H&M Personalized Fashion Recommendation

MSc in Big Data Management and Analytics
Applied Data Science Project 1- Part A

Group C

Kelvin Muindi Matingi

Shruti Sindhi

Yusuf Ibrahim

Background information:

- On exploring the online project details, description, and dataset, we uncovered that our topic doesn’t fall just under Fashion Domain, but it also falls under Retail Domain.
- The recommendation system will support fashion retailers in managing the store operations, supply chain management, and visual merchandising of the brand they are working for.
- To work on the available clothing-related dataset, one doesn’t need to be a fashion pioneer, wardrobe stylist, or fashionista.
- Some background knowledge of fashion and retail is sufficient.
- This project doesn’t take haute couture into consideration.

Problem Statement:

Shopping is an art. But all are not an artist. And with too many choices available now, customers may get dazed. They might not quickly find what interests them or what they are looking for, and eventually, they might not make a purchase. Thus, to sweeten the shopping experience, product recommendations are essential. In addition, assisting customers in making the right choices boosts business and has positive implications for sustainability. Sustainability reduces returns and minimizes emissions from transportation. H&M Group is endeavouring to develop product recommendations based on previous transactions and customer & product metadata. The main aim is to move away from the mass-market assortment and decide the best alternative from wide assortment, deep assortment, scrambled assortment, and localized assortment. The recommendation system will aid the brand in determining which assortment strategy is best at various locations.

What is the dataset?

The database is about the assortment of H&M shops. It consist of 3 dataset: articles, customer, and transaction.

- Articles Dataset:
  - Number of records:105542
  - Number of attributes:25
  - Attributes with numeric data type:11
  - Attributes with ordinal data type: 14
  - On exploration of the article dataset it was found that only one attribute had missing values. The total number of missing value detail_desc had is 416. It turns out to be 0.394%
- Transaction dataset:
  - Number of records: 31788323
  - Number of attributes: 5
  - Attributes with numeric data type: 3
  - Attributes with ordinal data type: 2
  - The transaction data had no missing values.

What is the dataset?

The database is about the assortment of H&M shops. It consist of 3 dataset: articles, customer, and transaction.

- Customer Datasets:
  - Number of records:1371980
  - Number of attributes:7
  - Attributes with numeric data type:3
  - Attributes with ordinal data type: 4
  - 5 features in the dataset have missing values. The number of missing values per column are:
    - FN: 895050
    - Active: 907576
    - Club Member Status: 6062
    - Age: 15861
    - Fashion News Frequency: 16009

Attributes detailed description:

- Articles Dataset:
  - Unique identifier of an article:
    - article_id :a unique identifier of the articles, it has 105542 unique values (total number of instances in the dataset).
  - 5 columns related to the product:
    - product_code: it is a 6 digit product code with 47224 unique values.
    - product_name: it includes the name of the product with 132 distinctive values.
    - product_group_name: it comprises of name of the product group, in total there are 19 groups.
    - product_type_name: the attributes mention about the name of the product type, it is equivalent of product_type_no.
    - product_type_no: it mentions about product type number and includes 131 unique values.
  - 2 columns related to the colour*:*
    - department_no: there are 299 unique departments present in the dataset.
    - department_name: department name, 299 unique values.
  - 1 column related to the description of the articles:
    - detail_desc: 43404 unique values.

Attributes detailed description:

- Articles Dataset:
  - 2 columns related to the garment group:
    - garment_group_n : garment group number, 25 unique values.
    - garment_group_name: garment group name, 25 unique values (accessories, shoes, outdoor and more).
  - 2 columns related to the section :
    - section_no : section number, 56 unique values.
    - section_name: section name, 56 unique values.
  - 4 columns related to the perceived color (general tone):
    - perceived_color_value_id: perceived color id, 8 unique values.
    - perceived_color_value_name: it includes 8 unique values of the perceived color.
    - perceived_color_master_id: perceived master color id, 20 unique values.
    - perceived_color_master_name: perceived master color name, it includes 20 unique values.
  - 4 columns related to the index, which is actually a top-level category.
    - index_code: index code, 10 unique values.
    - index_name: index name, 10 unique values.
    - index_group_name: index group name, 5 unique values.

Attributes detailed description:

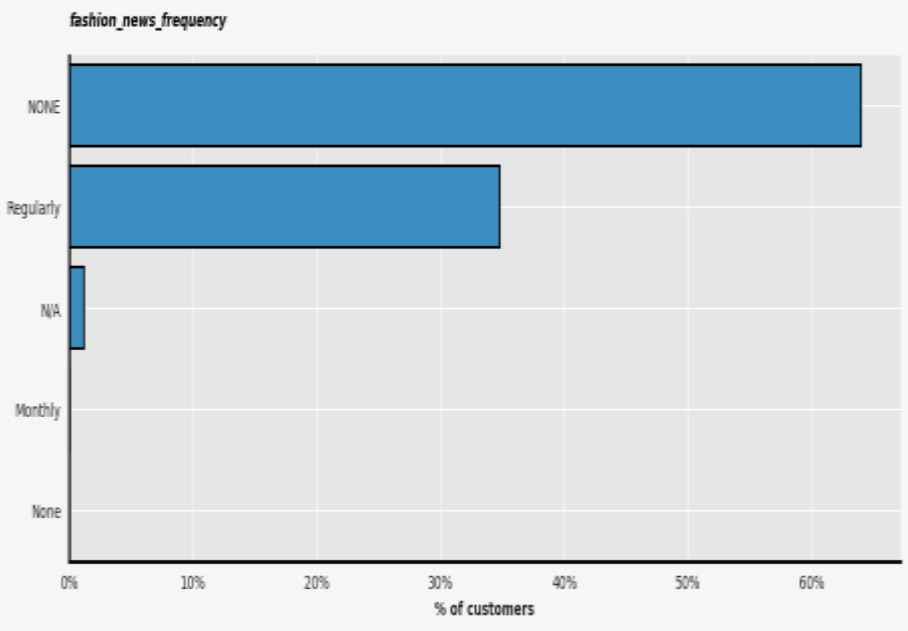

fig. 1: Fashion news frequency

- Customer Database:
  - Unique identifier of a customer:
    - customer_id : a unique identifier of the customer.
  - 6 columns related to the product:
    - FN: it is a binary feature (1 or NaN).
    - Active: it is a binary feature (1 or NaN).
    - club_member_status: it includes status in a club, 3 unique values (active, pe-create, and left club).
    - fashion_news_frequency: it consists of the frequency of sending a communication to the customer, 4 unique values. (Monthly, Regularly, NONE, and None)
    - age: age of the customers.
    - postal_code: postal code (anonymized), 352 899 unique values.

Attributes detailed description:

- Transaction Dataset:
  - 5 columns in the transaction dataset are:
    - t_dat: this feature includes the date of the transaction in the format YYYY-MM-DD, but it is provided as a string.
    - customer_id: it is the customer's identifier that can be mapped to the customer_id column in the customer table.
    - article_id: it is the product's identifier that can be mapped to the article_id column in the articles table.
    - price: the price paid by the individual.
    - sales_channel_id: sales channel includes 2 unique values. The values are 1 and 2. No detailed explanation or description could be concluded for 1 and 2 even after EDA.

How data was prepared?

- The dataset was acquired from H&M, a Swedish multinational clothing company.
- The H&M Group has 53 online markets and approximately 4,850 stores. It was unknown which store was the primary source of data.
- But on exploratory analysis, it was discovered that data was collected from 20-09-2018 to 22-09-2022.
- From the time series graph, we can see that there are distinct variations and spikes in the number of transactions per day.

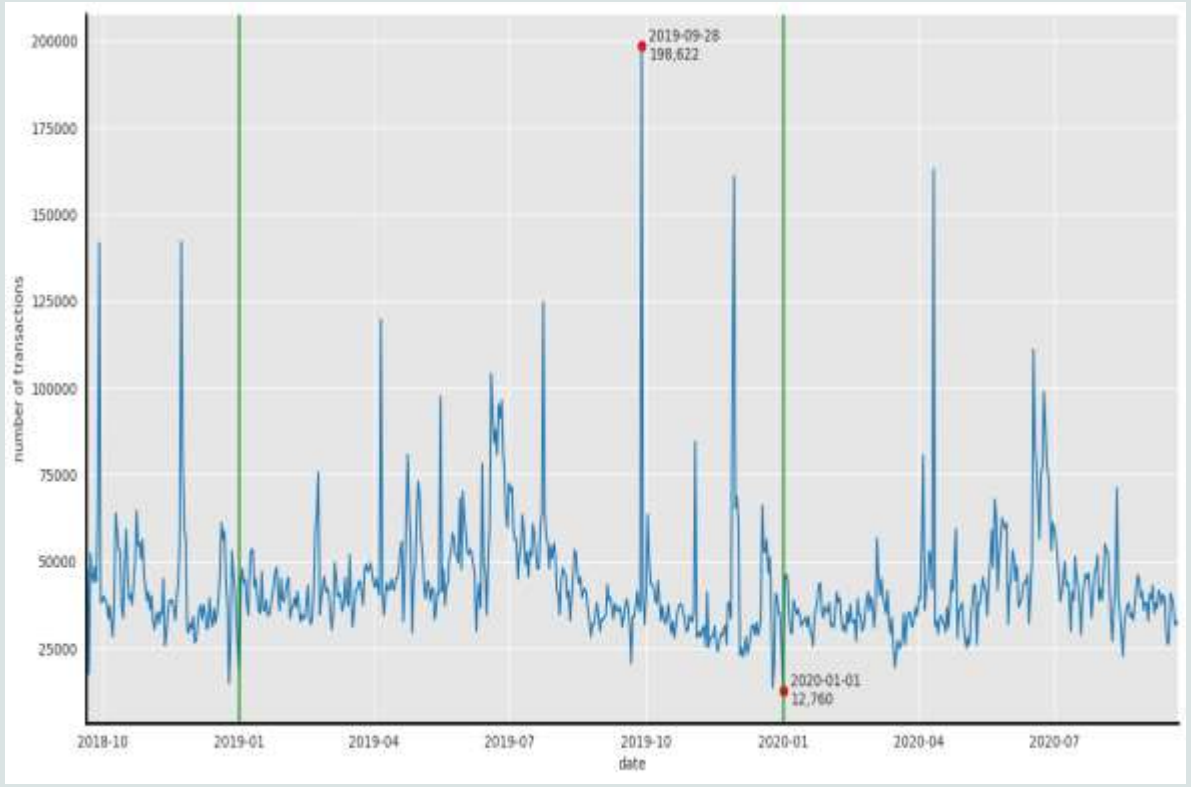

fig 2: number of transactions each day for period of 2 years.

What model was used to conduct the analytics:

- Collaborative filtering:
  - ➢ Predicting the future purchases using the metadata and purchase history. (in our scenario it is obtained using transaction related features of the customers).
  - ➢ It uses user-based nearest neighbour to find similar users.
  - ➢ The problem with user based neighbour formulation of collaborative filtering is the lack of scalability.
  - ➢ Recommendation of the product is done using the age group and gender attribute.
- Content-based methods:
  - ➢ It builds a baseline that uses product information.
  - ➢ Prediction is done using features related to products and images dataset.
  - ➢ The problem with this is that there is lack of novelty and diversity.
  - ➢ There is more to recommendation than just relevance.

What model was used to conduct the analytics:

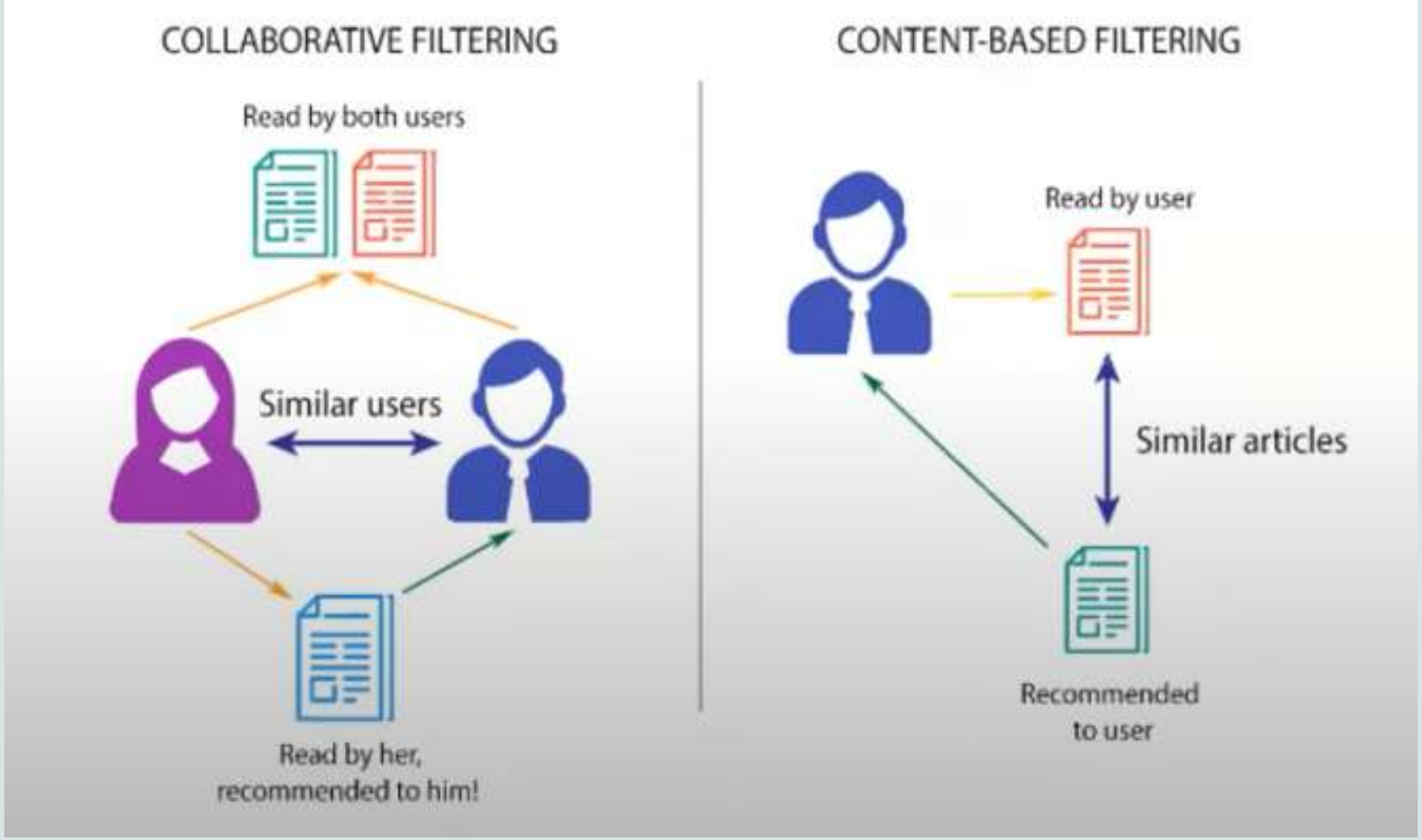

Fig. 3: Difference between user-based and item-based filtering algorithm

What model was used to conduct the analytics:

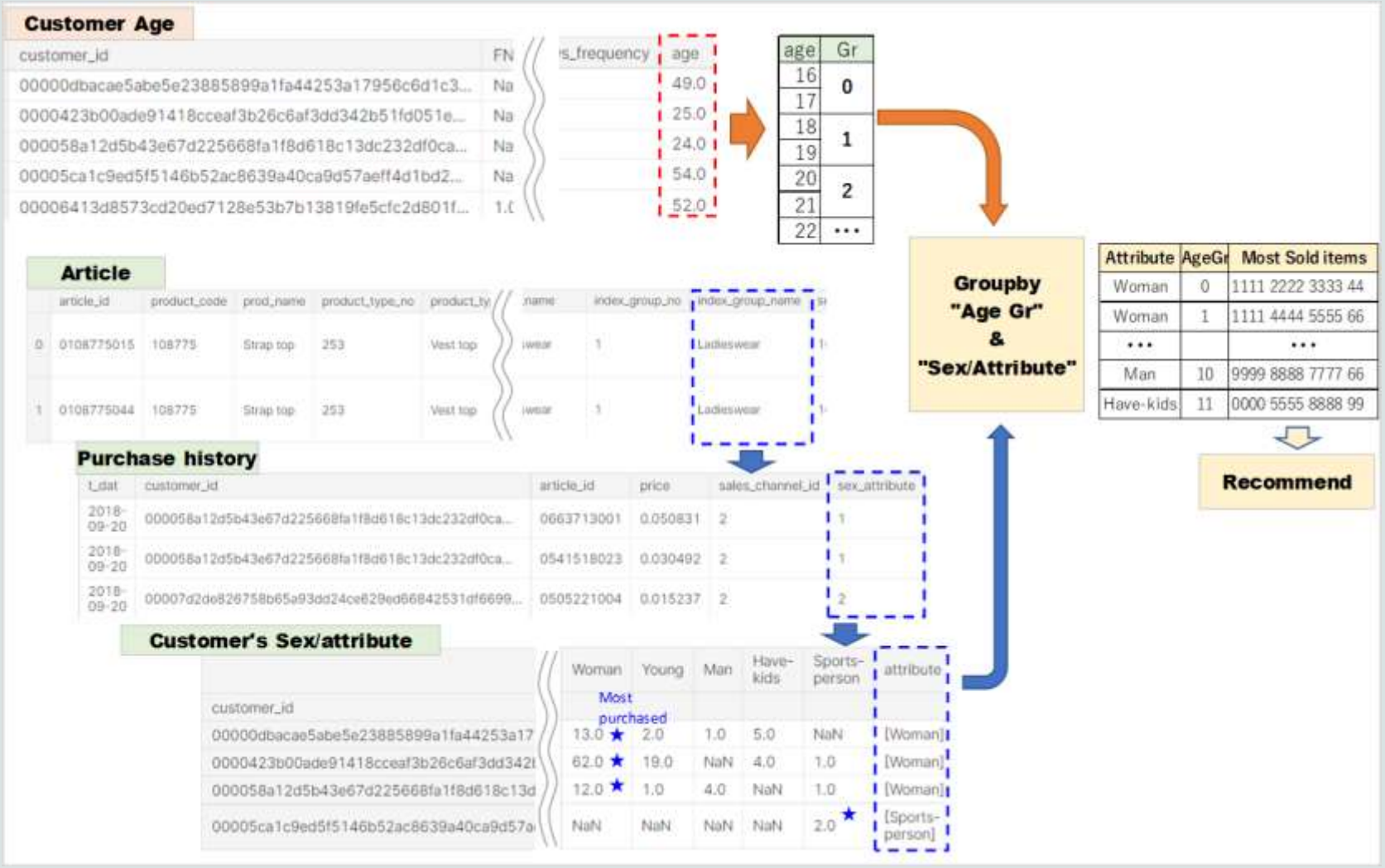

Fig. 4: Collaborative filtering approach overview.

Findings:

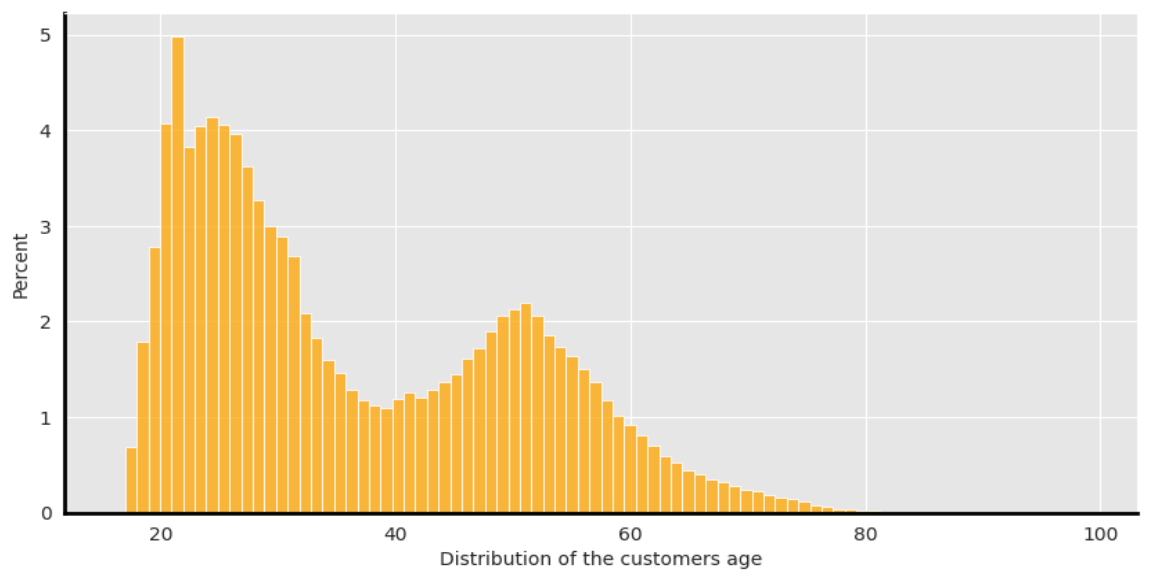

fig. 5: Distribution of the customer by age.

- The two main age-groups of customers are around 20-30 years old and 44-45 years old.

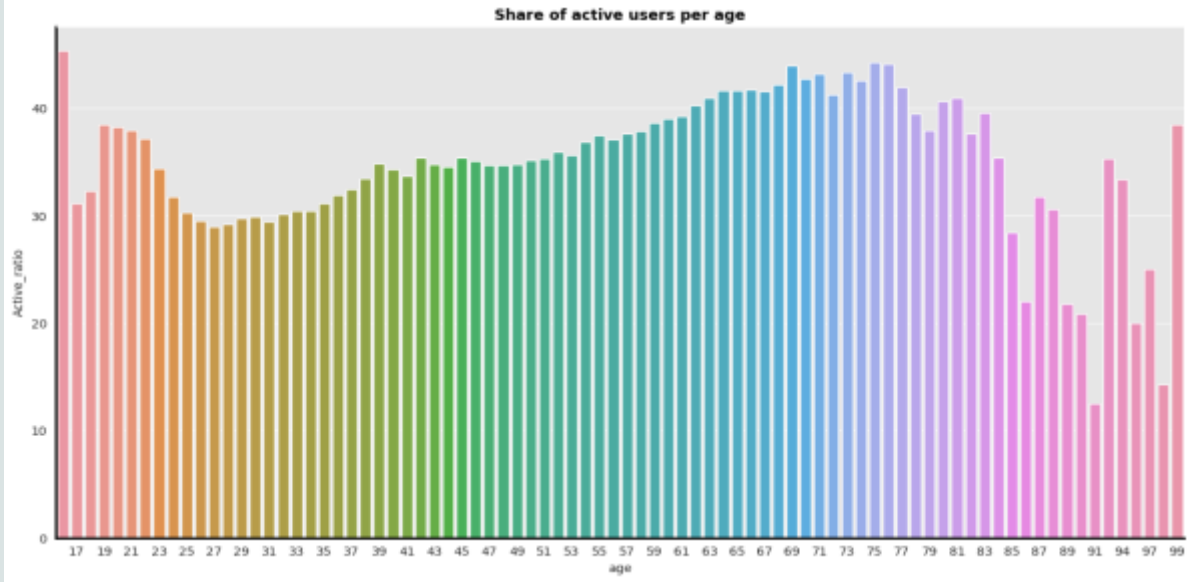

fig. 6: Share of active users per age.

- The oldest active customer was 99 years old.

Findings:

```
Number of subcategories in "Garment Upper body"" is 15.
Number of subcategories in "Underwear"" is 11.
Number of subcategories in "Socks & Tights"" is 3.
Number of subcategories in "Garment Lower body"" is 5.
Number of subcategories in "Accessories"" is 38.
Number of subcategories in "Items"" is 5.
Number of subcategories in "Nightwear"" is 4.
Number of subcategories in "Unknown"" is 1.
Number of subcategories in "Underwear/nightwear"" is 2.
Number of subcategories in "Shoes"" is 16.
Number of subcategories in "Swimwear"" is 6.
Number of subcategories in "Garment Full body"" is 6.
Number of subcategories in "Cosmetic"" is 2.
Number of subcategories in "Interior textile"" is 3.
Number of subcategories in "Bags"" is 6.
Number of subcategories in "Furniture"" is 1.
Number of subcategories in "Garment and Shoe care"" is 6.
Number of subcategories in "Fun"" is 1.
Number of subcategories in "Stationery"" is 1.
```

fig. 7: Subcategories of product group in product type.

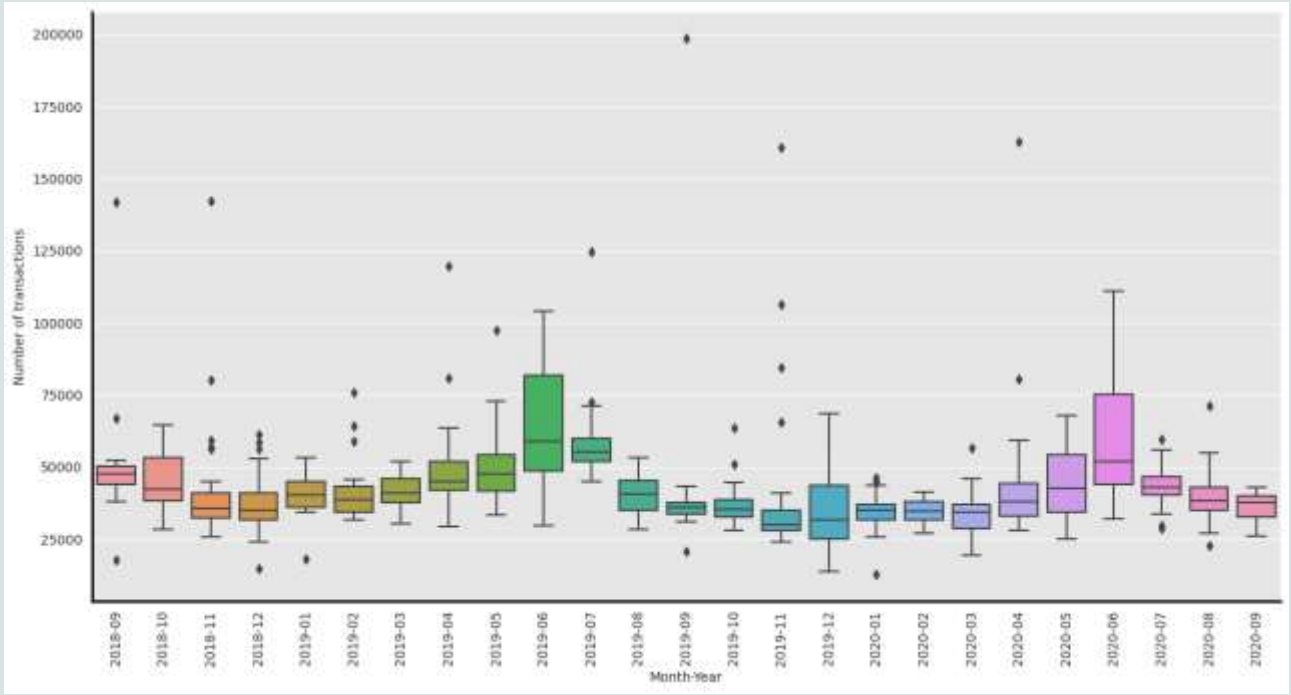

fig. 8: Number of transactions by month

- The accessories and shoes have the biggest number of subcategories.
- The other key takeaways are:
  - The hierarhy of categories is: index_group --> index --> group --> type
  - Over 80% of the products lays in 4 product groups.

- The box-plot indicates that sales spikes during summer time and drops during winter.
- It also indicates that per day usual number of transactions lay in range about between 25000 and 100000 transactions per day.

Findings:

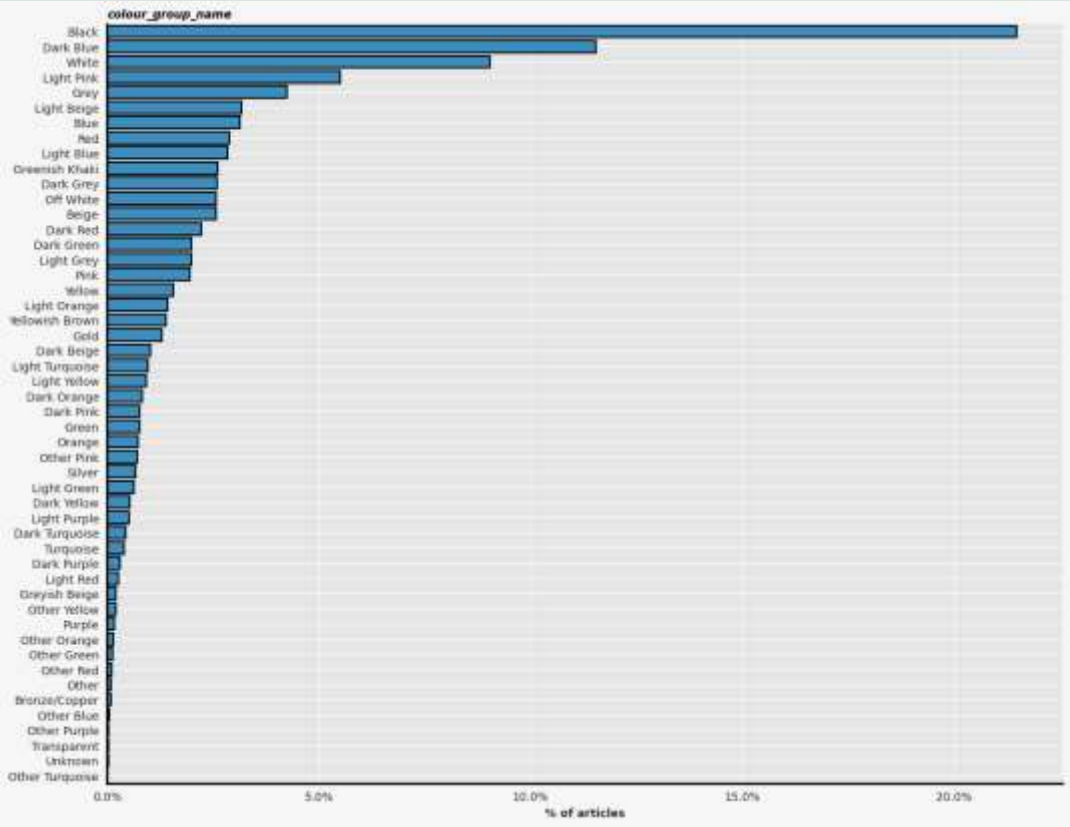

fig. 9: Colors offered in H&M Products

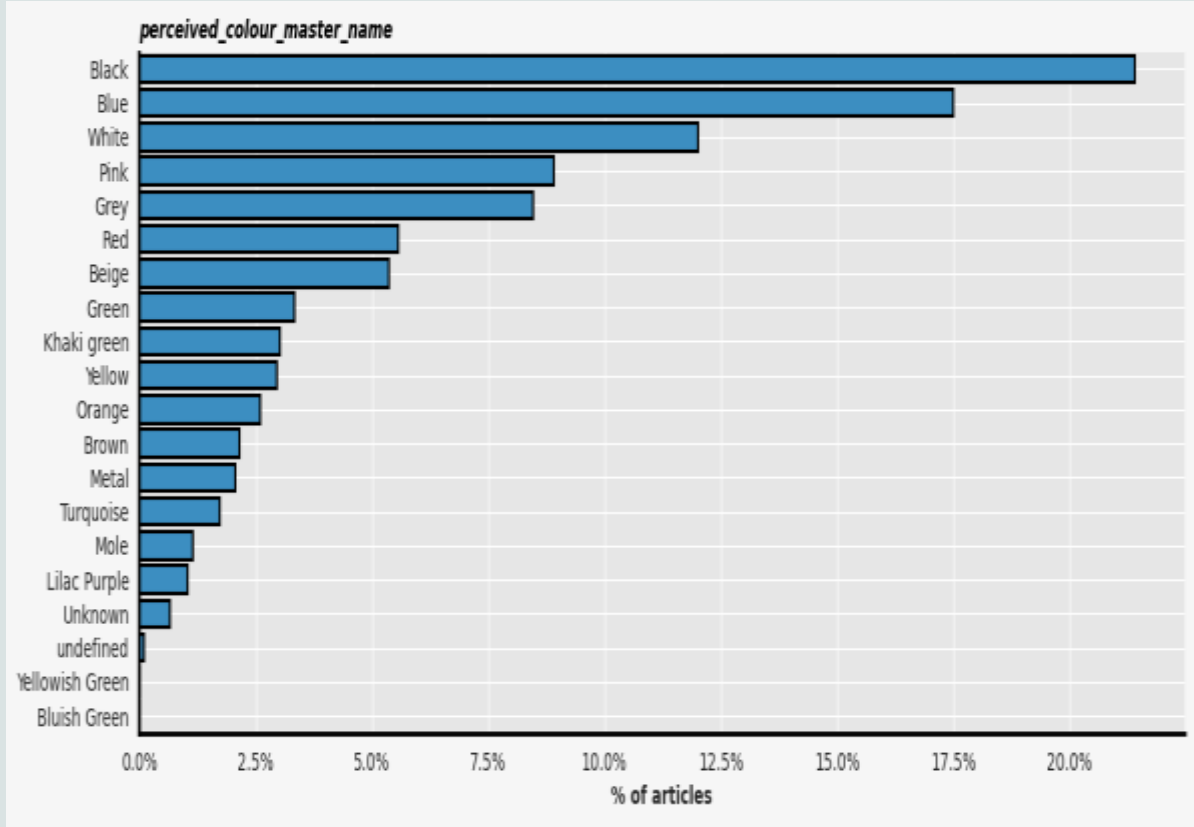

fig. 10: Perceived color of products

- Perceived color is more or less in the line with the color itself but it’s more generic.

Describe the lifecycle of the project:

- Getting the domain knowledge and understanding all 3 datasets.
- Combining the data for exploratory data analysis.
- Cleaning of the data (removing duplicate instances, checking for NA values, and replacing them using binning)
- Exploratory data analysis (with the help of data visualization)
- Feature Reduction/Selection
- Matrix factorization: it is done as our topic of selection is a recommendation system
- Performing collaborative filtering algorithm and content-based filtering model.
- The model's evaluation and comparison to find which product a customer is more likely to buy in the next 7 days.

Thankyou