Customer Segmentation of Online Retail Customers

Gillian Tatreau
Department of Data Science, Bellevue University
DSC680: Applied Data Science
Amirfarrokh Iranitalab
4 February 2024

Business Problem
All retailers wish to increase their profits, and therefore, their sales. Online retailers have the opportunity to reach a larger customer base because they are not limited to the individuals that are able to physically visit the store; therefore, understanding customer characteristics and transaction patterns within their virtual storefront is even more crucial in developing effective marketing strategies as well as allocating resources where they are most needed. In order to increase their sales, a retailer must first understand what their current customers are buying and what sales patterns exist within their own order history data. These sales patterns might include insight into their most-sold products, the geographic region that their customers are buying from, and what products are bringing in the most profit.

Background/History
Customer segmentation is the division of all customers into groups that share similar characteristics. Each group can be described by a persona or profile that is then used by marketing teams to best appeal to that customer segment.

Data Explanation
The data was found on Kaggle, but it originates from a dataset shared on Data.world. The data details information about transactions made on an online retailer site.

Data Dictionary:

- InvoiceNo: unique identifier for each transaction (int)
- StockCode: unique identifier for each item (string)
- Description: brief description for each product (string)
- Quantity: number of units sold of product (int)
- InvoiceDate: date and time transaction took place (datetime)
- UnitPrice: price per unit of product (int)
- Country: country where purchase was made (string)

The InvoiceDate variable loads as an object, but it was converted to datetime. NA values were dropped. This removed 135,080 rows, leaving 406,829 rows of viable data. Outliers were assessed using boxplots. Negative values in Quantity were removed; these represent either entry error or returns, but the focus of this project remains on sale transactions. Rows that had values of greater than 1,000 for the Quantity variable and greater than 500 for UnitPrice were removed. This does keep some outliers in the data, but it removes the most egregious.

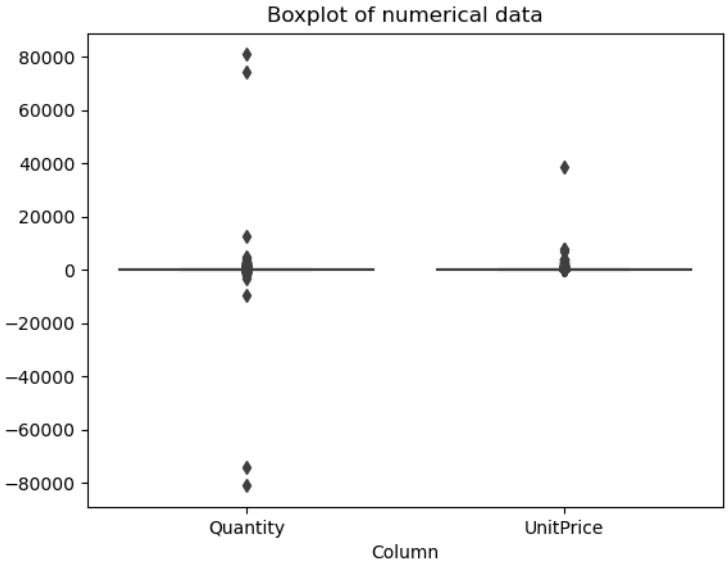


Figure 1. Boxplot of data with outliers.

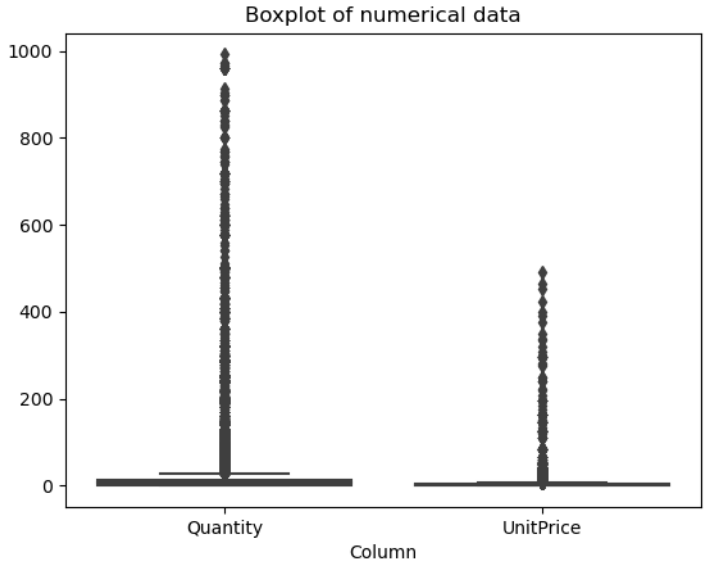


Figure 2. Boxplot of data after removal of some outliers.

Methods

The data will be examined with careful exploration and graphical analysis, or EDA. K-means clustering will be used to create customer segments. These clusters will be evaluated for being appropriate representations of segments within the data by looking at both the plot of the inertia value versus k-clusters, or the elbow method, and the silhouette score. For the Description variable, word clouds will be created for visual analysis of word frequency for preliminary product popularity assessment.

Analysis

First, some statistical exploration of the data was undertaken. There were 4,329 unique customers within the cleaned dataset, which gives us an average of 92 transactions per customer. There are 37 unique entries for the Country variable. The United Kingdom had, by far, the largest number of transactions and number of unique customers, followed by Germany and France. There are eight countries that have higher average transactions per customer than the overall average; these are France, Germany, Netherlands, Australia, Norway, EIRE (or Republic of Ireland), Iceland, and Singapore. The top three countries for average Quantity per transaction were Netherlands, Sweden, and Japan. The three countries with the highest average UnitPrice were Cyprus, Singapore, and Lebanon. It is possible that Singapore might be a growing market for this retailer, as it is the only country in the top five for both average Quantity per transaction and average UnitPrice, and it is also above the overall average for transactions per customer. The relationship between the two native numerical variables is examined.

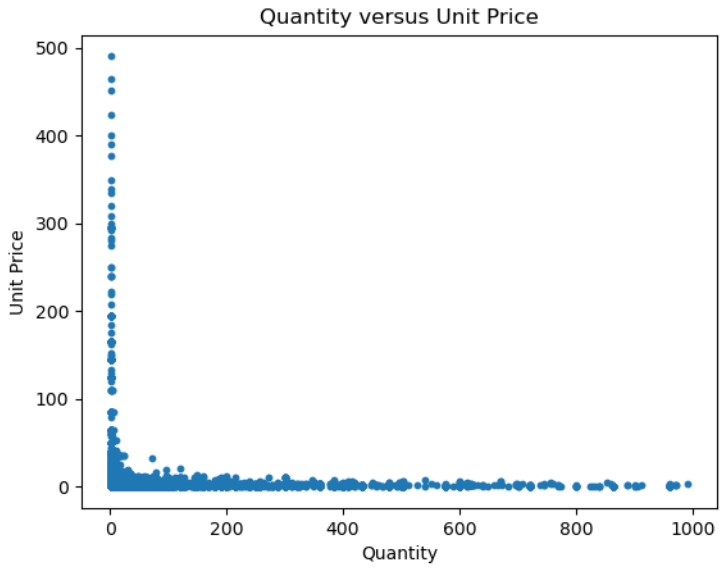


Figure 3. Relationship between UnitPrice and Quantity.

There appears to be a dichotomy between two extremes- in general, higher unit prices yield fewer quantities purchased and lower unit prices yield larger quantity purchased. Mathematically, the relationship appears to follow an exponential function.

Transaction Segmentation

K-means clustering was performed on the numerical variables native to the data, UnitPrice and Quantity, relaying information about the different transaction types within the data. Following the elbow method, the appropriate number of clusters is 5. The silhouette score for 5 clusters is 0.647.

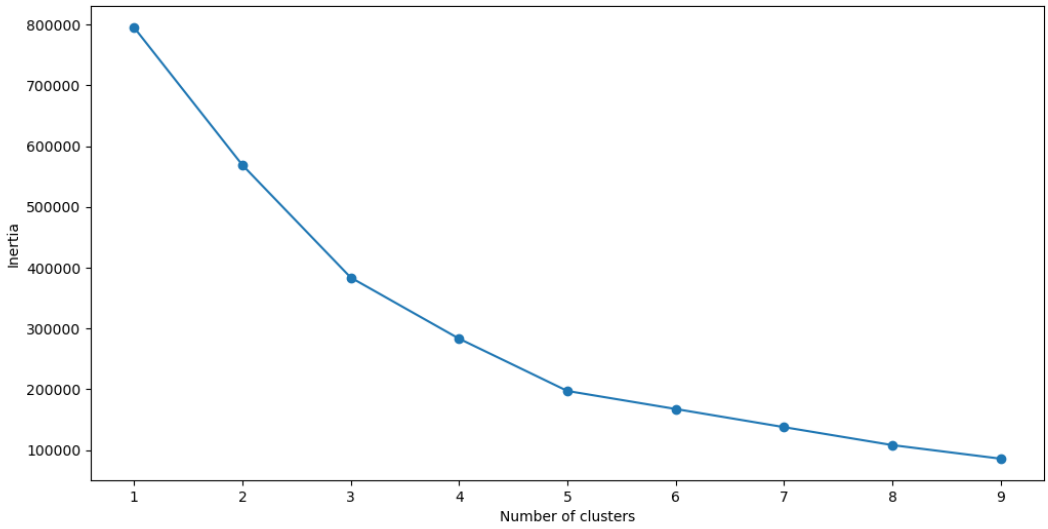


Figure 4. Elbow Method for Transaction Segmentation. k=5.

The five clusters represent five types of transactions within the data; office/classroom sized transactions, super bulk transactions, party-sized transactions, personal transactions, and family/home sized transactions. The office/classroom sized transactions represent cluster 0. These are transactions of less than 25 items of relatively low value. The super bulk transactions represent cluster 1. These are purchases of over 500 items of the cheapest priced items. The party-sized transactions represent cluster 2. They are transactions between 100 and 150 items of the cheapest priced items. The key difference between the super bulk and the party-sized transactions is the differences in quantity purchased. The family/home sized transactions represent cluster 3. These transactions consist of small quantities of mid-priced items. The personal transactions represent cluster 4. These transactions are very small quantities of very high-priced items (see Appendix 1 for boxplots, labeled scatterplot of the transaction segmentation).

Customer Segmentation

To achieve true customer segmentation, some feature engineering was performed to create features that represent the recency, frequency, and monetary value associated with each customer (*RFM analysis for Customer Segmentation)*. Recency was calculated from the InvoiceDate variable, where customers with more recent transactions are given a higher recency score. Frequency was calculated by summing the InvoiceDate by CustomerID. MonetaryValue is the product of the sums of UnitPrice and Quantity per CustomerID. K-means clustering was then performed, and the number of clusters was determined using the elbow method.

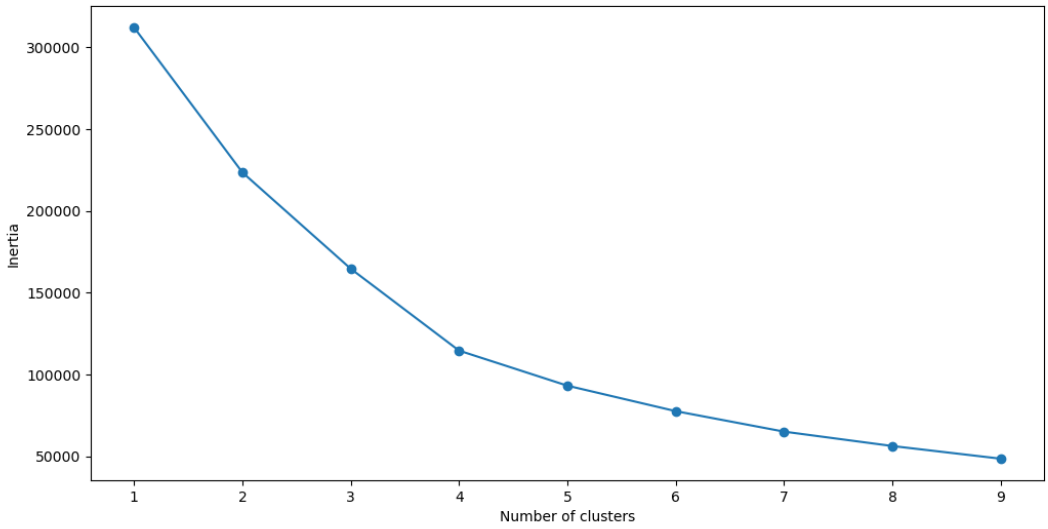


Figure 5. Elbow Method for Customer Segmentation. k=4.

The number of clusters is determined to be 4, with a silhouette score of 0.456. The barplots in figures 6, 7, and 8 below represent the different traits present in each cluster.

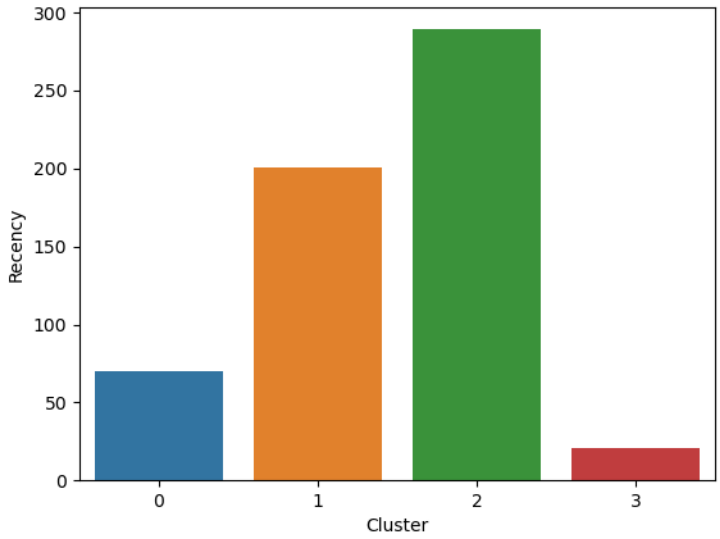


Figure 6. Barplot for cluster traits of Recency.

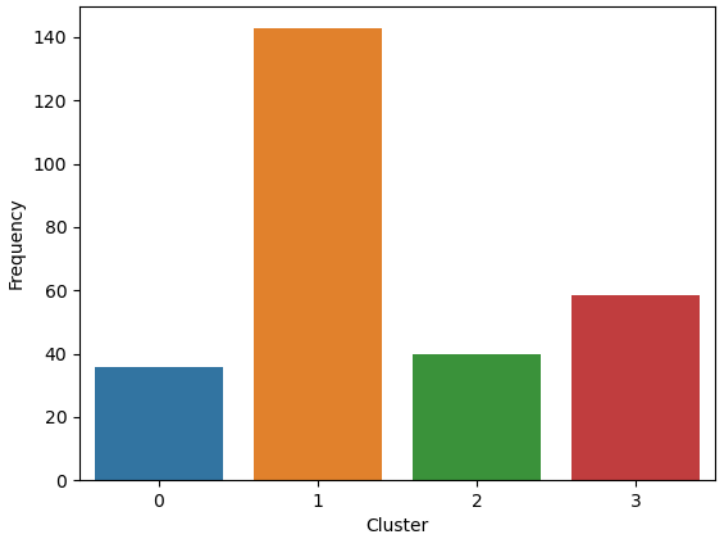


Figure 7. Barplot for cluster traits of Frequency.

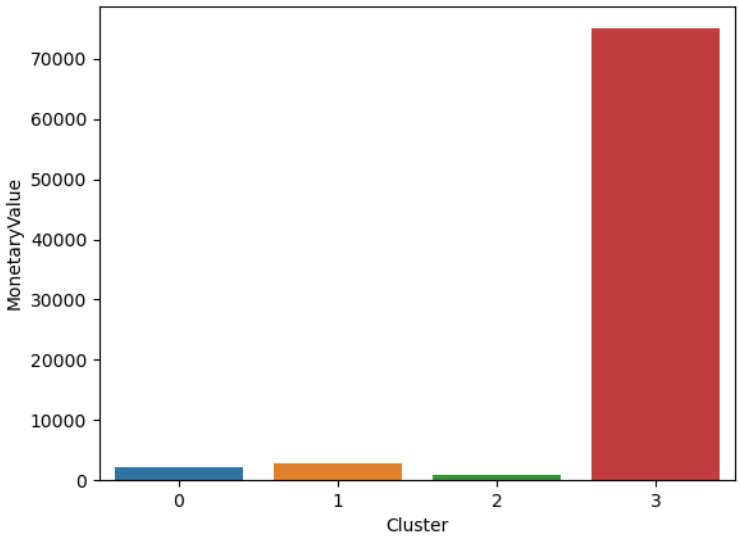


Figure 8. Barplot for cluster traits of MonetaryValue.

The four customer segments can be characterized as the Occasional Browsers (cluster 0), Regulars (cluster 1), New Customers (cluster 2), and Big Spenders (cluster 3). The Occasional Browsers do not shop frequently nor recently and bring very little monetary value to the retailer. The New Customers include customers with some of the most recent transactions, with low frequency scores and have not yet contributed a lot of monetary value to the retailer. The Regulars include customers that are both frequent and mostly recent, but they do not contribute too much monetary value to the retailer individually. The Big Spenders represent large monetary value to the retailer, but those transactions are not overtly frequent or recent.

Common Words in Product Description

The Description variable was used for some preliminary analysis of the most common words found within the text data. This allows the retailer to understand what product traits are more popular and how marketing can address this inherent popularity, or how operations can perhaps assure these products are stocked in higher volumes. Word clouds were created by creating a body of text from the Description variable for all transactions within the data.

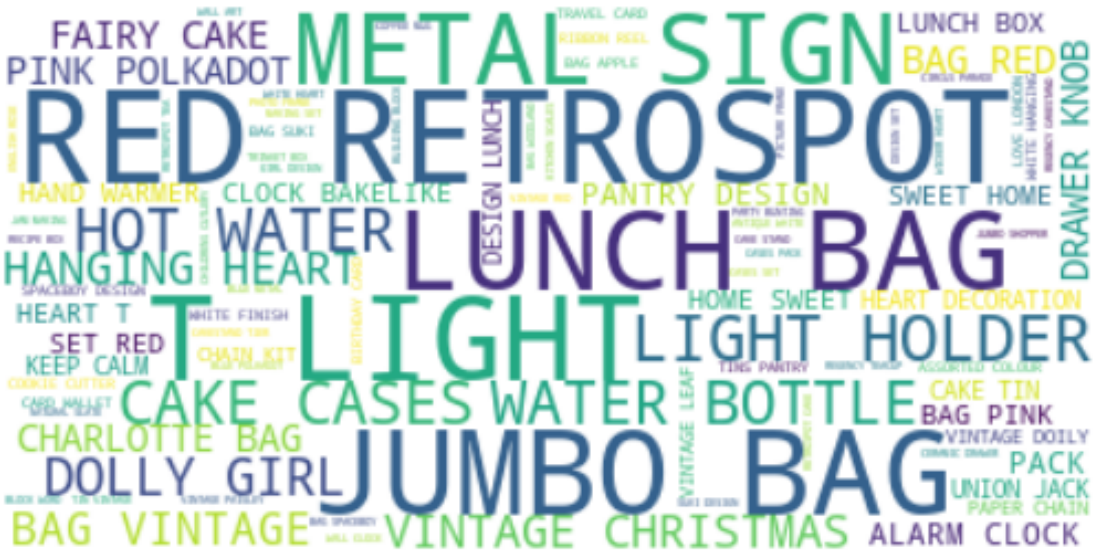


Figure 9. Word cloud of all product descriptions.

Some of the most frequent words include ‘red retrospot’, ‘lunch bag’, ‘T light’, ‘metal sign’, and ‘jumbo bag’. These words represent the traits of the products that have sold the most items of within this data. Analyzing this data by country gives insight into what an individual country’s market looks like (see Appendix 3 for word clouds for 10 individual countries).

Conclusion

There are four customer segments identified within this data. These customer segments each represent different archetypes of customers, which can be leveraged by marketing teams. The silhouette score for the customer segmentation suggests that these clusters do not contain too much overlap and that four clusters is appropriate for the purpose of marketing. The transaction segmentation does not provide nearly as much insight for future marketing campaigns, but it does provide insight into the different general types of transactions that this retailer experiences.

Assumptions

The primary assumption made is that the UnitPrice variable was reported in the same currency value for every entry.

Challenges

One possible challenge is the use of the Description variable. It provides a lot of opportunity for understanding market trends, in terms of product popularity, but true analysis of this variable will require term frequency-inverse document frequency vectorization and then subsequent analysis

of the TF-IDF vectorization; however, while this analysis would provide the stepping stones needed to create any kind of recommender systems that involve the products themselves (in the form of *"if you liked x item, maybe you would be interested in y item"* by looking at similar products or *"similar users also liked y"* by applying this analysis to the clusters from the customer segmentation), this step has not been attempted within the scope of this project. Instead, word clouds were created that visually display the most frequent words within selected bodies of text. This gives some rudimentary insight into the types of products that are currently selling well and still provides some insight for both marketing and operations teams.

Future Uses

The different archetypes identified in the customer segmentation analysis can be used for future marketing strategies. The insights gained from the word clouds could provide a beginning point from which to build recommendation systems in the future, using clusters formed in this project and further analysis of the Description variable.

Recommendations

We could create a special offer for the New Customers to help them become Regulars. Introducing some kind of loyalty program with special discounts after a specified number of transactions for all segments would reward current Regulars, help New Customers turn into Regulars, and possibly incentivize the Occasional Browsers to check back more often. We could check in with Big Spenders with email or other communication that runs along the lines of either "How do you feel about your last purchase?" or "We miss you- here's a special offer just for you". A variation of the latter suggestion could also be used for the Occasional Browsers to encourage them to return to the retailer site.

Ethical Assessment

Using consumer data requires careful ethical considerations, as to consumer privacy, transparency, and fairness. Companies must obtain consumer consent before using their data in any way. Consumers must also be made aware of how the company is using their data and must have the option to opt out of their data being used. Since the purpose of this customer segmentation is to provide insight for future marketing campaigns, careful attention must be paid such that there is no chance that discriminating factors, such as religion, sexual orientation, gender, or race, are being used in a way that would unfairly target or ignore certain customers. Within this data, there is only the Country variable that might possibly pose a problem, if this data is used in a way that highlights differences between developing versus developed countries, which is then used to market more expensive goods to developed countries; however, this variable can still be used to provide insight as long as marketing campaigns are not built in the previously mentioned way. The Country variable could be used to ethically to determine where more support and infrastructure is needed to support the customers within that country or region.

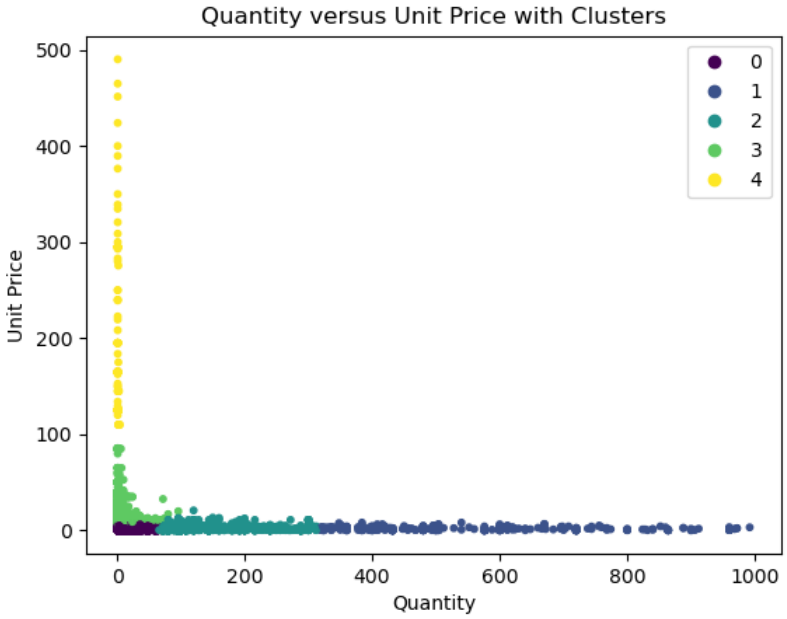
References

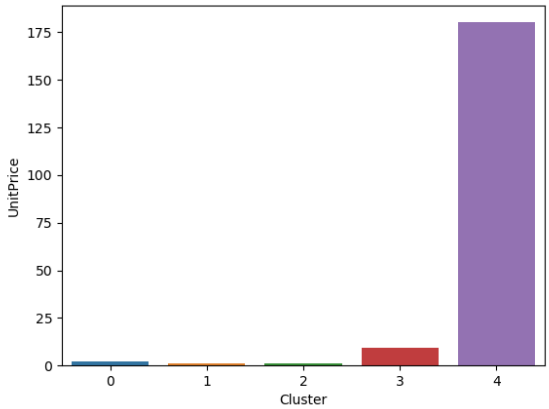
RFM analysis for Customer Segmentation. CleverTap. (2023, November 29). https://clevertap.com/blog/rfm-analysis/

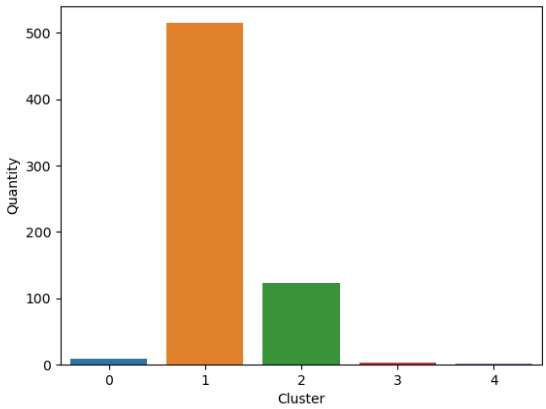
Szafraniec, M. (2017, June 22). *Online retail invoices dataset*. data.world. https://data.world/mszafraniec/online-retail-invoices-dataset

Appendix 1

Transaction Segmentation

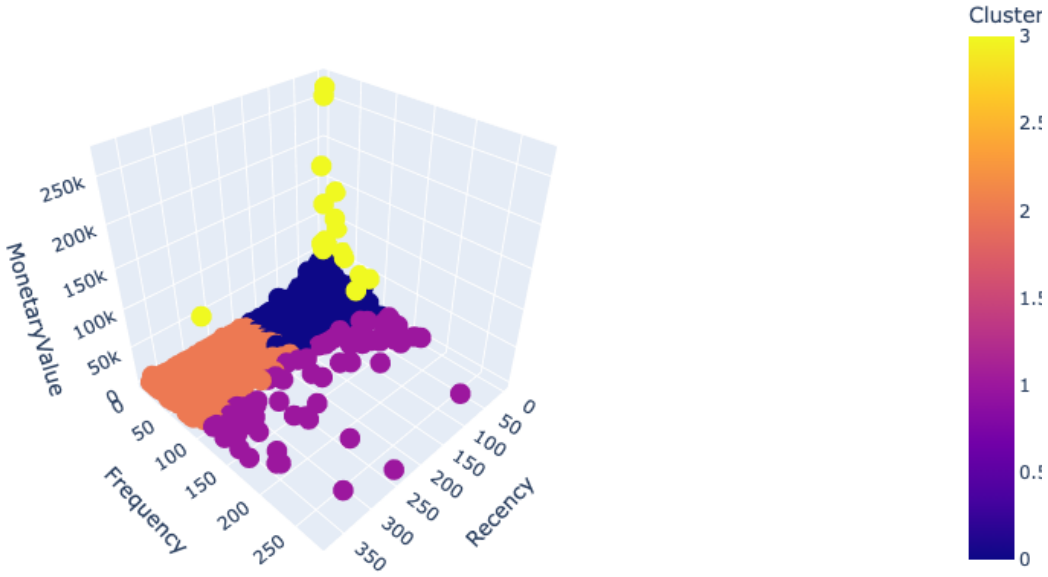






Appendix 2

Scatterplot of Features in Customer Segmentation



Appendix 3

Word Clouds for Individual Countries

Australia

EIRE

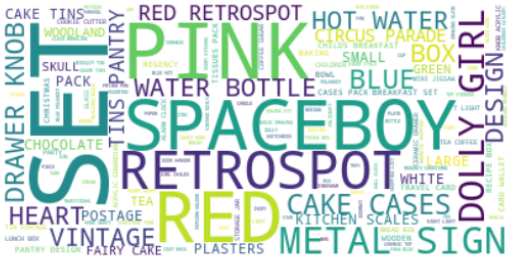
France

Germany

Iceland

Netherlands

Norway



Singapore

Spain

United Kingdom