CISC 451 – Assignment 3

Using RFM Analysis and Unsupervised Learning for Customer Segmentation

Due: November 15th, 2020

Gavin McClelland – 10211444
Marshall Cunningham - 20249991

1.0 Introduction

The increase in online retail sales over the past decade along with the volume of data collected has called for the study of consumer behaviour. By better understanding the behaviour of their customers, companies can insightfully target different kinds of people who exhibit different traits. This ultimately enables companies to conduct more efficient marketing and can lead to personalized shopping experiences for different consumers.

For this assignment, the Recency, Frequency, and Monetary (RFM) model was used to quantify the value different customers from the provided dataset of transactions from an online retailer in the UK. Then, unsupervised clustering was used to create meaningful customer segments on the RFM representation. This document details the software packages used, the analytics process undertaken, the results achieved, and the main characteristics of consumers in each segment.

2.0 Software Packages

All code written for this assignment was written in the Python programming language using Jupyter notebooks. The main software packages used in this assignment were, pandas, scikit-learn, and numpy. A tree structure of the submitted directory is shown below.

```
|    CISC 451 – Assignment 3 Report.pdf
|    requirements.txt
|    Mean_cluster_video.mp4, Max_cluster_video.mp4, Sum_cluster_video.mp4
+---code
|        customer_segmentation.ipynb
\---data
|        Online Retail.xlsx
```

3.0 Instructions

Instructions to replicate all work completed as submitted are listed below:

1. Install all dependencies by running the command “pip install -r requirements.txt” in the root directory of the submitted folder.
2. Before running either notebook in the code folder, add your working directory at the top where indicated (i.e. %cd “<your directory here>”). Then, run all cells in the customer_segmentation.ipynb notebook.

4.0 Methodology and Analytics Process

The analytics process employed consisted of the following steps:

1. Exploratory Data Analysis (EDA) and data cleaning (looking at null values, negatives, etc.)
2. Following the hints specified in the assignment outline to develop the RFM model, whereby records were grouped into one per customer, and aggregates were performed to resolve Recency, Frequency, and Monetary metrics.
3. Using clustering algorithms for customer segmentation (K-Means)
4. Evaluation of results and hyperparameter tweaking

This section discusses these steps in further detail.

4.1 EDA

When first exploring the dataset, it was found that there were very few records in which features were null, except for the CustomerID column, which was null in 24.92% of records. Since the provided dataset has more than 500,000 records, it seemed acceptable to drop these records in the pre-processing phase. Moreover, since the objective was to create one record for each customer, this decision was made easier. Simple analyses were also performed out of curiosity, such as the percentage of null customers by country, where 100% of customers from Hong Kong did not have a customer ID. Additionally, 45.3% of records with a null CustomerID also had a country of "Unspecified", which indicates that these records might be unreliable altogether.

Another thing that was noticed were negative values for both Quantity and UnitPrice. Initial thoughts indicated that these would also be removed from the initial dataset, but upon further investigation, these records typically contained product descriptions along the lines of "Missing", "Lost", or even more explicit phrases like "Adjust bad debt" or "Discount". These findings indicated that these products were likely returned or lost revenue that should be included as legitimate transactions when exploring consumer behaviour. For example, what if a particular customer frequently returned items or used discounts?

4.2 Modeling

Modeling was separated into two phases, where the initial dataset was first transformed into an RFM representation, which was then used in the construction of simple unsupervised clustering models using Scikit-Learn.

4.2.1 RFM

The hints in the assignment outline were used to create Recency, Frequency, and Monetary metrics from the provided online retail dataset. After doing some basic cleaning on the initial dataset, the dataset was grouped by each CustomerID, then Recency, Frequency, and Monetary aggregates were computed. Recency is expressed as the difference between the most recent date observed in the entire dataset ("Today", or 09/12/2011) and the most recent day a customer made a transaction, where lower is better. Frequency was simply calculated as the number of purchases each customer made (count of InvoiceNo by CustomerID). Lastly, a "TotalPrice" feature was created as the product of Quantity and UnitPrice for each record. For the Monetary metric, the sum, minimum, maximum, mean, and median of this number (grouped by CustomerID) were computed to add additional depth to the feature space.

Scores for each RFM metric were assigned from 1-to-5 by quantizing each record into one of 5 tiers, where the optimal score is 15. Since there are five different Monetary sub-metrics, each of these were quantized, and then the mean of the 5 "M" values was used as the final M value. Additionally, outliers were removed after constructing the RFM representation by computing the Z-Score for each data point, and removing any records containing a data point with a Z-Score greater than 3 (more than 3 standard deviations away from the mean for that feature). There were 82 such outliers removed. It was found that using the sum of "TotalPrice" achieved optimal performance, so this was used for the Monetary feature.

4.2.2 Unsupervised Clustering

With a cleaned RFM representation now created, unsupervised clustering was performed to establish meaningful customer segments. The first step in this process was to get each RFM metric on a logarithmic scale (base 10) and normalized accordingly. The elbow method was then

used to determine the number of clusters to be used. Here, the within cluster sum-of-squares (WCSS) was plotted for one-to-thirty potential clusters, and the "elbow" of the plot would be the number of clusters to use, which appeared to be 5 (WCSS is also called "inertia" in Scikit-learn). This elbow plot is shown below in Figure 1.

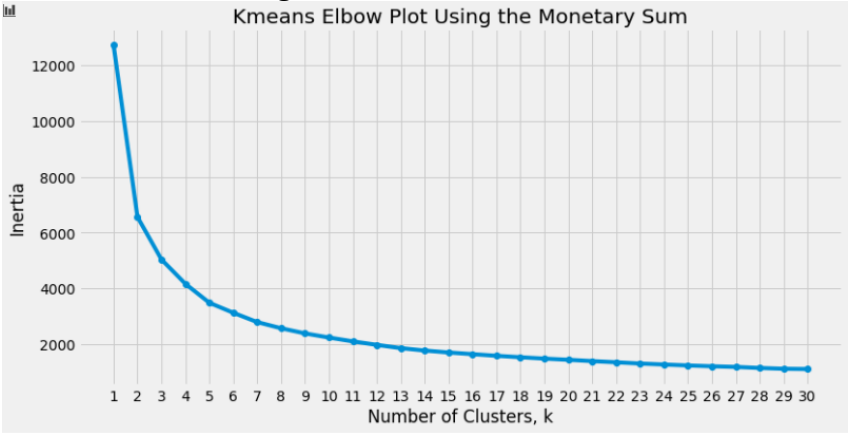

Figure 1 – Finding the WCSS for different numbers of clusters to determine the number of clusters to use.

Next, K-Means was used to create five clusters from the RFM data. The output is shown below in Figure 2, where very clear cluster boundaries can be seen.

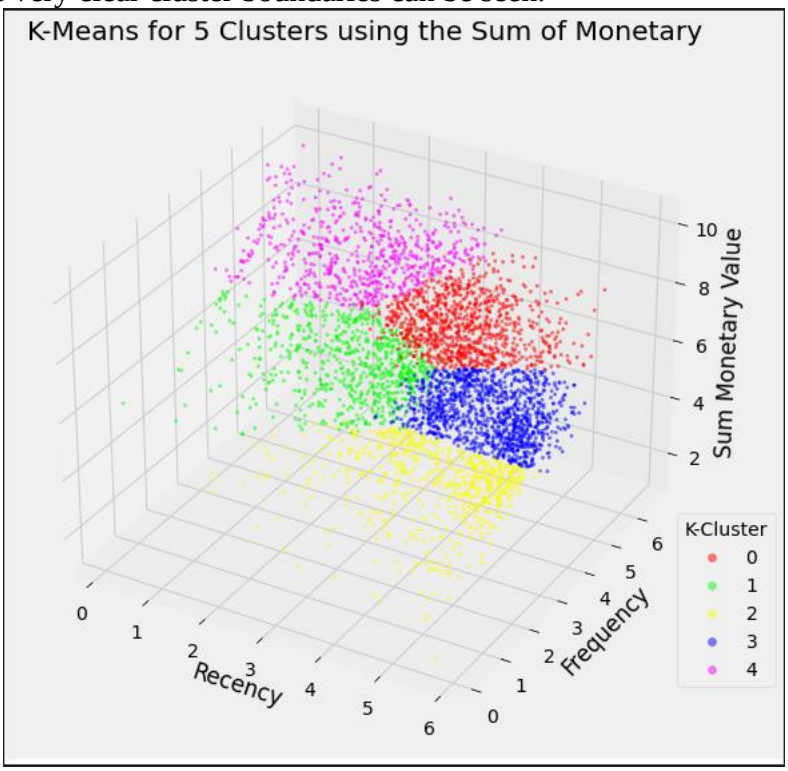

Figure 2 – K-Means using the sum of each customer's purchases as the Monetary feature.

This was the best result achieved out of all approaches, other efforts that were less successful will be detailed below.

4.3 Other Approaches Explored

In parallel with the creating the optimal clusters using the sum for the Monetary feature, the mean and maximum (by customer) were also used to explore different behaviour. It was found

that the mean created a much “tighter” dataset, which makes sense as this would “smooth” the dataset. Using the maximum of each customer’s purchases for the Monetary feature performed a bit better but did not perform as well as the sum. Clustering results using the mean and maximum for the monetary feature are shown below in Figures 3 and 4, respectively.

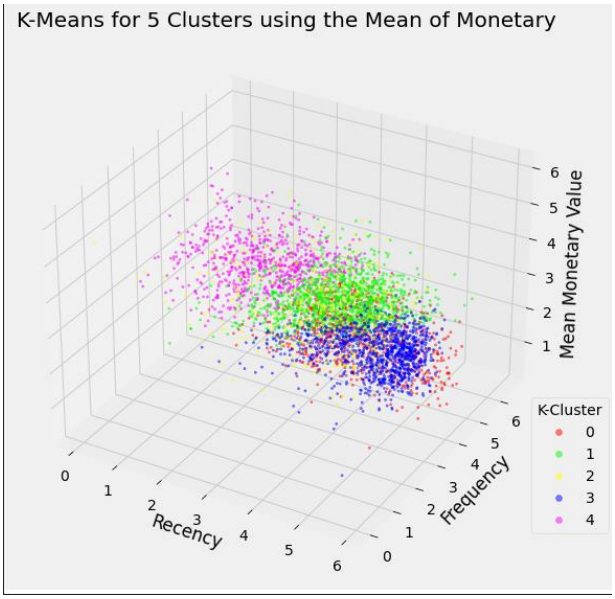

Figure 3 – K-Means using the mean of each customer’s purchases as the Monetary feature.

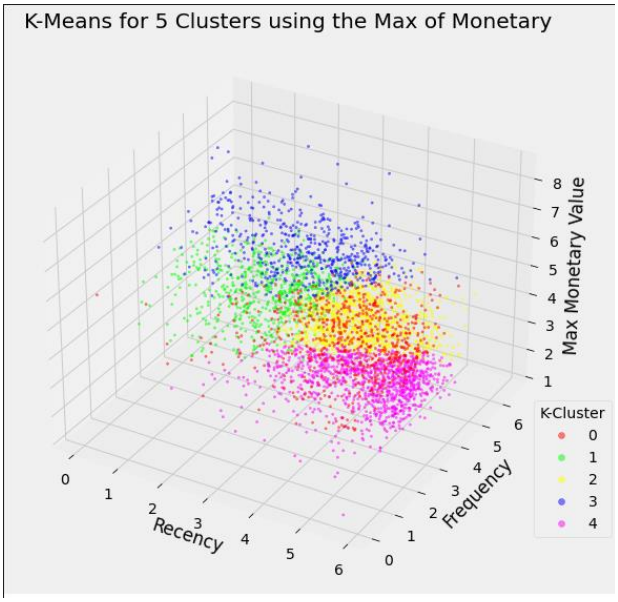

Figure 4 - K-Means using the maximum of each customer’s purchases as the Monetary feature.

Additional perspectives on the Recency and Frequency features in the RFM representation were considered but not explored. One such example was to create weekly and monthly features for Frequency to better understand periodic shopping behaviour. Furthermore, creating features for “time between purchases” could add additional dimensions to the Recency feature. This could help understand if certain customers shop at different times of the year. These features should be explored in future work.

5.0 Model Evaluation

5.1 Distribution of Customers and RFM Metrics in Each Cluster

The distribution of customers varied when using the mean, maximum, and sum for the Monetary feature, with the best observed distribution being observed when using the sum. The distribution for each of the three cases is shown below in Figure 5.

	Recency			Frequency			Monetary_mean			Percentage
	mean	max	median	mean	max	median	mean	max	median	count
Cluster_Mean										
0	78.0	374	41	129.0	756	90	6.0	13.0	6.0	730	17.213
1	48.0	330	36	72.0	421	55	22.0	138.0	18.0	1223	28.838
2	122.0	374	82	11.0	110	6	102.0	408.0	72.0	373	8.795
3	185.0	374	185	19.0	87	16	18.0	50.0	17.0	1190	28.059
4	6.0	24	4	185.0	757	137	20.0	171.0	17.0	725	17.095

	Recency			Frequency			Monetary_max				Percentage
	mean	max	median	mean	max	median	mean	max	median	count	
Cluster_Max											
0	142.0	374	111	18.0	116	14.0	180.0	3794.0	132.0	684	16.128
1	8.0	19	8	139.0	757	96.0	53.0	159.0	50.0	678	15.987
2	68.0	369	47	98.0	674	72.0	55.0	164.0	50.0	1165	27.470
3	20.0	267	11	177.0	742	128.0	302.0	4161.0	200.0	555	13.087
4	167.0	374	161	20.0	85	17.0	31.0	68.0	30.0	1159	27.328
	Recency			Frequency			Monetary_sum				Percentage
	mean	max	median	mean	max	median	mean	max	median	count	
Cluster_Sum											
0	63.0	338	46	114.0	588	98	1859.0	11582.0	1491.0	970	22.872
1	17.0	45	16	41.0	175	35	585.0	3516.0	502.0	740	17.449
2	166.0	374	162	8.0	31	7	177.0	816.0	151.0	701	16.529
3	165.0	374	147	32.0	132	28	505.0	3276.0	417.0	1163	27.423
4	9.0	43	7	239.0	757	199	4338.0	25748.0	3172.0	667	15.727

Figure 5 – Distribution of customers using the mean (top), maximum (middle), and sum (bottom) of the Monetary feature.

5.2 Assessment of Cluster Features

The distinct features of each of the five clusters is broken down in Table 1 based on empirical observations.

Table 1 - Distinct characteristics of each of the five clusters.

Cluster	Description
0	Customers who on average spend the second most, shop somewhat frequently, and have shopped somewhat recently.
1	Recent shoppers who do not make frequent purchases and spend moderately
2	Infrequent shoppers who have made 8 purchases on average in the dataset, and do not spend much on average. The least-profitable customer segment.
3	Shoppers who have not made purchases recently but shopped somewhat frequently in the past and spent considerably.
4	Customers who spend the most, shop often, and have shopped recently. The most-profitable customer segment.