Clustering of UK Online Retail Data

1. Introduction

In this project, I performed K-means clustering and hierarchical clustering on a dataset that provides information on transactions from a UK-based online retail company. I found that the K-means clustering algorithm produced clusters that were noticeably different from those produced by the hierarchical clustering algorithm. In both methods, I found that the clusters produced were not incredibly homogenous and did not produce distinctively well-defined classes.

2. Data

The dataset provides information on transactions occurring between 1/12/2010 and 9/12/2011 for a UK-based and registered non-store online retail. The company mainly sells unique all-occasion gifts, and many of their customers are wholesalers. The original dataset contains 541,909 observations and 8 variables. A description of the 8 original variables are provided in Table 1.

Variable	Description
InvoiceNo	Nominal, a 6-digit integral number uniquely assigned to each transaction. If this code starts with letter 'c', it indicates a cancellation.
StockCode	Product (item) code. Nominal, a 5-digit integral number uniquely assigned to each distinct product.
Description	Product (item) name. Nominal.
Quantity	The quantities of each product (item) per transaction. Numeric.
InvoiceDate	Invoice Date and time. Numeric, the day and time when each transaction was generated.
UnitPrice	Unit price. Numeric, Product price per unit in sterling.
CustomerID	Customer number. Nominal, a 5-digit integral number uniquely assigned to each customer.
Country	Country name. Nominal, the name of the country where each customer resides.

Table 1. Variables in the original dataset. Descriptions are sourced from https://archive.ics.uci.edu/ml/datasets/Online%20Retail#

The original dataset underwent the following clean-up steps prior to analysis:

1. Removed any observations for which Country was not specified.
2. Removed transactions that were returns or cancellations.
3. Aggregated all transactions for each customer in the dataset. This produced an aggregate amount spent per customer.
4. Determined the first month of customer interaction, as well as the last month of interaction. These two variables were used to calculate the duration of the customer's interaction with the company.
5. Calculated the Amount per Purchase, Purchases per Month, and Amount per Month for each customer.

All variables that were considered in the analysis are provided in Table 2.

Variable	Description
Country	Country name. Nominal; the name of the country where each customer resides.
Amount	Numeric; the total amount spent by customer.
LastMth	Nominal; the last month of customer interaction
Months	Numeric; the duration of the customer's interaction with the company
Purchases	Numeric; the number of purchased made by the customer during the time frame
Amount.per.Purchase	Numeric; Amount/Purchases
Purchases.per.Month	Numeric; Purchases/Months
Amount.per.Month	Numeric; Amount/Months

Table 2. Variables considered in the analysis.

2.1 Sources and References

The dataset was obtained from the University of California Irvine Machine Learning Repository. The source of the data is listed below:

Source

1. Daqing Chen, Sai Liang Sain, and Kun Guo, Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining, Journal of Database Marketing and Customer Strategy Management, Vol. 19, No. 3, pp. 197-208, 2012 (Published online before print: 27 August 2012. doi: 10.1057/dbm.2012.17).

3. Analyses, Plots, and Tables

I attempted to identify whether this online retail company's consumers can be segmented meaningfully by the recency and duration of their interaction with the company, the frequency of their purchases, and the amount they spent. We were also curious about whether we could segment the customers by their country of residence.

3.1 Exploratory Data Analysis

After the original dataset was cleaned using the steps described in Section 2, we explored whether there were any strong associations between the quantitative variables. The left panel of Figure 1 shows a correlation plot of all of the variables that were considered for inclusion in the analysis. I found strong associations between the following variables:

- `Amount.per.Month` and `Amount.per.Purchase` (positive association)
- `FirstMth` and `Months` (negative)
- `Purchases` and `Purchases.per.Month` (positive)

Based on the collinearity between the variables, I made the following decisions regarding which variables to include (or exclude) in the analysis:

- Exclude `FirstMth`.
- Exclude `Purchases`.
- Include `LastMth`. Despite the relatively strong association between `LastMth` and `Months`, the `LastMth` variable will serve as a measure of the recency of interaction between the customer and the online retailer.

- Include `Amount`. Despite the relatively strong associations between `Amount` and (1) `Amount.per.Month` and (2) `Purchases`, the `Amount` variable will serve as a measure of the volume of purchases by the customer.

Table 1 provides a description of the variables that were actually included in the analysis. The right panel of Figure 1 shows a correlation plot of these variables, and Table 3 provides summary statistics for these variables. We see that, with the exception of `Amount.per.Month` vs. `Amount.per.Purchase`, there are no Pearson correlation coefficients with magnitudes greater than 0.8 in the final dataset. Figure 2 provides histograms for the continuous variables in the dataset. We see that these variables all have right-skewed distributions.

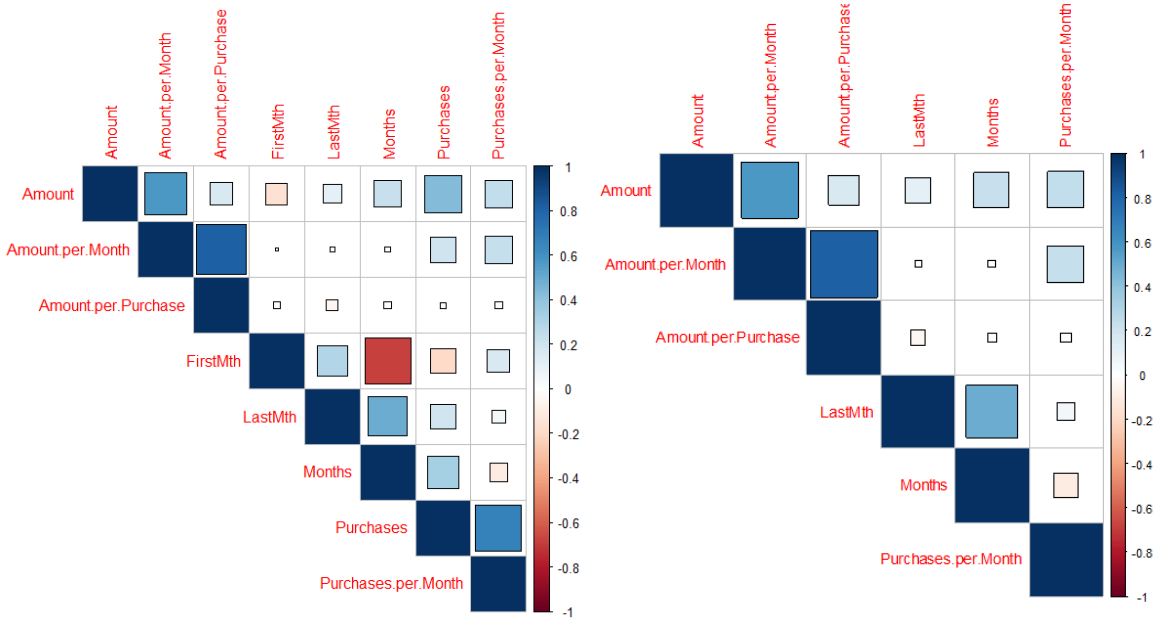

Figure 1. Left panel: Correlation plot of all variables that were considered for inclusion in the analysis. Right panel: Correlation plot of the variables that were, indeed, included in the analysis.

Amount	FirstMth	LastMth	Months
Min. : 0.0	Min. : 1.00	Min. : 1.00	Min. : 1.00
1st Qu.: 303.9	1st Qu.: 2.00	1st Qu.: 8.00	1st Qu.: 1.00
Median : 656.6	Median : 5.00	Median :11.00	Median : 4.00
Mean : 1952.1	Mean : 5.45	Mean : 9.47	Mean : 5.02
3rd Qu.: 1593.3	3rd Qu.: 9.00	3rd Qu.:12.00	3rd Qu.: 9.00
Max. :268478.0	Max. :12.00	Max. :12.00	Max. :12.00
Purchases	**Amount.per.Purchase**	**Purchases.per.Month**	**Amount.per.Month**
Min. : 1.00	Min. : 0.00	Min. : 0.1818	Min. : 0.0
1st Qu.: 17.00	1st Qu.: 12.34	1st Qu.: 6.6667	1st Qu.: 122.0
Median : 40.00	Median : 17.69	Median : 13.0000	Median : 221.5
Mean : 88.54	Mean : 54.56	Mean : 20.4697	Mean : 400.3
3rd Qu.: 97.00	3rd Qu.: 24.82	3rd Qu.: 24.0000	3rd Qu.: 388.4
Max. :7376.00	Max. :77183.60	Max. :1145.5000	Max. :77183.6

Table 3. Summary statistics of the quantitative variables considered in the analysis.

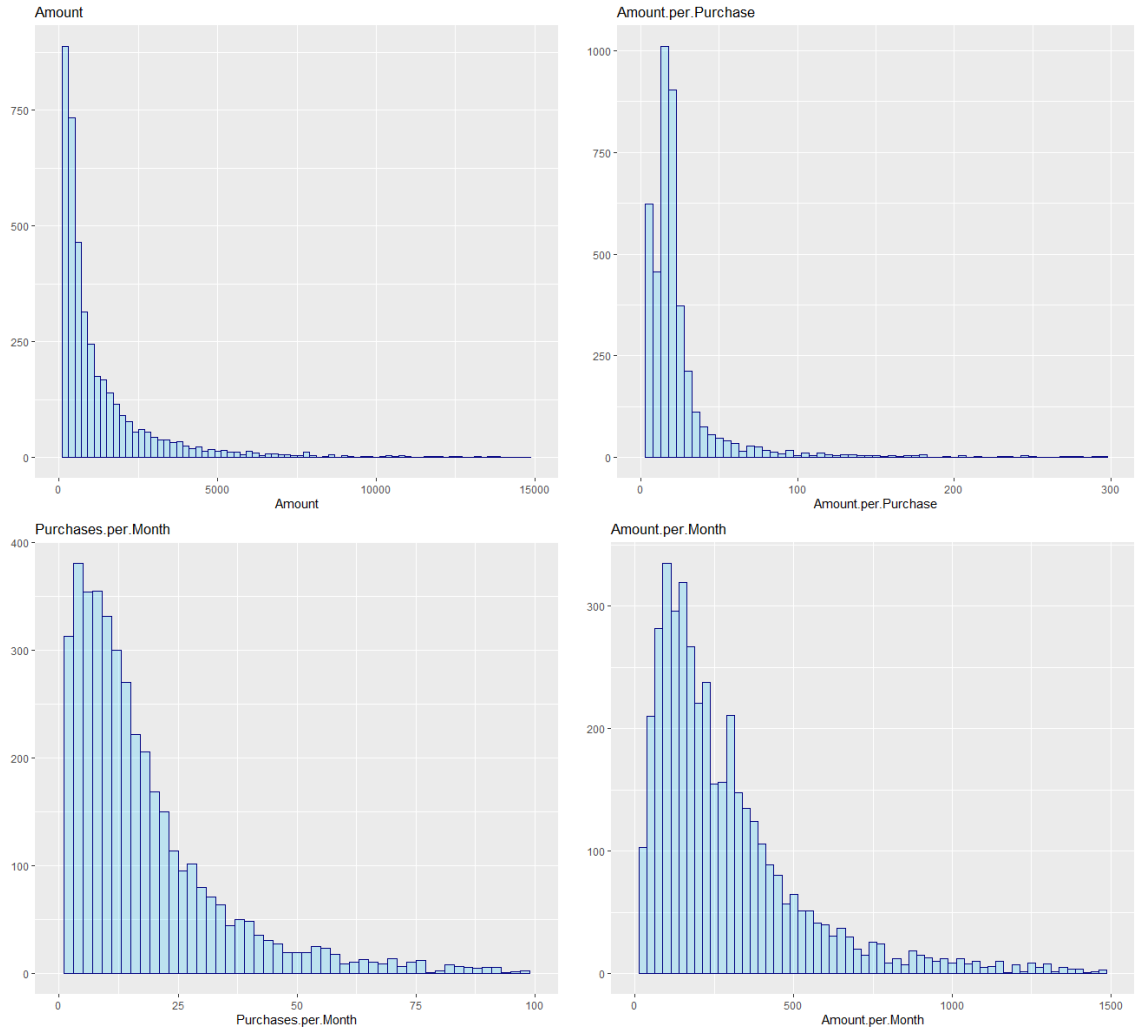

Figure 2. Histogram of the continuous variables included in the analysis.

3.2 K-Means Clustering

For both the K-means clustering and hierarchical clustering algorithms, the variables were scaled. The K-means algorithm was run with a start variable of 50 to obtain the local optimum. Values of K between 1 and 10 were considered. `Country` was not included in either clustering algorithm, considering it is a categorical factor. Calculating the Euclidean distance between categorical factors (even if they are treated as numerical) is not appropriate. Nevertheless, I considered `Country` after the K-means algorithm was run with the remaining numerical variables to determine if the algorithm found any patterns related to `Country`.

In Figure 3, we see an elbow in the value of the within cluster sum-of-squares at K=5. This determination is subjective, considering we could also conclude that the elbow exists at K=6. However, I proceeded with K=5, and fit the K-means clustering model to the dataset using 5 clusters.

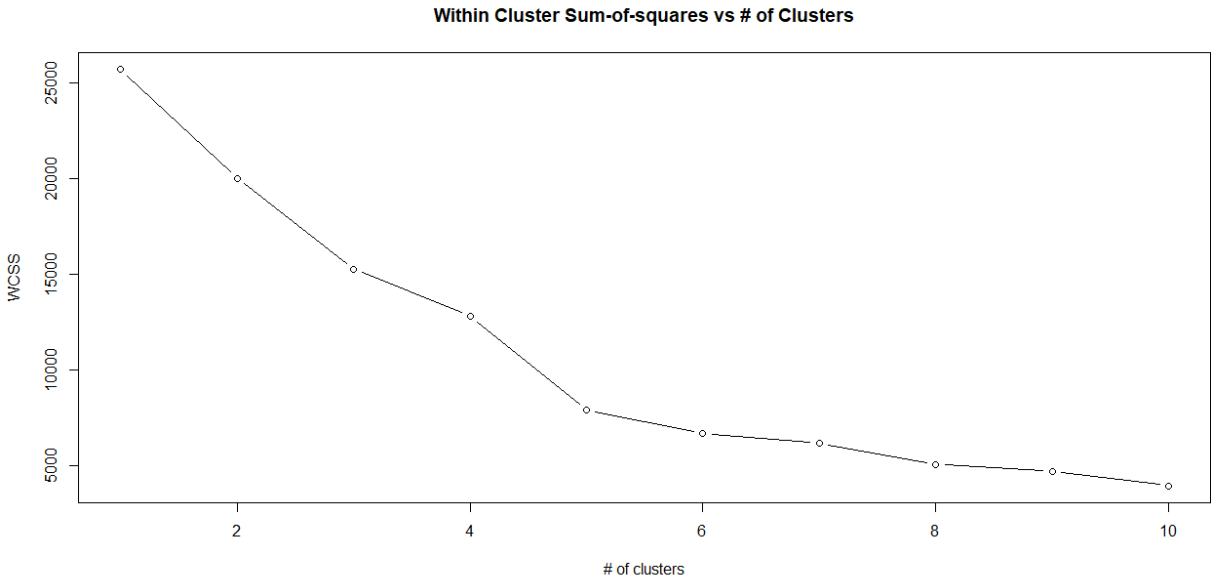

Figure 3. Plot of the within cluster sum-of-squares for each value of K between 1 and 10.

I did not find any overwhelmingly distinct segments in the clustering. First, I considered `Months`, which measures the duration of the customer's interaction with the online retailer. As summarized in Table 4, we can note a few findings:

- The majority of observations for which `Months` ≥ 6 are in Cluster 1. This suggests that customers that have had a longer interaction time with the company have similar purchasing habits.
- Customers with lower values of `Months` (`Months` < 5) are typically found in Clusters 3, 4, and 5.
- Cluster 2 captures only 9 customers.

Months	Cluster 1	2	3	4	5	Sum
1		2	771	360	559	**1,692**
2			134	47	38	**219**
3			128	55	27	**210**
4		1	62	71	20	**154**
5	12		71	104	1	**188**
6	130		17	60		**207**
7	181			34		**215**
8	221					**221**
9	252					**252**
10	233					**233**
11	251	1				**252**
12	435	5	3			**443**
Sum	**1,715**	**9**	**1,186**	**731**	**645**	**4,286**

Months	Cluster 1	2	3	4	5	Sum
1		0.1%	45.6%	21.3%	33.0%	**100.0%**
2			61.2%	21.5%	17.4%	**100.0%**
3			61.0%	26.2%	12.9%	**100.0%**
4		0.6%	40.3%	46.1%	13.0%	**100.0%**
5	6.4%		37.8%	55.3%	0.5%	**100.0%**
6	62.8%		8.2%	29.0%		**100.0%**
7	84.2%			15.8%		**100.0%**
8	100.0%					**100.0%**
9	100.0%					**100.0%**
10	100.0%					**100.0%**
11	99.6%	0.4%				**100.0%**
12	98.2%	1.1%	0.7%			**100.0%**

Table 4. The Months variable (which had discrete values between 1 and 12) was referenced against the 5 clusters produced by the K-means clustering algorithm. Top Panel: Cluster 1 and Cluster 4 had the most observations. Cluster 2 had the fewest observations (9). Bottom Panel: The proportion of observations within each value of Months is referenced against the Cluster determined by the K-means clustering algorithm. For example, 45.6% of observations with Months=1 were classified in Cluster 3.

Furthermore, I did not find any overwhelming segments when considering the recency of the customer's interaction with the company, which is measured by the `LastMth` variable. A larger value for `LastMth` indicates that the customer had a more recent interaction with the company. As summarized in Table 5, we can note a few findings:

- The majority of customers for which `LastMth` ≤ 5 are in Cluster 5. This suggests that customers that have not had recent interactions with the online retailer have similar purchasing habits.
- The majority of customers for which $6 \leq$ `LastMth` ≤ 8 are in Cluster 4.

- Cluster 1 is composed of only customers for which `LastMth` ≥ 8.

	Cluster					
Lastmth	1	2	3	4	5	**Sum**
1					124	**124**
2		1			96	**97**
3					113	**113**
4					174	**174**
5				20	125	**145**
6				160	13	**173**
7		1	4	185		**190**
8	26		17	151		**194**
9	78		29	125		**232**
10	194		162	87		**443**
11	355		380	3		**738**
12	1,062	7	594			**1,663**
Sum	**1,715**	**9**	**1,186**	**731**	**645**	**4,286**

	Cluster					
Lastmth	1	2	3	4	5	**Sum**
1					100.0%	**100.0%**
2		1.0%			99.0%	**100.0%**
3					100.0%	**100.0%**
4					100.0%	**100.0%**
5				13.8%	86.2%	**100.0%**
6				92.5%	7.5%	**100.0%**
7		0.5%	2.1%	97.4%		**100.0%**
8	13.4%		8.8%	77.8%		**100.0%**
9	33.6%		12.5%	53.9%		**100.0%**
10	43.8%		36.6%	19.6%		**100.0%**
11	48.1%		51.5%	0.4%		**100.0%**
12	63.9%	0.4%	35.7%			**100.0%**

Table 5. Top Panel: The LastMth variable (which had discrete values between 1 and 12) was referenced against the 5 clusters produced by the K-means clustering algorithm. Bottom Panel: The proportion of observations within each value of LastMth is referenced against the Cluster. For example, 1.0% of observations with LastMth=2 were classified in Cluster 2.

Finally, I explored whether the clusters showed any patterns among the customers' countries of residence. I did not find much homogeneity in the clusters. However, I do note that for UK customers (which make up the largest customer base for the online retailer), 40.3% of customers fall within Cluster 1 and 27.4% fall within Cluster 3. This is summarized in Table 6.

	Cluster					
Country	1	2	3	4	5	Sum
Australia	4	1		1	1	7
Austria	1		2	3	3	9
Bahrain				2		2
Belgium	11		3	3	5	22
Brazil					1	1
Canada			1	2	1	4
Channel Islands	3		1	3	2	9
Cyprus	1		2		2	5
Czech Republic	1					1
Denmark	3		2	1		6
EIRE	1	2				3
European Community				1		1
Finland	4		6		1	11
France	37		25	9	15	86
Germany	37		33	12	11	93
Greece				1	2	3
Iceland	1					1
Israel			2	1		3
Italy	4		6		4	14
Japan	3		1	2	2	8
Lebanon					1	1
Lithuania					1	1
Malta	1		1			2
Netherlands	4	1	1	2	1	9
Norway	3		6		1	10
Poland	2			4		6
Portugal	5		7	2	5	19
RSA			1			1
Saudi Arabia					1	1
Singapore	1					1
Spain	12		8	5	2	27
Sweden	4		1	2	1	8
Switzerland	4		9	2	4	19
United Arab Emirates			1		1	2
United Kingdom	1,568	5	1,065	673	576	3,887
USA			2		1	3
Sum	1,715	9	1,186	731	645	4,286

	Cluster					
Country	1	2	3	4	5	Sum
Australia	57.1%	14.3%		14.3%	14.3%	100.0%
Austria	11.1%		22.2%	33.3%	33.3%	100.0%
Bahrain				100.0%		100.0%
Belgium	50.0%		13.6%	13.6%	22.7%	100.0%
Brazil					100.0%	100.0%
Canada			25.0%	50.0%	25.0%	100.0%
Channel Islands	33.3%		11.1%	33.3%	22.2%	100.0%
Cyprus	20.0%		40.0%		40.0%	100.0%
Czech Republic	100.0%					100.0%
Denmark	50.0%		33.3%	16.7%		100.0%
EIRE	33.3%	66.7%				100.0%
European Community				100.0%		100.0%
Finland	36.4%		54.5%		9.1%	100.0%
France	43.0%		29.1%	10.5%	17.4%	100.0%
Germany	39.8%		35.5%	12.9%	11.8%	100.0%
Greece				33.3%	66.7%	100.0%
Iceland	100.0%					100.0%
Israel			66.7%	33.3%		100.0%
Italy	28.6%		42.9%		28.6%	100.0%
Japan	37.5%		12.5%	25.0%	25.0%	100.0%
Lebanon					100.0%	100.0%
Lithuania					100.0%	100.0%
Malta	50.0%		50.0%			100.0%
Netherlands	44.4%	11.1%	11.1%	22.2%	11.1%	100.0%
Norway	30.0%		60.0%		10.0%	100.0%
Poland	33.3%			66.7%		100.0%
Portugal	26.3%		36.8%	10.5%	26.3%	100.0%
RSA			100.0%			100.0%
Saudi Arabia					100.0%	100.0%
Singapore	100.0%					100.0%
Spain	44.4%		29.6%	18.5%	7.4%	100.0%
Sweden	50.0%		12.5%	25.0%	12.5%	100.0%
Switzerland	21.1%		47.4%	10.5%	21.1%	100.0%
United Arab Emirates			50.0%		50.0%	100.0%
United Kingdom	40.3%	0.1%	27.4%	17.3%	14.8%	100.0%
USA			66.7%		33.3%	100.0%

Table 6. Top Panel: The Country variable was referenced against the 5 clusters produced by the K-means clustering algorithm. Bottom Panel: the proportion of observations within each Country is referenced against the Cluster. For example, 57.1% of the observations with Country=Australia were classified in Cluster 1.

To visualize the clusters, I projected the observations onto the plane created by the first two principal components of the dataset. These two principal components explain 61.89% of the variation. In Figure 4, we can see that Cluster 2 is the least dense, and is obviously separate from

the other four clusters. Clusters 1, 3, 4, and 5 are extremely close to each other in this space. We can see a distnict separation between Clusters 1 and 5. However, there is overlap between Clusters 1, 3, and 4. There is also overlap between Clusters 3, 4, and 5.

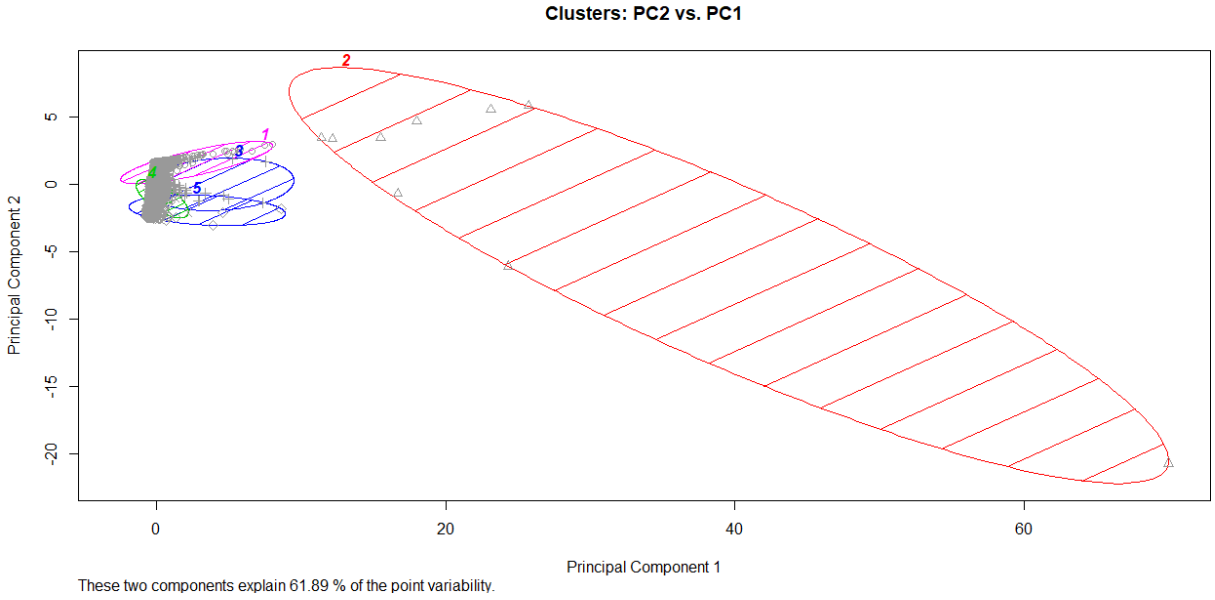

Figure 4. Visualization of the clusters, with observations projected onto the plane created by the first two principal components of the dataset. The gray points represent observations projected onto this plane. This plot was produced using the clusplot() function within the cluster library.

3.3 Hierarchical Clustering

The variables were also scaled before proceeding with the hierarchical clustering algorithm. I first observed the differences in the dendogram between the complete linkage, average linkage, and single linkage methods. I found that the complete linkage method produced the most distinct clusters (as shown in Figure 5), and proceeded with this method. Using the length of the branches in the dendogram, I subjectively decided to use 4 clusters to fit the model.

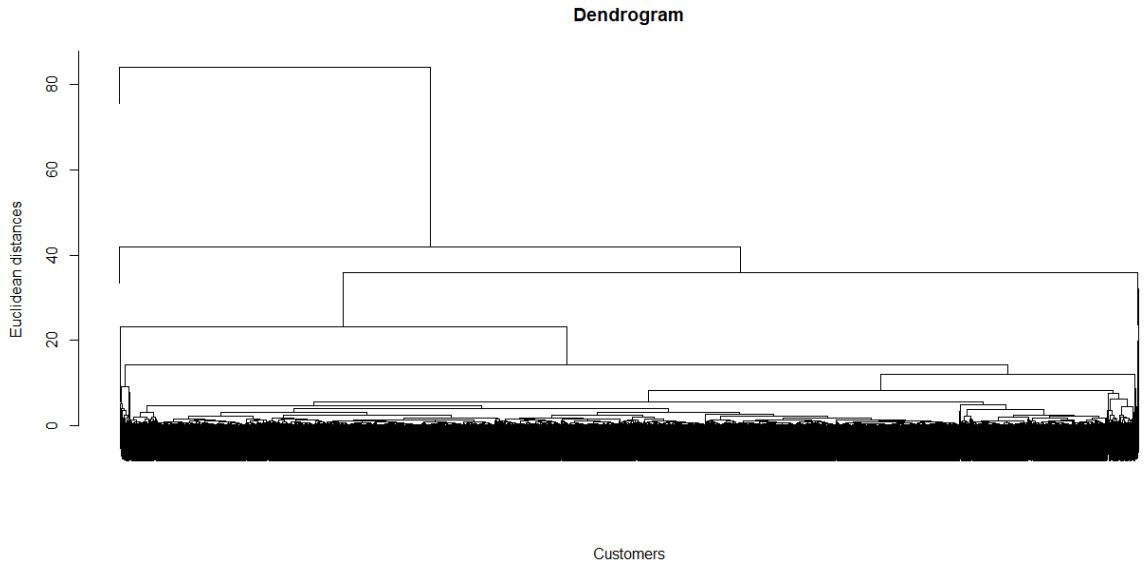

Figure 5. Dendogram produced by hierarchical clustering using complete linkage.

Just as in the K-means clustering discussed in Section 3.2, I considered the value of `Months` against the clusters produced by the hierarchical clustering algorithm. In Table 7, we see that Cluster 2 captures 99% of the observations. In fact, only customers for which `Months`$\in$ {1, 4, 11, 12} were found in the other clusters.

	Cluster				
Months	1	2	3	4	**Sum**
1	1	1,690	1		**1,692**
2		219			**219**
3		210			**210**
4		153		1	**154**
5		188			**188**
6		207			**207**
7		215			**215**
8		221			**221**
9		252			**252**
10		233			**233**
11		251	1		**252**
12		439	4		**443**
Sum	**1**	**4,278**	**6**	**1**	**4,286**

	Cluster				
Months	1	2	3	4	Sum
1	0.1%	99.9%	0.1%		100.0%
2		100.0%			100.0%
3		100.0%			100.0%
4		99.4%		0.6%	100.0%
5		100.0%			100.0%
6		100.0%			100.0%
7		100.0%			100.0%
8		100.0%			100.0%
9		100.0%			100.0%
10		100.0%			100.0%
11		99.6%	0.4%		100.0%
12		99.1%	0.9%		100.0%

Table 7. The Months variable was referenced against the 4 clusters produced by the hierarchical clustering algorithm. Top Panel: Cluster 2 had 99% of all observations. Bottom Panel: The proportion of observations within each value of Months is referenced against the Cluster.

In Table 8, we see a similar pattern when referencing the value of `LastMth` against the cluster produced by hierarchical clustering.

	Cluster				
Lastmth	1	2	3	4	Sum
1		124			124
2	1	96			97
3		113			113
4		174			174
5		145			145
6		173			173
7		189	1		190
8		194			194
9		232			232
10		443			443
11		738			738
12		1,657	5	1	1,663
Sum	1	4,278	6	1	4,286

	Cluster				
Lastmth	1	2	3	4	Sum
1		100.0%			100.0%
2	1.0%	99.0%			100.0%
3		100.0%			100.0%
4		100.0%			100.0%
5		100.0%			100.0%
6		100.0%			100.0%
7		99.5%	0.5%		100.0%
8		100.0%			100.0%
9		100.0%			100.0%
10		100.0%			100.0%
11		100.0%			100.0%
12		99.6%	0.3%	0.1%	100.0%

Table 8. Top Panel: The LastMth variable was referenced against the 4 clusters produced by the hierarchical clustering algorithm. Bottom Panel: The proportion of observations within each value of LastMth is referenced against the Cluster.

In Table 9, we see that only customers from Australia, EIRE, Netherlands, and the UK were found outside of Cluster 2.

	Cluster				
Country	1	2	3	4	Sum
Australia		6	1		7
Austria		9			9
Bahrain		2			2
Belgium		22			22
Brazil		1			1
Canada		4			4
Channel Islands		9			9
Cyprus		5			5
Czech Republic		1			1
Denmark		6			6
EIRE		2	1		3
European Community		1			1
Finland		11			11
France		86			86
Germany		93			93
Greece		3			3
Iceland		1			1
Israel		3			3
Italy		14			14
Japan		8			8
Lebanon		1			1
Lithuania		1			1
Malta		2			2
Netherlands		8	1		9
Norway		10			10
Poland		6			6
Portugal		19			19
RSA		1			1
Saudi Arabia		1			1
Singapore		1			1
Spain		27			27
Sweden		8			8
Switzerland		19			19
United Arab Emirates		2			2
United Kingdom	1	3,882	3	1	3,887
USA		3			3
Sum	1	4,278	6	1	4,286

	Cluster				
Country	1	2	3	4	Sum
Australia		85.7%	14.3%		100.0%
Austria		100.0%			100.0%
Bahrain		100.0%			100.0%
Belgium		100.0%			100.0%
Brazil		100.0%			100.0%
Canada		100.0%			100.0%
Channel Islands		100.0%			100.0%
Cyprus		100.0%			100.0%
Czech Republic		100.0%			100.0%
Denmark		100.0%			100.0%
EIRE		66.7%	33.3%		100.0%
European Community		100.0%			100.0%
Finland		100.0%			100.0%
France		100.0%			100.0%
Germany		100.0%			100.0%
Greece		100.0%			100.0%
Iceland		100.0%			100.0%
Israel		100.0%			100.0%
Italy		100.0%			100.0%
Japan		100.0%			100.0%
Lebanon		100.0%			100.0%
Lithuania		100.0%			100.0%
Malta		100.0%			100.0%
Netherlands		88.9%	11.1%		100.0%
Norway		100.0%			100.0%
Poland		100.0%			100.0%
Portugal		100.0%			100.0%
RSA		100.0%			100.0%
Saudi Arabia		100.0%			100.0%
Singapore		100.0%			100.0%
Spain		100.0%			100.0%
Sweden		100.0%			100.0%
Switzerland		100.0%			100.0%
United Arab Emirates		100.0%			100.0%
United Kingdom	0.0%	99.9%	0.1%	0.0%	100.0%
USA		100.0%			100.0%

Table 9. Left Panel: The Country variable was referenced against the 4 clusters produced by the hierarchical clustering algorithm. Right Panel: The proportion of observations within each Country is referenced against the Cluster.

To visualize the clusters, I again projected the observations onto the plane created by the first two principal components of the dataset. In Figure 6, we see that Cluster 2 is the most dense. There is some overlap between Clusters 2 and 3. In this space, Cluster 4 is within Cluster 3. Cluster

1 is distinctly different from the remaining three clusters, considering it is far from the other clusters in this space. We can see that the clusters produced by hierarchical clustering using complete linkage are incredibly different from the ones produced by K-means clustering using K=5.

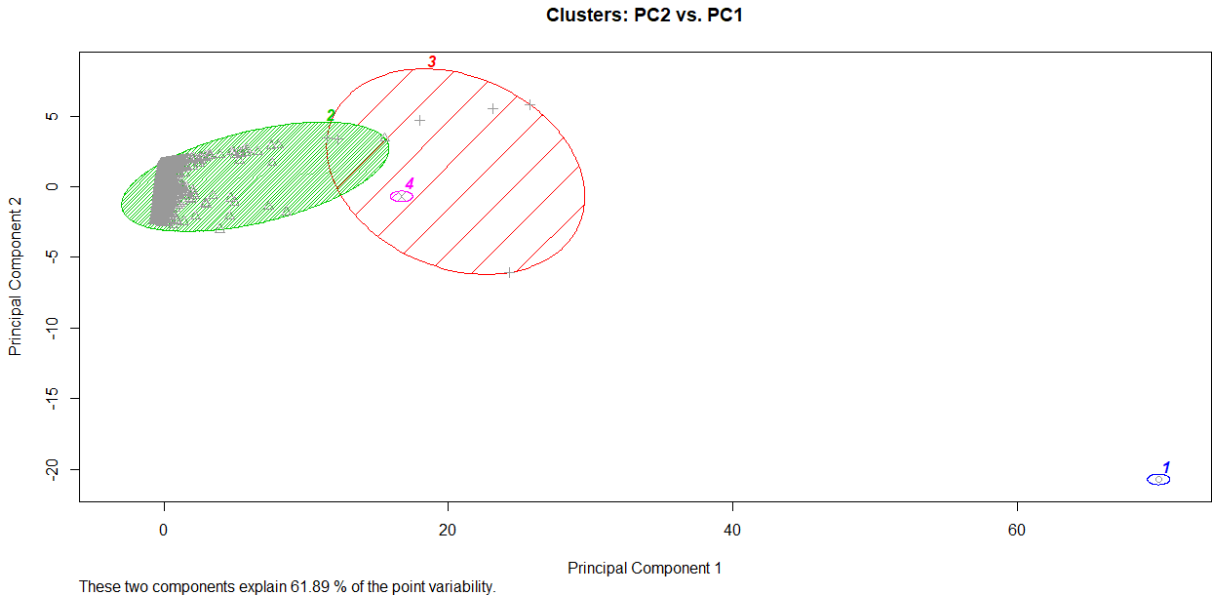

Figure 6. Visualization of the clusters, with observations projected onto the plane created by the first two principal components of the dataset. The gray points represent observations projected onto this plane.

4. Conclusion

In this project, I found that the K-means clustering algorithm and hierarchical clustering algorithm produced significantly different clusters for the UK online retailer dataset. I found more distinct segmentation in the K-means clustering. However, this segmentation did not provide us much meaningful, intuitive information about the customers. I noticed loose trends related to the duration and recency of the customer's interaction with the retailer. However, I also saw a large amount of overlap between the clusters. This illustrates that, in practice, clusters are often nonhomogenous and it is rare to find incredibly well-defined groups.