Kevin Maher

Dr. Siripun Sanguansintukul

MSDS680

February 23, 2017

K-means Clustering on the UCI Wholesale Customers Data Set

The assignment is to perform a K-means clustering analysis on the University of California, Irvine, Wholesale Customers Data Set (Cardoso, 2014). This is a customer data set for a wholesale distributer of groceries and related products. We can view this wholesale distributor as our data science client. The goal is to find out whether there are patterns in their customer’s purchasing habits that might be of use to improve marketing the company’s products. We do not have classifications for the customers of the type that we are seeking, these need to be discovered from the data using K-means cluster analysis.

There are no missing values in the data. Most of the data consists of numeric fields that give the customer’s purchases in monetary units by category. Given the date of the data set this is probably Euros since the data is from Portugal, but this information is not confirmed. Five features describe this, “MILK”, “GROCERY”, “FROZEN”, “DETERGENTS\_PAPER”, and “DELICATESSEN”. There are also categorical features for the distribution channel, whether retail or a combination of hotels, restaurants and cafes and one for sales region, either Lisbon, Porto or “Other”. I created one additional numeric feature, a total customer size feature that is the sum of all of the customer’s purchasing categories. This I felt would help separate large customers from smaller ones.

The channel and region features were made more human readable by casting their numeric values to R factors with meaningful strings. These were then one hot encoded. One can be left out when using one hot encoding, but I felt that this made the K-means tables less readable so I left all the values in (Lantz, 2015). By leaving all of the variables in the model there is no need for the user interpreting the R table output to have to impute the missing value.

I applied scaling to the numeric values to prevent any fields with unusually large values from dominating the model. This is especially important given the wide range of the data, outliers and my added total feature which could dominate the model if not scaled. As it is, even scaled, the outliers may come to have a major impact on the model. This can be observed easily with a table, as all of the numeric variables contain apparently contain outliers much higher than the mean, large customers who buy much more than an average amount of product (see figure 1).

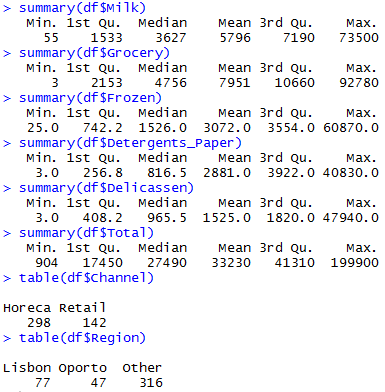


Figure 1. Exploratory Data Analysis

A sum of squares plot was made for different numbers of clusters, see figure 2 (Ben, 2014). This plot is somewhat susceptible to the random seed used to create it. However, it still is useful. Looking for a knee in the plot is one recommended way to find an optimal value for the number of clusters. From this plot it appears that 5 or 6 might make a good choice without too much model complexity. I experimented with the K-means algorithm and found that the small bend at 3 also seems to produce an interesting model. The model for k = 3 will be tackled first since it is the simplest.

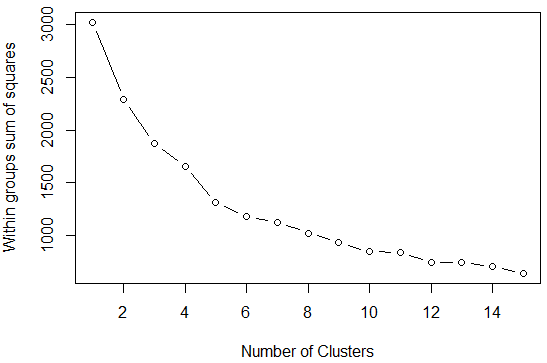


Figure 2. Sum of Squares Plot for Different Numbers of Clusters.

The model for k = 3 is striking. It segments our market into a large group of smaller customers and two groups of large customers. These two groups of large customers are very different in their buying habits (see figure 3). We can tell that the segmentation into groups 2 and 3 is based in part on customer size because of the “Total” field that was added. Groups 2 and 3 show much higher total purchases than group 1. What is also important to notice is that our two large customer groups have very different buying habits. For example, group 2 buys a lot of fresh goods from our client but not much milk, group 3 buys a lot of milk but not many fresh goods. This pattern is repeated for all of the goods types that our client sells, many sales to one of our large customer groups but low sales to the other. These two groups also segment by distribution channel, but not so much by region, showing that distribution channel is an important predictor of what our client is currently selling to its largest customers.

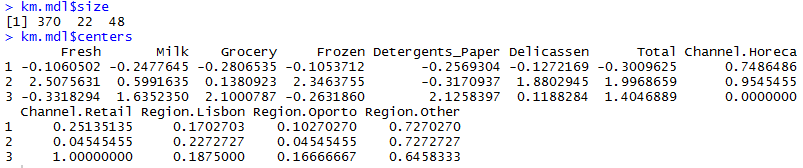


Figure 3. K-means Cluster Table for K = 3.

These insights might open up opportunities, for example, why would retail customers buy lots of milk but very little in the way of similarly perishable fresh goods from our client? We cannot answer this from the data, but what is important about the data analysis is that we learned to ask this question from doing the clustering analysis. I believe that this clearly presented finding makes a case for k = 3 as being an excellent model to show to a management group. For example, consider figure 4. It shows clearly that Groceries are sold highly in the retail channel but not to the “Horeca” (hotel, restaurant, café) channel. I believe that it is a clear and easy to understand graph of the results of the model. It also shows that fresh goods are sold to the Horeca channel, but not so much to the retail channel. While it is possible that these are characteristics of the channel, it is also possible that there are marketing opportunities to sell the under-represented goods into the channels where they currently do not sell well.

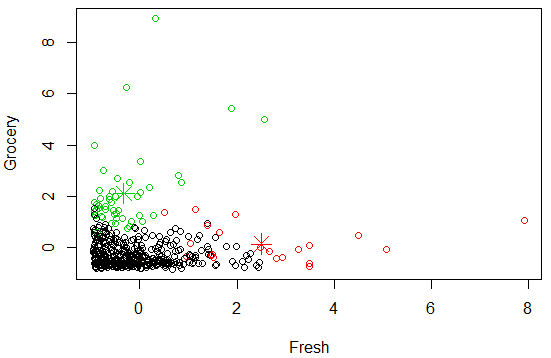


Figure 4. Groceries vs. Fresh Goods for Large Customers.

Next to be tried was to set k = 5. This did not work as well as hoped since one cluster had only 3 members and a second cluster had only 11. This means that setting k = 5 in essence found 3 meaningful clusters. Clusters 2 and 3 contain the largest customers. While there is some evidence of segmentation by purchasing types, this is not as clear as when k was set to 3. In fact, looking at the “Total” column, this model seems to have clustered by customer size rather strongly. There are some clear patterns, for example, Delicatessen (misspelled in the data set as Delicassen) sales appear to track by customer size as might be expected since the model segments fairly strongly by customer size.

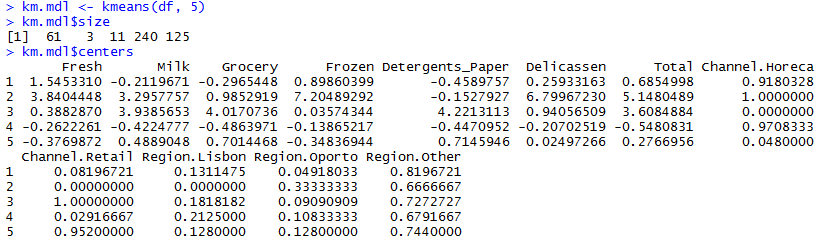


Figure 5. Clustering Result for K = 5.

Segmentation also shows up in the charts for k = 5 where the groups tend to buy one, but not another product type from our client (see figure 6). This shows that we have two groups of customers who tend to buy detergents and paper, but not to buy frozen products from our client. We also have two groups that are the inverse, buying frozen goods but not detergents and paper.

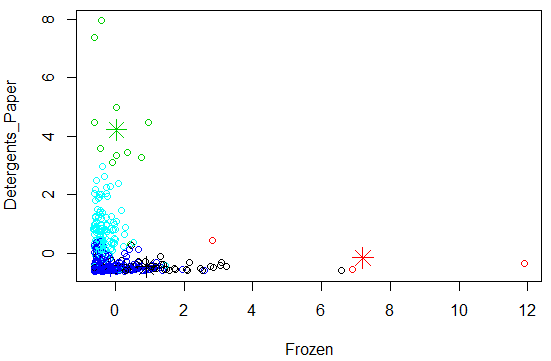


Figure 6. Segmentation of Clusters by Frozen Products and Detergents and Paper.

Another k value that might be tried based on a small knee in figure is k = 12. This produces a complex chart that is more difficult to interpret, but still offers useful insights. Some groups consist of as little as one or two points. These are large customers and in my view are now excessively segmented by the model. However, we can now gain some insight into the behavior of our smaller customers.

For example, consider cluster 12, our smallest customers according to the “Total” column. Relatively speaking these customers buy delicatessen products but not fresh goods. They are all in the “Horeca” channel. It could be an interesting marketing exercise to investigate why they do not buy much from the fresh goods category. Perhaps they don’t need these items, or perhaps we don’t do a good job of selling to these customers. Our largest cluster in the k = 12 model is cluster 10, a largely retail group of customers. Again, the model asks questions for us. Such as why we sell perishable milk in this channel but not fresh goods which are similarly likely to be perishable.

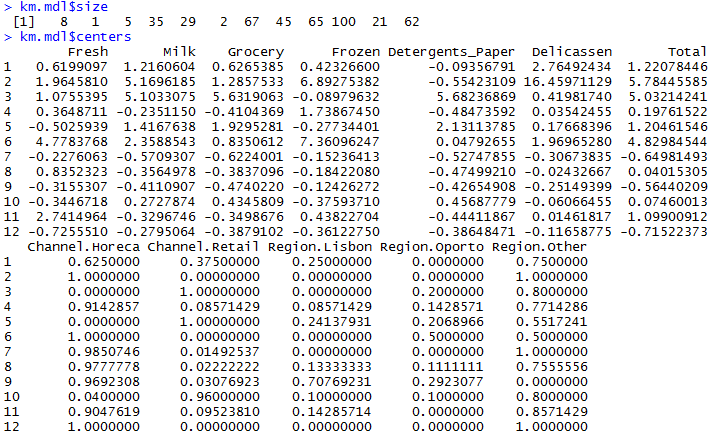


Figure 7. Cluster Chart for K = 12.

Scatter charts become very hard to interpret with this many clusters, especially where most of the data is concentrated in the lower left corner of the chart. Still if we look at one such chart we can observe a similar pattern to that observed when k = 3 (see figure 8). Specifically that we have a group of customers who but some goods from our client but not others. Note the cyan cluster to the left, the green, grey and yellow clusters towards the lower part of the chart. While more work to interpret, the model for k = 12 also gives useful insights for smaller customers that would recommend targeted market research based on the model’s results. This can be seen especially in figure 9 which zooms in on the smaller customers to show their groupings.

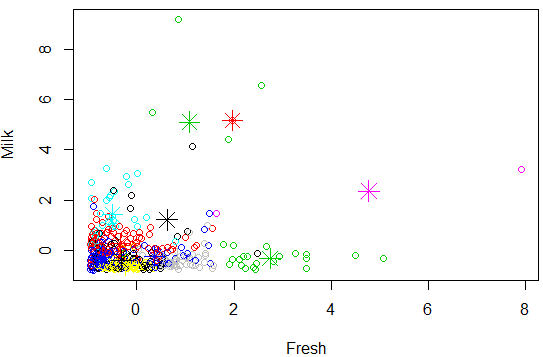


Figure 8. Milk vs. Fresh Goods for K = 12

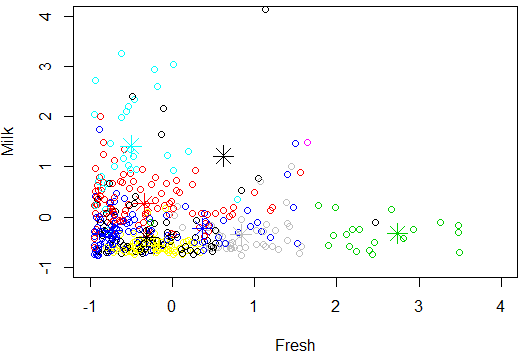


Figure 9. Zoom in Lower Left of Figure 8.

One issue with the models is that there are a large number of smaller customers that are difficult to segment and graph because of the size and influence of the larger customers. This is true even with Z-score scaled data. I tried taking the log of the numeric fields in an attempt to make the data more normal, and then scaling this result. Then a sum of squares plot was done in an attempt to find a new optimal value for k (see figure 10). This chart did not help much, the only knee in the chart that I observed was at k = 2 and since this segments the data strongly by customer size without dividing those large customers into segments, I did not see any benefit in this approach so it was not continued further.

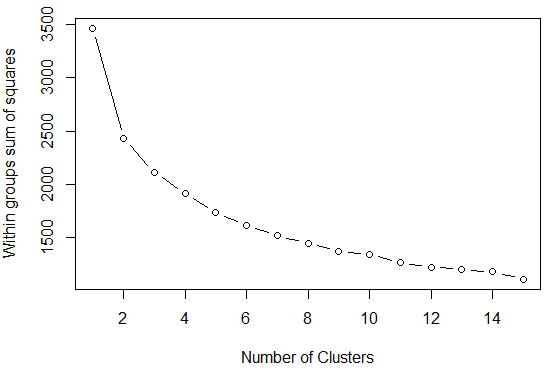


Figure 10. Sum of Squares Graph for Log Taken on Numeric Fields.

Clusters of the smaller customers by total purchases are difficult to detect and plot well. A future version of this model could be enhanced by splitting off the large customers from the smaller ones as detected by the k = 3 model. This would offer plotting advantages for the smaller customers as then they would be plotted across the entire range of the plot axis rather than being confined to the lower left corner which is being caused by the effect of the large outlier customers.

One of the goals was to find an optimal value for k for this model. I do not believe that there is a single value of k that is best for all circumstances with this model. K = 3 gave insights into two classes of large customers, k = 12 gave some information about smaller customers but overly segmented the large customers so that some clusters for them had only 1 or 2 members. I believe that this model should help the client’s marketing department know what further research to conduct, for this purpose k = 3 and k = 12 both produce good models. I would use them both together if I had the resources to conduct the marketing research that the model indicates might offer insights and opportunities for the client’s business.

References:

Cardoso, Margarida (2014). Wholesale Customers Data Set. University of California, Irvine, Machine Learning Repository. Retrieved from: <https://archive.ics.uci.edu/ml/datasets/Wholesale+customers>.

Lantz, Brett (2015). *Machine Learning with R, 2nd Edition*. Packt Publishing. Birmingham, UK.

Ben (2014). Cluster analysis in R: determine the optimal number of clusters. Stack Overflow. Retrieved from: <http://stackoverflow.com/questions/15376075/cluster-analysis-in-r-determine-the-optimal-number-of-clusters>.

Attachments:

Kmeans-exercise.R: R code file for the exercise