CSE 4/587 Project Phase 1

Fashion Recommender System

Problem statement:

Our main objective in this project is to design and develop a recommender system that uses machine learning and advanced analytics techniques to solve many prediction use cases used in the fashion industry to increase sales. The customers are targeted using this recommender system to promote the products that are likely to be purchased by the customer. Our system should be able to predict future possible purchases of a customer and also to predict the products that are more likely to be purchased together. This is mainly useful for e-commerce platforms to potentially increase sales.

1. Articles Data (articles.csv):

This dataset contains information about the available articles of clothing and accessories on H&M’s e-commerce website. It includes columns such as article id, product description, categories, prices, and most importantly all the article images linked to this data.

1. Customers Data (customers.csv):

The customer dataset provides information about H&M registered customers. It contains columns with the following data customer id, activity, membership, age, postal region, etc. This data set is primarily used to get customer segmentation. We can get an idea of what sort of customers are participating in the shopping and their purchase behavior.

1. Transaction Data (transactions\_train.csv):

This dataset records all the transactions that have taken place on the online platform of H & M. It contains the time-series data of all the purchases relating them to the customers and articles data. It contains the following columns of data timestamp, customer\_id, article\_id, price, and sale channel id. The transaction data helps us understand the history of customers' purchases and also gives us insight into the seasonal purchase of the items.

1. Sample Submission Data (sample\_submission.csv):

This dataset is used to test our model performance in some scenarios such as predicting what customers intend to buy in the future using the past data of the transactions. This contains the customer id and what items we predict that be purchased in the future by the customer.

In this project, our main idea is to create a recommendation system to predict what customers would tend to buy in the next fixed period considering seasonal events, etc, based on the analysis. We are trying to create a model that recommends a combination of the products that customers tend to get recommendations for when they purchase one product from the website. We further follow the following steps of implementation:

1. Data Collection: We gathered the article’s data, customers’ data, transaction data, and sample submission files from the Kaggle website.
2. Data Cleaning / Processing: Based on the data we try to clean the data using multiple standard cleaning techniques which further help to increase the efficiency of the model we build in the further phases. This involves multiple tasks like handling duplicate rows, missing data, and one hot encoding of the categorical data and normalizing the data.
3. Feature Extraction: Extract features from the available data for exploratory data analysis. This gives us an insight into comparing data output with each feature and helps us in deciding the techniques that help in applying further.
4. Model Building: This happens further in the next phase where with the existing processed data we obtained in the first two steps we apply advanced statistical techniques and machine learning algorithms to predict the output we need. We have the scope to use clustering algorithms to cluster the data, we can use the cosine similarity technique to cluster together similar products. So, we are exploring many such techniques that are compatible with the dataset and output that we want to achieve.
5. Evaluation: We use standard metrics to evaluate the models that we built to compare the models that we build based on this data.
6. Deployment: Using MLOps we can deploy the model

Data Cleaning / Processing:

We have three datasets that we need to clean and process using pandas and Numpy. To do that we are using famous data-cleaning techniques to make our dataset more efficient for building models. Firstly, we need to understand the data available in the CSV files. Article data has to be loaded using pandas into the data frame. Using the shape command we understand the dimension of the data available which is (105542, 25) along with datatypes of elements in each column. We use the unique function to get the unique elements present in the column.

We apply the drop duplicate function to the entire dataset to drop any duplicate rows in the data. When we implement a model to build the model we would like to use the description of each article for clustering and other purposes so we remove unnecessary spaces in the elements in the description column. Then convert the string into list words which we use as tags in the future.

Now we also find that there are a lot of columns in the article data which is redundant and couldn’t be used for any sort of model building. We combine two columns into a single column and create a new column while dropping the two previous columns. We are using this technique on eight columns which are garment\_group\_name, garment\_group\_no, section\_name, section\_no, index\_group\_name, index\_group\_no, index\_name, and index\_code We are creating new columns garment\_group, section, index\_group and index.

We can observe that the product\_code is the part of article\_id so the first six digits of article\_id represents product\_code. So, we are dropping that column too. Now we could target customers.csv file where the data contains customers data. This data contains seven columns and 1371980 customer data as rows. Here we applied a new column to segregate each customer into different age\_group segments to update that data into a new column known as age\_group. We try to understand how many values in the total are present in the fashion\_news\_frequency column. We we run the unique pandas command on the entire column we get five distinct values.

We tried removing the duplicates in the customer data frame. The columns FN and Active contain NaN values which we replaced with 0 for convenience in the upcoming stages. This is because we have only two unique values in that entire column. We are categorizing the fashion\_news\_frequency so we are replacing NaN values with None to ignore that data in future models. As this might be incomplete data.

Finally, we are loading transactions\_train.csv which contains time series data of transaction data which contains sale data of the articles for a period of time. This contains five columns and 31788324 rows. On this data, we dropped duplicate rows. We checked whether there were any NaN values in the entire data. For further analysis, I separated the month from the timestamp for future monthly analysis.

Exploratory Data Analysis:

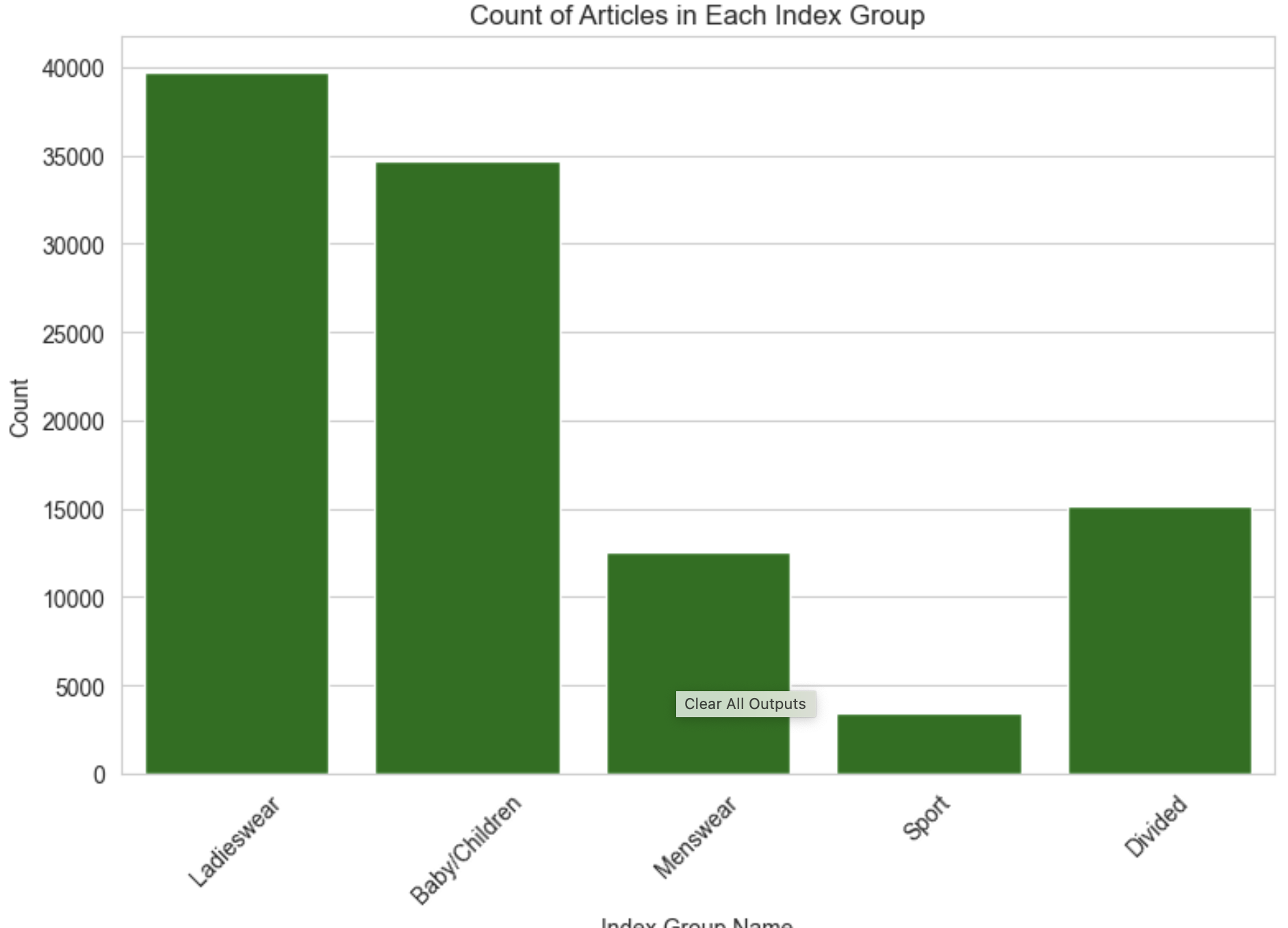


Fig 1.

The index\_group\_name is a categorical variable representing different categories of products. Simultaneously, the index\_group\_no appears to be a numerical encoding or label encoding of these product categories. In other words, each unique index\_group\_name corresponds to a specific numerical value in index\_group\_no.

Upon analysis, it is evident that the most prevalent products fall under the "Ladieswear" category, indicating that this category has the highest count of products. On the other hand, the "Sportswear" category has the lowest count of products, signifying its scarcity compared to other categories.

This observation highlights the distribution of products across various categories, with "Ladieswear" being the most prominent and "Sportswear" being the least represented in the dataset. This information can be valuable for understanding the product portfolio's composition and the relative popularity of different product categories within the dataset

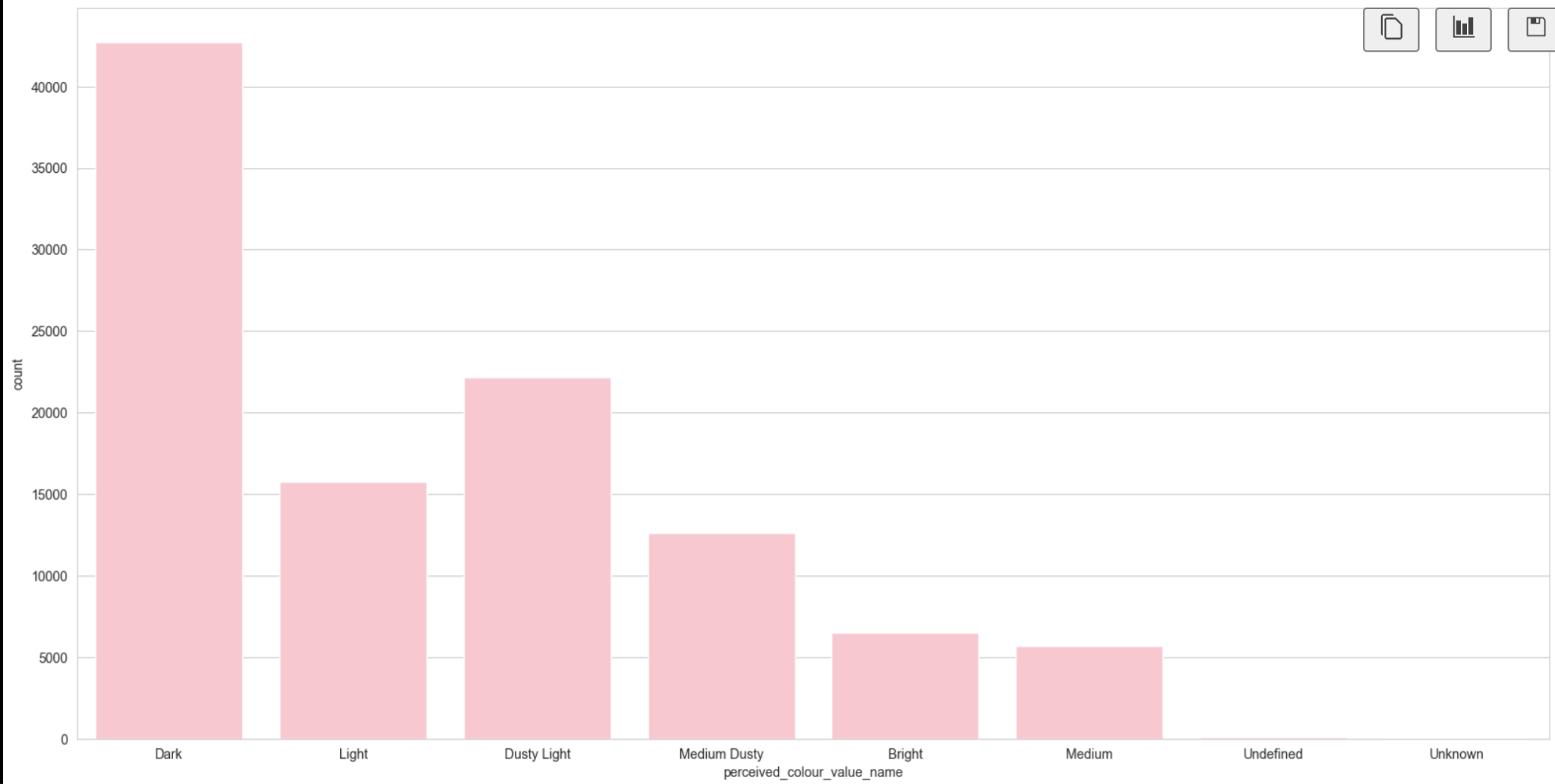


Fig 2.

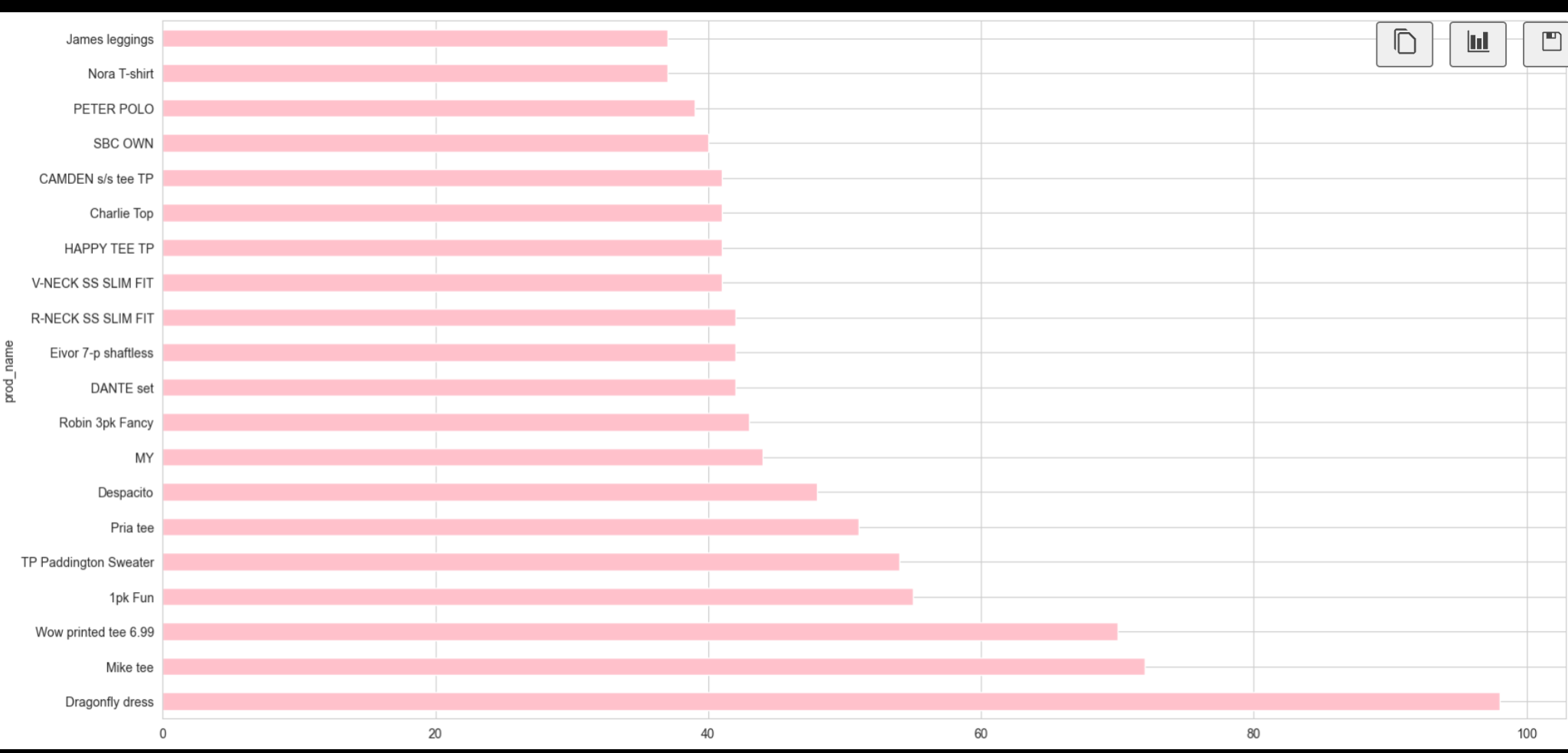


Fig 3.

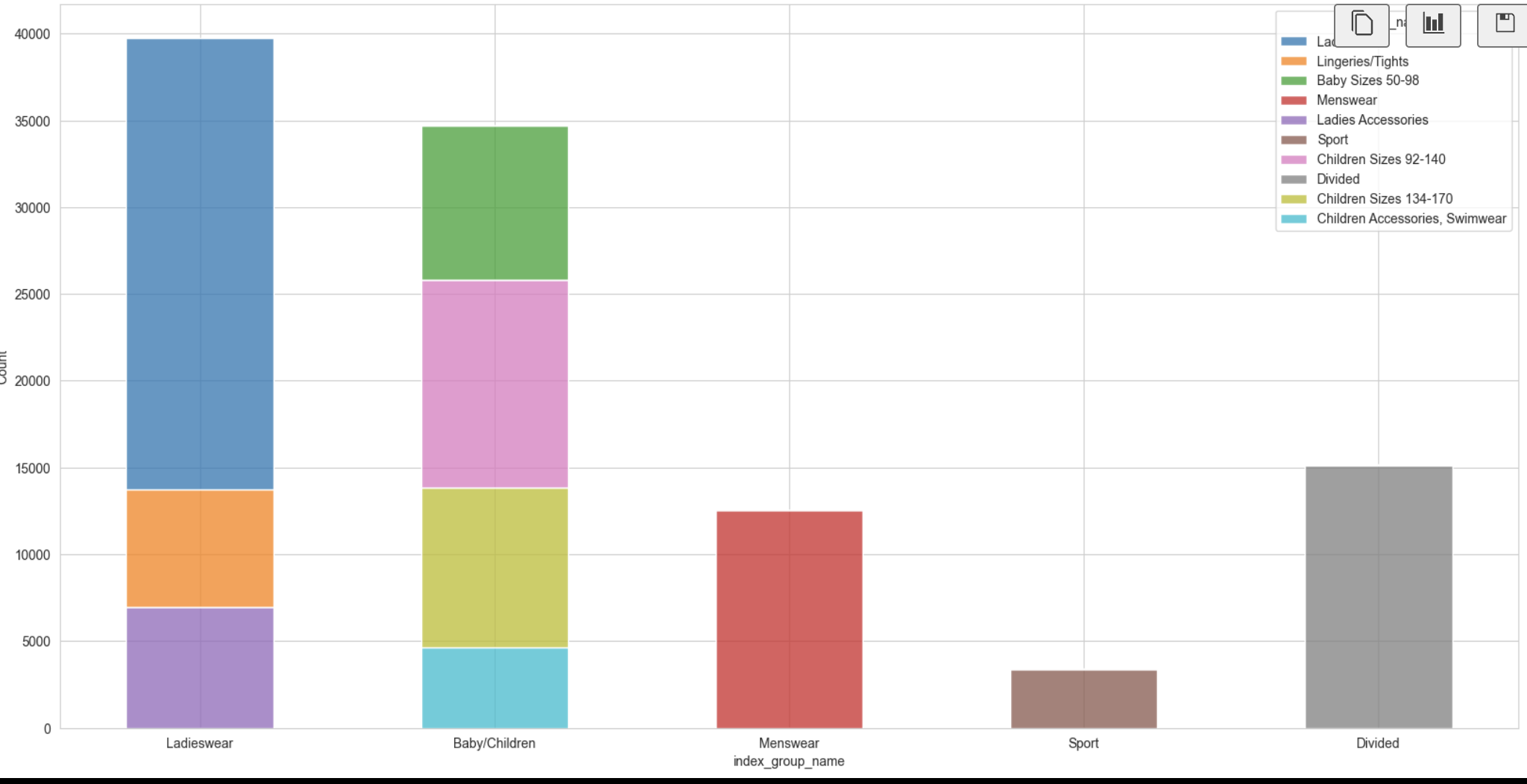


Fig 4.

The product category "Ladieswear" is notably diversified, comprising three distinct sections or subcategories. In contrast, the "Baby/Children" category exhibits an even greater level of subdivision, containing four distinct sections or subcategories within it. In these cases, the category label is not a monolithic representation but is further detailed into specific sections, each potentially representing a different set of products or product types.

On the other hand, the remaining product categories outside "Ladieswear" and "Baby/Children" are characterized by a more straightforward structure, with each of them having just one section. These categories are not further subdivided but are instead represented as a single, unified entity within their respective categories.

This insight provides a clearer understanding of the hierarchical organization of product categories within the dataset, with some categories being more granular and segmented, while others maintain a singular, undivided structure.

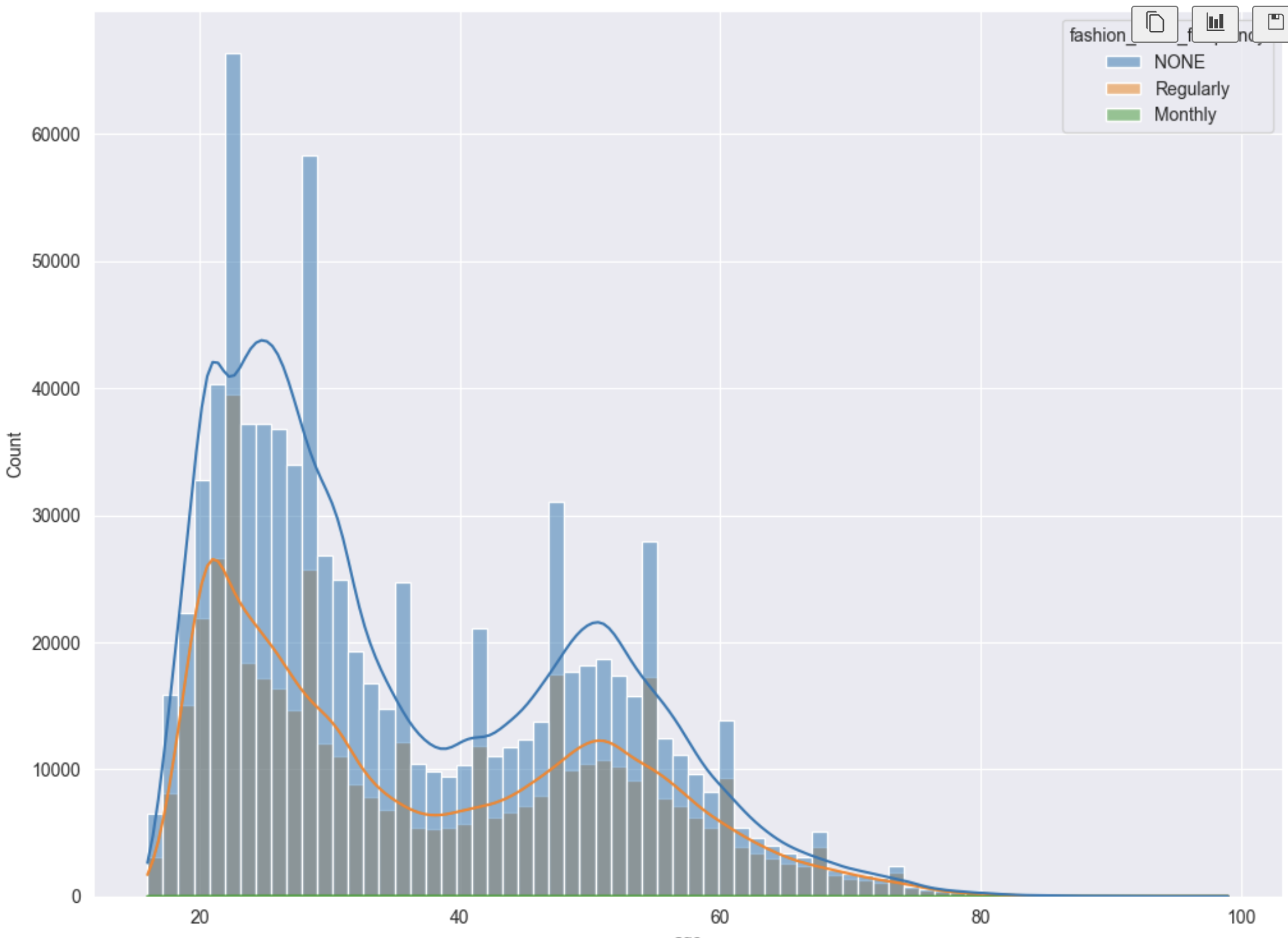


Fig 5.

The age distribution of customers reveals two notable peaks, one in the age group of 20 to 30 years and another in the age range of 45 to 60 years. These specific age groups are also coincidentally associated with a higher level of engagement with fashion news, as reflected by the update frequency.

However, it's worth noting that the age groups corresponding to "Monthly" and "None" fashion news update frequencies exhibit relatively lower counts. These groups are visually less prominent in the graphical representation due to their lower frequency, making them less visible in the graph.

In summary, the data indicates that certain age groups, specifically those between 20-30 and 45-60 years, are not only well-represented but also display a stronger connection to fashion news updates, while other age groups, such as those updating "Monthly" or "None," have comparatively fewer individuals

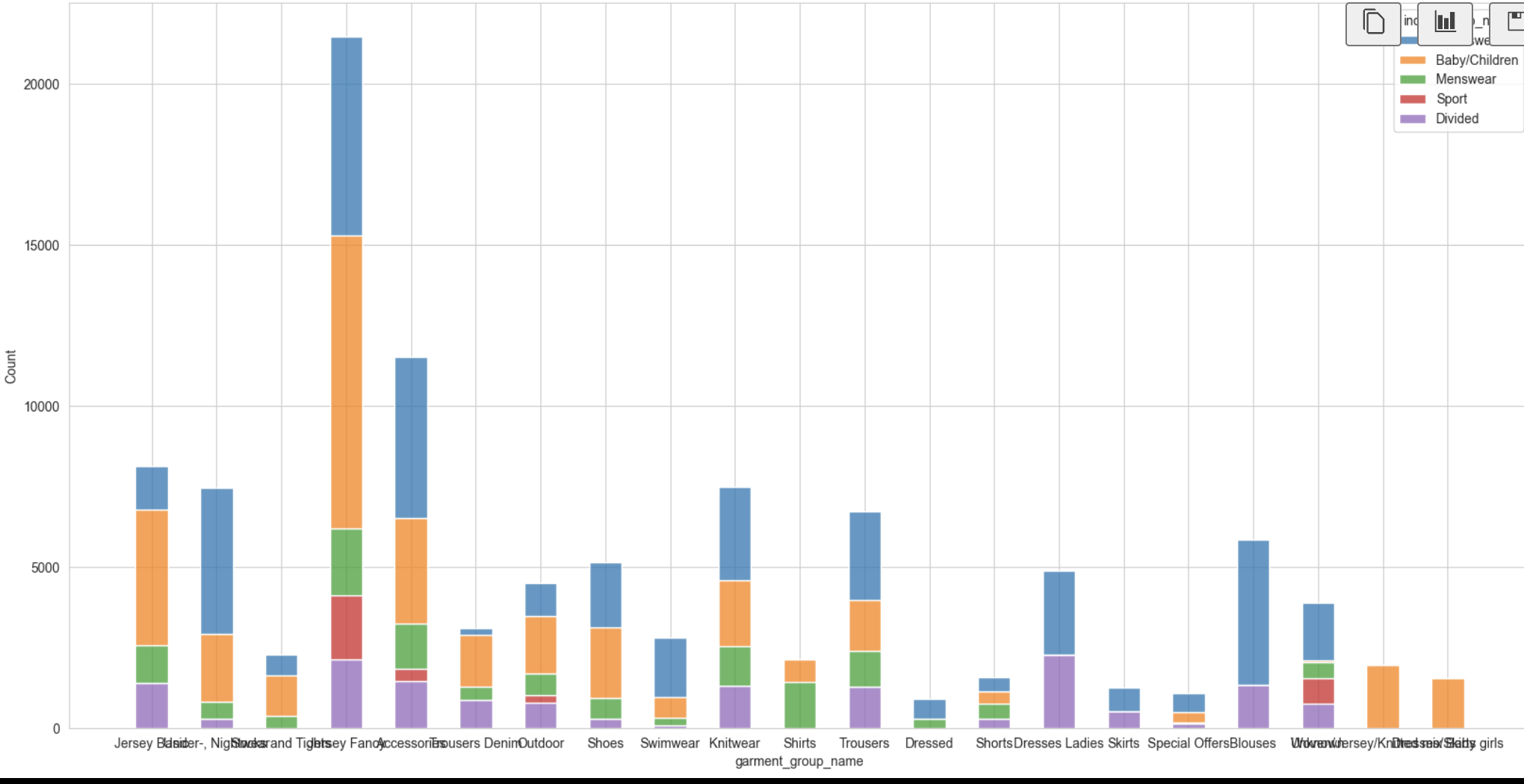


Fig 6.

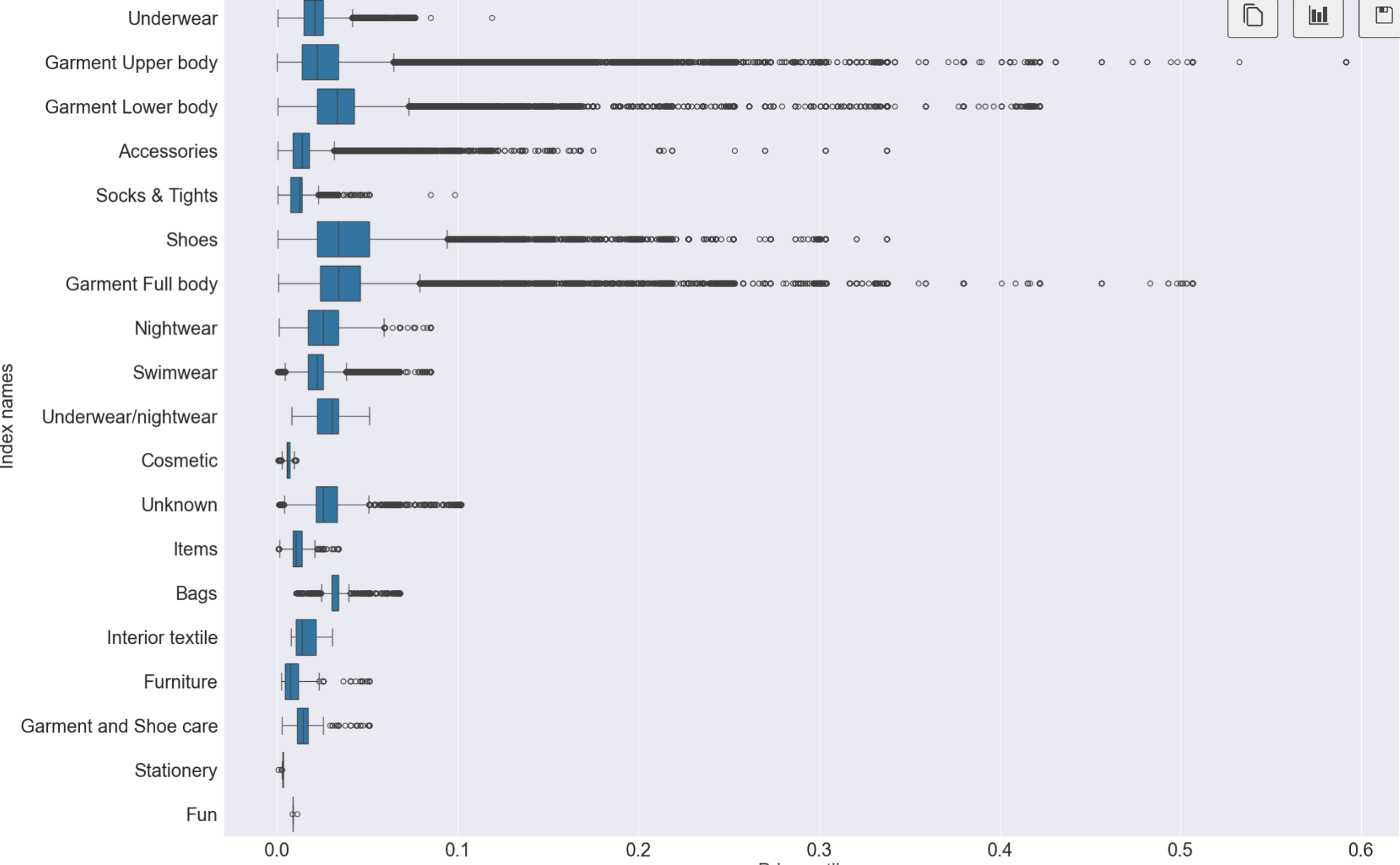


Fig 7.

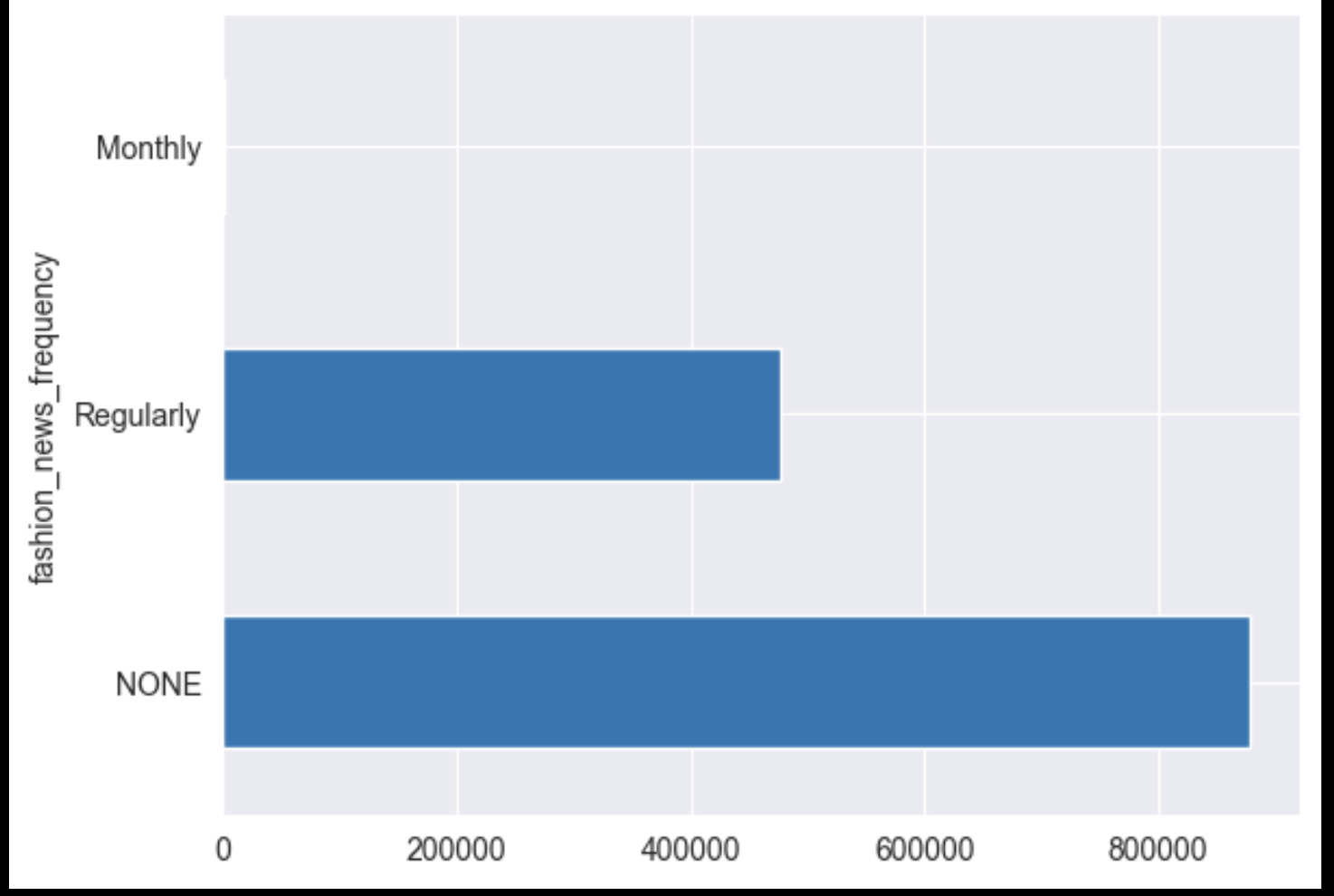


Fig 10.

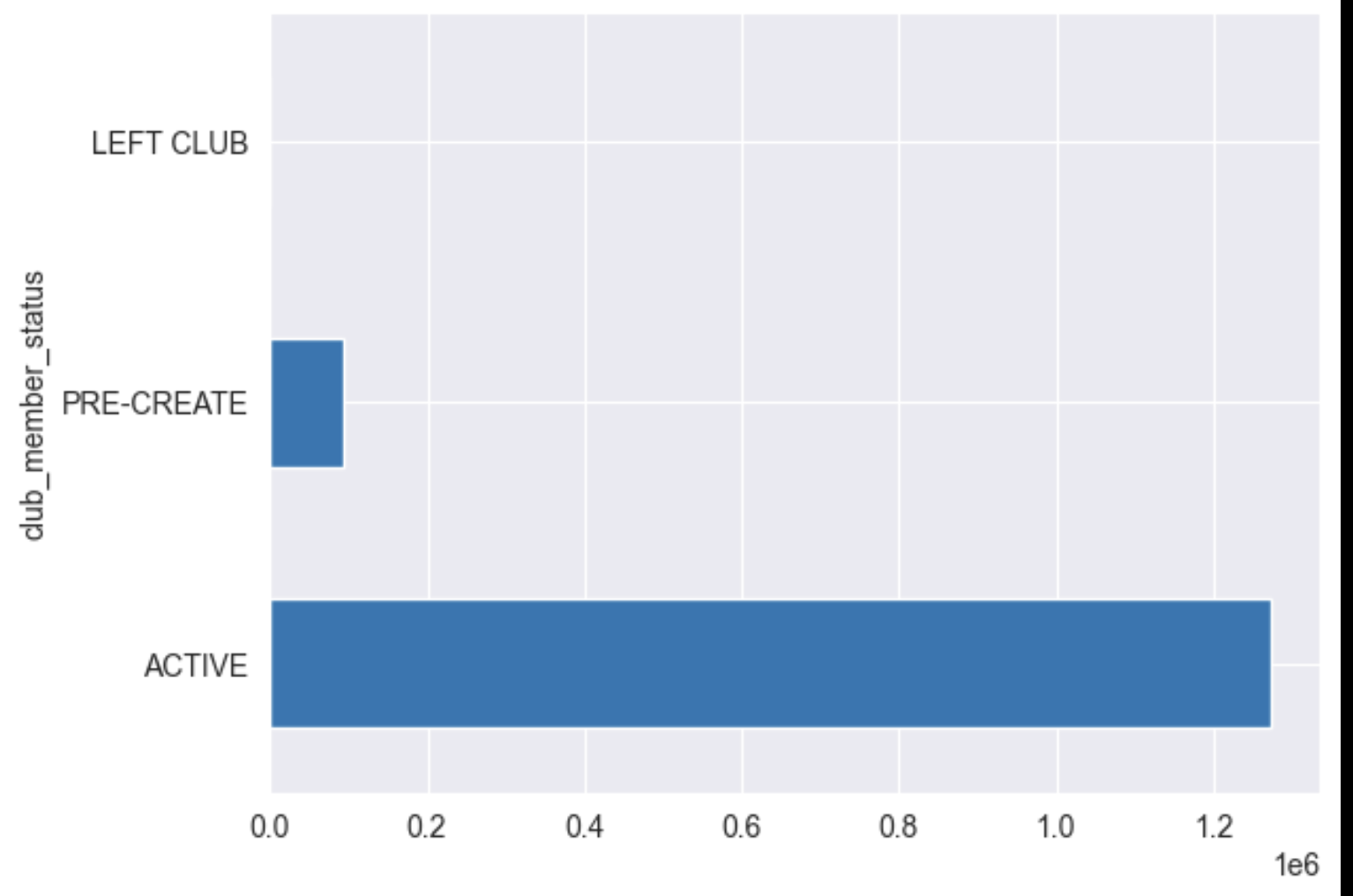


Fig 8.

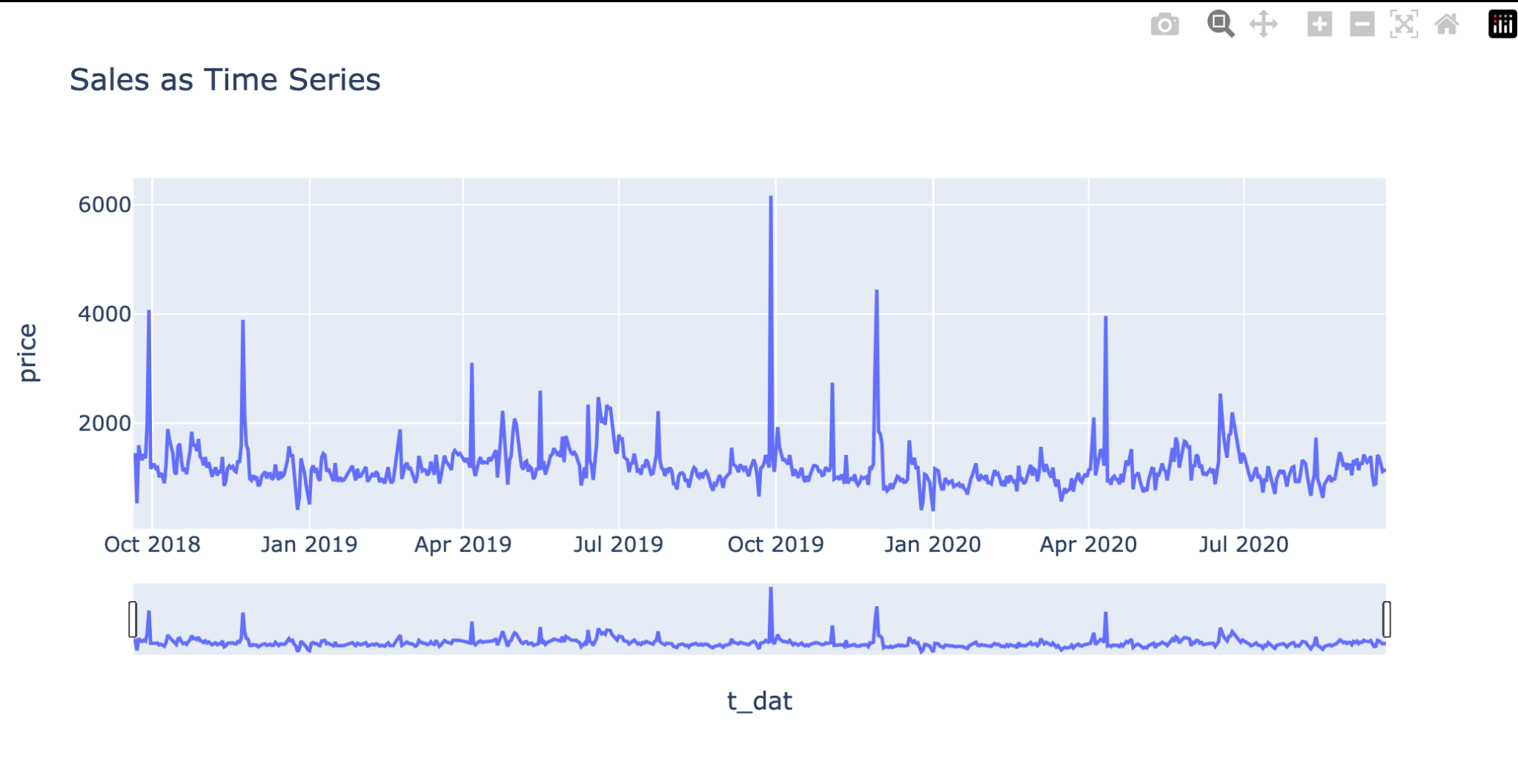


Fig 9.

Distinct and substantial peaks are evident in the data for April, September, and November. These specific months stand out with notably higher occurrences or significant events compared to the rest of the months.

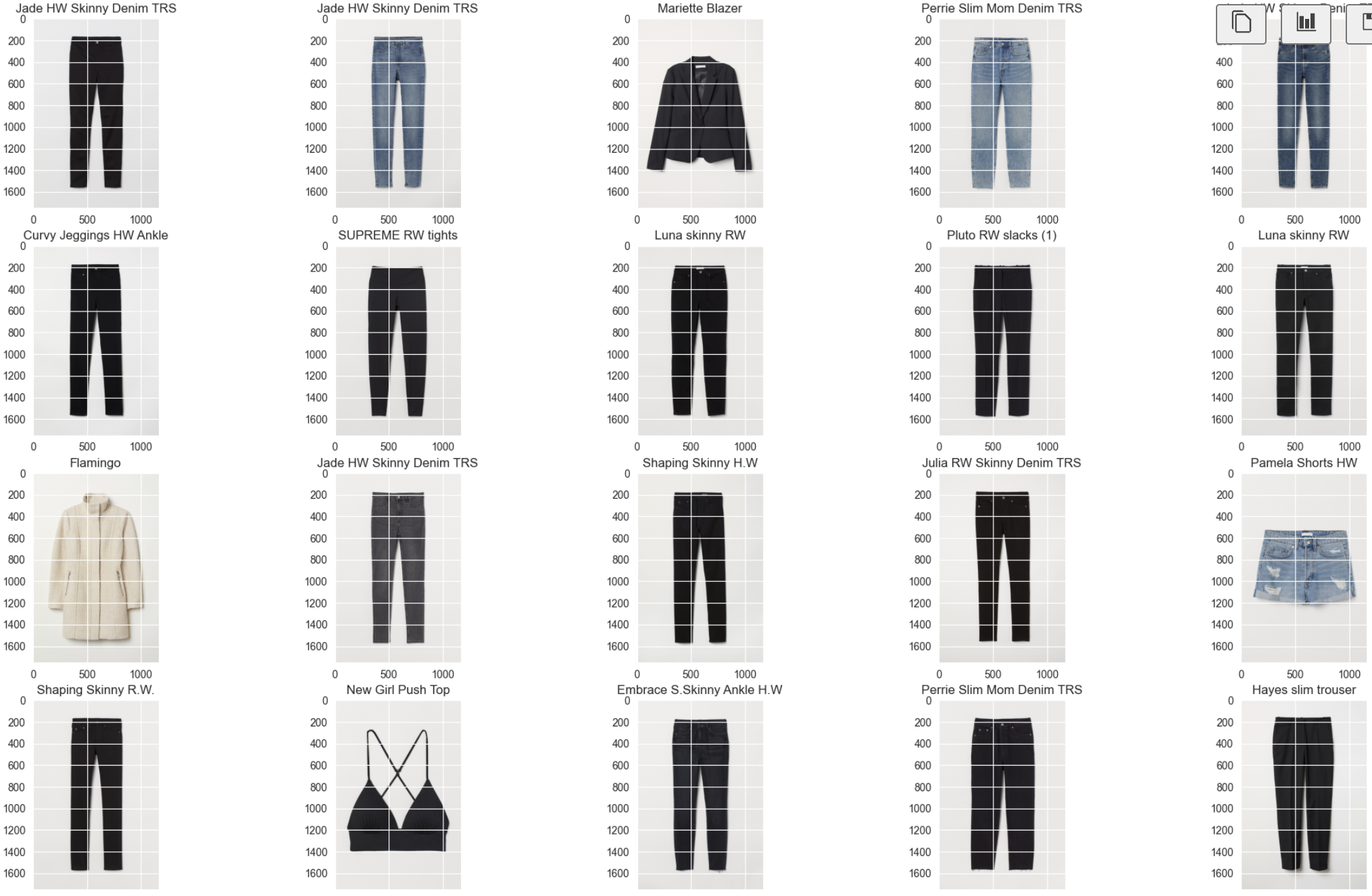


Fig 11.

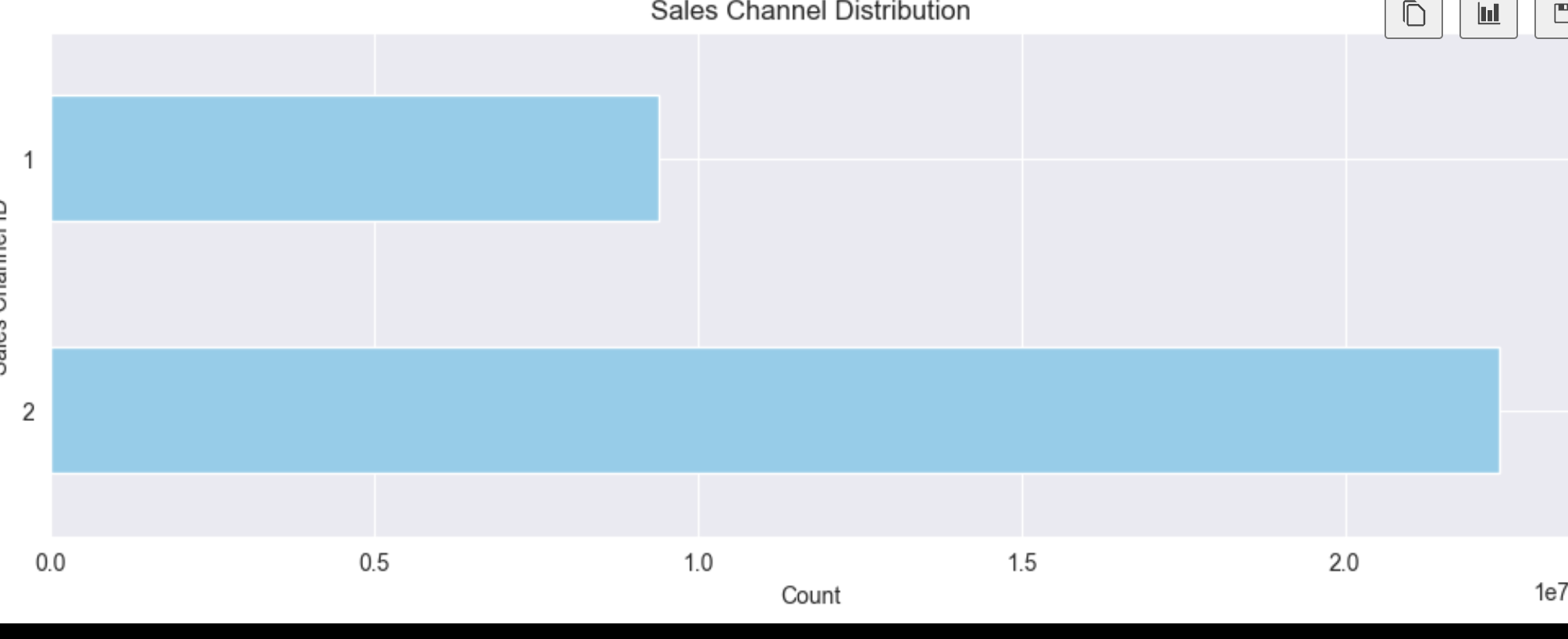
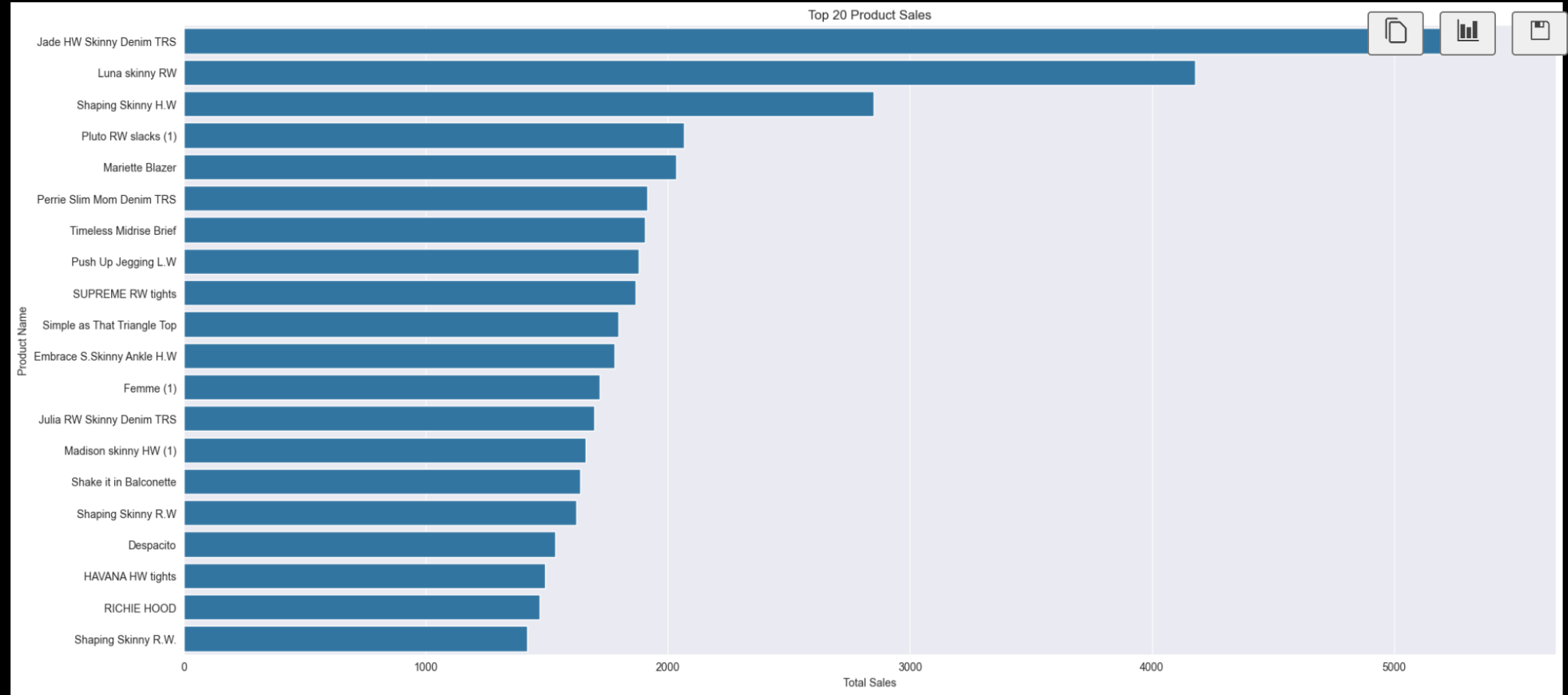
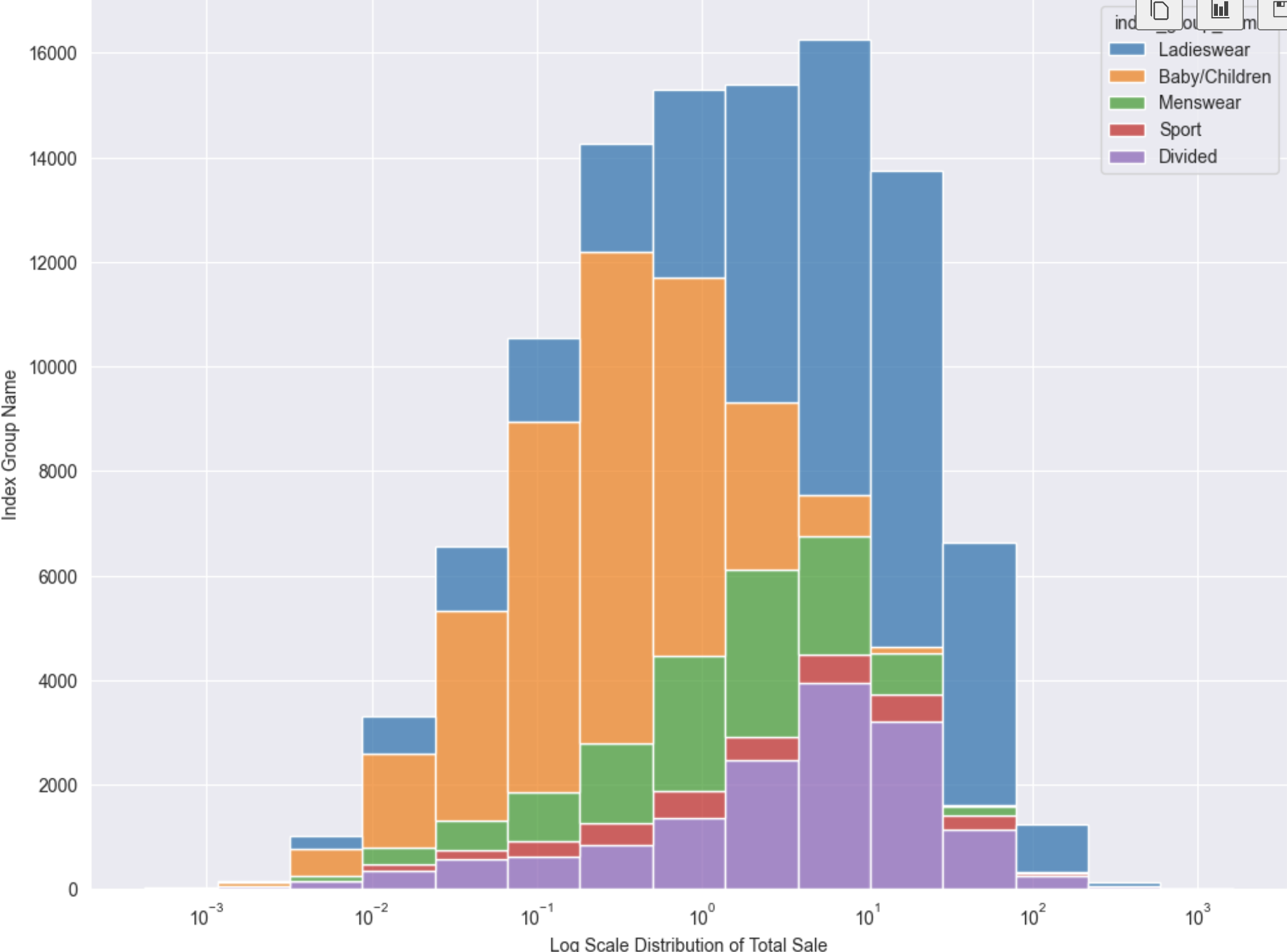


Fig 12.



## Fig 13. Top 20 Product Sale



## Fig. 14 Log Scale Distribution of Total Sales

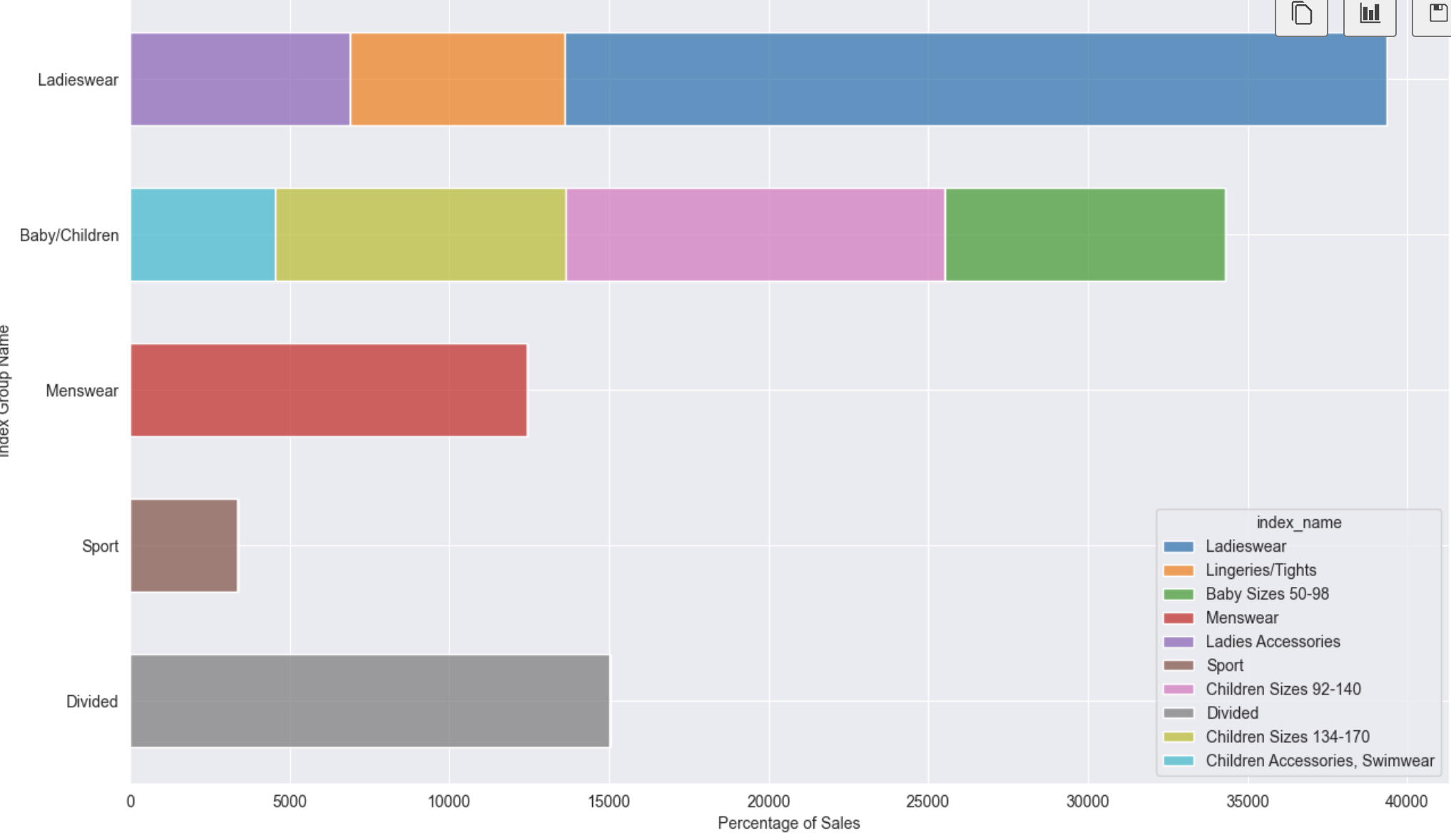


Fig 15.

It's evident that "Ladieswear" consistently maintains a strong sales performance, which can be expected given its broader appeal and wide-ranging offerings. Moreover, within the "Children" section, it's notable that sizes 92-140 tend to drive

higher sales, indicating a specific range within this category that garners more consumer attention and demand

References:

<https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data>

<https://youtu.be/l0KNPQn5EmE>

<https://www.youtube.com/watch?v=l0KNPQn5EmE&ab_channel=ChaiTimeDataScience>

<https://www2.hm.com/en_us/index.html>