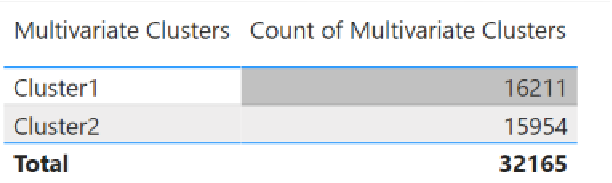
Assignment3

Author: Jingyi Wu (jingyiw2)

PART A:

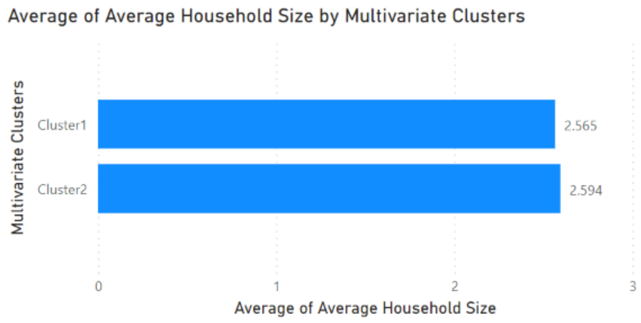
a) Here we removed region_id for clustering, since it makes no sense and is not an attribute of people.

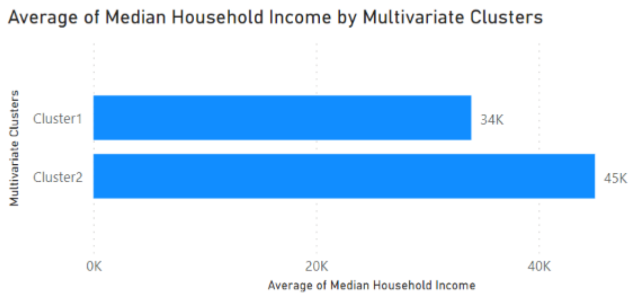
b) Use PowerBI to do clustering and the result is as below. 1013 instances contain missing values in their attributes. Therefore, only 32165 instances are considered here for clustering and PowerBI automatically clustered them into 2 groups.

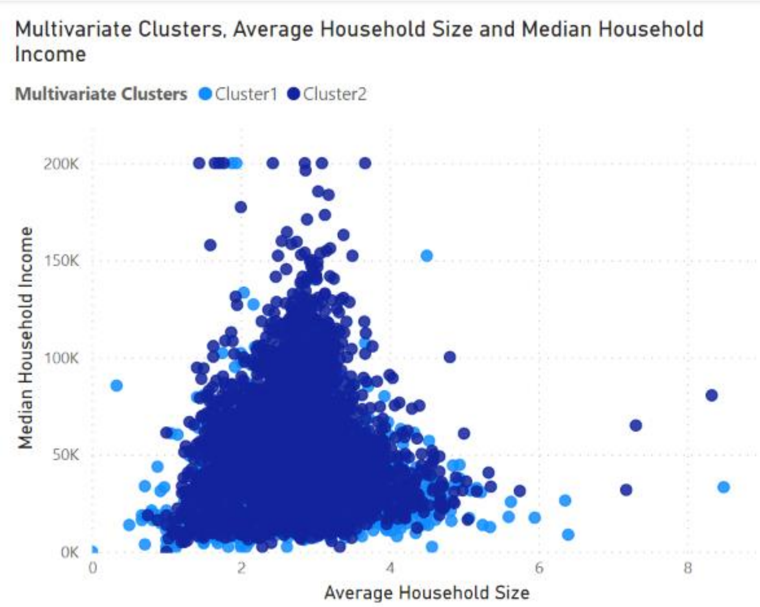


Multivariate Clusters	Count of Multivariate Clusters
Cluster1	16211
Cluster2	15954
Total	**32165**

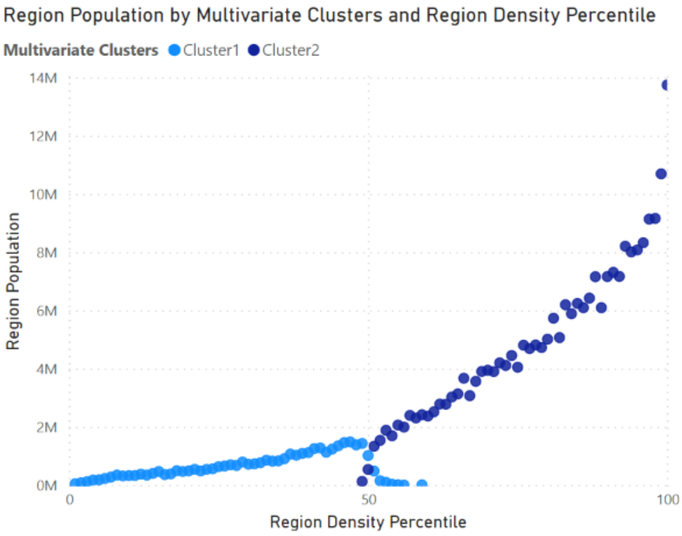
Here we can have a look at the feature distribution for each cluster to find out the features about clusters.



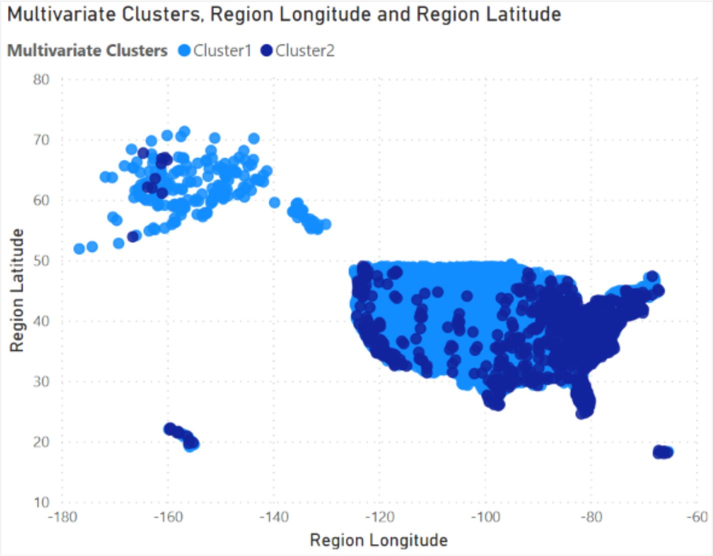




From the above charts, we can see that regions in cluster 1 generally have a slightly smaller household size and lower median household income.



As for the aspect of population size and density, regions in cluster1 have a lower population density and smaller population on average.



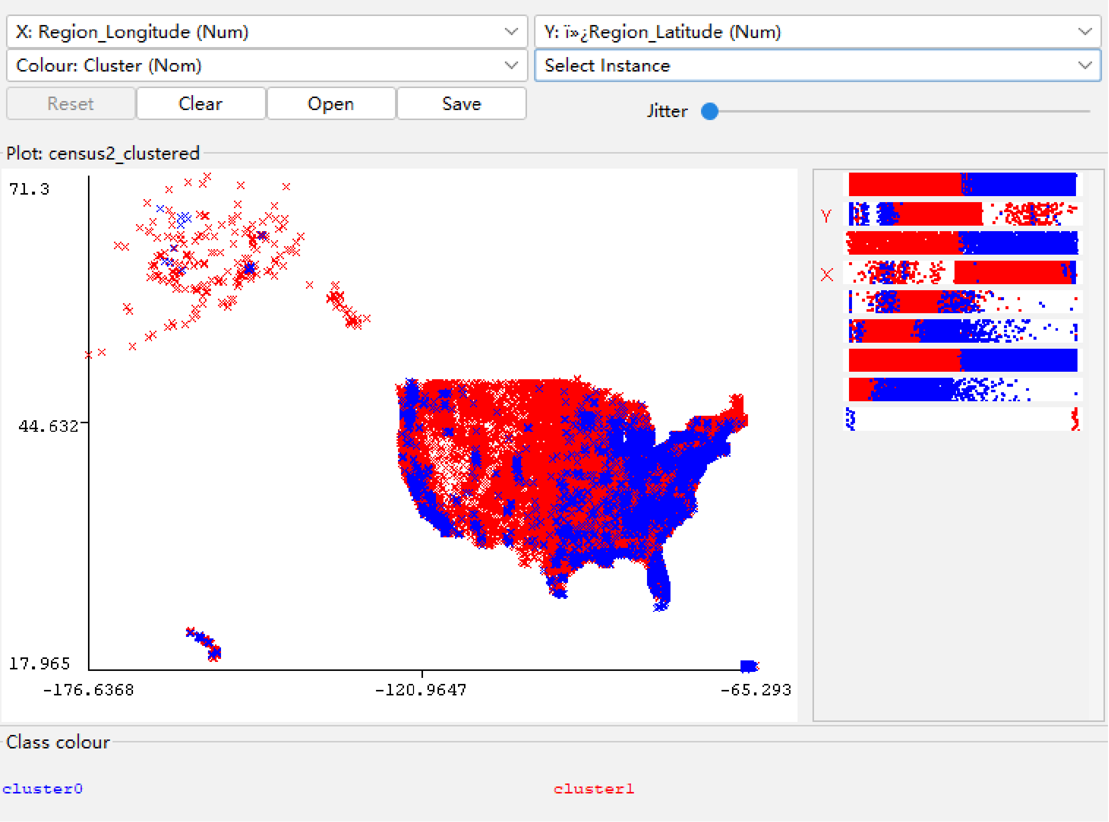
As for location side, regions in cluster 1 tend to have a higher latitude and mostly locate in the northern and middle USA. The east coast of USA and west coast of USA are mostly in cluster2.

In general, these charts show that cluster 2 include mostly regions from coastal areas of USA, which have a larger population and higher population density and people here have a slightly bigger family and higher household income. This is in accord with reality.

c) Apply simple k-means in weka and we will start with k=2, the result is as below.

```
Clustered Instances
0        15957 ( 50%)
1        16208 ( 50%)
```

Within cluster sum of squared errors: 34188.93. The visualization is similar to PowerBI's result.



X: Region_Longitude (Num)

Y: ï»¿Region_Latitude (Num)

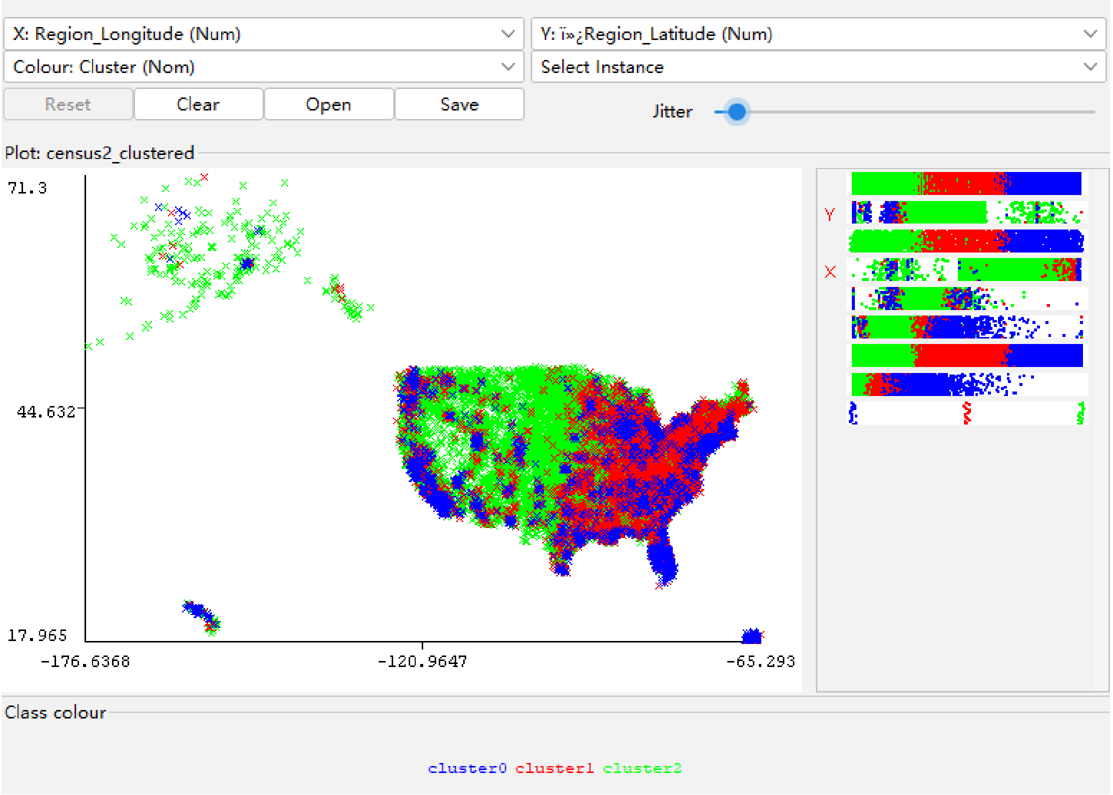
Colour: Cluster (Nom)

Select Instance

Try k=3

Clustered Instances, within cluster error here is 33738.91.

0	10818 (34%)
1	11223 (35%)
2	10124 (31%)



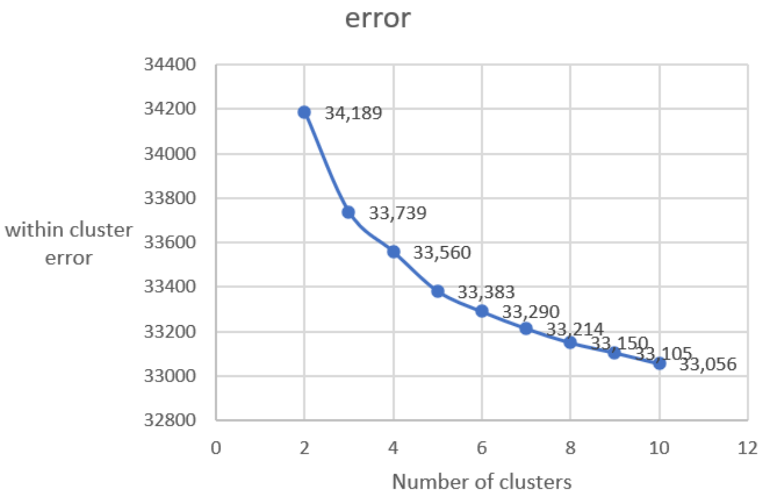
X: Region_Longitude (Num)

Y: ï»¿Region_Latitude (Num)

Colour: Cluster (Nom)

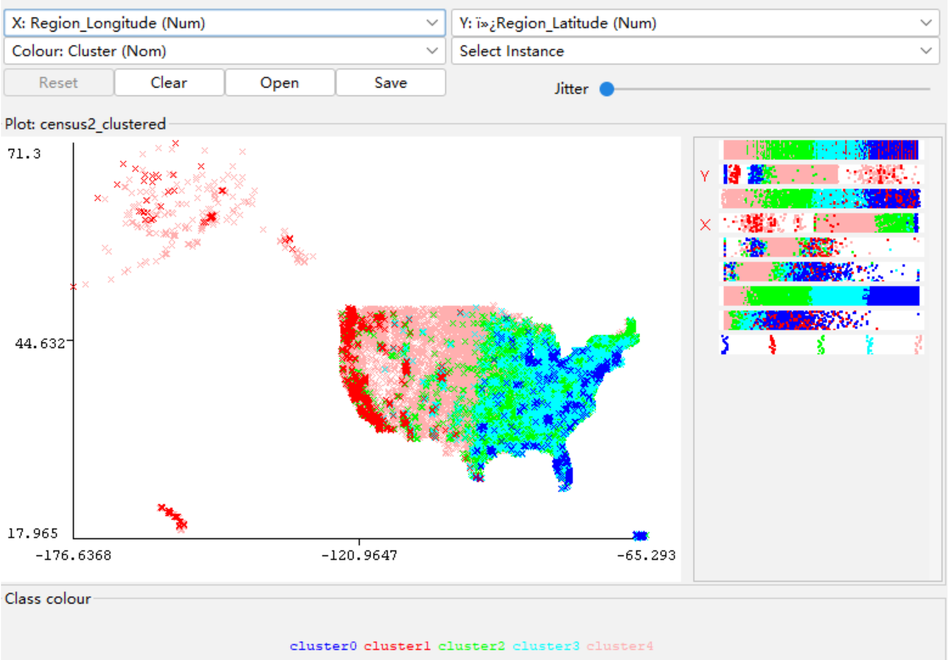
Select Instance

Similarly, we can get a plot of k number and within cluster sum of squared error, and as we increase k, some cluster get really small and include only a small portion of instances, like 5%.



As we can see, increasing the number of clusters will decrease within cluster error but the heterogeneity between different clusters will decrease as well. Here choose a k which makes the error drop most quickly, so k=2 is indeed the best choice.

K=5 may also work. We can get more nuanced clusters.


X: Region_Longitude (Num)

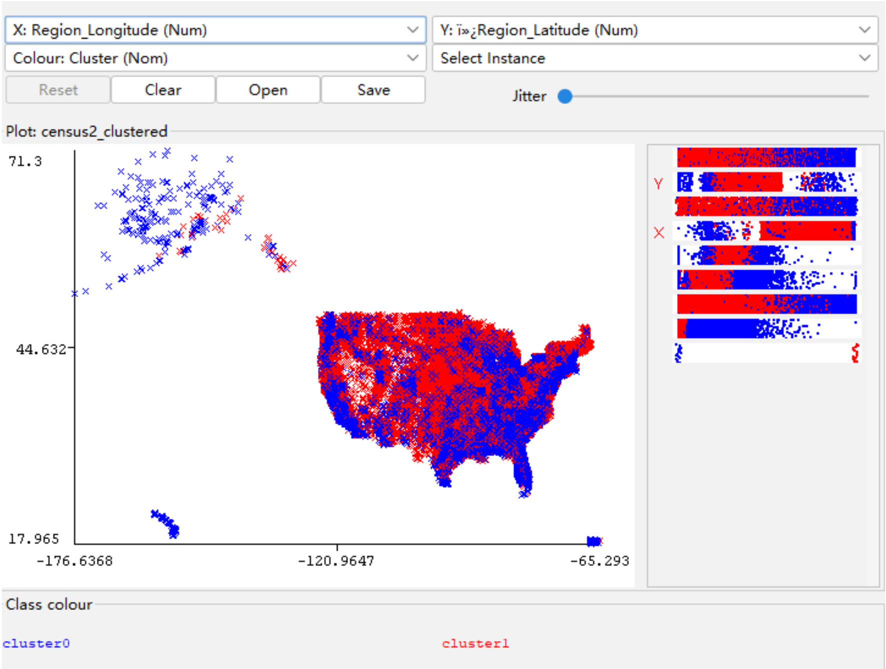
Y: ï»¿Region_Latitude (Num)

Colour: Cluster (Nom)

Select Instance

Plot: census2_clustered

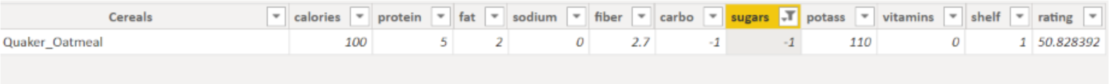
d) Try EM algorithm for clustering. K=2, we can see the result is mostly similar to k-means, but some minority instances are classified into different cluster.



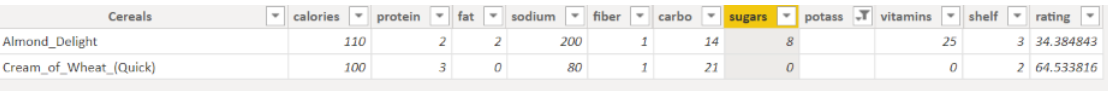
Combined with reality, the upper left instances are from Alaska, the third-least populous and mostly sparsely populated state. It should be classified into cluster0, which include less populated areas. Therefore, in this case, k-means is a better option.

PART B:

a) There are missing values in potass column and wrong value in carbon and sugar columns.



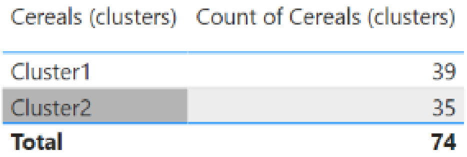
Cereals	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	rating
Quaker_Oatmeal	100	5	2	0	2.7	-1	-1	110	0	1	50.828392



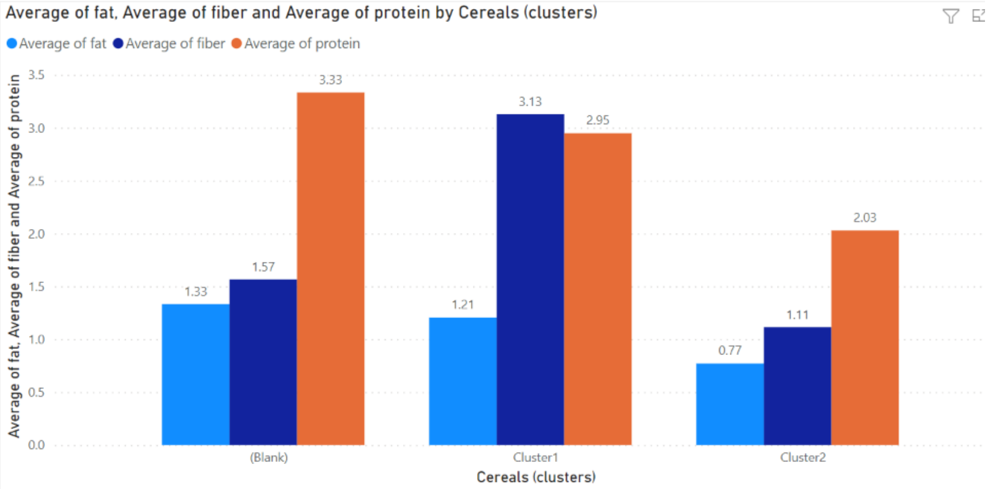
Cereals	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	rating
Almond_Delight	110	2	2	200	1	14	8		25	3	34.384843
Cream_of_Wheat_(Quick)	100	3	0	80	1	21	0		0	2	64.533816

Since this data has no apparent numeric relationship with existing data and there is no way for us to forcast a reasonable value here, for clustering I will use what weka automatically does – replace missing values with mean/mode. As for the wrong values(it's impossible that sugar and carbon here are -1), I will set them to nan value also and do the same processing.
Also transform shelf to nominal value.

b) Use PowerBI, we temporarily ignore instances with missing values and do the clustering, and the result is as follows.

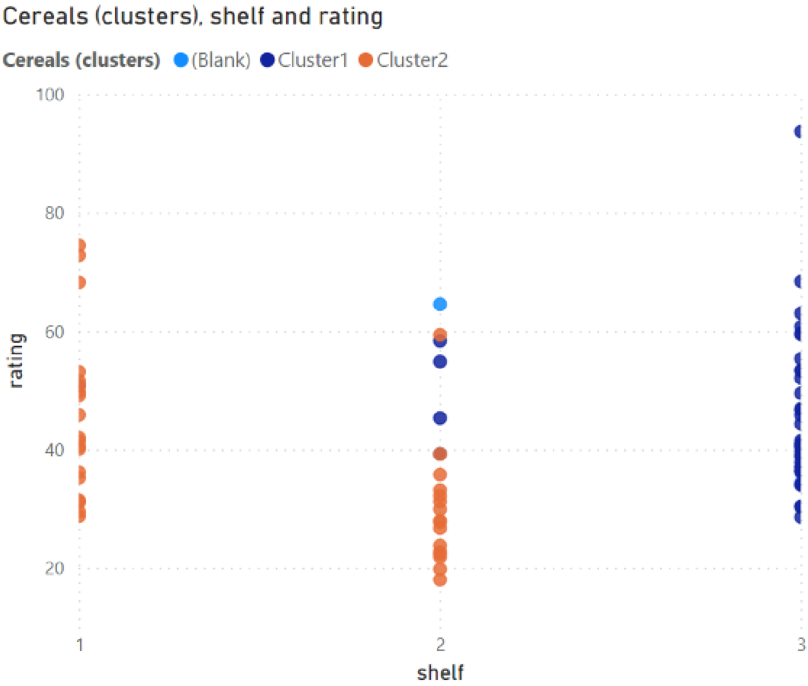


Cereals (clusters)	Count of Cereals (clusters)
Cluster1	39
Cluster2	35
Total	**74**

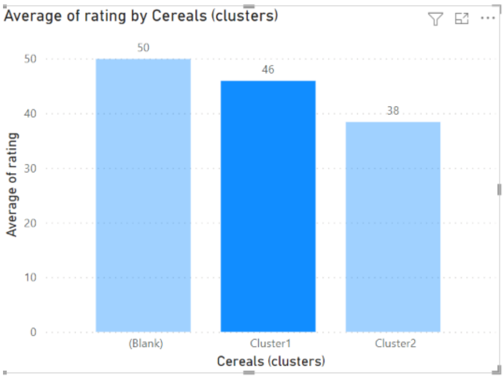


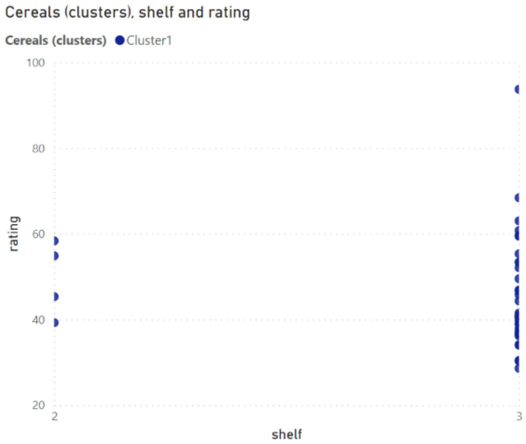
Average of fat, Average of fiber and Average of protein by Cereals (clusters)

we can see that for cluster1 cereals, they have more protein and fiber, which are more beneficial.

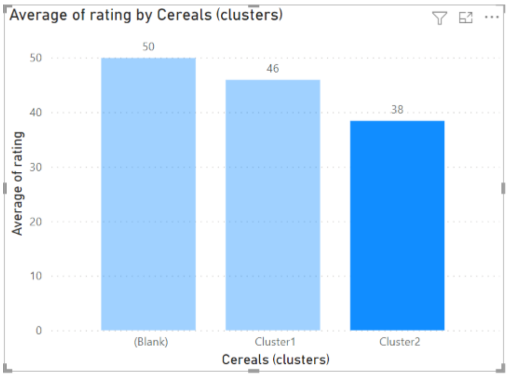


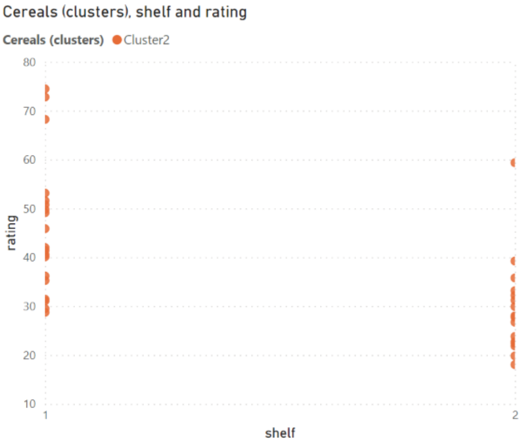
Cereals (clusters), shelf and rating





We can see that cereals in cluster 1 tend to have better ratings and mostly are placed on 3rd display shelf.

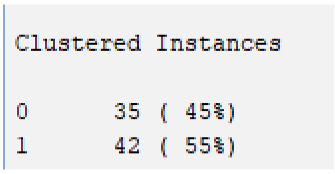




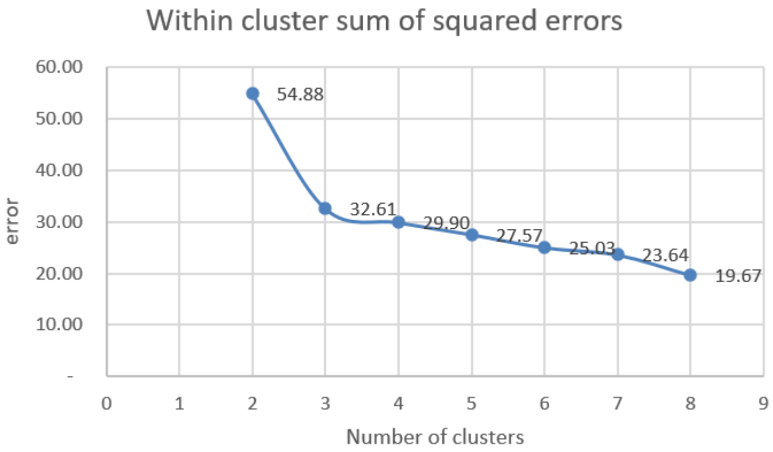
Cluster2.: lower rating and both 1 and 2 shelves have them.

c) Use K-means in weka and we start from k=2 also. The result is as below. (Missing values globally replaced with mean/mode)

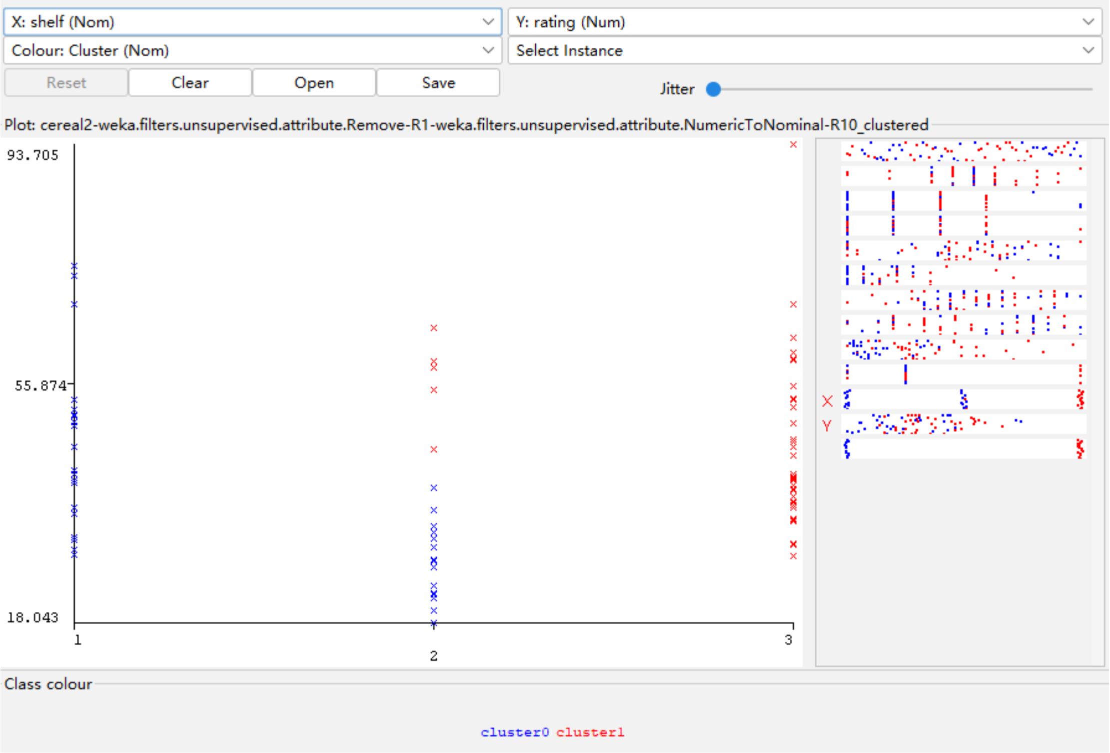
Clustered Instances



```
Clustered Instances

0       35 ( 45%)
1       42 ( 55%)
```



Here k=2 is the best choice.



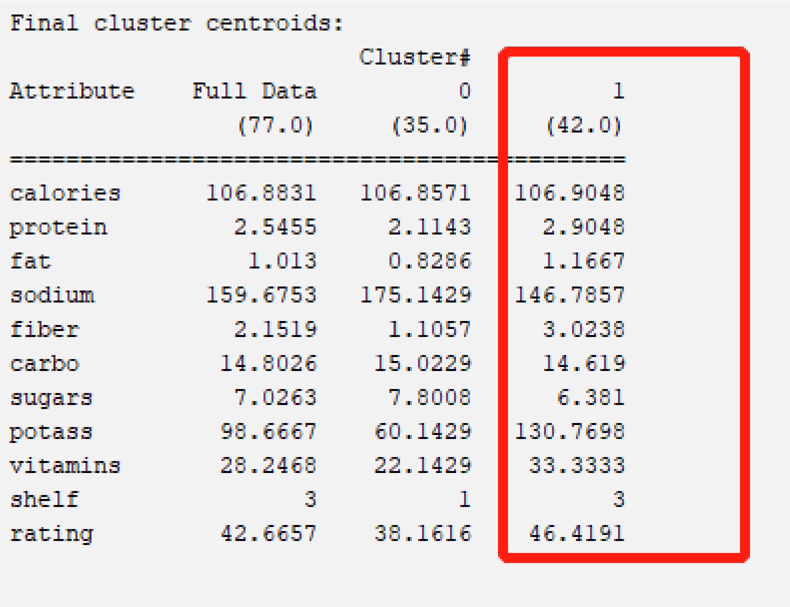
X: shelf (Nom)

Y: rating (Num)

Colour: Cluster (Nom)

Plot: cereal2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.NumericToNominal-R10_clustered

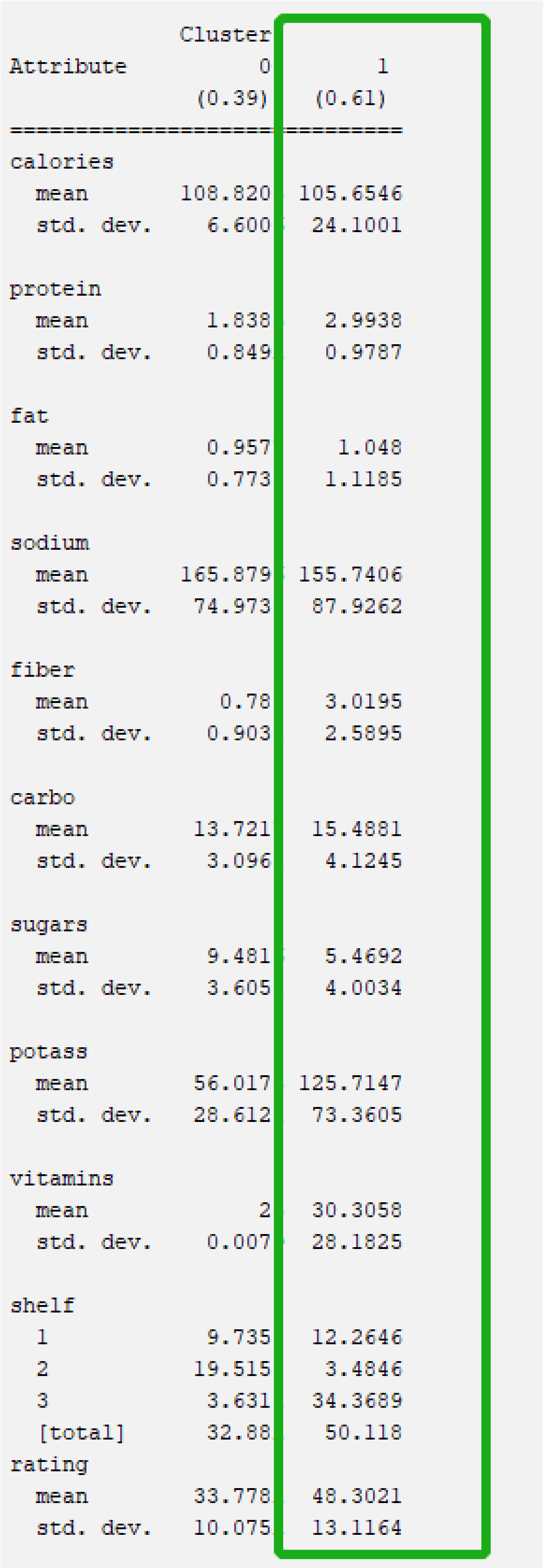
Similar to what we got in PowerBI.



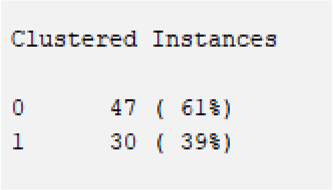
```
Final cluster centroids:
                         Cluster#
Attribute    Full Data          0          1
               (77.0)     (35.0)     (42.0)
============================================
calories     106.8831   106.8571   106.9048
protein        2.5455     2.1143     2.9048
fat             1.013     0.8286     1.1667
sodium       159.6753   175.1429   146.7857
fiber          2.1519     1.1057     3.0238
carbo         14.8026    15.0229     14.619
sugars         7.0263     7.8008      6.381
potass        98.6667    60.1429   130.7698
vitamins      28.2468    22.1429    33.3333
shelf               3          1          3
rating        42.6657    38.1616    46.4191
```

For k=2, we can see centroid cereal in cluster 1 has higher rating.
High-rating cereals are expected to generally have high protein, low sodium, high fiber, low carbohydrates, low sugar, high potassium, high vitamin and tend to be placed on 3rd shelf.
This k separates the cereals well.

Use EM clustering, also apply k=2



```
                Cluster
Attribute             0          1
                 (0.39)     (0.61)
====================================
calories
  mean          108.820   105.6546
  std. dev.       6.600    24.1001

protein
  mean            1.838     2.9938
  std. dev.       0.849     0.9787

fat
  mean            0.957      1.048
  std. dev.       0.773     1.1185

sodium
  mean          165.879   155.7406
  std. dev.      74.973    87.9262

fiber
  mean             0.78     3.0195
  std. dev.       0.903     2.5895

carbo
  mean           13.721    15.4881
  std. dev.       3.096     4.1245

sugars
  mean            9.481     5.4692
  std. dev.       3.605     4.0034

potass
  mean           56.017   125.7147
  std. dev.      28.612    73.3605

vitamins
  mean                2    30.3058
  std. dev.       0.007    28.1825

shelf
  1               9.735    12.2646
  2              19.515     3.4846
  3               3.631    34.3689
  [total]         32.88     50.118
rating
  mean           33.778    48.3021
  std. dev.      10.075    13.1164
```



```
Clustered Instances

0       47 ( 61%)
1       30 ( 39%)
```

Here cereals in cluster 1 with the higher mean rating generally have low calories, high protein, low sodium, high fiber, high carbohydrates, low sugar, high potassium, high vitamin and tend to be placed on 3rd shelf. (I have highlighted difference feature using red color) In general, we get pretty much the same result using two methods.

d) Here we use the clustering result from k-means method, and cluster 1 cereals are healthy cereals. As is explained above, these cereals have high amount in protein/fiber/potassium/vitamin, which are beneficial to students. At the same time, low carbohydrates/sodium/sugar will not cause obesity problem. Also these cereals are popular among customers and are likely to be students' favorites.

Since we calculate Euclidian distance for measuring similarity between instance, it is important to keep these features at similar scale, otherwise features with larger scale will have greater impact than ones with small scale, and this is against the algorithm. For example, the calories values are much bigger than protein values. Without standardization, the cluster result will be influenced majorly by calories.