**––**

**American International University-Bangladesh**

**DATA WAREHOUSING AND DATA MINING**

**Fall 2017-18**

**Project 1: Unsupervised Learning**

**Name:** Chowdhury, Jahiruddin

**ID:** 15-28397-1

**Section:** A

**Project Definition**

**Purpose**

The purpose of this project is to find the most suitable cereal having the right ingredients from the cereal data set and find any possible co-relation between ingredients of cereals that may lead to a better knowledge for healthy breakfast.

**Problems/Opportunities Addressed**

The problem can be addressed is the lack of proper quantity of all the ingredients in the cereals intended for healthy breakfast.

**Project Scope**

A Dendrogram with the clusters summarized and a knowledge discovery from the Dendrogram.

**Key Stakeholders**

The key stake holders are all individuals who consume cereal on regular basis.

**Literature Survey**

In this report, the use of data mining tools on a pre-compiled dataset cereals using the tool ‘Weka’. There have been a major improvement and increase of concern among individuals for a perfect diet and a healthy life. So, this report is to satisfy that need.Data mining is the process of selecting, exploring and modeling large amounts of data. This process has become an increasingly extensive activity in all areas, specially in medical science research. Data mining has resulted in the discovery of useful hidden patterns from massive databases. Data mining problems are often solved using different approaches from both computer sciences, such as multi-dimensional databases, machine learning, soft computing and data visualization; and statistics, including hypothesis testing, clustering, classification, and regression techniques.

It is found through various medical researches that the breakfast has a lot impact on an individual’s regular life and a healthy breakfast can enhance one’s daily activities. So, as the starting meal of the day everyone should maintain consistency in both food intake timings and the types of food they choose. Dietary consistency helps peoples to prevent diabetics, high blood pressure and heart diseases. Meal planning includes choosing nutritious foods and eating the right amount of food at the right time.

**Methods**

Step 1: I made a .arff file of the given data set and I open it on weka (data mining application).

Step 2: With the filter option, I added an id attribute to the data set for better consideration in the later steps.

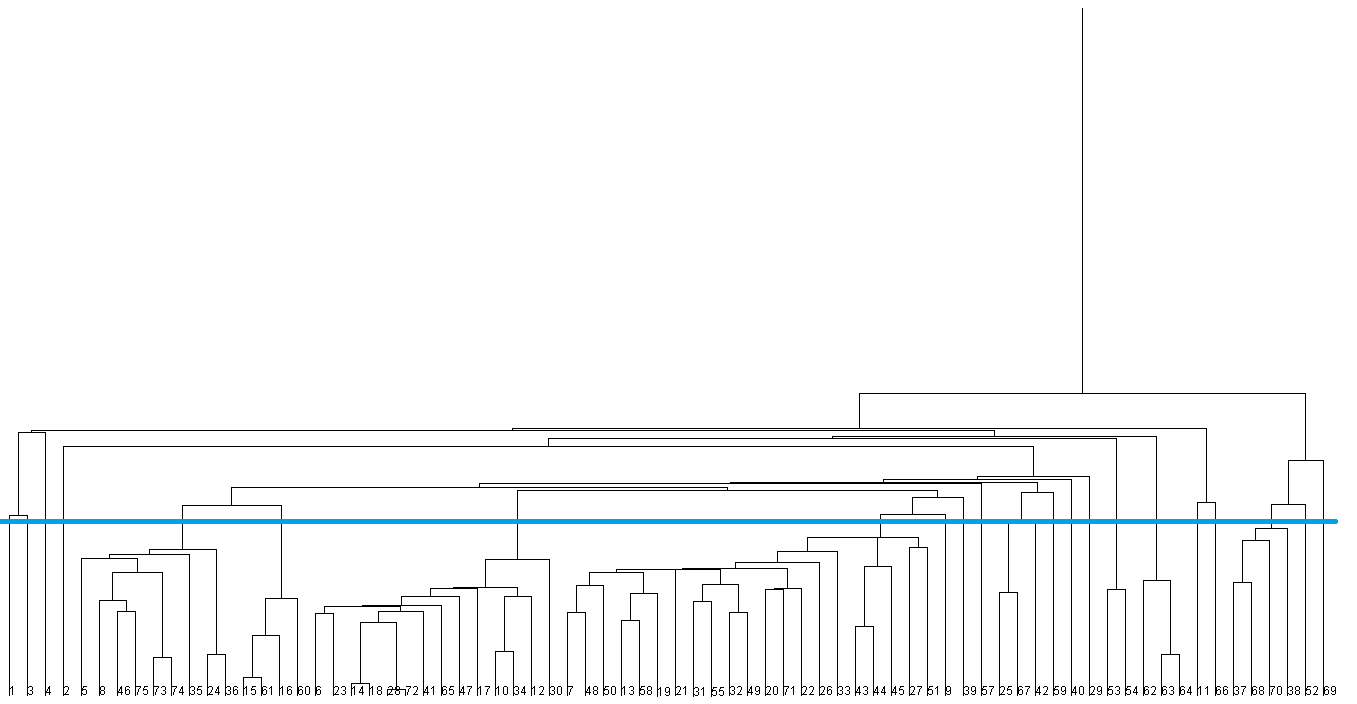
Step 3: From the Cluster tab I have chosen hierarchical clustering and applied it on the data set using training set.

Step 4: Right click on the result list and go to the ‘visualize tree’ option

Step 5: Took a screenshot and with that applied the cutting point on the tree

Step 6: Separate the clusters.

**Hierarchical cluster tree with cut points**

****

Data Set Structure:

@relation cereal

@attribute cerealname string

@attribute calories numeric

@attribute protein numeric

@attribute fat numeric

@attribute sodium numeric

@attribute dietaryfiber numeric

@attribute complexcarbohydrates numeric

@attribute sugars numeric

@attribute displayshelf numeric

@attribute potassium numeric

@data

**Cluster 1:**

**(Low Calorie, Medium Protein, Low Fat, Medium Sodium, Medium Fiber, Low Carbohydrte,Low Sugar,High Potassium, Low Vitamin, Medium Rating )**

100%\_Bran 70 4 1 130 10 5 6 280 25 3 68.402973

**Cluster 2:**

**(Low Calorie, Medium Protein, Low Fat, High Sodium , Medium Fiber, Low Carbohydrte ,Low Sugar ,High Potassium,Low Vitamin, Low rating)**

All-Bran 70 4 1 260 9 7 5 320 25 3 59.425505

**Cluster 3 :**

**(Low Calorie ,Medium Protein, Zero Fat, Medium Sodium, High Fiber, Medium Carbohydrted ,No Sugar ,High Potassium, Low Vitamin,Medium rating)**

All-Bran\_with\_Extra\_Fiber 50 4 0 140 14 8 0 330 25 3 93.704912

**Cluster 4:**

**(Medium Calorie ,Medium Protein, High Fat, Low Sodium, Low Fiber, Medium Carbohydrted ,Low Sugar ,Medium Potassium, Low Vitamin,Low rating)**

100%\_Natural\_Bran 120 3 5 15 2 8 8 135 0 3 33.983679

**Cluster 5:**

**(Low Fiber ,Low Potassium, Low Vitamin)**

Almond\_Delight 110 2 2 200 1 14 8 25 3 34.384843

Apple\_Cinnamon\_Cheerios 110 2 2 180 1.5 10.5 10 70 25 1 29.509541

Basic\_4 130 3 2 210 2 18 8 100 25 3 37.038562

Cocoa\_Puffs 110 1 1 180 0 12 13 55 25 2 22.736446

Corn\_Chex 110 2 0 280 0 22 3 25 25 1 41.445019

Double\_Chex 100 2 0 190 1 18 5 80 25 3 44.330856

Great\_Grains\_Pecan 120 3 3 75 3 13 4 100 25 3 45.811716

Honey\_Graham\_Ohs 120 1 2 220 1 12 11 45 25 2 21.871292

"Muesli\_Raisins,\_Peaches,\_&\_Pecans" 150 4 3 150 3 16 11 170 25 3 34.139765

Raisin\_Nut\_Bran 100 3 2 140 2.5 10.5 8 140 25 3 39.7034

Raisin\_Squares 90 2 0 0 2 15 6 110 25 3 55.333142

Triples 110 2 1 250 0 21 3 60 25 3 39.106174

Trix 110 1 1 140 0 13 12 25 25 2 27.753301

Wheat\_Chex 100 3 1 230 3 17 3 115 25 1 49.787445

**Cluster 6:**

**(Low Fat, Low fiber, Low Sugar,Low Potassium, Low Vitamin)**

Cocoa\_Puffs 110 1 1 180 0 12 13 55 25 2 22.736446

Corn\_Chex 110 2 0 280 0 22 3 25 25 1 41.445019

Raisin\_Nut\_Bran 100 3 2 140 2.5 10.5 8 140 25 3 39.7034

Raisin\_Squares 90 2 0 0 2 15 6 110 25 3 55.333142

**Cluster 7:**

**(Low Fat,High Potassium, Low Vitamin,Low rating)**

Apple\_Cinnamon\_Cheerios 110 2 2 180 1.5 10.5 10 70 25 1 29.509541

Bran\_Flakes 90 3 0 210 5 13 5 190 25 3 53.313813

Cheerios 110 6 2 290 2 17 1 105 25 1 50.764999

Clusters 110 3 2 140 2 13 7 105 25 3 40.400208

Corn\_Flakes 100 2 0 290 1 21 2 35 25 1 45.863324

Corn\_Pops 110 1 0 90 1 13 12 20 25 2 35.782791

Crispy\_Wheat\_&\_Raisins 100 2 1 140 2 11 10 120 25 3 36.176196

"Fruit\_&\_Fibre\_Dates,\_Walnuts,\_and\_Oats" 120 3 2 160 5 12 10 200 25 3 40.917047

Fruity\_Pebbles 110 1 1 135 0 13 12 25 25 2 28.025765

Grape-Nuts 110 3 0 170 3 17 3 90 25 3 53.371007

Kix 110 2 1 260 0 21 3 40 25 2 39.241114

Mueslix\_Crispy\_Blend 160 3 2 150 3 17 13 160 25 3 30.313351

Shredded\_Wheat\_'n'Bran 90 3 0 0 4 19 0 140 0 1 74.472949

Total\_Whole\_Grain 100 3 1 200 3 16 3 110 100 3 46.658844

**Cluster 8:**

**(Medium Calorie ,Medium Protein, Low Fat, Low Fiber, Medium Carbohydrted ,Low Potassium, Low Vitamin,Medium rating)**

Apple\_Jacks 110 2 0 125 1 11 14 30 25 2 33.174094

Cinnamon\_Toast\_Crunch 120 1 3 210 0 13 9 45 25 2 19.823573

Count\_Chocula 110 1 1 180 0 12 13 65 25 2 22.396513

Cracklin'\_Oat\_Bran 110 3 3 140 4 10 7 160 25 3 40.448772

Cream\_of\_Wheat\_(Quick) 100 3 0 80 1 21 0 0 2 64.533816

Crispix 110 2 0 220 1 21 3 30 25 3 46.895644

Frosted\_Flakes 110 1 0 200 1 14 11 25 25 1 31.435973

Frosted\_Mini-Wheats 100 3 0 0 3 14 7 100 25 2 58.345141

Golden\_Crisp 100 2 0 45 0 11 15 40 25 1 35.252444

Golden\_Grahams 110 1 1 280 0 15 9 45 25 2 23.804043

Grape\_Nuts\_Flakes 100 3 1 140 3 15 5 85 25 3 52.076897

Lucky\_Charms 110 2 1 180 0 12 12 55 25 2 26.734515

Maypo 100 4 1 0 0 16 3 95 25 2 54.850917

"Muesli\_Raisins,\_Dates,\_&\_Almonds" 150 4 3 95 3 16 11 170 25 3 37.136863

Multi-Grain\_Cheerios 100 2 1 220 2 15 6 90 25 1 40.105965

Nut&Honey\_Crunch 120 2 1 190 0 15 9 40 25 2 29.924285

Nutri-Grain\_Almond-Raisin 140 3 2 220 3 21 7 130 25 3 40.69232

Nutri-grain\_Wheat 90 3 0 170 3 18 2 90 25 3 59.642837

Puffed\_Rice 50 1 0 0 0 13 0 15 0 3 60.756112

Quaker\_Oatmeal 100 5 2 0 2.7 -1 -1 110 0 1 50.828392

Total\_Raisin\_Bran 140 3 1 190 4 15 14 230 100 3 28.592785

**Cluster 9:**

**(Low Calorie ,Medium Protein, Low Fat, Medium Sodium, Low Fiber, Medium Carbohydrted ,Medium Potassium, Low Vitamin,Medium rating)**

Bran\_Chex 90 2 1 200 4 15 6 125 25 1 49.120253

**Cluster 10:**

**(Medium Calorie ,Medium Protein, Low Fat, Medium Sodium, Low Fiber, Medium Carbohydrted ,Low Sugar ,Low Potassium, Low Vitamin,Medium rating)**

Just\_Right\_Crunchy\_\_Nuggets 110 2 1 170 1 17 6 60 100 3 36.523683

**Cluster 11:**

**(High Calorie, High Protein ,Medium Fat, High Sodium, Low fiber, High Carbohydrte , Low Sugar, Low Potassium)**

Quaker\_Oat\_Squares 100 4 1 135 2 14 6 110 25 3 49.511874

**Cluster 12:**

**(Medium Calorie ,Low Protein, Low Fat, Low Sodium, Low Fiber, Medium Carbohydrted ,High Sugar ,LowPotassium, Low Vitamin,Low rating)**

Froot\_Loops 110 2 1 125 1 11 13 30 25 2 32.207582

Smacks 110 2 1 70 1 9 15 40 25 2 31.230054

**Cluster 13:**

**(Low Calorie ,Medium Protein, Zero Fat, Medium Sodium, High Fiber, Medium Carbohydrted ,No Sugar ,High Potassium, Low Vitamin,Low rating)**

Life 100 4 2 150 2 12 6 95 25 2 45.328074

**Cluster 14:**

**(Medium Calorie ,Medium Protein,Low Fat, High Sodium, Medium Fiber, High Potassium, Low Vitamin,Low rating)**

Raisin\_Bran 120 3 1 210 5 14 12 240 25 2 39.259197

**Cluster 15:**

**(Medium Calorie ,Medium Protein, Low Fat, Medium Sodium, Low Fiber, High Carbohydrted ,No Sugar ,Low Potassium, Low Vitamin,Low rating)**

Just\_Right\_Fruit\_&\_Nut 140 3 1 170 2 20 9 95 100 3 36.471512

**Cluster 16:**

**(Medium Calorie ,Medium Protein, Zero Fat,High Sodium, Medium Fiber, Medium Carbohydrted ,high Sugar ,Medium Potassium, Low Vitamin,Medium rating)**

Fruitful\_Bran 120 3 0 240 5 14 12 190 25 3 41.015492

**Cluster 17:**

**(Medium Calorie ,Medium Protein, Low Fat, Medium Sodium, Low Vitamin,Medium rating)**

Post\_Nat.\_Raisin\_Bran 120 3 1 200 6 11 14 260 25 3 37.840594

Product\_19 100 3 0 320 1 20 3 45 100 3 41.50354

**Cluster 18:**

**(Medium Calorie ,Low Protein, Zero Fat, Low Fiber, High Carbohydrted ,Low Sugar ,Low Potassium, Low Vitamin,High rating)**

Rice\_Chex 110 1 0 240 0 23 2 30 25 1 41.998933

Rice\_Krispies 110 2 0 290 0 22 3 35 25 1 40.560159

Shredded\_Wheat 80 2 0 0 3 16 0 95 0 1 68.235885

**Cluster 19:**

**(Medium Calorie ,Low Protein,Low Fat, High Sodium, Low Fiber, Medium Carbohydrted ,High Sugar ,Low Potassium, Low Vitamin,Low rating)**

Cap'n'Crunch 120 1 2 220 0 12 12 35 25 2 18.042851

**Cluster 20:**

**(Low Calorie ,Medium Protein, Zero Fat, Low Sodium, Low Fiber,High Carbohydrted ,No Sugar ,Medium Potassium, Low Vitamin,Medium rating)**

Shredded\_Wheat\_spoon\_size 90 3 0 0 3 20 0 120 0 1 72.801787

**Cluster 21:**

**(Medium Calorie ,Medium Protein, Low Fat, High Sodium,Low Fiber, Medium Carbohydrted ,Medium Sugar ,Medium Potassium, Low Vitamin,Low rating)**

Honey\_Nut\_Cheerios 110 3 1 250 1.5 11.5 10 90 25 1 31.072217

Honey-comb 110 1 0 180 0 14 11 35 25 1 28.742414

Special\_K 110 6 0 230 1 16 3 55 25 1 53.131324

Total\_Corn\_Flakes 110 2 1 200 0 21 3 35 100 3 38.839746

**Cluster 22:**

**(Medium Calorie ,Medium Protein, Low Fat, Medium Sodium, Low Fiber, Medium Carbohydrted ,High Sugar ,Medium Potassium, Low Vitamin,Low rating)**

Oatmeal\_Raisin\_Crisp 130 3 2 170 1.5 13.5 10 120 25 3 30.450843

**Cluster 23:**

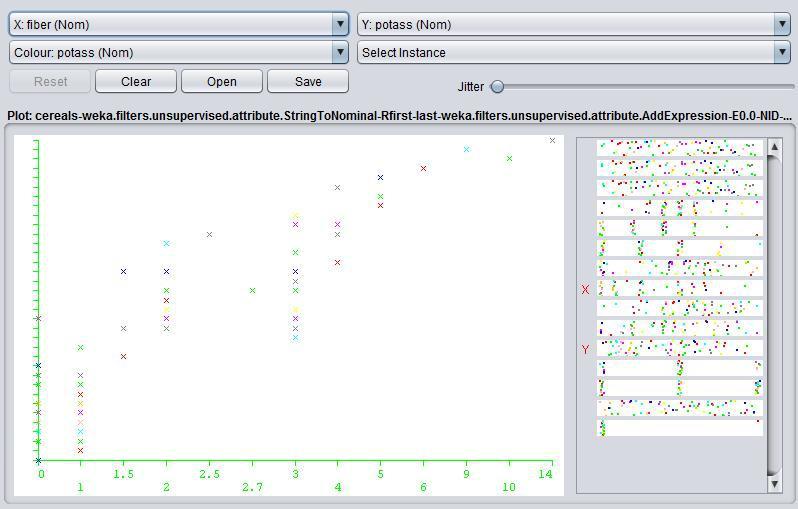
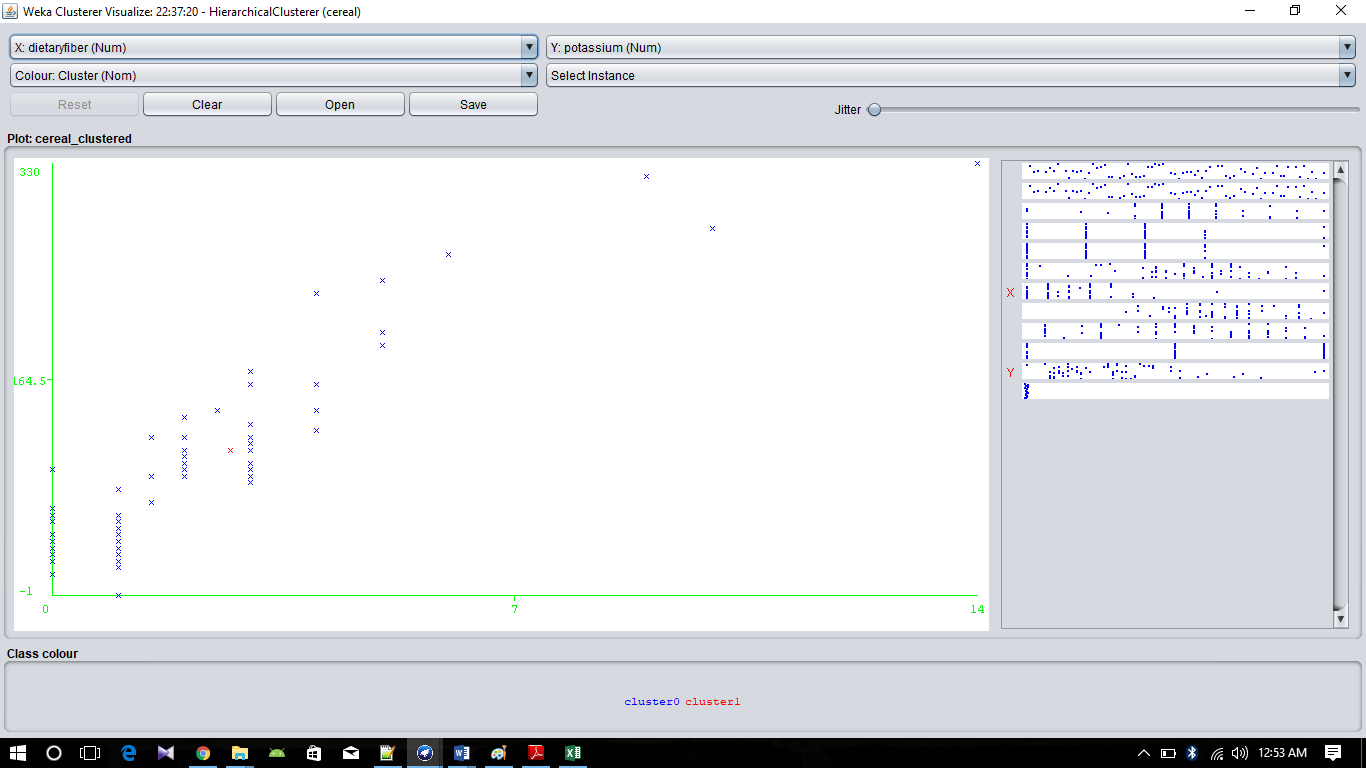
**(Low Calorie ,Medium Protein, Zero Fat, Low Sodium, Low Fiber, Medium Carbohydrted ,Low Sugar ,Low Potassium, Low Vitamin,Medium rating)**

Strawberry\_Fruit\_Wheats 90 2 0 15 3 15 5 90 25 2 59.363993

**Question answer:**

Question: Is a strong correlation between dietary fiber and potassium?

Answer: Yes, there is a strong correlation between dietary fiber and potassium. That is because the correlation of the fiber is proportional to potassium. We can also see from the plot area in the graph which is given below that the position of dietary fiber and potassium increased linearly or proportionally.

****

Question: Are groups of cereals from which we can choose according to our preferences?

Answer: Yes, from the 23 clusters anyone can get any combination of he/she likes such as-

1) People who want low calorie and no fat and no sugar can choose cluster 13.

2) People can take high potassium and high fiber one can have better cereals on cluster 3 & 13.

3) People who want high carbohydrates one can choose cereal from cluster 15,18,20.

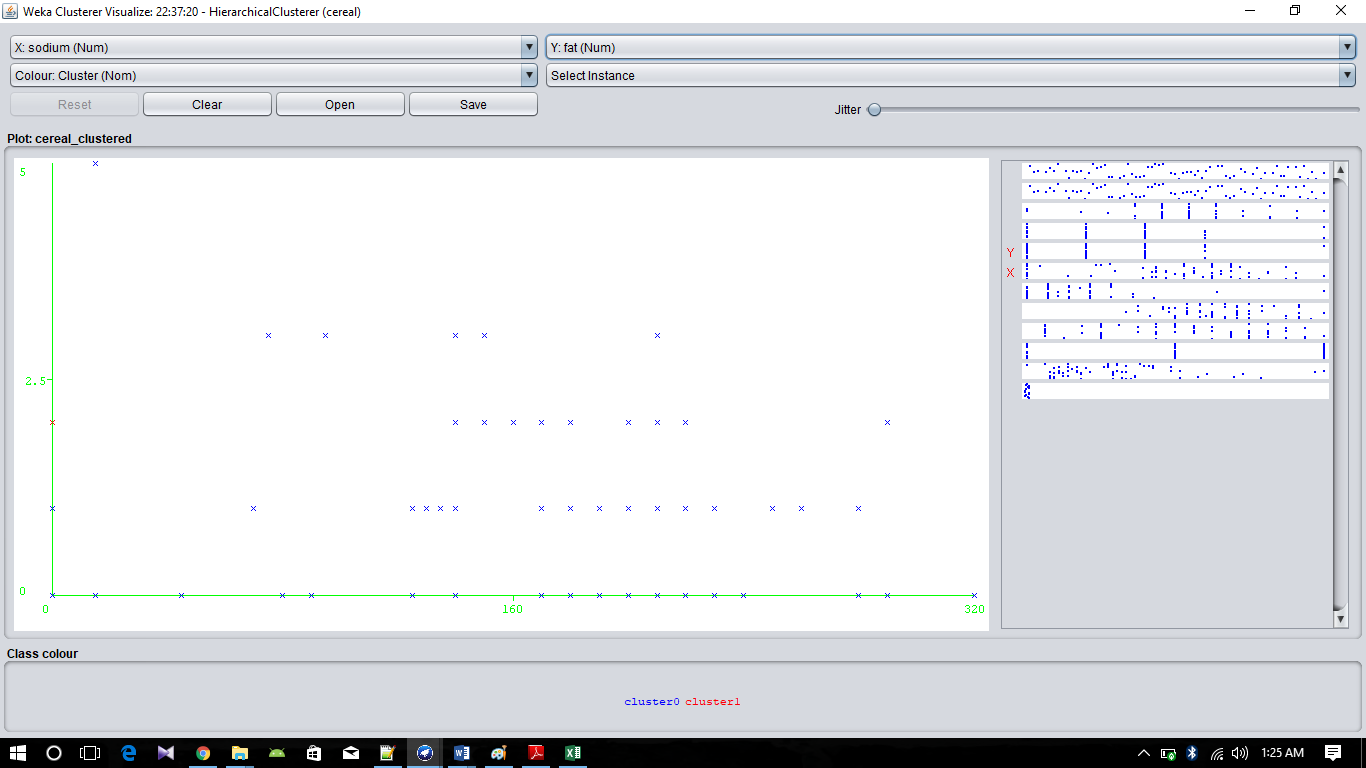
4) People who want high calories and protein with low sugar can pick a cereal from cluster 11.

Etc.

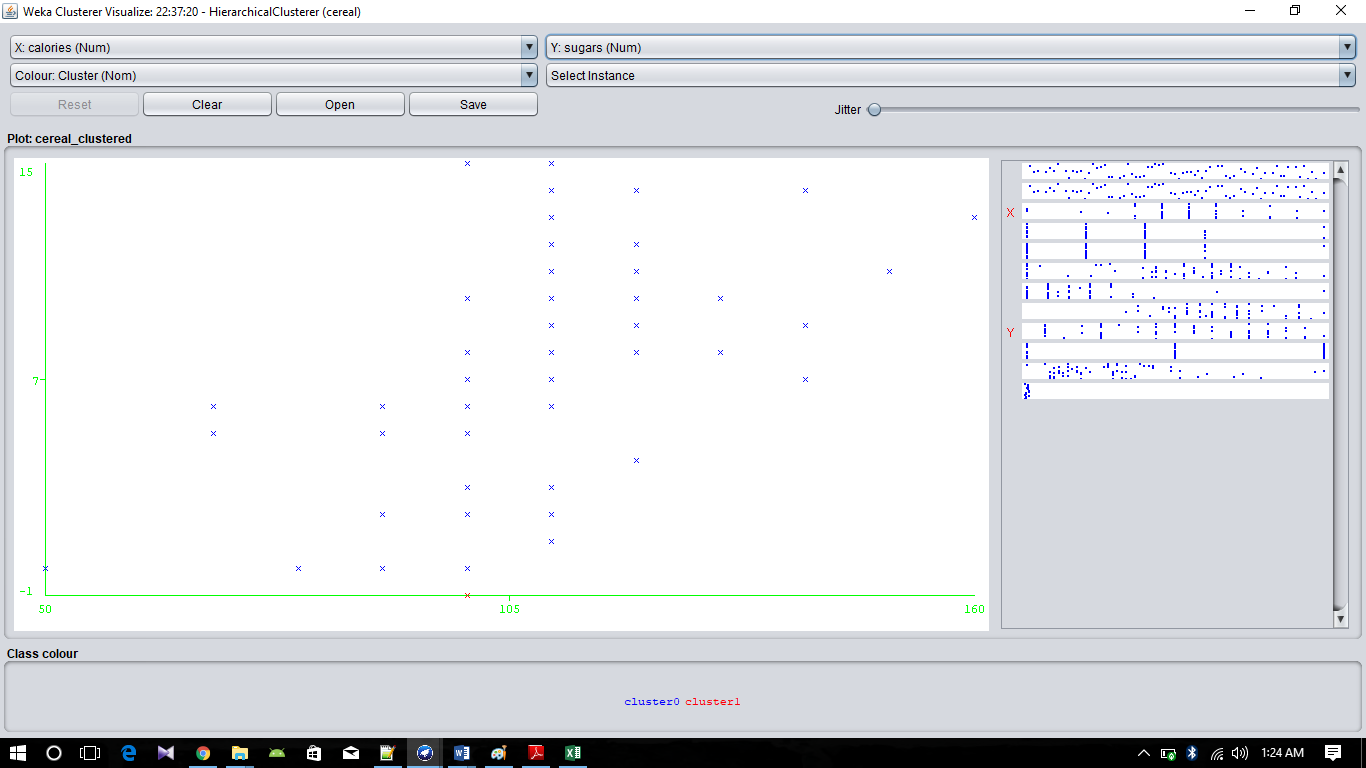
Question: See other correlation between the data given in the files.

Answer:

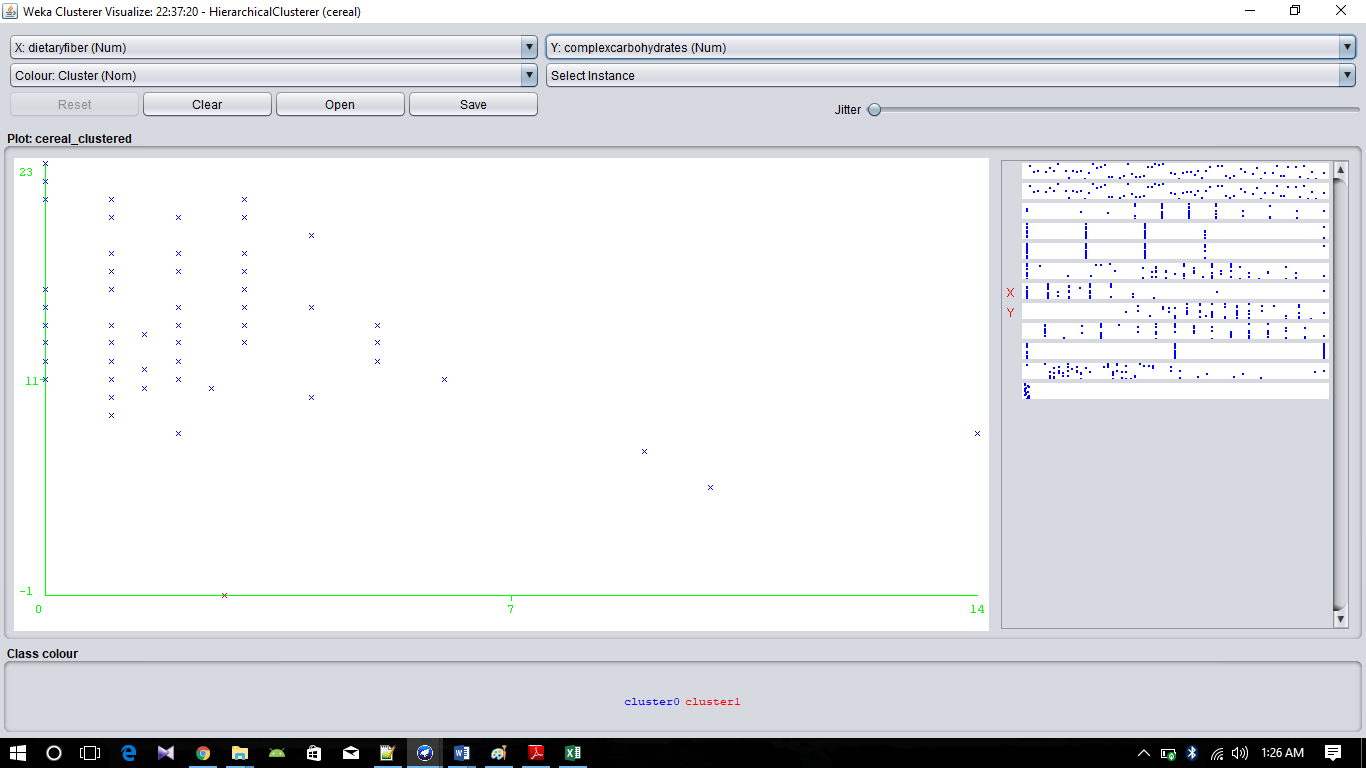
🡪Sodium and Fat corelation



🡪Calorie and Sugar co-relation



🡪Fiber and Carbohydrte co-relation



Question 4:

Answer:

In our given dataset, we also find that where calorie is more there is less protein. And where is more calorie.

Let’s we take 100\_Natural\_Bran where the calorie level is 120g and protein is 3g. There is invert relation between them.

I also observe where the calorie is less, the potassium level is also high. For example, All-Bran cereal there is 70g calorie and 320g of potassium. It has also an invert relation.

There is positive relation between calorie and carob. If one increases then other increases too. As example, Almond Delight has 110g of calorie and 14g of carbo.