H&M Dataset Presentation

Team:
Sai Krupa Jangala (sj3140), David Heagy (Dh2868), Karunakar Gadireddy (kg2911), Jugal Shah (js5950), Jiayuan Cui (jc5670)

Overview & Data Description

Overview:

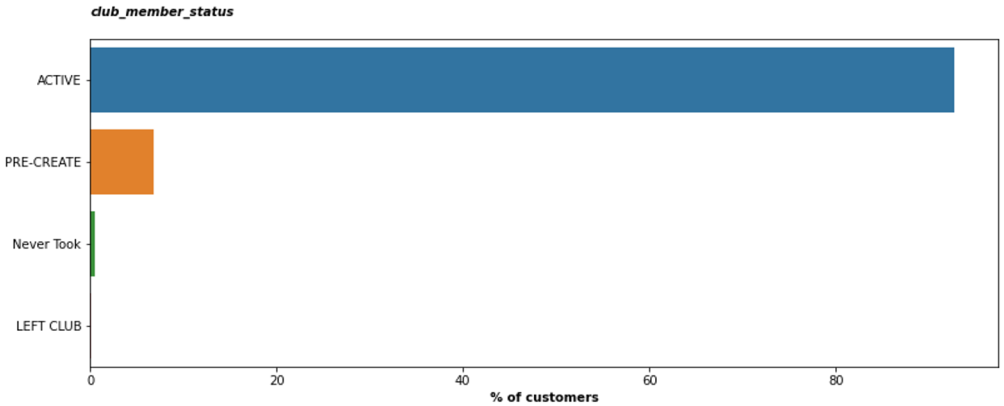
- Many online shopping sites face the issue of customer's scrolling through their page but not making a purchase. We decided to build a product recommendation system based on H&M's data from previous transactions, as well as the customer's and product meta data.
- The recommendation system will help the customer decide which product to buy easily from the plethora of options.
- An efficient recommendation system reduces the risk of return and thereby reducing the cost the firm pays for the logistics on the return policy.

Data Description:

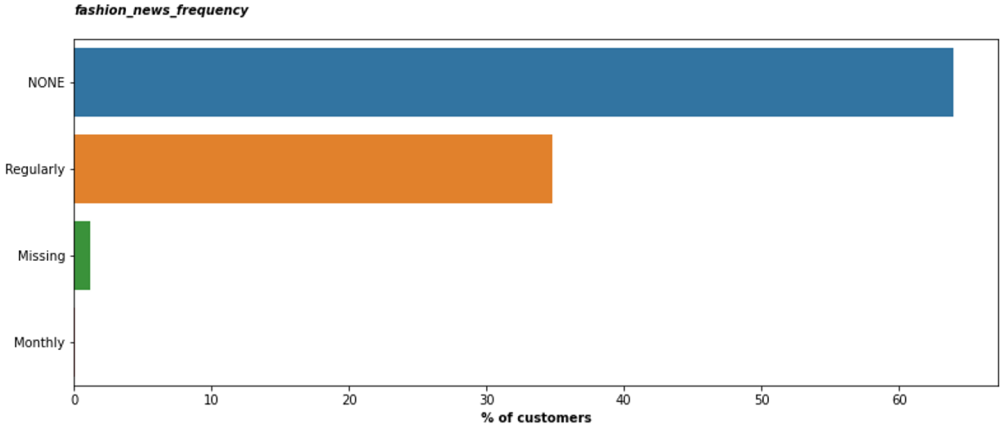
- We have three csv files with different information along with the image of each article.
  - File articles.csv contains article_id, which shows available articles that can be purchased.
  - Folder images/ contains images corresponding to each article_id.
  - File customers.csv contains customer_id, which has the detailed information of customers.
  - File transactions_train.csv contains the training data, consisting of the purchases of each customer for each date.

Link to the dataset.

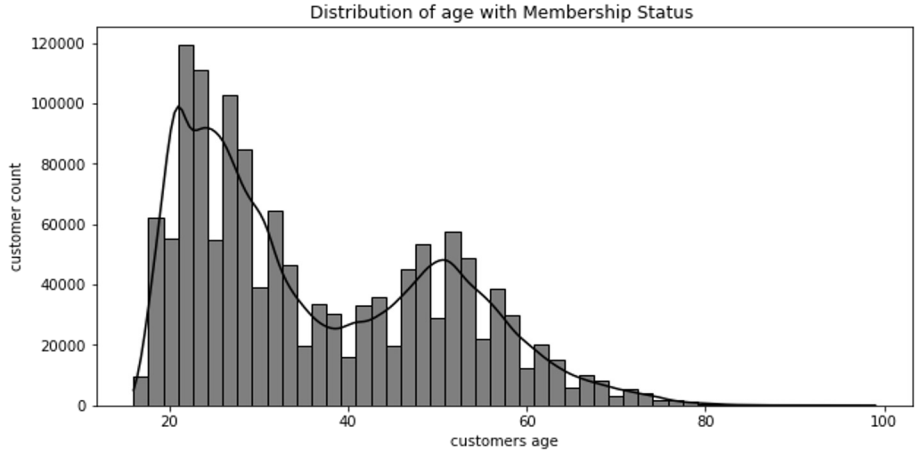
Customer Data

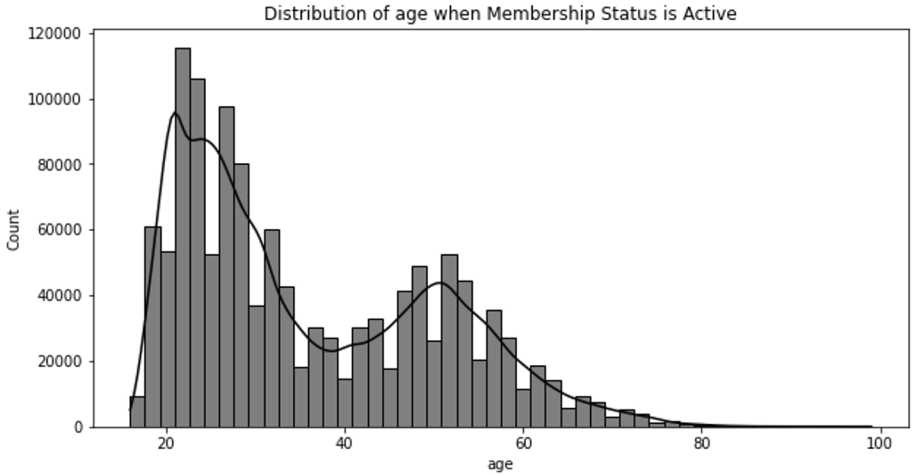


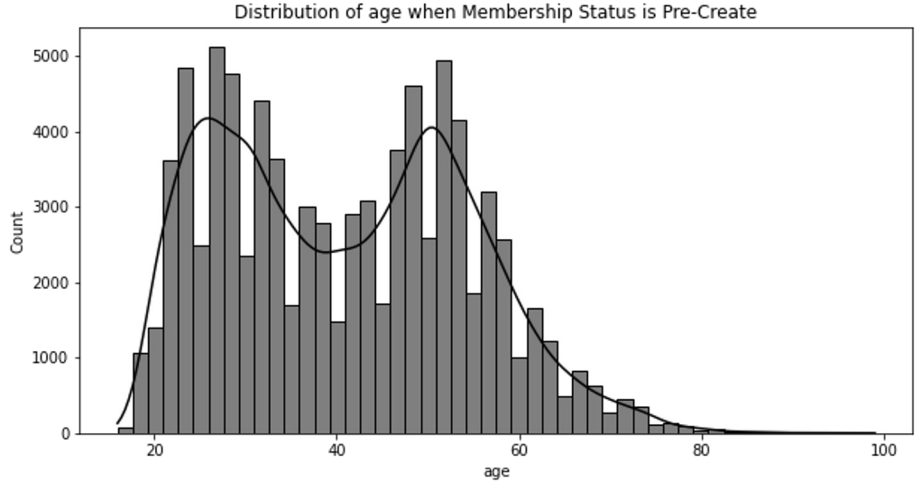
- H&M club status of the customer. Can be active, pre-active or left-club or none. If it is none we can assume that this customer has never taken a membership
- We can see that most of the customers are Active club members with a small minority still creating their membership or left the club. We can also see that the percentage of people with no club status is very small.



- Frequency of sending communication to the customer. Can be Regularly, Monthly, None, N/A,NONE.
- Since None and NONE represent the same data we can consider them the same value.
- For the missing data there is no way to infer this data and since it is a small percentage (less than 5) we can drop these rows before training if required.
- We can notice that majority of the customers receive no communication from H&M







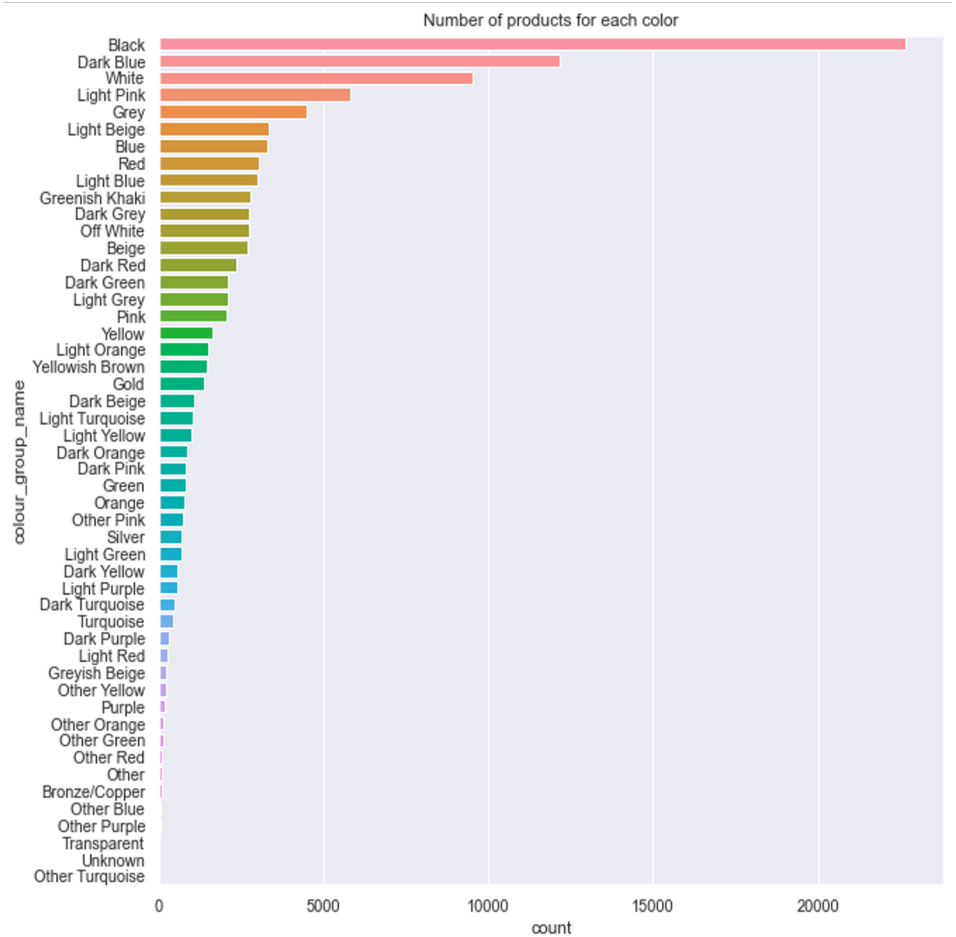
- The distribution of age with customer count shows the maximum age of the customer buying the products lies around 21-23.
- There is a drop in membership between the 30-40 age group.

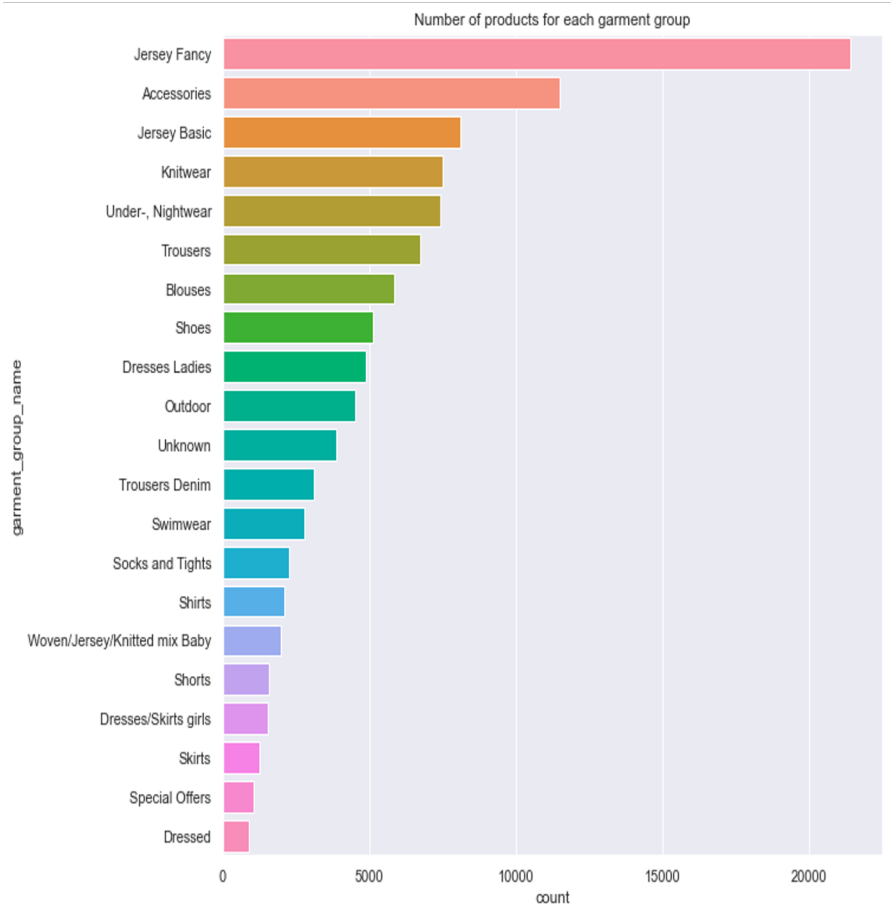
Articles Data

- The articles data contained 25 columns which are as follows:

```
Index(['article_id', 'product_code', 'prod_name', 'product_type_no',
       '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'],
      dtype='object')
```

- As a part of data cleaning, we dropped columns which are IDs namely, product code, product_type_no, graphical_appearence_no, perceived_colour_value_id, perceived_color_master_is, department_no, index_code, index_group_no, section_no, garment_group_no
- Others are categorical variables and target variable(prod_name). There are no missing values found in the data.
- We plan to do target encoding for these categorical features. Target encoding is not based on the product name(target variable), but we considered the usage of a particular color or product. For example, if a particular color: say black is predominant over all the products, it would be given a higher weightage.



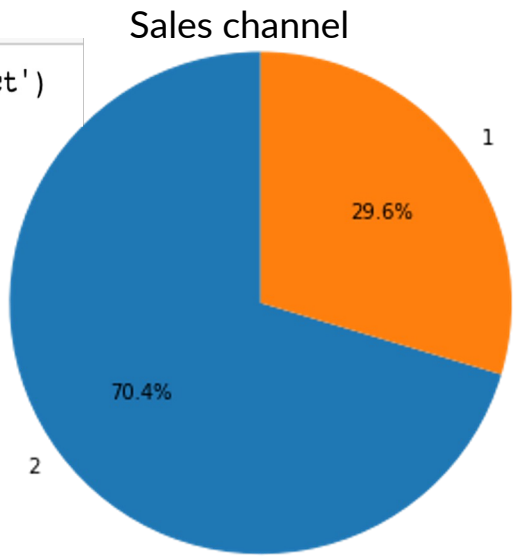


As we can observe, Black is the most used color and Jersey Fancy is the most used garment group. These two groups would have higher weightage compared to other groups in our encoding process.

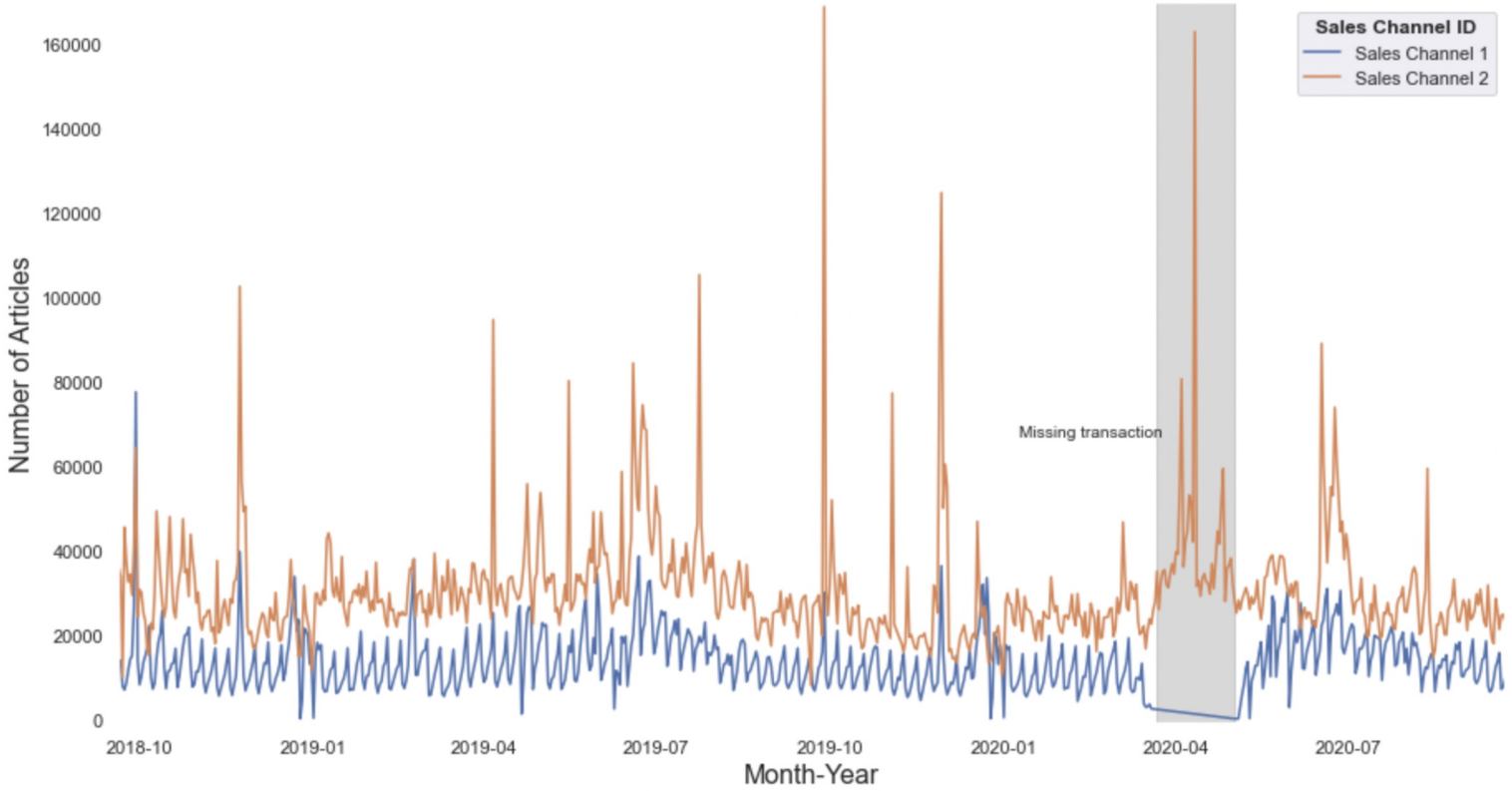
Transactions

- The transactions data consists of the following columns:

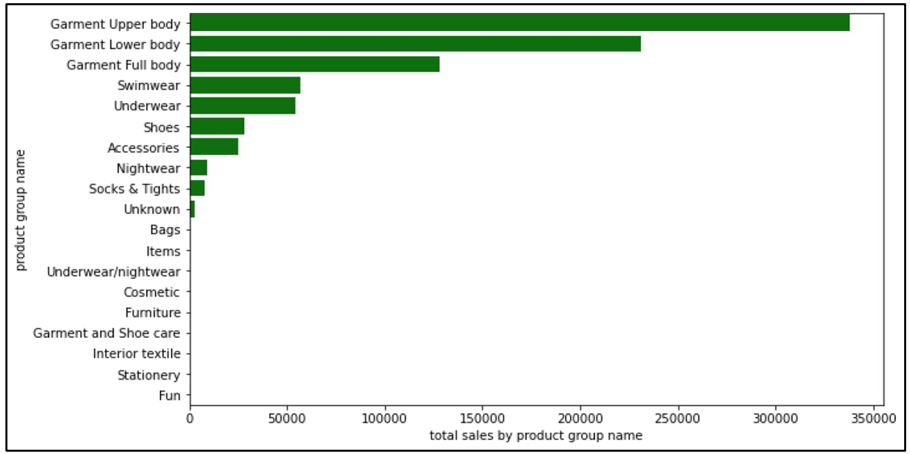
```
Index(['t_dat', 'customer_id', 'article_id', 'price', 'sales_channel_id'], dtype='object')
```

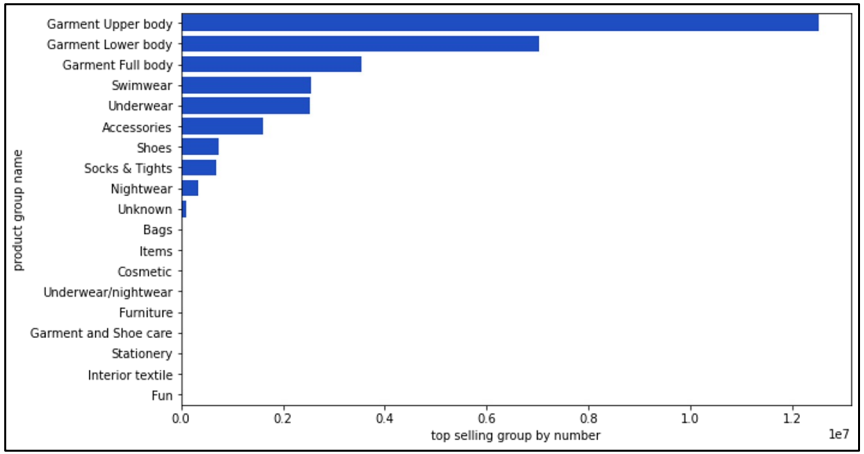


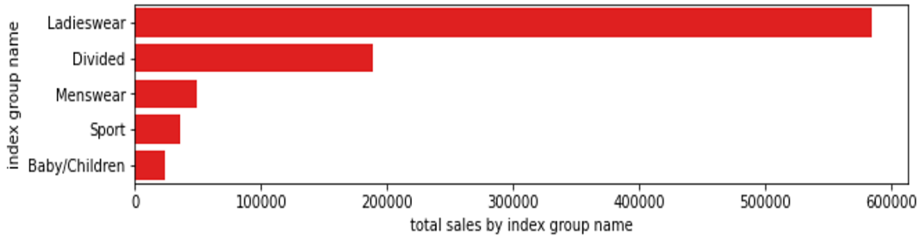
- This data is gives us information about different transactions, the time of purchase, customer who purchased it, etc. There is no missing information here as well
- There are two major sales channels which are used. As seen in the pi chart we can observe that a lot of transactions took place in sales channel 2 which is almost 70% of the total transactions.
- Though sales channel 2 is widely used, we had no confirmation that a lot of products will be bought if a particular channel was used. Hence, we believe that the product purchase is independent of the the channel. So we used one hot encoding for this particular feature.
- Other features like customer_id, article_id was used to merge with the previous datasets described. Price was used as it is without any encoding.

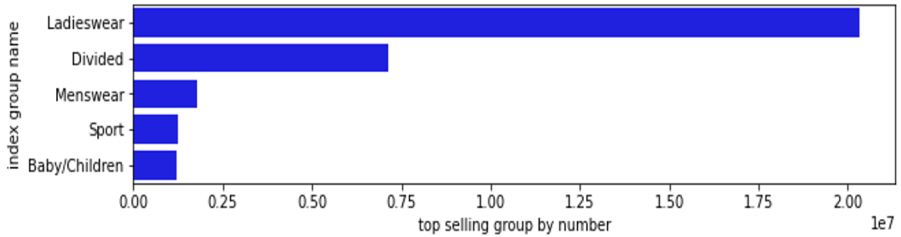


- The grey period indicates the missing transactions.
- Sales Channel 1 has missing transactions for few months.
- We plan to impute 0 for missing transactions in sales channel 1.





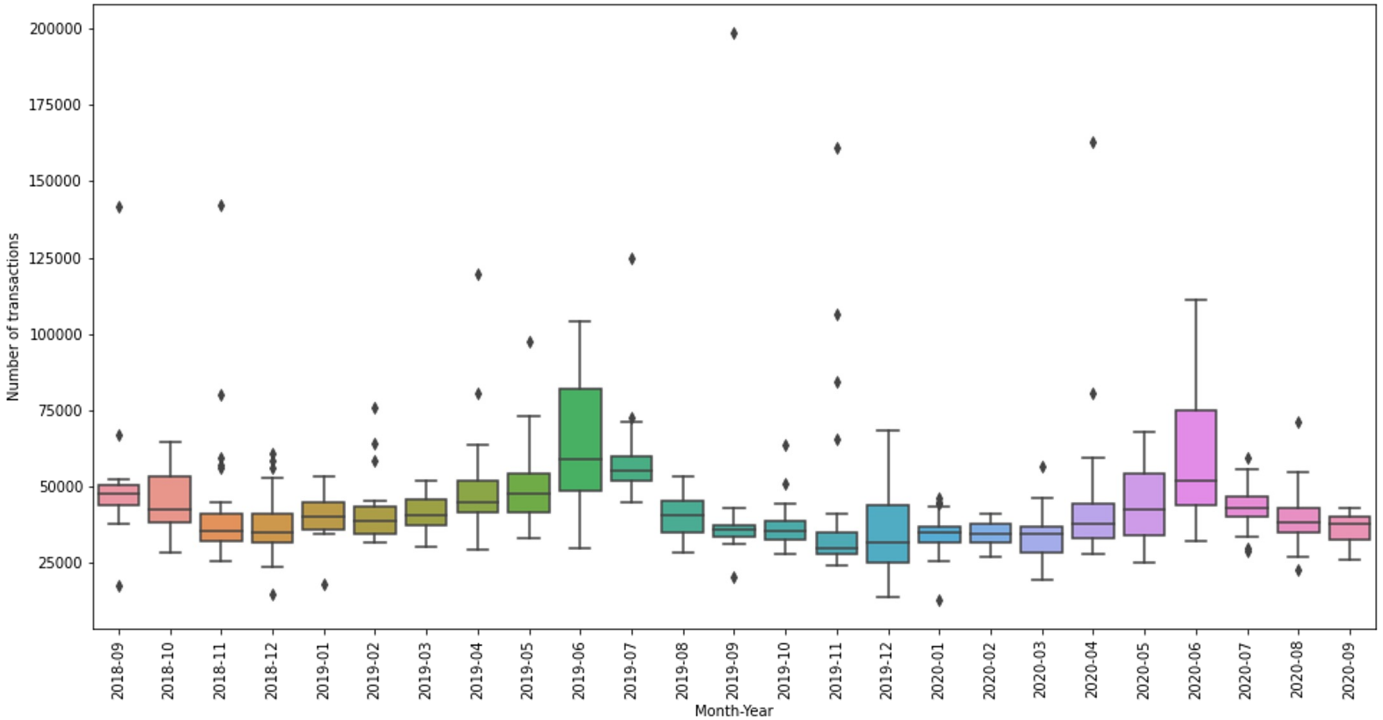




When analyzing the merge data, we found the corresponding groups and product groups that were the highest selling and were generating the highest sales and are most popular.

- In Product Group Garment Upper body was the highest sold and revenue generating merchandise
- Ladieswear generated the highest sales among the different groups present.

Distribution of sales for each month-year



- We have transactions for almost two years(Nov 2018 - Nov 2020).
- As we can see, the width of the box plots and the range is different for different months. We have a lot of transactions for the months of June in 2019 and 2020.
- We plan to create a new set of features for the twelve months and add weights accordingly based on the number of transactions in that particular month.

Proposed ML Techniques

- Collaborative Filtering: Build a model from past transactions as well as similar decisions made by other users
  - Nearest Neighborhood (KNN): Use the similarity distance (cosine) of user/item interactions in a sparse matrix of popularity ratings to find the k most similar items; (may need to reduce dimensions of dataset because of computational complexity)
  - Alternating Least Squares (ALS): Matrix factorization technique that decomposes sparse matrix of user-item ratings into separate user and item latent factors in lower dimensional spaces. ALS runs gradients descent on two loss functions: one holding user matrix fixed and the other running item matrix fixed.
- We plan to use Collaborative Filtering for our baseline models. We plan to advance with Neural Networks and Deep Learning if time permits.