H&M Personalized Fashion Recommendations

Group name: Orion

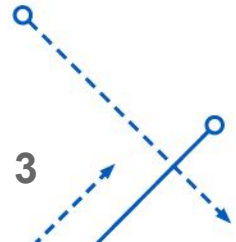
Presenters: Yash Pandey, Mohammed Numan Rizwan

Problem Statement and Research Questions

- Understanding fashion trends that influence buying surge in the clothing segment.
- Picking up on users buying patterns and analyzing it based on users feature like age, location, weather, previous purchases and other features to extract meaningful insight.
- Picking the best hypothesis to predict the garments to be bought by the customer.
- Develop a clothing recommender system that increases the selling of clothing article based on all the above analysis.

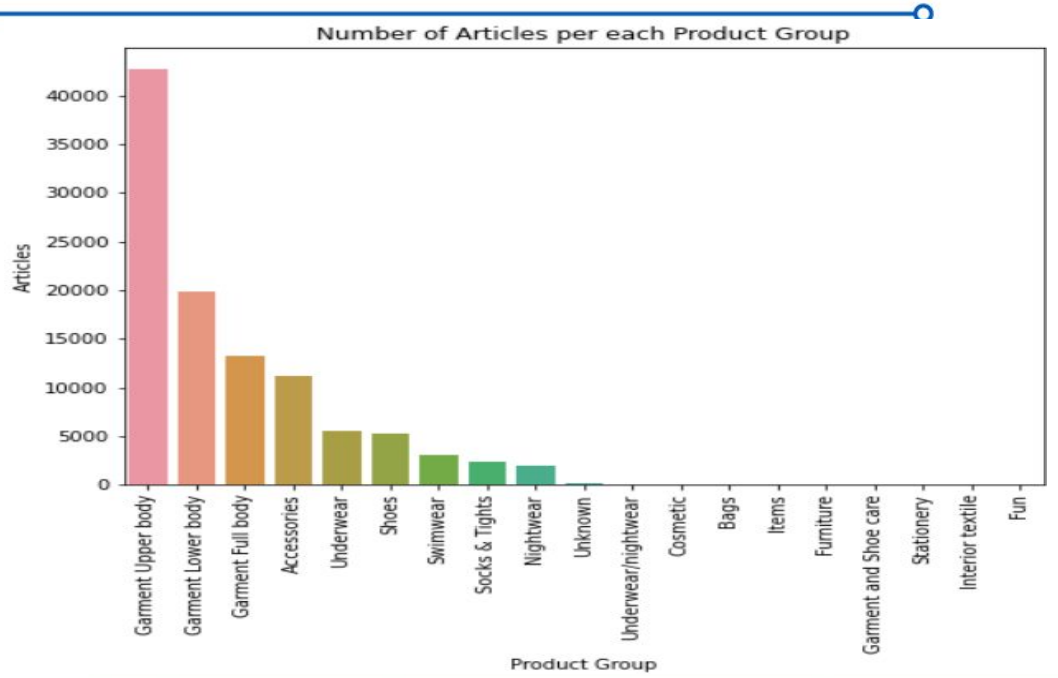
Data

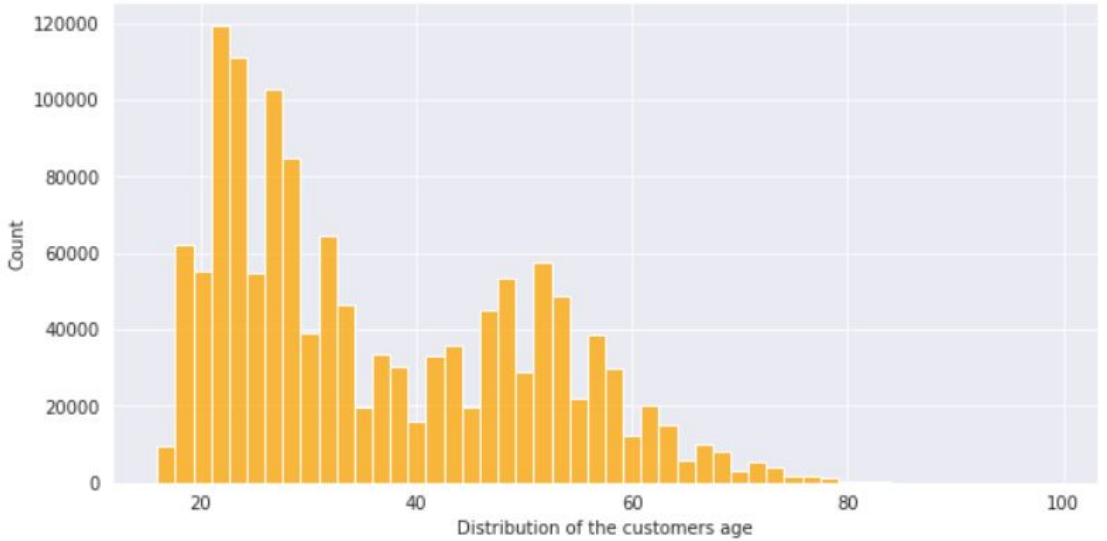
- We have three tabular datasets namely articles, customers and transactions tables. All the tables were the official data provided by H&M.
- The Article table (having 106 thousand observations) contains details (25 columns) about the items, eg article_id, product_code, prod_name, product_type, product_type_name, graphical_appearance_no, and more. This table contains clothing items bought and the features of those items which can be used to extract information on the buying trends.
- The customer table (having 1.37 million observations) contains all the information about the customer which can be used to understand the customers taste, environment and behavior that can be beneficial to understanding and predicting customers buying pattern.
- It has 7 columns: customer id, FN, is active, club member status, fashion news frequency, age and postal code.
- The transaction table (having 31.8 million observations) in a way joins the Article table to the Customer table, as it helps understand the buying patterns and behavior of customer in regards to the clothing item. It also gives insight on frequency of purchase and cost of purchase which are very useful indicators in predictions.
- It has 5 columns: transaction date, customer id, article id and price.



Proposed Approaches

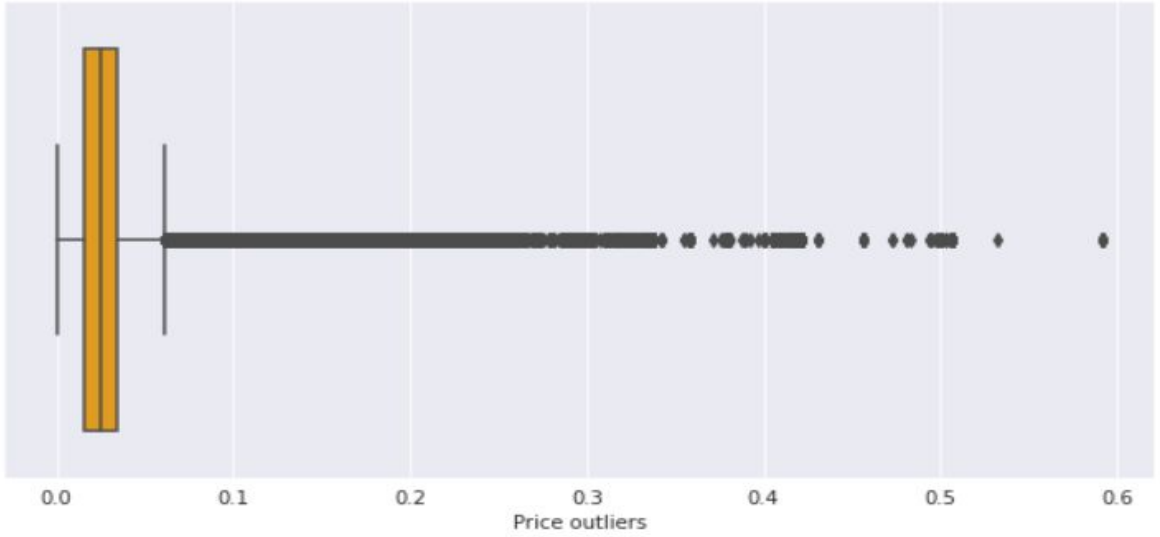
- For this project, we are referring to the purchase history of customers across time, in conjunction with supporting metadata like articles ids of the articles that were purchased, customer ids, purchase days, purchase weeks, number of purchases per item, etc.
- We start by checking the summary of data and clearing the missing data and normalizing the data and analyze the datasets to understand patterns.
- Summary statistics and graphs for exploratory data analysis which consists of the initial analysis of data in order to discover patterns, spot anomalies, test hypotheses, and examines assumptions.
- Word clouds to get a visual representations of textual features in which greater importance is given to the words that appear more frequently.
- Histograms are plotted to get the number of articles per each product group, product type, department and more.
- Histplots are used to get the most and least frequent items in the dataset and the distribution of customers by age.
- Pie Charts can be used to get the distribution of fashion news frequency to understand how frequently customers follow the fashion news and make purchases.
- Identifying if customers buy the same product multiple times.

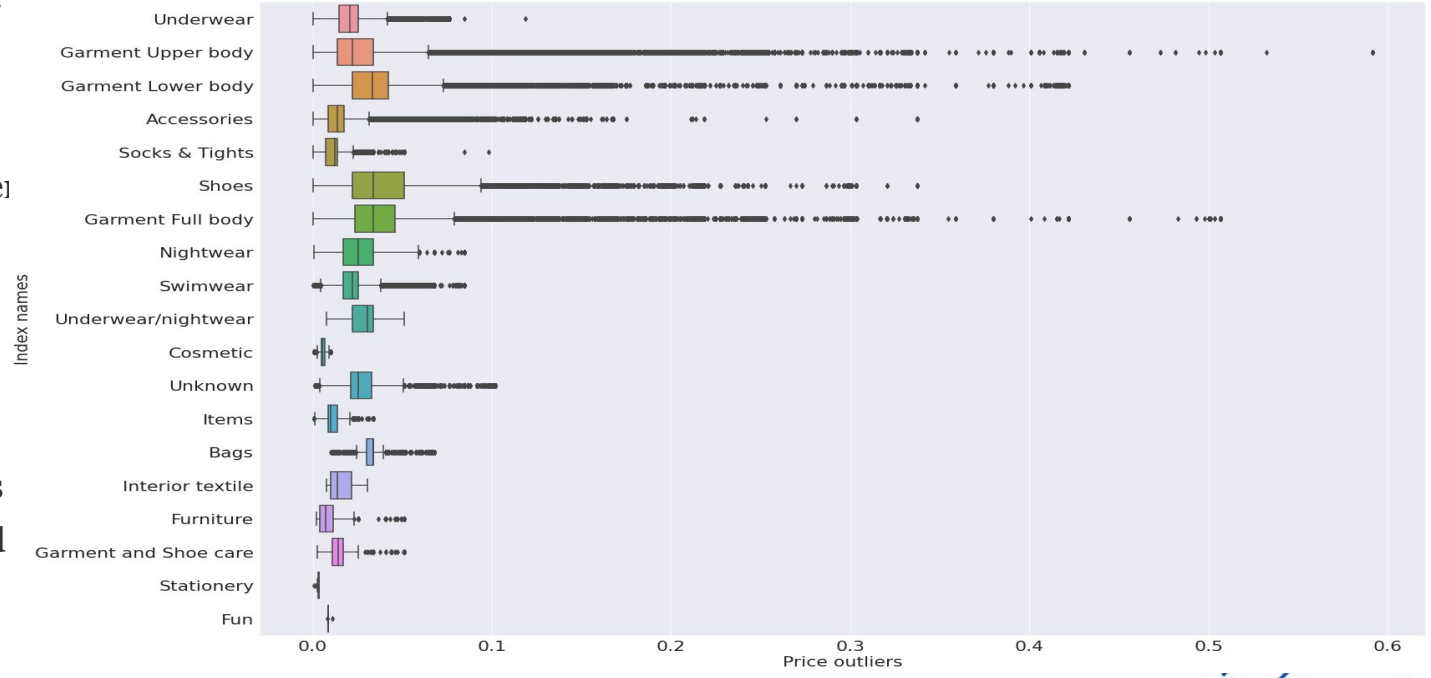




Proposed Approaches continued..

- Box plots are used to find price outliers. Outliers may lead to overfitting of the model which will in turn perform poorly on the new dataset.
- Further analysis is done to find out if customer buys the same items multiple time and the frequency of buying the article. The result is obtained in terms of the percentage of customers that buy the same items again in one, two, or three weeks
- Bar plots are used to visualize the groups of a specific number of top items from the dataset in terms of their mean prices and we can also infer how the mean prices vary over a time period. This gives a clear picture of the articles whose prices keep varying over a period of time and how customers react to the varying prices and change their buying patterns.
- We also try to find out which items were bought together. This analysis can help sellers in determining in the placement of recommended items in the order of likelihood that the two items can be purchased together, which can help in increasing sale and improving recommendations.





University at Buffalo The State University of New York

Intermediate Findings and Discussion

- Found the items purchased together. The top 3 items were considered to be purchased together as they are the most frequent.
- Found each customer's last week of purchases.
- The time series line plot showed that every day, the second sales channel resulted in more unique number of articles sold. From March 2020 to May 2020, there was a steady decrease in the unique number of articles sold for sales channel with id 1.
- The customer with the highest number of transactions has spent $1895.
- 20.7% of customers will buy the same product in one week, 28.1% will buy the same product within two weeks, 32.6% will buy the same product within three weeks In other words, most customers who purchase a product again purchase the same product within three weeks
- Ladieswear accounts for a significant part of all dresses. Sportswear has the least portion. The garments grouped by index: Jersey fancy is the most frequent garment, especially for women and children.
- The most common age of customers is about 21-23. Almost every customer has an active club status, some of them begun to activate it (pre-create). A tiny part of customers has left the club.
- Customers prefer not to get any messages about the current news. Lower, upper and full body garments have a huge price variance.