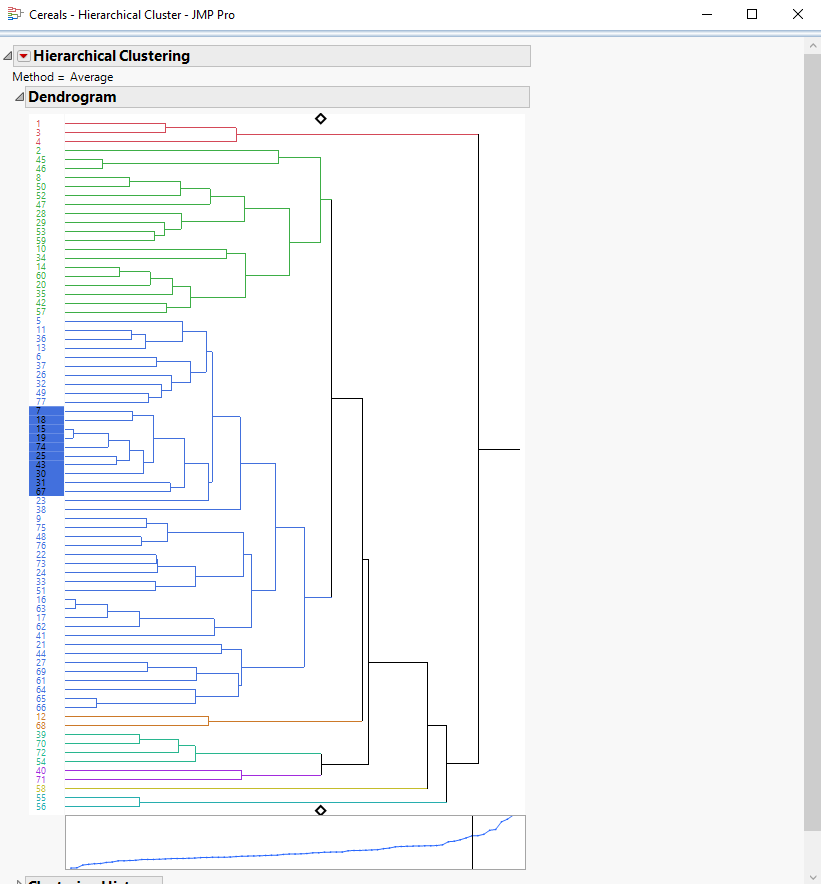
1. **Create clusters using hierarchical clustering and k-means clustering.  Include screen shots of your cluster results.  You must try at least one hierarchical and one k-means technique, but you'll probably want to try a few different settings to determine your best result. Do not include Rating.**

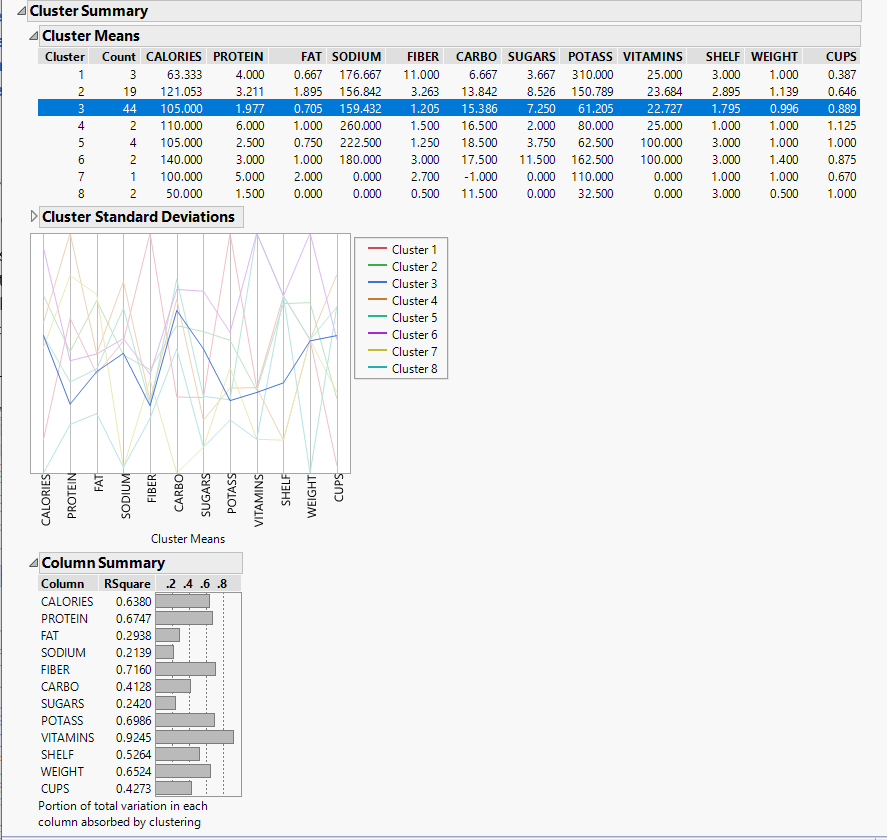
**Hierarchical:**

**Analyze -> Clustering -> Hierarchical Cluster**

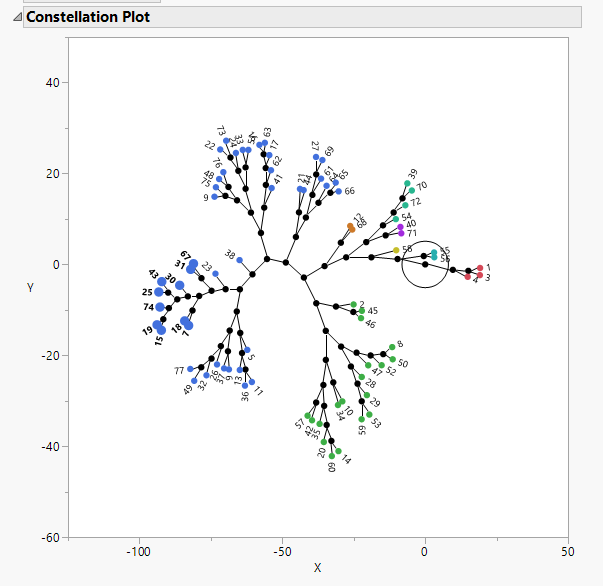
**Select Number of Clusters and input 8.**

Number of clusters depends on how much differentiation you want to show in the data. Each cluster is divided based on its own characteristics. The dendrogram shows how the clustering is conducted. The clustering process can be viewed by reading the dendrogram from left to right. Each step consists of combining the two *closest* clusters into a single cluster.





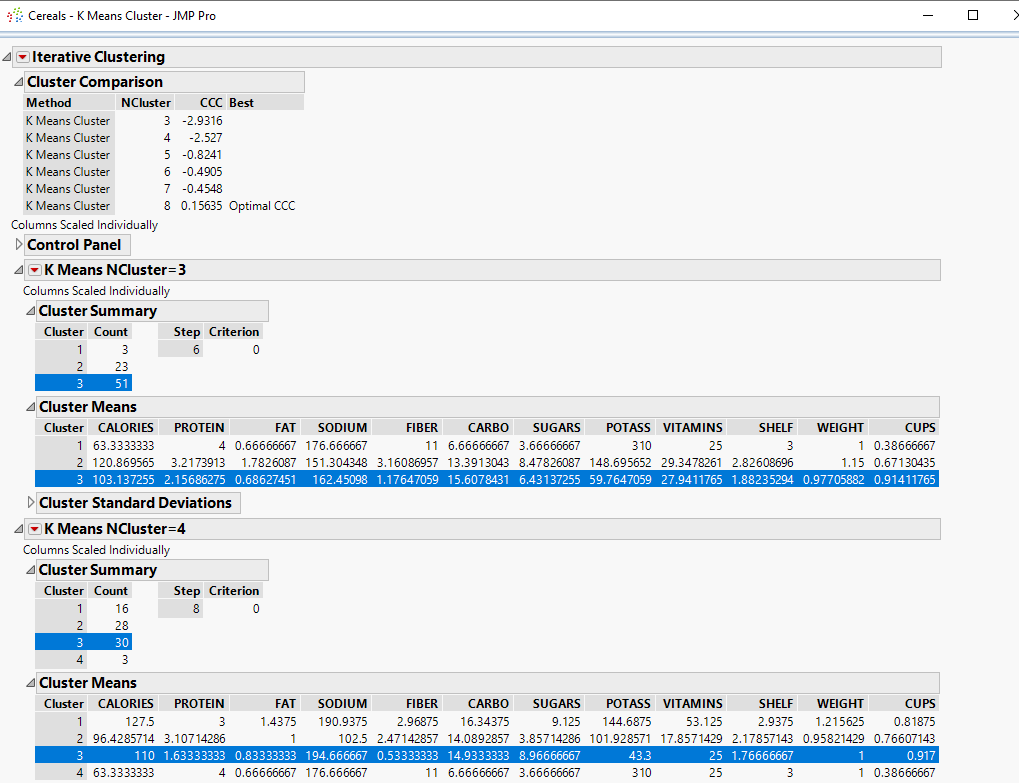
**Constellation Plot is another way of looking at the clusters.**

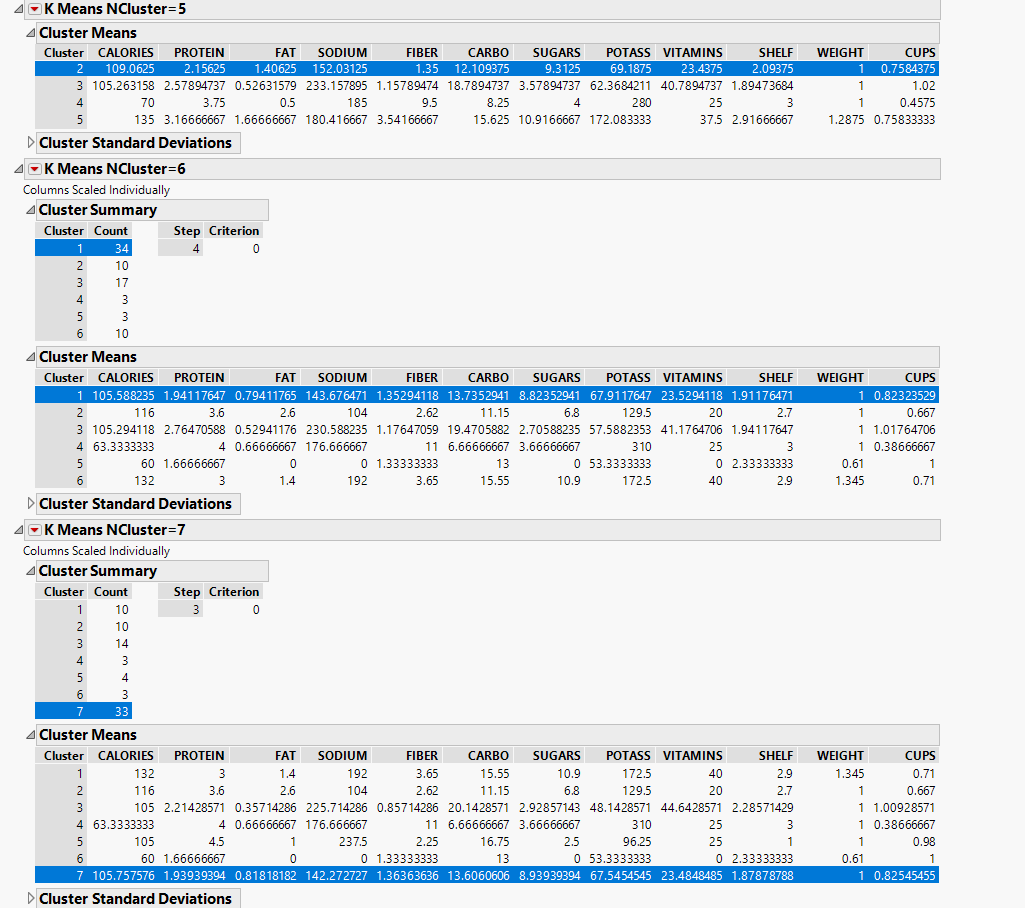


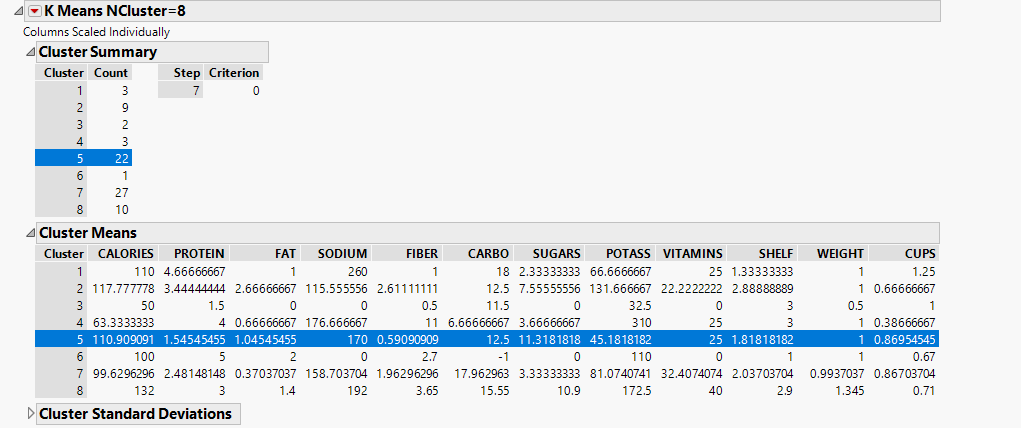
**K-Means:**

**Analyze -> Clustering -> K Means Cluster**

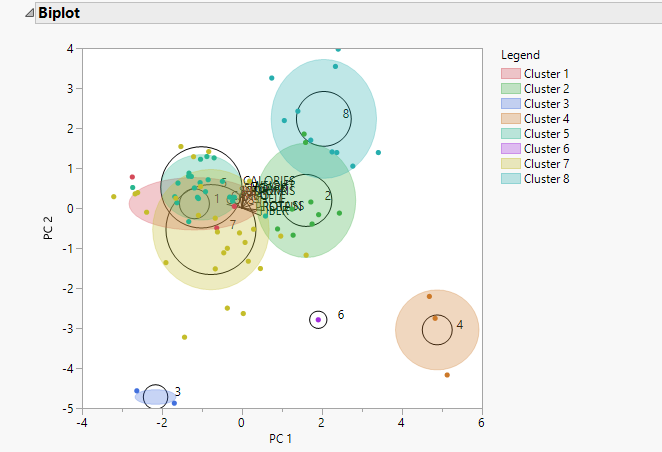
**The Number of clusters(=3) along with the range(=8) can be entered.**

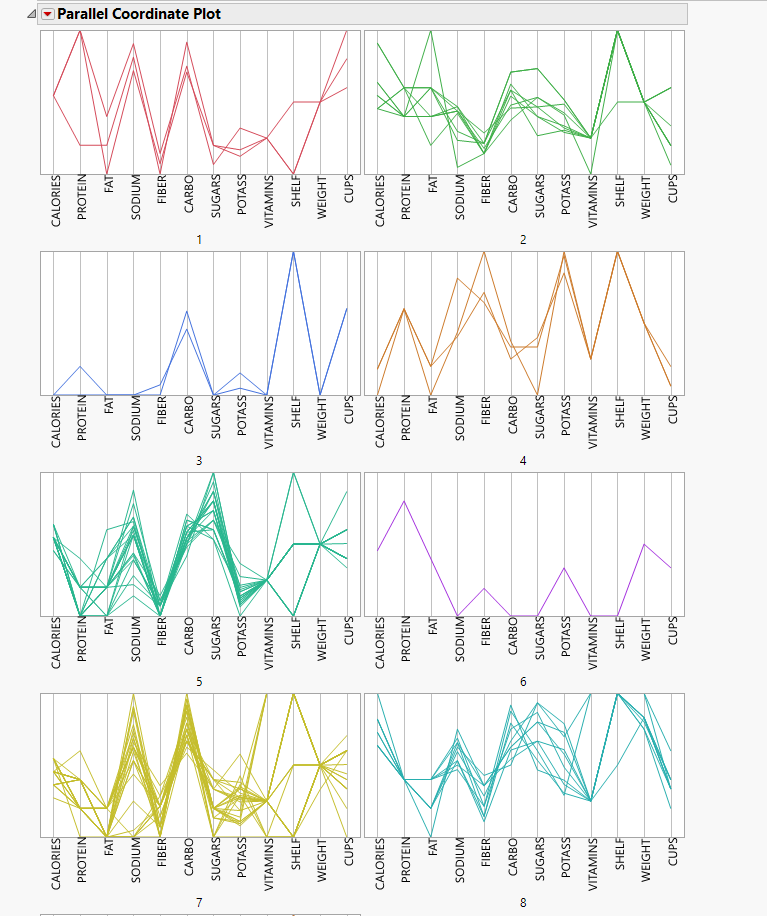




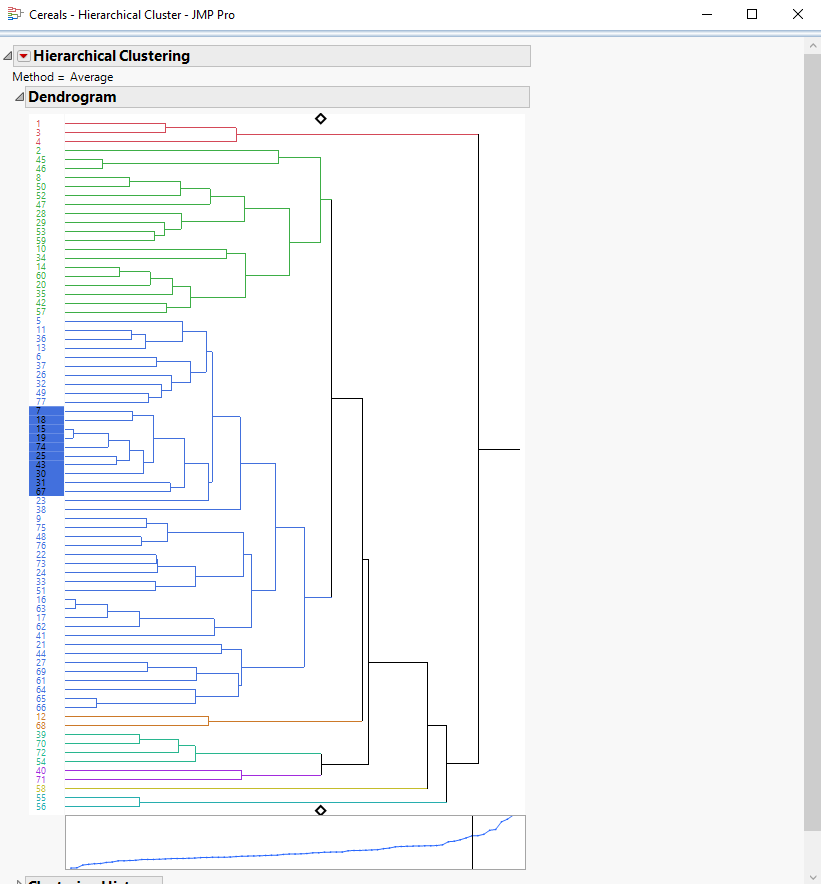


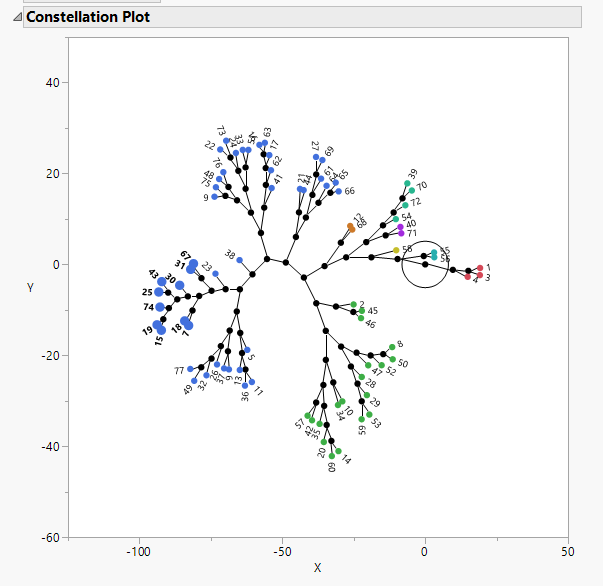
**Biplot :** Shows a plot of the points and clusters in the first two principal components of the data.

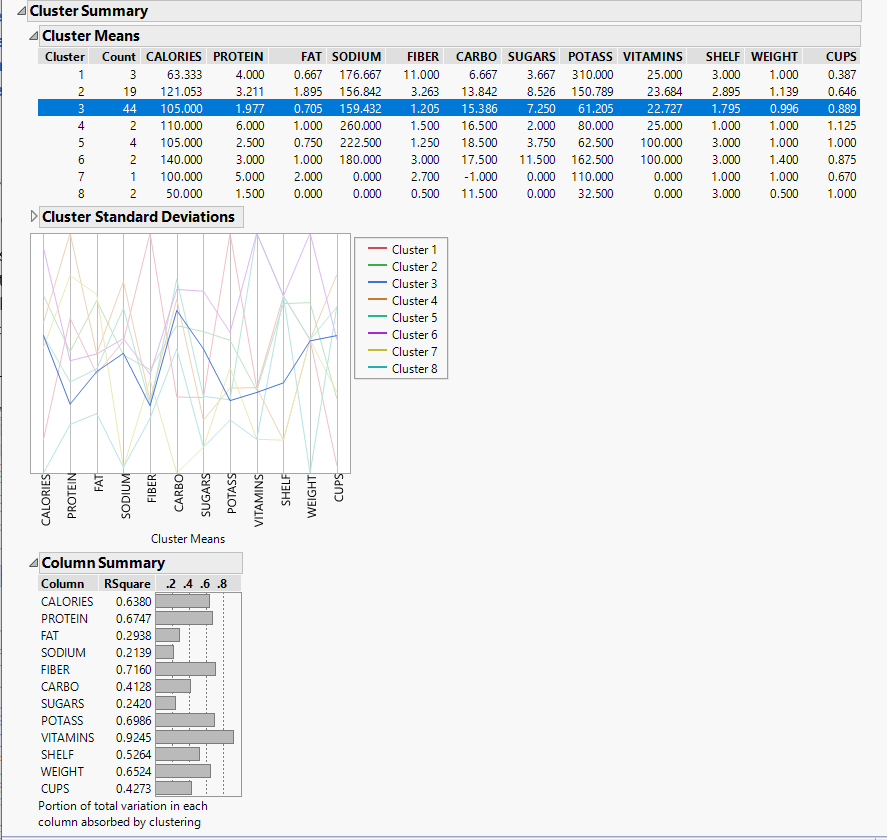




1. **Which cluster technique yielded the best results?  How did you choose it?**







Hierarchical clustering yields better results since the dataset is small. Hierarchical clustering outputs a hierarchy, i.e. a structure that is more informative than the unstructured set of flat clusters returned by k-means. Therefore, it is easier to decide on the number of clusters by looking at the dendrogram.

**Validation Strategy:**

**• Compactness** measures evaluate how close are the objects within the same cluster. A lower within-cluster variation is an indicator of a good compactness (i.e., a good clustering). The different indices for evaluating the compactness of clusters are based on distance measures such as the cluster-wise within average/median distances between observations.

• **Separation** measures determine how well-separated a cluster is from other clusters. The indices used as separation measures include:

1. Distances between cluster centers
2. The pairwise minimum distances between objects in different clusters

• **Connectivity** corresponds to what extent items are placed in the same cluster as their nearest neighbors in the data space. The connectivity has a value between 0 and infinity and should be minimized.

Dendrogram:

The main use of a dendrogram is to work out the best way to allocate objects to clusters.

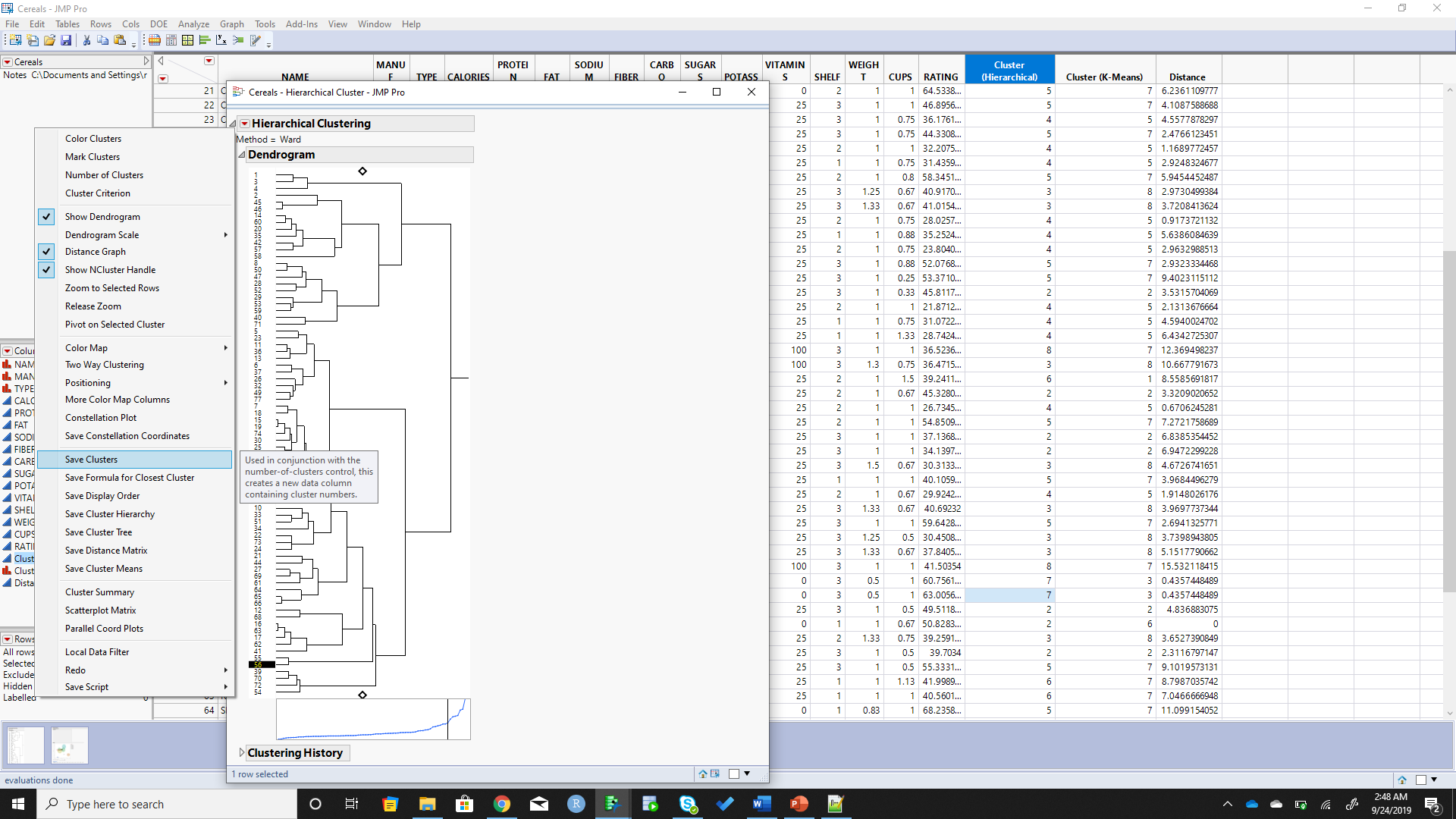
Constellation Plot:

From the plot we can observe that, the average distance within cluster is small;(inter cluster) and the average distance between clusters is large (intra cluster).

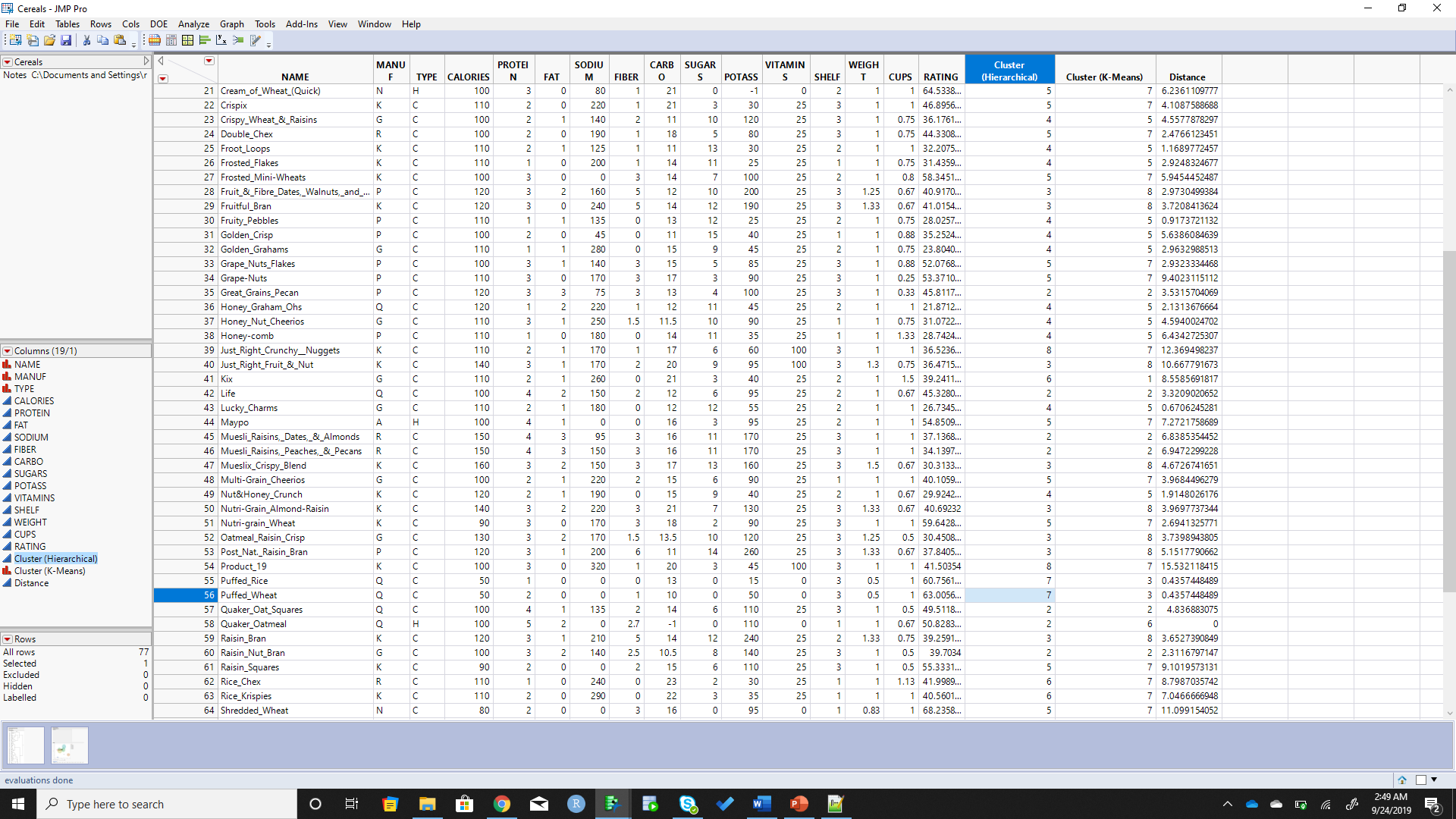
The advantages of choosing Hierarchical clustering includes,

* It is both more flexible and has fewer hidden assumptions about the distribution of the underlying data.
* The results are reproducible in Hierarchical clustering.
* Embedded flexibility concerning the extent of granularity. It is easier to implement. (We don’t require domain knowledge /prior knowledge of the number of clusters)
* The clusters correspond to meaningful taxonomies.
* The applicability to any attribute varieties.

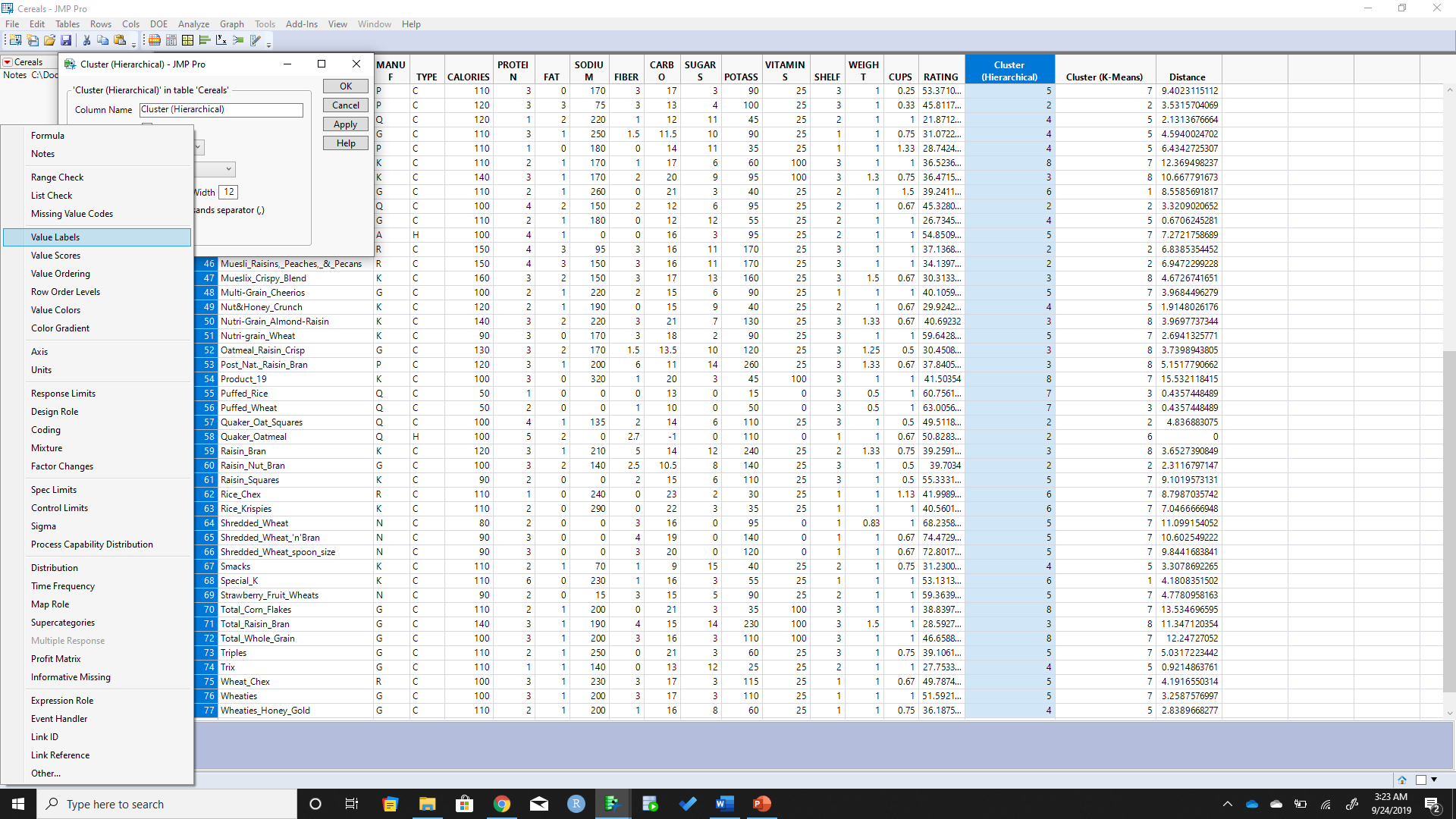
1. **Save your final clusters to the data table. Use graphics and descriptive tools to characterize the nature of the clusters.  Assign names to the clusters.**



Saving the cluster details adds a column to the data table with the cluster details.



Labels can be assigned to the clusters by selecting the column properties of the saved cluster column and editing the column properties -> **Value Labels**



Based on the nature from the clustering graphs, I have assigned labels to the clusters using the following description,

Cluster 1 – High fiber and potassium; Low calories.

Cluster 2 – High in sugar and Low in fiber, complex carbohydrates and protein.

Cluster 3 – Low in calories but little nutritional value.

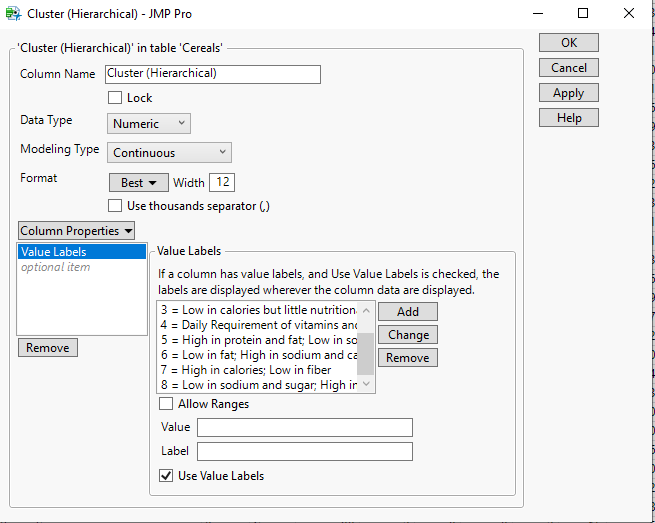
Cluster 4 – Daily Requirement of vitamins and minerals; Low in fat, fiber and sugar.

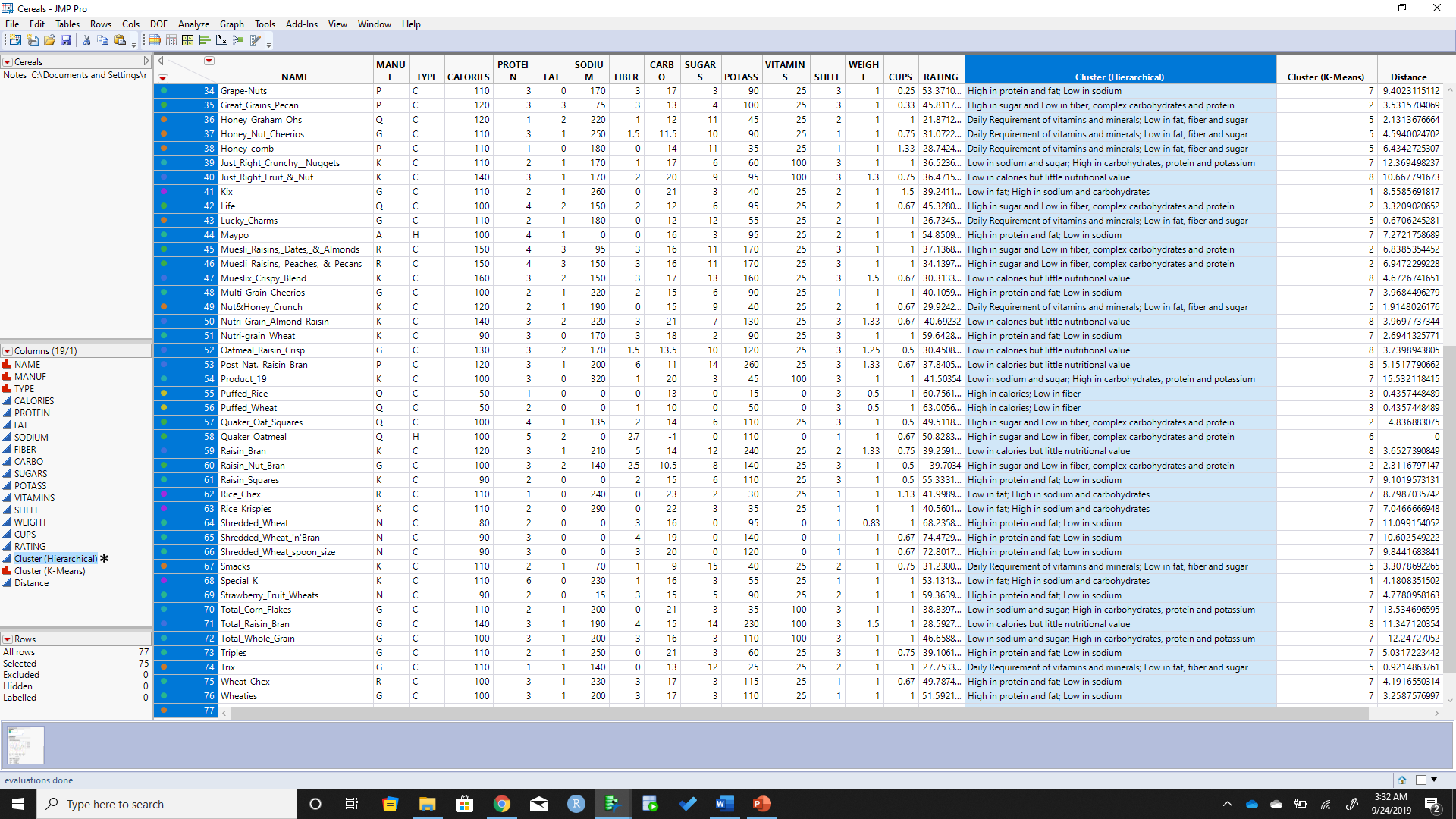
Cluster 5 – High in protein and fat; Low in sodium.

Cluster 6 – Low in fat; High in sodium and carbohydrates.

Cluster 7 – High in calories; Low in fiber.

Cluster 8 – Low in sodium and sugar; High in carbohydrates, protein and potassium.





**4.  Which cluster of cereals is the “healthiest”?  Does one cluster stand out as being the healthiest?**

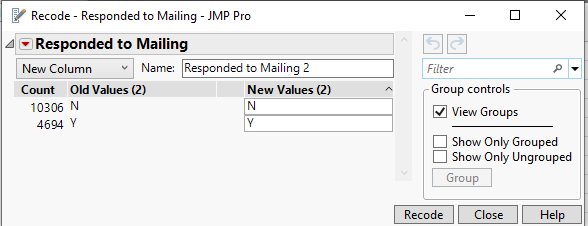
From the analysis, we learn that a high-fiber, high-protein, low-fat, and low-sugar cereal would make a nutritious breakfast. Choosing the right one for our breakfast will definitely make us healthy. Therefore, choosing cereals from **cluster one** is most healthy.

In cluster one, 100% Bran and All-Bran with extra fibers are healthy cereals with much dietary filters.

Avoid cereals in cluster three with little nutritional value.

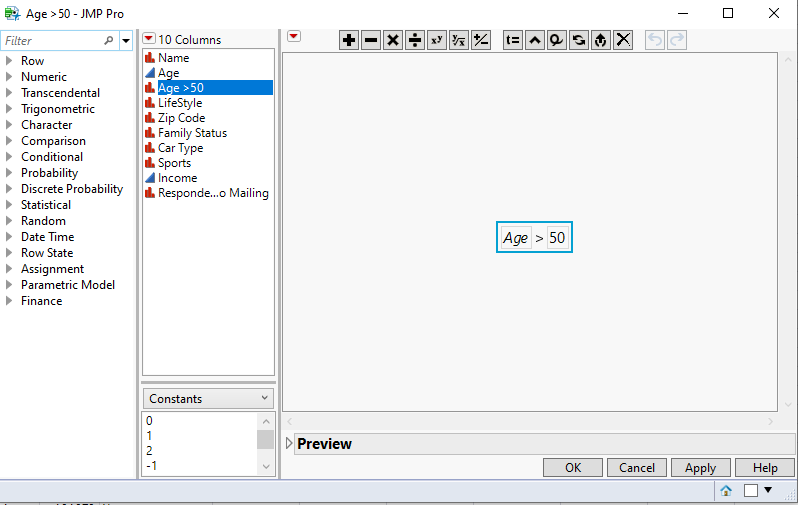
Cereals are highly nutritional in cluster four.

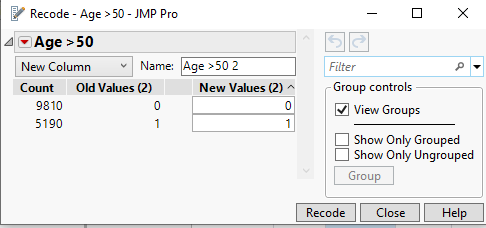
**5.  Refer to slides 12 - 14 in OPIM 5604 - Pattern Discovery-Basket of goods Analysis.ppt.  Write up your recommendation to the scenario on slide 14.**



1. **Total observations = 15000**
2. **Observations with response = 4694**
3. **Base response rate = (4694/15000) = 31.29%**
4. **Observations with age > 50:**

**Created a new row for Age > 50 with the formula Age > 50 and Recoded the value for the column.**



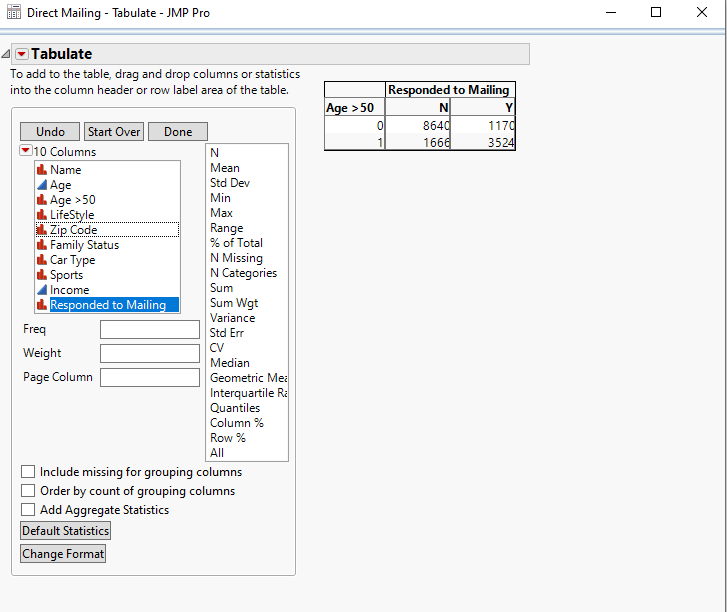


**Observations with age > 50 = 5190**

1. **Observations with response and age > 50:**

**Change Age >50 field to Nominal**

**Analyze -> Tabulate -> Drag Age > 50 & Responded to Mailing field to the matrix**





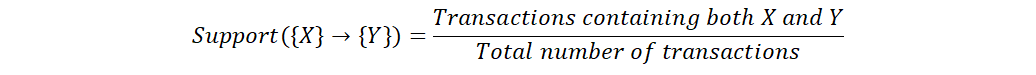
**Observations with response and age > 50 = 3524**

**Association Rules:**

**Support: observations with age > 50 & Response**

**X – Age >50**

**Y – Responded to the email**



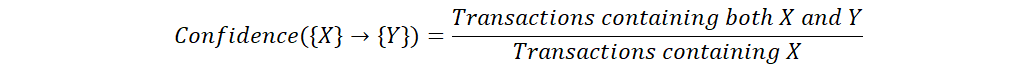
**= 3524/15000**

**= 0.2349**

**Confidence = observations with response and age > 50/observations with age > 50**

**X – Age >50**

**Y – Responded to the email**



**= 3524/5190**

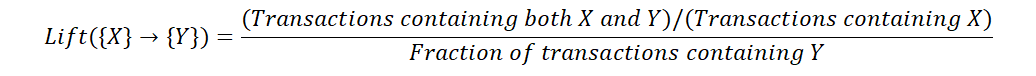
**= 0.6789**

**Expected confidence = observations with response/total observations**

**= 4694/15000**

**= 0.3129**

**Lift = Confidence/Expected Confidence**



**= 0.6789/0.3129**

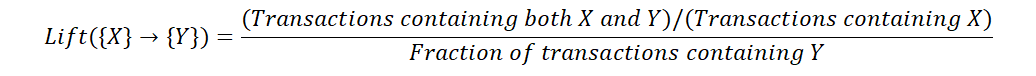
**= 2.1697**

**Consider**

**1. “age > 50”**

**2. “income > $100,000”**

**3. “age > 50” and “income > $100,000”**



1. **Age> 50 and Response was given,**

**Lift = (3524/5190) / 0.3129**

**= 2.1697**

1. **Income > $100,000 and Response was given**

**Lift = (2280/5754) / 0.3129**

**= 1.2663**

1. **Age > 50 and income > $100,000 and Response was given**

**Lift = (1393/1998)/ 0.3129**

**= 2.2281**

**The Lift (Association) is high for people who are above 50 and income higher than $100,000. It is a wise decision to send the offers to people whose age is above 50 and income greater than $100,000.**