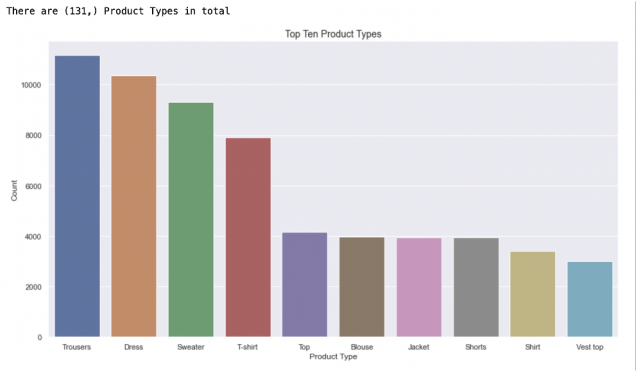
PERSONALIZED FASHION RECOMMENDATIONS

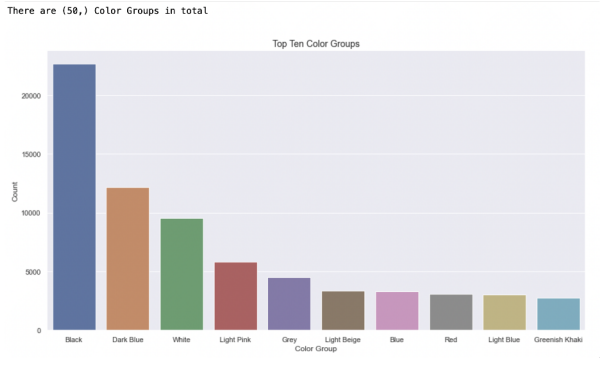
Data Analysis and Visualization

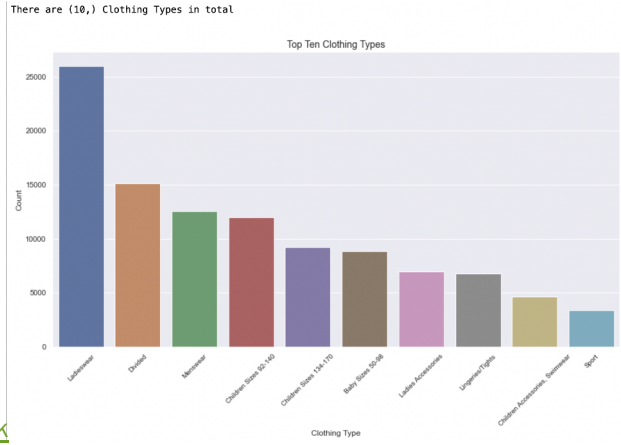
Hongtao Jiang – hj2554, Ling Sun - ls3759, Mohit Medini Singh - mms2308, Zeyu Jin - zj2325, Arvind Kanesan Rathna - ak4728

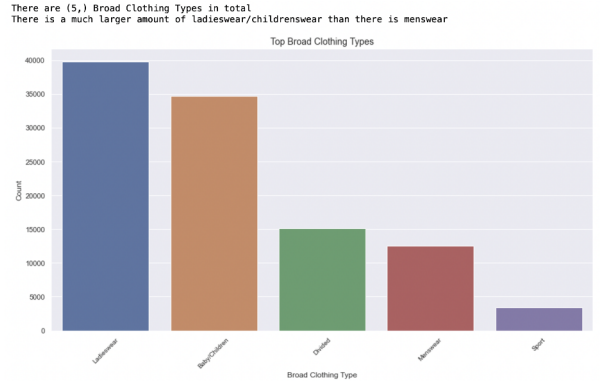
Initial Data Exploration: Articles of Clothing

There are 10 Clothing Types that encompass 131 underlying product types. Ladieswear makes up the majority of the clothing sold by H&M, followed by clothing for children. Menswear pales in comparison. Trousers, Dresses and Sweaters are the most frequent articles of clothing. Most of the clothing sold is black.



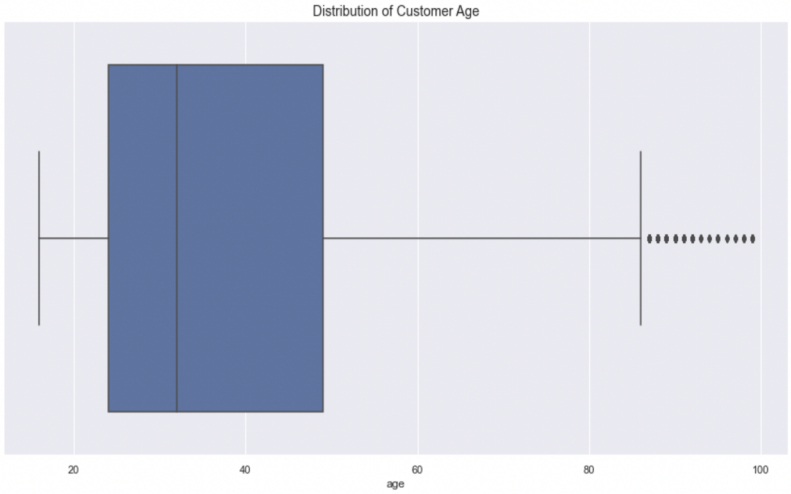




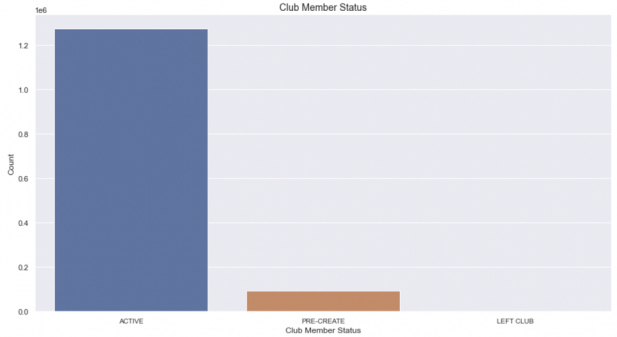


References: Dataset source, EDA notebook

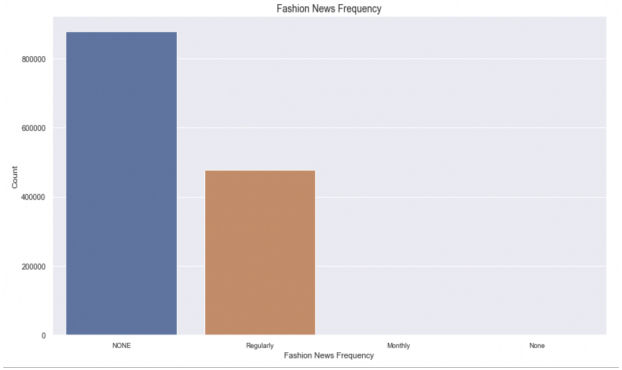
Initial Data Exploration: Customer Information




```
ACTIVE        1272491
PRE-CREATE      92960
LEFT CLUB         467
Name: club_member_status, dtype: int64
There are (3,) Club Member Status types in total
Most customers in this dataset are active members in the rewards program
```




```
NONE         877711
Regularly    477416
Monthly         842
None              2
Name: fashion_news_frequency, dtype: int64
There are (4,) Fashion News Frequency Categories in total
Most customers in this dataset do not subscribe to fashion news
```

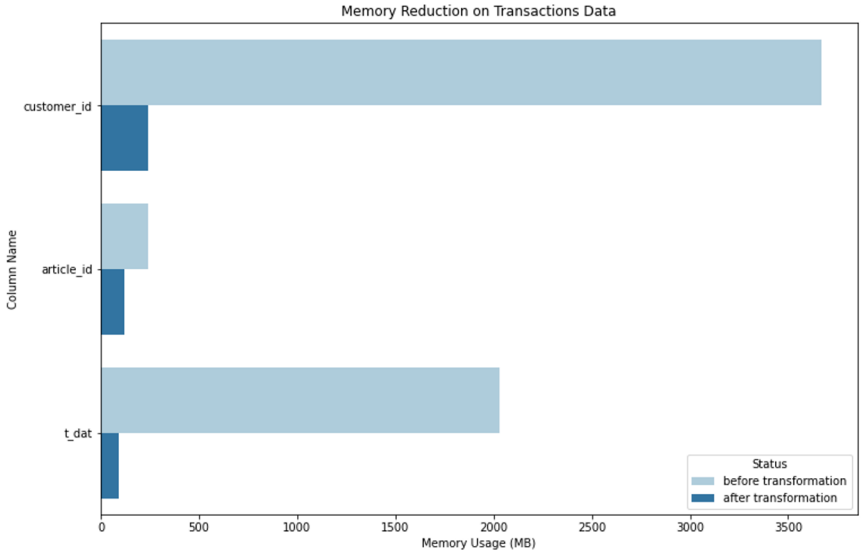


The average H&M customer age is around 37. Outlier customers are beyond 82 years of age which makes logical sense. Over 93.1% of the customers are active Club Members (1272491) in the rewards program. Mere 467 members (0.03%) have left the club. This means the club membership program is successful in general. However, a majority of the customers (64.72%) do not subscribe to the fashion news.

Data Cleaning: Memory Reduction

The transaction dataset is so large that it costs about 6.28 GB of memory. We transformed the “customer_id”, “article_id”, “t_dat” columns so that we reduced the total memory usage to only 0.92 GB.

- customer_id: the id of the customer that makes the transaction
  - map the hexadecimal string to int64
- article_id: the id of the article that is purchased
  - cast int64 to int32
- t_dat: the timestamp of the transaction
  - split into year, month and day, each being int8 type



References: Data Cleaning notebook

Data Cleaning: Memory Reduction

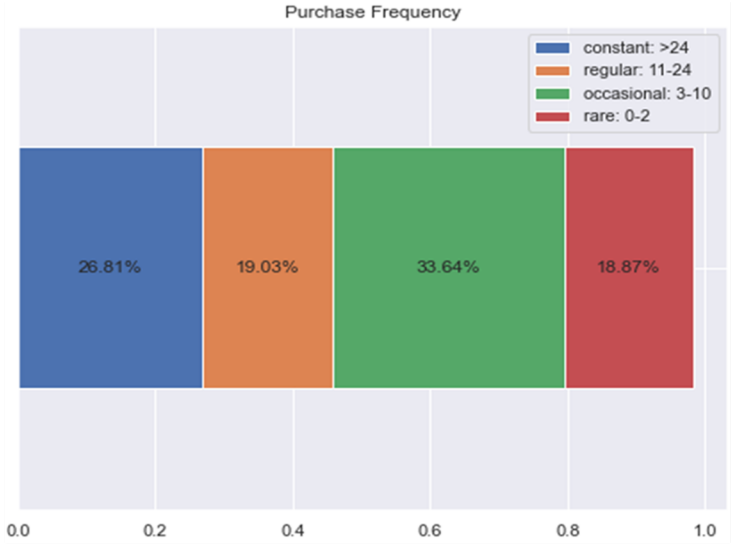
- Transaction Dataset:
  - *No missing value*
- Article Dataset:
  - *Only the description column has missing value and it is impossible to make imputation*
- Customer Dataset:
  - *“FN” and “Active” columns: Both columns have values of either NA or 1, so we replace NA with 0.*
  - *“age” column: Only 1% of the data is missing, so we imputed with the mean of the data.*

Data Cleaning: Encoding on Categorical Feature

The “club_member_status” and “fashion_news_frequency” columns of the customer dataset are both categorical, so we apply one-hot-encoding and ordinal encoding on the two columns respectively.

- club_member_status:
  - *unique values: Active, Pre-create, Left-club, NA*
  - *one-hot-encoding*
- fashion_news_frequency:
  - *unique values: NA, None, Monthly, Regularly*
  - *ordinal encoding: map [NA, None, Monthly, Regularly] to [-1, 0, 1, 2]*

Customer Data Insights: Recency & Frequency



Most recent customers: 19.96% of customers made purchases within 6 month

Difficulty: 14.5% of customers never made purchases in the last 2 years

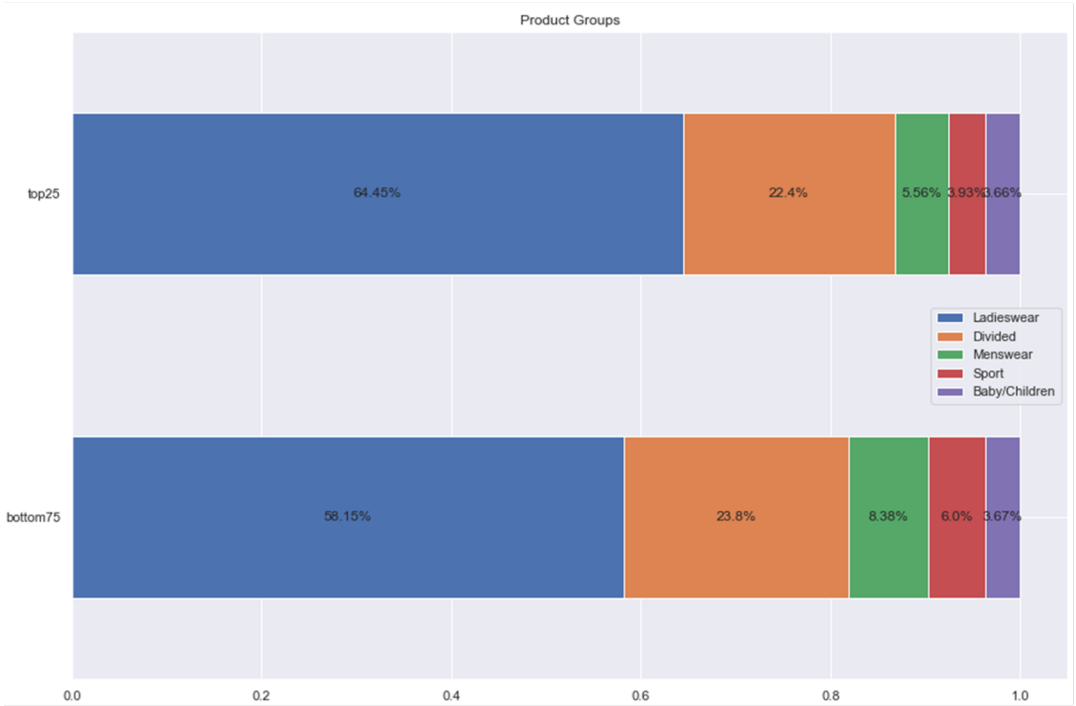
Regular customers: 1-2 purchases each season, 11-24 purchases in total over 3 years

45.84% of regular+ customers in total

Difficulty: 18.87% of customers <1 purchase each year

References: Data Exploration Insights notebook

Customer Data Insights: Monetary & Purchase Habits



2018 Jan - 2020 Dec:

1.4 million customers, $885 million total sales

The top 25% of customers who spend the most account for 75% of total sales

Purchase habits:

Ladieswear is the most popular in both groups, even more within Top25 (64.45%)

Shares for Divided, Menswear, and Sports go up a bit within Bottom75 (38.18% in total)

customer_id	item_perchased	total_amount($1k)
5854009424779598107	1895.0	57.676407
3407358910964148684	1361.0	50.921186
1135991499650384534	1570.0	49.967169
-6466893749761682	1522.0	47.682017
7398229172292340849	1441.0	47.662000
...	...	...
-8872730345030261744	0.0	0.000000
-4280122082666584064	0.0	0.000000
4415134887239278817	0.0	0.000000
5324495867649106638	0.0	0.000000
3086006858415218905	0.0	0.000000

Top25

Bottom75

Customer Data Insights: Top Associations

Top25

>30% times people buy multiple:
Dress, Trousers, Sweater, Socks

Times when types are bought together:
Blazer Trousers 16.84%
Costumes Trousers 15.98%
Jumpsuit Dress 16.64%

Sarong Swimwear 15.24%

Heels Dress 17.74%
Wedge Dress 16.98%
Heeled sandals Dress 16.30%
Sandals Dress 15.37%
Ballerinas Dress 15.11%

Bottom75

>30% times people buy multiple:
Dress, Trousers, Sweater, T-shirt, Shorts, Underwear, Pajama

Times when types are bought together:
Blazer Trousers 19.56%
Hoodie Sweater 15.73%

Sarong Swimwear 20.94%

Heeled sandals Dress 17.10%

Differences:

Top25 - buy different types of products at a time
Bottom75 - buy multiple items of the same type

Top25 - fashion choices
Bottom75 - cozy choices

Similarities:

Popular items - Dress, Trousers, Sweater

A special match with Sarong - Swimwear

A popular match with Dress - Heeled sandals

ML Techniques Proposal: Baseline Model

Main Idea: recommending items that are frequently purchased together is effective.

- Uses other orders data to group products that were bought together.
- Aims at increasing the user's cart size and average order value.

Potential Issue: some products are generally popular and they would appear as recommendations more often which can lead to low conversions.

ML Techniques Proposal: Adding variable “Quotient”

Main Idea: If a customer buys the same product for a short period of time, and if the growth rate of that product is high, it is expected that they will buy more.

- “Quotient” is a variable that is the quotient of the number of units sold in the week of 9/22 and the number of units sold in each week for a given article_id.

Explanation:

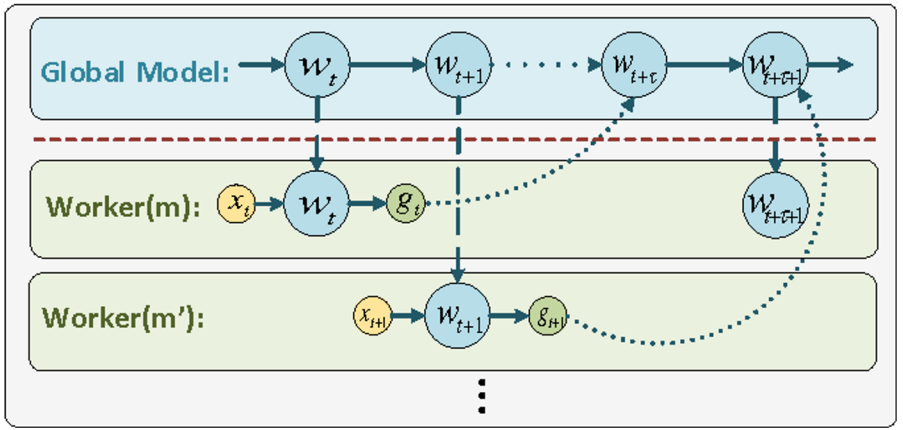
- If sales units in the week of 2020-09-22 are higher than sales in the week before 2020-09-22, the value will be larger, which is an indicator of the growth rate.
- Quotient is calculated for all transactions. Thus, the "sum of quotients" is an indicator that has a larger value when the number of units sold is higher, and an indicator that has a larger value when there is a high growth rate.

ML Techniques Proposal: Lightfm

Algorithm Explanation: A Python implementation of asynchronous stochastic gradient descent algorithms for both implicit and explicit feedback.

Steps:

1. The parameters of the model are distributed on multiple parameter servers.
2. Multiple workers process a mini-batch of data in parallel and communicating with the parameter servers independently of the other ones to compute gradient and update the model.



Choosing Loss Function - Bayesian Personalised Ranking 1 pairwise loss

- BPR maximises the prediction difference between a positive example and a randomly chosen negative example.
- Since the prediction goal of the project is to optimize the ROC-AUC (prefers the positive item over all other non-observed items.), this particular loss function will be chosen.