Online Retail II - Final Report

Vanessa Garretson Capstone 2

Introduction

Customer segmentation is considered the backbone of marketing. By grouping customers by common features, companies can tailor their marketing message to better reach their intended audience and meet their needs.

The Online Retail II dataset contains all online sales transactions for a UK based company between 01/12/2009 and 09/12/2011. The dataset includes 1,067,371 line items across 53,628 transactions and the following eight columns: 'Invoice', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'Price', 'Customer ID', and 'Country'.

Problem statement

Identify three to seven customer segments from the transaction data to focus ongoing marketing efforts while answering the below Key Questions about segment characteristics.

Key Questions:

- How often do each segment make purchases?
- How much have they spent over time?
- How recently have they made a purchase
- What does their basket look like?
- Are there differences in seasonality of purchases between segments?

Data wrangling

Online Retail II dataset was available from the UC Irvine Machine Learning Repository. The dataset was downloaded and saved as an Excel (.xlsx) file. The data is split amongst two sheets in the file, each with over 520,000 entries. The two sheets were imported into a Jupyter Notebook as a pandas dataframe and combined to create a new dataframe called 'df'.

Data wrangling started by removing all cancelled orders from the dataset which were denoted with 'C' before the invoice number. Additionally, all duplicated entries were removed. A 'Sales' column was created by multiplying 'Price' times 'Quantity'.

Numeric anomalies were explored by looking at values with non-positive 'Price' or 'Quantity'. The negative 'Price' values had an item 'Description' of 'Adjust bad debt'. I explored that description and found one additional positive 'Price' entry with that description that appears to be done in error as it was followed by two negative 'Price' entries with the same value to offset. All entries with 'Adjust bad debt' in the 'Description' were removed.

Non-positive 'Quantity' values were explored next. The entries with negative 'Quantity' were a mixture of damages, missing, store use, and other unsellable items based on their 'Description'. Any entry with a non-positive 'Quantity' was removed. Finally, entries with 'Price' equal to zero were explored. There was no clear pattern to entries with a 'Price' equal to zero. The 'Description' for the items was a mixture of adjustments to other entries and product descriptions. The dataset description says the company works with many wholesalers so it was assumed that the entries with 'Price' equal to zero that were not adjustments to other entries were product samples. All entries with 'Price' equal to zero were removed.

During the Initial EDA, it was discovered that 22.7% of entries were missing a 'Customer ID' value. The dataset was split into 2 datasets based on whether or not the 'Customer ID' value was present. 'has_id' includes all entries with a known 'Customer ID' and 'no_id' includes all entries with a missing 'Customer ID'. The 'rfm' dataset used in modeling was based on recency, frequency and monetary (RFM) calculations made by grouping the 'has_id' dataset by 'Customer ID' and calculating the respective RFM values. The RFM calculations process will be expanded on in the next section.

Initial EDA and RFM calculations

'Customer ID' was the only column with missing values. The 'no_id' dataset represented 22.7% of entries and 15.1% of 'Sales' but only 7.8% of 'Invoice' values. The top 'Description' for unknown customers was 'DOTCOM POSTAGE' indicating a large number of online purchases that were possibly checked out without registering. Since the primary objective of this project is customer segmentation for known customers, the 'no_id' dataset was not explored further.

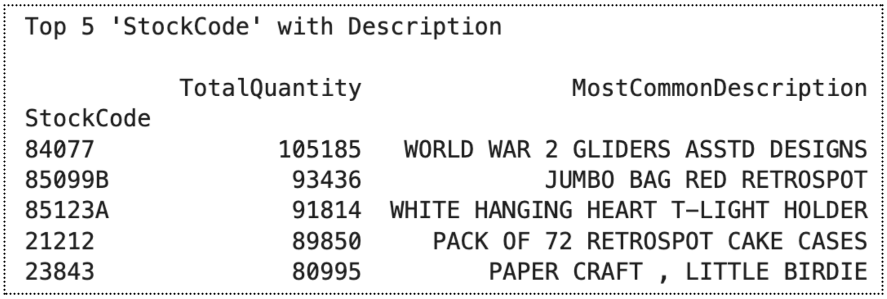
The 'has_id' Dataset with known customers had 779,425 entries from 36,969 unique 'Invoice' values and 5,878 'Customer ID' values. The data was explored in three ways: 1. Line Items, 2. Grouped by 'Invoice' 3. RFM by 'Customer ID'.

1. Basic Feature exploration by line

The basic features of the data set were first explored individually to look for any trends or anomalies.

'StockCode' and 'Description'

'StockCode' had 4,631 unique values and 'Description' had 5,283 unique values as some stock codes had multiple variations on the item description. The analysis focused on the 'StockCode' instead of the 'Description'. Figure 1 shows the top 5 items sold by total quantity with the most common description associated with the given stock code.



```
Top 5 'StockCode' with Description

           TotalQuantity              MostCommonDescription
StockCode
84077             105185   WORLD WAR 2 GLIDERS ASSTD DESIGNS
85099B             93436             JUMBO BAG RED RETROSPOT
85123A             91814  WHITE HANGING HEART T-LIGHT HOLDER
21212              89850     PACK OF 72 RETROSPOT CAKE CASES
23843              80995          PAPER CRAFT , LITTLE BIRDIE
```

Figure 1: Top 5 items sold by total quantity with the most common description associated with the given stock code

'Price'

The majority of items sold (99%) had a 'Price' below £14.50 with a mean of £3.22 and a median of £1.95. The top 'Price' line items had a 'Description' of 'Manual', 'DOTCOM POSTAGE' or 'POSTAGE'. Exploration of the entries with a 'Description' of 'Manual' was mixed with some being included as a single line in larger orders and some being large single line orders with a 'Quantity' of 1 and a high 'Price'. The latter suggests that there were entire orders represented in one line.

Items with 'Price' greater than 50 were explored to see if there were other unique, non-product items. There were 27 unique 'StockCode' values with a price greater than 50, of them the non-product stock codes included 3 for shipping, 2 for adjustments, one for discounts and the manual code mentioned above. Removing the non-product items did not greatly affect the descriptive statistics for 'Price'. The median stayed at £1.95 and the mean lowered slightly to £2.93.

99.9% of line items had a 'Price' less than £50 but there were several stock codes with higher prices plus the unique non-product items mentioned above. Figure 2 shows the distribution of 'Price' with all data on the left and the distribution of 'Price' with 'Price' capped at £50 on the right. The company sells mostly low price items with 86% of items purchased under £5.

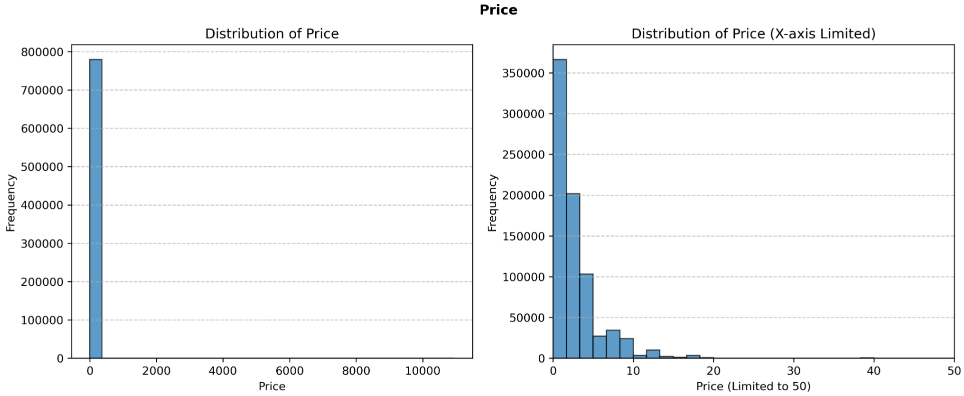


Figure 2: (left) Distribution of all 'Price' values; (right) Distribution of 'Price' limited to Price < £50 which represents 99.9% of data.

'Quantity'

The majority of line items sold (99%) had a 'Quantity' below 144 with a mean of 13.5 and a median of 6. Figure 2 shows the distribution of 'Quantity' with all data on the left and the distribution of 'Quantity' with 'Quantity' capped at 144 on the right. *Most items purchased are purchased in small quantities with the occasional large quantity order.*

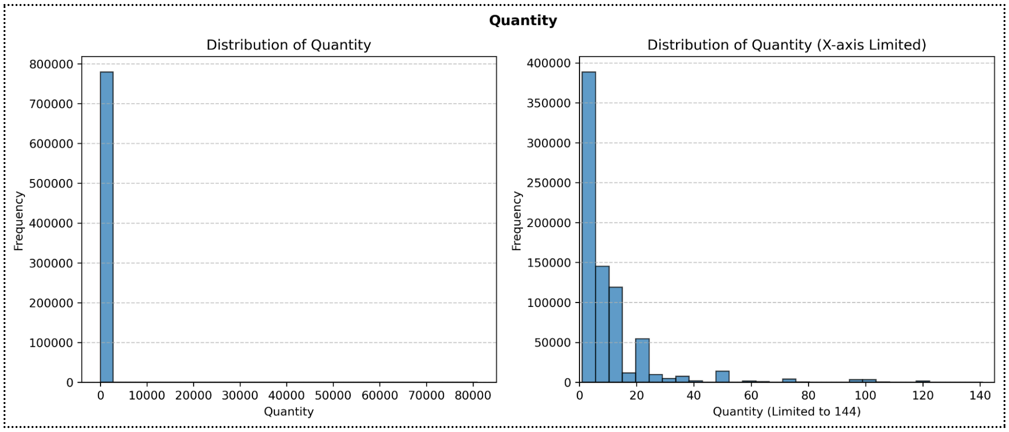
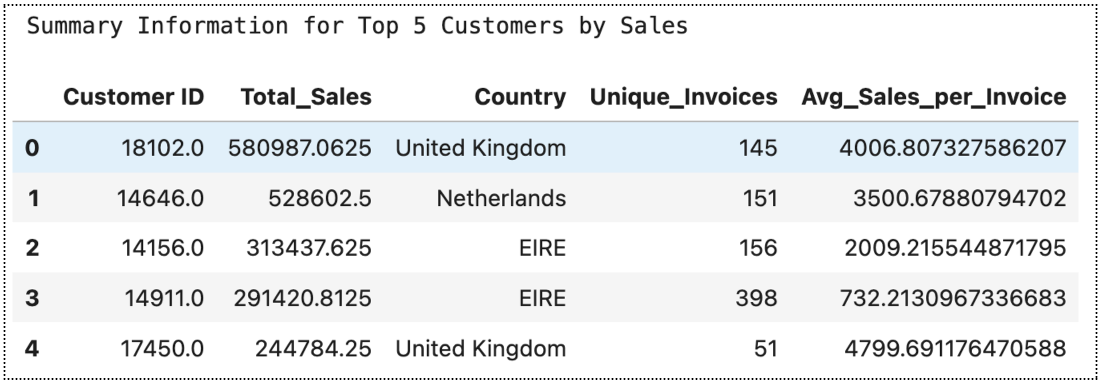


Figure 3: (left) Distribution of all 'Quantity' values; (right) Distribution of 'Quantity' limited to Quantity < 144 which represents 99% of data.

'Customer ID'

'has_id' has 5,878 unique customers. Figure 4 shows summary data for the top five customers when sorted by total 'Sales'. The top five customers account for 11.3% of total 'Sales'.



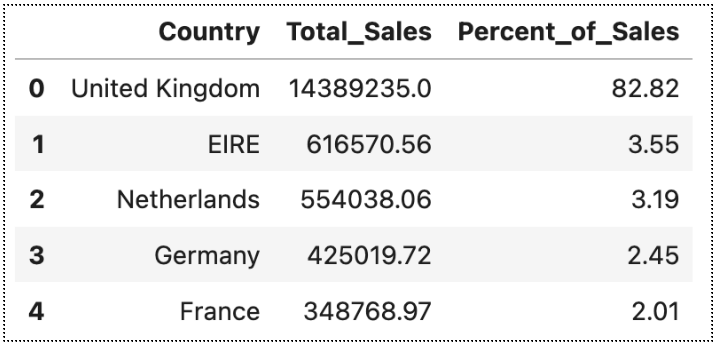
Summary Information for Top 5 Customers by Sales

	Customer ID	Total_Sales	Country	Unique_Invoices	Avg_Sales_per_Invoice
0	18102.0	580987.0625	United Kingdom	145	4006.807327586207
1	14646.0	528602.5	Netherlands	151	3500.67880794702
2	14156.0	313437.625	EIRE	156	2009.215544871795
3	14911.0	291420.8125	EIRE	398	732.2130967336683
4	17450.0	244784.25	United Kingdom	51	4799.691176470588

Figure 4: Summary data for top five customers ranked by total 'Sales'

'Country'

There are 41 countries represented in the 'has_id' dataset. The company is based in the United Kingdom and 90.7% of invoices are from the UK but only 82.8% of sales come from the UK. The top five countries ranked by 'Sales' are shown in Figure 5.



	Country	Total_Sales	Percent_of_Sales
0	United Kingdom	14389235.0	82.82
1	EIRE	616570.56	3.55
2	Netherlands	554038.06	3.19
3	Germany	425019.72	2.45
4	France	348768.97	2.01

Figure 5: Top 5 entries for 'Country' ranked by 'Total_Sales' and percent of total sales each country represents

'InvoiceDate' and 'YearMonth' Sales Trends

'YearMonth' was created to indicate the year and month the transaction took place. It was then used to look at sales trends. Figure 6 looks at the total number of 'Invoice' entries per month and total 'Sales' per month.

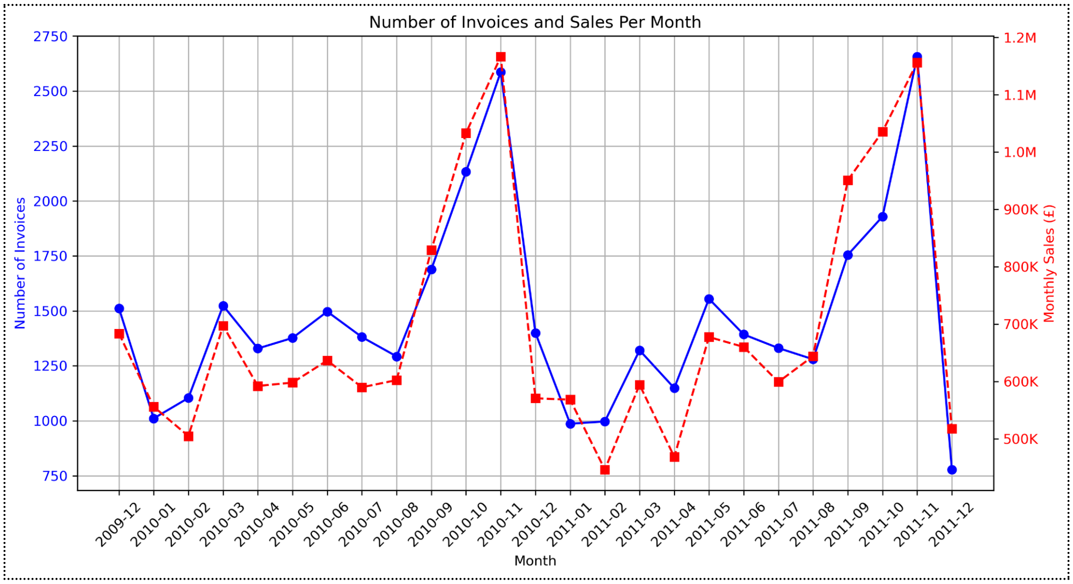


Figure 6: Total 'Sales' and number of 'Invoice' per month

Monthly sales increase in the lead up to the holidays with a lull in sales in January and February which is consistent with the company's primary products being gift items.

2. Feature exploration grouped by 'Invoice'

The features were re-examined after grouping by the 'Invoice' number. This didn't provide much additional information as most features were still heavily skewed towards low values with numerous high values.

Summary statistics give some additional insight into purchase behavior but not much.

- 'StockCode':
  - There were an average of 20.5 unique 'StockCode' values per invoice and a median of 15.
  - There was a range of 1 to 541 unique 'StockCode' entries per 'Invoice'
- 'Quantity':
  - There was an average of 484.4 total units sold per 'Invoice' and a median of 151 units sold.
  - There was a range of 1 to 87,167 units sold per 'Invoice'.
- 'Sales':
  - The average 'Sales' per invoice was £469.98 and a median of £303.40
  - 'Sales' per invoice ranged from £0.38 to £168,469.59

The correlation between features was also examined. Figure 7 shows the correlation heatmap between the numeric features. There is a slight positive correlation between the number of unique 'StockCode' and total 'Sales' and total 'Quantity' as you would expect the number of items and total sales to increase with each additional unique item purchased. The largest correlation was between 'Quantity' and 'Sales'. There was a strong positive (0.65) correlation between the total 'Quantity' of items purchased in the invoice and and the total 'Sales' of the invoice.

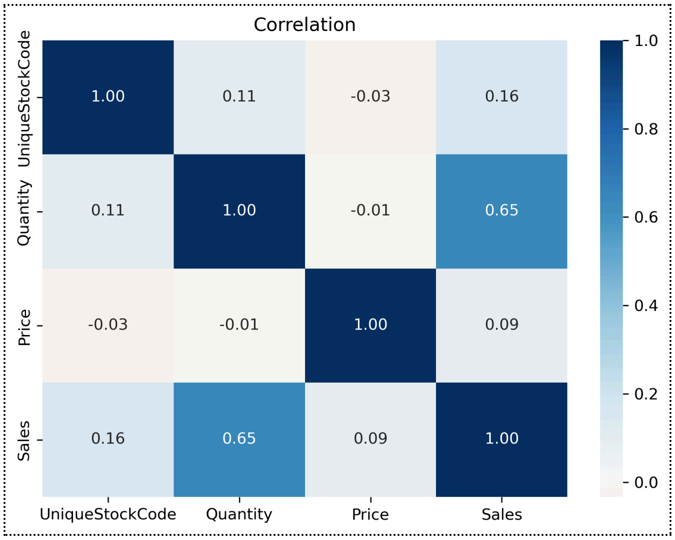


Figure 7: Correlation of features grouped by 'Invoice'

3. Recency, Frequency, Monetary (RFM)

Recency, frequency, and monetary values were calculated after grouping by 'Customer ID'. Recency was calculated by subtracting the 'InvoiceDate' from the day after the last date in the dataset. Frequency was calculated by counting the number of unique 'Invoice' values in the dataset. 'Monetary' could have also been called lifetime value as it was calculated by the total sum of 'Sales'. Recency has an inverse relationship to the other two measures as a lower value is better since it means the customer purchased more recently.

There are 5,878 unique 'Customer ID' in the data set.

Recency - days since last purchase

There was a range of 1 day to 739 days since the customer's last purchase. The average number of days since last purchase was 201.3 and a median of 96 days. The last day in the dataset was December 9, 2011 which is right at the end of the holiday buying rush. Figure 8 shows that a large number of customers have made a purchase within the last 100 days which is the holiday buying season with another bump around 400 days ago which would be the buying season for last year.

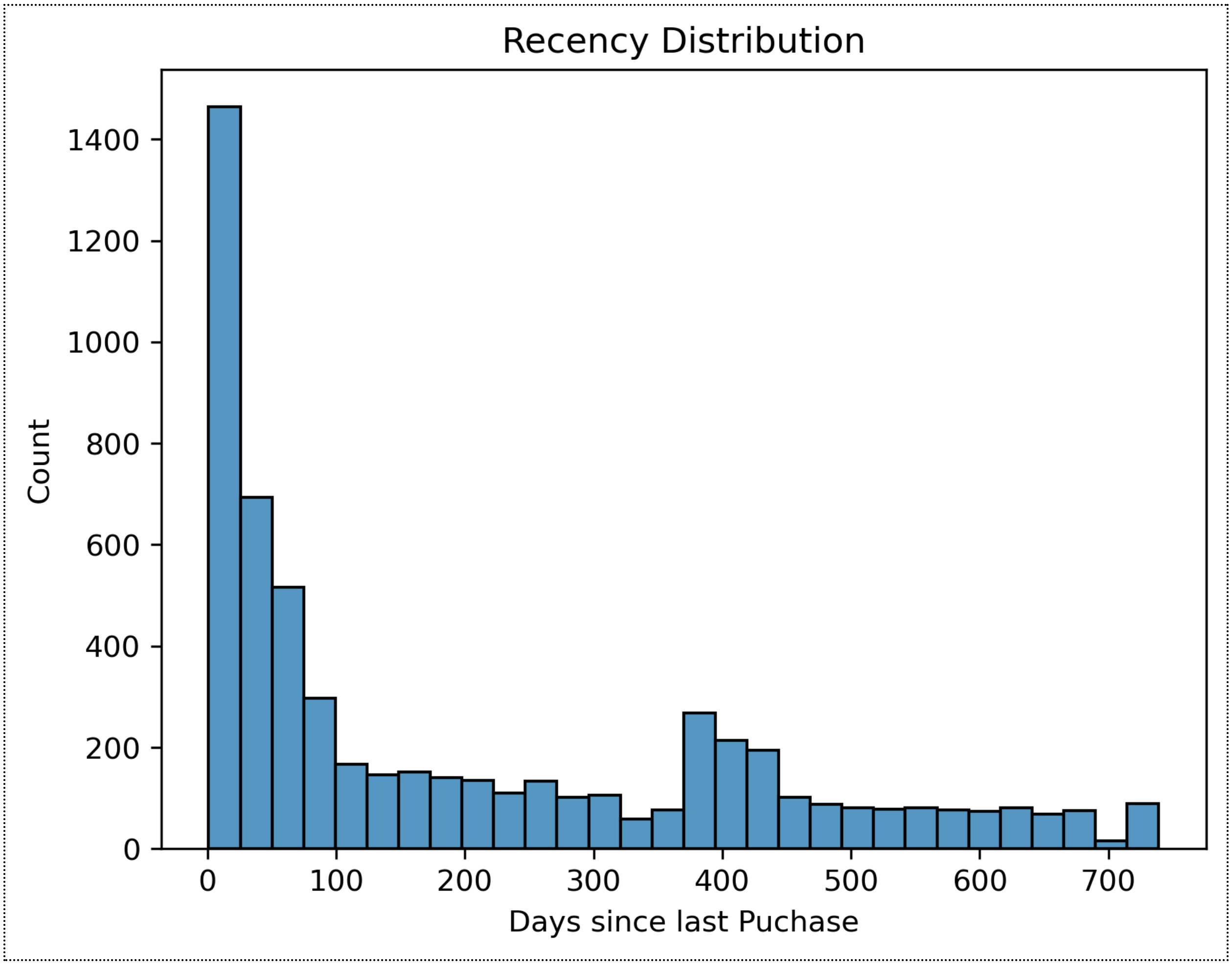


Figure 8: Regency distribution representing days since last purchase

Frequency - number of purchases

There was a range of 1 to 398 purchases with an average of 6.38 purchases per 'Customer ID' and a median of 3 purchases. Only 1% of 'Customer ID' (58 customers) had more than 46 purchases and 85% of customers had 10 or less purchases. Figure 9 shows the distribution of 'Frequency' values.

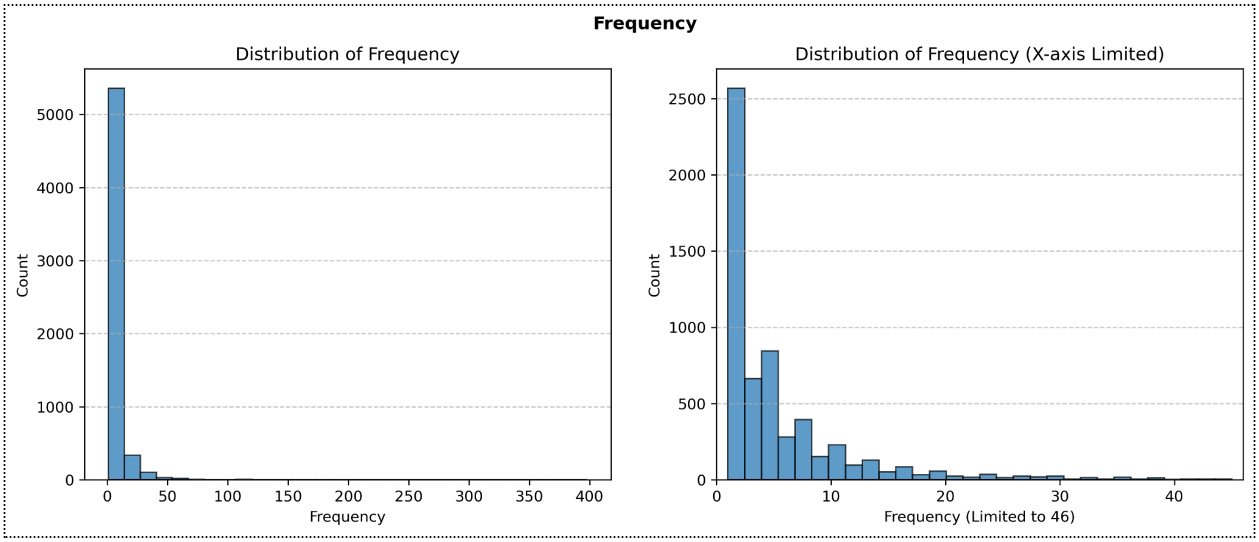


Figure 9: (left) Distribution of all 'Frequency' values; (right) Distribution of 'Frequency' limited to 46 which represents 99% of data.

Monetary - total amount spent

The total amount spent by 'Customer ID' ranged from £2.95 to £580,987.06. The average amount spent by 'Customer ID' was £2,955.90 and a median of £867.74. The distribution of 'Monetary' values in Figure 10 shows that the majority of customers are in the lower values of Monetary with 96% being under £10,000 in sales.

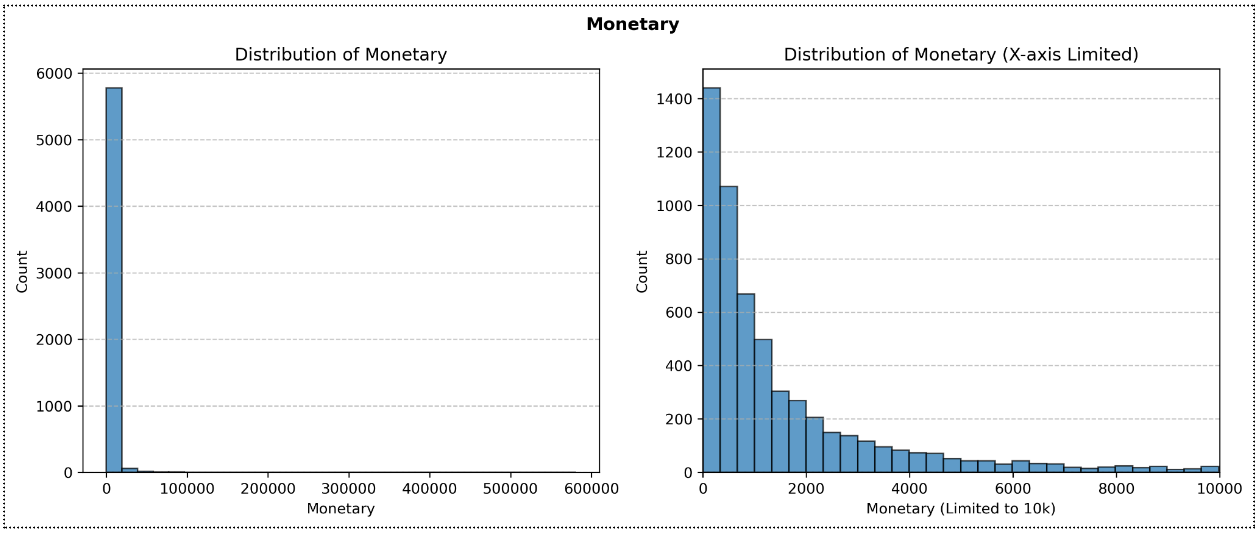


Figure 10: (left) Distribution of all 'Monetary' values; (right) Distribution of 'Monetary' limited to 10,000 which represents 96% of data.

Figures 11 and 12 look at the relationships between the RFM features. Figure 11 is a pairplot that has the distribution on the diagonal and scatterplots to show the relationship between the given features.

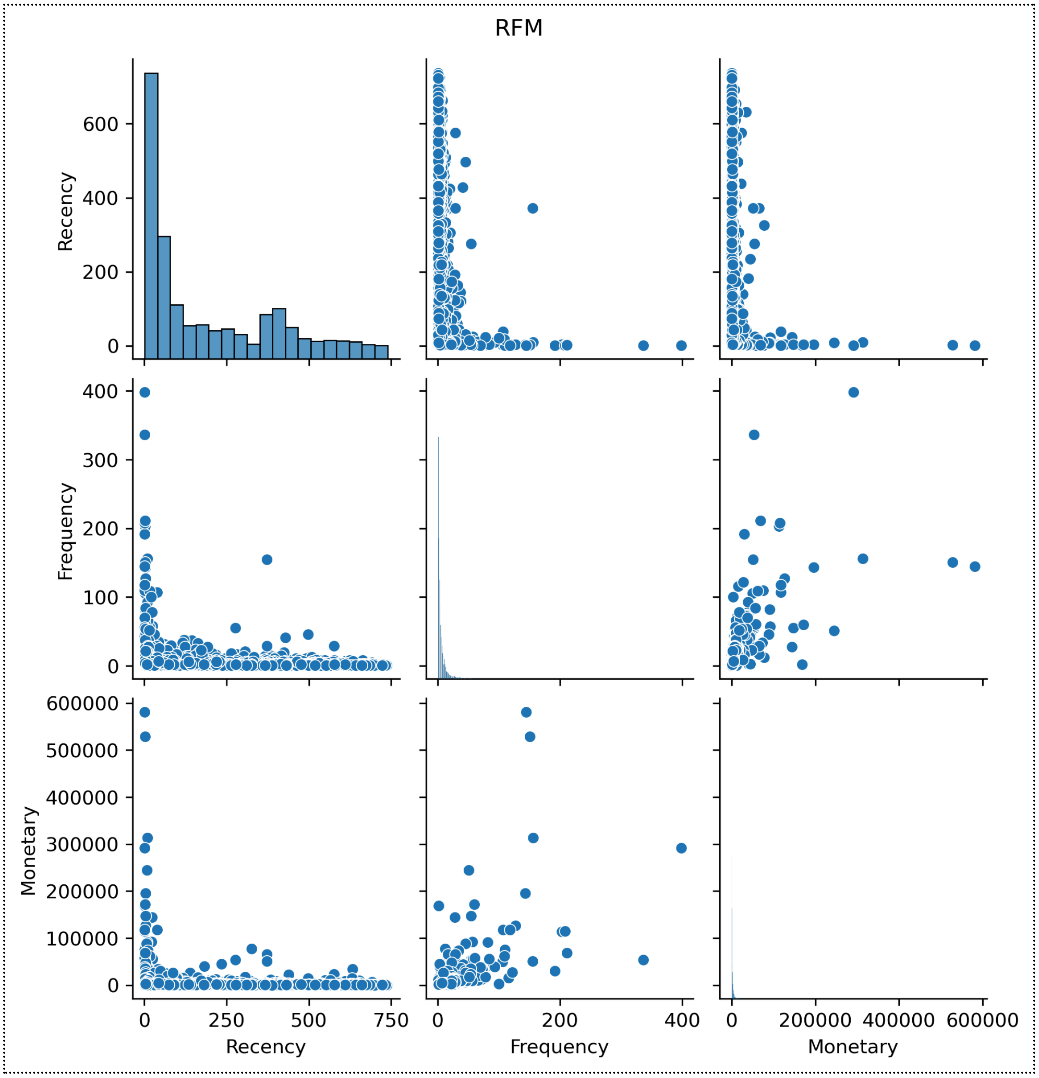


Figure 11: RFM pairplot with scatter plots of features and distribution on the diagonal.

Figure 12 shows the correlation between the features. The correlation between 'Frequency' and 'Monetary' suggests that when customers make more purchases they are going to spend more money *but the 0.63 values means that there are other factors that will impact that relationship.*

The negative correlation for 'Recency' is caused by the inverse relationship it has with the other two variables so a negative correlation is actually a positive relationship. The 0.26 (inverse) correlation between 'Recency' and 'Frequency' suggests that customers with more recent purchases tend to have more frequent purchases but it is greatly impacted by other factors.

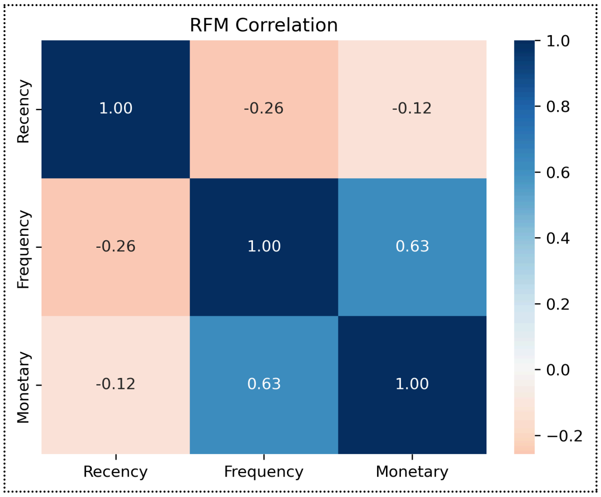


Figure 12: RFM correlation

Modeling

Model Preparation

All models and packages used were from sklearn.

The initial modeling used KMeans with two datasets. Model 1 used the 'rfm' dataset created by RFM calculations. Model 2 used an expanded version of the 'rfm' dataset called 'rfm_expoanded'.

'rfm_expoanded' created three additional features: Average Order Value ('AOV'), Customer Lifetime ('Lifespan'), and Customer Lifetime Value ('CLV'). 'AOV' was calculated from dividing 'Monetary' by 'Frequency' from the 'rfm' dataset. 'Lifespan' was calculated using the 'has_id' dataset, grouping by 'Customer ID' then subtracting the last 'InvoiceDate' from the first 'InvoiceDate' to get the number of days between the two values and adding 1. The addition of 1 was to account for the 1 that was added to the 'Recency' calculation in 'rfm'. CLV was calculated by multiplying 'Monetary' times 'Lifespan'. This was a mistake as 'Monetary' in this case was the equivalent of CLV in this case since all 'Sales' were included in the calculation of 'Monetary'.

For both models the data was preprocessed using PowerTransformer to help with the skewness of the data and StandardScaler to account for the difference in magnitudes of the data. Another error was made here by not inverting the scaled 'Recency' values since it has an inverse relationship to the other values. the scaled data was saved as a new dataset by adding '_scaled' to the end of the original dataset name and the scaled dataset was used in its respective model. Using values of k=2 through k=9, the elbow method and silhouette score were used to determine the optimal k value.

Model Findings

The initial findings showed that Model 1 using just RFM values with a k=2 was the best option for the data. However two market segments do not provide much information, so the next best option was picked which was Model 1 k=4. Figure 13 shows the Elbow method graph for Model 1 and Figure 14 shows the silhouette score for the respective k.

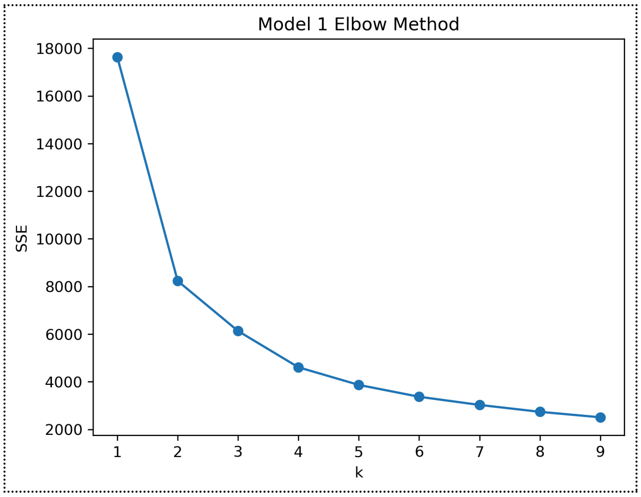
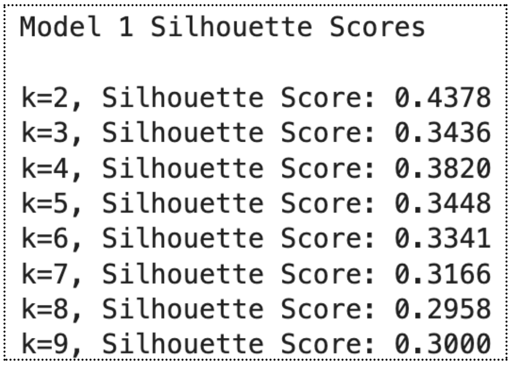


Figure 13: Model 1 Elbow method graph



```
Model 1 Silhouette Scores

k=2, Silhouette Score: 0.4378
k=3, Silhouette Score: 0.3436
k=4, Silhouette Score: 0.3820
k=5, Silhouette Score: 0.3448
k=6, Silhouette Score: 0.3341
k=7, Silhouette Score: 0.3166
k=8, Silhouette Score: 0.2958
k=9, Silhouette Score: 0.3000
```

Figure 14: Model 1 silhouette scores

A silhouette score of 0.44 for the optimal k=2 and 0.38 for the second best option of k=4 are not ideal. The scores suggest moderate clustering strength but that there is overlap within the clusters.

Note: Post-segmentation EDA (next section) confirmed that there is a lot of variation in the customer segments. I went back and tried to find a better model with no luck. I used different combinations of models and features including trying different 'Monetary' calculations for rolling 1, 3 and 6 month 'Sales' totals and averages. I also tried sub-clustering the clusters from Model 1 k=2 to see if there was an added benefit but there were no higher silhouette scores and multiple subclusters were nearly identical. The next best option was a hierarchical clustering model that showed the optimal number of clusters was 4 but it has a lower silhouette score (0.34) than either of the k values for Model 1.

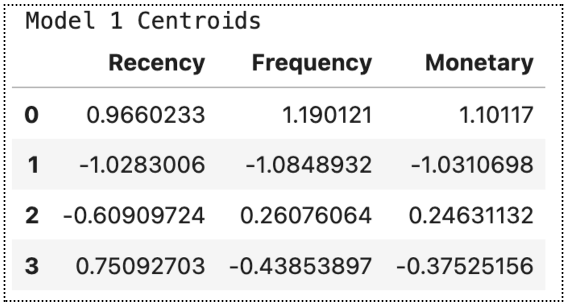
Post Segmentation EDA

Post segmentation EDA explored the variations in the clusters to develop customer profiles for each cluster. It also explored the Key Questions asked in the Problem Statement above

Centroids

The first values explored after segmentation were the centroids for the clusters found in Model 1. The goal was to get a preliminary idea of the characteristics for each cluster

Note on centroid values: Since the 'Recency' value has an inverse relationship to the other features and that was not accounted for earlier, 'Recency' values were multiplied by -1 to make them consistent with the other values and to make the relationships more logical. Because the data was standardized, the centroid values in Figure 15 represent the center of the cluster compared to the mean value in the dataset.



Model 1 Centroids

	Recency	Frequency	Monetary
0	0.9660233	1.190121	1.10117
1	-1.0283006	-1.0848932	-1.0310698
2	-0.60909724	0.26076064	0.24631132
3	0.75092703	-0.43853897	-0.37525156

Figure 15: Model 1 centroid values

Centroid analysis suggest the following preliminary characteristics of the clusters:

- Cluster 0 are the most desirable customers. They have the most recent purchases, the highest number of purchases, and have spent the most money.
- Cluster 1 are the least desirable customers. They have spent the least, made the fewest purchases, and haven't purchased very recently.
- Cluster 2 are lapsed customers who haven't purchased in a while but were semi-regular customers at one point as their frequency and monetary values were higher than average.
- Cluster 3 are newer customers. They have purchased second most recently but have fewer purchases and have a lower total spend.

Figure 16 confirms these initial characteristics from the centroids as well as the overlap and looser clusters expected from the lower silhouette scores. The top row represents all data and the bottom row has the axes limited to remove extreme values to better visualize the details.

Cluster 0, represented by the blue dots, has the highest 'Monetary' and 'Frequency' values with the lowest 'Recency' values (inverse relationship). Cluster 1, represented by orange, has the lowest 'Frequency' and 'Monetary' and the highest 'Recency'. Cluster 2, represented by green, is still low on 'Frequency' and 'Monetary' but higher than Cluster 1 and shifted higher on 'Recency' that Cluster 0 but still lower than Cluster 1. Cluster 3, represented by red, is focused in the bottom left corner indicating lower 'Monetary' and 'Frequency' but more recent purchases with the lower 'Recency'.

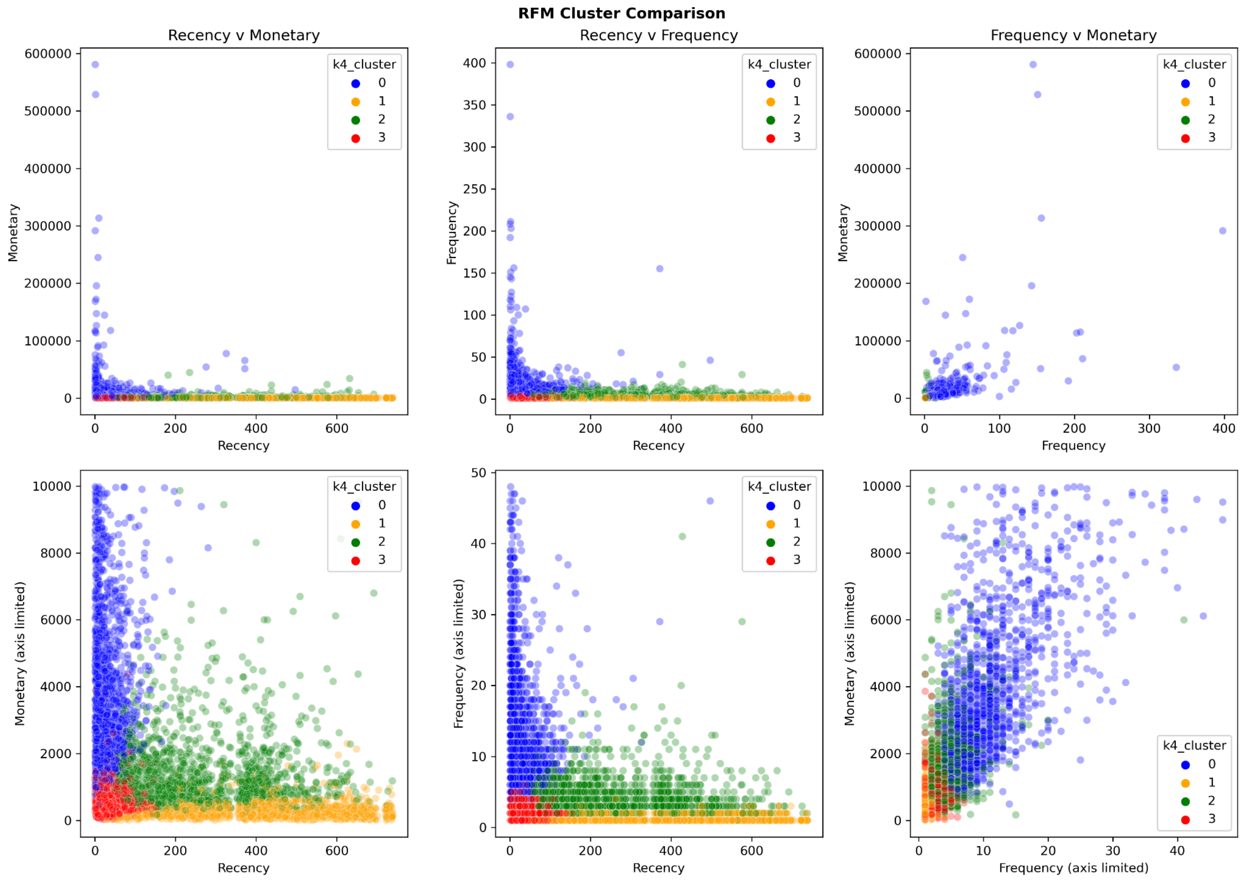


Figure 16: RFM scatterplots with colors representing Cluster . (top row) Scatter plots with all data. (bottom row) Scatter plots with axes limited to remove extreme values (data excluded: 0% of 'Recency', 1% of 'Frequency', and 4% of 'Monetary' values).

Key Questions from Problem Statement

Note: The variation in the loose segments makes giving definitive insights here difficult. Instead, a range of values that represents the middle 50% of customers (25th to 75th percentiles) will be given plus the minimum and maximum values, the mean and median values, and any notable values that are specific to each segment.

- How often do each segment make purchases? ('Frequency')
    - Cluster 0
        - 7 to 17 purchases for the middle 50% with median of 11 purchases and mean of 15.78 purchases.
        - 96.6% of customers have made 5 or more purchases and 55.9% made more than 10 purchases.
    - Cluster 1

    - 77.8% of customers made 1 purchase with a mean of 1.28 purchases
    - No customers have made more than 4 purchases.
  - Cluster 2
    - 3 to 5 purchases for the middle 50% with a median of 4 purchases and a mean of 4.49.
    - 37.8% of customers have made more than 5 purchases but only 4.2% made more than 10 purchases.
  - Cluster 3
    - 1-3 purchases for the middle 50% with a median of 2 purchases and a mean of 2.37.
    - 4.6% of customers have made more than 5 purchases but none have made more than 7 purchases

Figure 17 shows the distribution of the number of purchases by each cluster. Cluster 0 makes significantly more purchases than any other cluster.

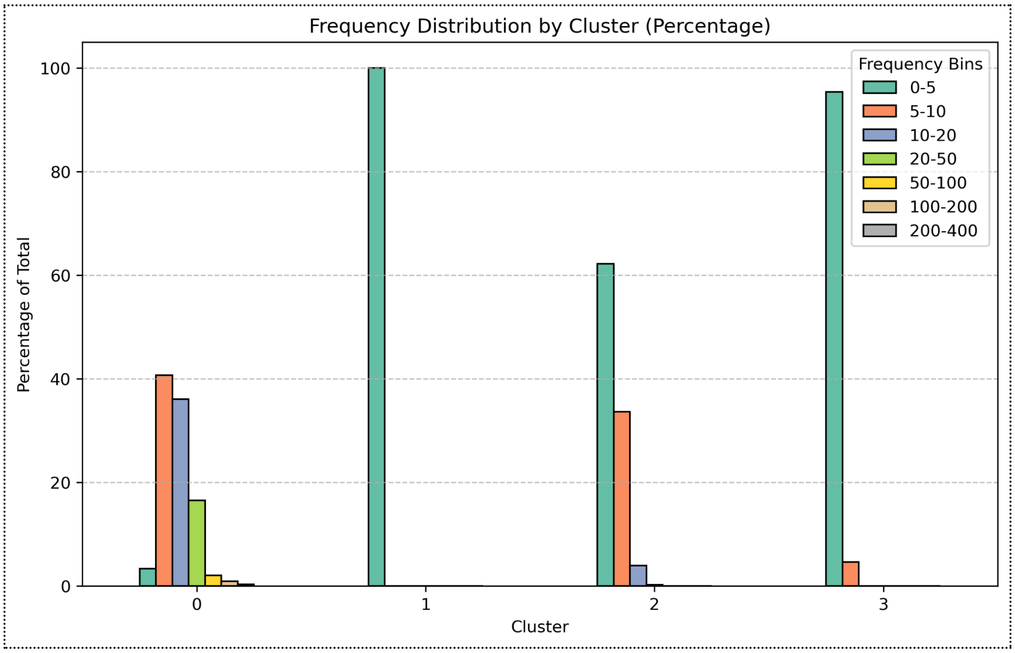


Figure 17: Frequency Distribution by Cluster

Figure 18 confirms this by showing the percent of total purchases each cluster made (by number of purchases).

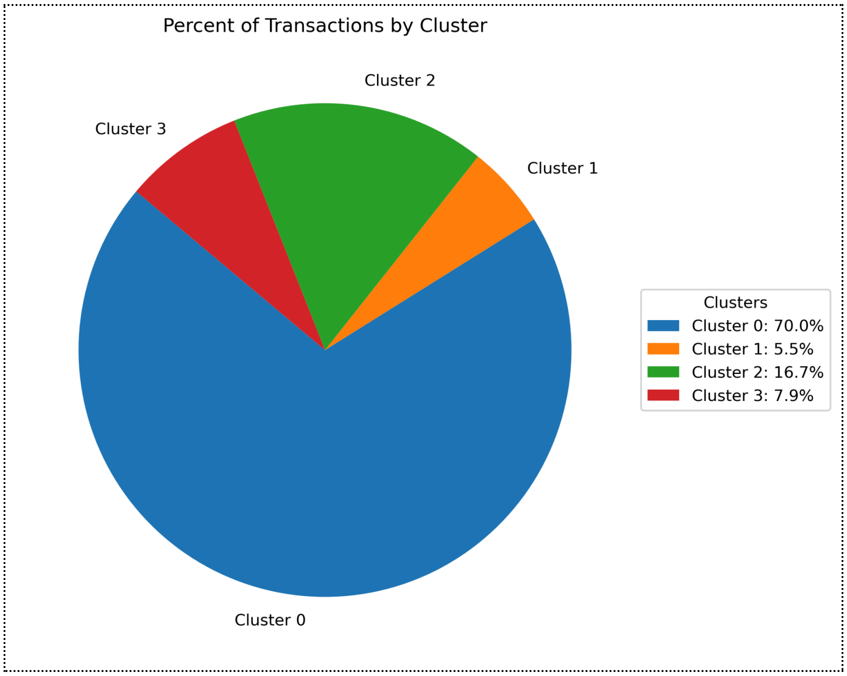


Figure 18: Percent of Transactions

- How much have they spent over time? ('Monetary')
  - Cluster 0
    - The middle 50% spent between £2,268 and £6,948 with a median of £3,770 spent and mean of £8,421.
    - 98.4% of customers have spent more than £1,000, 37.1% have spent more than £5,000, and 15.2% have spent more than £10,000.
    - 0.9% of customers have spent more than £100,000 with the highest amount being £580,987.
  - Cluster 1
    - The middle 50% spent between £142 and £374 with a median of £ 240 and mean of £294.
    - only 2.1% of customers have spent more than £1,000
  - Cluster 2
    - The middle 50% spent between £783 and £1,844 with a median of £1,168 spent and mean of £1,653.
    - 60.1% of customers have spent more than £1,000, 2.6% have spent more than £5,000, and 0.9% have spent more than £10,000
  - Cluster 3
    - The middle 50% spent between £342 and £887 with a median of £571 spent and mean of £662.
    - 18.8% of customers have spent more than £1,000

Figure 19 shows the portion of all sales each cluster represents. Cluster 1 represents 79.5% of all sales made.

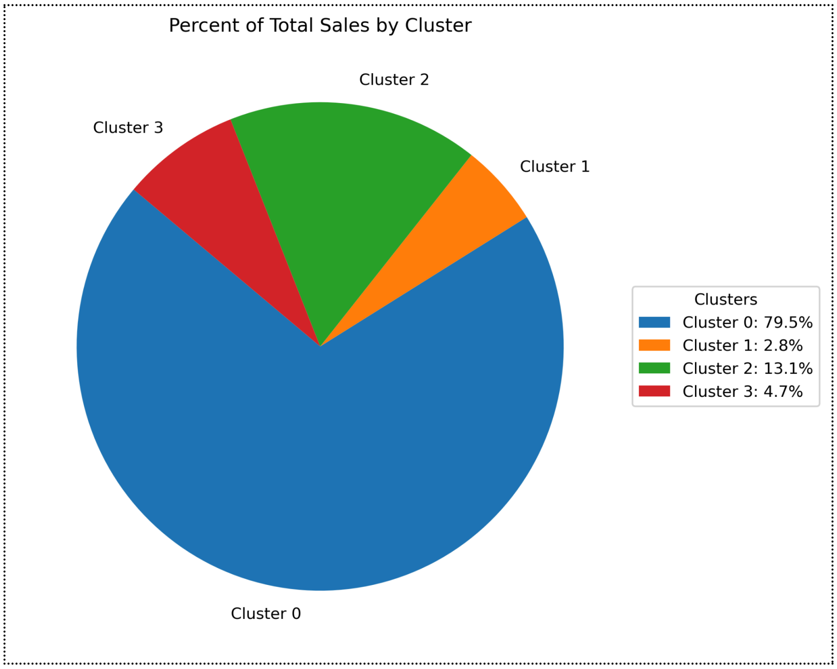


Figure 19: Percent of total Sales by Cluster

- How recently have they made a purchase ('Recency')
    - Cluster 0
        - The middle 50% have made a purchase in the last 8 to 42 days with a median of 20 days and a mean of 31 days.
        - 94.3% of customers have made a purchase in the last 90 days, 85.8% in the last 60 days and 64.7% in the last 30 days
    - Cluster 1
        - The middle 50% have made a purchase in the last 302 to 573 days with a median of 423 days and a mean of 431 days.
        - only 2.3% of customers have made a purchase in the last 90 days, 0.6% in the last 60 days and 0.1% in the last 30 days
    - Cluster 2
        - The middle 50% have made a purchase in the last 150 to 293 days with a median of 248 days and a mean of 276 days

        - 10.6% of customers have made a purchase in the last 90 days, 2% in the last 60 days and 0% in the last 30 days
    - Cluster 3
        - The middle 50% have made a purchase in the last 17 to 58 days with a median of 32 days and a mean of 40 days
        - 94.0% of customers have made a purchase in the last 90 days, 77.4% in the last 60 days and 47.6% in the last 30 days

Figure 20 shows the Recency distribution by Cluster. It is separated into groups of 30 days or less, 31-60 days, 61-90 days, 4 months to 1 year, and 1 plus year since last purchase. Clusters 1 and 3 have the most recent purchases with almost all purchases coming in the last 6 months whereas Cluster 1 has almost all purchases over 6 months.

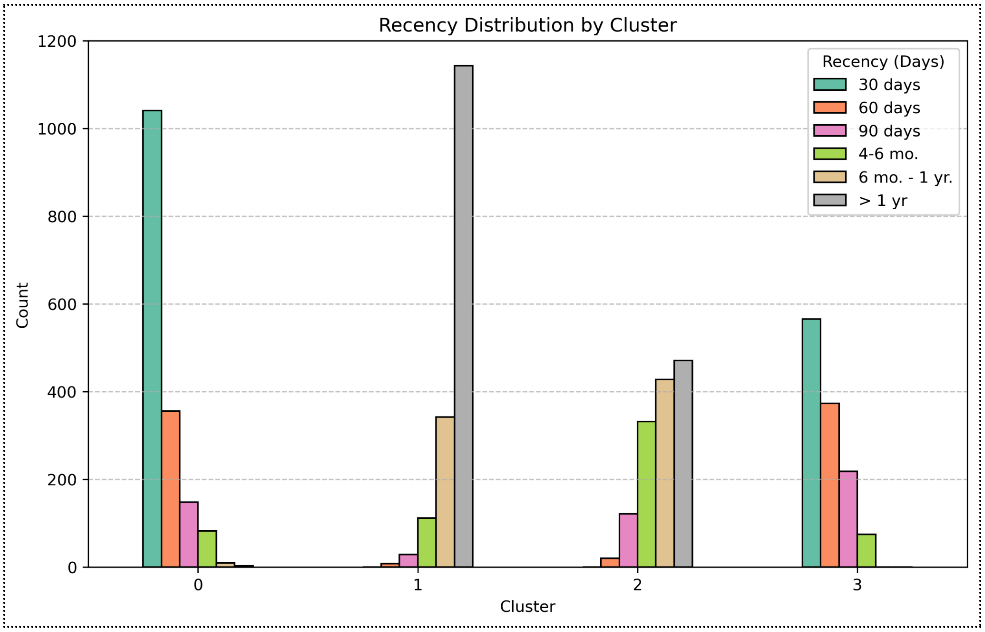


Figure 20: Recency distribution by Cluster

- What does their basket look like? ('AOV', 'Adjusted AOV', 'Price', 'Quantity')
    - ★ Note on 'Adjusted AOV': 'Adjusted AOV' is 'AOV' divided by 'Frequency'. It helps account for when the customer has a high frequency that is inflating the average order.
    - 'Price'

    - The middle 50%, median, and mean were nearly identical for all Clusters
- 'Quantity'
    - 'Quantity' purchased was nearly identical across all clusters for 80% of customers. For the top 20% of customers it stayed similar for Clusters 1 and 3 and for Clusters 0 and 2 with the latter purchasing more.
- Cluster 0
    - 'AOV'
        - The middle 50% spend £249 to £509 on average per purchase with a median AOV of £351 and a mean AOV of £515. All of these represent the highest values amongst the clusters.
    - 'Adjusted AOV'
        - The middle 50% have an 'Adjusted AOV' between £18 and £55 with a median 'Adjusted AOV' of £32 and a mean of £76. All of these represent the lowest values amongst the clusters.
- Cluster 1
    - 'AOV'
        - The middle 50% spend £124 to £311 on average per purchase with a median AOV of £197 and a mean AOV of £255. All of these represent the lowest values amongst the clusters.
    - 'Adjusted AOV'
        - The middle 50% have an 'Adjusted AOV' between £98 and £309 with a median 'Adjusted AOV' of £159 and a mean of £239. All of these represent the highest values amongst the clusters.
- Cluster 2
    - 'AOV'
        - The middle 50% spend £199 to £445 on average per purchase with a median AOV of £306 and a mean AOV of £443.
    - 'Adjusted AOV'
        - The middle 50% have an 'Adjusted AOV' between £41 and £146 with a median 'Adjusted AOV' of £72 and a mean of 167£.
- Cluster 3
    - 'AOV'

        - The middle 50% spend £169 to £371 on average per purchase with a median 'AOV' of £256 and a mean of £320.
    - 'Adjusted AOV'
        - The middle 50% have an 'Adjusted AOV' between £65 and £219 with a median 'Adjusted AOV' of £119 and a mean of £200.

- Are there differences in seasonality of purchases between segments? ('Sales' by 'MonthYear')

  Figure 21 shows Monthly sales for each Cluster. Cluster 1 has the most seasonality with sales growing and peaking in the lead up to the December holiday season. Cluster 3 being the newest customers shows similar peaks for the most recent holiday season but at a lower magnitude.

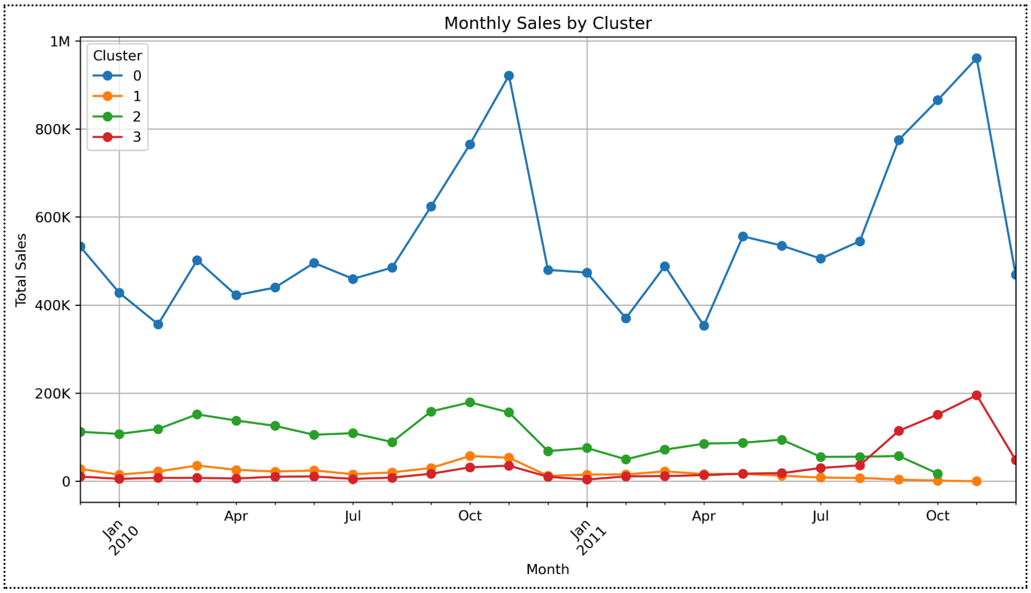


Figure 21: Monthly sales by Cluster

Customer Segments

Cluster 0: Loyal, Frequent Customers

These customers are the company's main customers. They represent 70% of transactions and 79.5% of sales. Nearly all customers (99.8%) in this segment have made a purchase in the last year with 94.3% making a purchase in the last 90 days. These customers have the highest portion of repeat business with 96.6% of customers making 5 or more purchases and 55.9%

making more than 10 purchases. They have the lowest adjusted AOV which indicates that while they make frequent purchases the average amount they spend per purchase is actually lower.

Cluster 1: One time customers

These customers are mostly one time customers who do not return. 77.8% have only made one purchase and 98.9% make 2 or less. Only 30.1% of these customers have made a purchase in the last year and that drops to 2.3% for purchases in the last 90 days. They have the lowest AOV and lowest total amount spent but the highest adjusted AOV indicating that the few purchases they did make were larger in value.

Cluster 2: Occasional Customers / Churn Risk

These customers are repeat customers but haven't purchased in a while and are at risk of churning or have already churned. 66.2% of these customers have made a purchase in the last year but only 34.5% in the last 6 months and that drops further to 10.6% in the last 90 days. These customers have the second highest AOV and second lowest Adjusted AOV indicating they used to be regular customers but no longer.

Cluster 3: Seasonal or New Customers

These are a mix of seasonal and newer customers. 53% of these customers made their first purchase in the last 6 months. The last 6 months also coincide with the lead up to the holiday buying season and 100% of customers in this segment have made a purchase over that timeframe.

Recommendations and Follow-up

- Focus on clusters 0 and 3.
  - Cluster 0 are loyal customers with the highest spend and account for 70% of sales. Building and maintaining these relationships are the key to company success.
  - Cluster 3 provides the greatest opportunity for growth. If the company can build a relationship with the new customers in this segment, there is a possibility of turning them into year round customers. Alternatively, they could focus their marketing efforts to these customers in the July and August in the build up to next year's holiday season to ensure they return.

    - The original Online Retail II dataset has limited data about customer demographics and the types of items they are buying. Further segmenting customers with the full set of information would allow for more nuanced recommendations based on more specific details of the customer.
- Explore customers that purchase year round.
    - Most of the company's sales come in the lead up to the holiday season. Building the relationship with customers that purchase year round can help hedge against the lulls in the sales.
- With the amount of variability in the data, a DBSCAN model may provide better clustering options than either the K-Means and Hierarchical models tried.