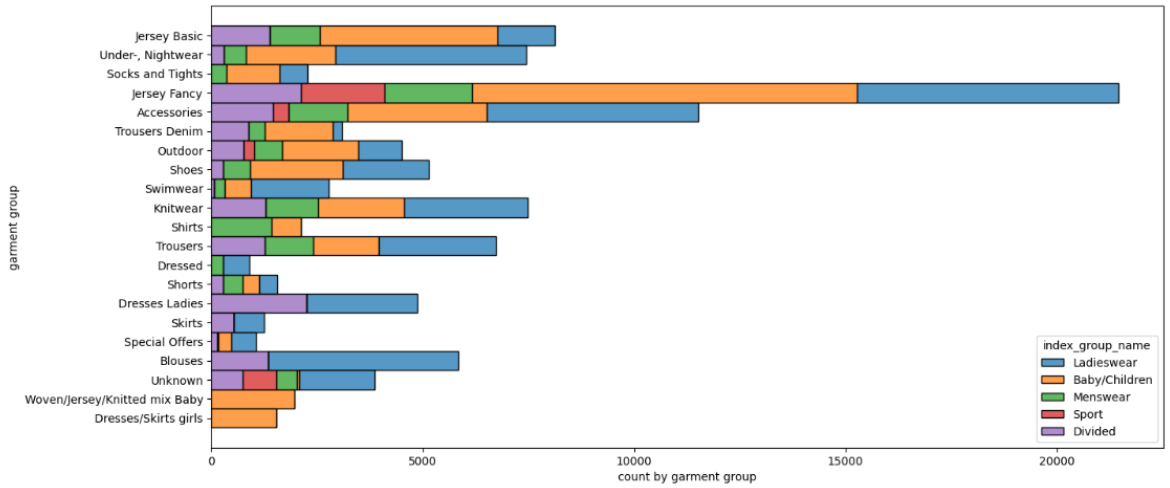
```
/opt/conda/lib/python3.10/site-packages/seaborn/_oldcore.py:1119: FutureWarning: use_inf_as_na option is deprecated and will be removed
in a future version. Convert inf values to NaN before operating instead.
  with pd.option_context('mode.use_inf_as_na', True):
```



Как мы видим, из распределения индексов товаров в данной таблице женские вещи составляют большую часть. На спортивные вещи приходится меньше всего товаров

Посмотрим структуру индексной группы (index_group) с названием предметов одежды (garment_group_name)

```
data_subset = grouped_data.get_group(pd_key)
```



Как видно из графика, наиболее встречающийся элемент одежды – jersey Fancy (Модный трикотаж), особенно для женщин и детей. Следующее место по количеству занимают аксессуары, множество различных аксессуаров по низким ценам.

Сделаем объединение index_group_name и index_name на проверку дополнительных подгрупп:

```
index_group_name  index_name
Baby/Children     Baby Sizes 50-98                8875
                  Children Accessories, Swimwear  4615
                  Children Sizes 134-170          9214
                  Children Sizes 92-140          12007
Divided           Divided                        15149
Ladieswear        Ladies Accessories              6961
                  Ladieswear                     26001
                  Lingeries/Tights                6775
Menswear          Menswear                       12553
Sport             Sport                           3392
Name: article_id, dtype: int64
```

Как видно из объединения index_group_name и index_name, что лишь две категории имеют свои подгруппы. Это группа Дети/Малыши и Женская одежда.

Давайте посмотрим на структуру товарных групп

```
product_group_name     product_type_name
Accessories            Accessories set          7
                       Alice band               6
                       Baby Bib                 3
                       Bag                   1280
                       Beanie                  56
                       Belt                   458
                       Bracelet               180
                       Braces                   3
                       Bucket hat               7
                       Cap                     13
                       Cap/peaked             573
                       Dog Wear                20
                       Earring               1159
                       Earrings                11
                       Eyeglasses               2
```

	Felt hat	10
	Giftbox	15
	Gloves	367
	Hair clip	244
	Hair string	238
	Hair ties	24
	Hair/alice band	854
	Hairband	2
	Hat/beanie	1349
	Hat/brim	396
	Headband	1
	Necklace	581
	Other accessories	1034
	Ring	240
	Scarf	1013
	Soft Toys	46
	Straw hat	6
	Sunglasses	621
	Tie	141
	Umbrella	26
	Wallet	77
	Watch	73
	Waterbottle	22
Bags	Backpack	6
	Bumbag	1
	Cross-body bag	5
	Shoulder bag	2
	Tote bag	2
	Weekend/Gym bag	9
Cosmetic	Chem. cosmetics	3
	Fine cosmetics	46
Fun	Toy	2
Furniture	Side table	13
Garment Full body	Costumes	90
	Dress	10362
	Dungarees	309
	Garment Set	1320
	Jumpsuit/Playsuit	1147
	Outdoor overall	64
Garment Lower body	Leggings/Tights	1878
	Outdoor trousers	130
	Shorts	3939
	Skirt	2696
	Trousers	11169
Garment Upper body	Blazer	1110
	Blouse	3979
	Bodysuit	913
	Cardigan	1550
	Coat	460
	Hoodie	2356
	Jacket	3940
	Outdoor Waistcoat	154
	Polo shirt	449
	Shirt	3405
	Sweater	9302
	T-shirt	7904
	Tailored Waistcoat	73
	Top	4155
	Vest top	2991
Garment and Shoe care	Clothing mist	1
	Sewing kit	1
	Stain remover spray	2
	Washing bag	1

	Wood balls	1
	Zipper head	3
Interior textile	Blanket	1
	Cushion	1
	Towel	1
Items	Dog wear	7
	Keychain	1
	Mobile case	4
	Umbrella	3
	Wireless earphone case	2
Nightwear	Night gown	171
	Pyjama bottom	220
	Pyjama jumpsuit/playsuit	388
	Pyjama set	1120
Shoes	Ballerinas	372
	Bootie	31
	Boots	1028
	Flat shoe	165
	Flat shoes	10
	Flip flop	125
	Heeled sandals	202
	Heels	22
	Moccasins	4
	Other shoe	395
	Pre-walkers	1
	Pumps	188
	Sandals	757
	Slippers	249
	Sneakers	1621
	Wedge	113
Socks & Tights	Leg warmers	7
	Socks	1889
	Underwear Tights	546
Stationery	Marker pen	5
Swimwear	Bikini top	850
	Sarong	66
	Swimsuit	662
	Swimwear bottom	1307
	Swimwear set	192
	Swimwear top	50
Underwear	Bra	2212
	Bra extender	1
	Kids Underwear top	96
	Long John	30
	Nipple covers	19
	Robe	136
	Underdress	20
	Underwear body	174
	Underwear bottom	2748
	Underwear corset	7
	Underwear set	47
Underwear/nightwear	Sleep Bag	6
	Sleeping sack	48
Unknown	Unknown	121

Name: article_id, dtype: int64

Из этой таблицы видно, что аксессуаров больше всего по типу товаров. Тут и сумки, кепки, сережки, перчатки и многие другие вещи

Теперь проанализируем уникальные значения в столбцах:

```python
for col in articles.columns:
    if not 'no' in col and not 'code' in col and not 'id' in col:
        un_n = articles[col].nunique()
        print(f'count unic in col - {col}: {un_n}')
```

```
count unic in col - prod_name: 45875
count unic in col - product_type_name: 131
count unic in col - product_group_name: 19
count unic in col - graphical_appearance_name: 30
count unic in col - colour_group_name: 50
count unic in col - perceived_colour_value_name: 8
count unic in col - perceived_colour_master_name: 20
count unic in col - department_name: 250
count unic in col - index_name: 10
count unic in col - index_group_name: 5
count unic in col - section_name: 56
count unic in col - garment_group_name: 21
count unic in col - detail_desc: 43404
```

Итог по таблице: присутствует большая несбалансированность классов с явными выбросами. У некоторых групп есть свои подгруппы, которые стоит учитывать при осуществлении рекомендации пользователям.

Теперь рассмотрим таблицу customers более подробно. Данная таблица содержит в себе следующие столбцы:

customer_id – уникальный id покупателя;

FN – принимает значения 1 или NaN;

Active – принимает значения 1 или NaN;

club_member_status – статус покупателя в клубе H&M;

fashion_news_frequency – как часто пользователю H&M может посылать рекламные записи

age – возраст клиента;

postal_code – почтовый индекс клиента.

Произведем запрос head() к таблице:

[14]:

	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...

Посмотрим количество повторяющихся записей от покупателей:

```
customers.shape[0] - customers['customer_id'].nunique()
```

]: 0

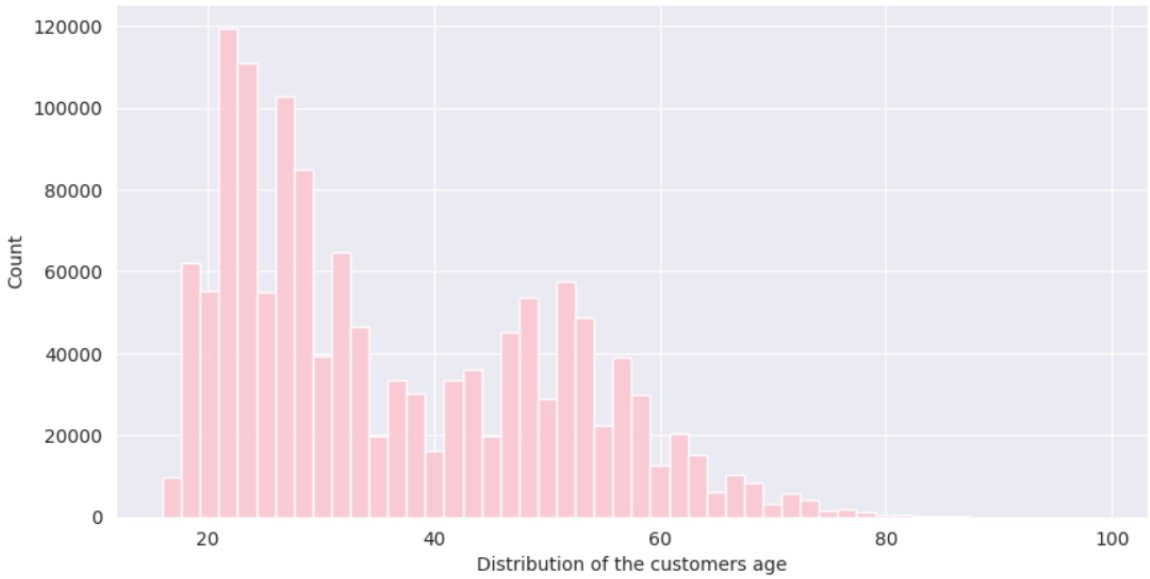
Записей с повторяющимися покупателями нету. Посмотрим подробнее на почтовые индексы отсортировав их:

[16]:

	postal_code	customer_id	FN	Active	club_member_status	fashion_news_frequency	age
61034	2c29ae653a9282cce4151bd87643c907644e09541abc28...	120303	42874	39886	118281	114377	118002
281937	cc4ed85e30f4977dae47662ddc468cd2eec11472de6fac...	261	109	104	261	261	260
156090	714976379549eb90aae4a71bca6c7402cc646ae7c40f6c...	159	90	88	159	159	158
171208	7c1fa3b0ec1d37ce2c3f34f63bd792f3b4494f324b6be5...	157	55	54	157	156	156
126228	5b7eb31eabebd3277de632b82267286d847fd5d44287ee...	156	42	41	156	156	155

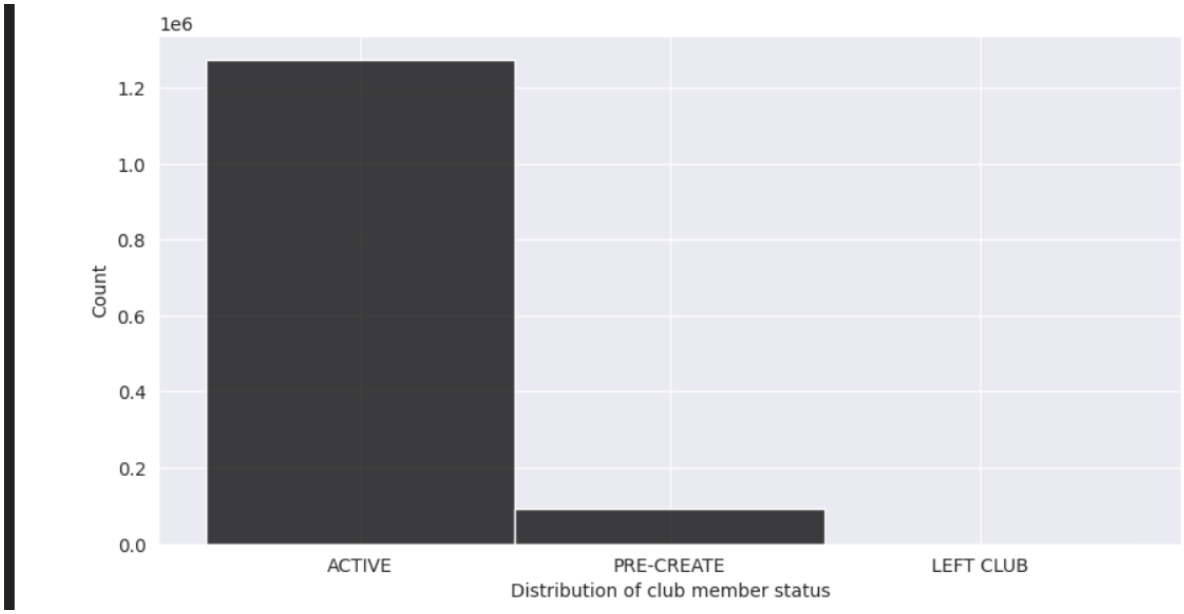
Как видно из таблицы, у нас есть один явно странный почтовый индекс. Он один имеет 120303 клиентов. Данная запись похожа на выброс.

Посмотрим распределение клиентов по их возрасту:



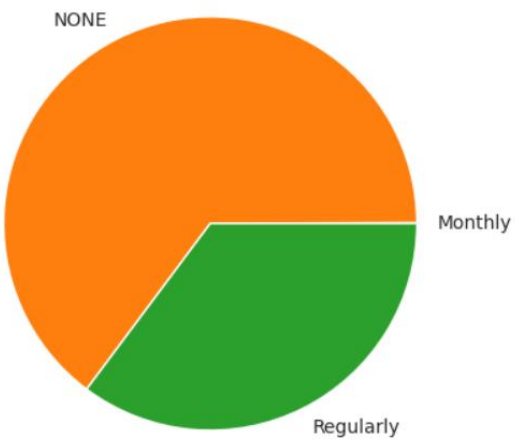
Как видно из графика, самый распространенный возраст клиента чуть больше 20 лет.

Давайте посмотрим, как много клиентов имеет статус в клубе H&M:



Большая часть покупателей имеет активный клубный статус, небольшая часть предварительно активируют клубный статус. Самая малая часть покупателей отказалась от клуба

Теперь посмотрим, как много клиентов подписано на рассылку по почте:



Distribution of fashion news frequency

Как видно из распределения, большая часть клиентов предпочитает не получать рекламные письма к себе на почту.

Теперь рассмотрим последнюю таблицу под названием transactions. Данная таблица содержит в себе следующие столбцы:

t_dat – дата покупки;

customer_id – уникальный id клиента;

article_id – уникальный id;

price – цена покупки;

sales_channel_id – канал продажи, который принимает значения 1 или 2;

Произведем запрос head() к таблице:

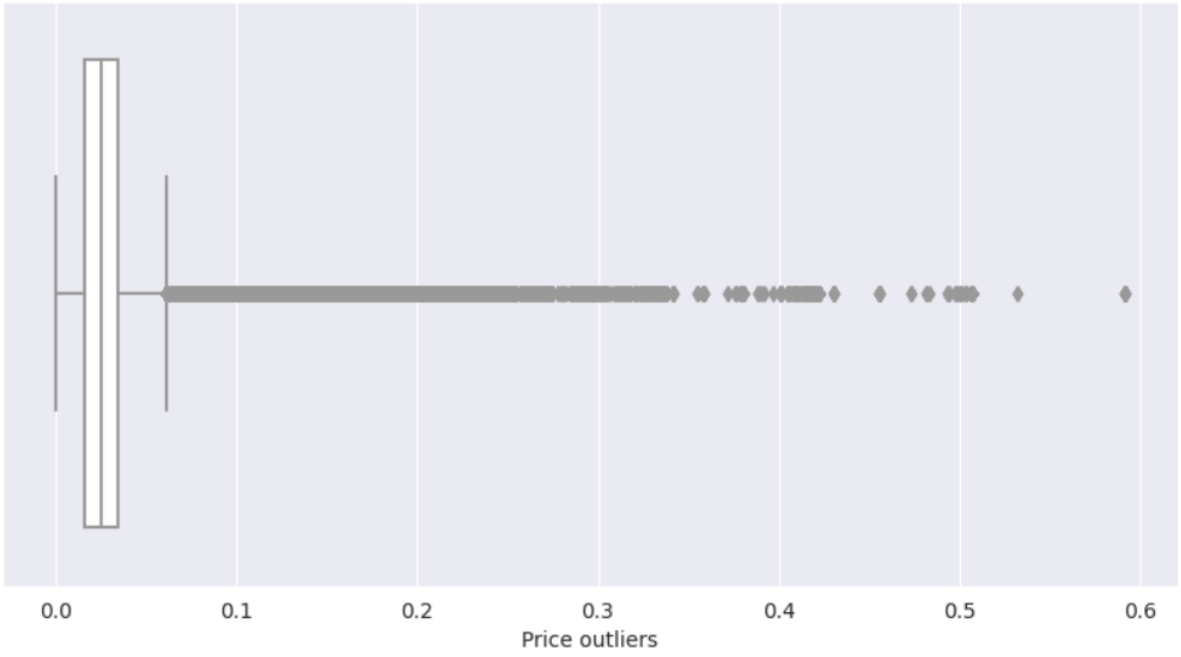
]:

	t_dat	customer_id	article_id	price	sales_channel_id
0	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	663713001	0.050831	2
1	2018-09-20	000058a12d5b43e67d225668fa1f8d618c13dc232df0ca...	541518023	0.030492	2
2	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	505221004	0.015237	2
3	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687003	0.016932	2
4	2018-09-20	00007d2de826758b65a93dd24ce629ed66842531df6699...	685687004	0.016932	2

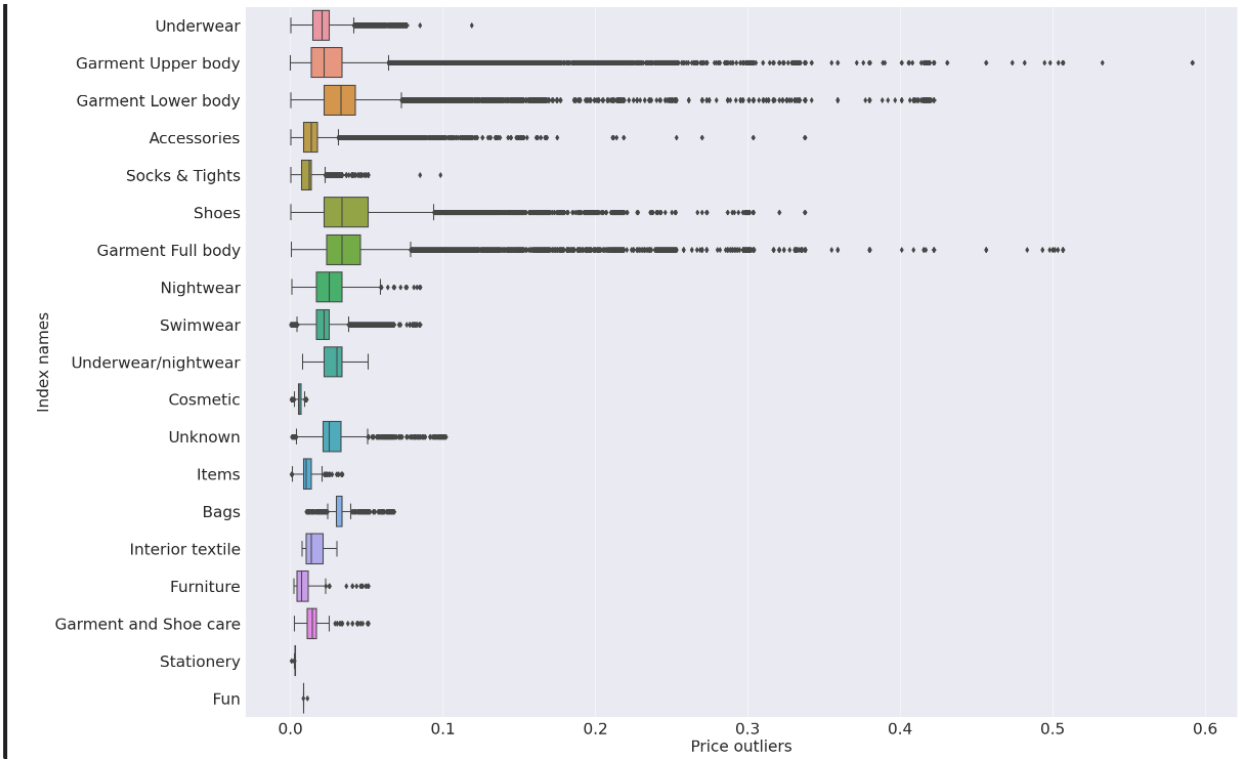
Выведем общую статистику по ценам:

```
[27]: count   31788324.0000
      mean            0.0278
      std             0.0192
      min             0.0000
      25%             0.0158
      50%             0.0254
      75%             0.0339
      max             0.5915
      Name: price, dtype: float64
```

Посмотрим отклонения цены с помощью box-plot

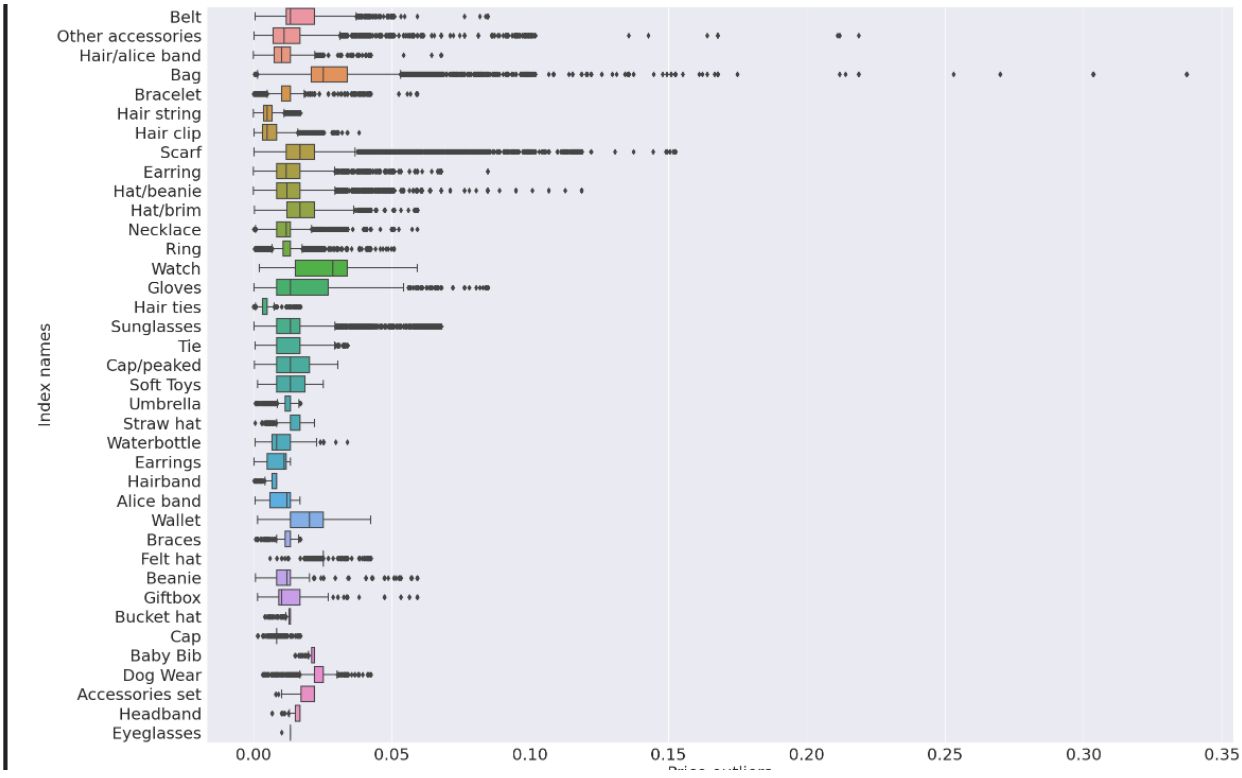


Теперь сравним отклонения цен внутри групп:



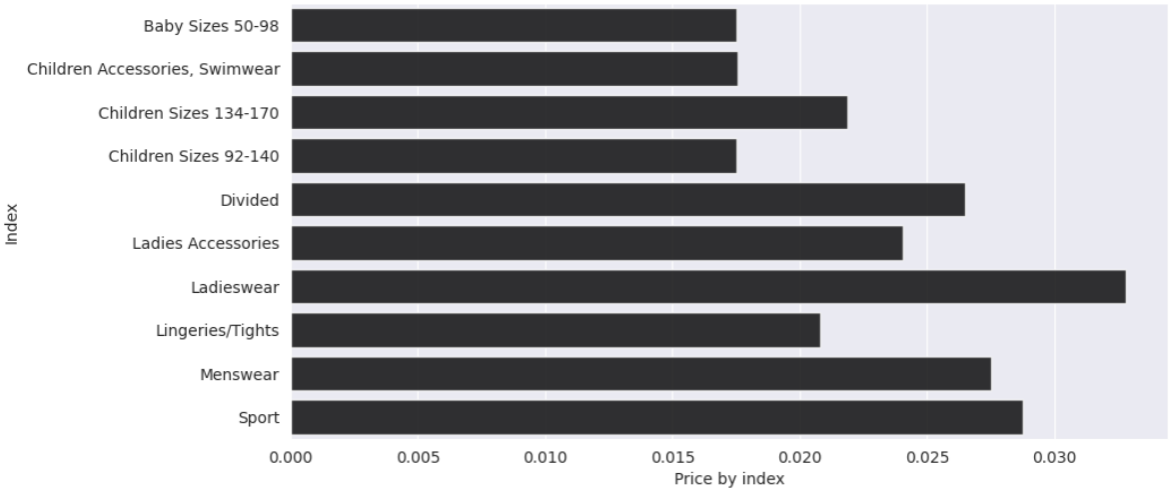
Видны большие выбросы по ценам, особенно в: аксессуарах, одежды на всю тело и одежды для верхней части тела. Скорее всего это связано с какими-либо дорогими коллекциями.

Теперь посмотрим ранжирование цен внутри самой большой категории товаров – аксессуаров:



На данном графике, видно сильное отклонение в ценах. Особенно это заметно с сумками, шапками, шарфами и категорией других аксессуаров.

Посмотрим распределение средней цены по индексу товаров:



Самая завышенная категория – женская одежда, а самая низка – детская

Итог: очень много выбросов, которые нужно обрабатывать, так как многие методы, которые используются для решения задачи рекомендации товара пользователю плохо работают с выбросами