The Product Recommendation for H&M

Course: BIA 679

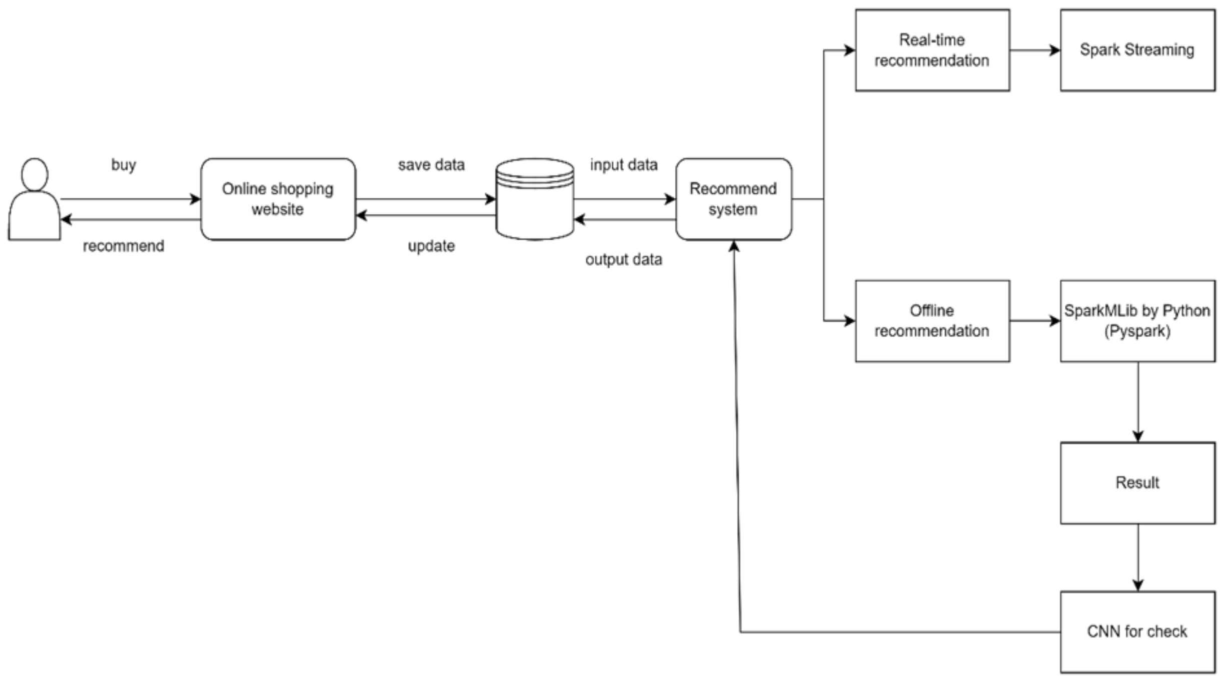
Member: Tengyue Chen, Hao Miao, Haoxing Zhang

Introduction

Online shopping has become one of the shopping channels for consumers with the development of the Internet. Total e-commerce sales in 2021 are expected to be $870.8 billion and grow 14.2% from 2020 (Young, 2022) in the U.S. Many online shopping companies have grown into well-known global companies, such as Amazon, Ebay, and Alibaba. However, with the increase in the number of items on online shopping platforms or websites. However, because the number of products on online shopping platforms continues to increase, the users may not choose their favorite products.

Recommend systems provide users with product information to help users make decisions and complete online purchases. The E-commerce recommendation system builds models that reflect user attributes and behaviors through the collected user information (Zhao, 2019). Online shopping platforms could use the E-commerce recommendation model in the backend to help users quickly find their favorite products.

This project attempts to build a model based on the dataset of products and user behavior provided by H&M. This model could predict the user's potential product selection and provide purchase suggestions through the user's previous purchase behavior or habits. Because real-time data cannot be obtained, the model of this project will focus on offline recommendation. The figure 1 shows the design of the recommendation model system.



Figue1. Model Design

Data collection

The dataset is collected from the Kaggle(https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data). The dataset contains the purchase history of H&M customers in the online store. The dataset contains three files, includes product pictures, product purchase records, and user information. Table 1, table2, and table 3 show the dataset columns.

Columns	Feature
article_id	The article id
product_code	The code of product
prod_name	The name of product
product_type_no	The number of product type

product_type_name	The name of product name
product_group_name	The group name of product
graphical_appearance_no	The number of graphical appearances
graphical_appearance_name	The name of graphical appearances
colour_group_name	The name of color group
perceived_colour_value_id	The value id for perceived color
perceived_colour_master_id	The master id for perceived color
perceived_colour_master_name	The master’s name for perceived color
department_no	The number of departments
department_name	The name of departments
index_code	The index code
index_name	The name of index
index_group_no	The number of index group
index_group_name	The name of index group
section_no	The number of sections
section_name	The name of section

garment_group_no	The number of garment group
garment_group_name	The name of garment group
detail_desc	The detail describes

Table 1. Columns of articles

Columns	Feature
customer_id	The customer id
FN	
Active	Active or not active
club_member_status	The status of club member
fashion_news_frequency	The frequency of fashion news
age	The customer age
postal_code	The customer’s postal code

Table 2. Columns of customer

Columns	Feature
t_dat	The date about the transactions
Customer_id	The customer id
Article_id	The article id
price	The price about the transactions
Sales_channel_id	The sales channel id

Table 3. Columns of transactions

Project Timeline

The project will be completed through 8 milestones. The Figure 2 show that the timeline about this project. In the week 1, the project's goal is established. In the second week, suitable datasets were searched. After, we clean the data in the third week. In weeks 4 to 5, we analyze the obtained dataset. We plan to present our results in week 6. In week 7 we review the project and check the error. Week 8, we finish the whole project and summarize it. The table 4 show the detail about the project timeline.

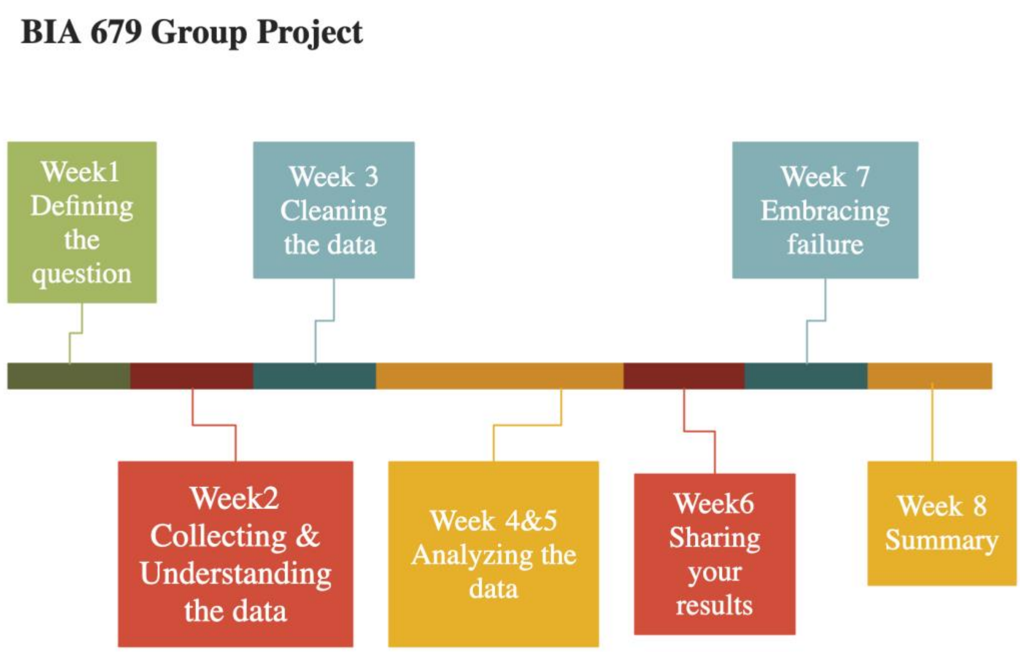


Figure 2. Project Timeline

Week	Detail
1	The group discuss about the topic and goal about the project.
2	A dataset is found or collected basic on the project. The group try to understand the dataset.
3	The group clean the dataset. The EDA is finished by the following step: 1. Statistics on product filter by time, amount, etc 2. Statistics on customer filter by id, age, etc 3. Transaction amount by date 4. Correlation between customer and product 5. Word Cloud
4	The group start to analyze the data. The Apache Spark is installed The group try to learn the Spark. Using the PySpark to build an ALS model.

5	Continued to build the model. Training and test the data. Output a result basic on the model
6	Using the result to output a recommendation dataset. Share the result in the class. Review the whole project and finished the project.
7	Check the mistake in the project. Correct any error or mistake in the project
8	Finish the whole project. Write the final white paper for the project

Table 4. Plan about the project

Data exploration and EDA analysis

The project uses the Python to analysis the dataset. The code could be found in our GitHub repository (https://github.com/tychen17/The-Product-Recommendation-for-H-M) . First, we check the dataset. The dataset includes 3 CSV files. The 3 CSV files do not have null values. Then, we check the outlier about datasets. The Figure3, Figure 4 and Figure 5 show the box plot about the dataset. Because the column about id has unique value, the distribution of the box plot about the column of id cannot indicate the existence of outliers. However, Age in the transactions may have outliers. Therefore, we checked the distribution about age. The results (Figure 6) show that the reason for the abnormal value is that some customers did not fill in the age. The distribution of age is acceptable.

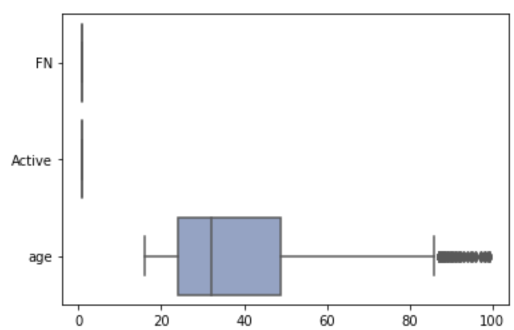


Figure 3. box plot of customers CSV

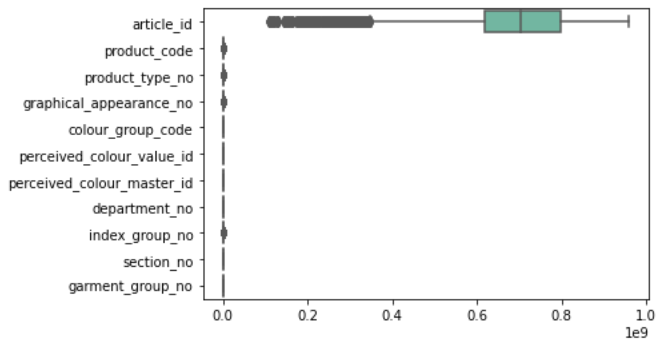


Figure 4. Box plot of articles CSV

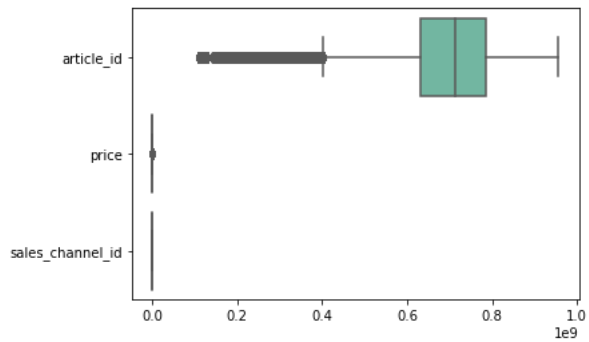


Figure 5. Box Plot of transaction CSV

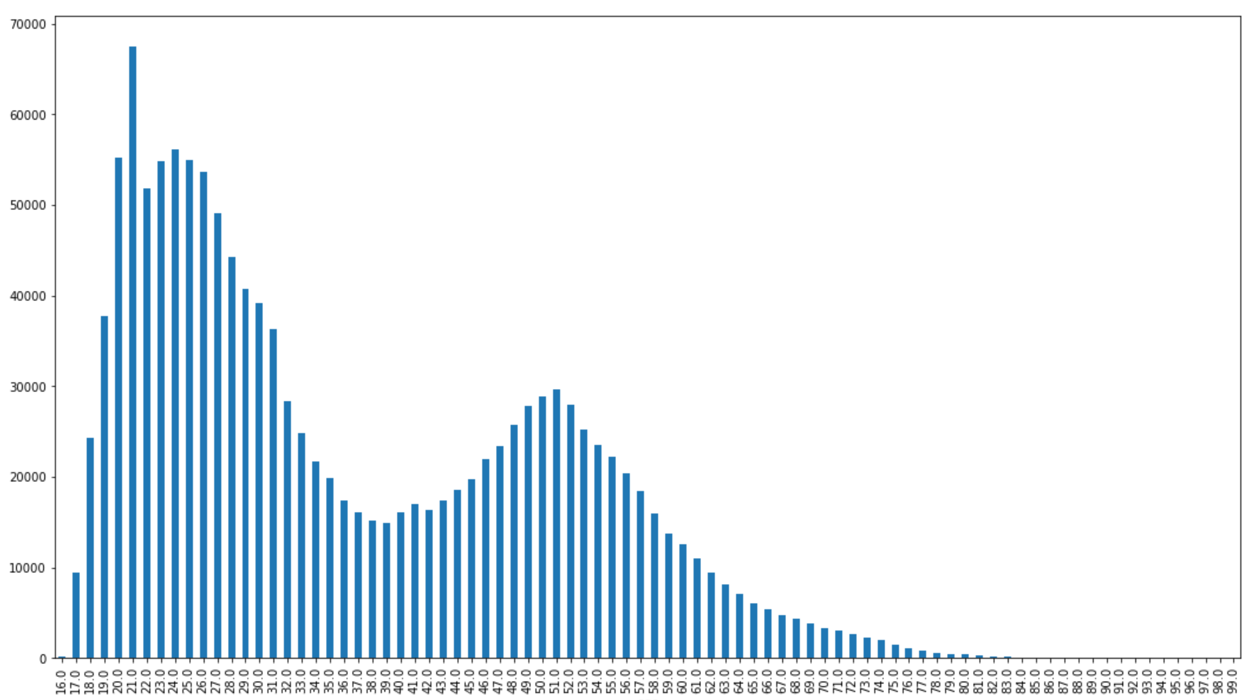


Figure 6. Age distribution

After examining the dataset, some interesting results are calculated. Table 5 show the total number of customers who purchase and not purchase. Table 6 show that the top 5 customers who purchase most product.

the number customer who purchases	1362281
the number of customers who do not purchase	9699

Table5. The total number of customers about purchases

Customer id	Total article
be1981ab818cf4ef6765b2ecaea7a2cbf14ccd6e8a7ee985513d9e8e53c6d91b	1895
b4db5e5259234574edfff958e170fe3a5e13b6f146752ca066abca3c156acc71	1441
49beaacac0c7801c2ce2d189efe525fe80b5d37e46ed05b50a4cd88e34d0748f	1364
a65f77281a528bf5c1e9f270141d601d116e1df33bf9df512f495ee06647a9cc	1361
cd04ec2726dd58a8c753e0d6423e57716fd9ebcf2f14ed6012e7e5bea016b4d6	1237

Table 6. The top 5 customer who purchase most products.

The Table 7 show the date that have most customer. The Figure 7 show the total articles sold by time. The Figure 8 show the total articles sold by age. The results show that dates around the holidays have the most consumers. The second possible reason is the implementation of discounts. Young consumers provide the most sales.

date	Total customer
2019-09-28	198622
2020-04-11	162799
2019-11-29	160875
2018-11-23	142018
2018-09-29	141700

Table 7. Total customer by time.

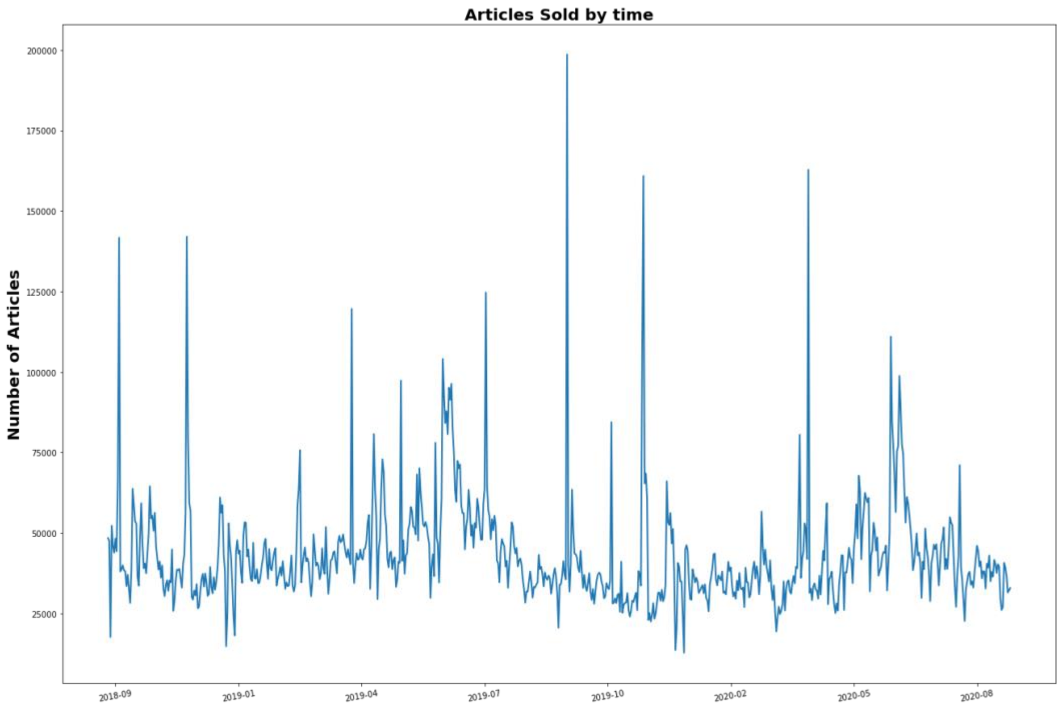


Figure 7. Articles Sold by time

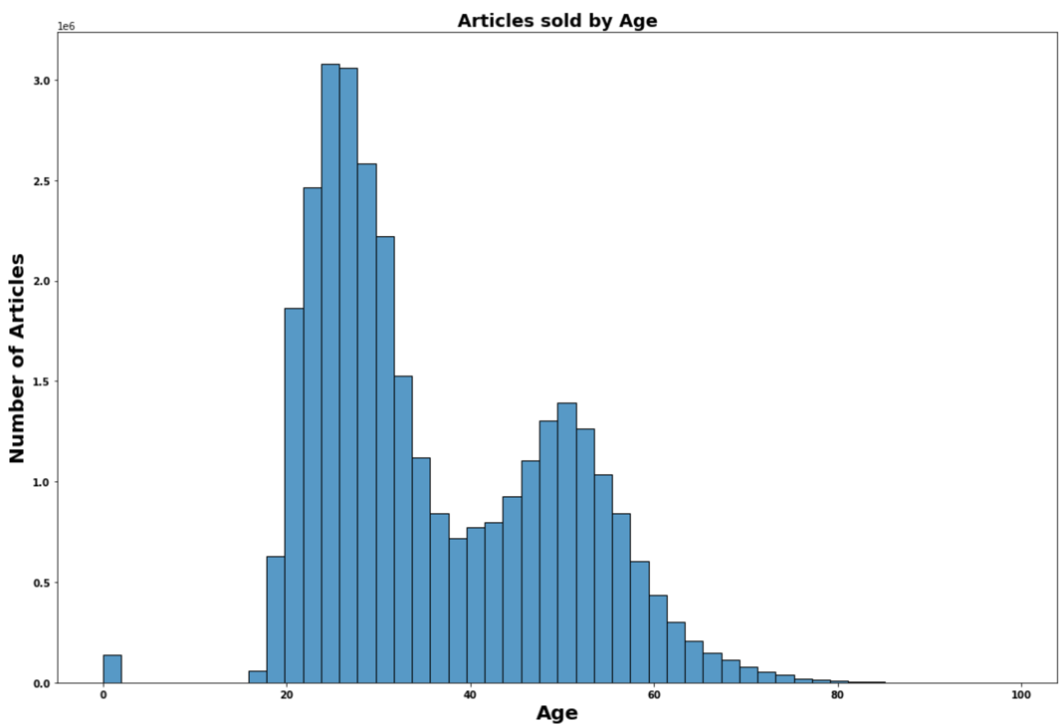


Figure 8. Articles Sold by age

The Figure 9 show that articles sold by product group and index group. Garment Upper body is the most popular product group, and in Garment Upper body, Ladieswear is the most chosen. Table 8 and Figure 10 show the total transaction amount by date. The result is basically the same as the previous total number of customers by date. The results show that days with more customers have higher sales.

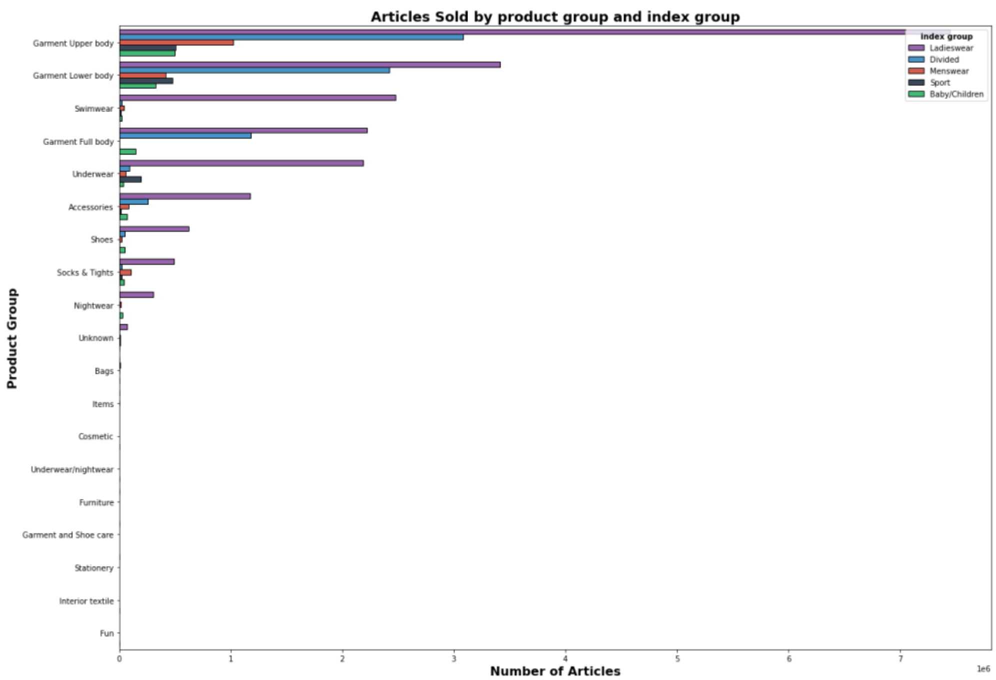


Figure 9. Articles Sold by product group and index group

date	Total transaction
2019-09-28	6161.603068
2020-04-11	4444.342390
2019-11-29	4071.381305
2018-11-23	3961.987763
2018-09-29	3891.939441

Table 8. Top 5 of total transaction by date

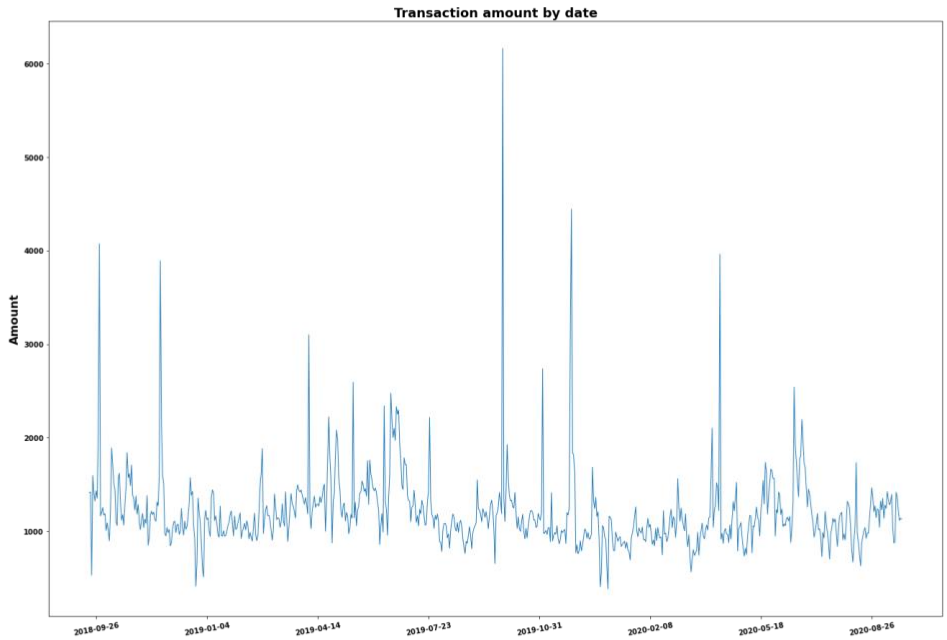


Figure 10. Transaction amount by date

Figure 11 shows the frequency of fashion news. Most customers choose none. Figure 12 shows the status of club member. Most customers are active members.

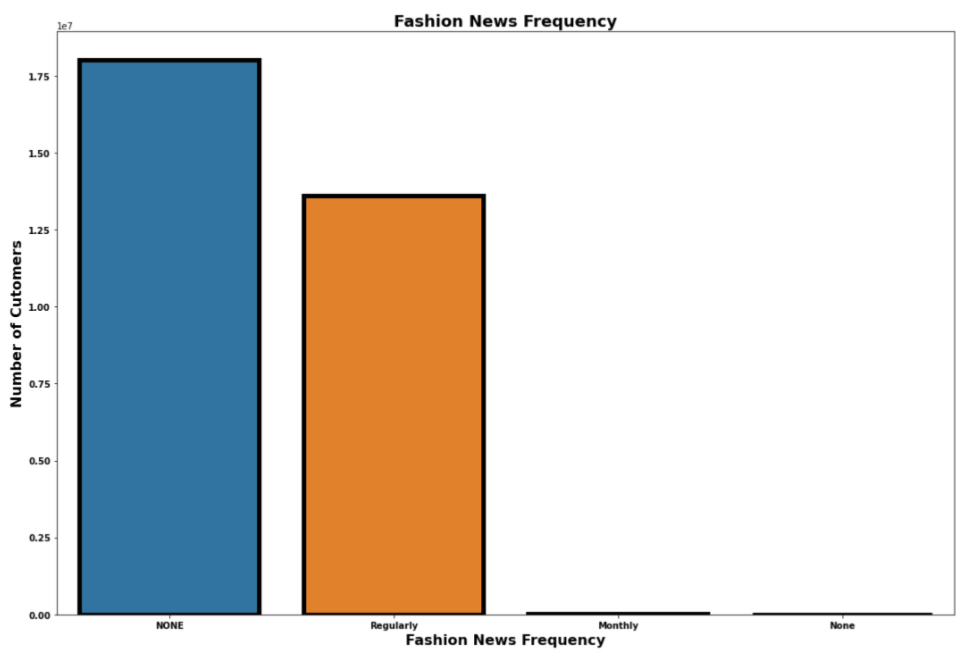


Figure 11. Fashion News Frequency

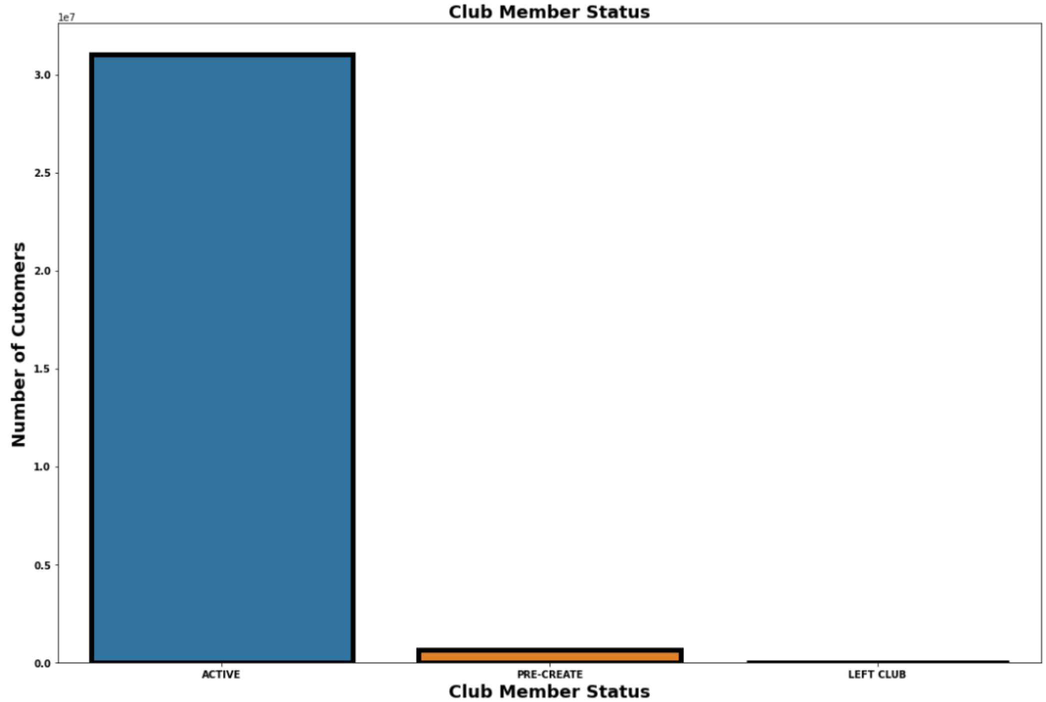


Figure 12. Club Member Status

To understand the reasons that affect the number of product purchasers, this project analyzes the correlations of each value. We first calculated the total number of customers for each item, and then exported a new data set table. We checked outliers again through the new data table (Figure13). Figure 14 and Figure 15 show the results about the correlation of the total customer table. The results show that product code is the most influential factor. Product code has a negative correlation with the number of customers, on the contrary, the price has a positive correlation with the number of customers.

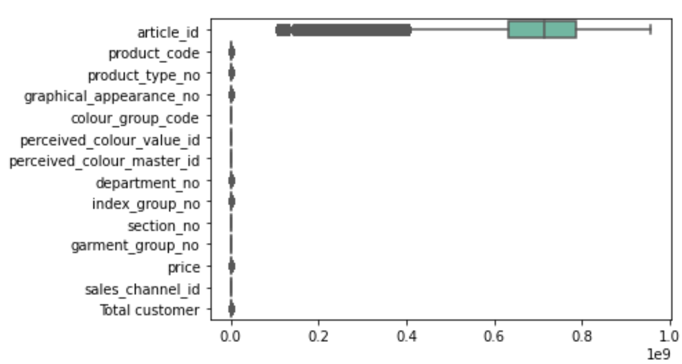


Figure 13. Box plot of total customer table

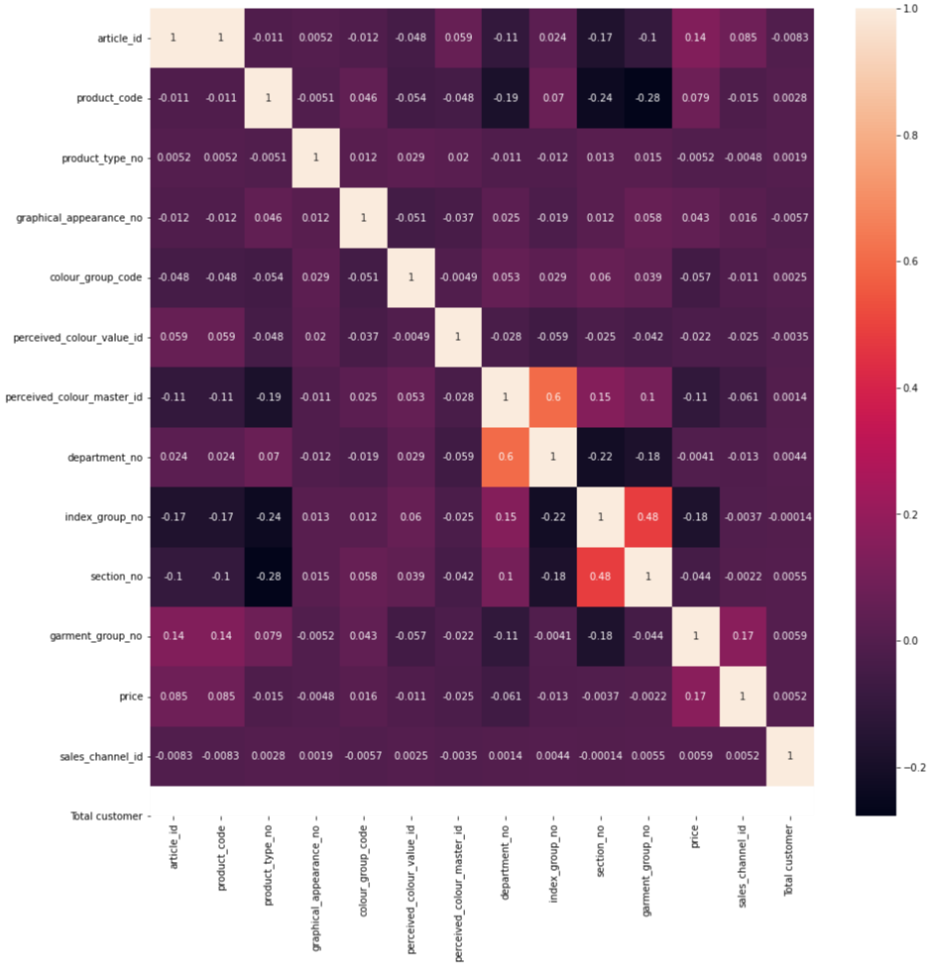


Figure 14. Correlations of total customer table

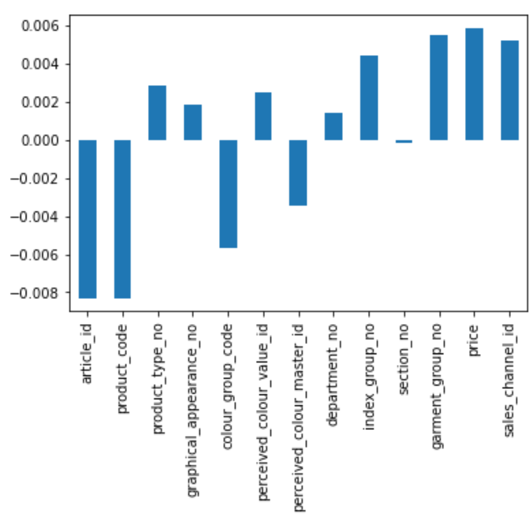


Figure 15. Correlations about total customer

Modeling

Apache Spark is an open-source distributed query and processing engine that provides the flexibility and scalability of MapReduce with faster speed, and users can read, transform, and aggregate data to train and design complex statistical models by Spark (Drabas et al., 2017). In this project, we use PySpark to design models and analyze data. PySpark allows interaction via notebooks like Jupyter or Databricks (Drabas et al., 2017). We use Jupyter and Python to complete the code part of this project.

Alternating Least Squares (ALS) is a matrix factorization algorithm that runs in parallel, ALS can be used in Apache Spark ML for large-scale collaborative filtering problems (Liao, 2018). ALS decomposes the given matrix R into two factors U and V such that $R \approx UTV$. The i column of the user matrix is denoted by u_i and R could be called the rating matrix with $(R)_{ij} = r_{ij}$, therefore the following problem is solved by (Apache Flink 1.2 Documentation, 2022):

$$argmin = \sum_{\{i,j|r_{i,j\neq 0}\}} (r_{ij} - u_i^T v_j)^2 + +\lambda(\sum_i n_{ui}\|u_i\|^2 + \sum_j n_{vj}\|v_j\|^2)$$

Reference:

Zhao, Xuesong. “A Study on e-Commerce Recommender System Based on Big Data.” *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, 2019, https://doi.org/10.1109/icccbda.2019.8725694.

Jessica Young | Feb 18, 2022, et al. “US Ecommerce Grows 14.2% in 2021.” *Digital Commerce 360*, 16 Sept. 2022, https://www.digitalcommerce360.com/article/us-ecommerce-sales/.

“H&M Personalized Fashion Recommendations.” *Kaggle*, https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations/data.

Drabas, Tomasz, et al. *Learning Pyspark: Build Data-Intensive Applications Locally and Deploy at Scale Using the Combined Powers of Python and Spark 2.0*. Packt Publishing, 2017.

Liao, Kevin. “Prototyping a Recommender System Step by Step Part 2: Alternating Least Square (ALS) Matrix...” *Medium*, Towards Data Science, 19 Nov. 2018, https://towardsdatascience.com/prototyping-a-recommender-system-step-by-step-part-2-alternating-least-square-als-matrix-4a76c58714a1.

“Alternating Least Squares.” *Apache Flink 1.2 Documentation: Alternating Least Squares*, 19 Oct. 2022, https://nightlies.apache.org/flink/flink-docs-release-1.2/dev/libs/ml/als.html.