School of Information Studies
Syracuse University

H&M PERSONALIZED FASHION RECOMMENDATIONS

FINAL PROJECT REPORT

IST 718 Big Data Analytics

Professor Dr. Jon Fox

Winter 2022

TEAM: MS. MELISSA M. MAVERICK | MR. NICHOLAS WAINE
MR. JOSHUA BIGGS-BAUER | MR. NOE FERNANDES

TABLE OF CONTENTS

1. INTRODUCTION

H&M Group, a large fashion company, opened a public competition in February 2022 to develop product recommendations with a top prize of $15,000. H&M Group is a family of brands and businesses with 53 online markets and approximately 4,850 stores. Their online store offers an extensive selection of products, but with too many choices, customers might not quickly find what interests them and ultimately not make a purchase. H&M Group wants to enhance their customer's shopping experience by making personalized product recommendations. H&M Group believes helping customers make better buying choices will have positive implications for sustainability, since it reduces returns which minimizes emissions from transportation.

H&M Group's competition is hosted on Kaggle, a crowd-sourced online machine learning and data science community website that allows users access to huge data repositories and publish data and code. H&M Group invited the Kaggle community to develop product recommendations based on data from previous transactions, customer data, and product metadata. Competitors are free to choose any methods of analysis and modeling, such as categorical algorithms, natural language processing, image processing, or deep learning.

Our project team decided to conduct a detailed exploratory data analysis of the H&M data, including merging datasets for a deeper understanding of the trends and patterns. Additionally, two models were selected to generate product recommendations, including a simple baseline model and a content-based filtering model. The simple baseline model used the most frequent recently purchased products to make product recommendations and the content-based filtering model used product scoring to make product recommendations. The analysis and modeling were performed using Python in Juptyer Notebook and Google Colaboratory.

2. THE DATA

H&M provided four comma separated values (CSV) files and a folder of product images uploaded to Kaggle (available here). The CSV files included detailed metadata for each article available for purchase (*articles.csv*), customer metadata (*customers.csv*), transaction data consisting of customer purchases by date (*transactions_train.csv*), and a sample submission file with the required format (*sample_submission.csv*). The folder (*images*) contained images corresponding to articles available for purchase, although not every article had a corresponding image.

The articles data consisted of 25 columns and 105,542 unique articles. The columns contained categorical and numeric values. Most of the numeric values are identification codes associated with the categorical data. For example, a product code of 108775 corresponds to the product name "strap top," and a color group code of 10 corresponds to the color group “black.” The dataset contained 416 missing values exclusively in the product description column, which accounted for only 0.4% of the entire dataset. The column names and descriptions are displayed in Table 1.

Table 1. Articles Data Column Names

Column Name	Description
article_id	Unique identification number for each product
product_code	Numeric designation for product name (not unique)
prod_name	Product name (e.g., “strap top”)
product_type_no	Numeric designation for product type name (not unique)
product_type_name	Product type (e.g., “vest top”)
product_group_name	Product group (e.g., “garment upper body”)
graphical_appearance_no	Numeric designation for graphical appearance name (not unique)
graphical_appearance_name	Product graphical (pattern) appearance (e.g., “solid”)
colour_group_code	Numeric designation for color group name (not unique)
colour_group_name	Product color group (e.g., “black”)
perceived_colour_value_id	Numeric designation for perceived color value name (not unique)
perceived_colour_value_name	Product perceived color value (e.g., “dark”)
perceived_colour_master_id	Numeric designation for perceived color master name (not unique)
perceived_colour_master_name	Product perceived color master name (e.g., “black”)
department_no	Numeric designation for department name (not unique)
department_name	Product department (e.g., “jersey basic)
index_code	Alphabetic designation for index name (not unique)
index_name	Product index (e.g., “lingeries/tights”)
index_group_no	Numeric designation for index group name (not unique)
index_group_name	Product index group (e.g., “ladieswear”)
section_no	Numeric designation for section name (not unique)
section_name	Product section (e.g., “womens everyday basic”)
garment_group_no	Numeric designation for garment group name (not unique)
garment_group_name	Product garment group (e.g., “jersey basic”)
detail_desc	Product description (e.g., “Jersey top with narrow shoulder straps)

The customer data consisted of 7 columns and 1,371,980 unique customers. The columns contained categorical, hash code, and numeric values. The customer identification and postal code columns did not have any missing values. Age and Fashion News frequency had only about 1% missing value, but Fashion News enrollment (FN) and the numeric value for Fashion News enrollment (Active) had about 65% missing values. The missing values represented a lack of enrollment in Fashion News. The column names and descriptions are displayed in Table 2.

Table 2. Customer Data Column Names

Column Name	Description
customer_id	Unique hash code for each customer
FN	Numeric value representing customer enrollment in Fashion News
Active	Numeric value for club member status
club_member_status	Club member status (e.g., ACTIVE)
fashion_news_frequency	Frequency of Fashion News enrollment
age	Numeric age of customer
postal _code	Hash code of customer zip code

The transaction data consisted of 5 columns and 31,788,324 transactions. The columns contained hash code and numeric values. The transaction data did not have any missing values. The column names and descriptions are displayed in Table 3.

Table 3. Transaction Data Column Names

Column Name	Description
t_dat	Transaction date (YYYY-MM-DD)
customer_id	Unique hash code for each customer
article_id	Unique identification number for each product
price	Price of article
sales channel	Numeric value corresponding to in-store or online purchase

The sample submission file was an example of how H&M required the output for the competition submission files. H&M wanted one column with the customer identification hash code and one column with the recommended article identification numbers.

The images folder contained 86 subfolders which contained a total of 105,100 product images. A folder of images corresponded to each article identification. Images were placed in subfolders starting with the first 3 digits of the article identification. Each image file name was its article identification number. Not all article identification values had a corresponding image (total of 442 missing images). Figure 1 displays examples of the image files.

Figure 1. Image files from the H&M data.

3. EXPLORATORY DATA ANALYSIS

The articles, customers, and transactions data sets were investigated to understand the data and any individual patterns or trends. The articles and transactions data were merged and explored, as well as the images and articles data, to get a broader understanding of the associations between the data.

3.1 ARTICLES

The articles data consisted of 105,542 unique articles. Since this data is only products offered for sale, the main investigative technique was to look at the number of items offered in each product group and all of subgroups to get a better understanding of what H&M currently offers to customers. Grouping techniques and counts were used on various pairings. The articles data had 12 main groups, including product name, product type, product group, graphical appearance, color group, perceived color, department name, index name, index group, section name, garment group, and detailed description. The top label of the hierarchy appeared to be product type, which had 131 unique values. Figure 2 displays the number of articles in the top 50 product types by count.

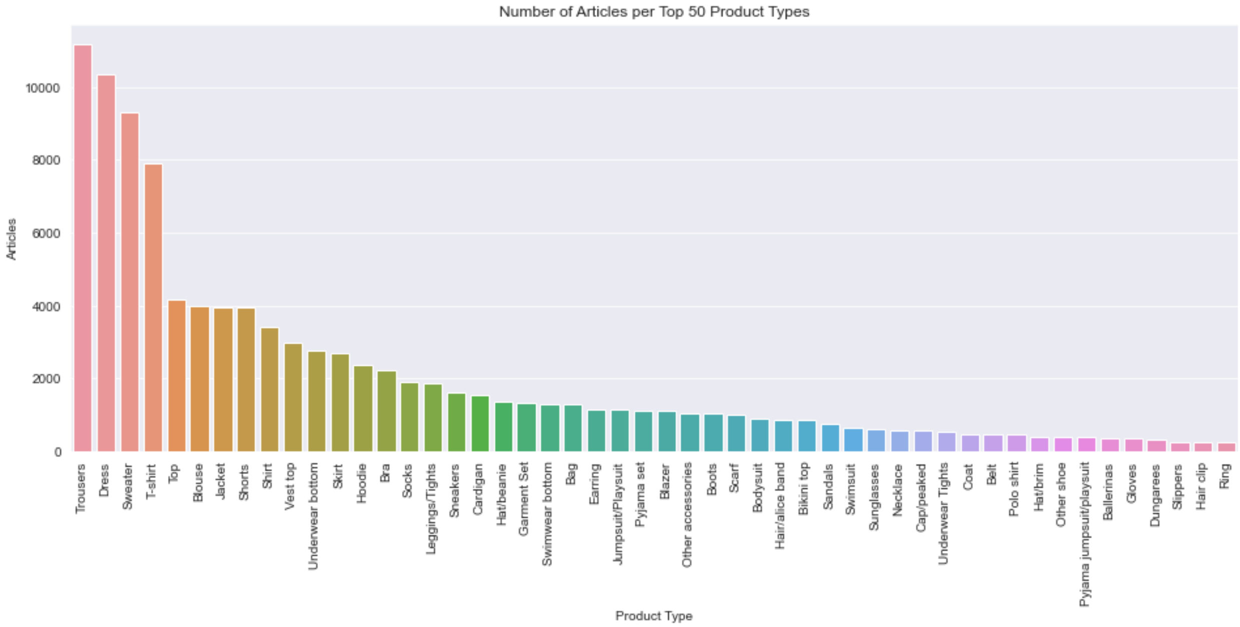


Figure 2. Top 50 product types by number of articles, highest to lowest.

The top four product types were trousers, dress, sweater, and T-shirt. There was then a steep drop in products offered, from ~8,000 to ~4,000, after the top four groups. The next label explored was articles by department. Figure 3 displays the number of articles in the top 50 departments by count.

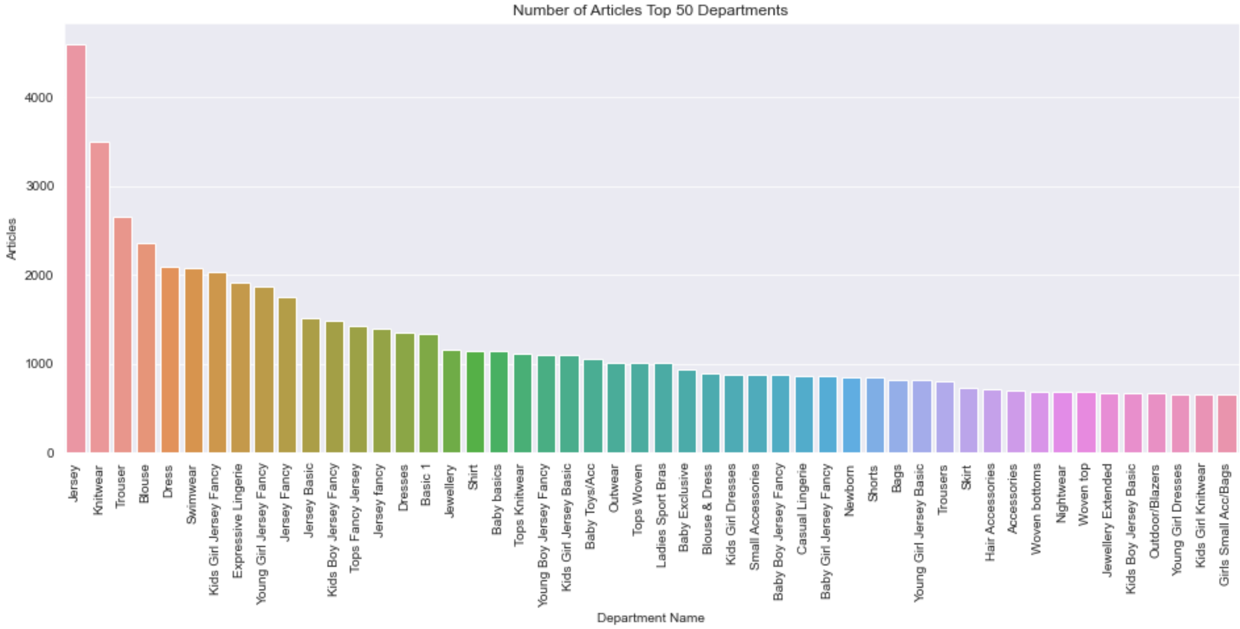


Figure 3. Top 50 product types by number of articles, highest to lowest.

The top five departments were jersey, knitwear, trouser, blouse, and dress. Again, there was a steep drop in numbers within just the top five groups, from ~4,500 to ~2,000. The next label explored was articles by color group. Figure 4 displays the number of articles in all color groups.

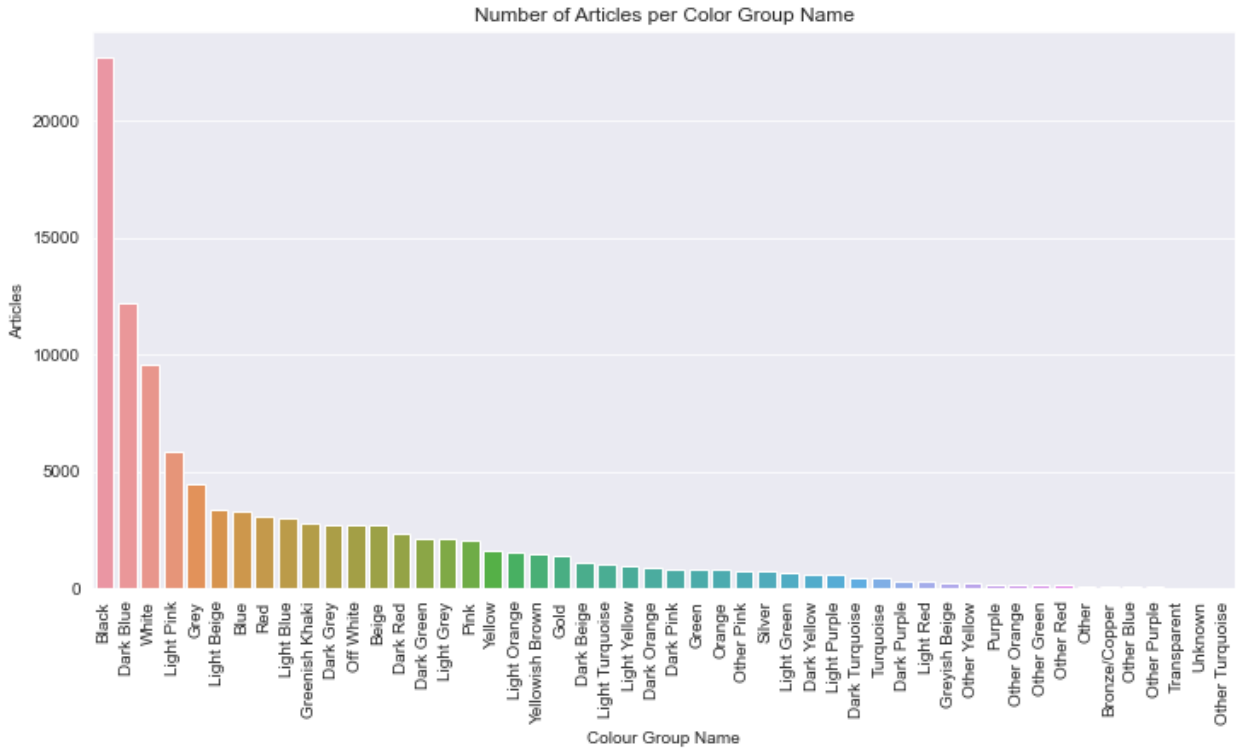


Figure 4. Number of articles in each color group, highest to lowest.

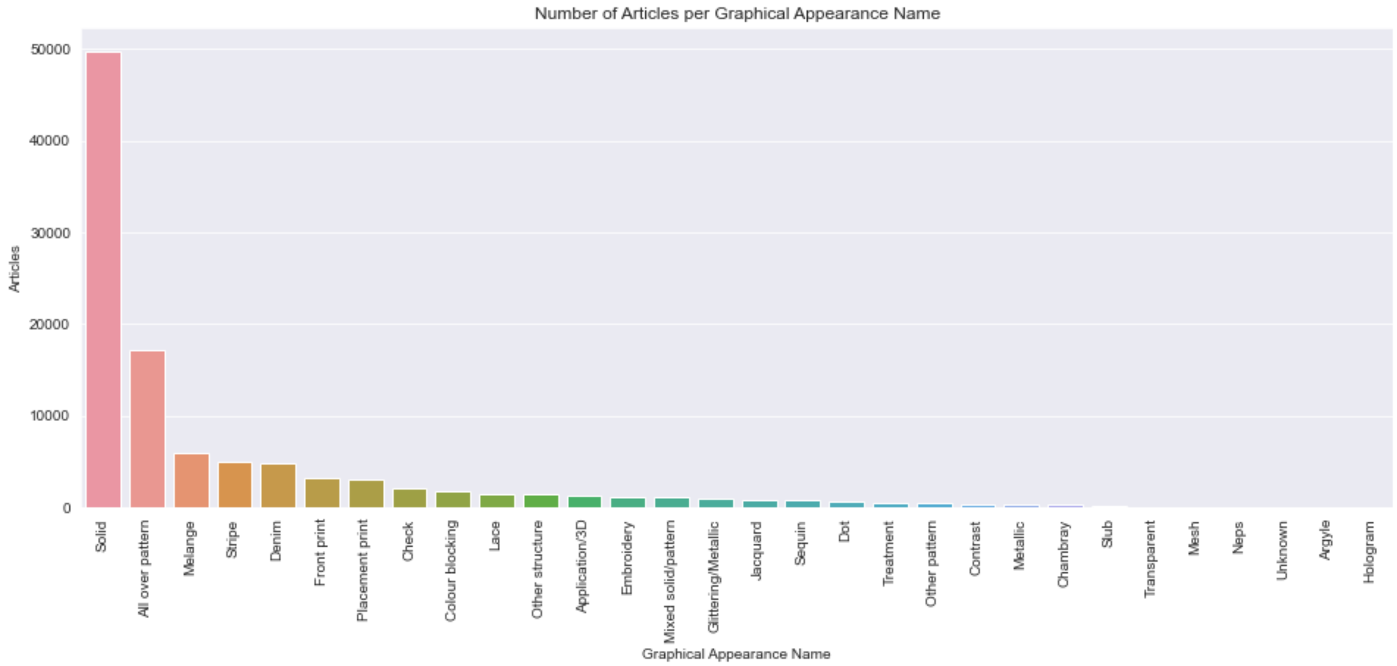
The top five color groups were black, dark blue, white, light pink, and grey. Black was by far the highest number at almost 230,000 products, while dark blue had roughly 150,000, followed by white at just under 100,000. Lastly, we looked at articles by graphical appearance. Figure 5 displays the number of articles in all graphical appearance groups.

Figure 5. Number of articles in each graphical appearance group, highest to lowest.

The top five graphical appearance groups were solid, all-over pattern, mélange (i.e., mixture or medley), stripe, and denim. Solid was by far the highest number at just under 50,000, followed by all-over pattern at roughly 17,000 products. This concluded exploratory analysis of the articles data.

3.2 CUSTOMERS

The customer data consisted of 1,371,980 unique customers. The customers data did include many variables that allowed for statistical analysis. We explored customer ages, H&M club member status, enrollment in H&M's Fashion News, and



frequency of Fashion News preference. The vast majority of customers were enrolled in H&M's club program while also *not* being enrolled in the Fashion News program. Of those that were enrolled, the majority received Fashion News "regularly" while a very small portion opted for "monthly." Figure 6 displays customers by age.

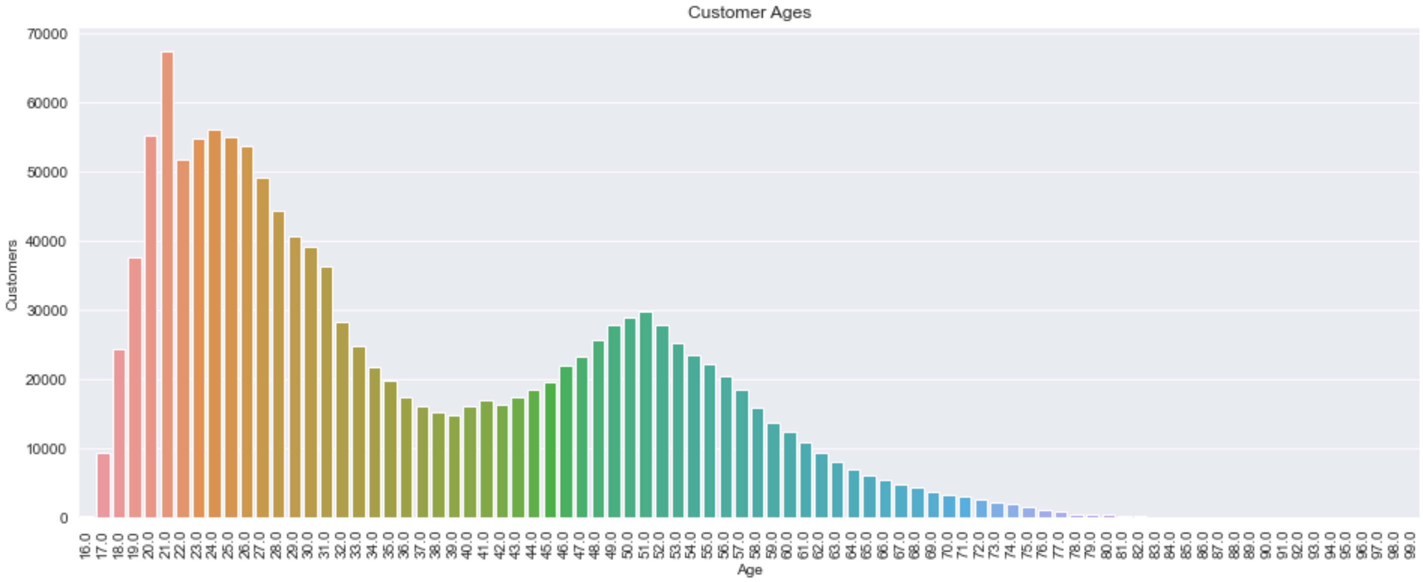


Figure 6. H&M customers by age.

The highest age group was 21-year-olds, with the majority of customers between 19 to 31 years old. A second peak of customer ages fell between 45 to 56 years old. Since H&M is normally marketed towards younger clientele, we speculate that the 45-56-year-old group of customers may be purchasing from H&M for children or grandchildren, although there is no data to support this assumption. This concluded exploratory analysis of the customers data.

3.3 TRANSACTIONS

The transaction data consisted of 31,788,324 customer transactions. This data had robust numerical data for price per article, purchase date, and method of purchase. *Sales Channel 1* represented in-store purchases and *Sales Channel 2* represented online purchases. We explored transactions by sales channel, price, article, and time series analysis. First, we looked at the logarithmic distribution of price frequency by sales channel. Since the prices in the transaction data were already transformed by H&M to keep that information private from competitors, a logarithmic transformation provided better insight into any patterns or trends in the data. We used a sample size of 100,000 transactions for the distribution. The logarithmic distribution reveals there is not much difference overall between the two sales channels and that the distribution is slightly bimodal. Figure 7 displays the logarithmic distribution of price frequency by sales channel.

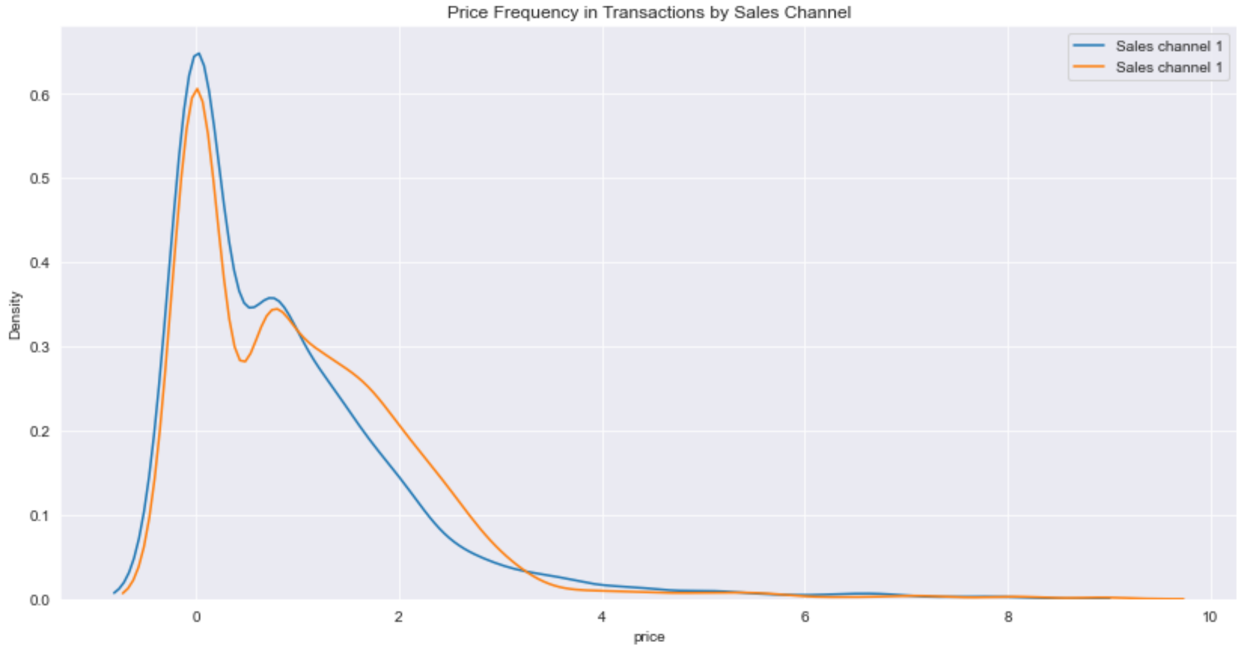


Figure 7. Logarithmic distribution of price frequency by sales channel.

Next, we looked at the number of transactions per day, number of transactions per day by sales channel, and the number of unique articles purchased per day by sales channel. There is a clear difference between the transaction by sales channel: online sales were significantly higher than in-store sales, even when looking at unique articles. Since the data provided includes 2020, there is a gap for in-store sales due to pandemic store closures. Figure 8 displays the transactions per day by sales channel using a sample of 300,000 transactions. Figure 9 displays the unique articles per day by sales channel.

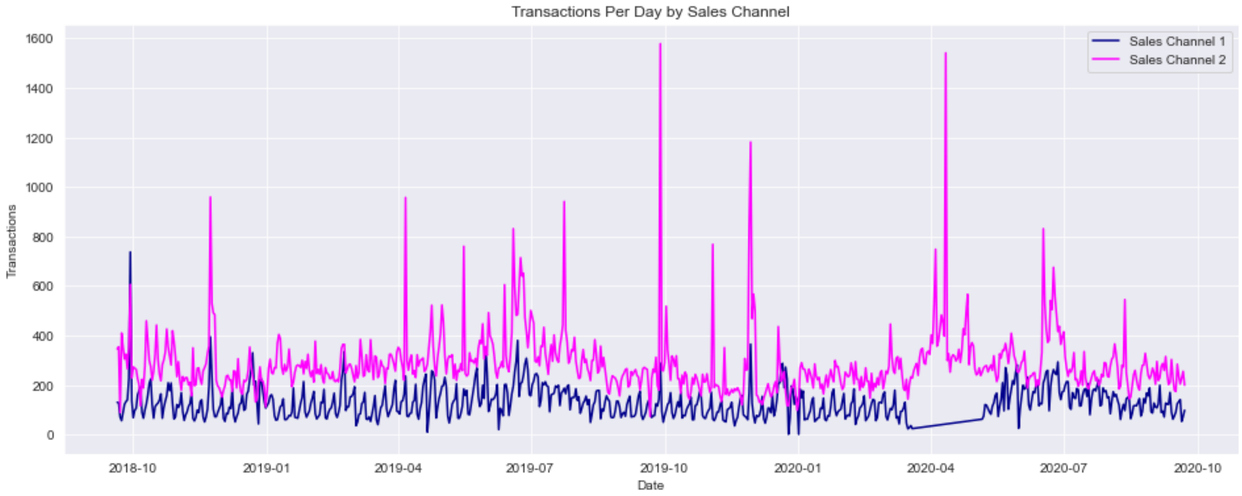


Figure 8. Number of transactions per day grouped by sales channel.

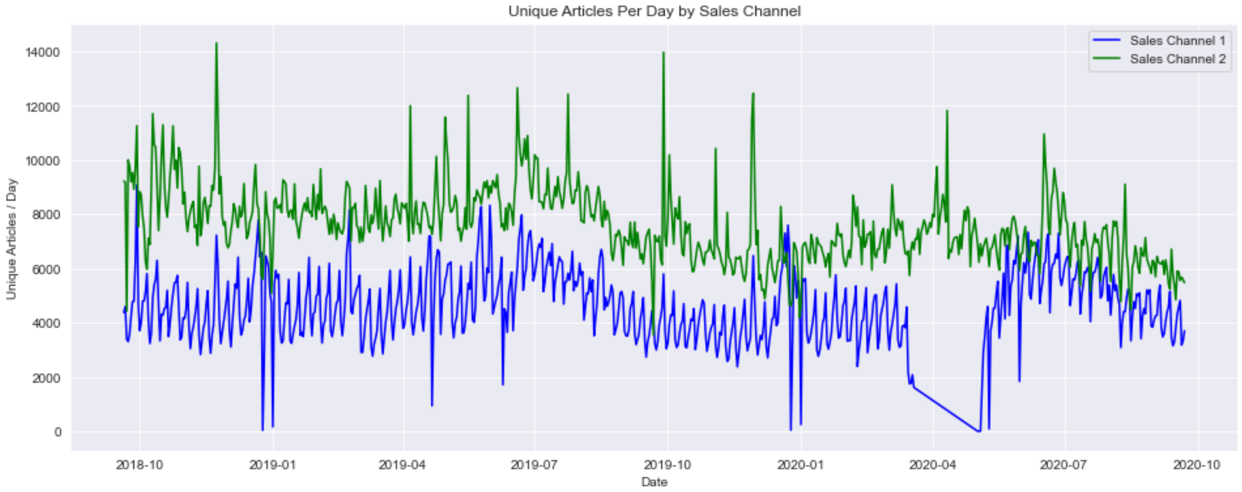


Figure 9. Number of unique articles purchased per day grouped by sales channel.

3.4 MERGED DATA

After exploring the data separately, the articles and transactions data were merged, as well as the images and articles data for further data exploration.

3.4.1 ARTICLES & TRANSACTIONS

The articles and transaction data were merged by using the article identification. Boxplots were generated to explore the price ranges of each product group and look for outliers. Figure 10 displays the boxplots for the product groups.

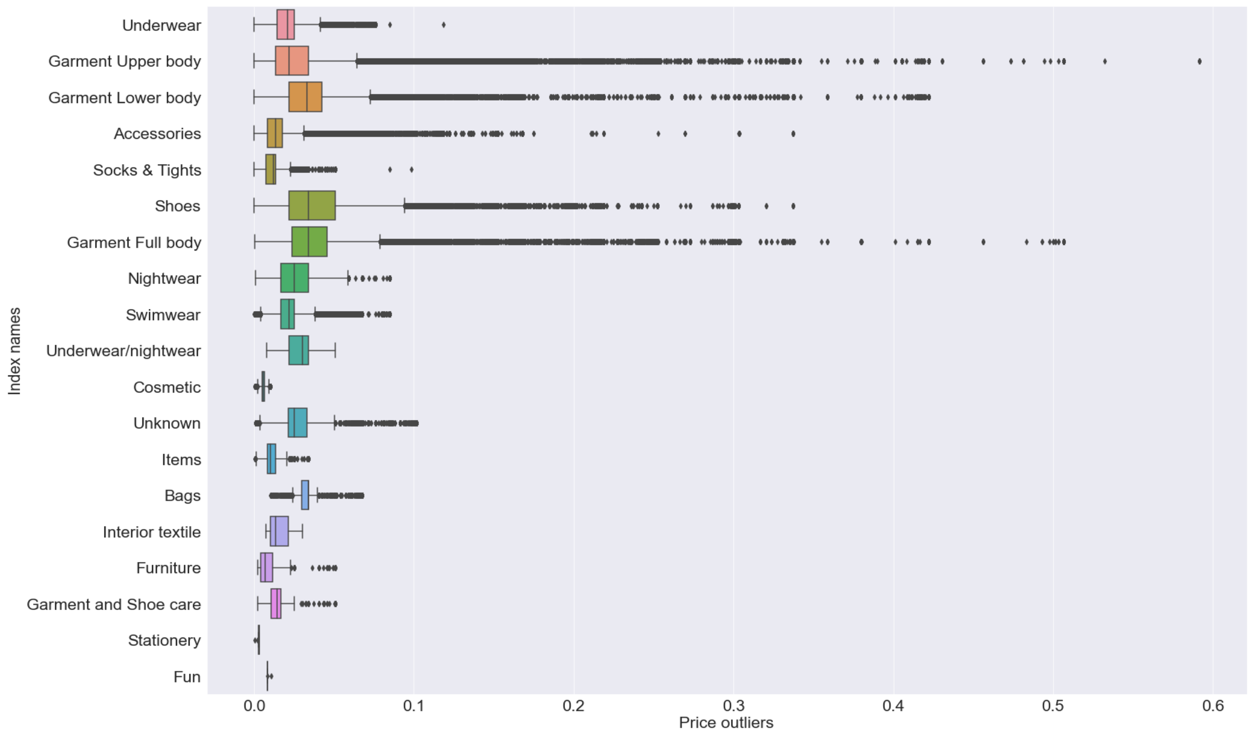


Figure 10. Price boxplots for all product groups.

Based on the outlier exploration, we generated a boxplot for the product group with the largest outliers, which was *garment upper body*, to look at the item sub-groups within that product group and their price ranges. Figure 11 displays the boxplots for the *garment upper body* product group.

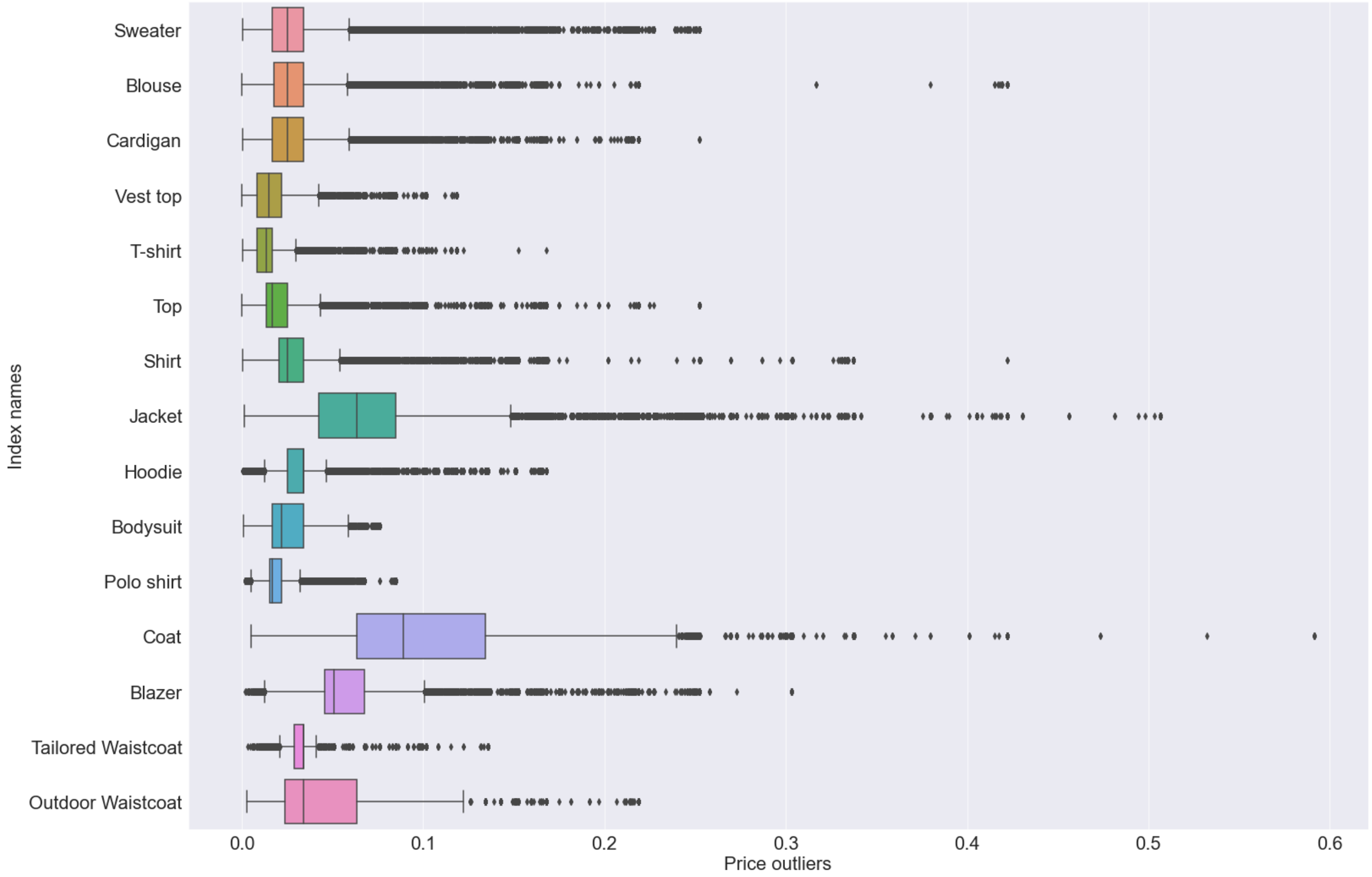


Figure 11. Price boxplots for *garment upper body* product group.

Next, we looked at mean prices by index and product group. The top five highest mean price indexes were (in order) ladieswear, sport, menswear, divided, and lingeries/tights. The top five highest mean prices by product group were for shoes, garment full body, bags, underwear/nightwear, garment lower and body. For a deeper understanding, we looked at the mean price changes over time for the top five product groups. Figure 12 displays the mean price changes over time, including the standard deviations around the trendlines.

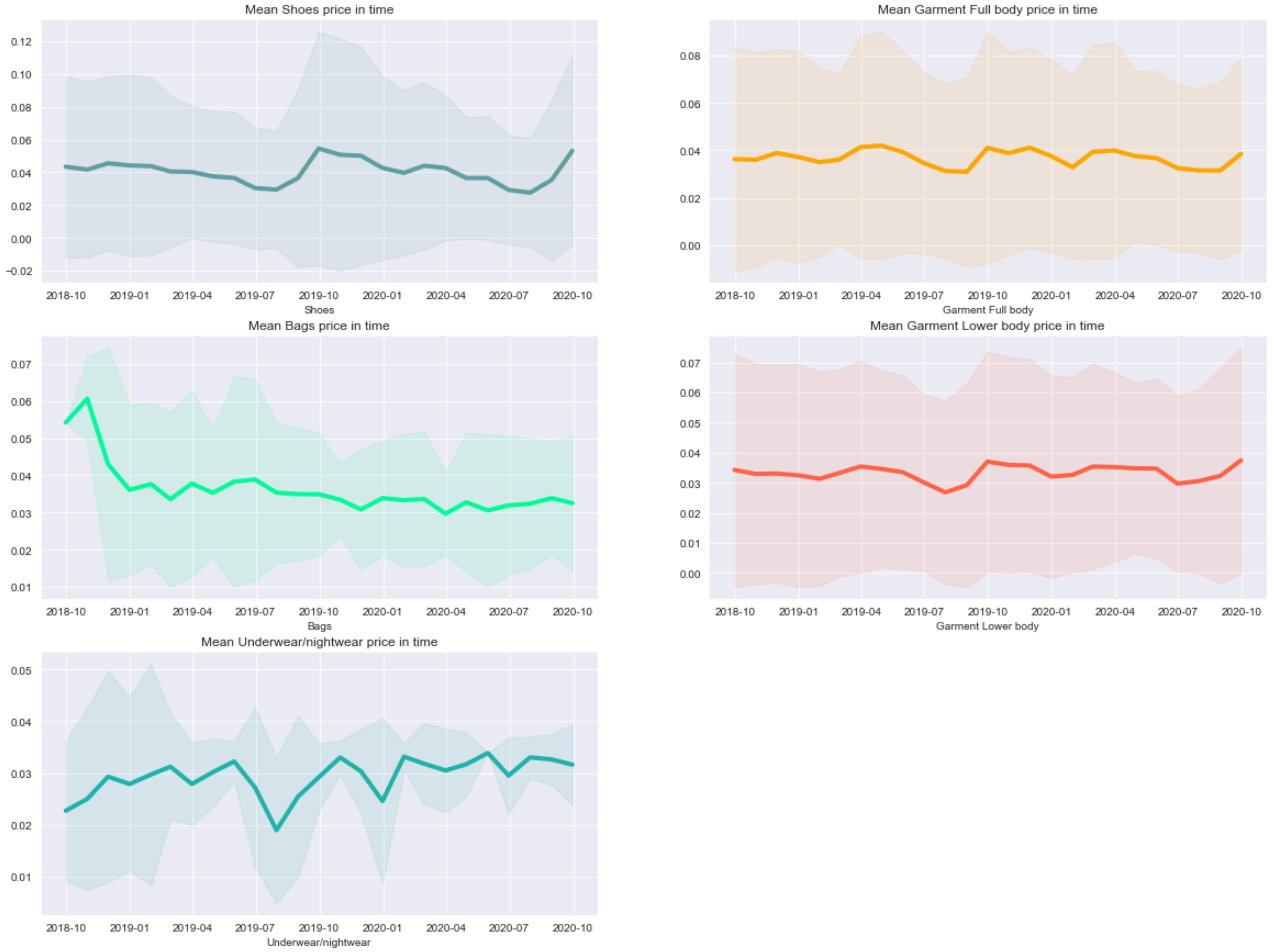


Figure 12. Mean price changes over time by product group.

3.4.2 IMAGE & ARTICLES DATA

Lastly, we explored merging the articles and image data in order to visualize the products based on some criteria. The first step was to merge the articles data by article identification to the individual images. Once this was accomplished, we could the images by product groups. Figure 13 displays some items from the *garment lower body* product group.

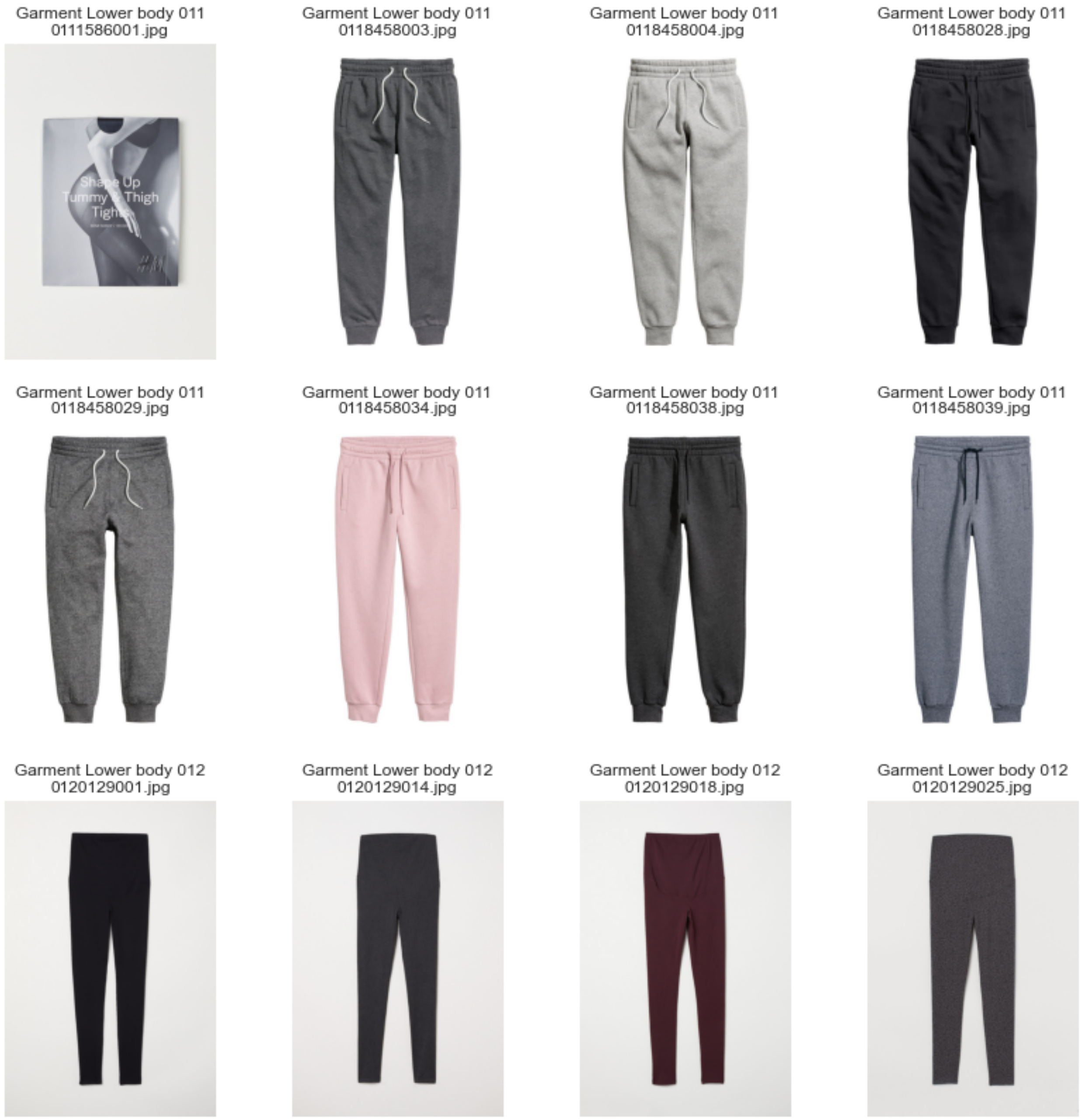


Figure 13. Images of articles in the *garment lower body* product group.

Figure 14 displays some items from the *cosmetics* product group.

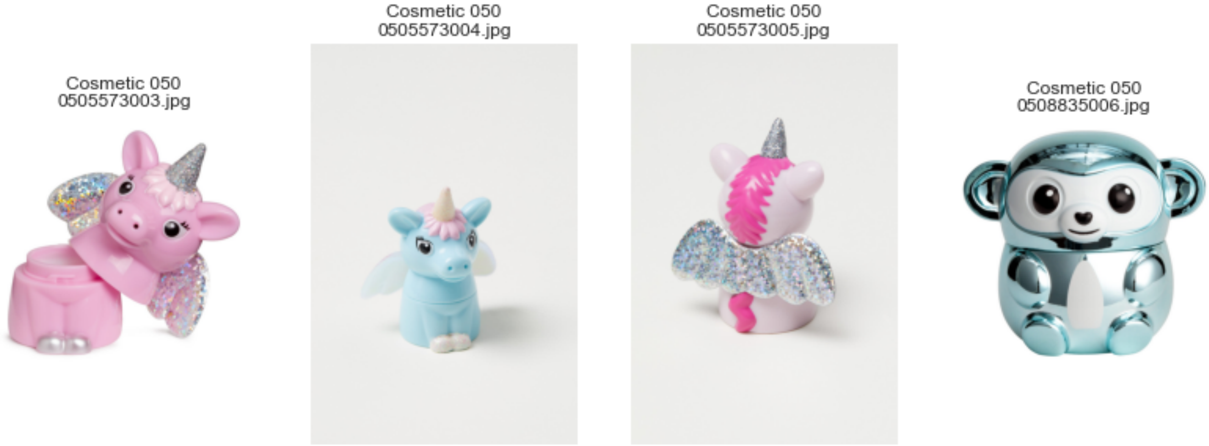


Figure 14. Images of articles in the *cosmetics* product group.

Next, we looked at the top five most expensive items and the last five lowest price items. Figure 15 displays the top five items and Figure 16 displays the bottom five items, as well as their price and item descriptions.

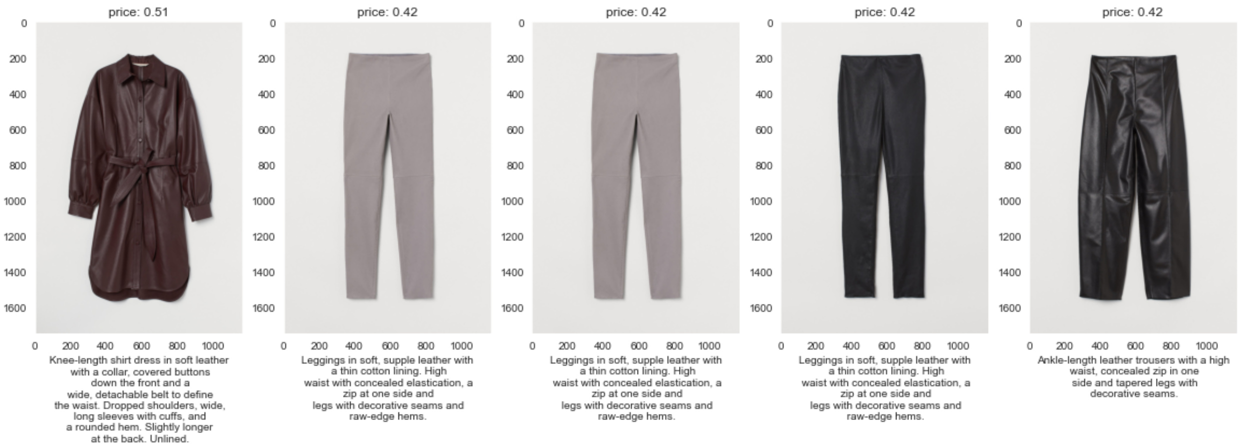

price: 0.51
price: 0.42
price: 0.42
price: 0.42
price: 0.42
0
200
400
600
800
1000
1200
1400
1600
0 200 400 600 800 1000
Knee-length shirt dress in soft leather with a collar, covered buttons down the front and a wide, detachable belt to define the waist. Dropped shoulders, wide, long sleeves with cuffs, and a rounded hem. Slightly longer at the back. Unlined.
Leggings in soft, supple leather with a thin cotton lining. High waist with concealed elastication, a zip at one side and legs with decorative seams and raw-edge hems.
Leggings in soft, supple leather with a thin cotton lining. High waist with concealed elastication, a zip at one side and legs with decorative seams and raw-edge hems.
Leggings in soft, supple leather with a thin cotton lining. High waist with concealed elastication, a zip at one side and legs with decorative seams and raw-edge hems.
Ankle-length leather trousers with a high waist, concealed zip in one side and tapered legs with decorative seams.

Figure 15. Images of articles in the *cosmetics* product group.

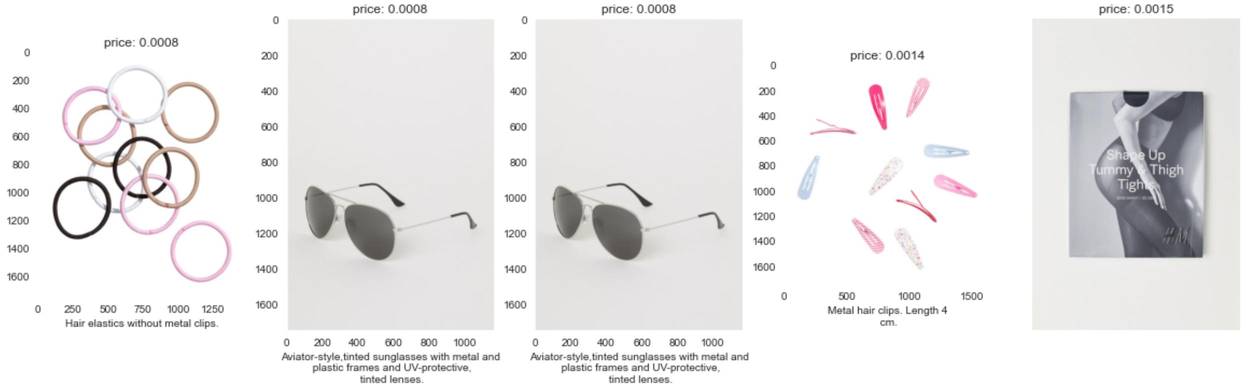

price: 0.0008
price: 0.0008
price: 0.0008
price: 0.0014
price: 0.0015
0
200
400
600
800
1000
1200
1400
1600
0 250 500 750 1000 1250
Hair elastics without metal clips.
0 200 400 600 800 1000
Aviator-style,tinted sunglasses with metal and plastic frames and UV-protective, tinted lenses.
Aviator-style,tinted sunglasses with metal and plastic frames and UV-protective, tinted lenses.
0 500 1000 1500
Metal hair clips. Length 4 cm.
Shape Up
Tummy & Thigh
Tights
H&M

Figure 16. Images of articles in the *cosmetics* product group.

4. PREDICTION MODELS

Two models were used to generate product recommendations, including a simple baseline model and a content-based filtering model. The details of these models are described in the following paragraphs.

4.1 SIMPLE BASELINE MODEL

The simple baseline model is based on the most frequent recently purchased items. A training set was created using the most frequent recently purchased items and then aggregated to compare to the sample submission file. Missing values were replaced with existing ones in the aggregated data. In total, 9,699 missing values were replaced. This model was set to generate 12 item predictions per customer.

4.2 CONTENT-BASED FILTERING MODEL

The content-based filtering model is based on historical information that an organization has on a particular individual. This model requires there be some background information on which to base the filtering or recommendations. Using this method, based on the previous items an individual purchased at H&M, we chose to approach the recommendations using customers who had a minimum of two and five transactions.

First, we created a new merge of the articles and transactions datasets so the historical data was available for the model. We then went through and selected a set of features to include within the model. Next, we had to select the number of minimum items previously bought for a particular customer. We then normalized the user features and dropped any potential duplicates within the article identifications. Then, we produced scores for the different articles that might be bought by someone that has bought a similar item previously. Lastly, we brought in the images of all the various articles to matched those previously bought and those recommended.

5. RESULTS

5.1 SIMPLE BASELINE RESULTS

The model generated 12 predictions per customer. The predictions were separated into 12 columns and the most frequent articles were counted per column. Two of the items were the same. The least frequent articles were also explored and there were numerous items with a single prediction. The top 11 most predicted items are displayed in Table 4, including the product name and short description.

Table 4. Top 11 Predicted items

Predicted Article	Product Name	Product Description
924243002	Ohlsson	Relaxed-fit sweater vest in a soft rib knit.
751471001	Pluto RW slacks	Ankle-length cigarette trousers in a stretch.
448509014	Perrie Slim Mom Denim	5-pocket, ankle-length jeans in washed, sturdy.
918522001	Jackie cable vest	V-neck slipover in a soft cable knit with ribbed.
866731001	LANA seamless HW tigths	Ankle-length sports tights in fast-drying functional fabric.
714790020	Mom Fit Ultra HW (trousers)	5-pocket, ankle-length jeans in washed, stretch.
788575004	Maja cargo Slim HW Denim (trousers)	Jeans in washed, stretch denim with a high waist.
915529005	Liliana	Jumper in a soft, fine knit.
573085028	Madison skinny HW	5-pocket jeans in washed stretch denim.
918292001	STRONG HW seamless tights	Sports tights in fast-drying functional fabric.
850917001	Sadie Shirt	Gently fitted shirt in a stretch weave.

5.2 CONTENT-BASED RESULTS

The model produced recommendations based on previously bought items for specific customers. The results provided six recommendations based on both two and five previous transactions by a customer. The recommendations fall within similar types of articles as well as similar colors. Figure 17 displays the previously purchased articles for a specific customer with a minimum of two transactions. Figure 18 displays the recommendations for the same customer. Figure 19 displays the previously purchased articles for another specific customer with a minimum of five transactions. Figure 20 displays the recommendations for the same customer.

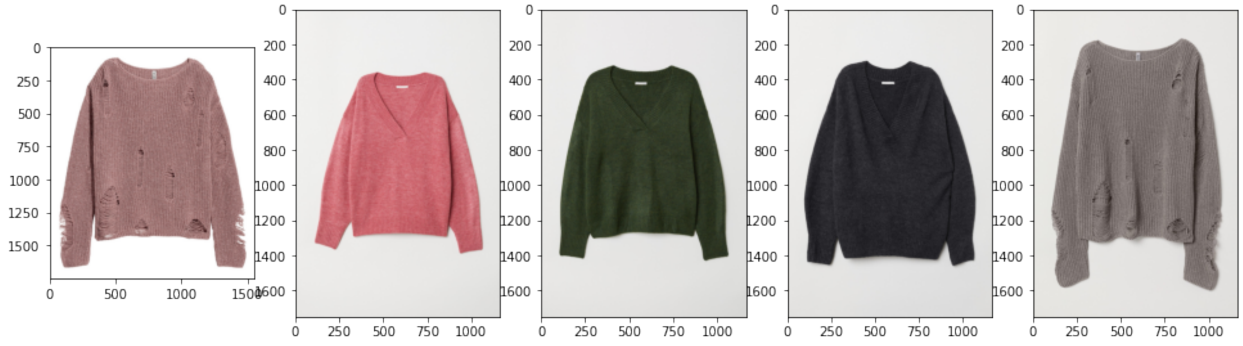


Figure 17. Previously purchased articles (minimum of two purchases).

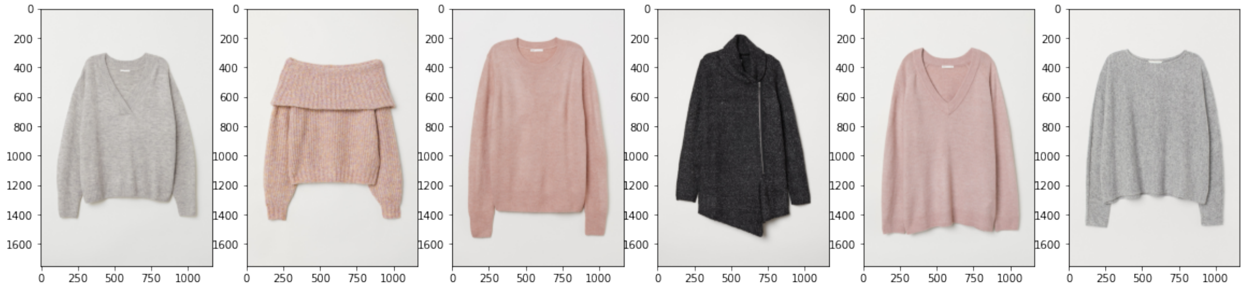


Figure 18. Recommended articles based on minimum of two purchases.

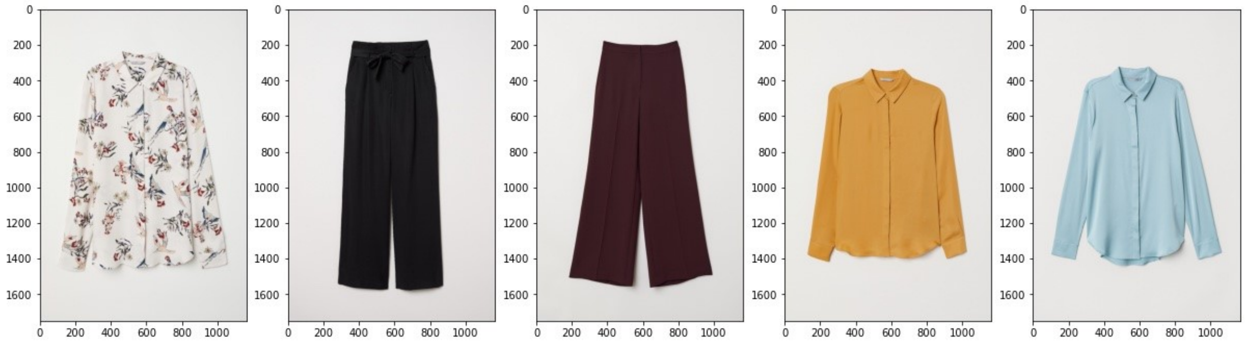


Figure 19. Previously purchased articles (minimum of five purchases).

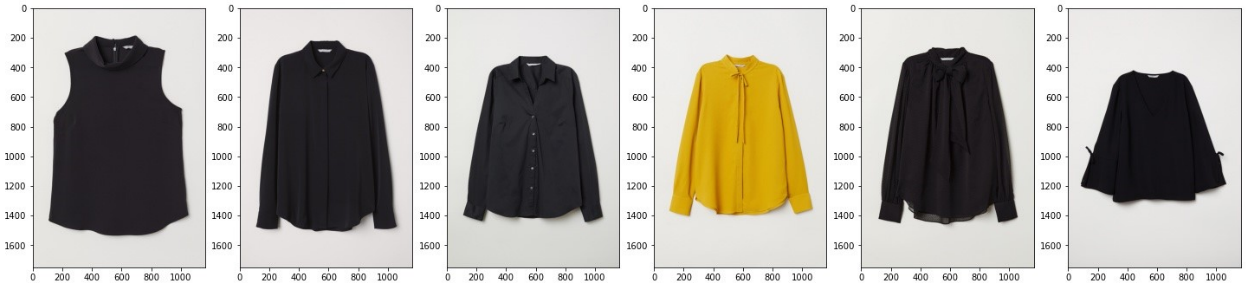


Figure 20. Recommended articles based on minimum of five purchases.

6. CONCLUSIONS

Using the findings from the predictive models developed, H&M has access to data to inform its decision-making surrounding future customer purchase behavior. The company can utilize this information to optimize inventory and marketing strategy, which would increase business efficiency and reduce its carbon footprint.

The simple baseline model identified the most likely future purchases given the aggregated purchase data among all customers, which could be used to guide marketing and ad campaign efforts. The model used the most frequent recently purchased products to predict future purchases. The content-based filtering model is the recommended method for identifying purchase predictions at the individual customer level. This model used similarities in product item features to generate recommendations for other items that the customer may be interested in.

H&M group should continue to incentivize customers with loyalty programs that offer exclusive discounts to loyal customers; the company should also explore new ways of attracting loyalty. Exploring the data revealed that H&M has two distinct age groups that the company could use different marketing strategies with. Older customers may respond more favorably to printed ad materials, while digital media may be the way to target the younger customer base. Increased loyalty would have the compound effect of sales and additional customer data for feeding future models to give H&M Group a competitive advantage in the fashion market.

REFERENCES

Google Recommendation Systems. 2021, Feb 5. *Content-based Filtering*. https://developers.google.com/machine-learning/recommendation/content-based/basics#:~:text=Content%2Dbased%20filtering%20uses%20item,previous%20actions%20or%20explicit%20feedback.

H&M Hennes & Mauritz GBC AB. 2022, Feb 7. *H&M Personalized Fashion Recommendations.* Kaggle. https://www.kaggle.com/c/h-and-m-personalized-fashion-recommendations

Karpov, Daniil. 2022, Feb 28. *H&M EDA FIRST LOOK*. Kaggle. https://www.kaggle.com/code/vanguarde/h-m-eda-first-look

Obeidat, Mohammed. 2022, Mar 16. *Content-Based Filtering with PCA*. https://www.kaggle.com/code/mohammedobeidat/content-based-filtering-with-pca/notebook

Preda, Gabriel. 2022, Feb 9. *H&M EDA and Prediction.* Kaggle. https://www.kaggle.com/code/gpreda/h-m-eda-and-prediction/comments