Problem Set 9

QMBE 3740: Data Mining

Module: Clustering with K-means

Part I: Segmenting Colleges

Exercise 1: Using the college dataset from chapter 12, perform clustering that looks at the average faculty salary and annual tuition rates for schools located in Indiana. Choose $k = 3$ and produce a visualization of your clusters.

Exercise 2: Use the techniques described in Chapter 12 to select two possible optimal values for k for the clustering problem you coded in Question 1. Justify your answer.

Exercise 3: Generate cluster diagrams for the two values of k that you selected in Exercise 2. Which one of these do you believe is the best result? Why?

Part II: Segmenting the Cereal Market

Exercise 4: Using the `Cereals.csv` dataset, read in the data and drop any missing values using `drop_na()`. Then select all variables except `name, mfr, type, weight, shelf, cups, rating` to create a subset of several features we will use to cluster the different cereals. Finally, use the 3 methods we've discussed to determine a good number of clusters for the data, making sure to generate and include the visualizations (`fviz_nbclust`) in your final document.

Exercise 5: Using the number of clusters you determined in Exercise 4, perform k-means clustering and printout the cluster centers. Then create a new column in the data called `cluster` which contains the cluster each observation belongs to. Then create graphic which plots the cluster centers as points in the sugars-calories space. (Fill in the appropriate parts of the code below to do this)

```
ggplot(data = ___,
       aes(x = scale(sugars), y = scale(calories), color = factor(cluster), size=2.5)) +
  geom_point()
```

Exercise 6: Based on the location of the cluster centers in Exercise 5, give each cluster an intuitive name such as "low calorie - average sugar", etc.

Part III: Segmenting Bathsoap Customers

Exercise 7: Read the data `BathSoapHousehold.csv` and determine the number of clusters to use for segmenting the customers on the basis of the `CHILD` and `Affluence Index` features, making sure to scale the data appropriately. Make sure to use all three methods covered in class and the book and include the necessary visualizations and justification for your final choice.

Exercise 8: For your chosen number of clusters, visualize the clusters (include the `fviz_cluster` plot) and instead of the option `repel = TRUE` use the option `geom = c("point")` which will plot only points to make things less cluttered. Then for each of your clusters, write down a short and business-intuitive description of the group represented by each cluster.

Exercise 9: Create a table of the average `Value` and `Total Volume` for each cluster (use the provided code below) and then determine which cluster has the highest average `Value` to the business and which cluster represents the largest purchasing `Total Volume`. Are they the same cluster? If not, then using your business understanding of the types of customers in these clusters, offer an explanation as to why they are not the same.

```
# I used dataset named "soap"
# I named the kmeans output "k_clust"
# You may need to adapt for your names
soap %>%
  mutate(cluster = k_clust$cluster) %>%
  group_by(cluster) %>%
  summarise_at(vars(Value, `Total Volume`), funs(mean))
```