Dong Nguyen a180770 – Github: https://github.com/dongerouss178/BigDataProjectB

1. Data Description

Our research goal is to develop a recommendation system that returns the top N categories each customer might purchase. Although the existing Order Item table contains much of the data we need, it does not have enough detailed user or product attributes. To address this problem, we will join two additional tables into a single dataset for our model. First, we will combine the Order Item table with the Product table to obtain information about each product, such as category, product cost, and brand. Then, to obtain customer information for each order, we will merge with the User table. This will provide us with customer location, gender, and age, which might be helpful features.

The new, merged dataset will allow us to see which categories frequently occur together. With user information included, we can identify and segment our customer base more effectively for tailored recommendations. Our combined data contains a total of 35 columns, but many are duplicate identification fields, such as product ID and inventory item ID, which can be removed as non-features. The key features we have identified in our dataset for the recommendation system are:

1. **Order Created Date** and **User Created Date**: We chose these two date columns as key features, compared to shipping date or return date. The reason is that these dates show the actual times when users took action. Our recommendations will be based on the time the user made a purchase.
2. **User age**, **gender**, and **country**: This information tells us about the demographics of our users and potentially allows us to segment our users for better insights. Although our dataset includes usernames, states, and addresses, we believe this information is redundant and will not add much value in determining customer demographics.
3. **Product cost** and **price**: This information is crucial to identify which categories are most profitable when making recommendations.
4. **Product category**, **brand**, and **department**: This information can help us explore which categories customers tend to buy more frequently on our website. A product name or ID might not be useful in our case, since fashion trends can change very quickly; it is better to focus on general categories or brands.
5. **User ID** and **Order ID**: Although these two pieces of information do not show any attributes, they are crucial for grouping and aggregating our dataset.

2. Explore Data Analysis

Our task is to generate a recommendation list for both new and existing users who have made purchases on our website. This task is classified as predictive, involving both clustering and classification. Firstly, we need to cluster customers into different groups to identify similarities, as this can help predict which categories a particular group of customers is likely to buy. Additionally, some categories often appear together in the same transaction, so we will explore the association relationships between items.

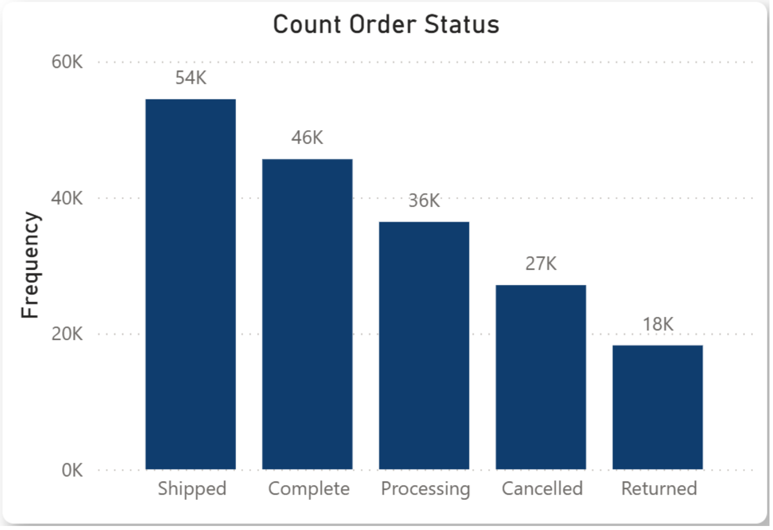
Before building the recommendation system, we will explore and visualise our dataset to gain better insights. After merging the tables and removing non-essential features, we have approximately 18 columns remaining. In this section, we will use Power BI to visualise our dataset and use Python along with other libraries to assist with advanced grouping and analysis.

2.1 Understanding Order Overview

Understanding the order information is a crucial first step for our project. This provides an overview of data distribution and overall business activity. We will start investigate the distribution of order statuses. Orders can have four different statuses, which are as follows:

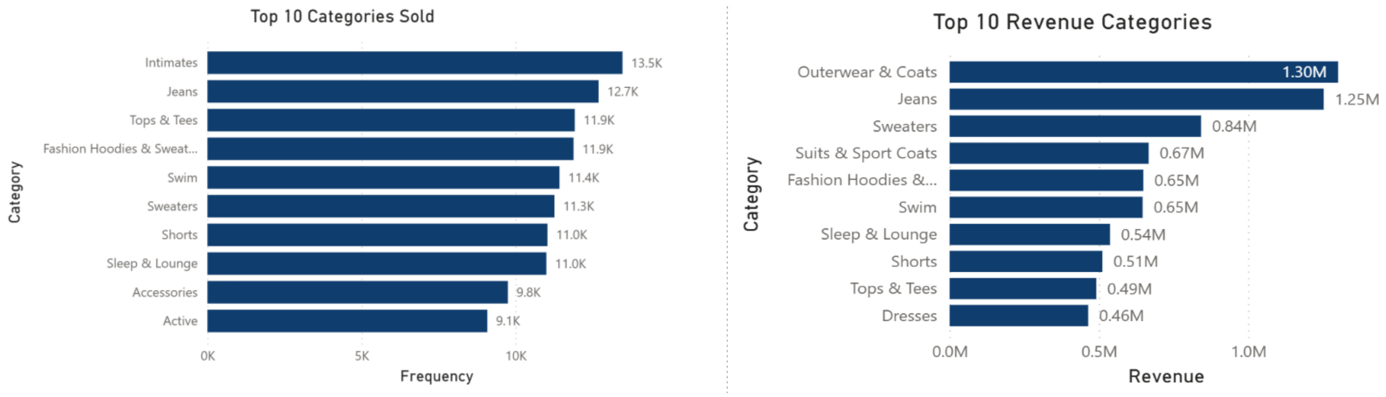
1. Shipped: The order is on its way to the customer.
2. Complete: The order has been delivered to the customer.
3. Processing: The order has been purchased and is waiting to be shipped.
4. Cancelled: The order was cancelled before it was shipped.
5. Returned: The order was delivered but returned by the customer.

Figure 1: Count Order Status



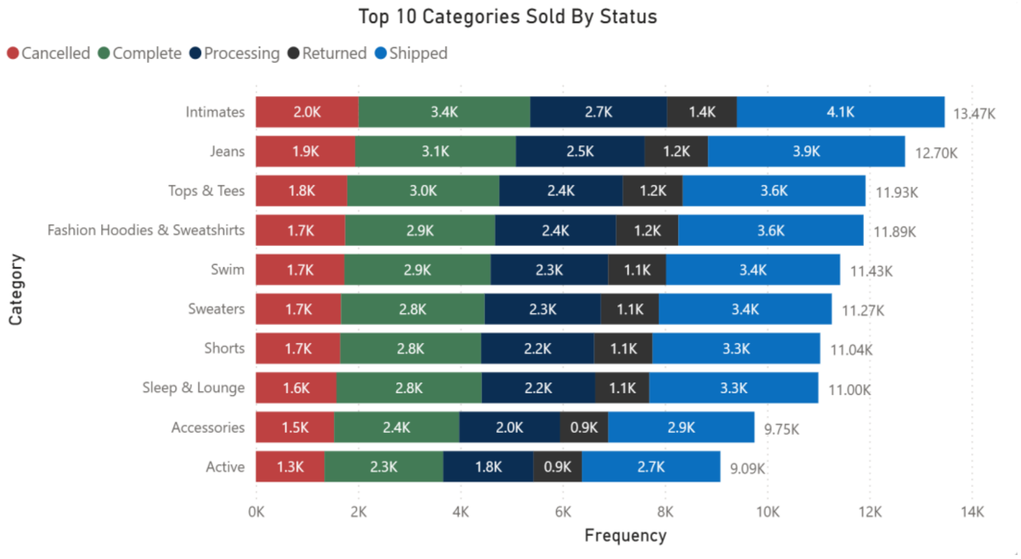
From Figure 1, we can see that our orders are distributed across these statuses. To improve the recommendations for our users, we will focus only on orders that are Shipped, Complete, and Processing, as these statuses reflect positive outcomes. It is also important to identify the top-selling products, since these can significantly influence our pattern mining. The more frequently an item appears, the more likely it is to have an associative relationship with other items.

Figure 2: Top 10 Categories Sold and Top 10 Revenue Categories



From Figure 2, we can see that the categories with the highest sales do not always generate the most revenue, so this is something we need to consider when building our recommendation system. We will aim to get a balance between profitability and the likelihood of purchase. By combining the insights from Figure 1 and Figure 2, we can compare which categories are more likely to be returned or cancelled. This analysis will help us determine whether we should recommend those categories to our customers (Figure 3).

Figure 3: Top 10 Categories by Status



We now have an overview of our orders and categories. It is important to understand characteristics such as revenue trends, profits, and which categories yield the most revenue and profit. Additionally, we can identify time series information, such as when customers usually make purchases and which quarters or months have the highest order volumes. These characteristics allow us to observe customer patterns and make more appropriate recommendations.

Figure 4: Revenue by hours

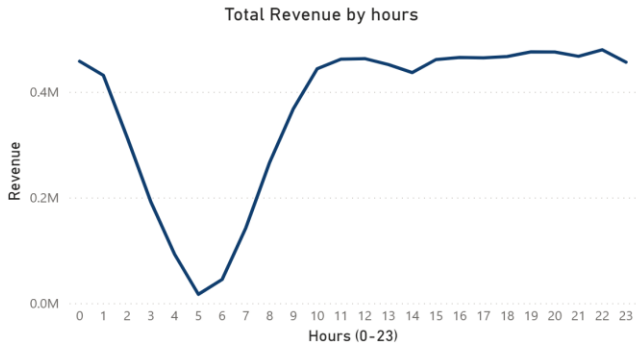
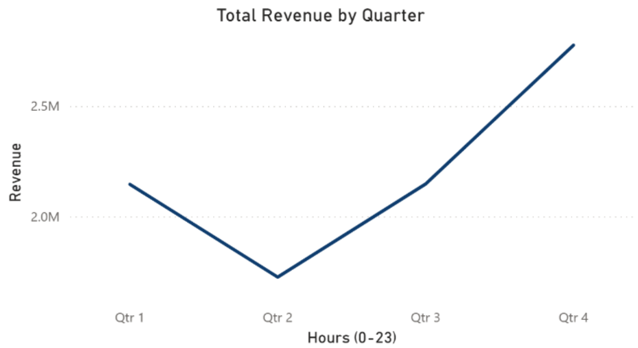


Figure 5: Revenue by Quarter

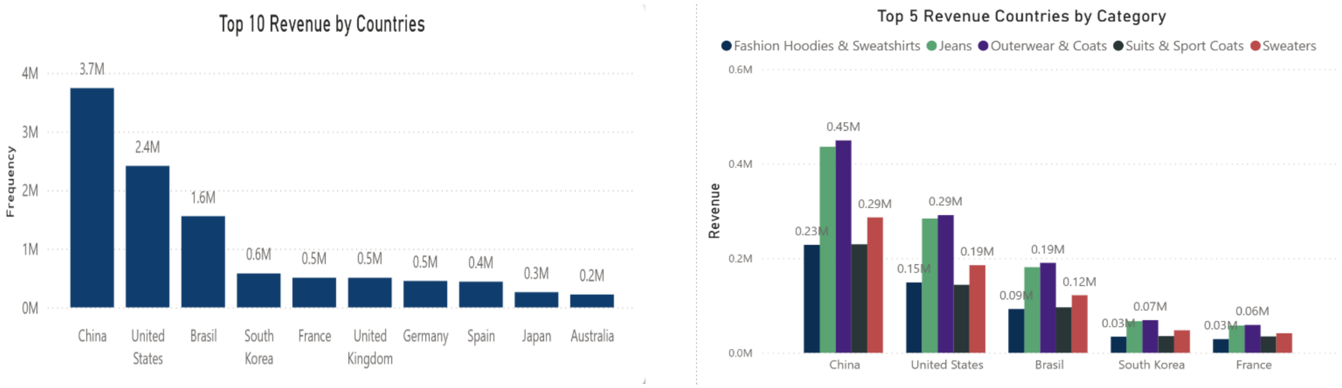


Based on Figure 4 and Figure 5, we can determine that our customers typically make purchases between 10 am and 11 pm, with Q3 and Q4 being the periods of highest sales. This insight can help us tailor our marketing strategies and promotions. We might consider bundling and recommending more items during these peak times.

2.2 Understanding Customer Behaviours

We will further explore our customers' buying patterns and demographic trends. This can help us provide the right recommendations to specific customer groups. Typically, customers within the same group are likely to purchase similar products (Tan et al., 2016). We begin by examining which countries generate the most revenue and what categories are most popular in those countries. Since our dataset contains 26 unique countries, it is difficult to visualize all of them, so we display only the top 5 or top 10 countries.

Figure 6: Revenue by countries and categories by countries



From Figure 6, we can see that China, the US, and Brazil are our top countries, and their buying patterns are quite similar. To have better visualisation, we display only the top 5 categories in each country. This information suggests that recommendations generated for a random person in one country may also be similar in another country. While clustering by country or gender is one approach, we consider grouping based on additional attributes.

One robust technique that incorporates multiple attributes is called K-Nearest Neighbour (KNN). This unsupervised learning method does not require us to manually analyse each attribute or attempt to segment users into groups ourselves (Li et al., 2009). Before applying this algorithm, we need to create a new table that aggregates our customers into a single profile, containing the key attributes we want the algorithm to use for clustering. We use Python code to generate these customer profiles from our merged table, then apply the KNN library to cluster our customers. Table 1 below describes each column in this new customer profile.

Table 1: New Customer Profile Table

Column	Description
user_id	This column is used to identify each customer.
total_spent	The user's total amount spent on all orders.
total_profit	The profit from all the orders.
order_frequency	The number of orders the client has placed, or how many times the client has repurchased.
items_purchased	The number of unique categories the person has bought from.
avg_order_value	The average value of their orders.
age	The age of the customer.
gender	Male or Female.
traffic_source	The platform that led the customer to visit our website.
country	The customer's country.
days_since_last_purchase	The number of days since the customer's last purchase. We used the maximum day in the dataset and then subtracted the maximum day the customer purchased.
user_lifetime_days	The number of days since the customer first registered on our website.
no_return	The total number of products returned.
no_cancel	The total number of products cancelled.
no_complete	The total number of products delivered.
no_shipped	The total number of products shipped.

no_process	The total number of products waiting to be shipped.

It is important to examine the distribution of each attribute to detect outliers before applying KNN, as clustering algorithms are sensitive to them. We created distribution graphs for all features to identify potential outliers (Figure 7). We can observe the data is quite skewed for total spent, which also affect profit and average order value in similar ways. Figure 8 shows that the outliers for total spent are those above $500. We decided not to remove these outliers but to keep them in mind when we begin building the recommendation model. The main goal of this phase is simply to explore the insights the data can provide.

Figure 7: Distribution of Customer Features

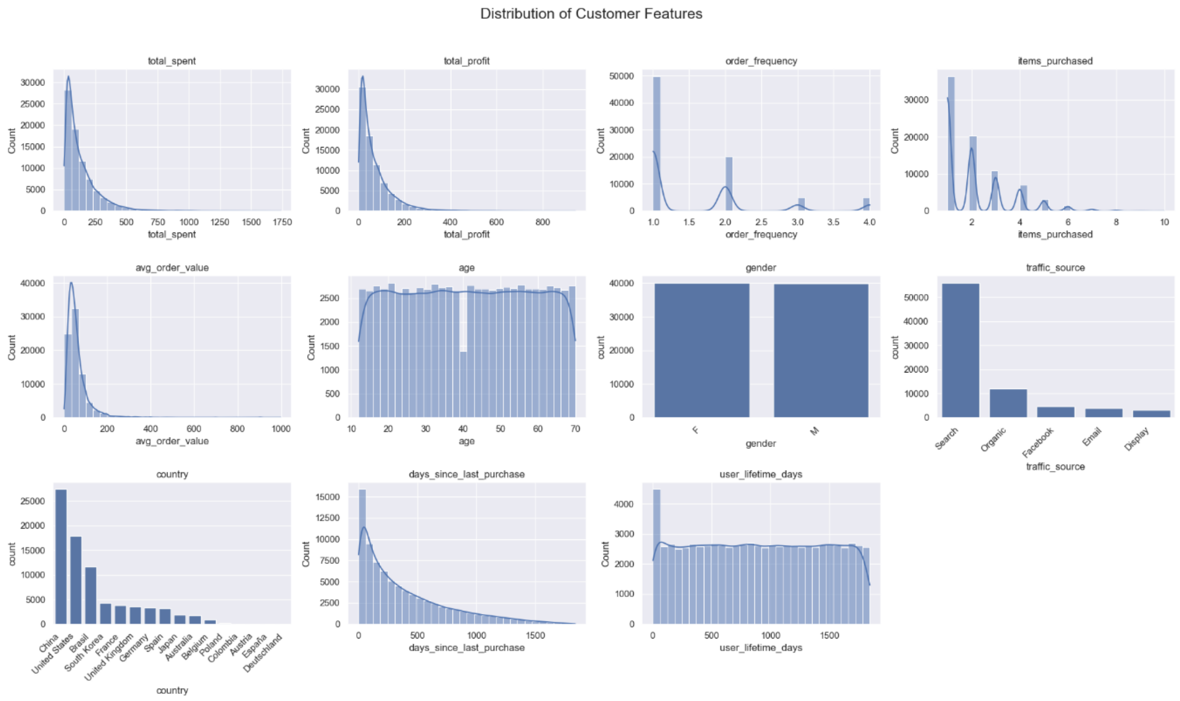
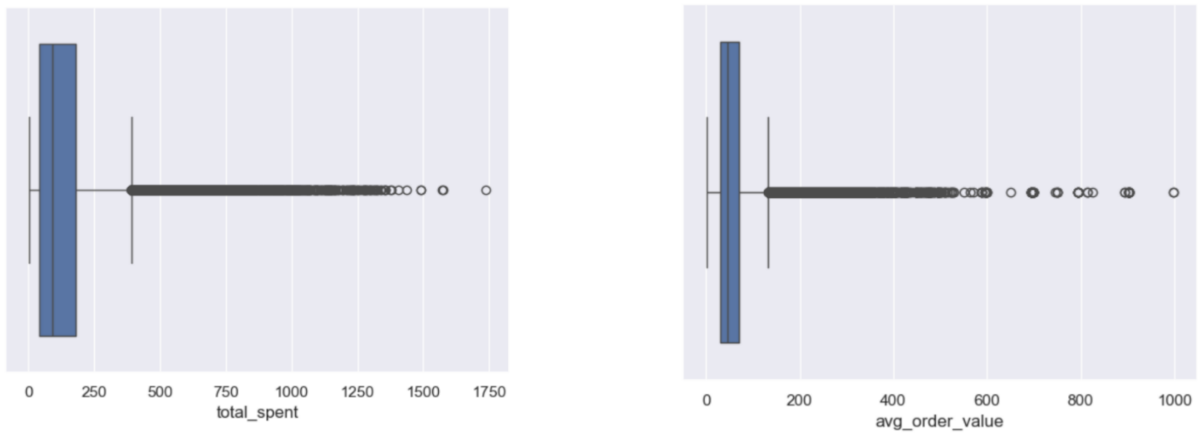
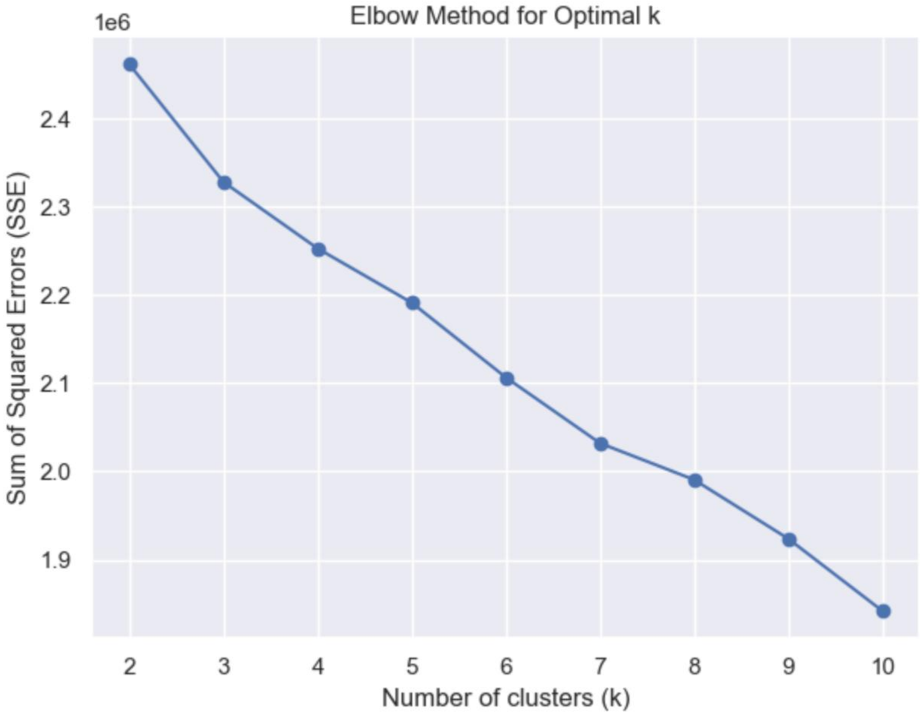


Figure 8: Customer Total Spent and Average Order Value Box Plot



The KNN method requires all columns to be numeric, so we convert gender, traffic source, and country into one-hot encoded features. In this algorithm, we also need to determine how many clusters to use, and one strategy for this is the Elbow method (Zhang et al., 2018). Based on Figure 9, the optimal number of clusters would be either 3 or 4, so we chose 4 for experimentation in this phase.

Figure 9: Elbow Graph to determine optimal K



After determining the optimal value of K, we used the KMeans function from sklearn to cluster our customers and merged the clusters into our customer profile DataFrame (Darshith & Puttaswamy, 2024). We used violin plots to display each attribute by cluster, allowing us to observe their distributions and differences. From Figure 10, it is clear that cluster 1 contains our “VIP” customers, as they spend the most and make purchases regularly compared to other groups. Additionally, their last purchase dates are quite recent. The next most loyal group is cluster 3, while groups 0 and 2 consist largely of one-time customers. Figure 11 illustrates the distribution of each cluster across countries. This information helps us identify where our VIP customers are located and whether there are specific countries to target for improving revenue and sales.

Figure 10: Violin Plots for each attribute by Clusters

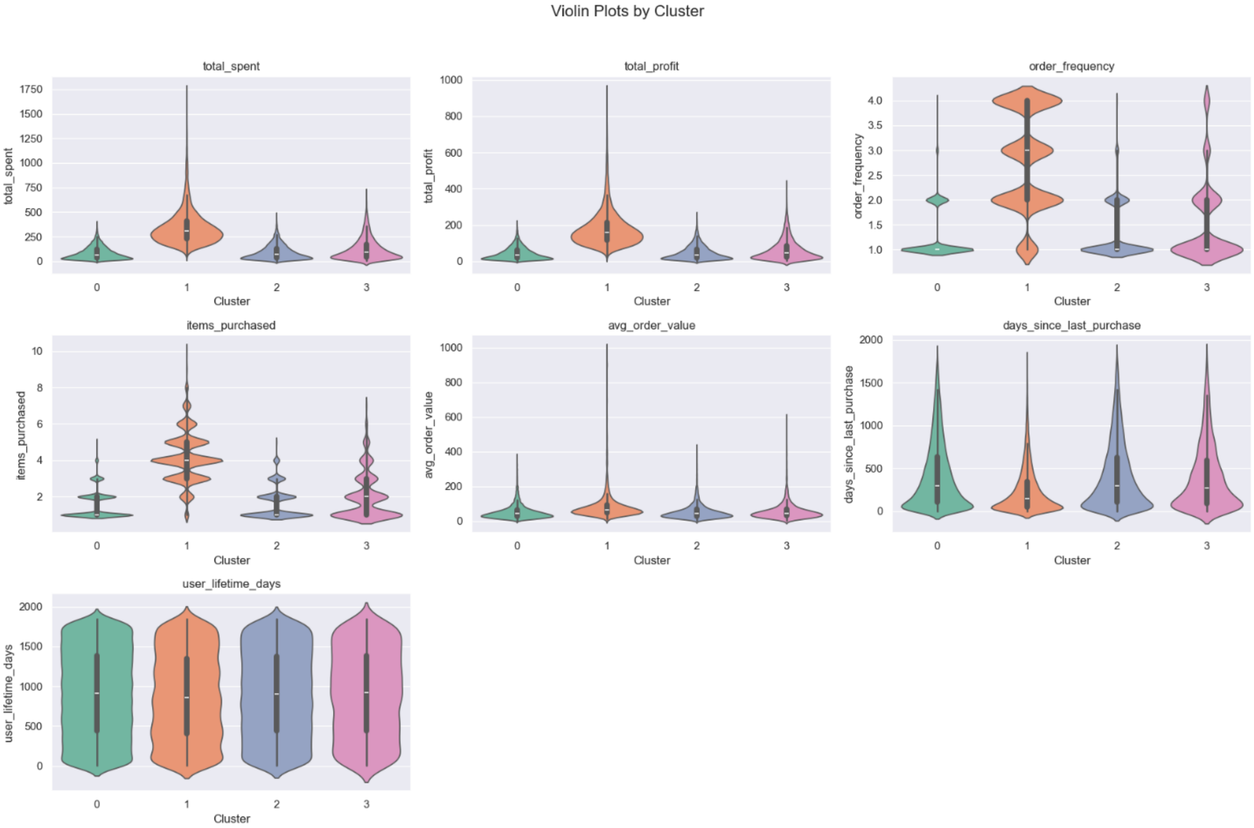
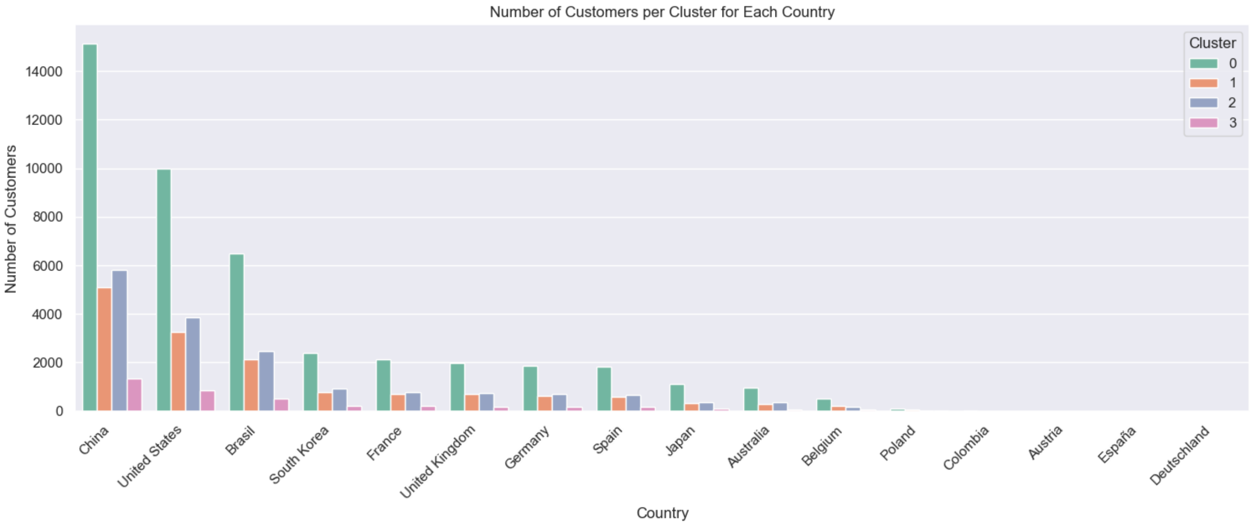


Figure 11: Number of Customers per Cluster for each country



3. Problem Refinement:

After exploring our dataset, we decided to refine our research problem. From Figure 1, we observed many orders that were cancelled or returned. To reduce the dataset and avoid recommending products that have been returned or cancelled, we will only consider orders that are not returned or cancelled. Also, since there are many countries that contribute little value and the buying patterns are similar across all countries, we will focus on the top 10 countries that generate the most revenue. Moreover, since we have identified four different customer groups from our clustering exploration, we will create separate recommendations for each group if our initial model does not perform well.

Our final research question is:

Build a recommendation system for each customer group in the top 10 countries. This system should work for both cold cases (customers who have never purchased on our website) and existing customers.

To fine-tune the model: Given the purchase time and country, will our model show improvement?

4. Reference

Darshith, T.N. & Puttaswamy, H., 2024. Comparative Analysis of K-Means Algorithm and Gradient Boosting Algorithm for E-commerce Platform. *ResearchGate*.

Li, X., Shi, D., Charastrakul, V. & Zhou, J., 2009. Advanced P-Tree based K-Nearest neighbors for customer preference reasoning analysis. *Journal of Intelligent Manufacturing*, 20, pp. 569–579.

Tan, P.N., Steinbach, M. & Kumar, V., 2016. *Introduction to Data Mining*. New Delhi: Pearson Education India.

Zhang, S., Li, X., Zong, M., Zhu, X. & Wang, R., 2018. Efficient kNN Classification With Different Numbers of Nearest Neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5), pp.1774–1785.