🛍 Sales and Customers Analysis 📈

Contents

1. Data Overview

1.1 Articles

This table contains all H&M articles with details such as a type of product, a color, a product group and other features

Articles dataset contains **25 columns** and **105542 rows:**

article_id **: A unique identifier of every article.**
product_code, prod_name **: A unique identifier of every product and its name (not the same).**
product_type, product_type_name **: The group of product_code and its name**
graphical_appearance_no, graphical_appearance_name **: The group of graphics and its name**
colour_group_code, colour_group_name **: The group of color and its name**
perceived_colour_value_id, perceived_colour_value_name, perceived_colour_master_i, perceived_colour_master_name **: The added color info**
department_no, department_name: **: A unique identifier of every dep and its name**
index_code, index_name: **: A unique identifier of every index and its name**
index_group_no, index_group_name: **: A group of indeces and its name**
section_no, section_name: **: A unique identifier of every section and its name**
garment_group_no, garment_group_name: **: A unique identifier of every garment and its name**
detail_desc: **: Details**

	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

Figure 1. Sample of Articles dataset

1.2 Customers

Customers dataset contains **7 columns** and **1371980 rows:**

customer_id **: A unique identifier of every customer**
FN **: 1 or missed**
Active **: 1 or missed**
club_member_status **: Status in club**
fashion_news_frequency **: How often H&M may send news to customer**
age **: The current age**
postal_code **: Postal code of customer**

	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...

Figure 2. Sample of Customers dataset

1.3 Transactions

Transactions dataset contains **5 columns** and over **31 milion rows:**

t_dat **: A unique identifier of every customer**
customer_id **: A unique identifier of every customer (in** customers **table)**
article_id **: A unique identifier of every article (in** articles **table)**
price **: Price of purchase**
sales_channel_id **:** 1 or 2, 1 is ONLINE and 2 is OFFLINE

	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

Figure 3. Sample of Transactions dataset

2. Data Preprocessing

2.1 Articles

Each *code/id* column represents for *name* column, so I will remove all *code/id* columns (except *article_id*) to save memory. This dataset doesn't contain any Nan values.

```
article_id                    0
prod_name                     0
product_type_name             0
product_group_name            0
graphical_appearance_name     0
colour_group_name             0
perceived_colour_value_name   0
perceived_colour_master_name  0
department_name               0
index_name                    0
index_group_name              0
section_name                  0
garment_group_name            0
dtype: int64
```

Figure 4. Number of Null values of Articles dataset

2.2 Customers

To save memory:

- Drop the *postal_code* column
- Because *customer_id* is a long string, it will take up a lot of memory → encode it using row index

```
id_to_index_dict = dict(zip(df_customers["customer_id"], df_customers.index))
```

The remaining columns contain many Null values

```
customer_id                 0
FN                     895050
Active                 907576
club_member_status       6062
fashion_news_frequency  16009
age                     15861
dtype: int64
```

Figure 5. Number of Null values of Customers dataset

- *FN, Active*: this column only contains values 1 and Nan → Encode Nan values with 0
- Fill *age* with Mode value
- *club_member_status*: Replace Nan values with "None"

fashion_news_frequency: Here we have three types for NO DATA. Let's unite these values → Replace NONE, nan, None values with "None"

```
df_customers['fashion_news_frequency'].unique()
```

```
array(['NONE', 'Regularly', nan, 'Monthly', 'None'], dtype=object)
```

2.3 Transactions

```
t_dat               0
customer_id         0
article_id          0
price               0
sales_channel_id    0
dtype: int64
```

Figure 6. Number of Null values of Transaction dataset

- Similar to Customers dataset, we need to encode the *customer_id*
- Replace *sales_channel_id* column with *sales_channel*

```
df_transactions['sales_channel'] = df_transactions['sales_channel_id'].apply(lambda x: "ONLINE" if x==1 else "OFFLINE")
```

3. Articles analysis

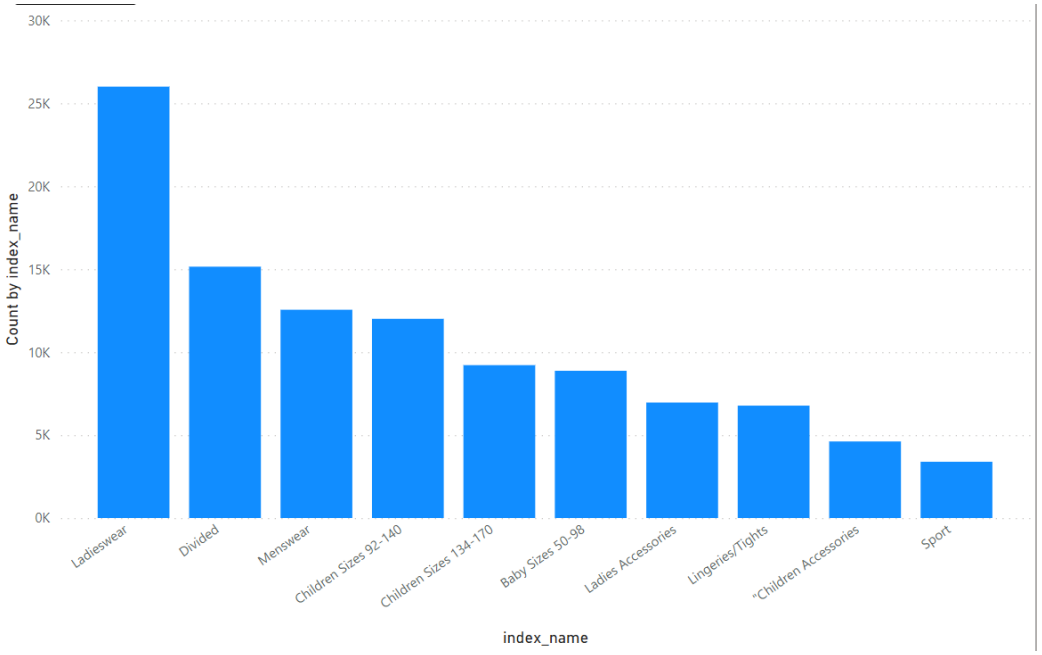

Figure 7. Number of articles_id by index_name

Ladieswear accounts for a significant part of all dresses. Sportswear has the least portion.

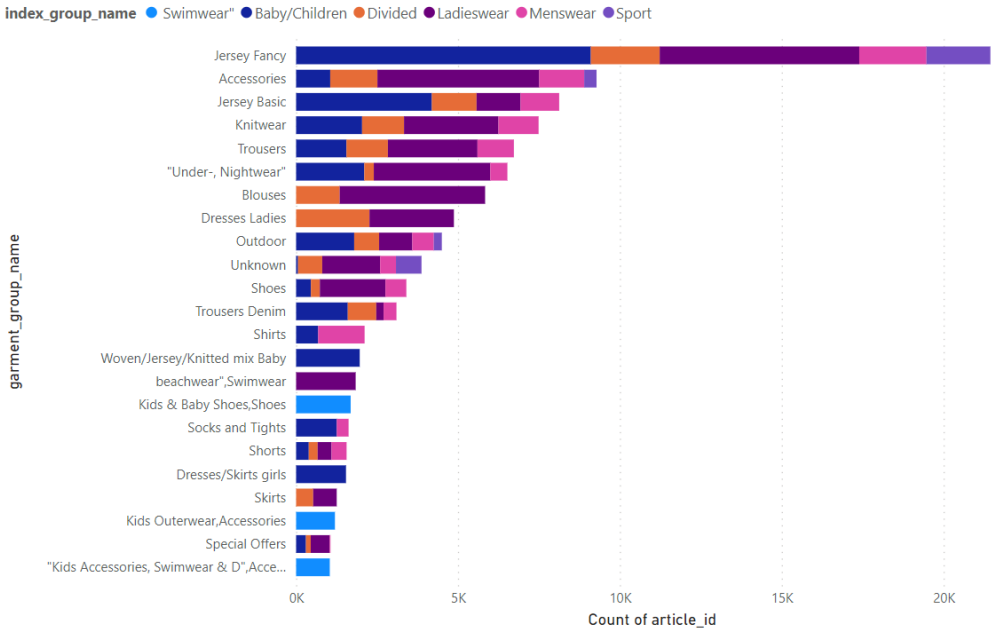

Figure 8. Number of article_id by garment_group_name

The garments grouped by index: Jersey fancy is the most frequent garment, especially for women and children. The next by number is accessories, many various accessories with low price.

4. Customers analysis

4.1 Club member status

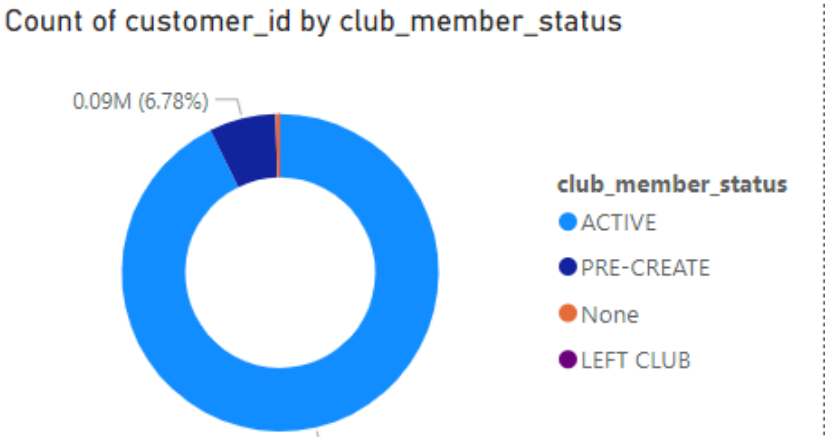

Figure 9. Number of customer by club_member_status

Status in H&M club. Almost every customer has an active club status, some of them begin to activate it (pre-create). A tiny part of customers abandoned the club.

4.2 Fashion news frequency

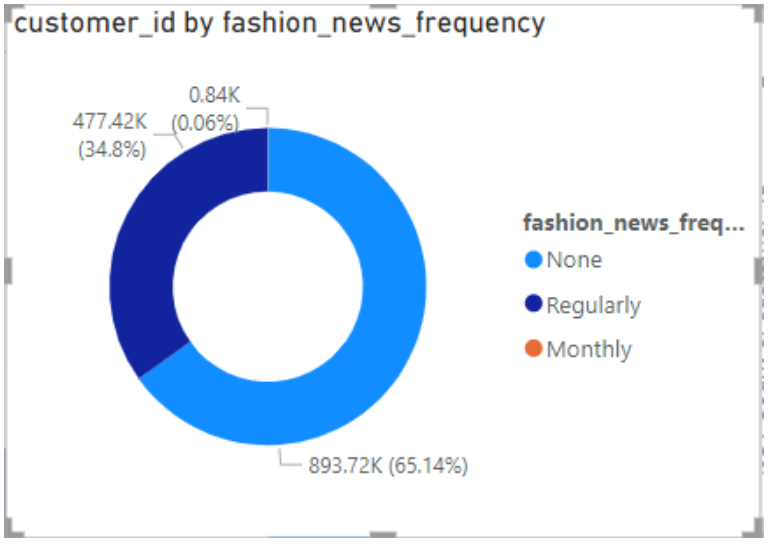

Figure 10. Number of customer by fashion_ news_frequency

65% Customers prefer not to get any messages about the current news.

4.3 Customer's Age

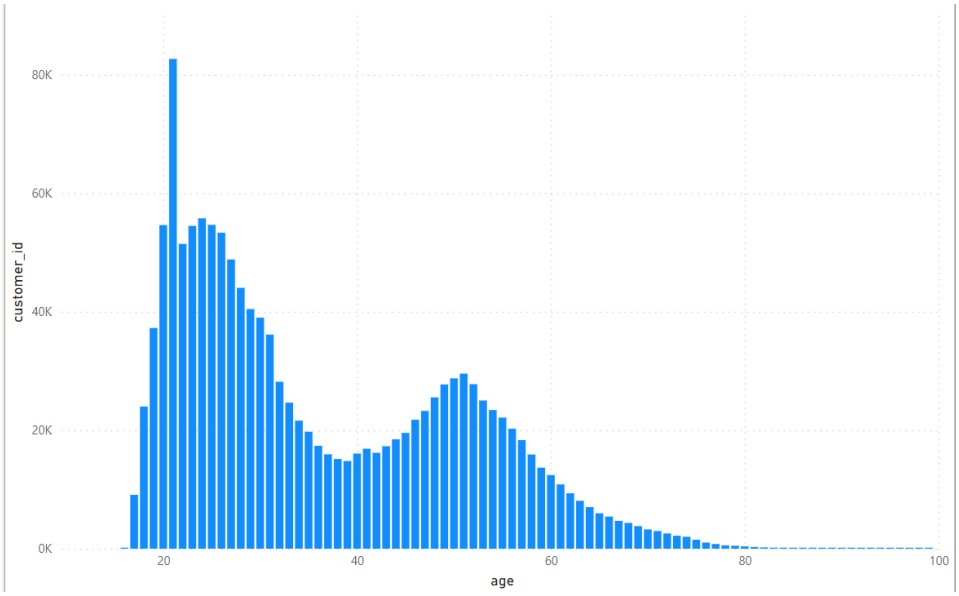

Figure 11. Customer's age distribution

The most common age is about **21-23**

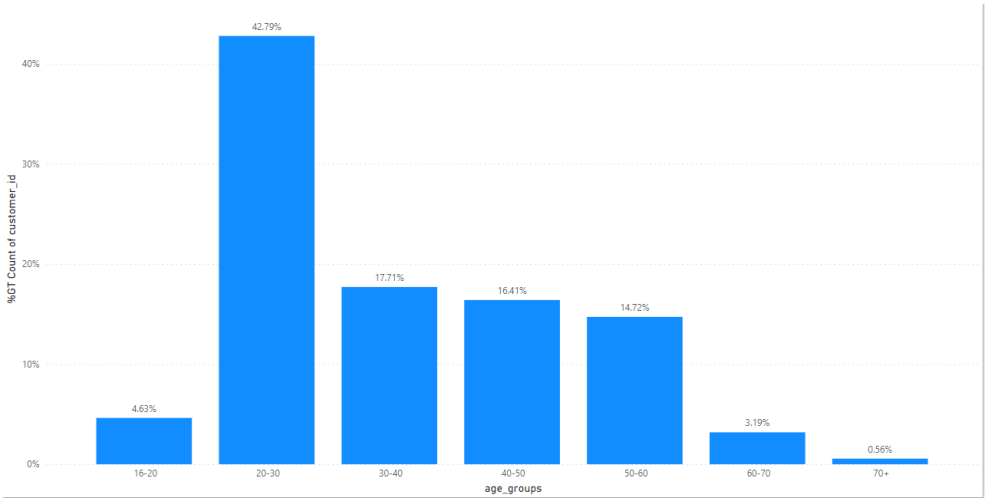

Figure 12. Customer age group distribution

Customers in the range **20-30** are responsible for more than **42%** of the total purchased products.

Customers in the range **16-20, 60-70 and 70+** are responsible for the **8%** of the total purchased products

Customers in the range **30-40, 40-50 and 50-60** are responsible for **16%** of purchased quantity each.

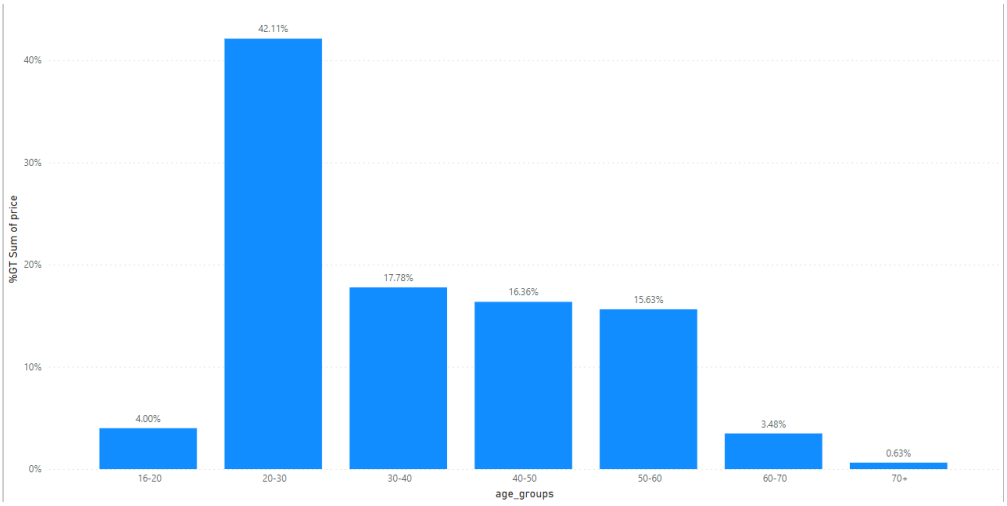

Figure 13. Customer's spend by age group

Indeed a very similar situation to the purchases quantity can be found in the earnings analysis, since customers who buys more, on average leads to higher earnings for the company.
The age group **20-30** is by far responsible for the highest earnings for the company (**41.9% of total earnings**).

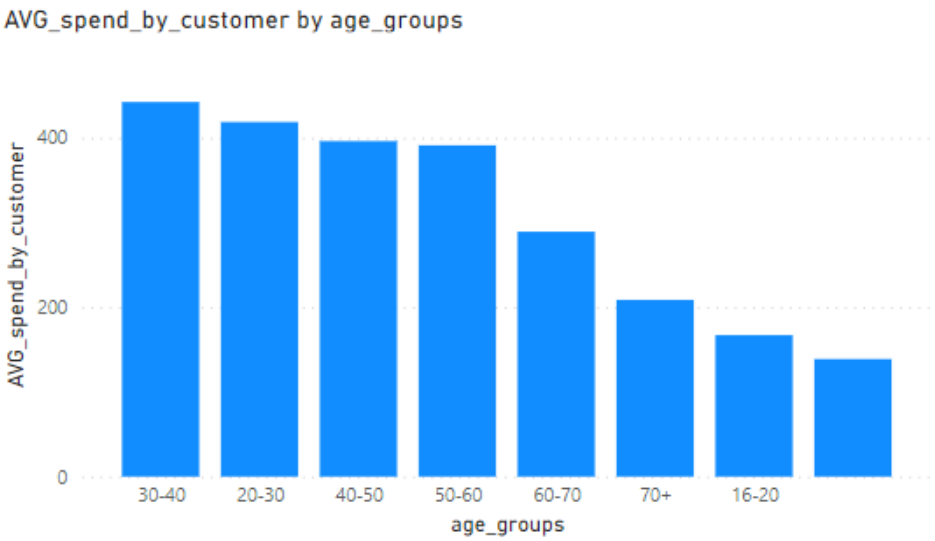

Figure 14. Average Customer Spend by age_group

Although the 20-30 age group bring in the highest revenue, the 30-40 age group has a higher average spend.

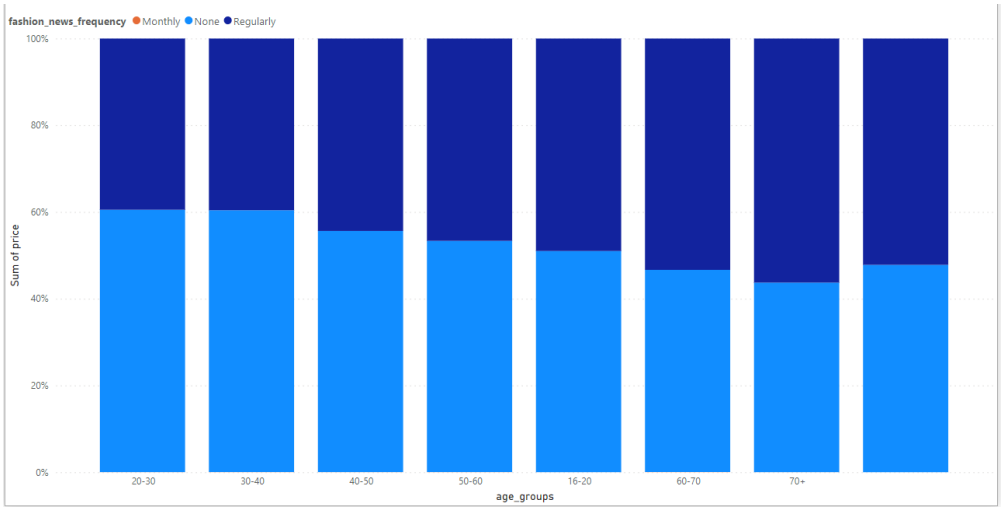

Figure 15. Customer's spend by age group and fashion new frequency

We can see that customers in the range 20-30 and 30-40 have the lowest percentage of fashion news frequency, while being the groups which buy the most**.**
Moreover**,** the frequency of customer that regulary check fashion news starts increasing from the range 40-50, with a peak value of 43.7% of regular/active users for customers in the range 70+ years old. *This means that checking fashion news seems to be more effective for older customers, who still represent a small percentage of total sold products, while younger customers do not need to check the news to buy new products***.**
It could be effective for the company to invite younger customers (range 20-40) to check the news more frequently in order to increase the sold items.

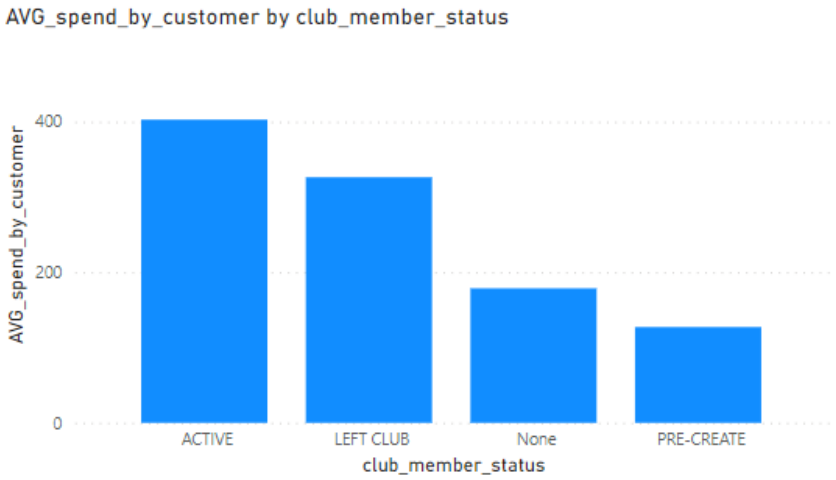

Figure 16. Average Customer Spend by club_member_status

Left club user have the **second highest average spend** although the percentage is very low. This indicates that customers with large purchases left for some reason.

5. Transactions Analysis

5.1 Overall business situation

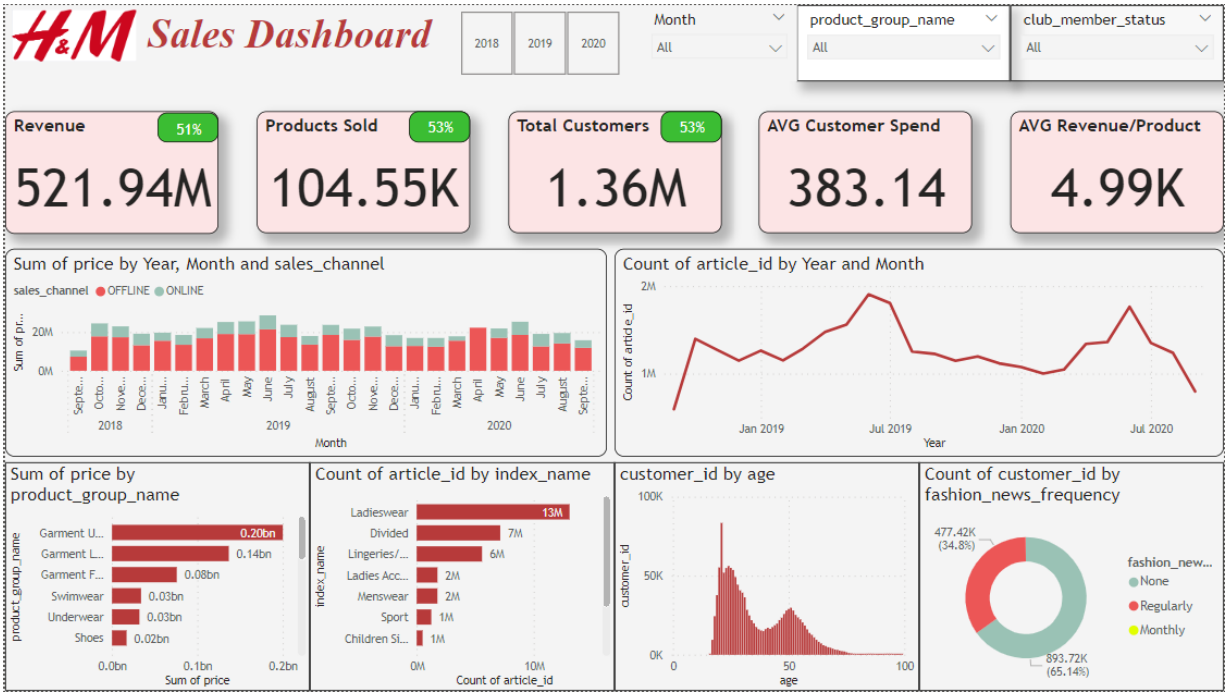

Figure 17. Overall dashboard

Total Revenue: **$521.94M** and YoY increase **51%** over the previous year.

Total Product Sold: **104.55K** and YoY increase **53%** over previous year.

Total Customer: **1.36M** and YoY increase **53%** over previous year.

Each customer spends around **$383.14** and each product brings the company about **$5K** in revenue.

The monthly sales have been quite consistent all year round, it averages about **11-12k orders per month**. The total sales start to pick up from April onwards and peak in June before declining back to the average level in August. Likewise, revenue averages around **$20M** per month before picking up April onwards and peak in June/July with an average of **$25-28M**, ***meaning that customers purchased more in the summer.***

We can observe that in **April 2020**, there is **no sales in ONLINE channel**, maybe the website is down at that time or this is an error when input the data.

Product group name: The majority of revenue comes from **Garment** (upper, lower, and full body).

Index name: **Ladieswear** is the most sold.

5.2 Top 50 most sold product and their characteristics

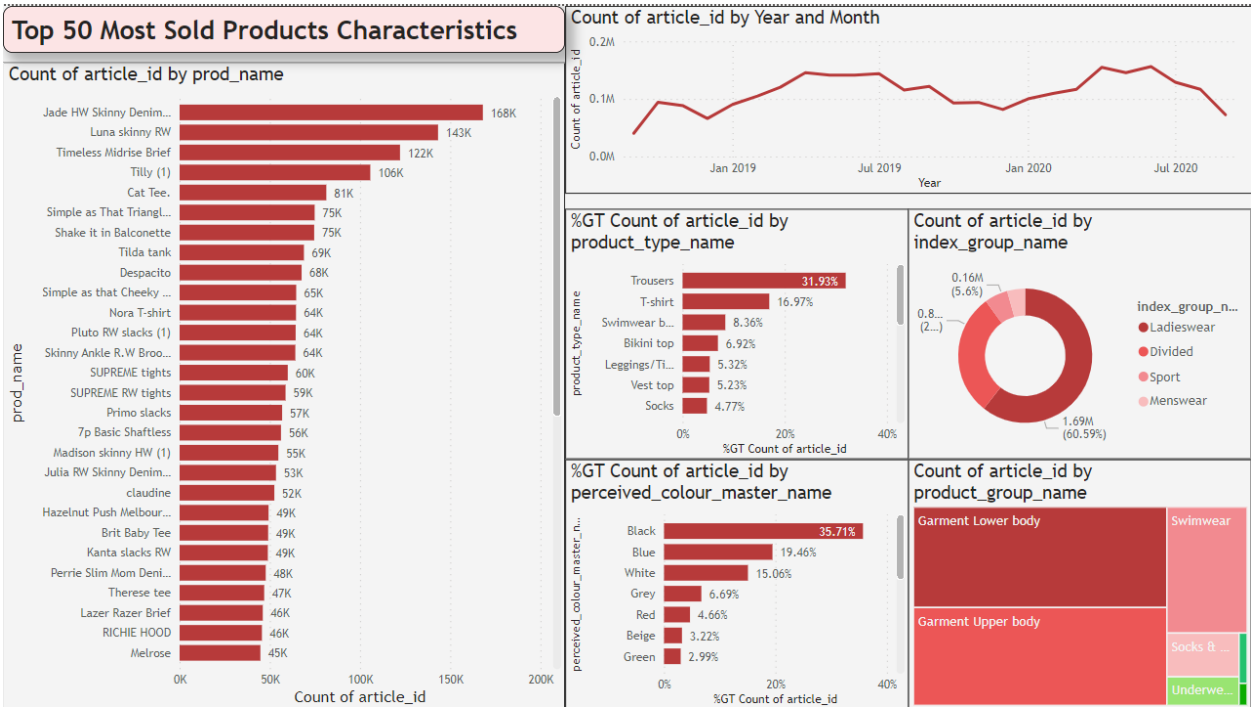

Figure 18. Top 50 most sold products characteristics

Among the TOP 50 of solds products:

Almost **32%** of sold products are **Trousers**

60% is **Ladieswear**, **30%** is **Divided**

Black, blue and white are the three most popular colors

Over **75%** are related to **Garment**

5.3 Top 50 highest revenue and their characteristics

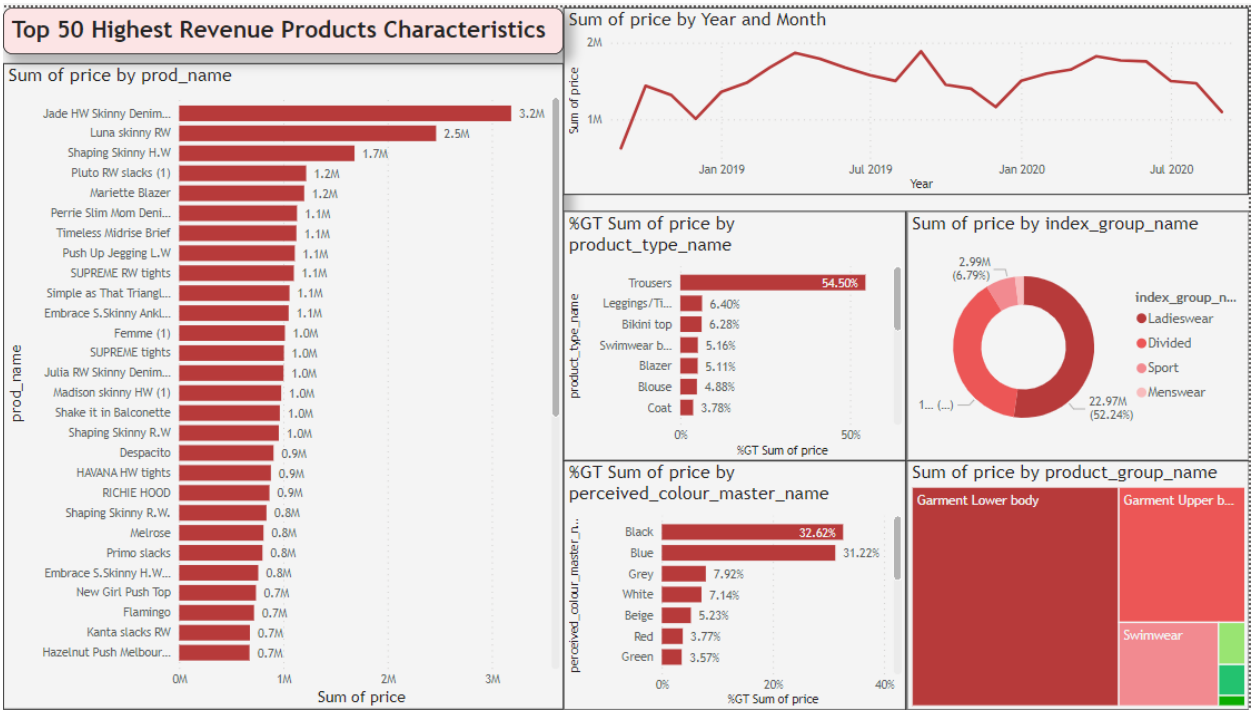

Figure 19. Top 50 highest revenue products characteristics

Among the TOP 50 highest revenue products:

54.5% of the TOP 50 products in terms of earnings are generated by selling **Trousers**

52% is **Ladieswear** and almost **40%** is **Divided** (*similar to top 50 most sold products*)

Over **60%** of the products are **black and blue**

62% of the products are related to **lower body**

Conclusion:

Trousers are the product that sells the most and also brings in the highest revenue

White products do not bring high revenue even though they have high sales volume

Garment lower body have the same sales volume as garment upper body but bring in 2.7 times more revenue.

5.4 Product Sales Seasonality

Are all items selling well in June?

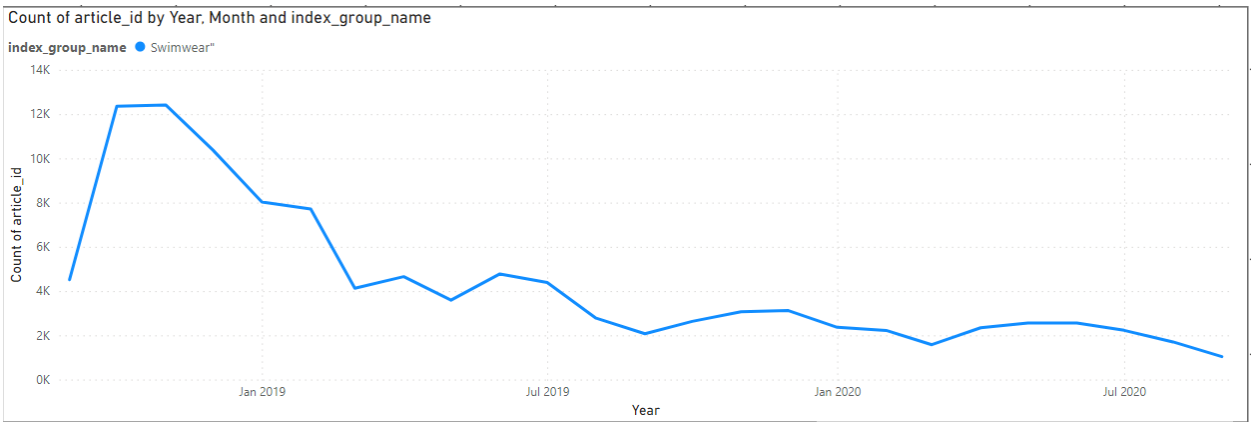

Figure 20. Swimwear sales sales by time

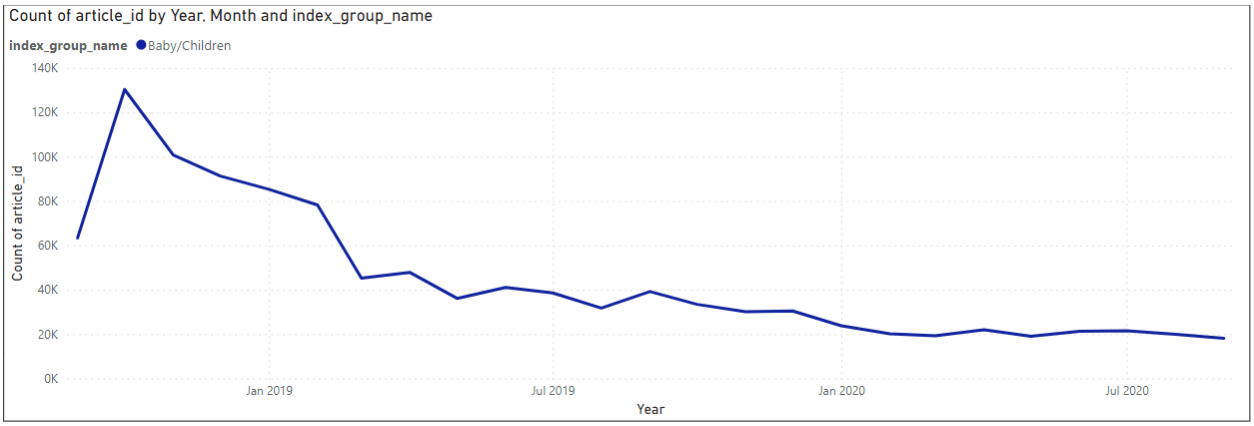

Figure 21. Baby/Children sales by time

Swimwear and Baby/Children are sold the most in the **forth quarter of 2018** and gradually descreased after that. This means that customers found another site/store and no longer buy these two items at H&M.

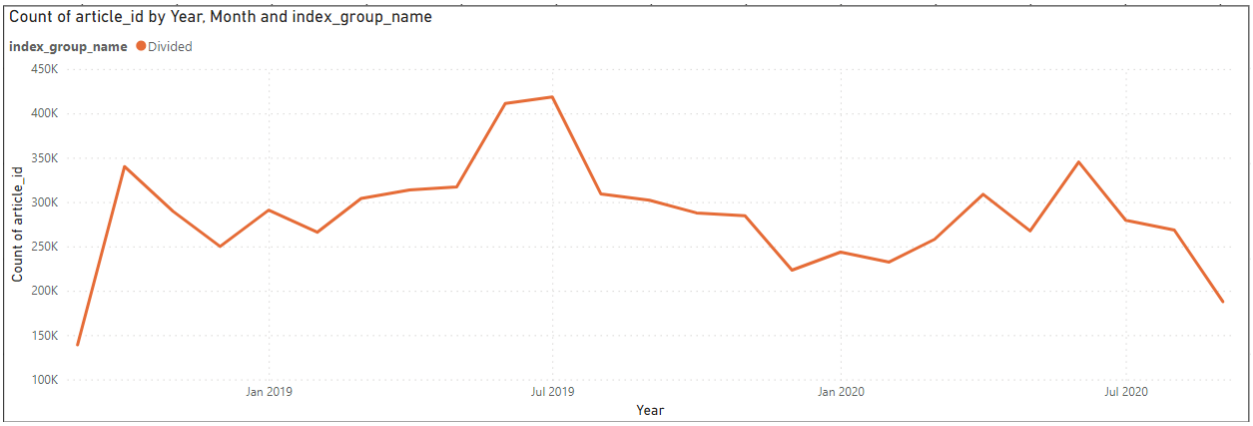

Figure 22. Divided sales by time

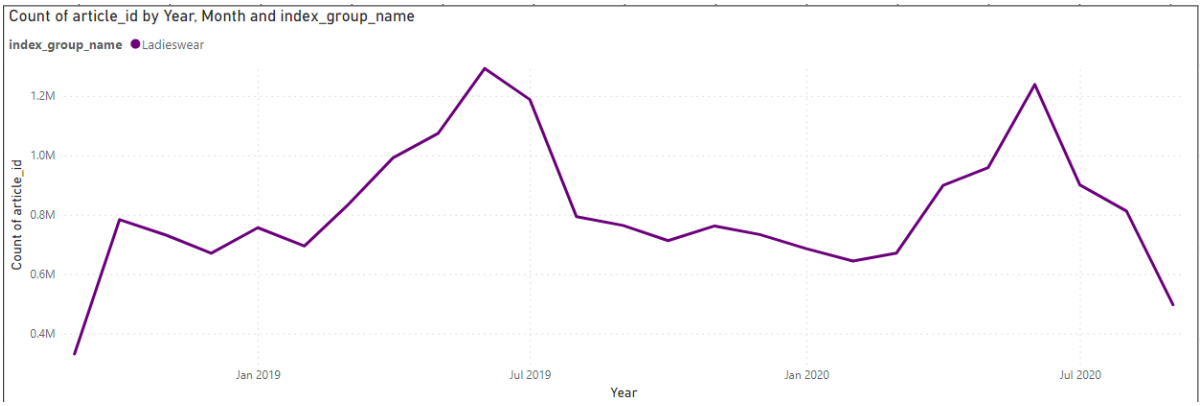

Figure 23. Ladieswear sales by time

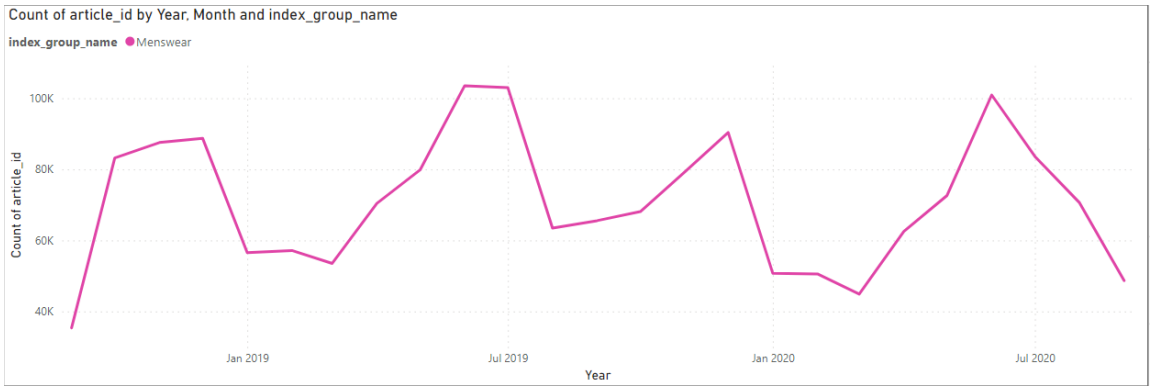

Figure 24. Menswear sales by time

Divided, Ladieswear and Menswear sales are most sold in **June** (*same with overall sales*), indicating that these products are highly correlated with overall sales. Actions to increase sales of this product will help the business revenue grow rapidly.

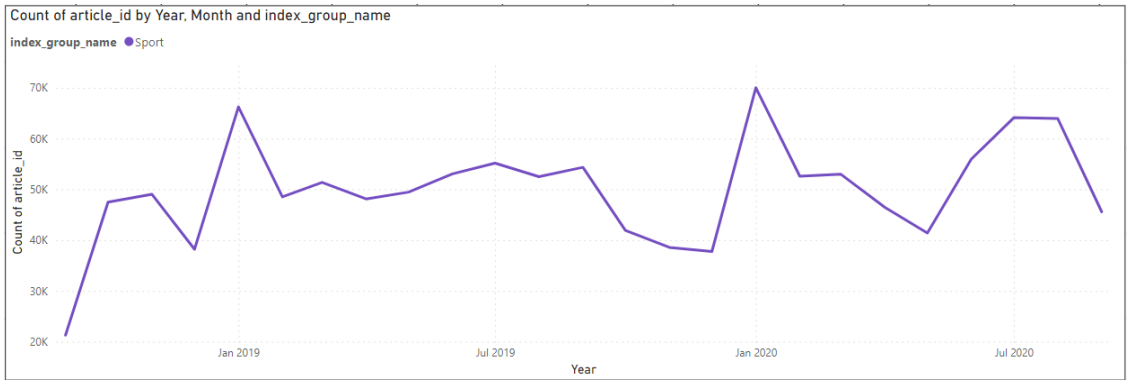

Figure 25. Sport sales by time

Sport is sold the most in **January** so we need to launch marketing campaigns focusing on Sportwear in December or earlier to sell more swimsuits in January.

5.5 The product is repurchased many times

Does customers buy the extract the same products (article_id) again in within a few weeks?

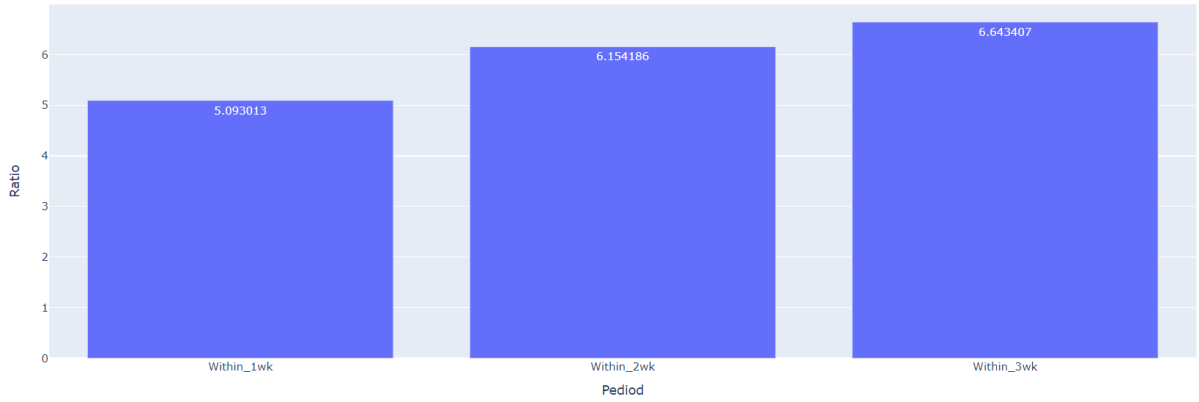

Figure 26. Percentage of customers who repurchase a product (article_id)

5.1% of customers will buy the same product in one week

6.2% will buy the same product within two weeks

6.6% will buy the same product within three weeks

Do customers buy different colors and sizes of the same product?

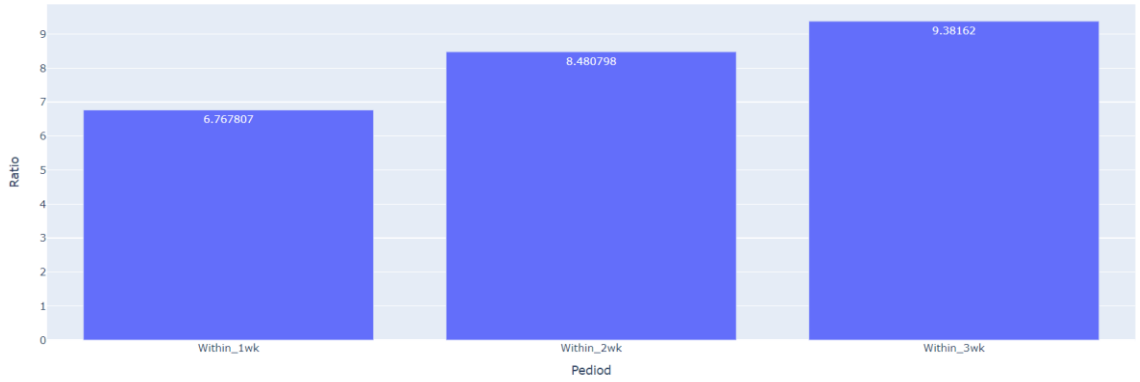

Figure 27. Percentage of customer who repurchase a product with different colors and sizes

6.8% of customers will buy the same product code item in one week

8.5% will buy the same product code item within two weeks

9.4% will buy the same product code item within three weeks

Customers seem to need a little more time when buying products with the same product code but different colors and sizes.

Do customers buy the same product type?

Same product type here is same index_group_name, index_name and product_type_name

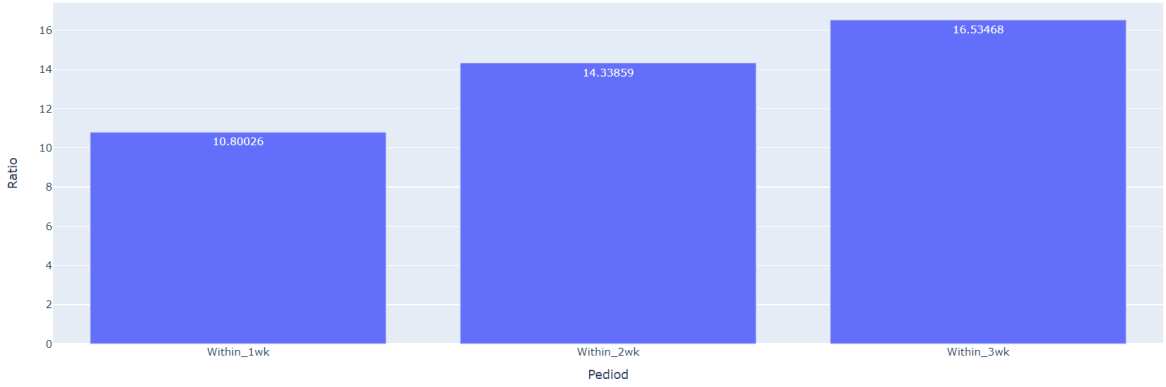

Figure 28. Percentage of customer who buy the same product type

10.8% of customers will buy the same product type item in one week

14.3% will buy the same product type item within two weeks

16.5% will buy the same product type item within three weeks

16.5% is huge! If we observe the purchasing behavior of customers over a short period of time, they seem to buy multiple similar products.

6. Customer Cohort Analysis

Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
January	242024	91634	95766	98617	99279	107084	103376	86416	88404	88811	90712	88544	1280667
February	147358	43885	45417	45671	49944	47909	38517	39475	39658	40670	39309		577813
March	104515	29604	29290	32151	30619	23883	25146	25247	26129	24714			351298
April	87237	22130	23539	22401	17197	17471	17981	18498	17708				244162
May	75184	19058	17246	13012	13206	13405	13902	13720					178733
June	73414	16794	10994	11047	11241	11619	11458						146567
July	58338	8983	7936	8278	8514	8708							100757
August	35889	5496	5319	5276	5278								57258
September	35865	6073	5530	5344									52812
October	38133	5970	5342										49445
November	43249	6530											49779
December	35595												35595
Total	976801	256157	246379	241797	235278	230079	210380	183356	171899	154195	130021	88544	3124886

Figure 29. Customer Retention

Month	1	2	3	4	5	6	7	8	9	10	11	12	Total
January	1.00	0.38	0.40	0.41	0.41	0.44	0.43	0.36	0.37	0.37	0.37	0.37	5.29
February	1.00	0.30	0.31	0.31	0.34	0.33	0.26	0.27	0.27	0.28	0.27		3.92
March	1.00	0.28	0.28	0.31	0.29	0.23	0.24	0.24	0.25	0.24			3.36
April	1.00	0.25	0.27	0.26	0.20	0.20	0.21	0.21	0.20				2.80
May	1.00	0.25	0.23	0.17	0.18	0.18	0.18	0.18					2.38
June	1.00	0.23	0.15	0.15	0.15	0.16	0.16						2.00
July	1.00	0.15	0.14	0.14	0.15	0.15							1.73
August	1.00	0.15	0.15	0.15	0.15								1.60
September	1.00	0.17	0.15	0.15									1.47
October	1.00	0.16	0.14										1.30
November	1.00	0.15											1.15
December	1.00												1.00
Total	12.00	2.48	2.21	2.04	1.86	1.68	1.48	1.26	1.09	0.88	0.64	0.37	27.99

Figure 30. Customer Retention Rate

Customer retention rate in the first quarter of 2018 is very high and consistent. In **January**, the retention rate reached nearly **40%** and remained the same throughout the remaining 11 months. However, the customer retention rate has dropped sharply since **June**, to only about **15%** even though this is the period that brings the highest revenue for the company.

Month	1	2	3	4	5	6	7	8	9	10	11	12
January	19735490	8207869	10090946	11230254	11052168	11926467	9921769	7885664	10408097	9067557	9501218	7375131
February	10334783	3735619	4290839	4277236	4745421	3874555	2935770	3882887	3442011	3537985	2769673	
March	8327359	2530465	2572432	2910564	2366011	1725177	2352438	2110729	2176193	1691317		
April	7215186	1756786	1990885	1634573	1184005	1579451	1431051	1485867	1176342			
May	5922307	1459159	1151411	843832	1069574	1013881	1012867	860149				
June	5670072	1049209	663032	849197	791842	815563	683156					
July	3830882	505059	571203	570462	585562	497119						
August	2279515	397313	354340	364885	319192							
September	2669234	404139	373764	320735								
October	2636743	403769	314942									
November	2666440	373472										
December	2066656											

Figure 31. Customer Retention Revenue

Month	1	2	3	4	5	6	7	8	9	10	11	12
January	1.00	0.42	0.51	0.57	0.56	0.60	0.50	0.40	0.53	0.46	0.48	0.37
February	1.00	0.36	0.42	0.41	0.46	0.37	0.28	0.38	0.33	0.34	0.27	
March	1.00	0.30	0.31	0.35	0.28	0.21	0.28	0.25	0.26	0.20		
April	1.00	0.24	0.28	0.23	0.16	0.22	0.20	0.21	0.16			
May	1.00	0.25	0.19	0.14	0.18	0.17	0.17	0.15				
June	1.00	0.19	0.12	0.15	0.14	0.14	0.12					
July	1.00	0.13	0.15	0.15	0.15	0.13						
August	1.00	0.17	0.16	0.16	0.14							
September	1.00	0.15	0.14	0.12								
October	1.00	0.15	0.12									
November	1.00	0.14										
December	1.00											

Figure 32. Customer Retention Revenue Rate

Revenue is similar to retention, months with high retention also bring in higher revenue and those who are retained tend to generate more revenue (**40% retention rate generated 50% revenue equal to the revenue that new customers (the remaining 60%) bring in January**).

7. Association rule

Association rule mining finds interesting associations and relationships among large sets of data items. This rule shows how frequently a itemset occurs in a transaction

Support

The probability that an item has been purchased.

The degree of support (A) can be calculated by n(A)/n(U).

When both A and B are purchased at the same time, n(A∩B)/n(U) can be calculated.

Calculate support using mlxtend library

```
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

```
support = apriori(df_matrix,min_support=0.1,use_colnames=True)
support
```

Explain the results:

The first row of below dataframe is (Trousers, Sweater) with support score is 0.34, this mean **Trousers and Sweater** are usually bought together and amount of these transaction accounts for **3.4%** of the total transactions. Likewise, the last row is **(Sweater, T-shirt, Bikini Top, Dress)** mean **1%** of total transactions contain these product.

	support	itemsets	length
125	0.343519	(Trousers, Sweater)	2
79	0.312334	(Trousers, Dress)	2
134	0.304123	(T-shirt, Trousers)	2
75	0.284765	(Sweater, Dress)	2
137	0.283035	(Trousers, Top)	2
...	...	...	...
577	0.100196	(Sweater, Dress, Shorts, Vest top, Trousers)	5
385	0.100043	(Trousers, Top, Bikini top, Dress)	4
322	0.100031	(Vest top, Top, Shirt)	3
230	0.100017	(Bra, Sweater, Skirt)	3
379	0.100002	(Sweater, T-shirt, Bikini top, Dress)	4

574 rows × 3 columns

The Top 5 of the result indicate that Trousers are most often first item purchased, so it should be placed in the fronf of the store. Sweater, Dress, T-Shirt and Top should be placed nearby.

The bottom 5 of the result show that there are many related item, so we should bundle overstock or items that don't sell well with our bestsellers.