Clustering for Cereal dataset

The Breakfast cereal data set (https://www.kaggle.com/code/jeandsantos/breakfast-cereals-data-analysis-and-clustering) contains information about a variety of breakfast cereal brands and their attributes. The data set aims to provide insights into the characteristics of different cereal brands. In this project, we try to cluster various types of breakfast cereal based on their nutritional content.

- Summary Statistics

	calories	protein	fat	sodium
X	Min. : 50.0	Min. :1.000	Min. :0.000	Min. : 0.0
X.1	1st Qu.:100.0	1st Qu.:2.000	1st Qu.:0.000	1st Qu.:130.0
X.2	Median :110.0	Median :3.000	Median :1.000	Median :180.0
X.3	Mean :106.9	Mean :2.545	Mean :1.013	Mean :159.7
X.4	3rd Qu.:110.0	3rd Qu.:3.000	3rd Qu.:2.000	3rd Qu.:210.0
X.5	Max. :160.0	Max. :6.000	Max. :5.000	Max. :320.0

Table 1: Summary Statistics

	fiber	carbo	sugars	potass
X	Min. : 0.000	Min. :-1.0	Min. :-1.000	Min. : -1.00
X.1	1st Qu.: 1.000	1st Qu.:12.0	1st Qu.: 3.000	1st Qu.: 40.00
X.2	Median : 2.000	Median :14.0	Median : 7.000	Median : 90.00
X.3	Mean : 2.152	Mean :14.6	Mean : 6.922	Mean : 96.08
X.4	3rd Qu.: 3.000	3rd Qu.:17.0	3rd Qu.:11.000	3rd Qu.:120.00
X.5	Max. :14.000	Max. :23.0	Max. :15.000	Max. :330.00

Table 2: Summary Statistics

- Matrix scatterplot and Correlation matrix

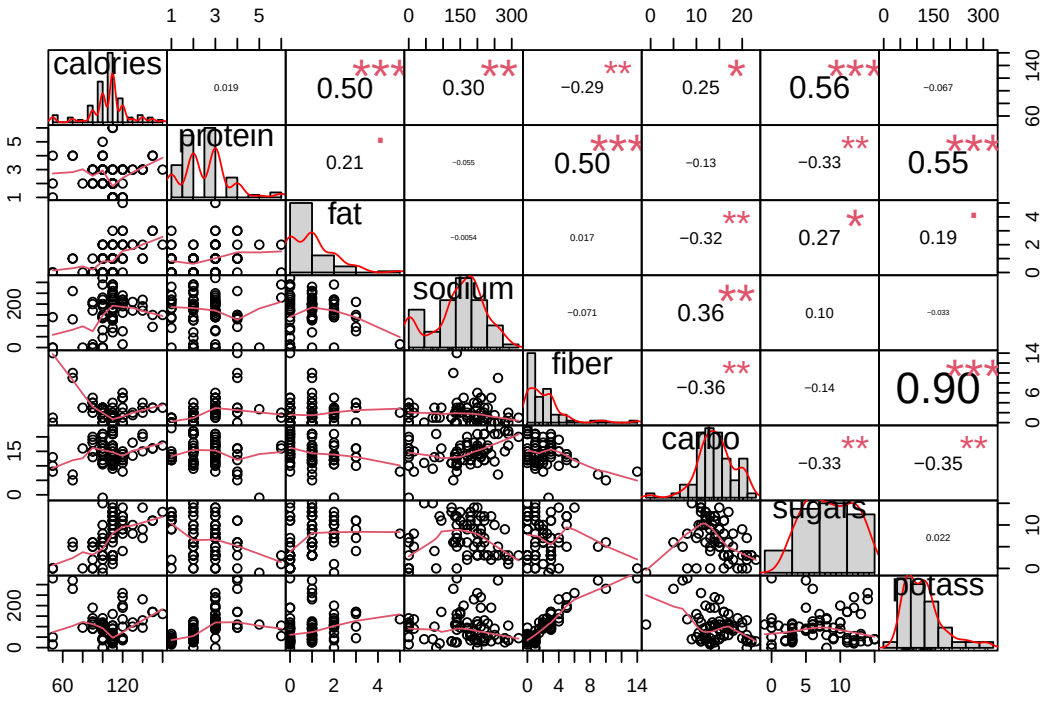

Figure 1: Matrix scatterplot

	calories	protein	fat	sodium	fiber	carbo	sugars	potass
calories	1.00	0.02	0.50	0.30	-0.29	0.25	0.56	-0.07
protein	0.02	1.00	0.21	-0.05	0.50	-0.13	-0.33	0.55
fat	0.50	0.21	1.00	-0.01	0.02	-0.32	0.27	0.19
sodium	0.30	-0.05	-0.01	1.00	-0.07	0.36	0.10	-0.03
fiber	-0.29	0.50	0.02	-0.07	1.00	-0.36	-0.14	0.90
carbo	0.25	-0.13	-0.32	0.36	-0.36	1.00	-0.33	-0.35
sugars	0.56	-0.33	0.27	0.10	-0.14	-0.33	1.00	0.02
potass	-0.07	0.55	0.19	-0.03	0.90	-0.35	0.02	1.00

Table 3: Correlation matrix

- As the correlation between most of the variables are very low, therefore, rather than going with the correlation-based distance, we can use matric-based distance for clustering.
- Also I would suggest standardizing the variables as they are in different scales and some have very high ranges.
- The panel histograms shows that the distributions of some of the variables are highly right skewed.

Hierarchical Clustering with Complete linkage

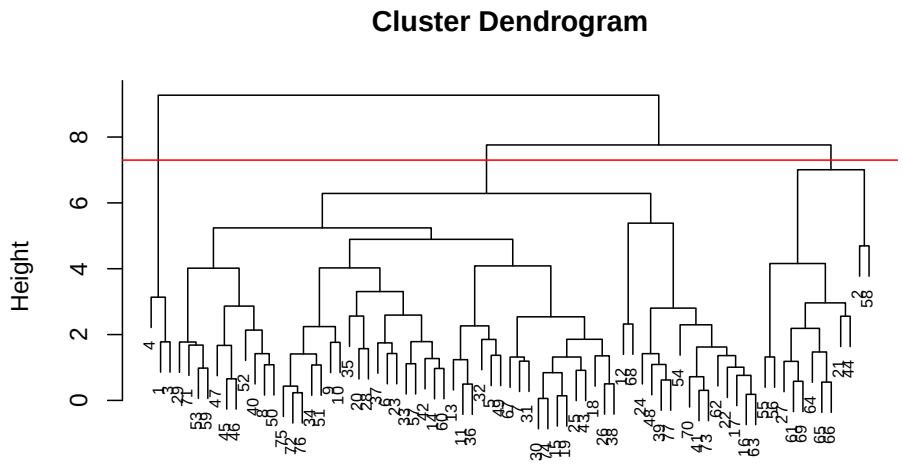

Figure 2: Hierarchical Clustering with Complete linkage

	1	2	3
1	3	12	62

Table 4: Number of observations within each cluster (Hierarchical Clustering)

	calories	protein	fat	sodium	fiber	carbo	sugars	potass
Cluster 1	63.333	4.000	0.667	176.667	11.000	6.667	3.667	310.000
Cluster 2	88.333	2.750	0.667	9.167	2.058	13.833	2.333	88.250
Cluster 3	112.581	2.435	1.097	187.984	1.742	15.129	7.968	87.242

Table 5: Cluster means of the variables (Hierarchical Clustering)

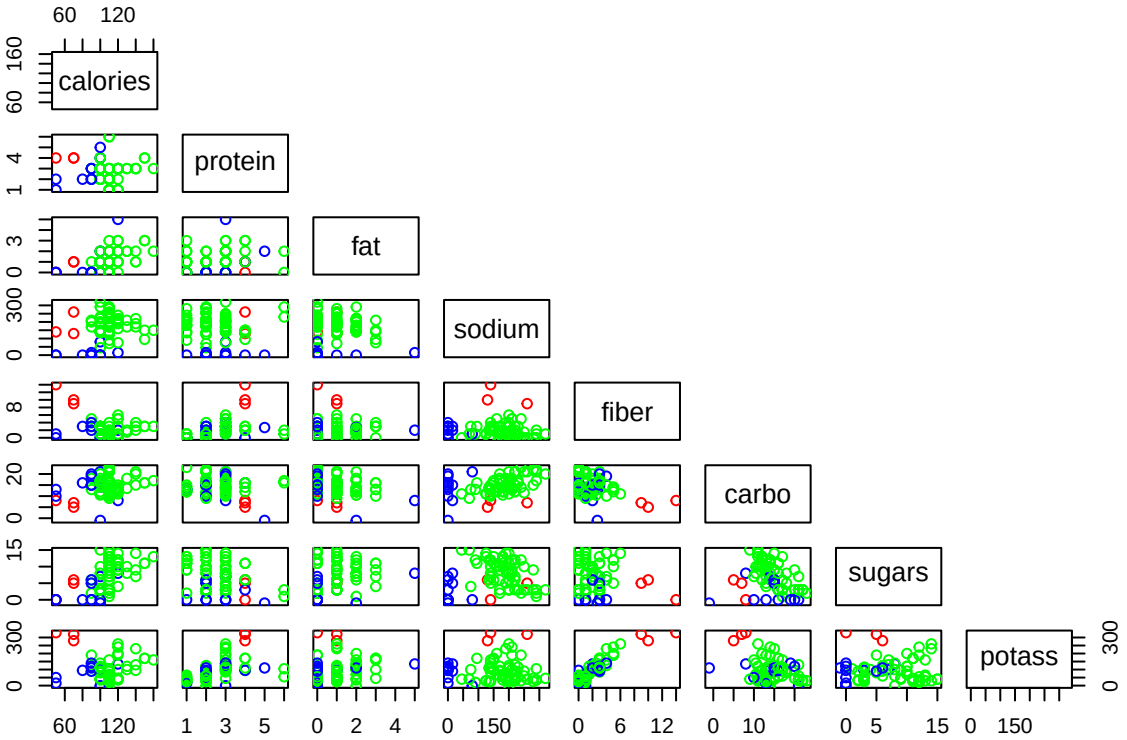

Figure 3: Matrix scatter plots

- We can see that Calories and Sodium are good variables to achieve 3 clusters as there is minimal overlap between the clusters.

K-means clustering with k = 3

	1	2	3
1	31	24	22

Table 6: Number of observations within each cluster (k=3)

	calories	protein	fat	sodium	fiber	carbo	sugars	potass
Cluster 1	98.710	2.516	0.387	152.742	1.677	17.806	3.226	72.710
Cluster 2	113.750	3.500	1.792	159.167	4.196	12.375	7.667	172.917
Cluster 3	110.909	1.545	1.045	170.000	0.591	12.500	11.318	45.182

Table 7: Cluster means of the variables for k-means clustering (k=3)

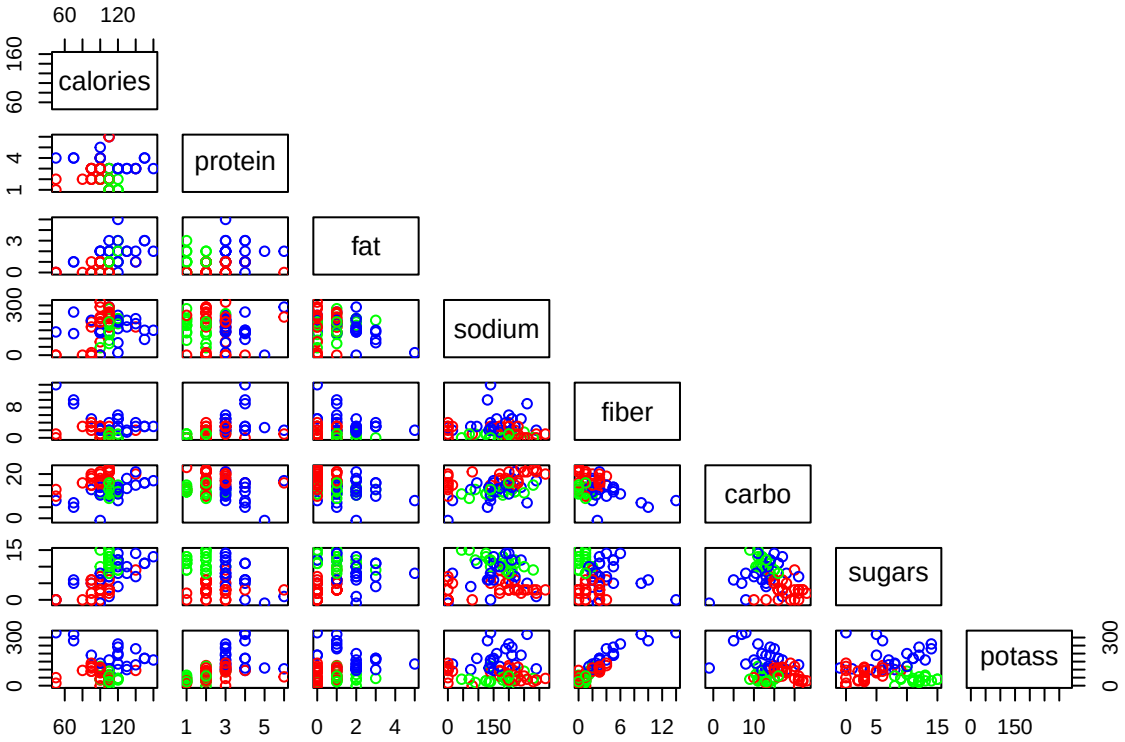

Figure 4: Matrix scatter plots

- The clusters seems to be overlapping.

K-means clustering with k = 4

	1	2	3	4
1	3	20	32	22

Table 8: Number of observations within each cluster (k=4)

	calories	protein	fat	sodium	fiber	carbo	sugars	potass
Cluster 1	63.333	4.000	0.667	176.667	11.000	6.667	3.667	310.000
Cluster 2	124.500	3.400	2.050	155.750	3.135	13.500	8.550	150.750
Cluster 3	97.188	2.562	0.375	153.438	1.781	17.469	3.188	76.844
Cluster 4	110.909	1.545	1.045	170.000	0.591	12.500	11.318	45.182

Table 9: Cluster means of the variables for k-means clustering (k=4)

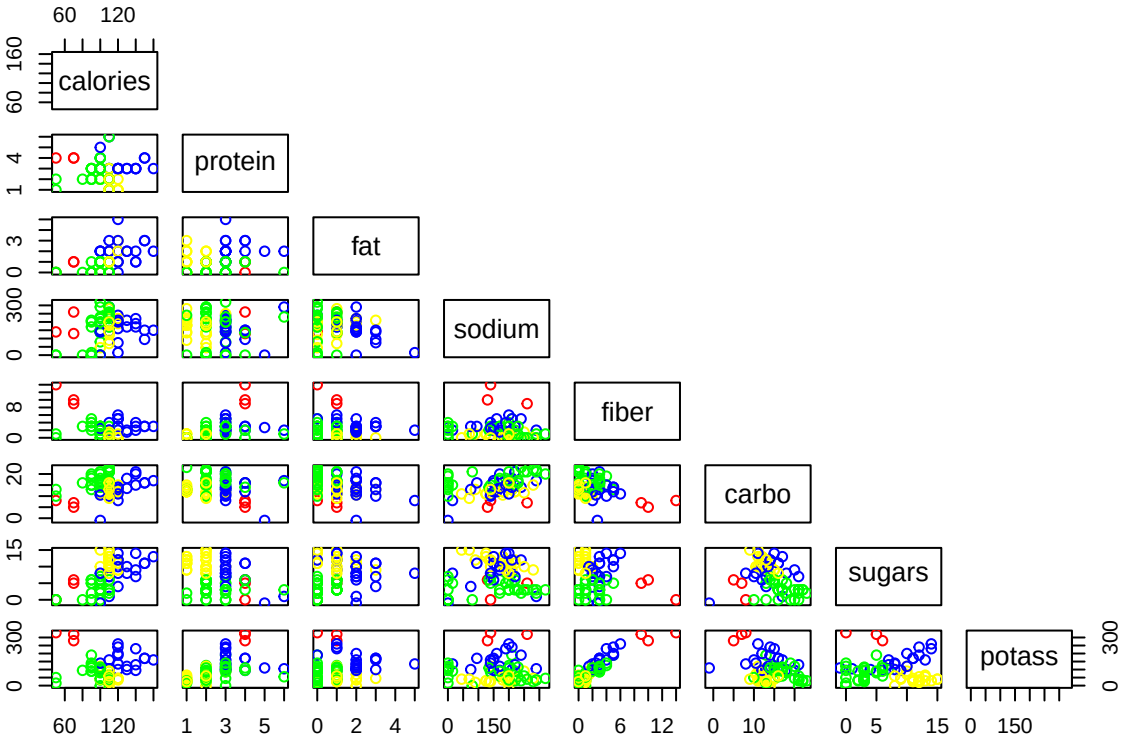

Figure 5: Matrix scatter plots

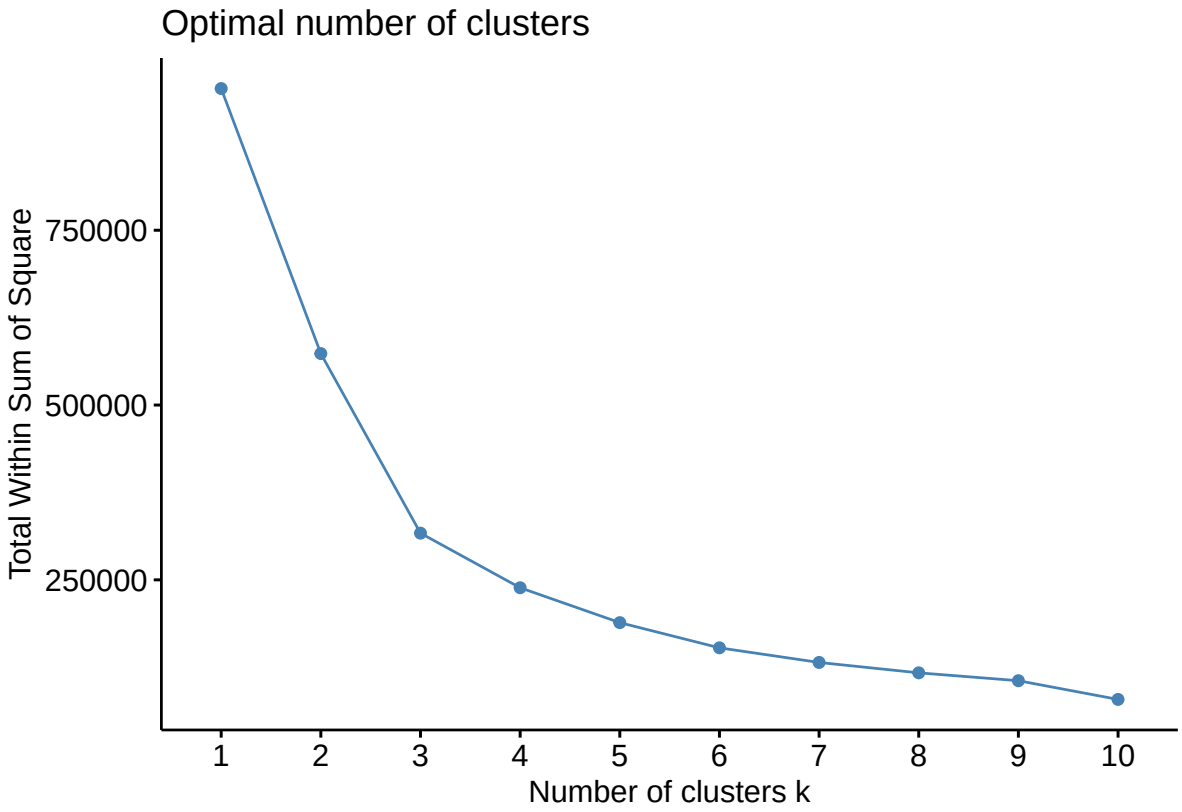

Figure 6: Plot of number of clusters vs total within cluster sum of squares

- The scree plot also shows that the optimal number of clusters is 4.
- If we compare the 2 methods Hierarchical clustering and K-means clustering, both methods seems to give good results for this data. The classification seems to be difficult due to the presence of some extreme values.