Group 2

CSN 515: Data Mining & Warehousing

RECOMMENDATION SYSTEM FOR APPAREL STORE

Team members

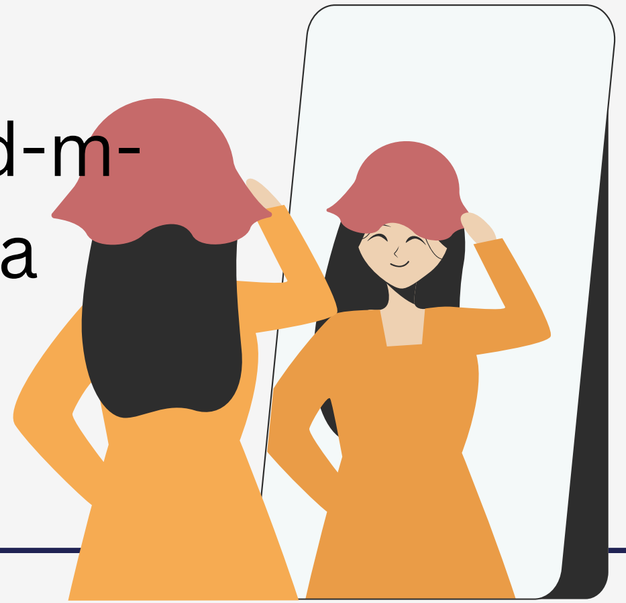
Enrollment No.	Name	Contribution	Page No.
20114013	Anshul Gharde	Association rule using Apriori	15
20114035	Dighiya Sanidhya	Last week's most popular articles	14
20114068	Patare Abhishek	Content based Recommendation	18
20114079	Rahul Kurkure	Customer wise most popular articles	13
20114082	Prabhas Sagiraju	Cluster based Recommendation	19

Problem Statement

P2 : Use data mining to make recommendations for items to buy from an apparel store.

Dataset

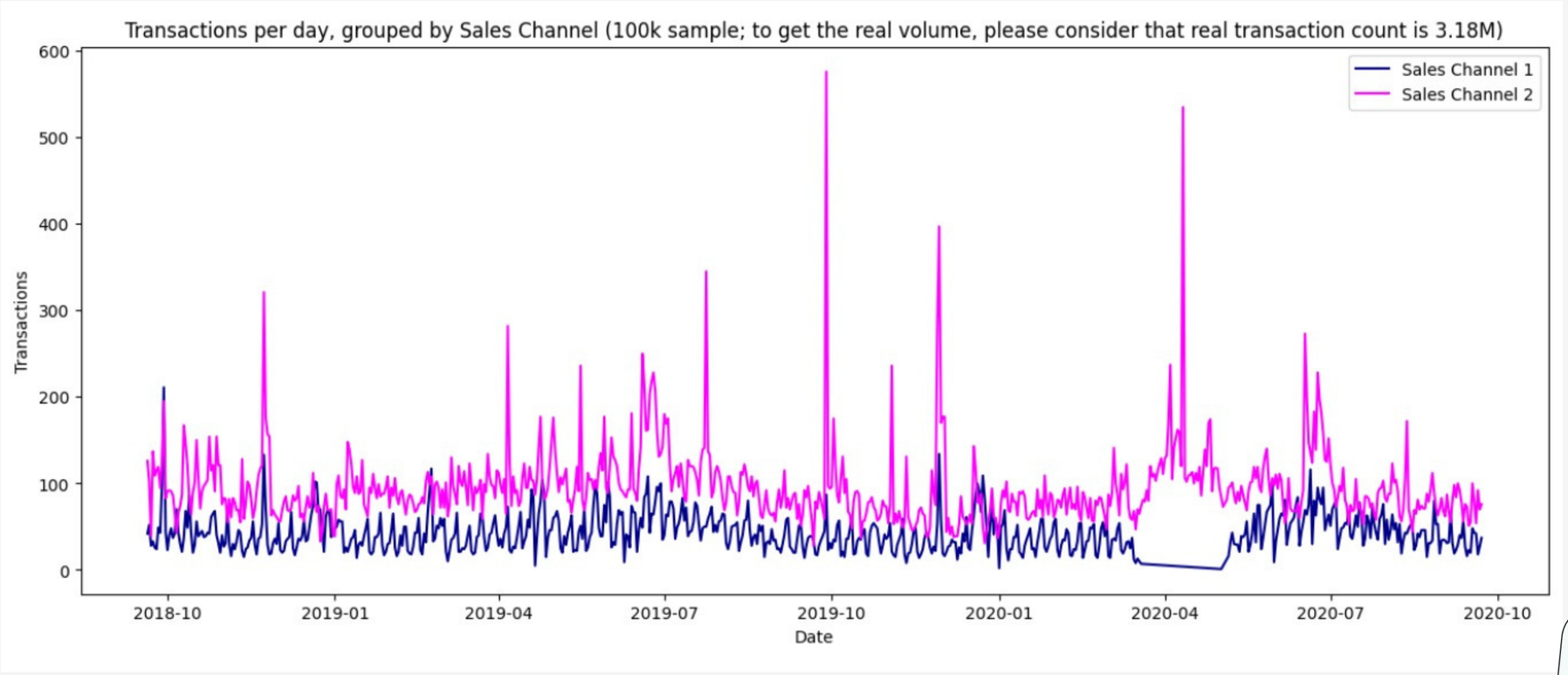
- We took the dataset from the Kaggle H&M competition.
- It provided the data regarding the articles, customers and all the transactions done in two years in the store.
- The data is available here: https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data



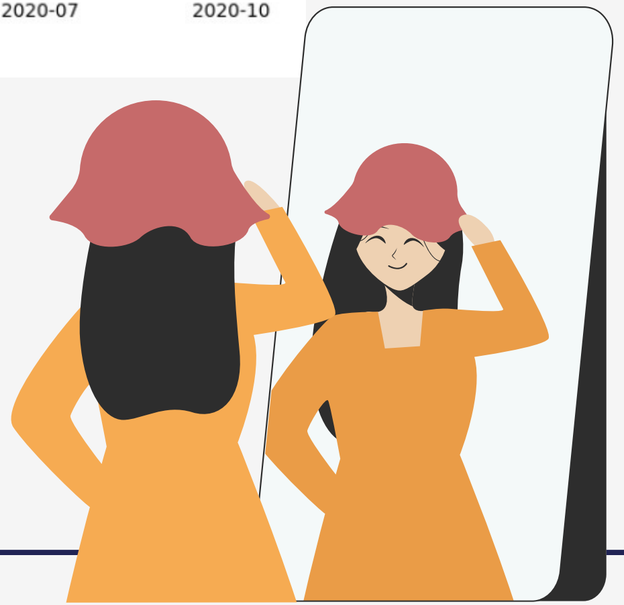
Exploratory Data Analysis

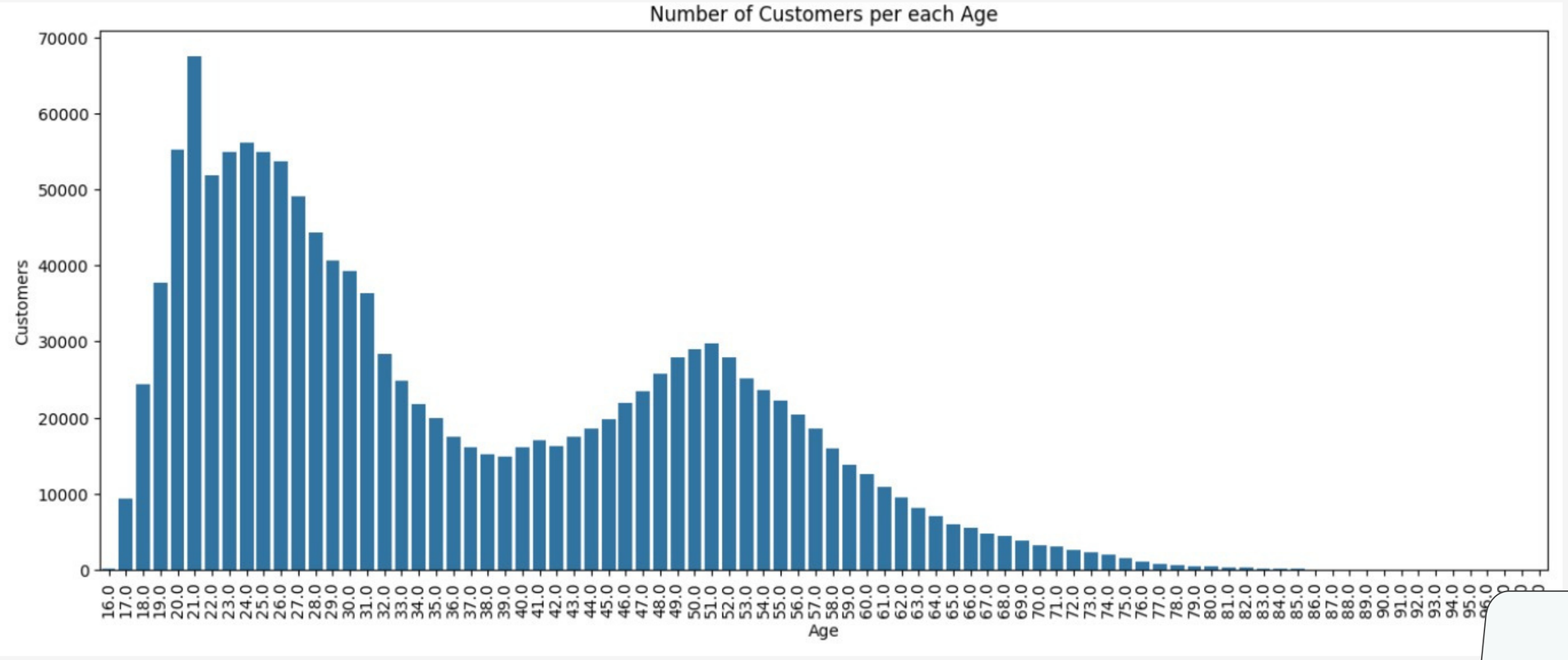
EDA serves as a critical exploratory phase in the data analysis lifecycle, providing analysts with a comprehensive understanding of the dataset's characteristics and informing subsequent analytical tasks, modeling approaches, and decision-making processes.

We did EDA on the given data and found got some insights about the dataset as a whole

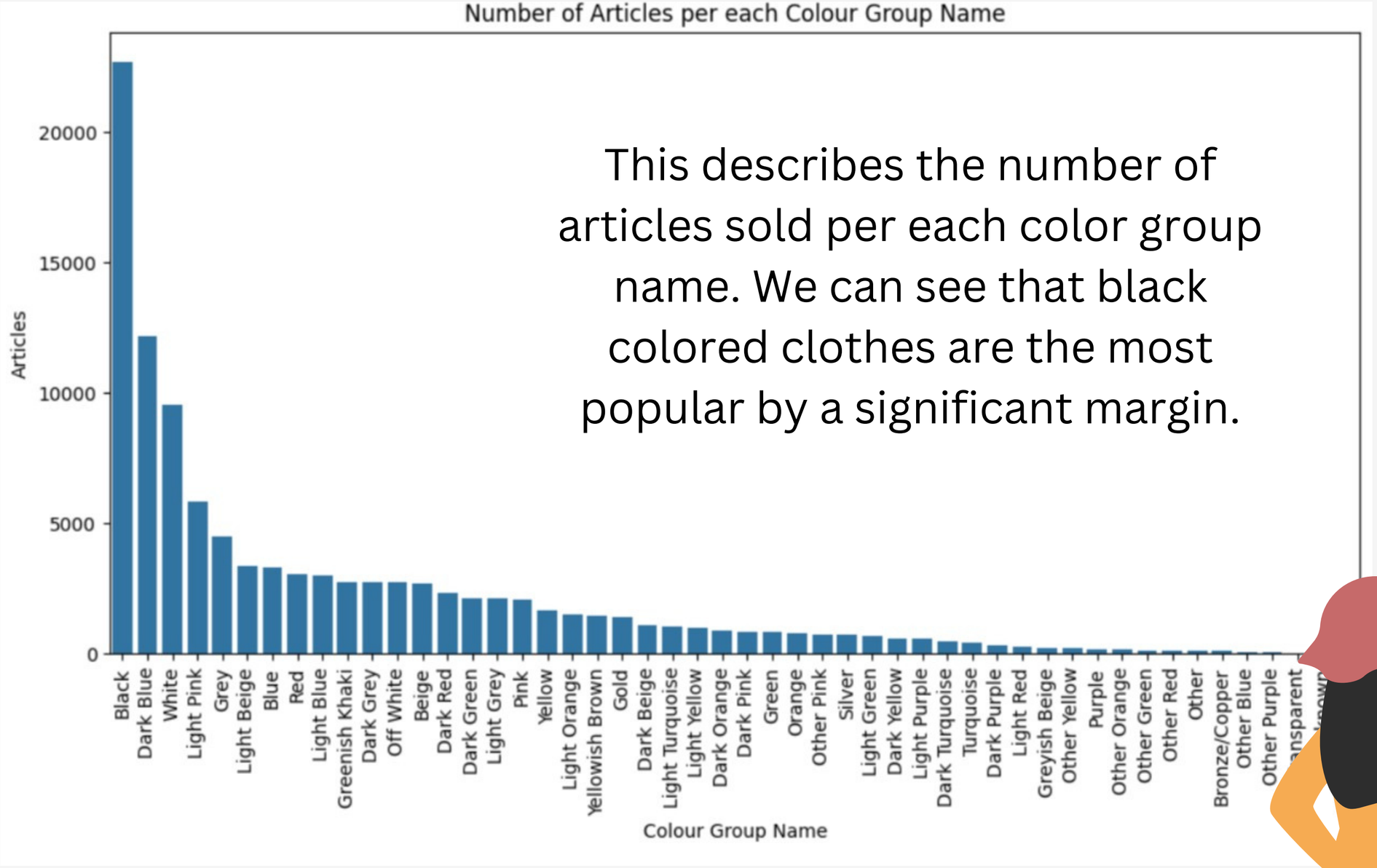


It shows the volume of sales over the 2 year period. We can see the sales have peaked during festive seasons of April(wedding season), October(Dussehra and diwali) and December(Christmas and new year).

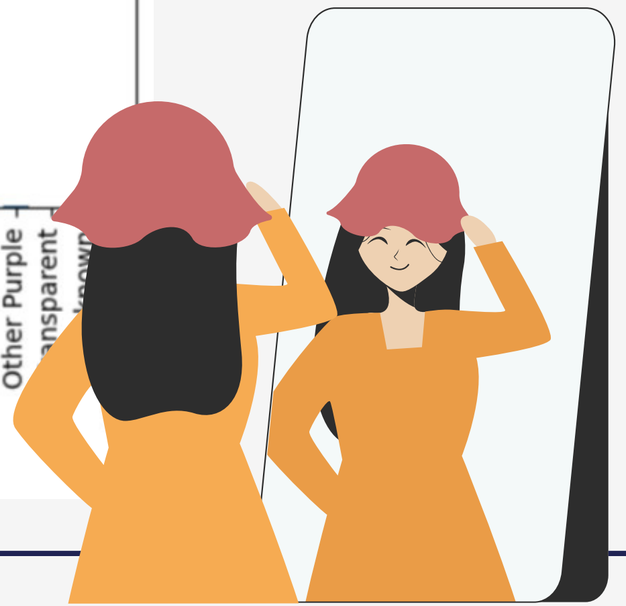


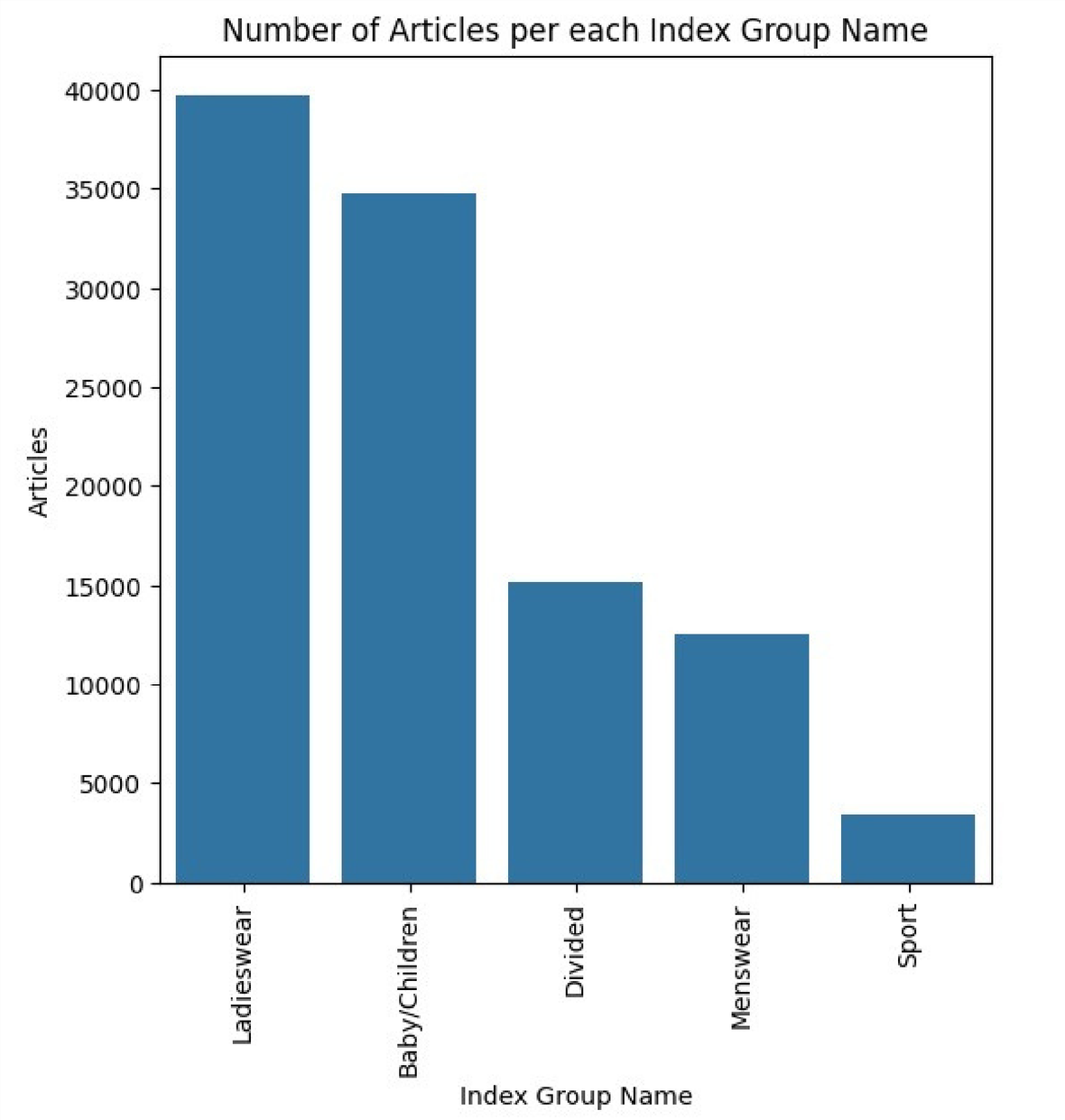


It describes the age distribution of the customers. As, H&M is a fashion brand, we can see that the majority of the customers are from the 20-30 year demographic.

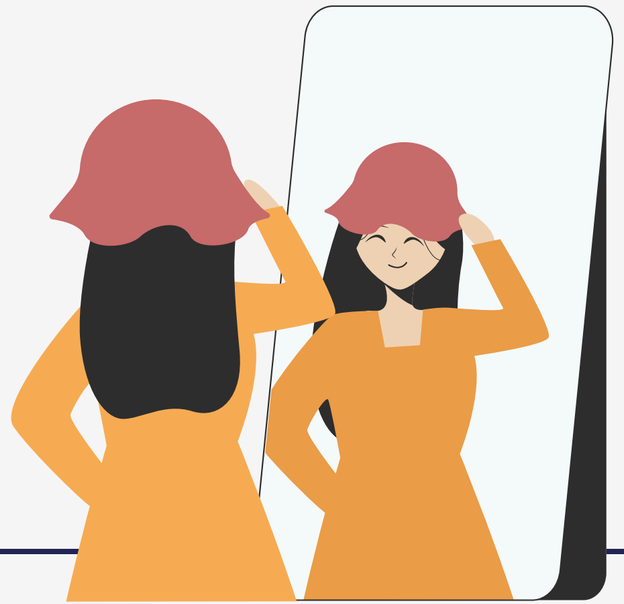


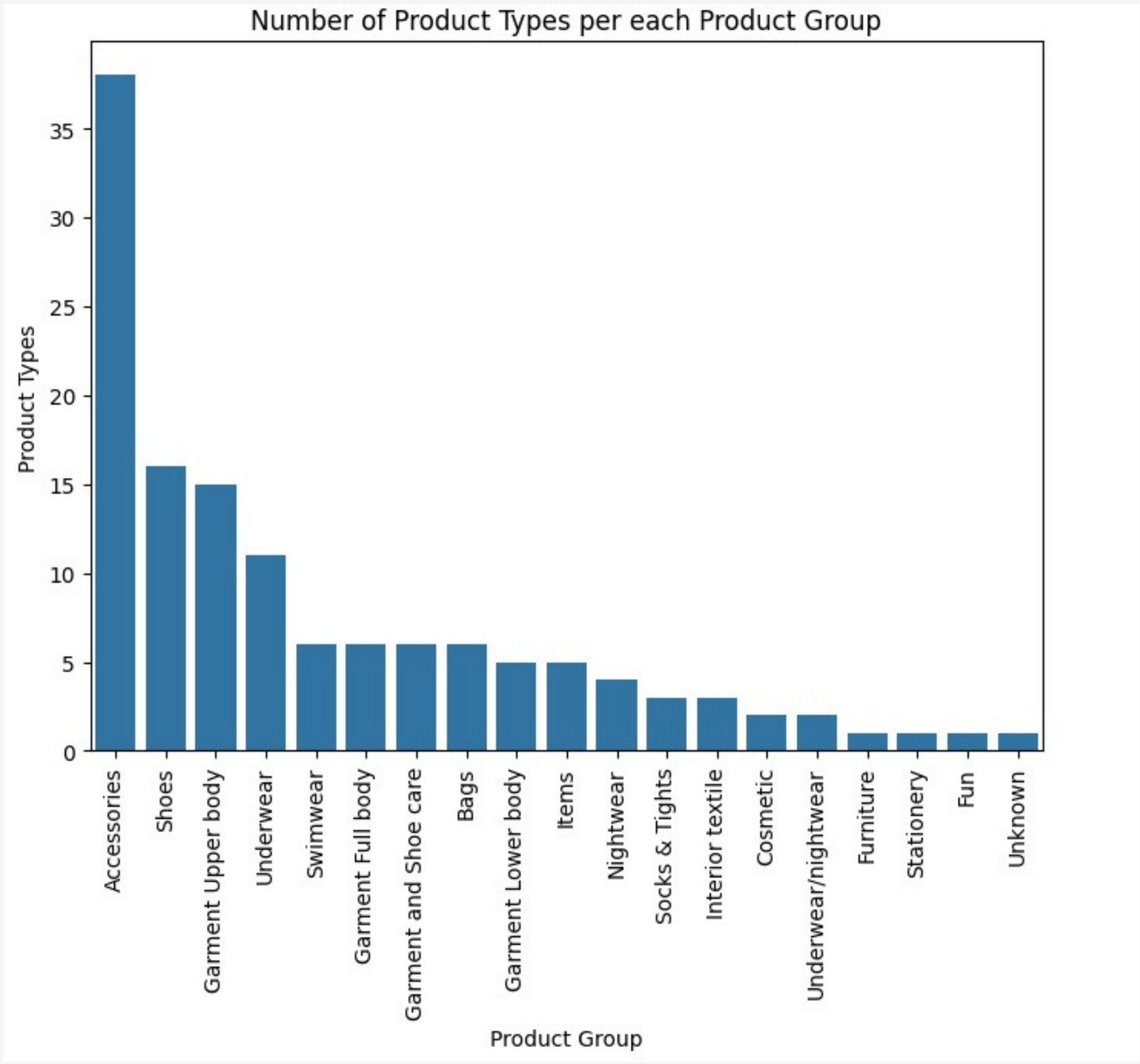
This describes the number of articles sold per each color group name. We can see that black colored clothes are the most popular by a significant margin.





It describes the number of articles per index group name. We can see that ladieswear leads the list followed closely by baby/children wear and then menswear.





It describes the number of product types per each product group. We can see that there are a wide variety of products in the accessories section ranging from spectacles to socks.

General Data Preprocessing

DATA CLEANING

DATA INTEGRATION

DATA/ATTRIBUTE REDUCTION

DATA TRANSFORMATION

Proposed Recommendation Methodology

General Ideas

1. Last week's most popular articles
2. Customer wise most popular articles
3. Making Association rules using Apriori algorithm

Novelty and Innovation

1. Content based recommendation
2. Clustering based on age group

Last week's most popular articles

- A general recommendation to all the customers about the product which is most popular during the last week.
- This gives an idea about the trending product of the apparel store and thus will help to restock the item.

Customer wise most popular articles

- A customer specific recommendation system where we recommend the items which the customer has bought several times in the previous years.

- By looking at the previous trends of a particular customer, we can recommend a new product of same category as well which a customer is likely to buy.

Making Association rules using Apriori algorithm

- An association rule based recommendation system.
- We fed data of one month as the algorithm is memory extensive.
- Generating association rule and predicting the article which customer can buy based on the articles he/she may have bought.

Vectorization

- Vectorization transforms each piece of text into a fixed-length numerical vector representation. This allows NLP models to capture semantic meaning, relationships, and patterns in the text, enabling tasks such as sentiment analysis, document classification, and text generation.

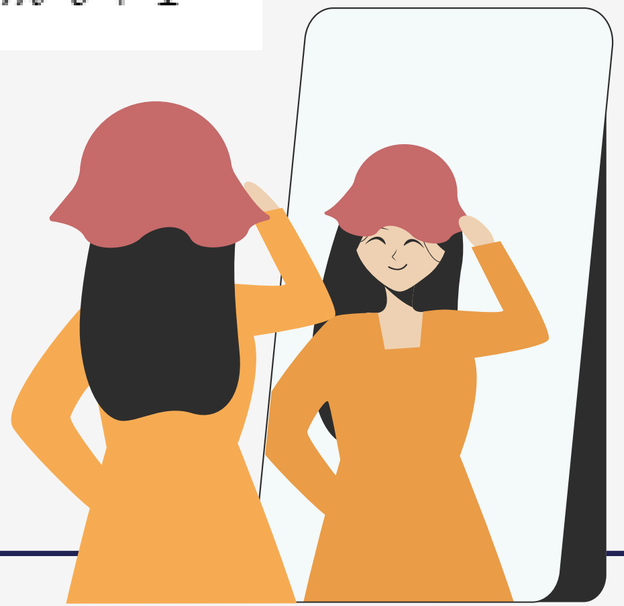
- TF-IDF vectorization: Words that appear frequently in a document but infrequently across all documents are given higher weights, capturing their importance in the document.

Vectorization

$$TF(t,d) = \frac{\text{Number of times term } t \text{ appears in document } d}{\text{Total number of terms in document } d}$$

$$IDF(t,D) = log\left(\frac{\text{Total number of documents in corpus } D}{\text{Number of documents containing term } t + 1}\right)$$

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$



Content based recommendation

- Content based recommendation system where we concatenate all text columns to create a document collection for all articles.
- Then we convert these text document to TF-IDF vectors.
- Now, these vectors are compared using cosine similarity and similar items are recommended to the customer.

Clustering based on age group

- K-means clustering algorithm is applied on the customers based on their age groups and 4 clusters are formed (using elbow method).

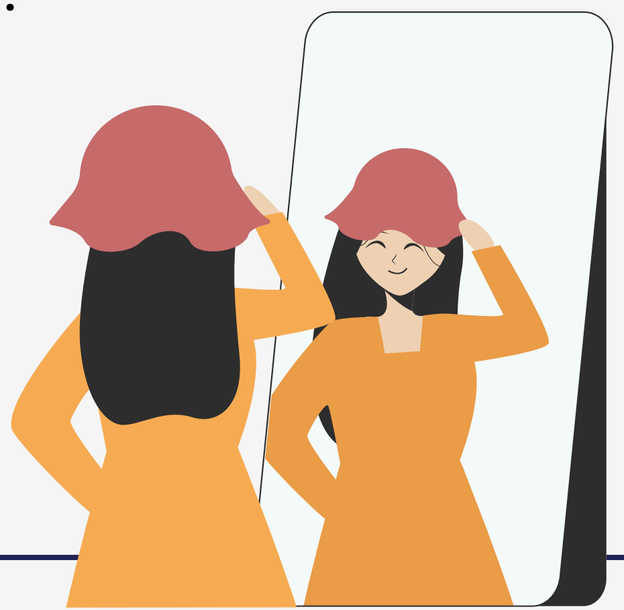
- Now we applied the initial three recommendation methods on each cluster separately to provide age appropriate recommendations.

Conclusion

Thus, our implementation of five distinct recommendation systems for the H&M apparel store dataset demonstrates a commitment to personalized customer experiences and innovative retail strategies.

By leveraging methods ranging from traditional popularity-based approaches to cutting-edge content-based and clustering techniques, we have curated tailored recommendations to meet diverse customer needs.

These systems not only enhance discoverability and engagement but also uncover hidden patterns and preferences, informing targeted marketing efforts and driving customer satisfaction.



Thank you!