Hybrid Recommender System for Fashion Articles

(DSAI Mini Project REPA)

Daniel Yang
Joel Lim
Rhys Lie

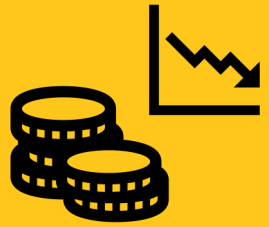
TABLE OF CONTENTS

01 PROBLEM DEFINTION

Background

The e-commerce market has experienced unprecedented growth in the last 5 years. The increasing significance of such a service in this digital age, have forced retailers to extract value from newly unlocked data points

PROBLEM STATEMENT



With the right analysis, the data can be used to better cater to customer preferences to combat declining sales, as well as more accurately predict inventory to minimize inventory costs.

But how should retailers, of any size, approach their data analysis effectively?

Our Solution

We answer the following question which is valuable to fashion retailers.

Which fashion articles will a particular customer be interested in purchasing next?

Customer ID (Input) → **List of Article IDs (Output)**

Key Considerations

1. Trending Market Purchases

2. Age Demographic Habits

3. Individual Historical Purchases

4. AI Recommendation

Hybrid Recommender System

We carry out the filtering of our recommendations in a systematic manner

Collaborative Filtering

Trending Market Purchases

Age Demographic Habits

+

Top 100 Relevant Popular Purchases

Content Based Filtering

Individual Historical Purchases and Preferences

Most relevant few articles

DNN to suggest complementary items

Goal: List of recommended articles for the customer

02

EXPLORATORY DATA ANALYSIS

OUR DATASET

Our chosen dataset contains the following key details from a fashion retailer. The attributes sufficiently represents common data points that would typically be collected. Our dataset covers the time period of year 2020.

Articles

Detailed information on articles

35MB

Transactions

Information on customer transactions

1.24GB

Customers

Detailed information on customers

202MB

EDA APPROACH

Identify most relevant/useful data columns

Use the appropriate charts to visualize trends

Process and understand findings

Articles

What type of articles are trending?

- Fashion categories
- Colours
- Graphical appearances

Transactions

Where are customers spending on and how much?

- Transaction frequency
- Average quantum

Customers

How can we best profile customers to learn their preferences?

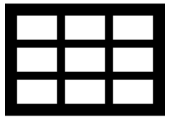
- Age

DATA WRANGLING METHODOLOGY

Fortunately, our dataset was relatively well populated and had almost no empty value that we had to drop or replace. Instead, we took a few other steps to make our subsequent data analysis more efficient

Conversion to more suitable data types

E.g., Pandas datetime



Selecting only necessary columns for queries

E.g., df = transactiondata[[‘price’, ‘customer_id’]]

Using Log scale when necessary for better visualizations

ARTICLES EDA

ARTICLES ATTRIBUTES

article_id

unique identifier of article

prod_name

product name

prod_type_name

product type name

graphical_appearance_name

graphical appearance of article

colour_group_name

colour group of article

perceived_colour_value_name

perceived colour of article

ARTICLES ATTRIBUTES

perceived_colour_master_name

perceived master colour of article

department_name

department of article

index_name

index name of article

index_group_name

index group name of article

section_name

section of article

garment_group_name

garment group of article

FIRST LOOK AT ARTICLES

Categories	Count
Ladieswear	39737
Baby/Children	34711
Divided	15149
Menswear	12553
Sport	3392

index_group_name
(5 rows)

Categories	Count
Ladieswear	26001
Divided	15149
Menswear	12553
Children Sizes 92-140	12007
Children Sizes 134-170	9214
Baby Sizes 50-98	8875
Ladies Accessories	6961
Lingeries/Tights	6775
Children Accessories, Swimwear	4615
Sport	3392

index_name
(10 rows)

Categories	Count
Garment Upper body	42741
Garment Lower body	19812
Garment Full body	13292
Accessories	11158
Underwear	5490
Shoes	5283
Swimwear	3127
Socks & Tights	2442
Nightwear	1899
Unknown	121
Underwear/nightwear	54
Cosmetic	49
Bags	25
Items	17
Furniture	13
Garment and Shoe care	9
Stationery	5
Interior textile	3
Fun	2

product_group_name
(19 rows)

Categories	Count
Trousers	11169
Dress	10362
Sweater	9302
T-shirt	7904
Top	4155
...	...
Bra extender	1
Blanket	1
Towel	1
Wood balls	1
Cushion	1

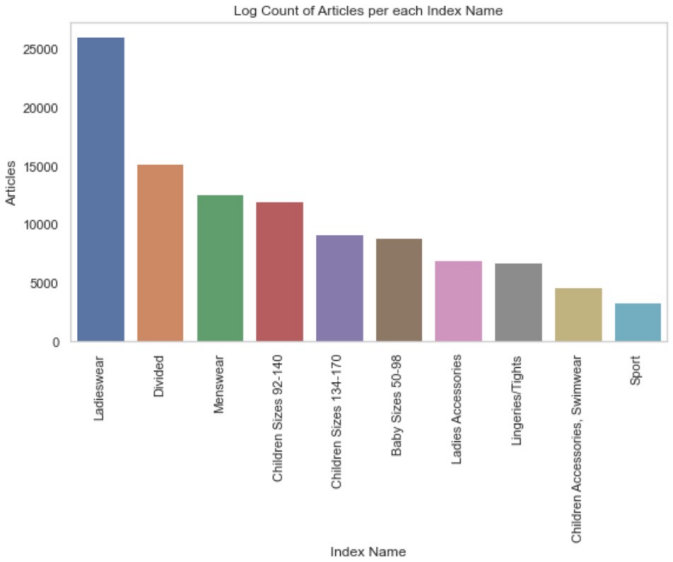
product_type_name
(131 rows)

OBVIOUS CHOICE OF PREFERRED INDEX GROUPS

Now, we have a better idea on the index group of articles in this fashion retailer.

We see that their articles cater more for baby/ children and ladies, with sports attire having the least number of articles

DELVING INTO INDEXES

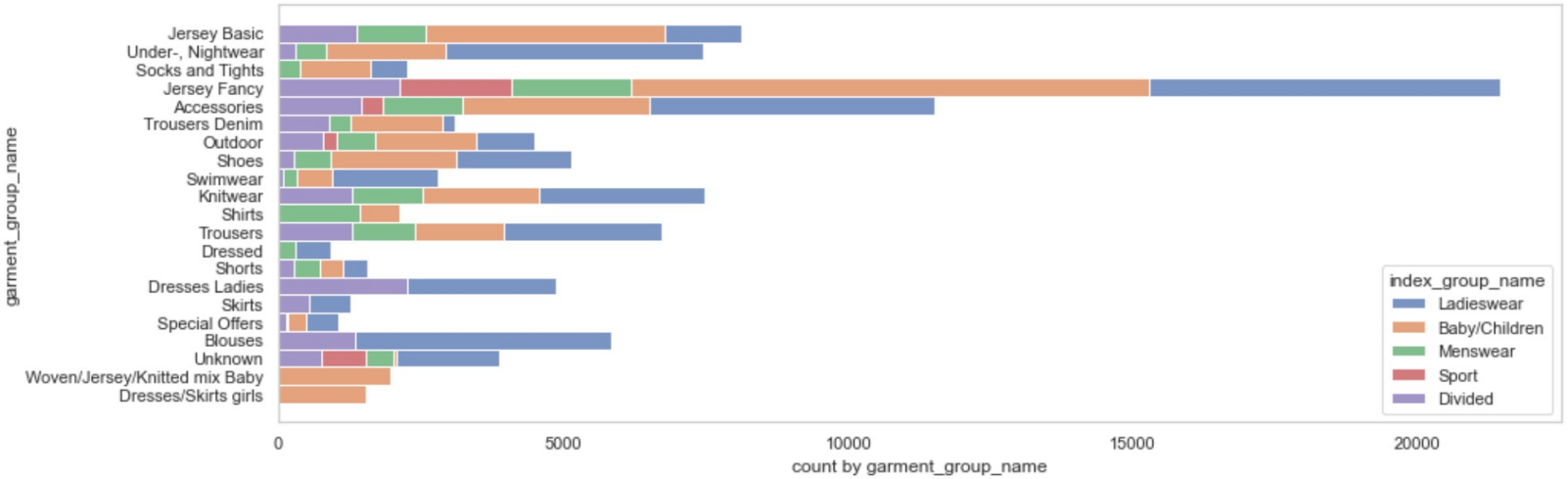


index_group_name	index_name	article_id
Baby/Children	Baby Sizes 50-98	8875
	Children Accessories, Swimwear	4615
	Children Sizes 134-170	9214
	Children Sizes 92-140	12007
Divided	Divided	15149
Ladieswear	Ladies Accessories	6961
	Ladieswear	26001
	Lingeries/Tights	6775
Menswear	Menswear	12553
Sport	Sport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

Delving deeper into the indexes, we can see that the baby/children index group is split into various indexes catering to babies, children of different sizes, and one for accessories and swimwear.

Similarly, Ladieswear is split into 3 indexes, one for accessories, one for ladieswear and one for lingeries/ tights.

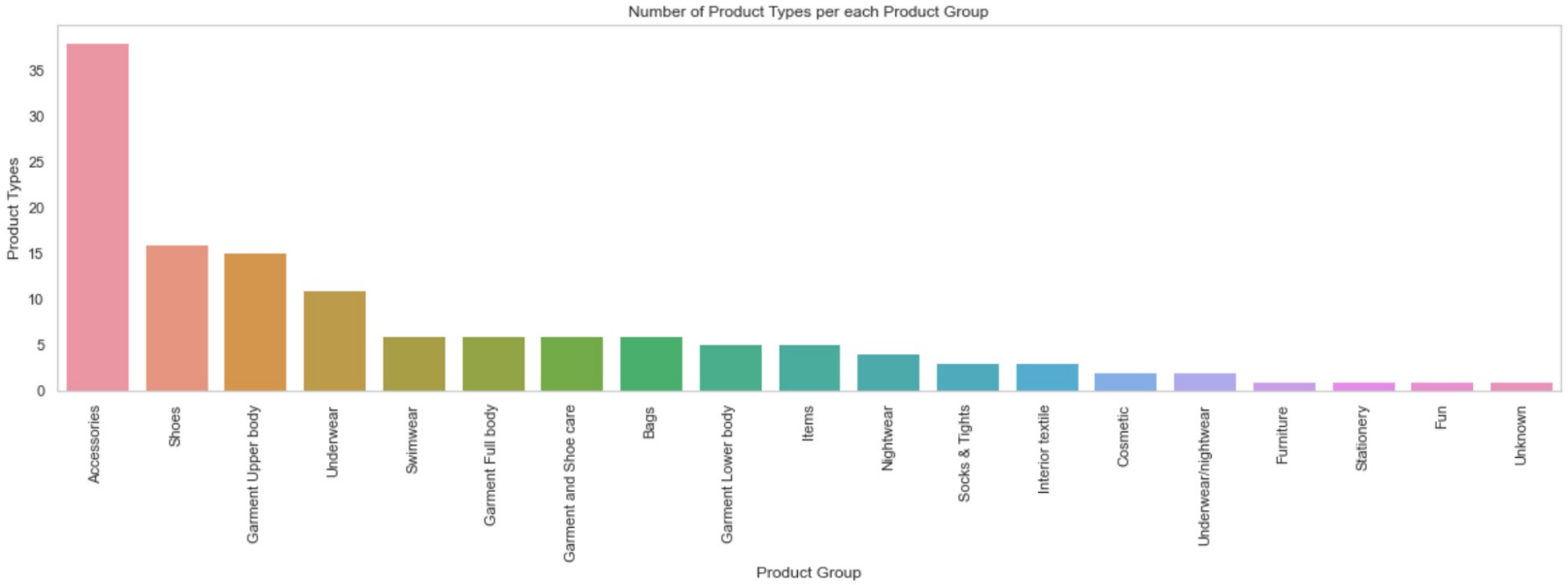
MORE ON GARMENT GROUPS



The garment group is also another interesting attribute to look at. We see that garment group defines the type of clothing the article refers to, and could span multiple categories. ‘Jersey Fancy’ and ‘Accessories’ are categories with the most unique articles.

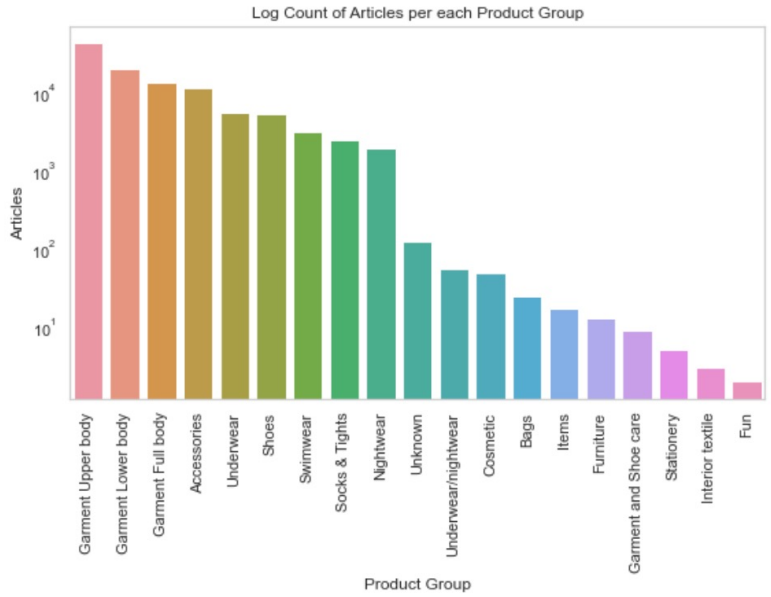
Some garment groups cater to specific index groups only (such as the baby/ girls articles).

PRODUCT TYPES WITHIN PRODUCT GROUPS

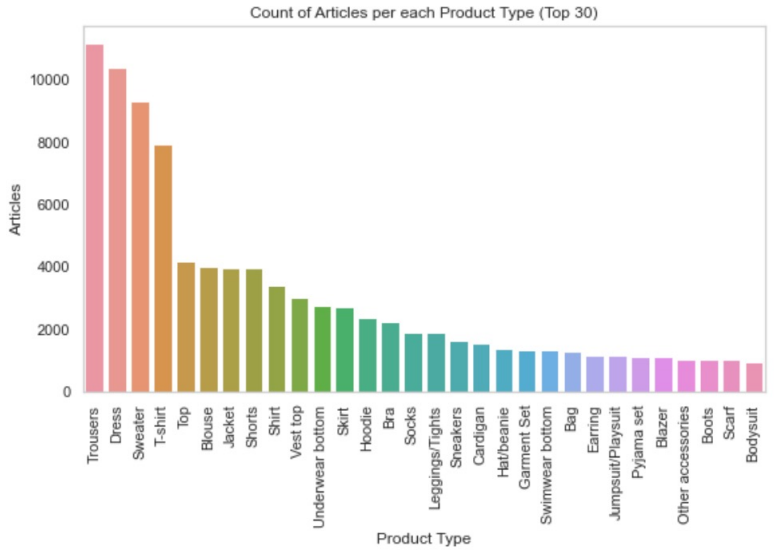


We see that within the product groups, accessories has the most number of unique product types.

COUNT OF ARTICLES PER PRODUCT GROUP/ TYPE



As for product groups, garments has the most number of unique articles.



In terms of product types, we see that trousers, dresses and sweaters make up the top 3 in terms of counts of articles for each product type.

TRANSACTIONS EDA

TRANSACTIONS ATTRIBUTES

t_dat

transaction date of purchases

customer_id

unique identifier of customer (foreign key)

article_id

unique identifier of article (foreign key)

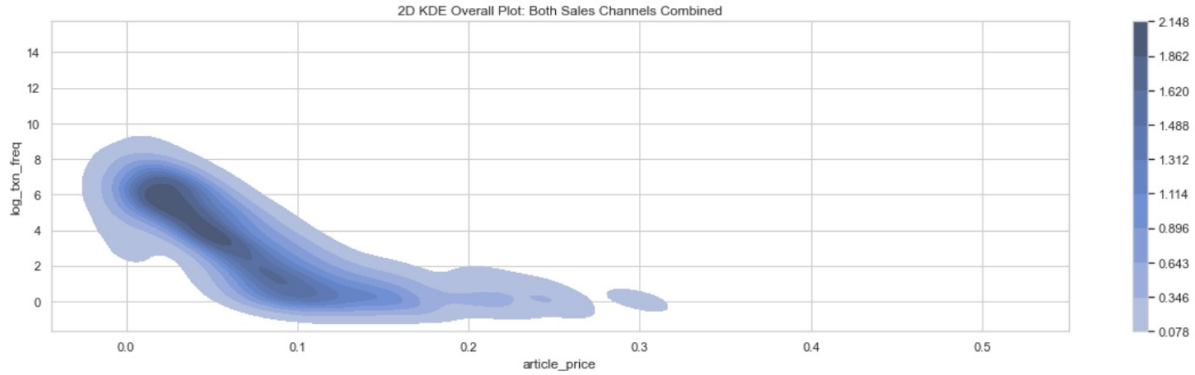
price

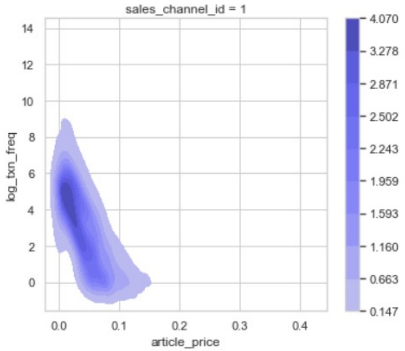
price of purchases

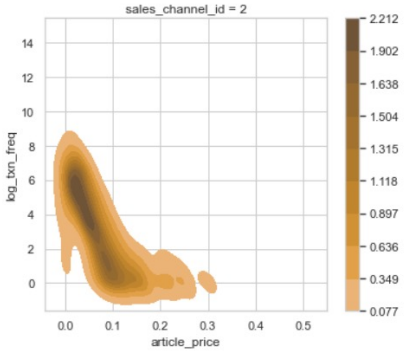
sales_channel_id

1 or 2 (channel in which purchases are made)

PRICE SPREAD OF ALL TRANSACTIONS



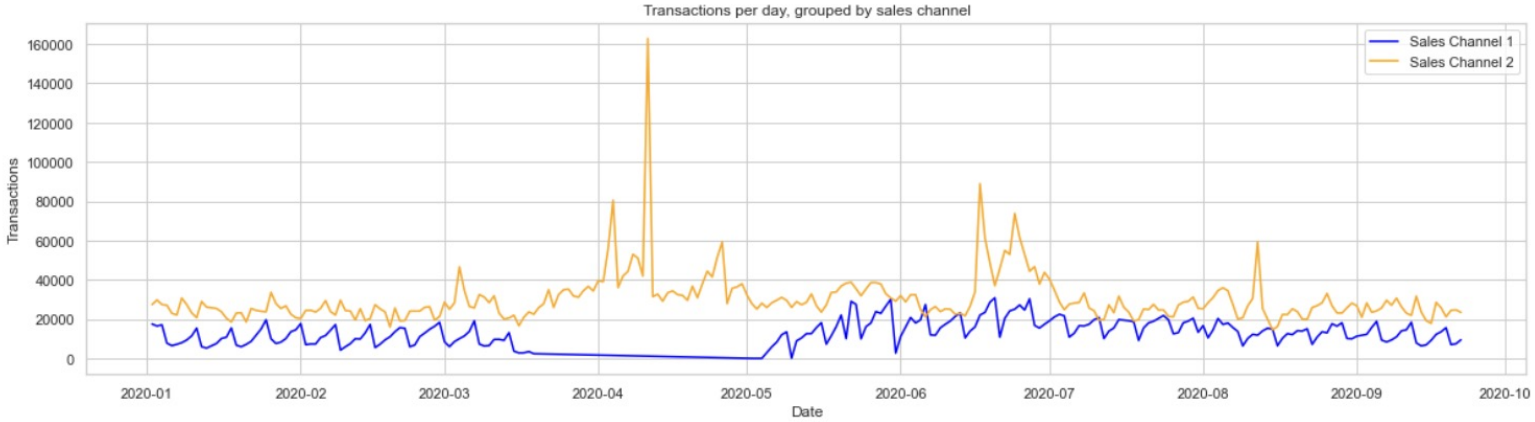




Using a 2D Kernel Density Estimation plot, we can clearly see the spread of prices for all the transactions.

We observe that most transactions have a price quantum of about 0.02-0.03. This applies for both sales channels

DISTRIBUTION OF TRANSACTIONS OVER TIME



Looking at the overall trend of number of transactions over time, we see a cyclically trend that increases and decreases almost consistently across each month. This will allow us to estimate the number of transactions that will likely take place given a specific time period.

At the same time, we see a spike in transaction in Sales Channel 2 from March to May, while there was a steep decline for Sales Channel 1. We can deduce that Sales Channel 1 could be an online channel that went offline for that period for maintenance etc.

CUSTOMERS EDA

CUSTOMER ATTRIBUTES

customer_id

unique identifier of customer

fashion_news_frequency

frequency that H&M sends fashion news to customers

Active

0 or 1

club_member_status

membership status of customer

FN

0 or 1

age

age of customer

postal code

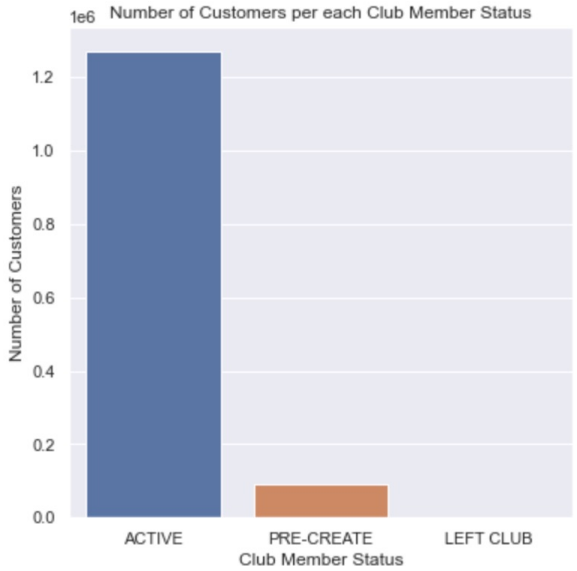
postal code of customer

POSTAL CODE

	Postal Code	Number of Customers
61034	2c29ae653a9282cce4151bd87643c907644e09541abc28...	120303
281937	cc4ed85e30f4977dae47662ddc468cd2eec11472de6fac...	261
156090	714976379549eb90aae4a71bca6c7402cc646ae7c40f6c...	159
171208	7c1fa3b0ec1d37ce2c3f34f63bd792f3b4494f324b6be5...	157
126228	5b7eb31eabebd3277de632b82267286d847fd5d44287ee...	156

Here we have abnormally large number of customers by one postal code Postal Code starting with "2c29ae653a9282cce4151bd87643c907644e09541abc28..." has 120303 customers, it could be the address of a huge distribution center, or secondary retailer.

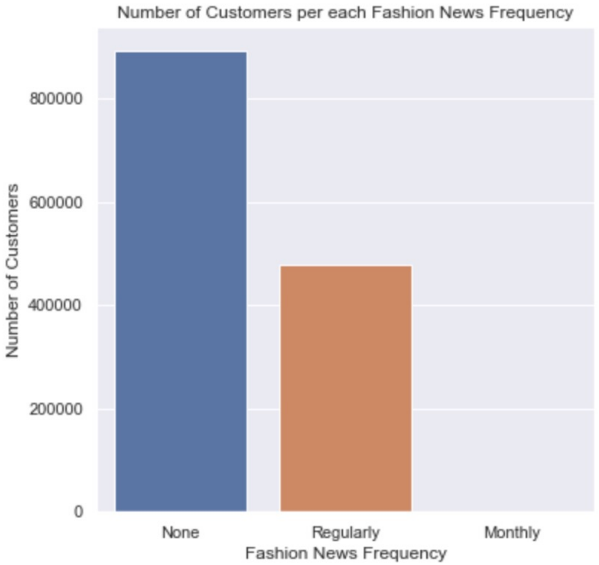
CLUB MEMBER STATUS



	Club Member Status	Number of Customers
0	ACTIVE	1272491
2	PRE-CREATE	92960
1	LEFT CLUB	467

Status of the H&M Club. Most of the customers are active members of the club, with a small proportion are in the pre-creation status. A very small minority have left the club

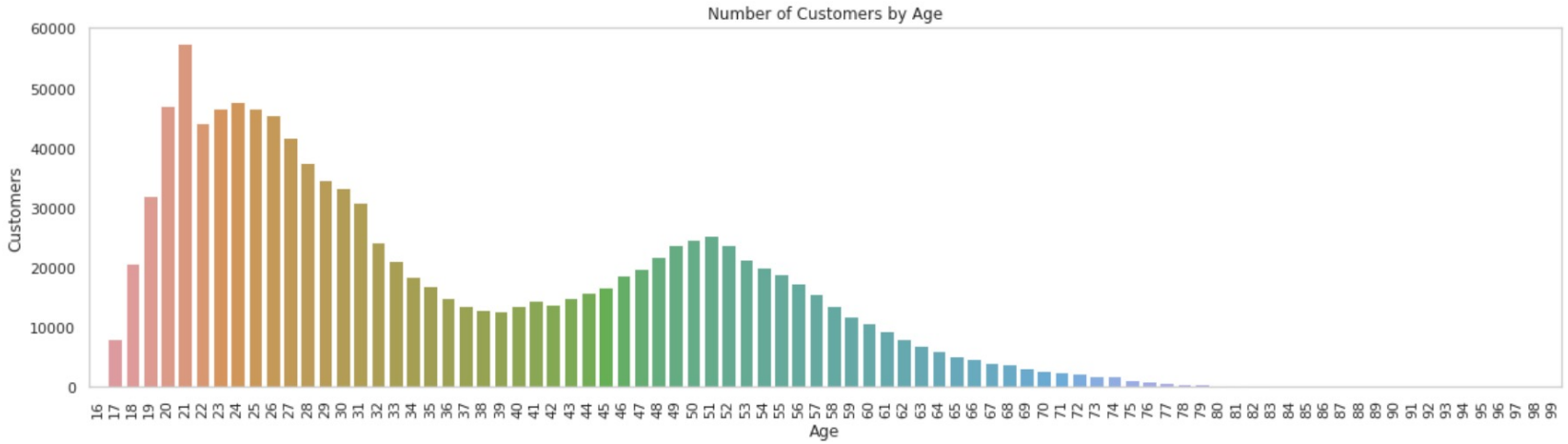
FASHION NEWS FREQUENCY



	Fashion News Frequency	Number of Customers
1	None	893722
2	Regularly	477416
0	Monthly	842

We see that most customers do not fancy fashion news, only an extremely small proportion want monthly fashion news, with many abstaining from it.

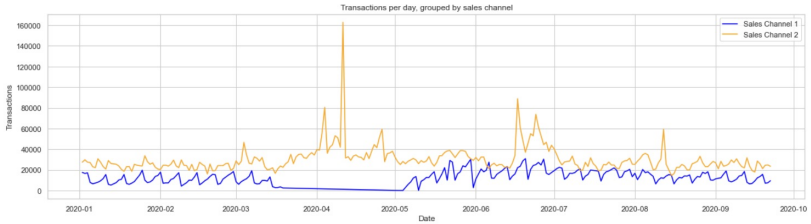
CUSTOMER DISTRIBUTION BY AGE



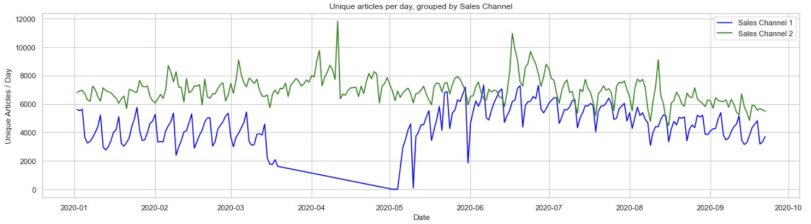
We see that the number of customers peak at early 20s and 50s.

03 IN-DEPTH ANALYSIS

ANALYZING TRANSACTIONS AND ARTICLES



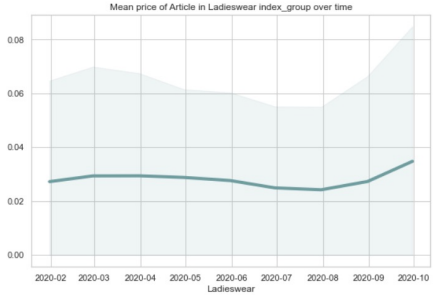
Unique Transactions over time

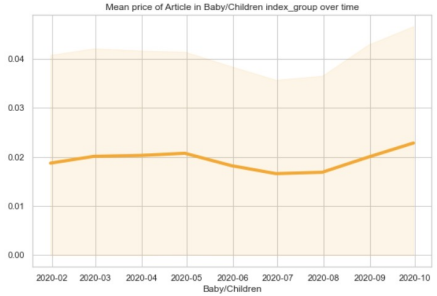


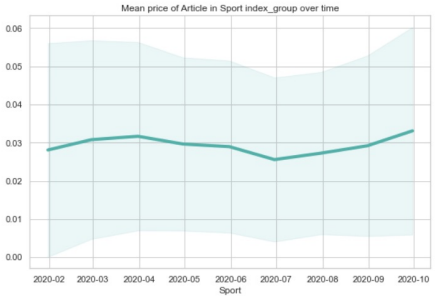
Unique Articles sold over time

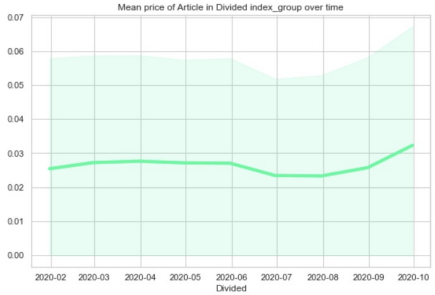
We observe a similar trend between the number of unique articles sold and transactions that take place over time. Generally, it makes sense and tells us that the transactions were of a diverse range of articles rather than a select few extremely popular items.

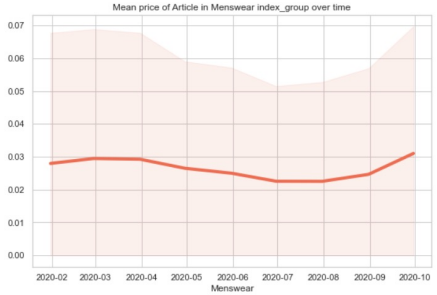
PRICE DISTRIBUTION OF ARTICLES





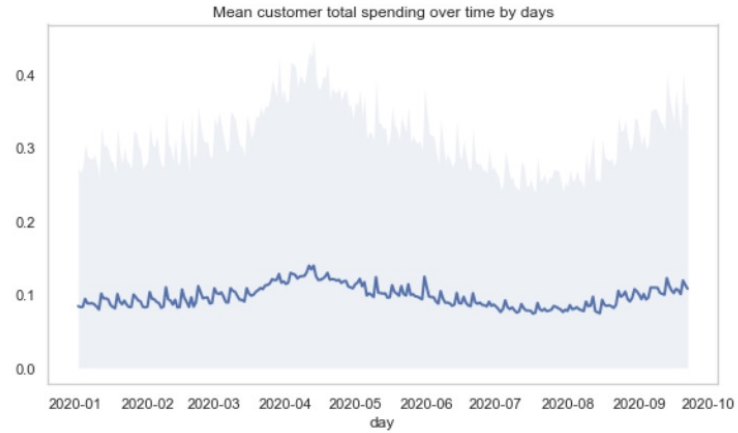




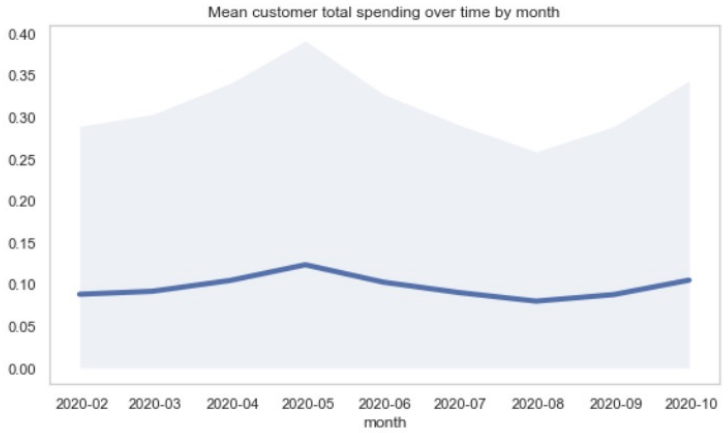


We observe that the average price of most articles, categorized by index_group moved in synch. Their prices did not fluctuate much apart from a dip in July of 2020 before trending back up. This rather stable prices throughout the year is expected and will also allow us to make proper estimations about future articles prices.

DISTRIBUTION OF CUSTOMER SPENDING OVER TIME



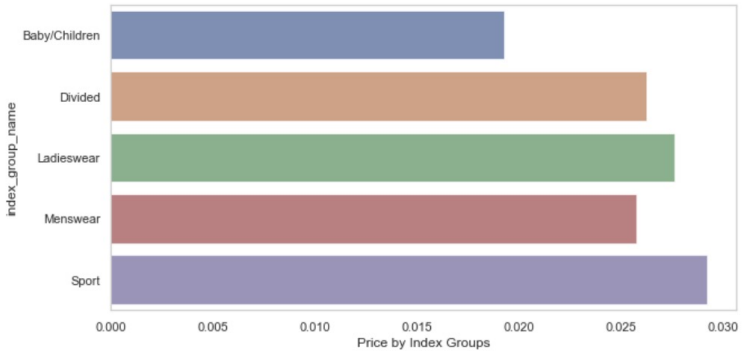
Filtered by Day

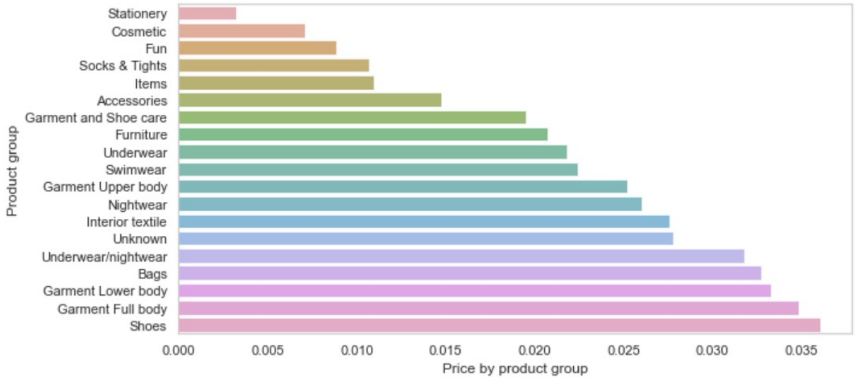


Filtered by Month

We see that customer spending stayed approximately constant throughout the year. It increased corresponding to the spike in unique transactions and decreased corresponding to the dip in article price as seen earlier. We were managed to link these observations to form a more cohesive picture of how the business of the fashion retailer is doing.

MEAN PRICE ACROSS INDEX AND PRODUCT GROUPS





We observe that sport and shoes products are slightly pricier.

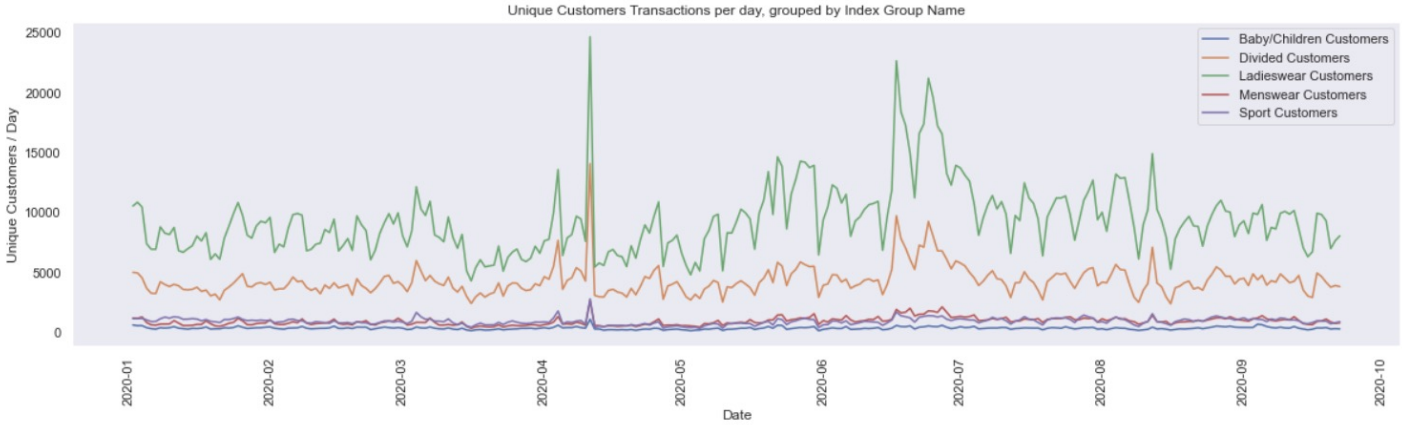
IDENTIFYING TOP PAYING AND TRANSACTING CUSTOMERS

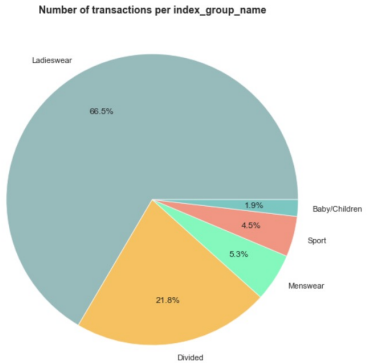
customer_id	Total Spending
863f0e03da282ae32a76775ce55d8a4605a85c84a26066e1ad0e9469e8c40e68	33.07593
b637a3e7d8b0caa947aaefd609b8d84a9ee962cf0a52a51bac507ffc2bf1b741	29.78998
be1981ab818cf4ef6765b2ecaea7a2cbf14ccd6e8a7ee985513d9e8e53c6d91b	21.62464
03d0011487606c37c1b1ed147fc72f285a50c05f00b9712e0fc3da400c864296	21.45612
77db96923d20d40532eba0020b55cd91eb51358885c2d698a2805e79481f64a1	21.38700
a65f77281a528bf5c1e9f270141d601d116e1df33bf9df512f495ee06647a9cc	20.37000
7f0ac4394297dc4a885d3b9277ba526cbbfbf7fb7cae465b256ed8e55b864f03	19.87612
b4db5e5259234574edfff958e170fe3a5e13b6f146752ca066abca3c156acc71	18.56707
e238725cbff3774b711407cc000f42c0ddabf6b07eb0e311ffb5fc72e862a34b	18.20598
a3ab708684132c6bbd3dad7aa41f9b9c7d1c95d7d5cb1a3a052905191e858566	17.84576

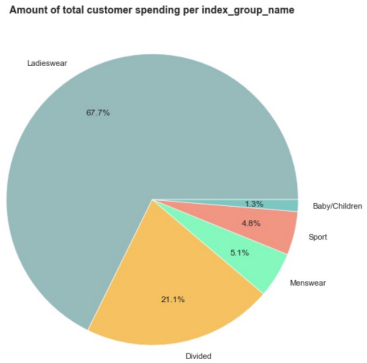
customer_id	Total Transactions
b637a3e7d8b0caa947aaefd609b8d84a9ee962cf0a52a51bac507ffc2bf1b741	790
be1981ab818cf4ef6765b2ecaea7a2cbf14ccd6e8a7ee985513d9e8e53c6d91b	738
4308983955108b3af43ec57f0557211e44462a5633238351fff14c8b51f16093	643
67931690bdf18d2e328854ae772cd5ce2505fdc11164693998b13e706db0bb56	616
a65f77281a528bf5c1e9f270141d601d116e1df33bf9df512f495ee06647a9cc	613
03fdb0bf2d9ff8ba23e1b4aef53709119aad5bc83691d89293a01a52b93d7370	590
f874a19b8d3417b8a7eeffa3cecd595a5f6383e0876da285390dd9c2727e905d	587
b4db5e5259234574edfff958e170fe3a5e13b6f146752ca066abca3c156acc71	584
7f0ac4394297dc4a885d3b9277ba526cbbfbf7fb7cae465b256ed8e55b864f03	543
863f0e03da282ae32a76775ce55d8a4605a85c84a26066e1ad0e9469e8c40e68	539

We see that there are a few customers which appear as the greatest number of total transactions as well as total spending. The lack of overlaps signify customers who may have fewer transactions but of a higher quantum!

DISTRIBUTION OF UNIQUE TRANSACTIONS BY INDEX GROUP

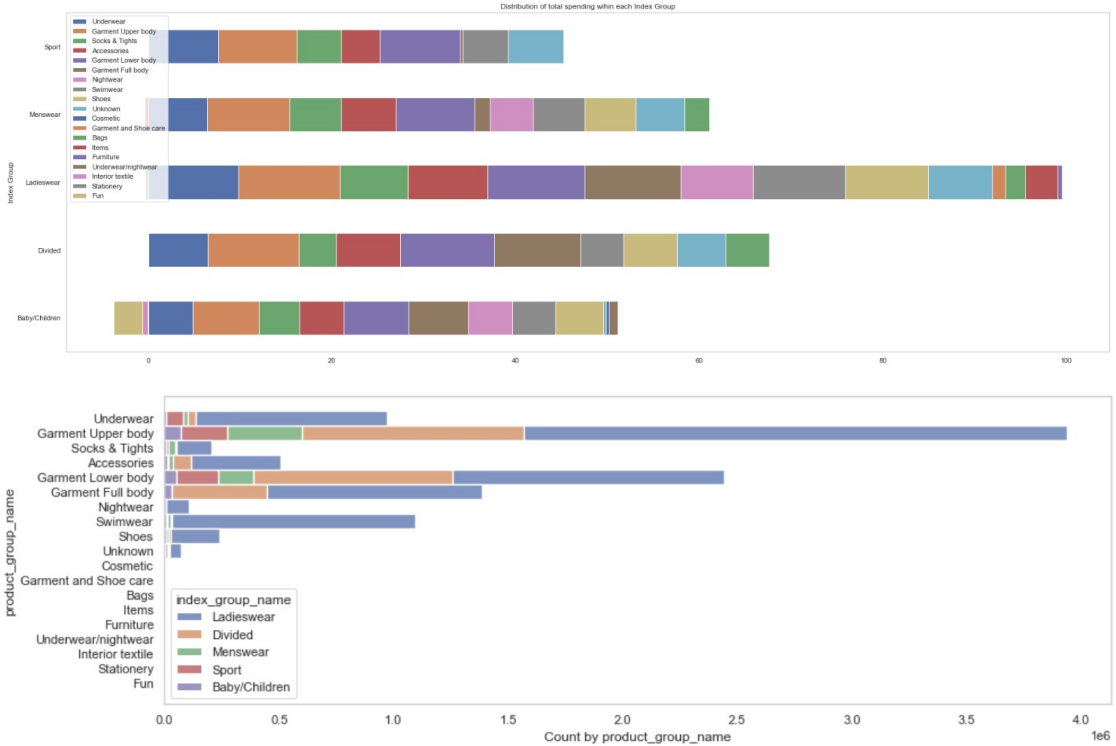






We see that ladieswear takes up the bulk of transactions, in terms of quantity and spending. This is followed by divided, which targets teens.

DIVING DEEPER INTO PRODUCT GROUPS



We see that ladieswear garment upper body, full body and underwear takes up a huge portion of spending and product quantity.

04

HYBRID RECOMMENDER SYSTEM - “TRY SOMETHING NEW!”

OUR PROCESS

Step 1

General Idea

Find out what's trending and likely to attract the customer

Step 2

General Idea

Limit the results based on the customer's personal preferences

Step 3

General Idea

Expand the result set with other complementary articles that might be unexpected but will also interest the customer

OUR PROCESS

Step 1

Top 100 most popular articles by customer age group

**We assume that the customer will likely be interested in the mass fashion trend for his/her demographic at that time. We also do not scope down the customer's gender as we do not have sufficient information and analysing transactions itself could be misleading*

Step 2

Choosing the most important and popular attributes based on the customer's transactions within past 3 months

*We assume that his past purchases are indicative of his preferences for his future purchases

Step 3

Use Image Processing to recommend similar articles for customer

STEP 1

a) Get CustomerDF within required age range (e.g. 60-70)

	customer_id	age
31	00018385675844f7a6babbed41b5655b5727fb16483b6e...	68.0
93	000412345109c7c085b5faec96afe864b19a172fa4cb9b...	63.0
106	00048f2f68760664d2d0fa1e7fbfe083f05287f342484c...	67.0
172	000747860042b94e85707605c2a627c6ba30c4117d025d...	62.0
309	000eae69313b4fc1824fa7e439f168cc140bf4c3f3a7e9...	67.0

b) Merge CustomerDF with TransactionDF on customer_id and find transactions within recent 90 days

	customer_id	age	t_dat	article_id	max_dat	diff_dat
30	000eae69313b4fc1824fa7e439f168cc140bf4c3f3a7e9...	67.0	2020-07-08	562245018.0	2020-09-22	76.0
31	000eae69313b4fc1824fa7e439f168cc140bf4c3f3a7e9...	67.0	2020-07-08	562245088.0	2020-09-22	76.0
32	000eae69313b4fc1824fa7e439f168cc140bf4c3f3a7e9...	67.0	2020-07-10	720125041.0	2020-09-22	74.0
33	000eae69313b4fc1824fa7e439f168cc140bf4c3f3a7e9...	67.0	2020-07-10	562245018.0	2020-09-22	74.0
37	000f1c4a03223d999d3c6d3703e247dba81d6dacb3dbfb...	62.0	2020-06-25	817472005.0	2020-09-22	89.0

c) Find top 100 articles from mergedDF based on purchase count for all customers within required age range

	Article ID	Count
14986	896152002	328
10920	856840001	321
14993	896169002	222
5784	783346001	205
14987	896152003	195

STEP 2

a) Get top attributes of articles bought by a specific customer within the past 3 months

b) Apply feature importance (Chi Squared coefficient) as weights to the absolute counts for each category

c) Updated feature ranking of each customer with new weights

	Category	Value	Count	Weights	Weighted Count
115	department_name	Jersey	30	0.647311	19.419316
65	colour_group_name	Black	156	0.062894	9.811446
206	garment_group_name	Jersey Fancy	60	0.144265	8.655917
39	product_group_name	Garment Upper body	107	0.055737	5.963870
49	graphical_appearance_name	Solid	181	0.026084	4.721114
99	perceived_colour_master_name	Black	156	0.022047	3.439334
179	section_name	Womens Everyday Collection	78	0.036577	2.853025
167	index_name	Ladieswear	176	0.002768	0.487197
93	perceived_colour_value_name	Dark	173	0.001336	0.231048
175	index_group_name	Ladieswear	219	0.000982	0.214998
0	product_type_name	Trousers	53	NaN	NaN

STEP 2

d) Combine Steps 1 and 2 to select Articles that are the top 100 in customer age group and possess the most characteristics of popular articles

e) Use Linear Regression to find find expected Price of Selected Articles

	article_id	expectedPrice
1	896152002	0.033778
2	863646001	0.033766
3	863595006	0.033229
0	915526002	0.032942

f) Find Mean Price of Transactions made by specific customer

Mean Price = 0.033971

g) Find article with mean price closest to expected price

	article_id	expectedPrice
1	896152002	0.033778

STEP 3

a) Suggest additional recommendations from previously trained feature map

```
In [*]: feature_vec = np.array(features)
        result=result_vector_cosine(main_model,feature_vec,preprocess_img(output[selected_articleID]))
        input_show(cv2.imread(output[selected_articleID]),output[selected_articleID])
        show_result(output,result)

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```

CHI SQUARED FEATURE SELECTION

Step 1

Encode categorical attributes of clothes

Apply OrdinalEncoder on all transactions within an age group range

Step 2

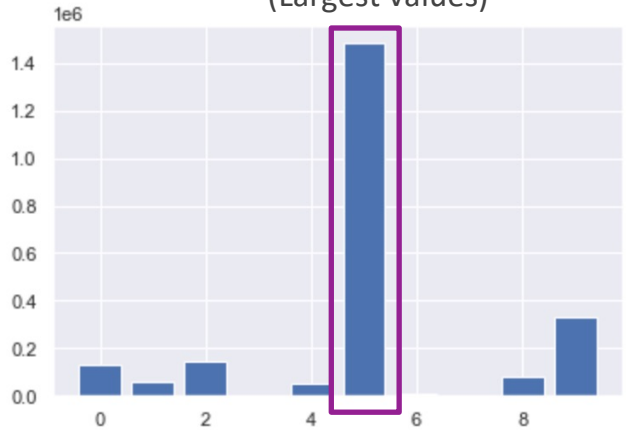
Perform Train-Test Split (33% test size)

Step 3

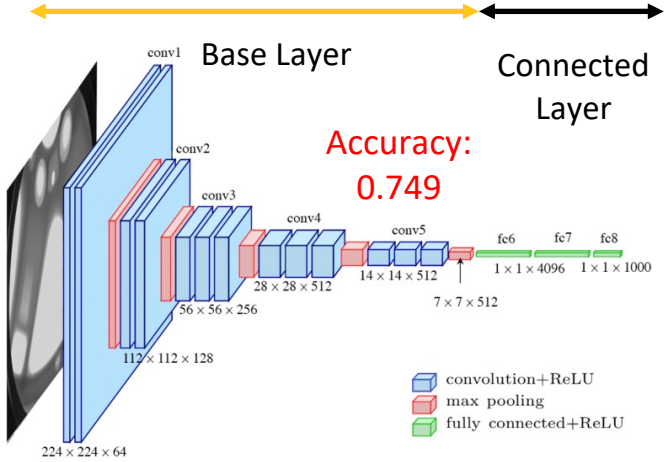
Obtain feature coefficients

Pearson's chi-squared statistical feature election strategy was used to select ***k*** **most relevant features** (Largest values)

```
Feature 0 product_group_name: 127669.496136
Feature 1 graphical_appearance_name: 59745.971223
Feature 2 colour_group_name: 144062.564468
Feature 3 perceived_colour_value_name: 3059.131154
Feature 4 perceived_colour_master_name: 50500.124313
Feature 5 department_name: 1482707.295581
Feature 6 index_name: 6340.658688
Feature 7 index_group_name: 2248.704987
Feature 8 section_name: 83782.587899
Feature 9 garment_group_name: 330449.100526
```



STEP 3 - Keras ResNet50




```
model = ResNet50(include_top=False, weights='imagenet')
```

Pretrained ResNet50 model (50 layers deep) on ImageNet image database

Borrowing the base convolution network, we extracted image features for all article images (specific to the H&M image database)

Nearest Neighbour — Distance Measures

- Given two feature vectors with numeric values

$$A = (a_1, a_2, ..., a_n) \text{ and } B = (b_1, b_2, ..., b_n)$$

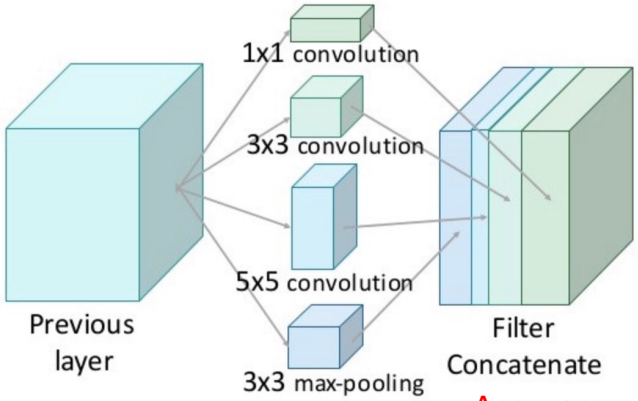
- Use the *distance measure*:

$$d = \sqrt{\sum_{i=1}^{n} \frac{(a_i - b_i)^2}{R_i^2}} = \sqrt{\frac{(a_1 - b_1)^2}{R_1^2} + \frac{(a_2 - b_2)^2}{R_2^2} + ... + \frac{(a_n - b_n)^2}{R_n^2}}$$

R_i is the *range* of the ith component

Measured the **Euclidean distance** between the input image and the features of other articles generated by our model

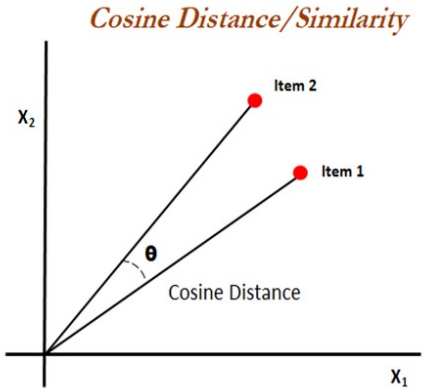
STEP 3 - Keras Xception



Accuracy: 0.790

```
model=Xception(weights='imagenet',include_top=False)
```

Pretrained Xception model (88 layers deep) on ImageNet image database.



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

Measured the **Cosine Similarity** between the input image and the features of other articles generated by our model. A more accurate means of finding similarity as **orientation is considered**, on top of distance.

Comparisons between models

Product Title: Kenzy Denim Dungaree
Euclidean Distance from input image: 10.931457

Product Title: Dungaree Dress
Euclidean Distance from input image: 12.176915

Product Title: Kenzy Denim Dungaree
Euclidean Distance from input image: 14.290119

Product Title: RYAN PINNAFORE
Euclidean Distance from input image: 14.563355

ResNet50 with Euclidean distance

========== input product image ======

Product Title: Dungaree Shorts

Xception with cosine similarity

05 SUMMARY OF FINDINGS

CONCLUDING KEY INSIGHTS

1. Understanding the state of the business	The company targets young adults and older mature adults
	Ladieswear and Teens product categories are their main revenue driver. Average sales transaction quantum is small
	Revenue and purchase transactions trends are consistent and steady
2. Creating Value with Hybrid Recommender	We developed a full data pipeline to make use of the available dataset effectively and efficiently
	Using hybrid filtering, we determine when a customer will make a purchase next and what they might purchase given their spending habits
	Leveraging a deep learning network ResNet, we use CNN to train the model on product images and KNN for evaluating an article recommendation

06 FUTURE WORKS

FUTURE WORKS

Estimating customer's average spending amount based on transactional history

The nature of this problem is can be rather unpredictable. There are many factors that may sway a customer's decision to spend more on an item that they like. What we have done so far is to make a fair estimate of the customer's typical budget which will drive their decision to purchase an item of interest. We could use an AI model to learn the customers purchase habits and make better predictions.

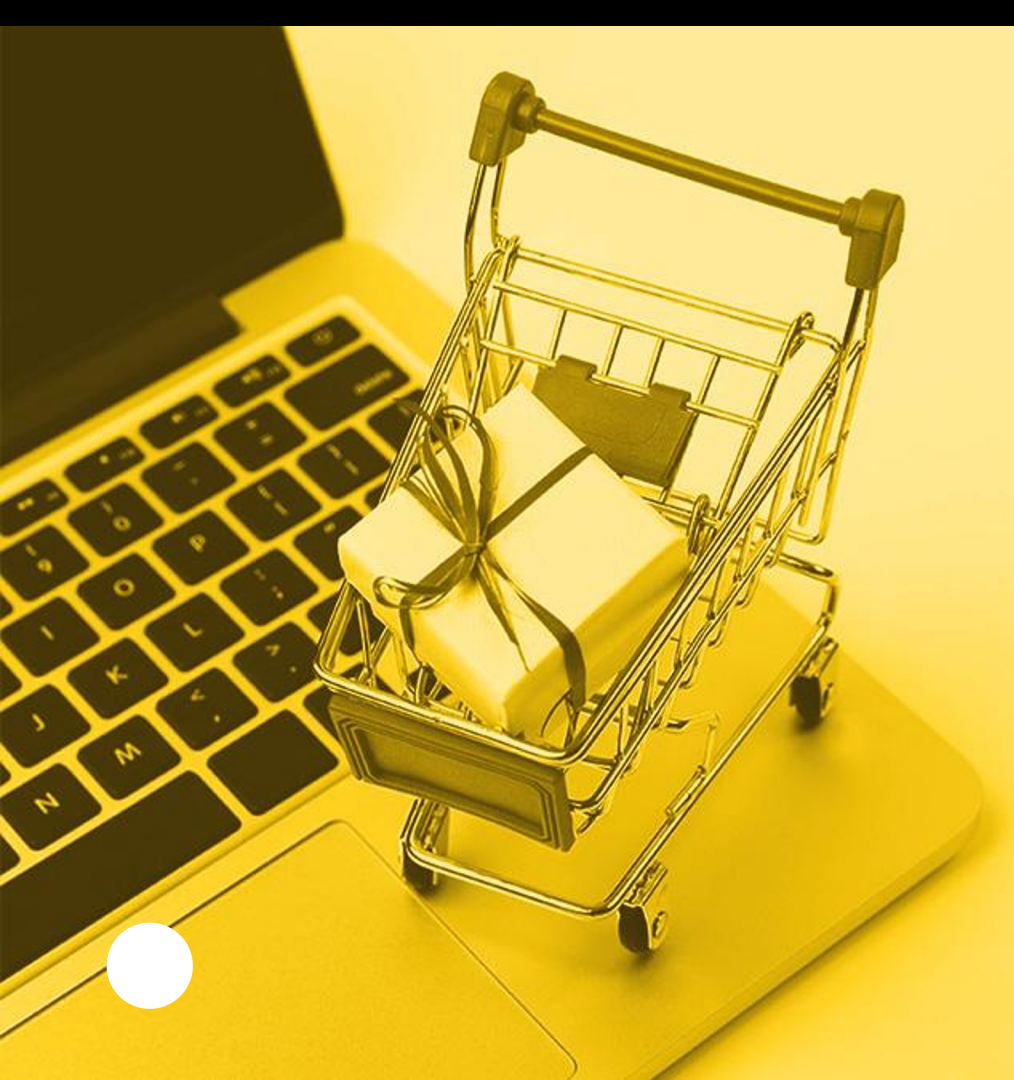
Quality of available data is important. We look to source for more data points that will allow us to rely less on some of the assumptions that we have made thus far.

FUTURE WORKS

Expanding final suggestion list with complementary items using AI

Improving image processing model through transfer learning. By integrating a custom connected layer, our image processing model could be significantly more accurate in extracting unique features of each clothing article.

Quality of training data is more important than data quantity.



THANKS!