RFM ANALYSIS ON ONLINE RETAIL DATASET

Group 2:

1- Ahmed Abdelaal
2- Amira Abu Issa
3- Motaz Habib
4- Yousef Mohamed

Abstract

We aimed to understand our customers' behavior, target the right demographics, and deliver value in the most efficient way possible. To this end, we utilized RFM (recency, frequency, and monetary) analysis, creating an RFM score for each customer to identify patterns and detect the most valuable clients. This analysis allowed us to segment our customers into nine distinct categories, each representing varying degrees of customer loyalty or risk of churn. To validate our manual segmentation, we applied k-means clustering to our RFM scores, using the elbow method and silhouette analysis to determine the optimal number of clusters. These methods suggested three and two clusters respectively, and their comparison with our manual segmentation demonstrated strong alignment. This integration of machine learning with traditional RFM analysis has effectively refined our customer segmentation process, driving targeted strategies for our varied customer segments.

Introduction

Understanding what the consumer wants and needs is important in every business. Because every customer is unique; some buy more, and others buy less. While some clients come back again, others might churn. Therefore, it's crucial to identify which clients are the most valuable and learn how to satisfy them. For our assignment, we examined an online retail dataset that contains transactions between 2010 and 2011. Two key tools were used to further our understanding of the customers. The Recency, Frequency, Monetary (RFM) model was first used. This is an easy approach to examine clients based on when they most recently made a purchase (Recency), how frequently they make purchases (Frequency), and how much money they spend overall (Monetary). We were able to identify the key clients and understand their purchasing patterns thanks to this methodology. K-means clustering was used after that. Based on similar purchase habits, clients are grouped together. We compared these groups to the RFM model to understand the customers better and come up with more effective and reliable approaches to satisfy them. By integrating the RFM model and k-means clustering, we hope to uncover novel customer insights that will benefit the company's success.

System’s Architecture

1. Data Collection:

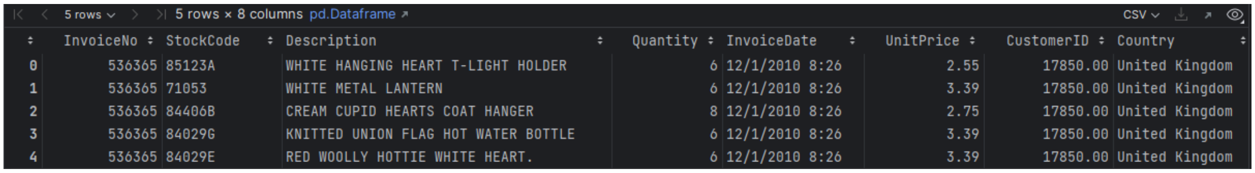
We used an online retail dataset from UCI Machine Learning Repository that contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts. Many customers of the company are wholesalers.

The dataset has eight features:

1. **InvoiceNo**: This is the unique number given to each transaction. If it starts with 'c', it means the transaction was cancelled.
2. **StockCode**: This is a unique code for each product. It's like an ID number so we know exactly what product we're talking about.
3. **Description**: This is the name of the product.
4. Quantity: how many of each product were bought in the transaction.
5. **InvoiceDate**: This is the date and time when the transaction happened.

6. **UnitPrice**: This is the price for each unit of the product.

7. **CustomerID**: This is a unique number for each customer. Like the StockCode, it's like an ID number for each customer.
8. **Country**: This is the country where the customer lives.



	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.00	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.00	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.00	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.00	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.00	United Kingdom

2. Data Preprocessing:

We took a close look at our data. We found that a significant portion of the data in the 'CustomerID' column was missing - almost 25%. That's quite a lot of missing information, and it's particularly challenging for us, as our analysis is focused on customers. Therefore, we made the decision to remove these entries. We also identified some missing values in the 'Description' column, but these were automatically addressed when we eliminated records with absent 'CustomerID' information.

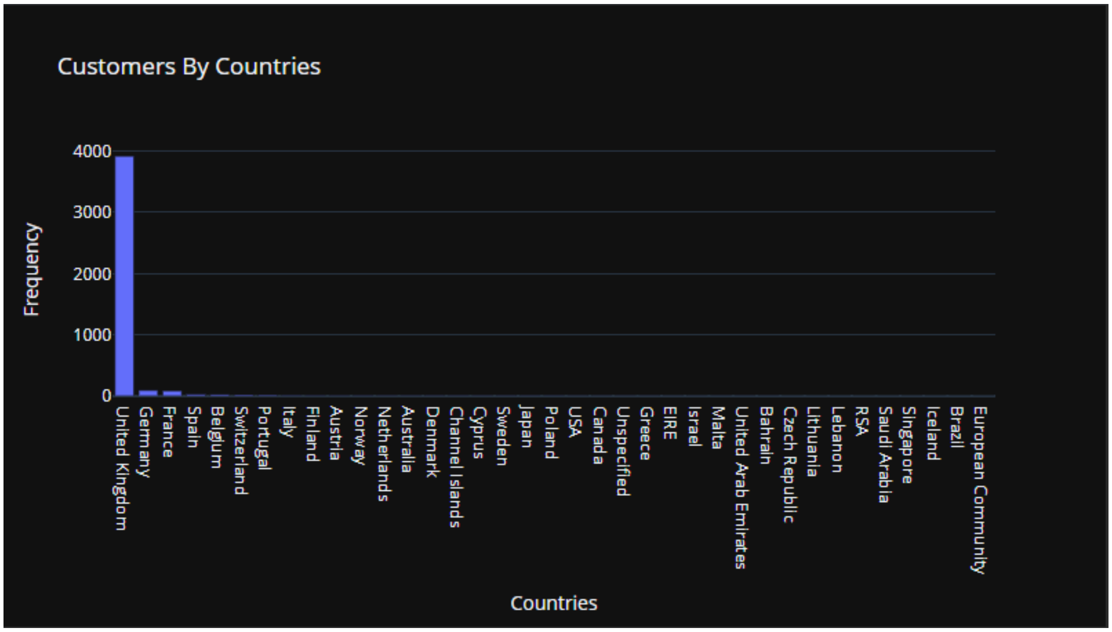
Next, we dealt with the issue of duplicate records. We identified 4879 of them, which accounted for less than 1% of the total records., we decided to remove these duplicates.

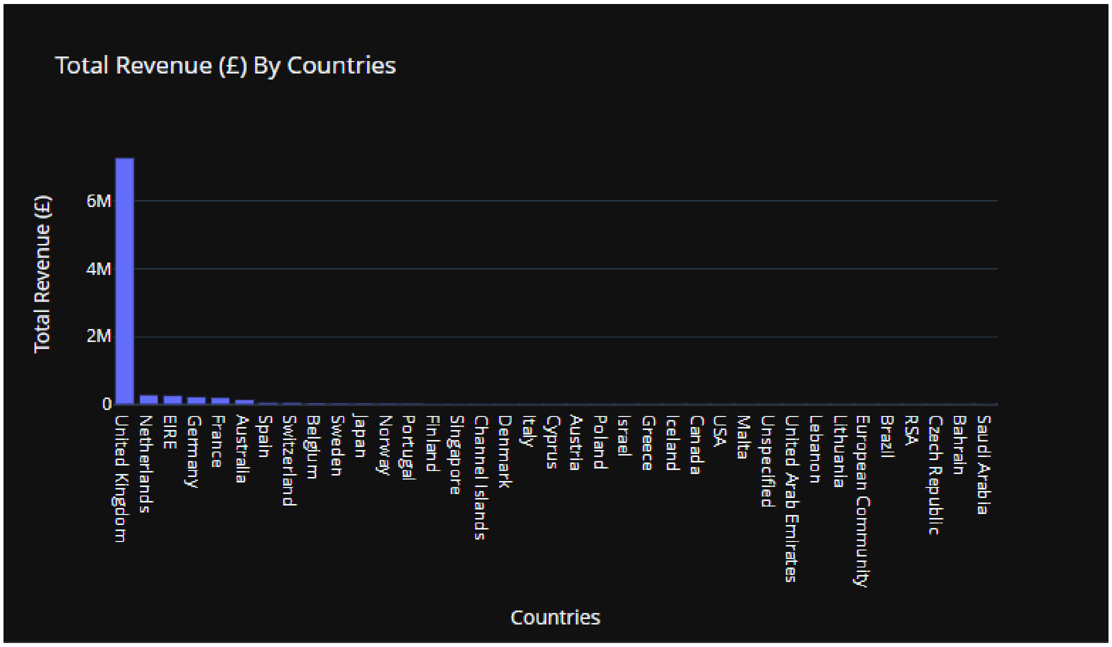
We noticed that some records had negative values in the 'UnitPrice' and 'Quantity' columns. On investigating further, we found that these anomalies were tied to cancelled orders, which were indicated by an 'InvoiceNo' starting with a 'C'. These cancelled orders made up 14% of all the orders - a substantial proportion. However, since our analysis focuses on successful transactions, we decided to exclude these records. Additionally, to avoid any misleading information, we checked for any regular orders that matched these cancelled ones, but we didn't find any such instances.

3. Data Exploratory**:**

When we first looked at our orders, we noticed something strange. The number of unique 'StockCode' items didn't match the number of unique 'Description' items. But since we are focusing on customers, not products, this mismatch won't change our analysis.

Next, we examined the demographic distribution of our customers and revenue. We found that 90% of our customers are based in the UK, and these customers generate 82% of our total revenue. Considering this strong demographic and economic concentration, we decided to focus our analysis on the UK market and set aside the other regions.





4. RFM Analysis:

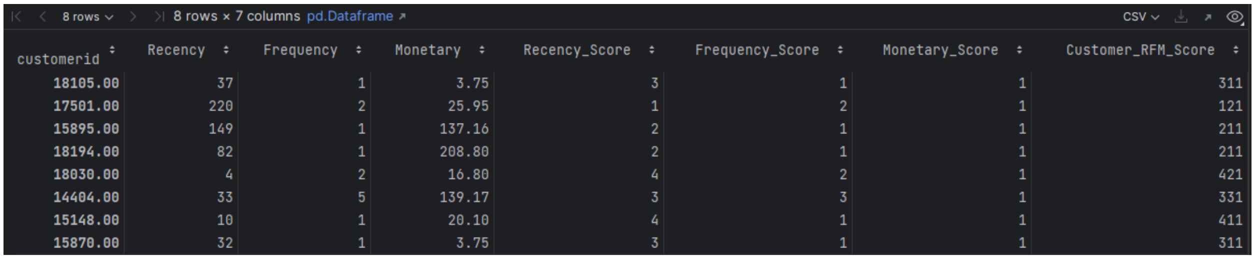
A key component of our consumer segmentation approach, RFM analysis enables us to analyze prior purchasing activity and categorize our clients accordingly.

1. Recency (R): This is about when the customer made their last purchase. Customers who have purchased more recently are more likely to purchase again than customers who have not purchased in a while. To calculate recency, you find the most recent purchase date for each customer and calculate the number of days between that date and the present.

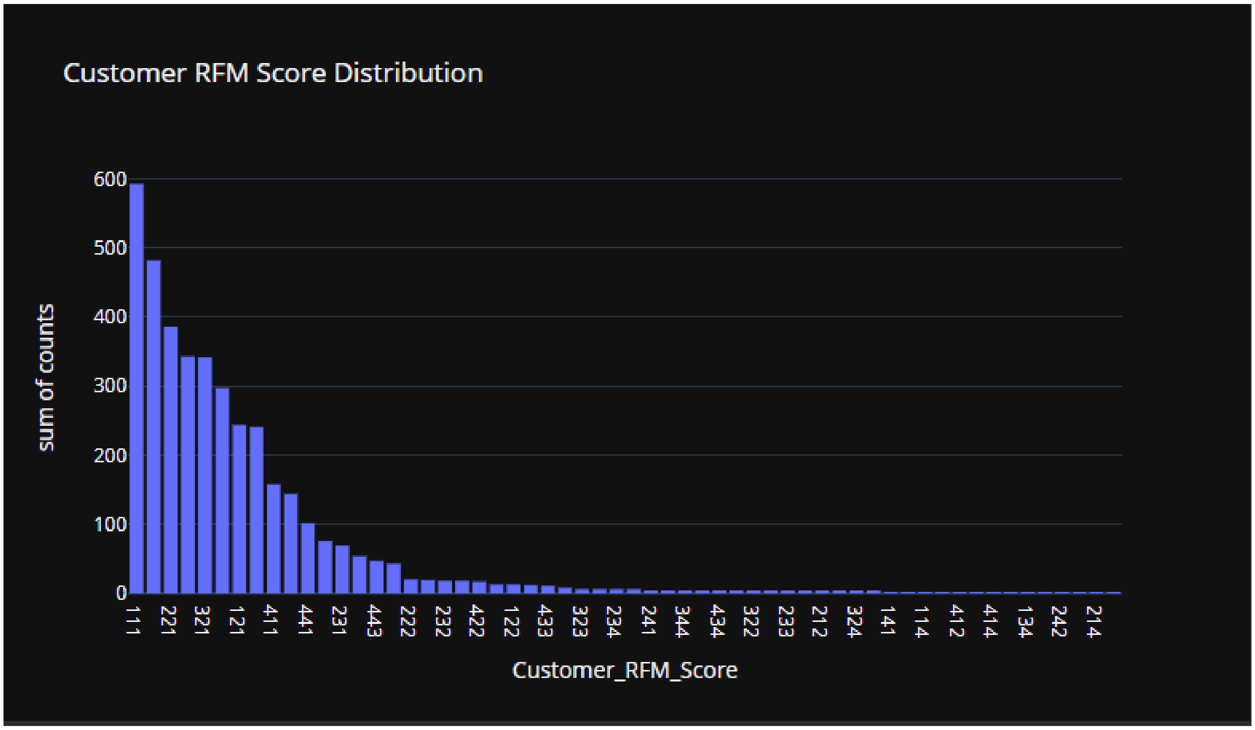
2. Frequency (F): This is about how often the customer makes a purchase. Customers who purchase frequently are more likely to do so again. To calculate frequency, you count the number of purchases each customer made in the period you're examining.

3. Monetary Value (M): This is about how much money the customer has spent on purchases. Customers who have spent a lot in the past are more likely to spend again. To calculate monetary value, you add up all the purchases for each customer.

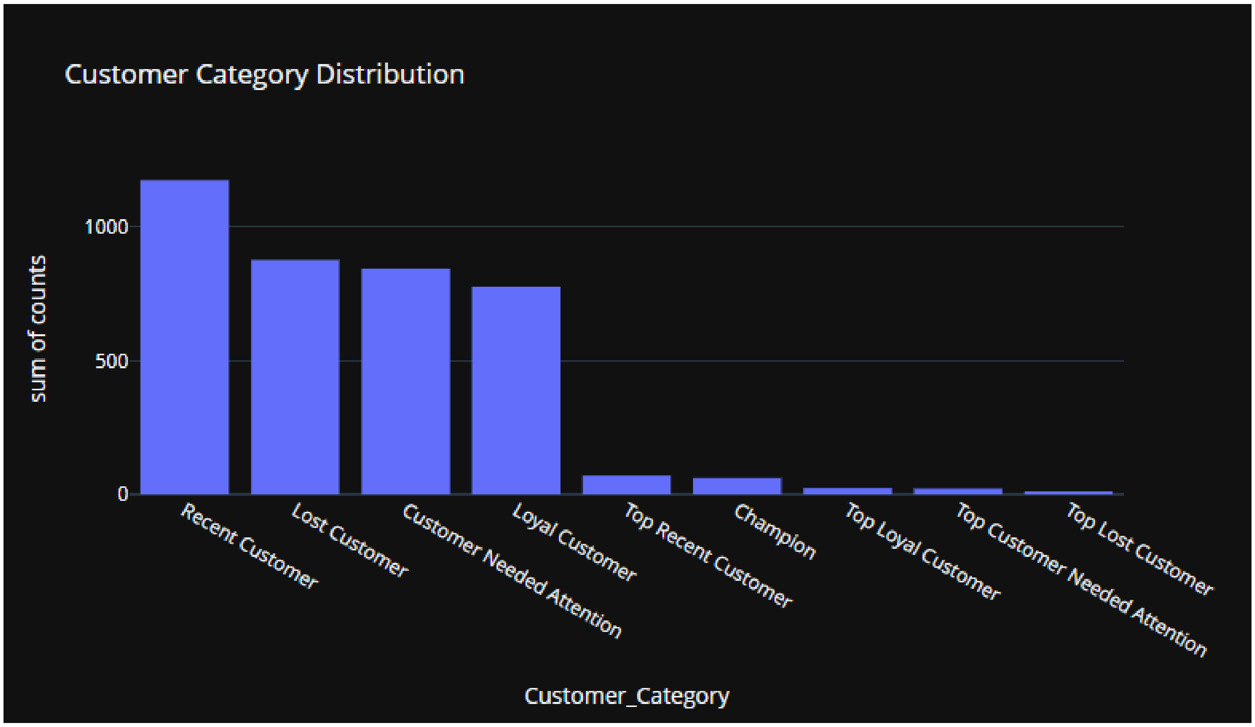
First, we calculated RFM score for each customer:

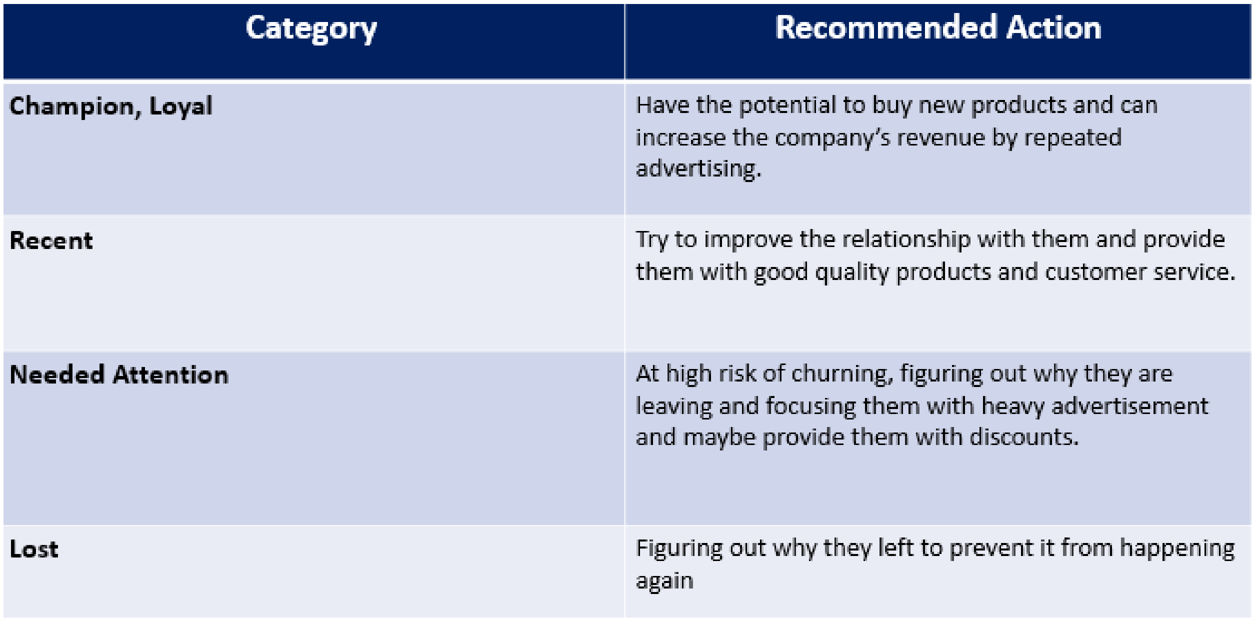


customerid	Recency	Frequency	Monetary	Recency_Score	Frequency_Score	Monetary_Score	Customer_RFM_Score
18105.00	37	1	3.75	3	1	1	311
17501.00	220	2	25.95	1	2	1	121
15895.00	149	1	137.16	2	1	1	211
18194.00	82	1	208.80	2	1	1	211
18030.00	4	2	16.80	4	2	1	421
14404.00	33	5	139.17	3	3	1	331
15148.00	10	1	20.10	4	1	1	411
15870.00	32	1	3.75	3	1	1	311



Then we Categorized our customers based on their scores:



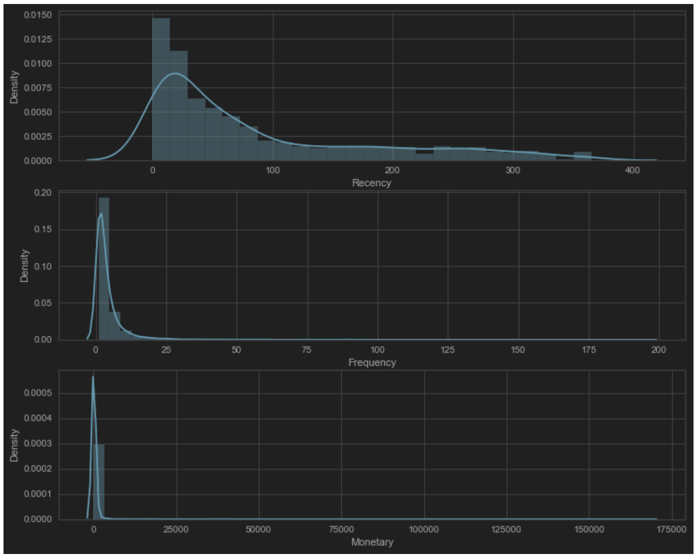


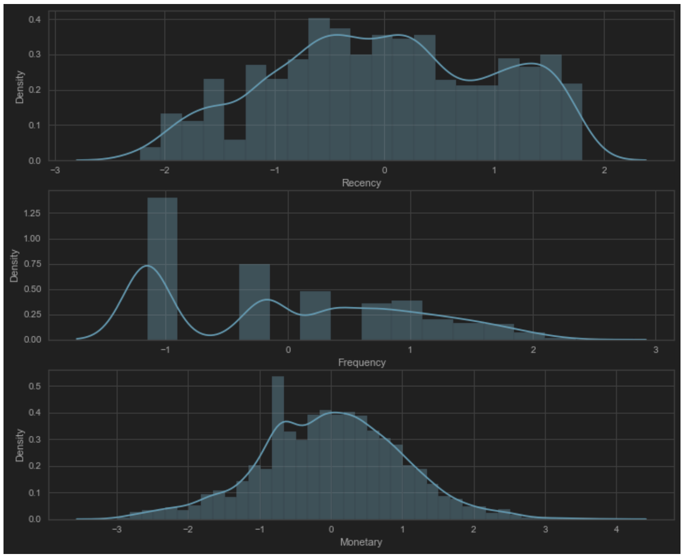
Category	Recommended Action
Champion, Loyal	Have the potential to buy new products and can increase the company's revenue by repeated advertising.
Recent	Try to improve the relationship with them and provide them with good quality products and customer service.
Needed Attention	At high risk of churning, figuring out why they are leaving and focusing them with heavy advertisement and maybe provide them with discounts.
Lost	Figuring out why they left to prevent it from happening again

Modeling and Performance Evaluation

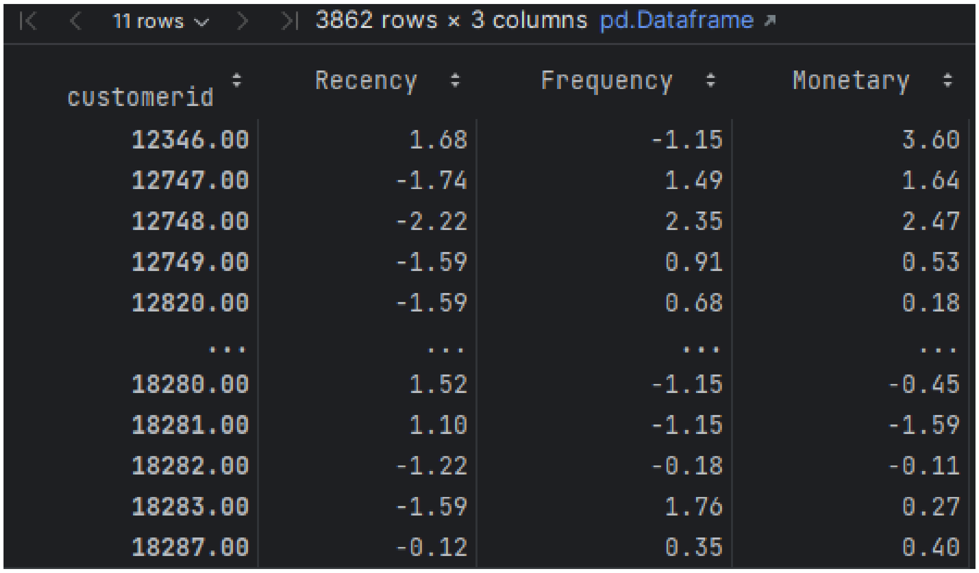
K-means is a popular choice for segmentation problems due to its simplicity, efficiency, and adaptability to large datasets. Its power lies in its ability to form distinct groups from a large pool of diverse entities—in our case, customers. By using this method, we aimed to cluster our customers into separate segments based on their purchasing behavior, as quantified by their Recency, Frequency, and Monetary (RFM) scores.

1. Preprocessing for Clustering:
   Given the diverse range and skewed nature of RFM scores, we employed a Power Transformer to make the distribution of these scores more Gaussian. This transformation stabilizes variance and minimizes skewness, enabling more accurate and meaningful clusters to emerge from our K-means algorithm.
   This is before and after using Power Transformer:





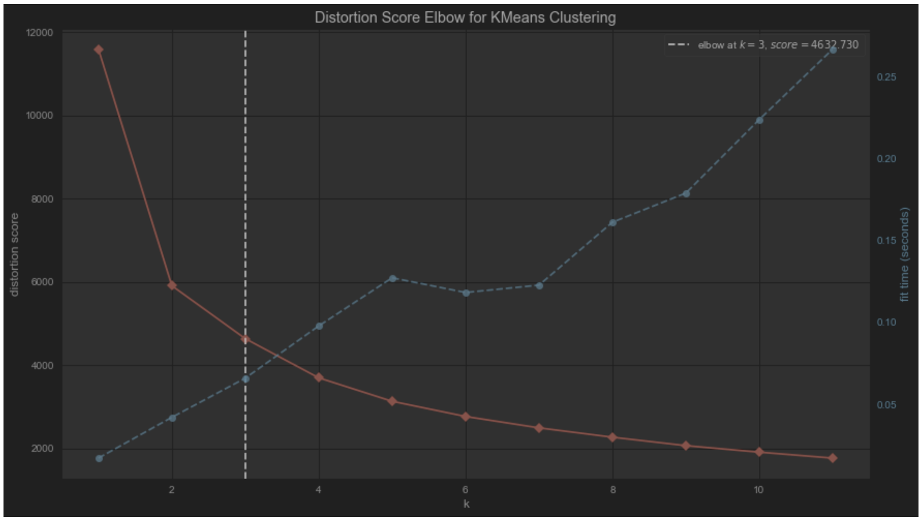
Next, we applied Standard Scaler to our data.it standardizes our features (RFM scores in this case) to have a mean of 0 and a standard deviation of 1. Standardization is particularly essential for K-means because this algorithm uses Euclidean distance to determine cluster membership. Differing scales of our RFM scores could lead to one feature dominating the distance calculations. With the standard scaler, we ensure that all features contribute equally to the final result.

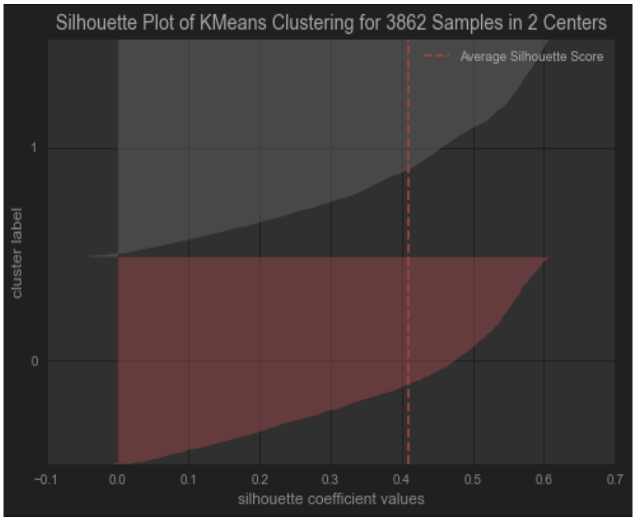


|< < 11 rows ∨ > >| 3862 rows × 3 columns pd.Dataframe ↗

customerid	Recency	Frequency	Monetary
12346.00	1.68	-1.15	3.60
12747.00	-1.74	1.49	1.64
12748.00	-2.22	2.35	2.47
12749.00	-1.59	0.91	0.53
12820.00	-1.59	0.68	0.18
...	...	...	...
18280.00	1.52	-1.15	-0.45
18281.00	1.10	-1.15	-1.59
18282.00	-1.22	-0.18	-0.11
18283.00	-1.59	1.76	0.27
18287.00	-0.12	0.35	0.40

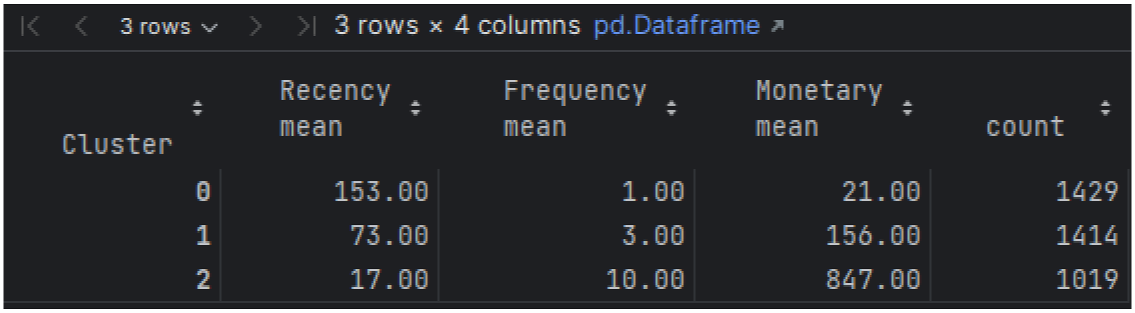
2. Modeling:
   With our data appropriately transformed and scaled, we proceeded with K-means clustering.
   To determine the optimal number of clusters. We applied two methods - the Elbow method and the Silhouette analysis.





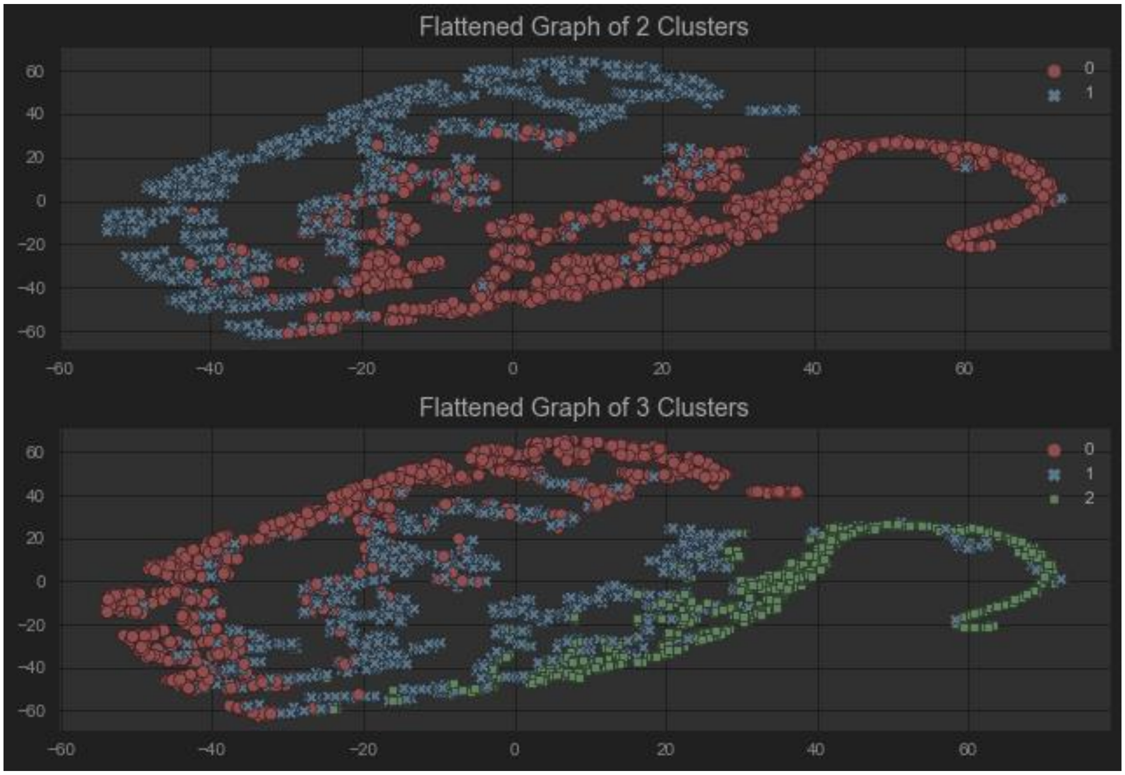
The Elbow method showed us a sharp bend at three clusters, indicating this might be a good number to use. At the same time, the Silhouette analysis, which helps us understand how well each customer fits into their cluster, suggested that two clusters could be better. We decided to try both options to see which one worked best.

In the end, we found that using three clusters gave us a better separation of our customers. This means that it was easier to tell the different groups apart when we divided them into three clusters.



Cluster	Recency mean	Frequency mean	Monetary mean	count
0	153.00	1.00	21.00	1429
1	73.00	3.00	156.00	1414
2	17.00	10.00	847.00	1019

From the results we can see:



1. Cluster 0 - 'Lost Customers': These are customers who haven't made a purchase in a while, signified by a high Recency value. Moreover, their Frequency and Monetary values are low, indicating that they don't shop often and when they do, their spend is relatively low. It is crucial for the business to delve into understanding why these customers are slipping away and what strategies can be implemented to re-engage them.

2. Cluster 1 - 'At-Risk Customers': These customers have a medium Recency value - they've made a purchase more recently than the 'Lost Customers', but it's been a while since their last purchase. They have higher Frequency and Monetary values, suggesting they shop more often and spend more than 'Lost Customers'. They are at risk of becoming 'Lost Customers' if not properly engaged. Strategies need to be put in place to incentivize them to shop more frequently.

3. Cluster 2 - 'Top Customers': These customers have the lowest Recency value, meaning they have made a purchase very recently. In addition, they have the highest Frequency and Monetary values, indicating that they shop very frequently and spend a lot when they do. These are our best customers, and maintaining a good relationship with them is key. The business should focus on retaining these customers and potentially using them as models for targeting or acquiring new customers.

Summary and conclusion

To better understand customer behavior, we combined the Recency, Frequency, Monetary (RFM) analysis, and K-means clustering on an online retail dataset spanning from 2010 to 2011. The RFM model allowed us to segment customers based on their shopping habits, while K-means clustering grouped these customers based on their similar attributes. After data cleaning and preprocessing, which included dealing with missing values, removing canceled orders, and handling duplicates, we focused our attention on customers from the UK market, as they represented the majority of our customers and revenue. We then used the RFM model to create initial customer segments. This was followed by applying K-means clustering on the RFM scores. The Elbow method and Silhouette analysis were used to determine the optimal number of clusters, which turned out to be three.

Our analysis revealed three distinct customer segments: Lost Customers (High Recency, Low Frequency, and Monetary values), At-Risk Customers (Medium Recency, Higher Frequency, and Monetary values), and Top Customers (Lowest Recency, Highest Frequency, and Monetary values). This segmentation can provide valuable insights into the customer base, helping in targeted marketing and improving customer relationship management. The combined approach of using RFM and K-means clustering brought together the best of both worlds: the simplicity and straightforwardness of RFM with the objective and detailed nature of K-means clustering. Moving forward, the business can use these insights to build more tailored marketing strategies, work on re-engaging Lost and At-risk customers, and maintain the loyalty of Top customers. This work has shown the power of data analysis in helping a business understand its customers better and make more informed decisions.

References

[1] Chen, D., Guo, K., Li, B. (2019). Predicting Customer Profitability Dynamically over Time: An Experimental Comparative Study. In: Nyström, I., Hernández Heredia, Y., Milián Núñez, V. (eds) Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications. CIARP 2019. Lecture Notes in Computer Science(), vol 11896. Springer, Cham. https://doi.org/10.1007/978-3-030-33904-3_16.

[2] E. Wong and W. Yan, "Customer online shopping experience data analytics: Integrated customer segmentation and customised services prediction model," International Journal of Retail & Distribution Management, vol. 46, (4), pp. 406-420, 2018. Available: https://login.proxy.bib.uottawa.ca/login?url=https://www.proquest.com/scholarlyjournals/customer-online-shopping-experience-data/docview/2034194183/se-2. DOI: https://doi.org/10.1108/IJRDM-06-2017-0130.

[3] P. P. Pramono, I. Surjandari and E. Laoh, "Estimating Customer Segmentation based on Customer Lifetime Value Using Two-Stage Clustering Method," 2019 16th International Conference on Service Systems and Service Management (ICSSSM), Shenzhen, China, 2019, pp. 1-5, DOI: 10.1109/ICSSSM.2019.8887704.

[4] Sokol, Ondřej, and Vladimír Holý. "The Role of Shopping Mission in Retail Customer Segmentation." International Journal of Market Research, vol. 63, no. 4, 2021, pp. 454–70, https://doi.org/10.1177/1470785320921011.

[5] Mitra, A., Jain, A., Kishore, A., Kumar, P. (2023). A Comparative Study for Machine Learning Models in Retail Demand Forecasting. In: Bhattacharyya, S., Banerjee, J.S., Köppen, M. (eds) Human-Centric Smart Computing. Smart Innovation, Systems and Technologies, vol 316. Springer, Singapore. https://doi.org/10.1007/978-981-19-5403-0_23.

[6] Online Retail, UCI Machine Learning Repository, 2015. [Online]. Available: https://doi.org/10.24432/C5BW33.