Noah Frank, 400050264
Online Shoppers - Cluster Analysis
frankn1@mcmaster.ca

1 Data Description

1.1 Data Source

The data used for this assignment is the 'Online Retail Data Set' retrieved from the UCI Machine Learning Repository.[1] The data contains a list of all transactions of an online retailer between December 1st 2020 and December 9th 2011 based in the UK.

1.2 Variable Details

The data set is originally made up of 8 fields, but was reduced to 3 fields after data cleaning and preparation. To clean the data, we removed fields that were not of interest for the cluster analysis such as the product description, stock code, invoice number, and country. The other features were prepared to make them more meaningful for analysis. A count of all purchases was summed up and grouped per customer to give a total number of purchases. The unit price and quantity for each order was then combined and summed over all orders to give a measure of total money spent. The difference between the last available date in the data set and the most recent purchase date was calculated to give a measure of how recent of a customer they are. The final features used in the analysis are restated here:

- **Total Spent:** Sum of all purchase value per customer
- **Number of Purchases:** Count of all purchases per customer
- **Purchase Recency:** Number of days since last purchase

Outlier detection was performed, using the interquartile rule. Any rows containing outliers were removed from consideration for the analysis.

1.3 Descriptive Analysis

After dealing with outliers, we are left with 3818 customers, each with 3 features which will be used for clustering. Some statistics of these features are summarized in Table 1 below.

Table 1: Summary of Customer Features

	Tot. Spent	Num. Purchases	Recency
Min.	80.2	3	1
Median	659.4	43	49
Mean	1167.0	70.0	83.9
Max.	9065.8	384	336

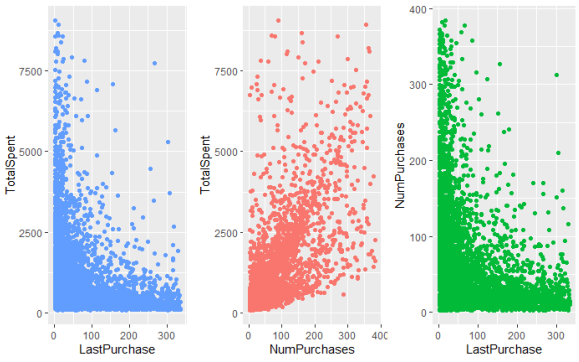

Figure 1: Scatter plot between each pair of features

In Fig. 1, we can see the relationships that exist between each pairs of features. The clearest relationship is between the 'total spent' and 'number of purchases' variables, where we can observe an overall trend showing that those with more purchases have spent more. Both variables 'total spent' and 'number of purchases' show a similar relationship with 'purchase recency'. We can see that those who have been customers more recently, have both spent more money overall, and also have made more purchases.

2 Problem Description

The aim of this analysis is to explore the data set from a clustering perspective in order to gain insights on customer's spending habits with this online retailer. Being able to group customers by the measured features would be beneficial knowledge for the online retailer's marketing, or any potential advertisers.

3 Technique Description

Clustering is used in order to group data points with the goal of having groups contain the most similar points. Clustering involves calculating measures of distance between points, which is called the dissimilarity.[2] One option for calculating the dissimilarity is using the Euclidean distance, as defined here:

$$d(x_i, x_j) = \sqrt{\sum_{m=1}^{M}(x_{im} - x_{jm})^2}$$

In hierarchical clustering, points are clustered into groups based on dissimilarities. Iteratively, cluster are merged together in what is known as agglomerative hierarchical clustering. The distance between clusters is needed to merge them, which is also known as linkage. An example of a type of linkage is complete linkage, given as:

$$d(A, B) = \max_{x \in A, y \in B} d(x, y)$$

[1] UCI Machine Learning Repository. *Online Retail Data Set*. 2015. URL: https://archive.ics.uci.edu/ml/datasets/online+retail.

[2] Prof. McNicholas. *CSE780: Introduction to Clustering*. McMaster University, School of Computational Science & Engineering. 2021.

Based on these measures, agglomerative hierarchical clustering will produce a dendrogram showing the distance between points as the height at which their branches join. These plots can be used to visually determine how many clusters to separate the groups into based on the distance between clusters.

Partitioning methods are another form of clustering that can be used to group data. In these methods, we cluster the points around a predefined k number of cluster centres. The cluster centres are learned and optimized to minimize the distance between the points belonging to a cluster, and the centre of that cluster. There are two main types of partitioning methods: k-means, where the cluster centres are means, and k-medoids, where the cluster centres are medoids. In order to choose the value of k, there exists two methods: the elbow method and the silhouette method. For each k, the elbow method calculates the sum of squares distance within each cluster (WSS) and sums them all together. As clusters become more accurate this measure should drop, and a value for k is chosen after the sharpest fall-off in WSS. Alternatively, the silhouette method calculates a coefficient for each point over each value of k depending on that points similarity to the cluster. The width of the coefficients are added up per cluster and the value of k which gives the largest mean width between all clusters is the optimal value.

The final method for clustering that will be investigated is mixture model-based clustering. For this analysis, Gaussian parsimonious clustering models (GPCMs) are used. These models are developed by parameterizing the covariance structure to give a variety of models.[3]. An eigenvalue decomposition of the covariance matrices for the Gaussian mixture model is given by:

$$\sum_g = \lambda_g \Gamma_g \Delta_g \Gamma_g'$$

λ_g is a constant,
Γ_g is a matrix of eigenvectors of $\sum_g$,
Δ_g is a unitary diagonal matrix with entries
proportional to the eigenvalues of $\sum_g$

This decomposition is used to construct a family of 14 GPCMs. The best model and number of clusters for the GPCMs are fitted using the Bayesian Information Criterion (BIC). For each number of clusters, every model from the GPCM family is fitted and the highest BIC throughout all models and all numbers of clusters is selected.

4 Result Description

Starting with agglomerative hierarchical clustering, we found the best results came from using Ward's method, implemented by the 'hclust' method of the 'cluster' R library. In Fig. 2, we can see the dendrogram colored according to the clusters when choosing 3 clusters based on the long height for which the 3 clusters are still distanced. Choosing 3 clusters, also allows us to more directly compare results with the k-means and k-medoids approaches. In Fig. 3, we can see the distribution of each cluster for each feature as a box plot. This figure allows us to visually examine where the cluster boundaries might lie for each feature.

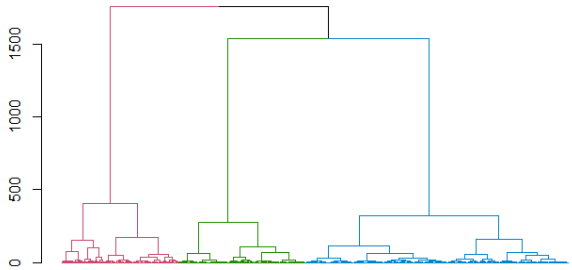

Figure 2: Dendrogram for Ward method clustering

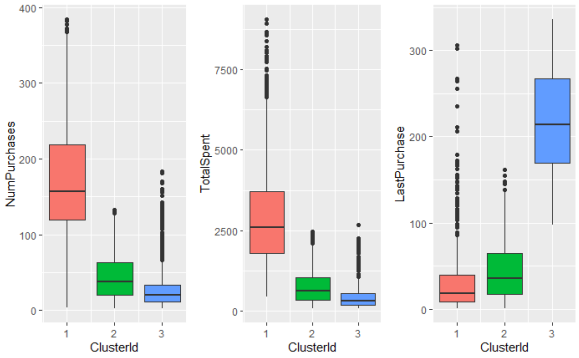

Figure 3: Hierarchical cluster summary statistics

Moving to the next method, k-means clustering, we can use the elbow method and the silhouette method to determine the optimal number of clusters 'k'. The plots in Fig. 4, generated by the 'factoextra' R package, show the optimal number of clusters to be 3 for both methods. We can then compute the k-means clustering with a k value of 3, giving Fig. 5, the box plot distribution of each cluster for each feature.

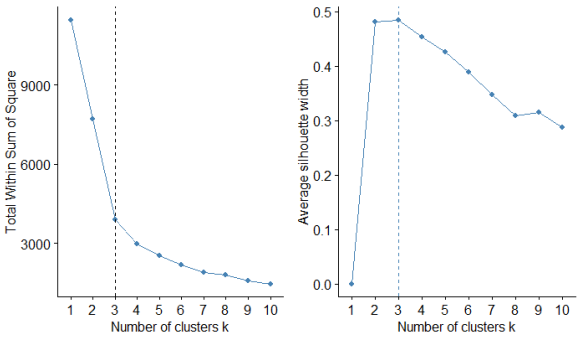

Figure 4: Elbow method and Silhouette method for k-means

[3]Prof. McNicholas. *CSE780: Model Based Clustering I*. McMaster University, School of Computational Science & Engineering. 2021.

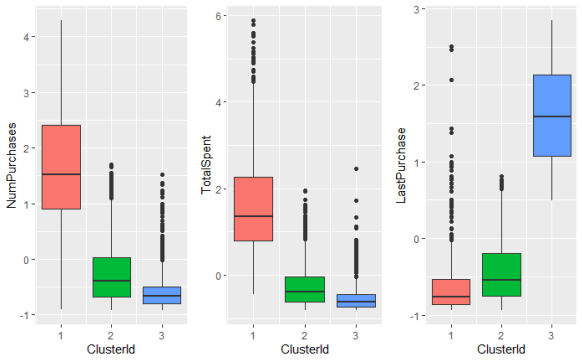

Figure 5: k-means summary statistics

Next, using k-medoids clustering is similar k-means clustering, where we can use the elbow method and the silhouette method to determine the optimal number of clusters 'k'. The plots in Fig. 6, show the optimal number of clusters to be also 3 from the elbow method and the silhouette method. The k-medoids clustering was then computed to give the box plots in Fig. 7, showing the distribution between clusters for the features.

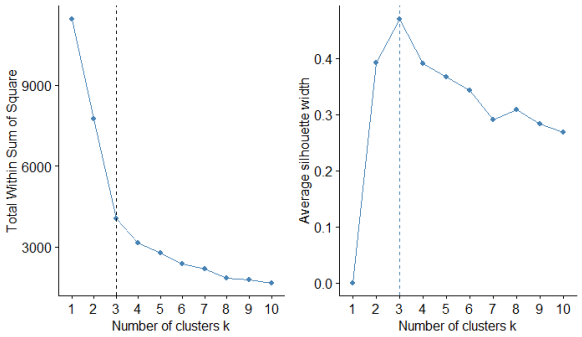

Figure 6: Elbow method and Silhouette method for k-medoids

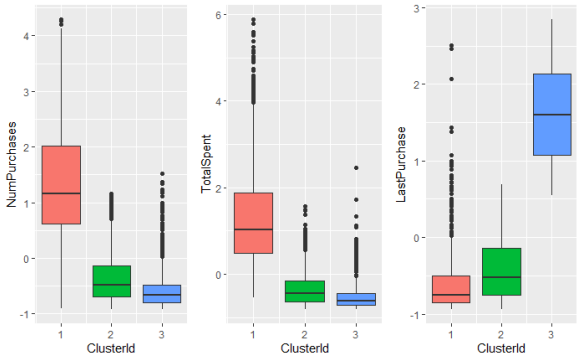

Figure 7: k-medoids summary statistics

The final method used for clustering is the GPCMs, through the use of the 'mclust' package in R. The BIC, determined that the best fitted model is the EVE model with 9 clusters. This model has equal volume between clusters, variable shape and equal orientation. The resulting box plot for this model clustering can be seen in Fig. 8. In order to compare to the previous models, the GPCMs were re-calculated for specifically 3 clusters. This resulted in an VVV (ellipsoidal, varying volume, shape, and orientation) model with 3 components. The box plots showing the distribution of the clusters across the features can be seen in Fig. 9.

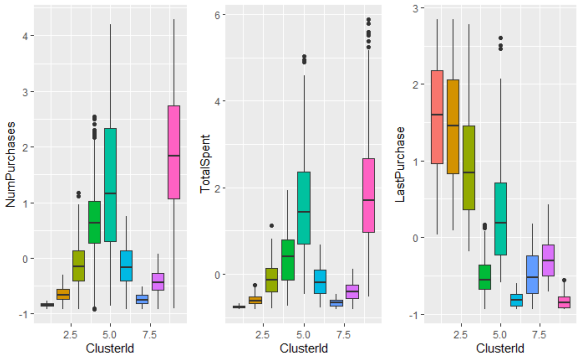

Figure 8: Scatter plot between each pair of features

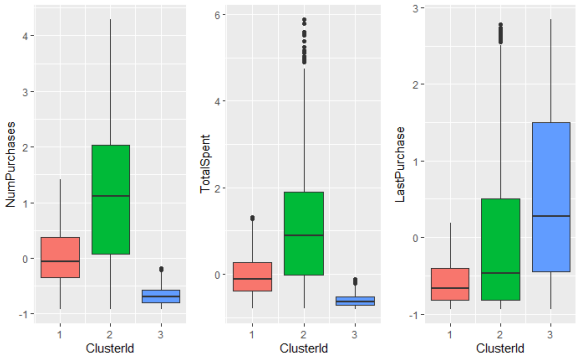

Figure 9: Scatter plot between each pair of features

5 Conclusions

The clustering analysis showed very similar results throughout the hierarchical, k-means, and k-medoids methods. All three methods grouped customers according to the same three groups. The GPCMs were not as similar in their grouping, likely due to the data not naturally forming well-defined clusters. By examining the summary statistics plots we can make some conclusions in the context of the customer data set. Cluster 1 represents customers who have the highest expenditures and purchases, and are the most recent customers. Cluster 2 represents customers who spend less money, make less purchases, but are still recent customers. Cluster 3 represents customers who have not done much transactions and are not recent customers. Based on this analysis, strategies for reaching out to each of these groups could be identified for advertisement purposes.