Online Retail Consumer Segmentation Through Cluster Analysis

Data Analysis Assignment Lesson 11

David Chen

April 14th, 2020

Introduction

For our analysis, we attempt to apply cluster analysis to an online company's customer transactions dataset, attempting to meaningfully segment customers into clusters. Using variables that represent recency of purchase, duration of transaction, amount of purchases, and total spending, we applied a hierarchical clustering model to identify a reasonable number of clusters. Using six clusters, we found that while the observations were relatively well divided between four of the clusters, "outliers" dominated the remaining two. While there does seem to be some segmentation, additional analysis may be required to produce actionable results.

Data

For this report, we will be considering an online retail data set, consisting of transaction records from a UK-based (online) store from December 1st, 2010, to November 30th, 2011. This store predominately sold all-occasion gifts to company wholesalers, so sales were often in bulk. After some initial data cleaning, we have the following variables:

1. `Country` - Country location of customer
2. `Amount` - Amount paid in total (Sterling pounds)
3. `FirstMth` - First month that the customer purchased goods
4. `LastMth` - Last month that the customer purchased goods
5. `Months` - Difference in months between first and last month of purchase
6. `Purchases` - Number of orders
7. `Amount.per.Purchase` - Average number of pounds spent per purchase
8. `Purchases.per.Month` - Average number of purchases per month
9. `Amount.per.Month` - Average amount paid per month

Additionally, since we wish to account for recency, we create an additional variable: `recency`. This variable ranges from 1-12, 12 indicating the most recent purchase being a year ago (Dec, 2010), while 1 indicates a purchase during the most recent month (November, 2011).

EDA

First, we observe summary statistics for the Purchases/Amount variables. We observe that there observes to be extreme outliers, where a company seemingly purchased far more than everyone else. Since there is so much of a skew, we will need to scale our variables. If we do not, variables such as Months (ranging from 1-12) will be completely overshadowed by these variables.

```
##    Purchases    Amount.per.Purchase Purchases.per.Month Amount.per.Month
##  Min.   :   1   Min.   :    0       Min.   :   0        Min.   :    0
##  1st Qu.:  17   1st Qu.:   12       1st Qu.:   7        1st Qu.:  122
##  Median :  40   Median :   18       Median :  13        Median :  221
##  Mean   :  89   Mean   :   55       Mean   :  20        Mean   :  400
##  3rd Qu.:  97   3rd Qu.:   25       3rd Qu.:  24        3rd Qu.:  388
##  Max.   :7376   Max.   :77184       Max.   :1146        Max.   :77184
```

Next, we examine the range and distribution of countries within this data set. There are 36 unique countries, where the UK contains significantly more observations than all the other countries.

Country	count
United Kingdom	3887
Germany	93
France	86
Spain	27
Belgium	22
Portugal	19
Switzerland	19
Italy	14
Finland	11
Norway	10
Austria	9
Channel Islands	9
Netherlands	9
Japan	8
Sweden	8
Australia	7
Denmark	6
Poland	6
Cyprus	5
Canada	4

Analysis

For our analysis, we will attempt to use cluster analysis to identify if the customers can be segmented meaningfully by recency (`recency`), duration (`Months`), frequency (`Purchases.per.Month`), and amount (`Amount.per.Month`). Note that variables such as `Amount` are not included, as it would be redundant as `Amount.per.Month` and `Months` are both present.

Recall that the variables are all standardized so that they all have equal weighting. If this was not included, variables such as `Months` and `recency`, which are on 1-12 scales, would become entirely irrelevant compared to purchases.

Hierarchical Clustering

First, we conduct a hierarchical clustering with complete linkage to form a baseline idea of how many clusters there are and what behaviors exist.

From Figure 1, we first note that since there are 4286 observations, the bottom text is illegible. However, we can still observe that there are some general trends. There are several observations on the left side of the graph that are very dissimilar from the rest of the graph. We can note that a bulk of the points are contained within the split located at height = 10, and that there are a few observations on the right side.

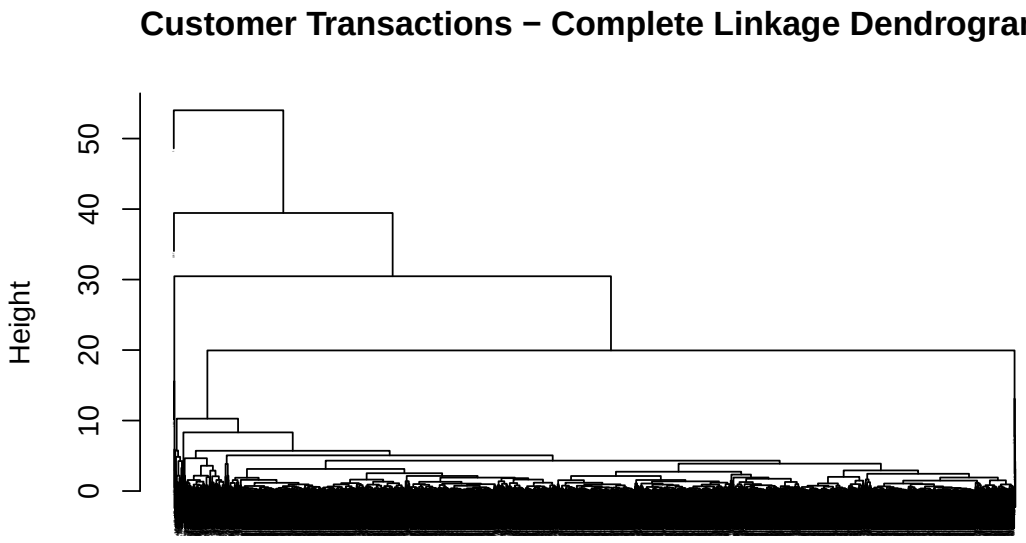

Figure 1: Customer Transcations - Hierarchical Clustering

To get a better idea of the graph, we proceed to cut the graph at height = 10, and observe the behaviors there. Here, it seems like there is about 6 clusters, and it seems to be fairly reasonable. Thus, we proceed forward to K-means analysis to analyze this further.

Post-submission note: we could have used the elbow method to decide the optimum number of clusters for k-means here.

Figure 2: Customer Transcations - Cut Dendrogram

K-Means

After conducting a K-means analysis with 6 clusters, we observe the following counts within each cluster:

```
## [1]   57 1047 1104  853 1223    2
```

We note that the first and sixth clusters both have much lower total counts than the other 4 observations (with 5 clusters, the same two low-count clusters appeared). Additionally, looking at the cluster means, we begin to understand more about these clusters. Clusters 1 and 6 have higher `Amount.per.Month` (standardized) values than the rest, especially cluster 6. It would seem that Cluster 6 captures the extreme outliers noted within the EDA section of the report.

Comparing all the predictors, we can note that clusters 2-5 cover every combination of positive and negative means for `recency` and `Months`. The `Purchases.per.Month` are all extremely similar, with just cluster 1 being far greater than the rest.

recency	Months	Purchases.per.Month	Amount.per.Month
0.521	-0.170	5.292	2.283
1.057	1.250	-0.133	-0.030
-0.934	-0.831	0.018	-0.006
-1.109	0.708	-0.307	-0.131
0.687	-0.804	0.066	-0.046
0.069	-0.969	-0.578	37.227

Countries

Lastly, we want to observe if there is segmentation according to `Country`. Examining the clusters assigned to observations of each country, we observe that there doesn't seem to be any major trends specific to certain countries. It would appear that clusters 1 and 6, which had the low total counts, are still predominately (or entirely) composed of customers from the UK. There does not appear to be any significant segmentation based on country.

```
##
##     Australia Austria Bahrain Belgium Brazil Canada Channel Islands Cyprus
##   1         1       0       0       0      0      0                0      0
##   2         1       0       0       7      0      0                2      0
##   3         1       3       2       3      0      3                3      1
##   4         3       1       0       6      0      0                2      1
##   5         1       5       0       6      1      1                2      3
##   6         0       0       0       0      0      0                0      0
##
##     Czech Republic Denmark EIRE European Community Finland France Germany
##   1              0       0    2                   0       0      1       2
##   2              0       1    0                   0       3     26      29
##   3              0       3    0                   1       4     19      25
##   4              1       2    1                   0       1     12      12
##   5              0       0    0                   0       3     28      25
##   6              0       0    0                   0       0      0       0
##
##     Greece Iceland Israel Italy Japan Lebanon Lithuania Malta Netherlands
##   1      0       0      1     0     0       0         0     0           1
##   2      0       0      0     4     2       0         0     1           3
##   3      1       0      1     1     3       0         0     1           3
##   4      0       1      1     0     1       0         0     0           1
##   5      2       0      0     9     2       1         1     0           1
##   6      0       0      0     0     0       0         0     0           0
##
##     Norway Poland Portugal  RSA Saudi Arabia Singapore Spain Sweden Switzerland
##   1      0      0        0    0            0         0     0      0           2
##   2      3      2        3    0            0         0     9      4           1
##   3      3      4        4    1            0         0     8      2           6
##   4      1      0        3    0            0         1     5      0           4
##   5      3      0        9    0            1         0     5      2           6
##   6      0      0        0    0            0         0     0      0           0
##
##     United Arab Emirates United Kingdom  USA
##   1                    0             47    0
##   2                    0            946    0
##   3                    1            995    2
##   4                    0            793    0
##   5                    1           1104    1
##   6                    0              2    0
```

Conclusion

In conclusion, there seems to be some segmentation possible, especially with 6 clusters. We observed that 2 clusters were contained much fewer counts than the rest, the means having substantially higher spending and purchases per month than the rest. The remaining clusters each had about 1,000 observations each, with varying combinations of recency and total months of purchases.

While these preliminary results do seem to indicate that some degree of segmentation exists, especially in identifying the large spenders/purchasers, additional analysis will likely be required for truly influential segmentation for marketing purposes. This lack of segmentation was highlighted in the country comparison, where there seemed to be no clear differences between countries.