AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Choose an item.

Assignment Cover Sheet

Assignment Title:	Project – Unsupervised Learning		
Assignment No:	02	Date of Submission:	15 May 2020
Course Title:	Data Warehousing and Data Mining		
Course Code:	CSC4139	Section:	A
Semester:	Spring 2019-20	Course Teacher:	Rahman Mohammod Hafizur

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.:

No	Name	ID	Program	Signature
1	Ahmed, Sinthia	17-33820-1	BSc [CSSE]	
2			Choose an item.	

Faculty use only

FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Project – Unsupervised Learning

Given, Dataset: Breakfast cereal dataset

Number of Instances: 77

Number of Attributes: 12

Number of Missing values: 2

Attribute Information:

1. cereal_Name: Name of cereal

2. calories: calories per serving

3. protein: grams of protein

4. fat: grams of fat

5. sodium: milligrams of sodium

6. fiber: grams of dietary fiber

7. carbo: grams of complex carbohydrates

8. sugars: grams of sugars

9. potass: milligrams of potassium

10. vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended

11. shelf: display shelf (1, 2, or 3, counting from the floor)

12. rating: a rating of the cereals (calculated by Consumer Reports)

Preparation of Dataset:

The dataset was downloaded from http://www.cs.umd.edu/hcil/hce/examples/cereal/cereal-updated.txt. For better dendrogram visualization I have removed the cereal name column from the dataset and used index column to uniquely identify each column and saved it as .csv file. Then I converted .csv to .arff file. I have dealt with the missing values in potass column by using the average of potass values. The attributes that I used in .arff-

```
@relation cereal

@attribute Index string
@attribute calories numeric
@attribute protein numeric
@attribute fat numeric
@attribute sodium numeric
@attribute fiber numeric
```

```
@attribute carbo numeric
@attribute sugars numeric
@attribute potass numeric
@attribute vitamins numeric
@attribute shelf numeric
@attribute rating numeric
```

Index is set to string so that Weka chooses it as instance name in dendrogram.

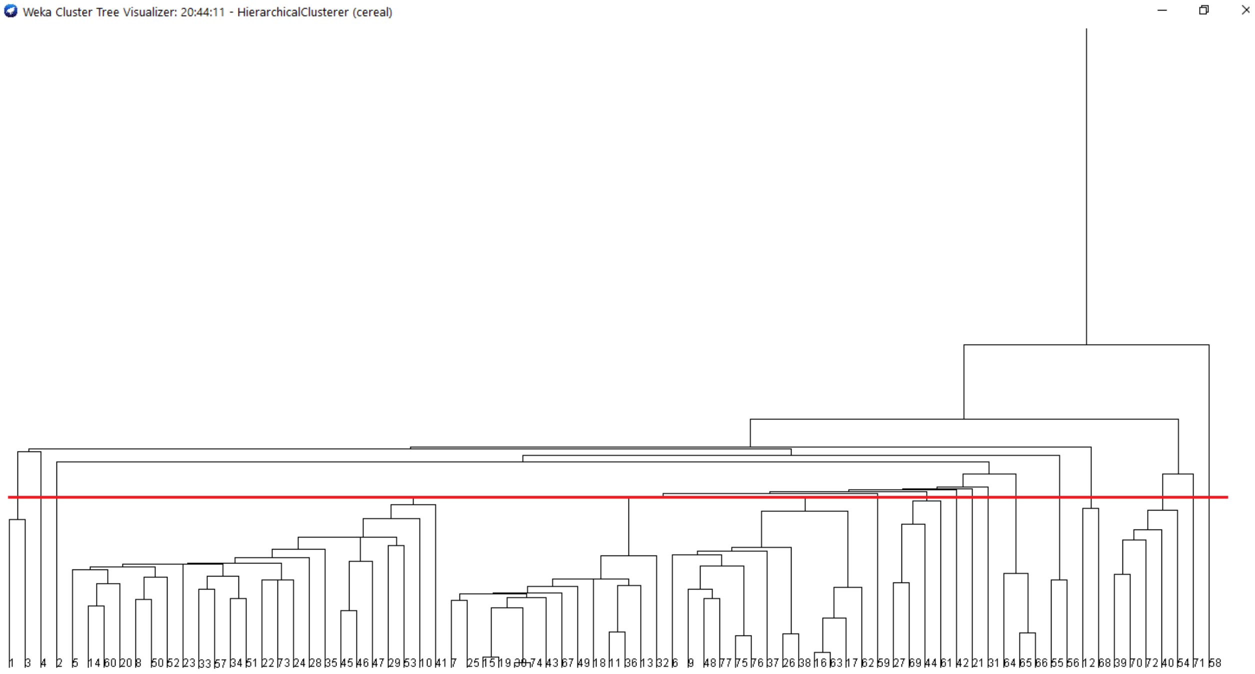
Objective: Objective is to analyze the dataset and answer following questions-

1. Is a strong correlation between dietary fiber and potassium?

2. Are groups of cereals from which we can choose according to our preferences?

3. See other correlation between the data given in the files.

Approach: In order to answer the above questions I used hierarchical clustering in weka to cluster the instances. The following of dendrogram was genrated.



From the cutting point I have choose there are 16 clusters-

Cluster No.	Serial no. of instances	Cereal
1	1,3	All_Bran_with_Extra_Fiber, Puffed_Wheat
2	4	100per_Bran
3	2	Puffed_Rice
4	5,14,60,20,8,50,52,23,33,57, 34,51,22,73,24,28,35,45,46,47,29,53,10,41	All_Bran, Corn_Flakes, 100per_Natural_Bran,

		Grape_Nuts_Flakes, Bran_Flakes, Kix, Rice_Chex, Multi_Grain_Cheerios, Apple_Jacks, Triples, Cheerios, Lucky_Charms, Maypo, Nutri_Grain_Almond_R,aisin, Product_19, Total_Whole_Grain, Clusters, Golden_Grahams, Grape_Nuts, Honey_Nut_Cheerios, Wheat_Chex, Rice_Krispies, Raisin_Squares,Crispix,
5	7,25,15,19,30,74,43,67, 49,18,11,36,13,32	Bran_Chex Quaker_Oat_Squares Cream_of_Wheat_Quick Golden_Crisp Wheaties Total_Raisin_Bran Frosted_Flakes NutnHoney_Crunch Just_Right_Crunchy__Nuggets Frosted_Mini_Wheats Shredded_Wheat_nBran Cocoa_Puffs Strawberry_Fruit_Wheats Apple_Cinnamon_Cheerios
6	6,9,48,77,75,76,37,26, 38,16,63,17,62	Shredded_Wheat Nutri_grain_Wheat Honey_comb Mueslix_Crispy_Blend Muesli_Raisins_Dates_n_Almonds Muesli_Raisins_Peaches_n_Pecans Corn_Chex Quaker_Oatmeal Corn_Pops Crispy_Wheat_n_Raisins Fruit_n_Fibre_Dates_Walnuts_and_Oats Double_Chex Cinnamon_Toast_Crunch
7	59	

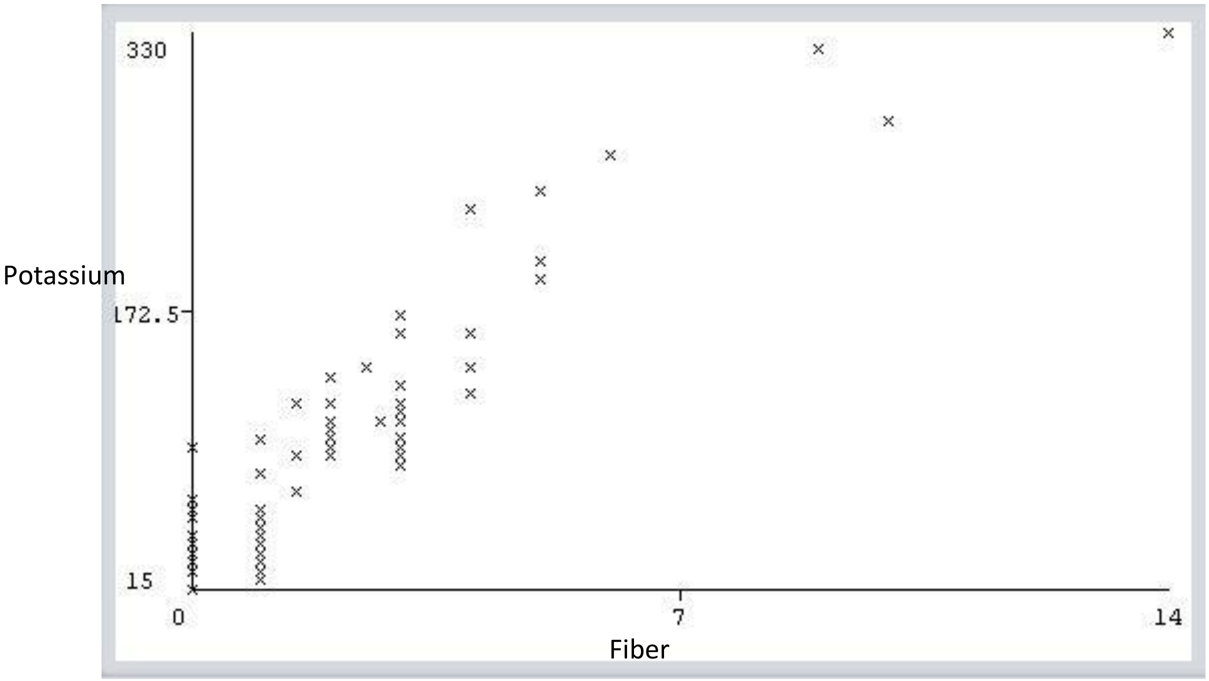
8	27,69,44,61	Raisin_Nut_Bran Raisin_Bran Fruity_Pebbles CapnCrunch
9	42	Froot_Loops
10	21	Life
11	31	Almond_Delight
12	64,65,66	Fruitful_Bran Great_Grains_Pecan Honey_Graham_Ohs
13	55,56	Special_K Total_Corn_Flakes
14	12,68	Shredded_Wheat_spoon_size Post_Nat._Raisin_Bran
15	39,70,72,40,54	Count_Chocula Basic_4 Just_Right_Fruit_n_Nut Cracklin_Oat_Bran Smacks
16	71	Oatmeal_Raisin_Crisp
17	58	Trix

Cluster Analysis:

Cluster	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	rating
1	Low	Mid	Low	High	Mid	Low	Low	High	Mid	High
2	Low	Mid	Low	Mid	High	Low	Low	High	Mid	High
3	Mid	Mid	High	Low	Low	Low	Mid	Mid	Low	Mid
4	Mid	Mid	Low	Mid	Low	Mid	Low	Mid	Mid	Mid
5	Mid	Low	Low	High	Low	Mid	High	Low	Mid	Low
6	Mid	Low	Low	High	Low	High	Low	Low	Mid	Mid
7	Mid	Mid	Low	High	Low	Mid	High	High	Mid	Mid
8	Low	Low	Low	Low	Low	Mid	Mid	Low	Mid	Mid
9	Mid	Mid	Mid	Mid	Low	Mid	Mid	Low	Mid	Mid
10	Mid	Mid	Low	Low	Low	High	Low	Low	Low	Mid
11	Mid	Low	Low	Low	Low	Mid	High	Low	Mid	Mid
12	Low	Mid	Low	Low	Low	High	Low	Mid	Low	High
13	Low	Low	Low	Low	Low	Mid	Low	Low	Low	Mid
14	Mid	High	Mid	High	Low	High	Low	Low	Mid	Mid
15	Mid	Mid	Low	High	Low	High	Low	Low	High	Mid
16	Mid	Mid	Low	High	Low	High	Low	Mid	High	Mid
17	Mid	Low	Low	Mid	Low	Mid	High	Low	Mid	Low

1. Is a strong correlation between dietary fiber and potassium?

Weka data visualization generated following graph-



From the graph we can see that Fiber and Potassium are proportionally correlated. Cereals having high fiber also has high potassium. From data we calculated correlation coefficient r=0.903. So we can say that Fiber and Potassium are strong and positively correlated.

2. Are groups of cereals from which we can choose according to our preferences?

There are different types preferences people have when choosing breakfast cereals.

- Diabetic patient – high fiber and low sugar
- Pregnant lady – high vitamins, high protein and high fiber
- Kid – high protein, high carbohydrate, high vitamins and mid sugar

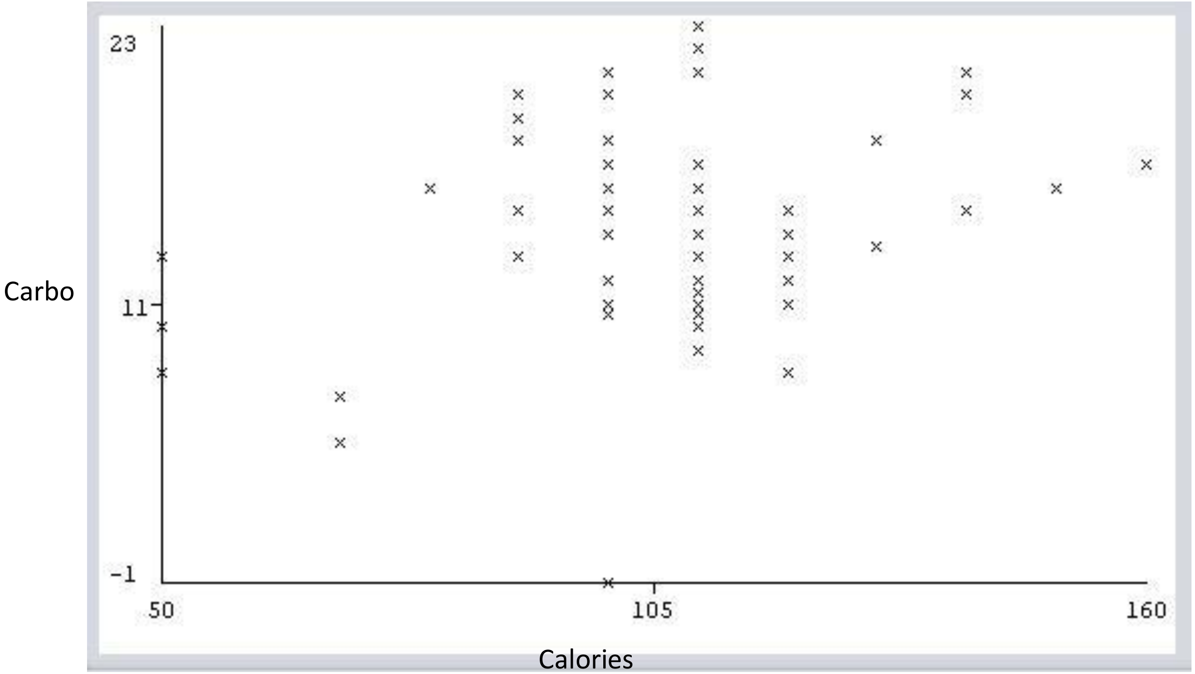
From the cluster analysis we can group the cereals which are according to above preference-

Diabetic patient – Food that are high in fiber and low in sugar are good for diabetic patients. So if they choose cereals they should choose from cluster 2 which fulfills both the criteria.

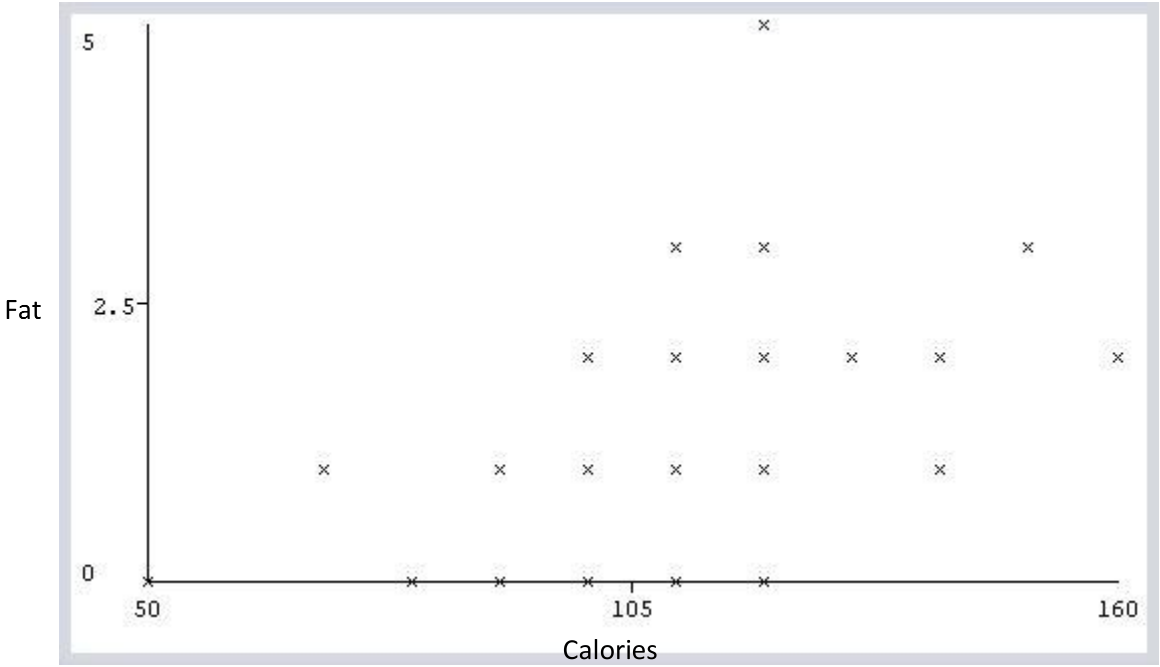
Pregnant lady – Though during pregnancy ladies need all kinds of nutrition but vitamins, protein and fiber are most important. For them cereals from cluster 2, 15 and 16 are suitable.

Kid – while a kid needs protein, carbohydrate and vitamins they also prefer cereal which have high or medium sugar. Cluster 7 and 9 are best cereal choices for kids.

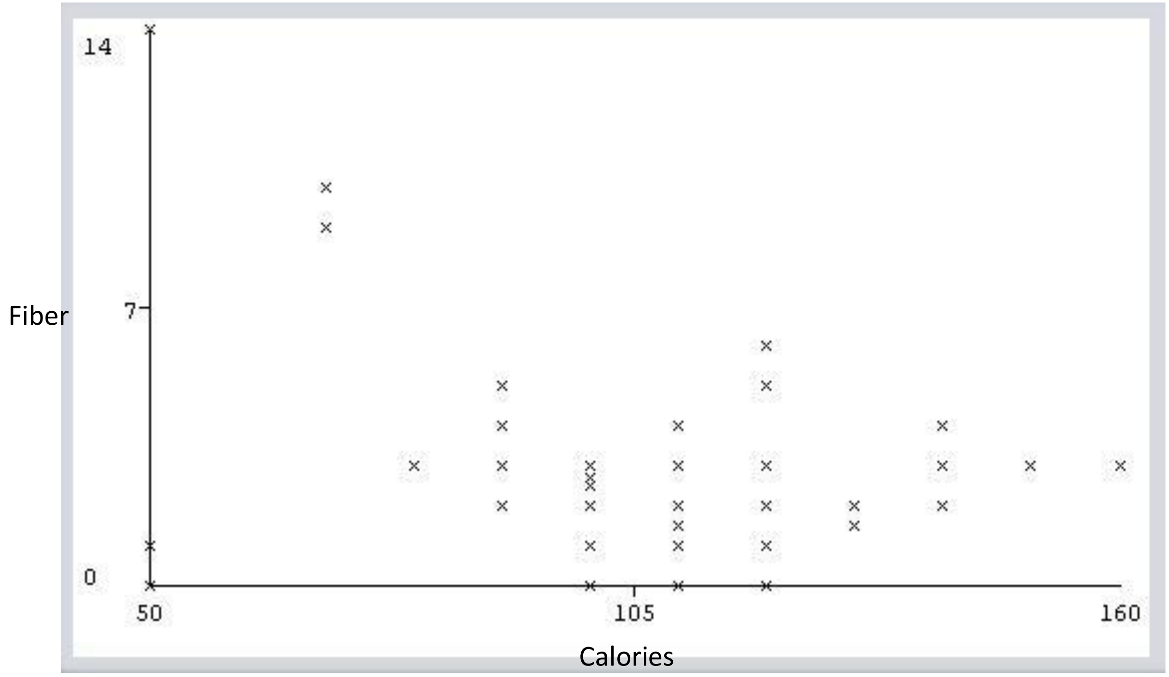
3. See other correlation between the data given in the files.



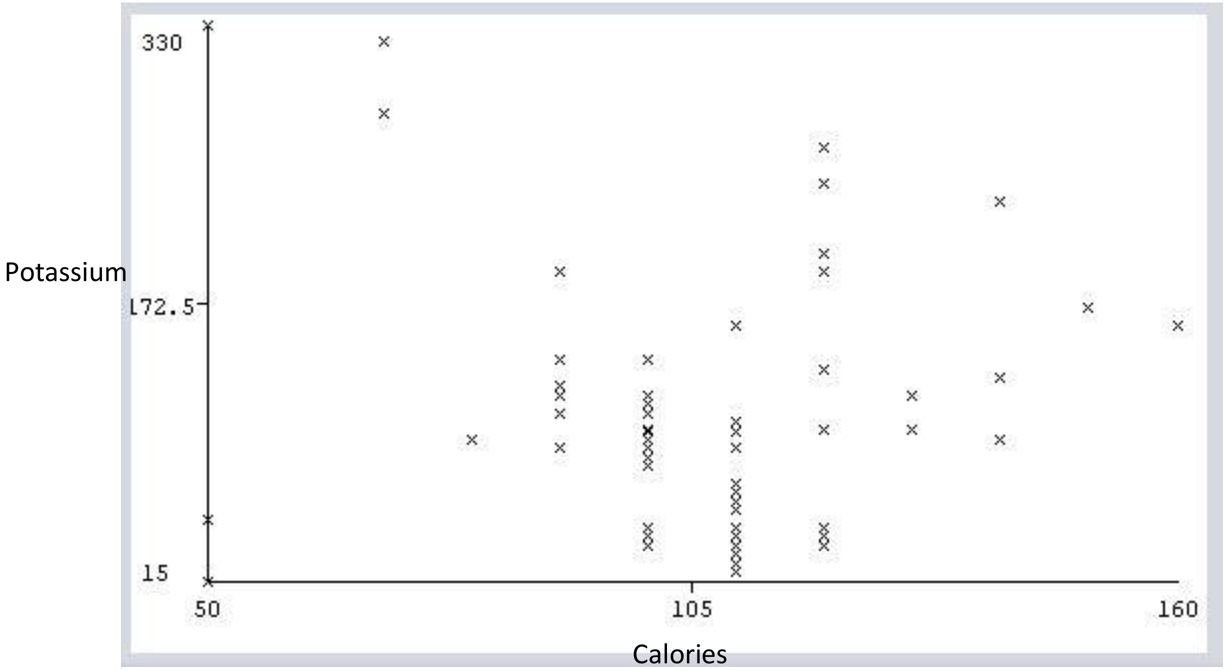
Calories_Carbo correlation – From data we calculated correlation coefficient r=0.251. So we can say that Calories and Carbohydrate are positively correlated.



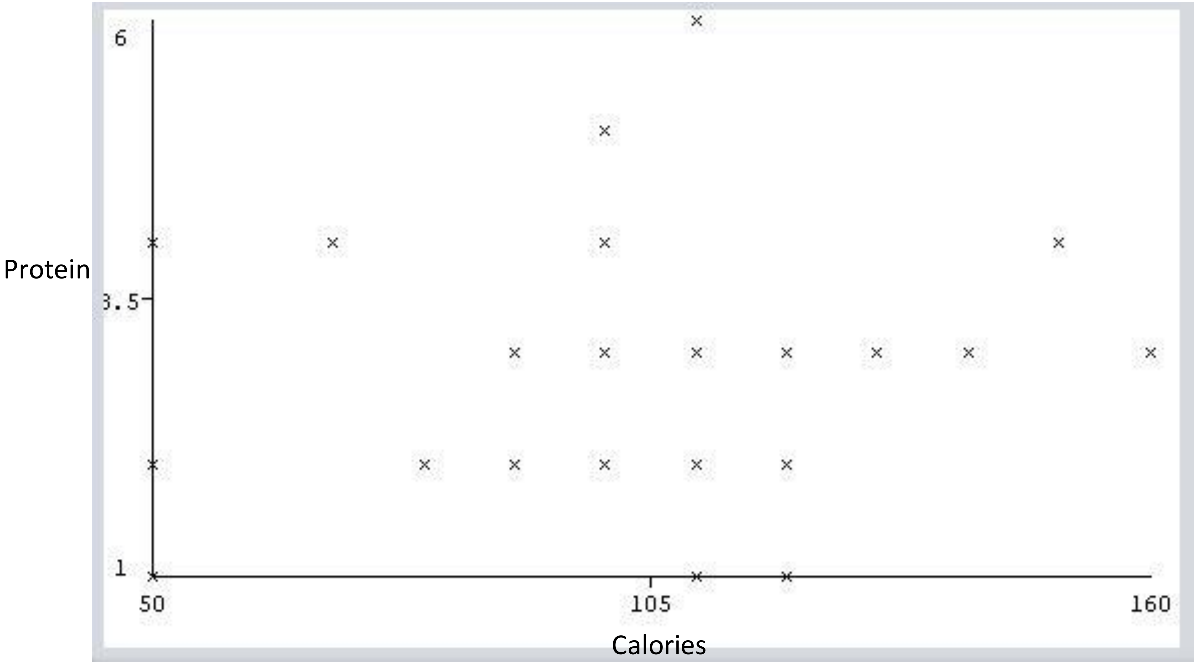
Calories_Fat correlation – From data we calculated correlation coefficient r=0.499. So we can say that Calories and Fat are positively correlated.



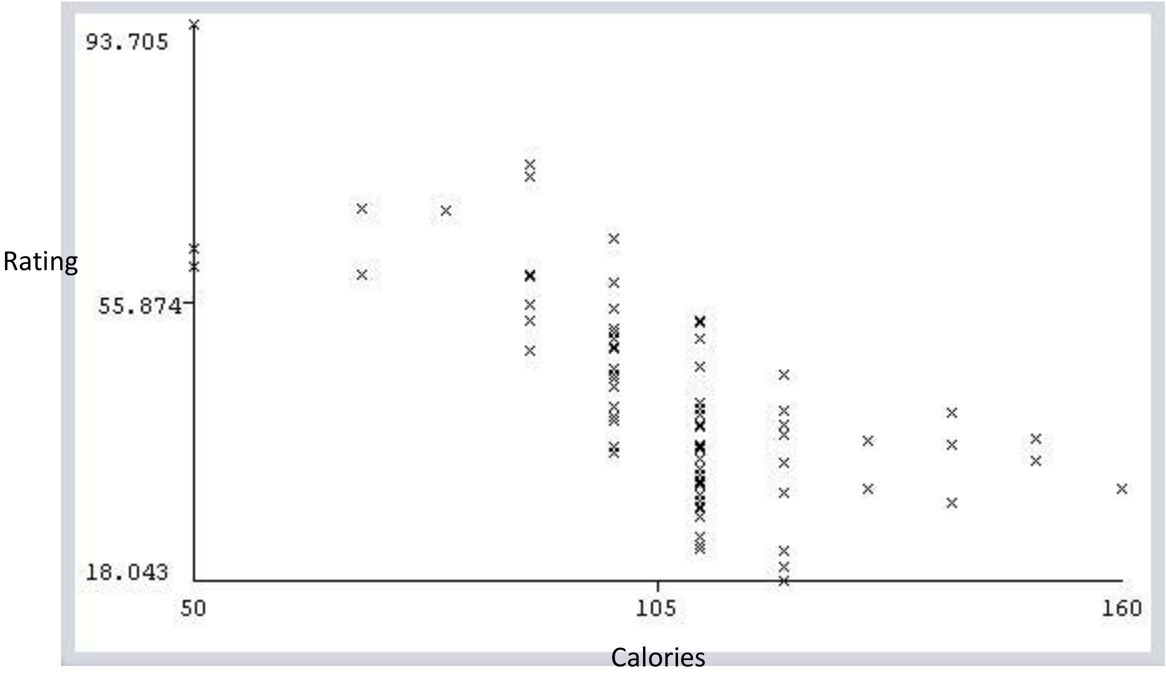
Calories_Fiber correlation – From data we calculated correlation coefficient r= -0.293. So we can say that Calories and Fiber are negatively correlated. Fiber is low on cereals which have high Calories.



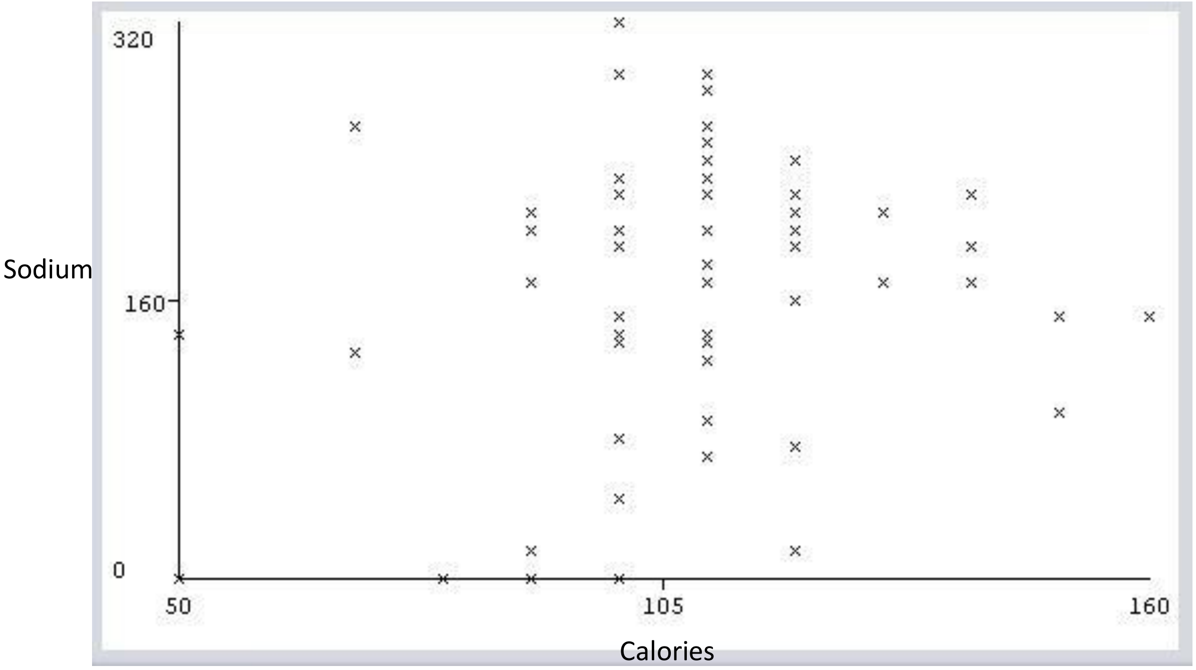
Calories_Potassium correlation – From data we calculated correlation coefficient r= -0.066. So we can say that Calories and Potassium are negatively correlated. Potassium is low when calories is high in cereals.



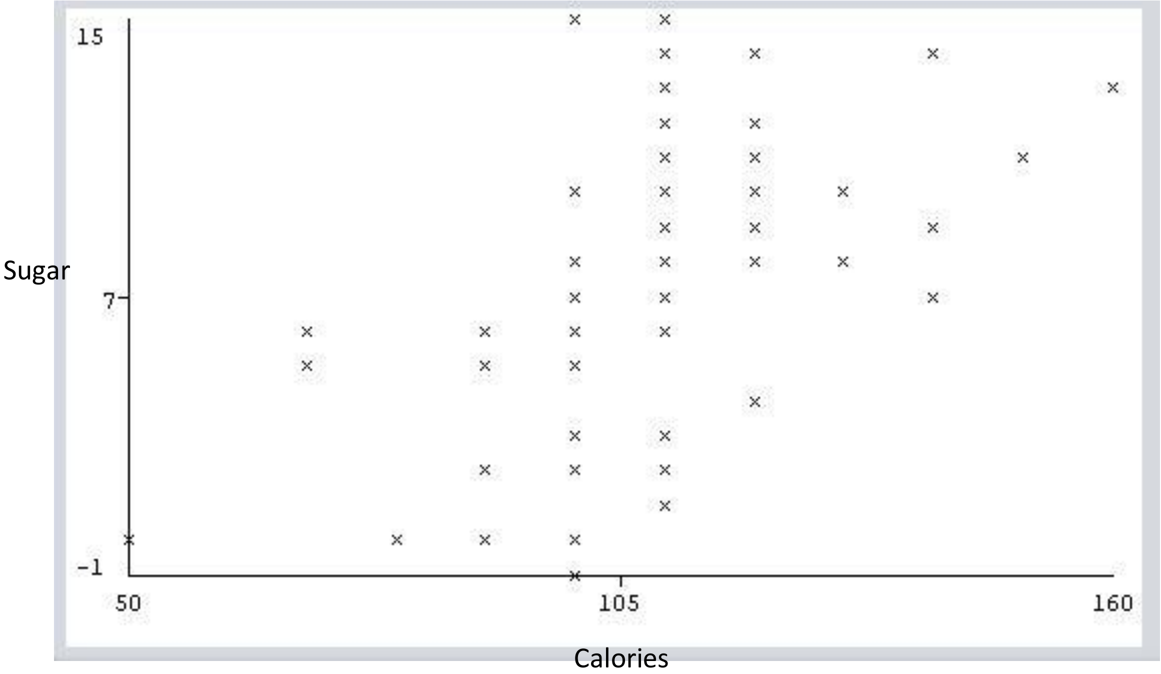
Calories_Protein correlation – From data we calculated correlation coefficient r= 0.019 which is close to zero. So we can say that Calories and Protein are not correlated.



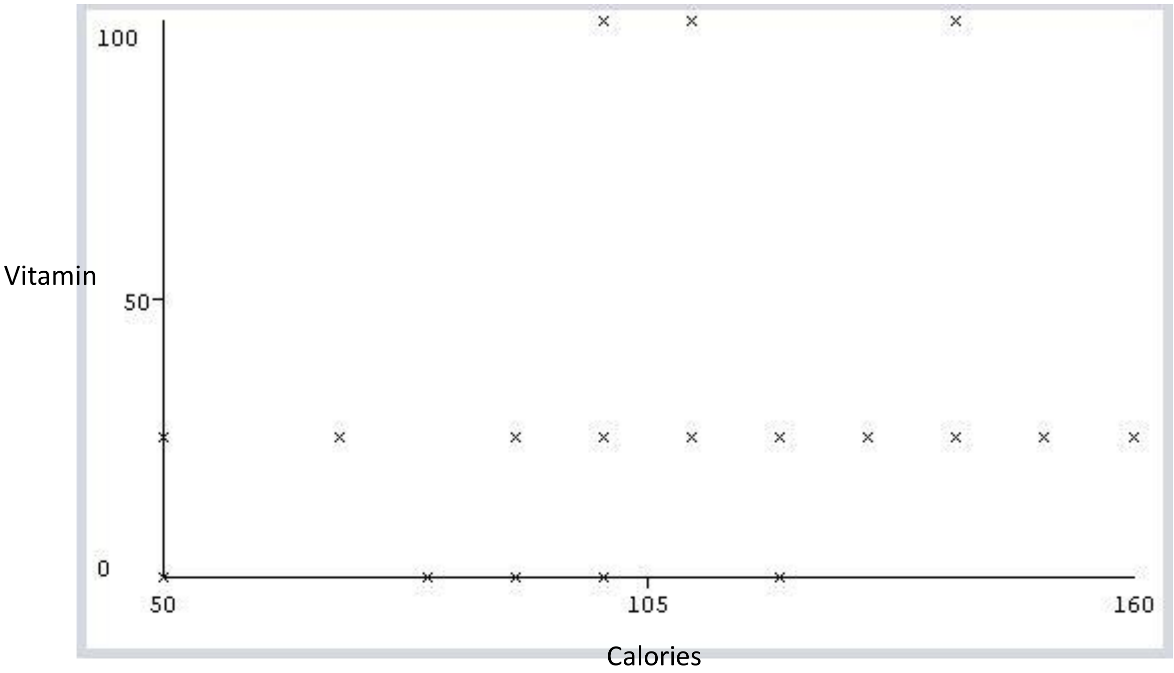
Calories_Rating correlation – From data we calculated correlation coefficient r= -0.689. So we can say that Calories and Rating are negatively correlated. Cereals which have high calories are not popular.



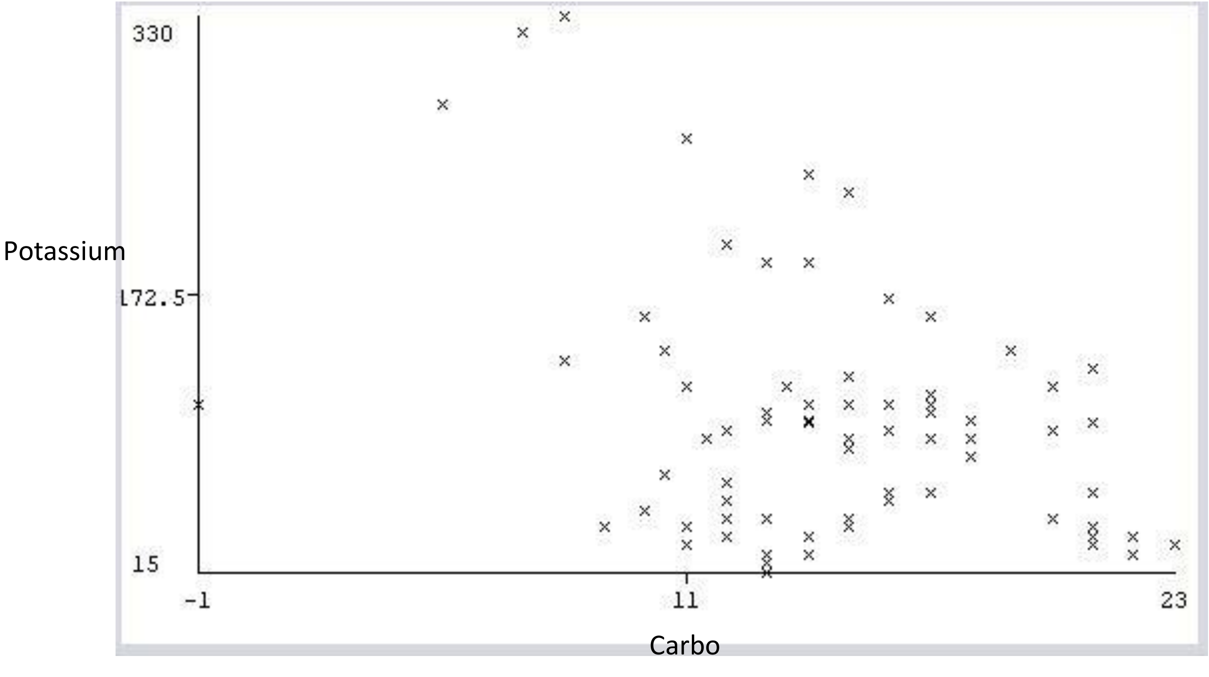
Calories_Sodium correlation – From data we calculated correlation coefficient r= 0.301. So we can say that Calories and Sodium are positively correlated. Cereals which have high calories also have high sodium.



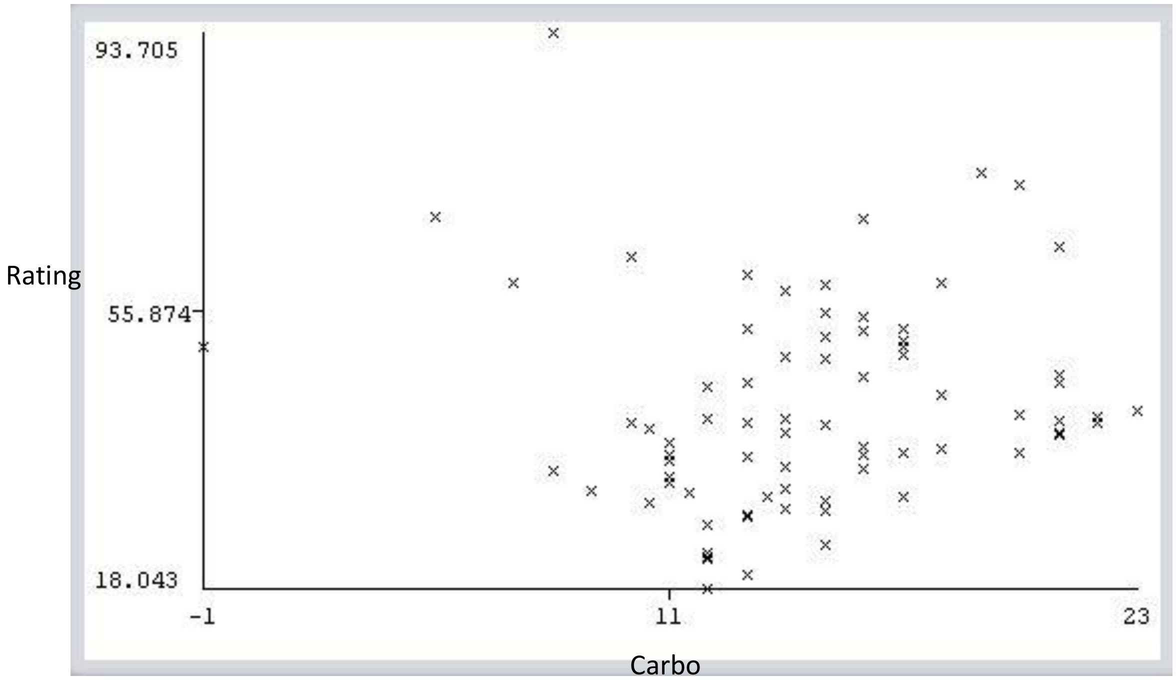
Calories_Sugar correlation – From data we calculated correlation coefficient r= 0.562. So we can say that Calories and Sugar are positively correlated. Cereals having high sugar have high calories.



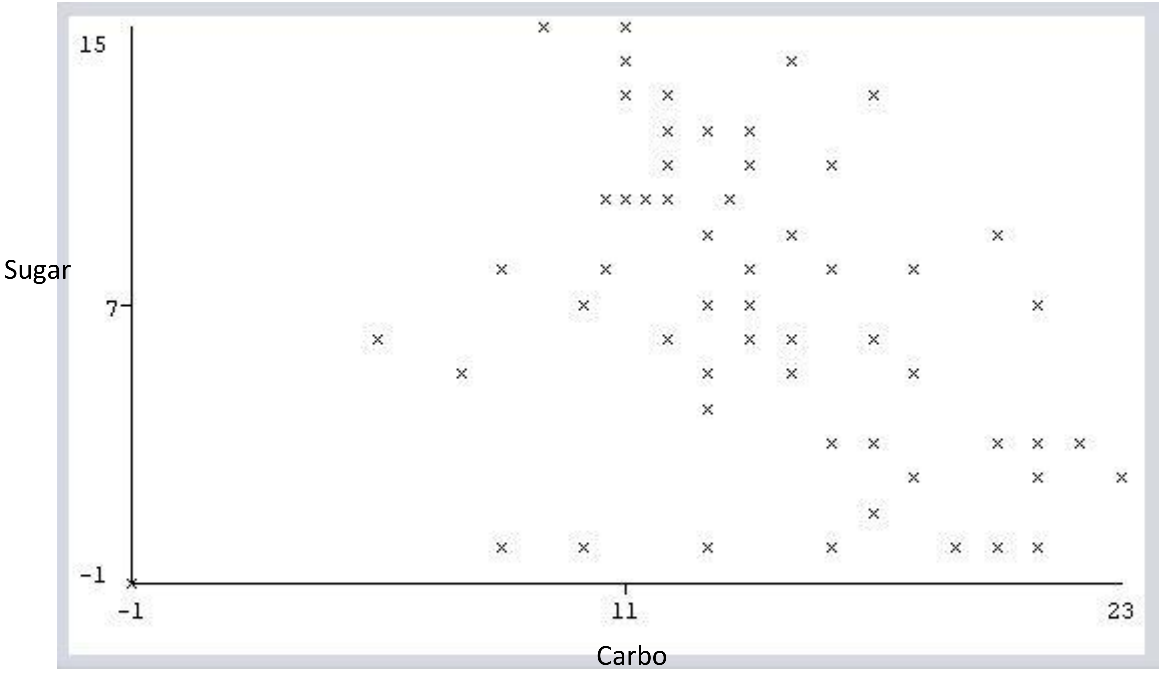
Calories_Vitamin correlation – From data we calculated correlation coefficient r= 0.265. So we can say that Calories and Vitamin are positively correlated.



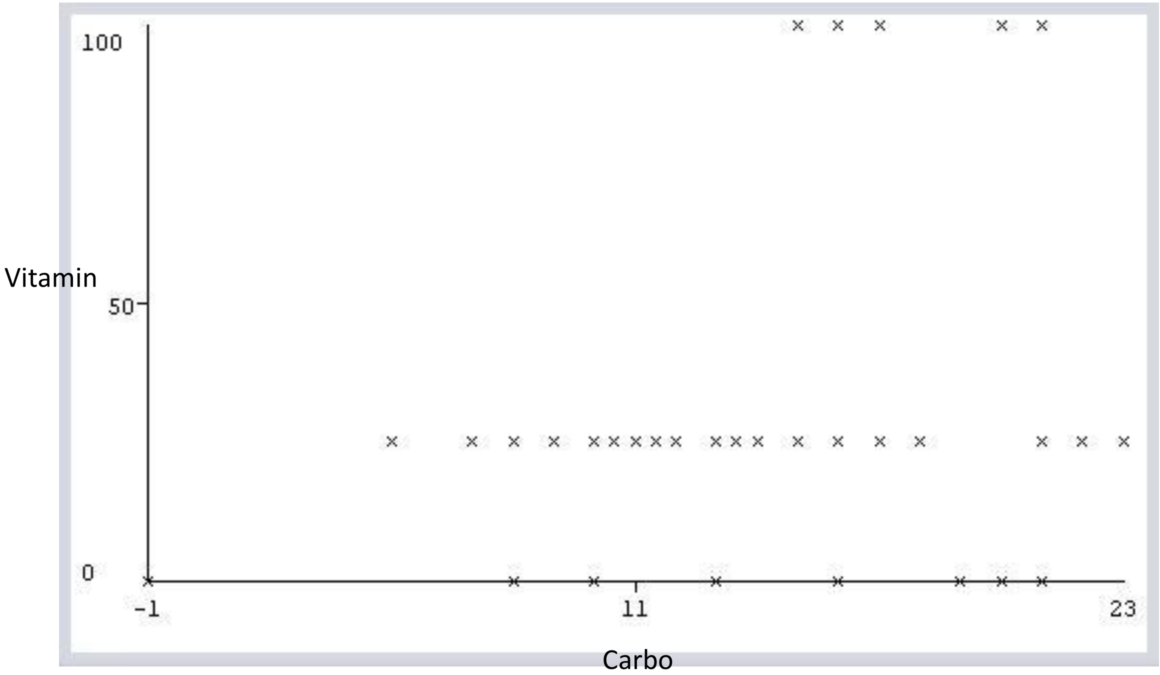
Carbo_Potassium correlation – From data we calculated correlation coefficient r= -0.349. So we can say that Carbohydrate and Potassium are negatively correlated. Cereals having low potassium have high carbohydrate.



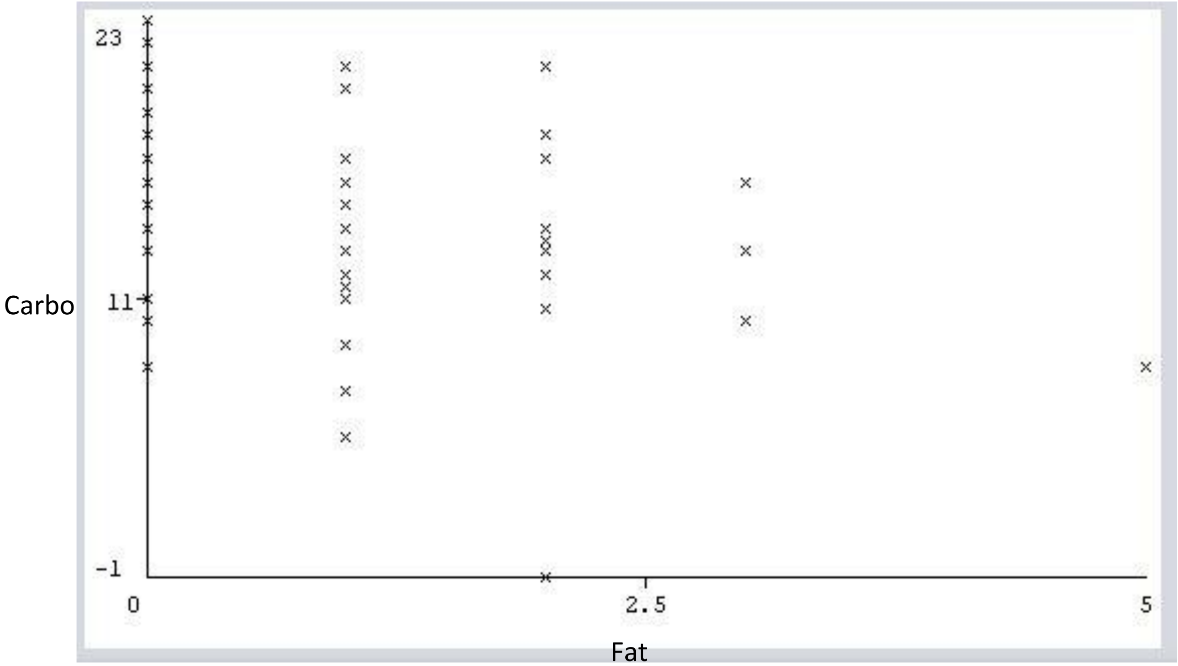
Carbo_Rating correlation – From data we calculated correlation coefficient r= 0.052. So we can say that Carbohydrate and Rating are not correlated.



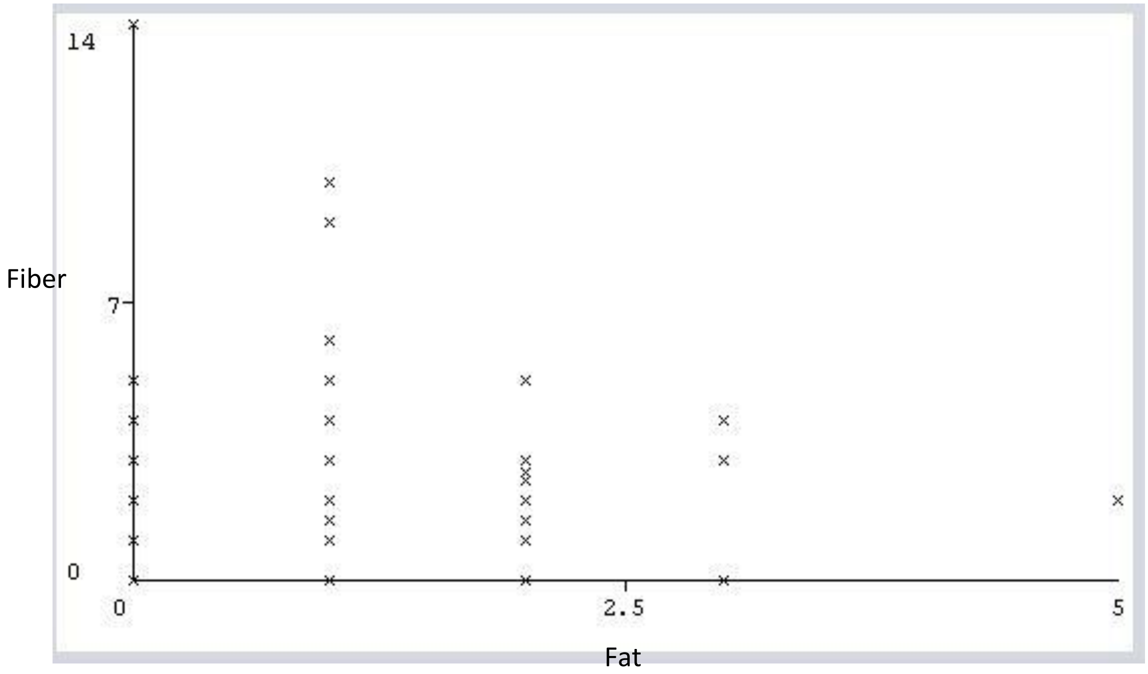
Carbo_Sugar correlation – From data we calculated correlation coefficient r= -0.331. So we can say that Carbohydrate and Sugar are negatively correlated.



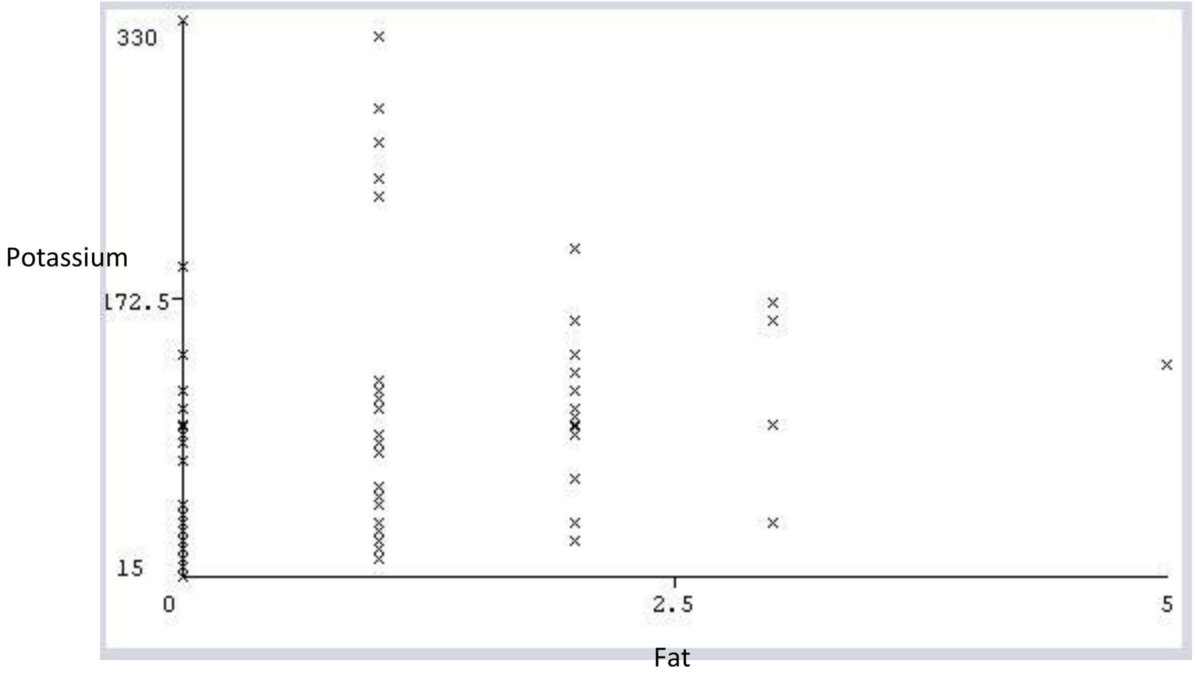
Carbo_Vitamin correlation – From data we calculated correlation coefficient r= 0.258. So we can say that Carbohydrate and Vitamin are positively correlated.



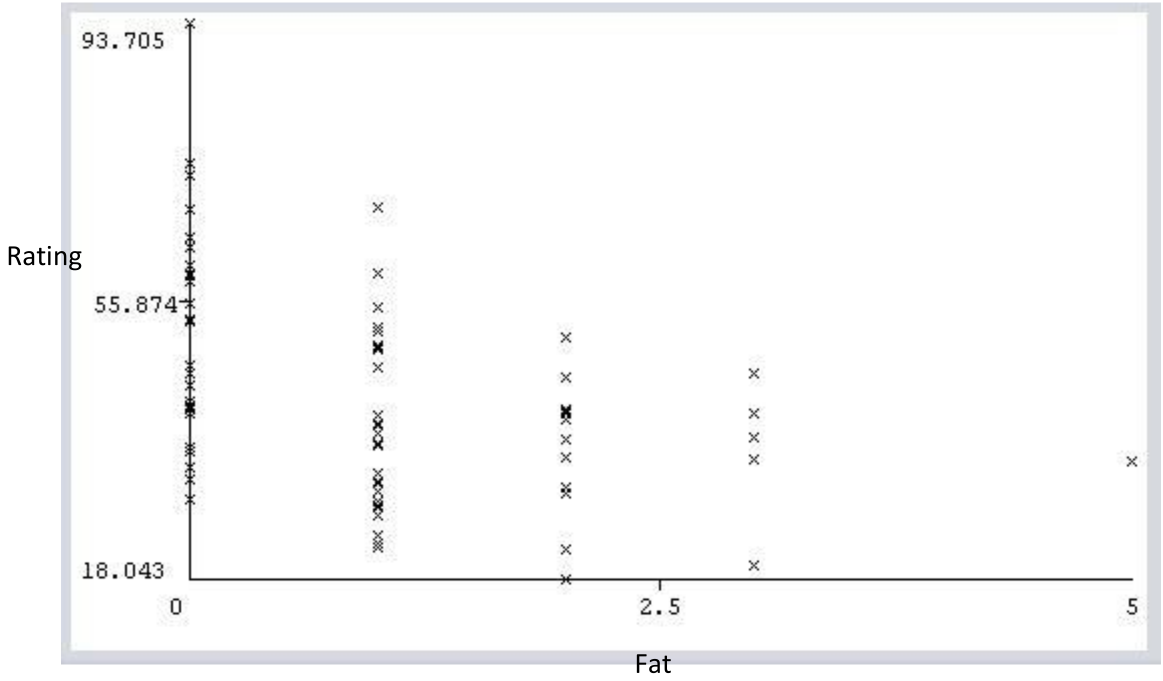
Fat_Carbo correlation – From data we calculated correlation coefficient r= -0.318. So we can say that Fat and Carbohydrate are negatively correlated. High carbo cereals have almost low fat.



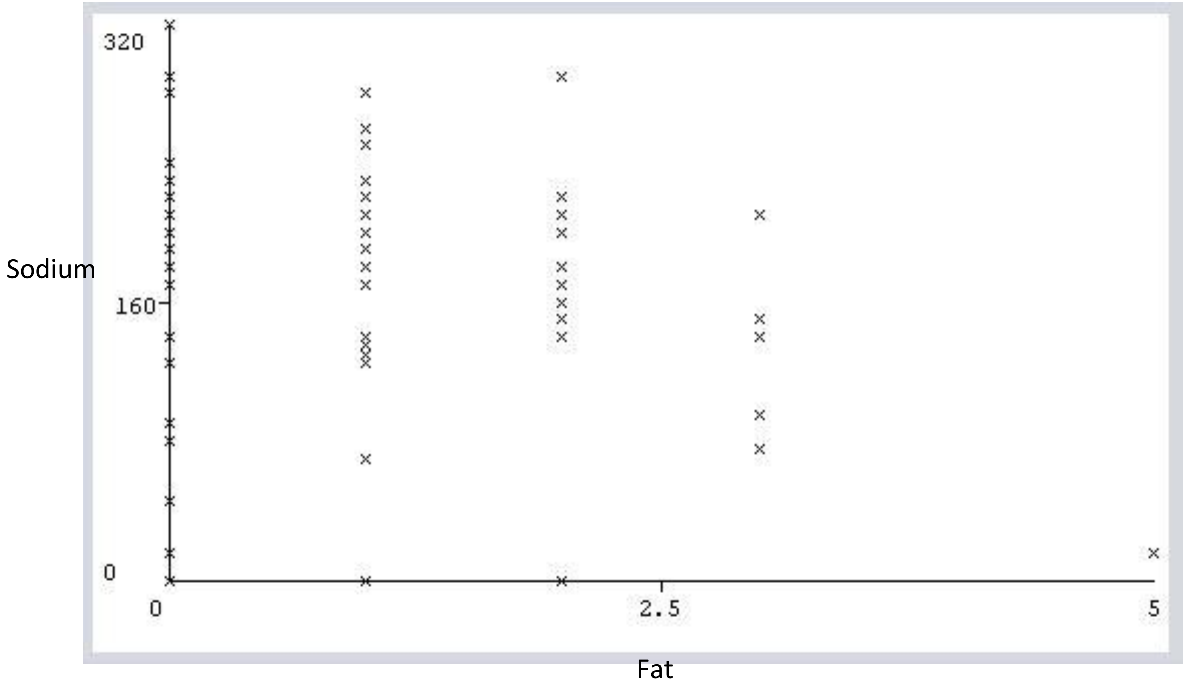
Fat_Fiber correlation – From data we calculated correlation coefficient r= 0.016 which is close to zero. So we can say that Fat and Fiber are not correlated.



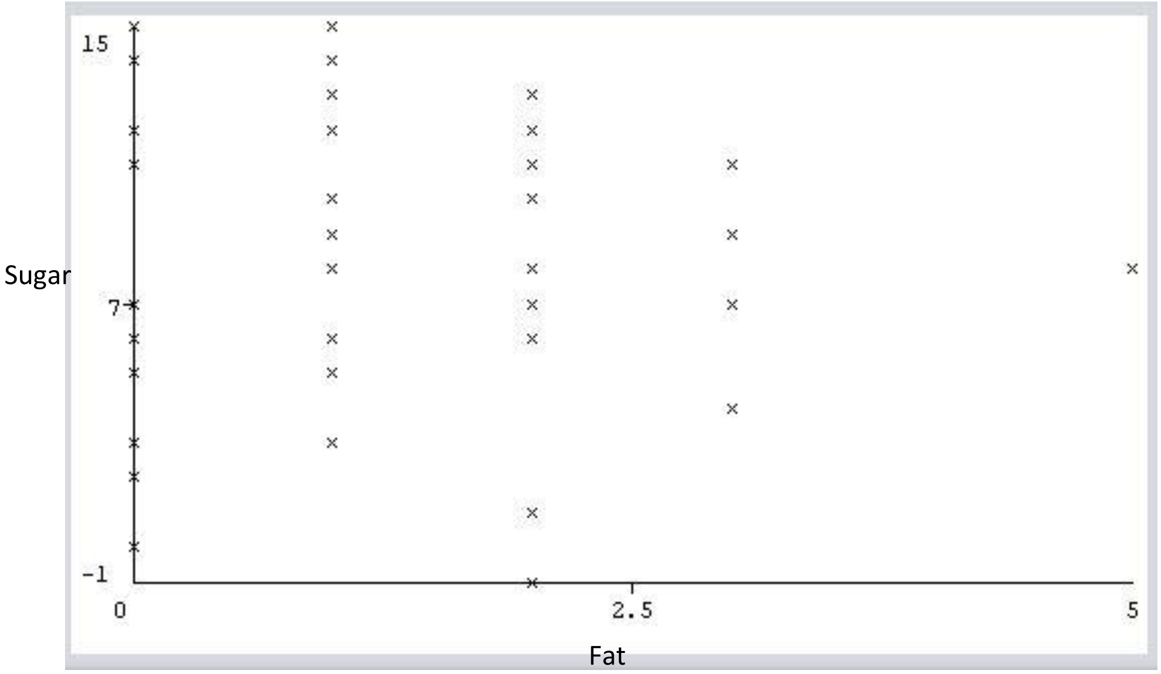
Fat_Potassium – From data we calculated correlation coefficient r= 0.193 which is close to zero. So we can say that Fat and Potassium are positively correlated.



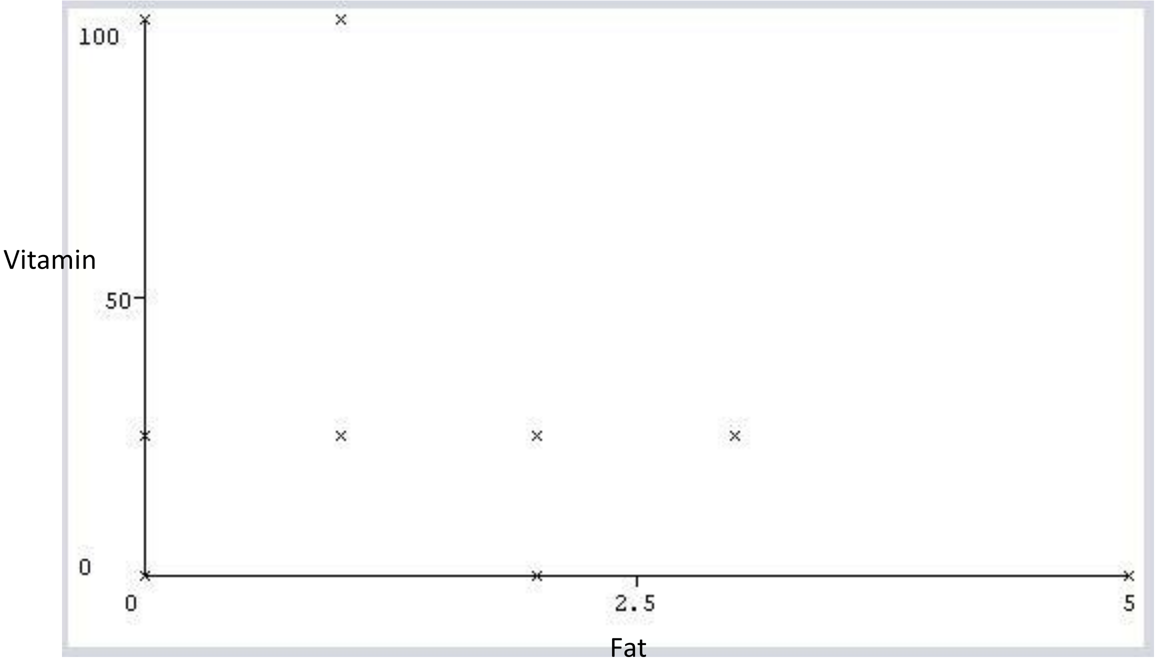
Fat_Rating correlation – From data we calculated correlation coefficient r= -0.409. So we can say that Fat and Rating are negatively correlated. Cereals containing fat are unpopular.



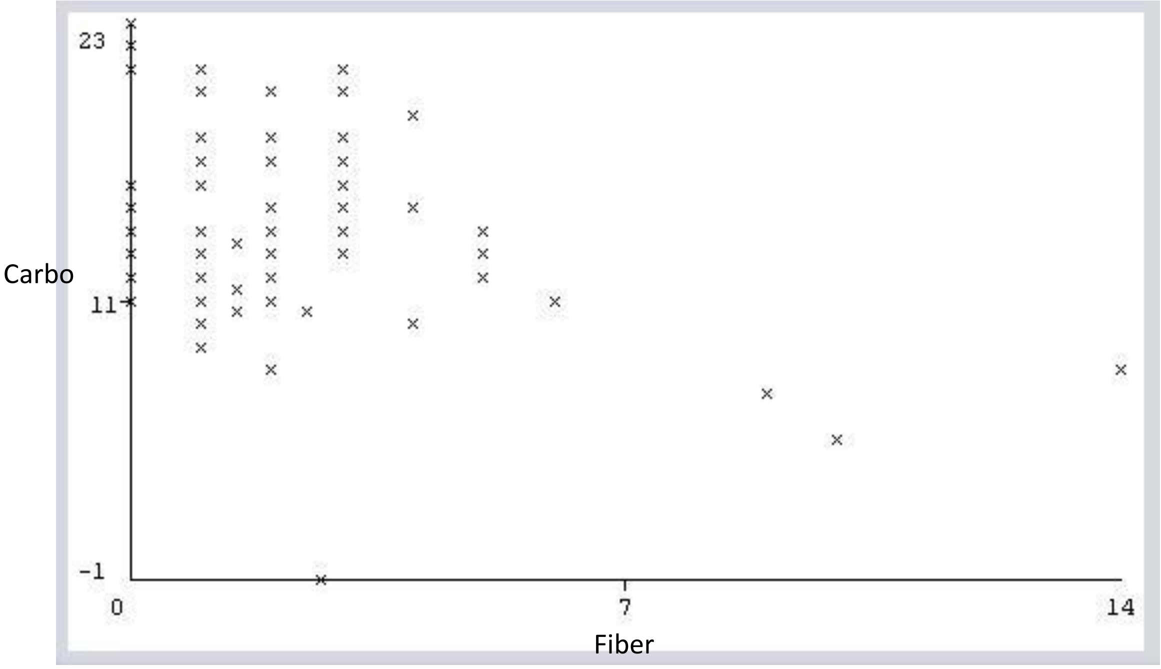
Fat_Sodium correlation – From data we calculated correlation coefficient r= -0.005 which is close to zero. So we can say that Fat and Sodium are not correlated.



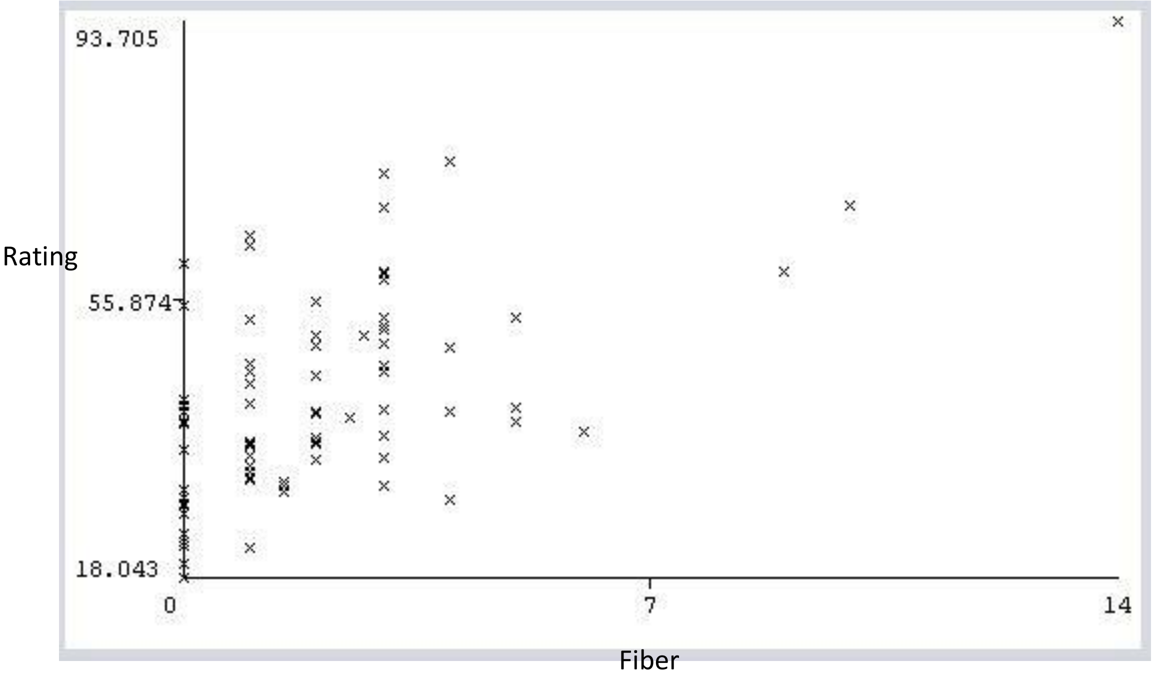
Fat_Sugar correlation – From data we calculated correlation coefficient r= 0.271. So we can say that Fat and Sugar are positively correlated. Cereals containing sugar also have fat.



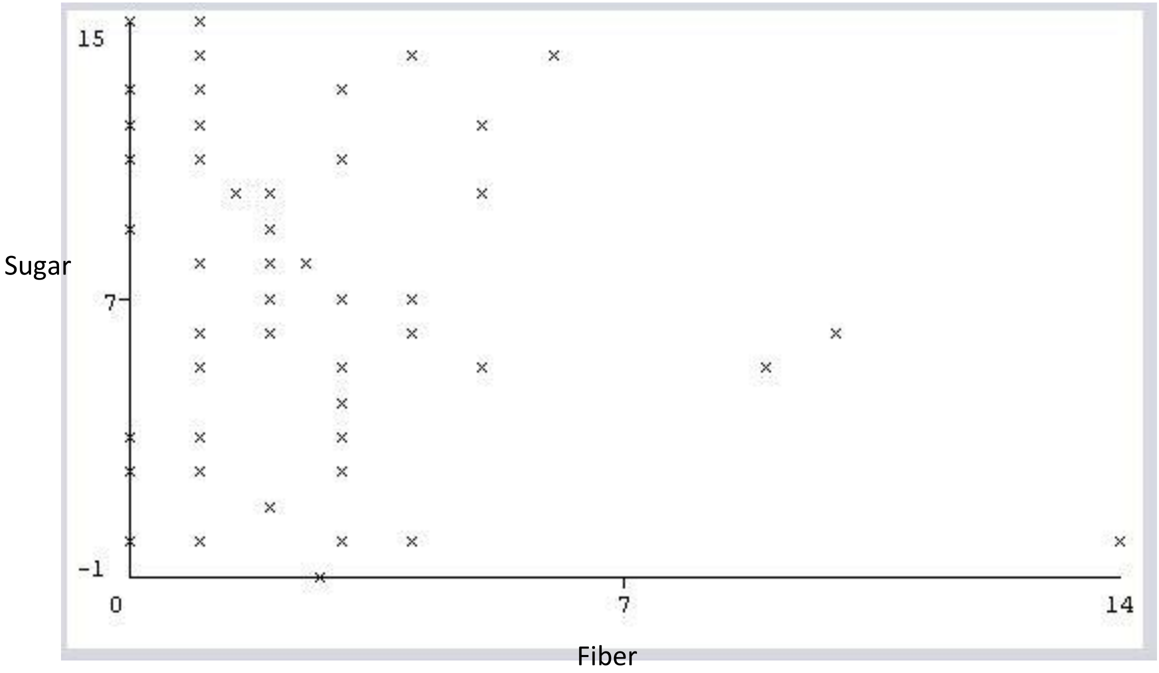
Fat_Vitamin correlation – From data we calculated correlation coefficient r= -0.031 which is close to zero. So we can say that Fat and Vitamin are not correlated.



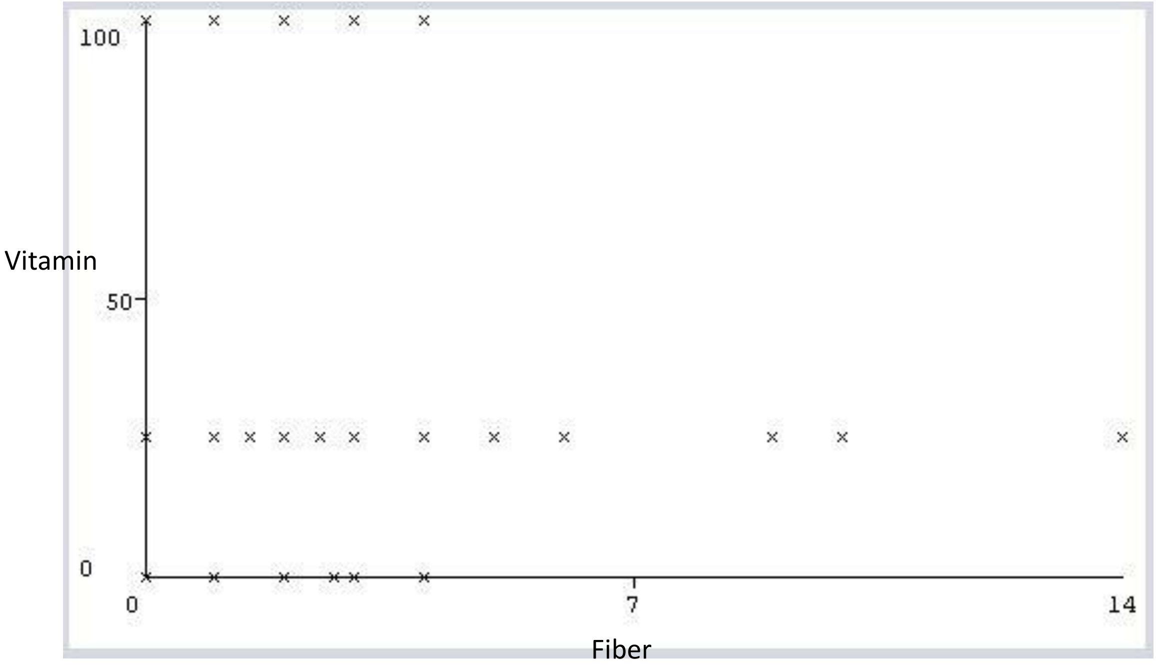
Fiber_Carbo correlation – From data we calculated correlation coefficient r= -0.356. So we can say that Fiber and Carbohydrate are negatively correlated. Cereals having High carbohydrate have low fiber.



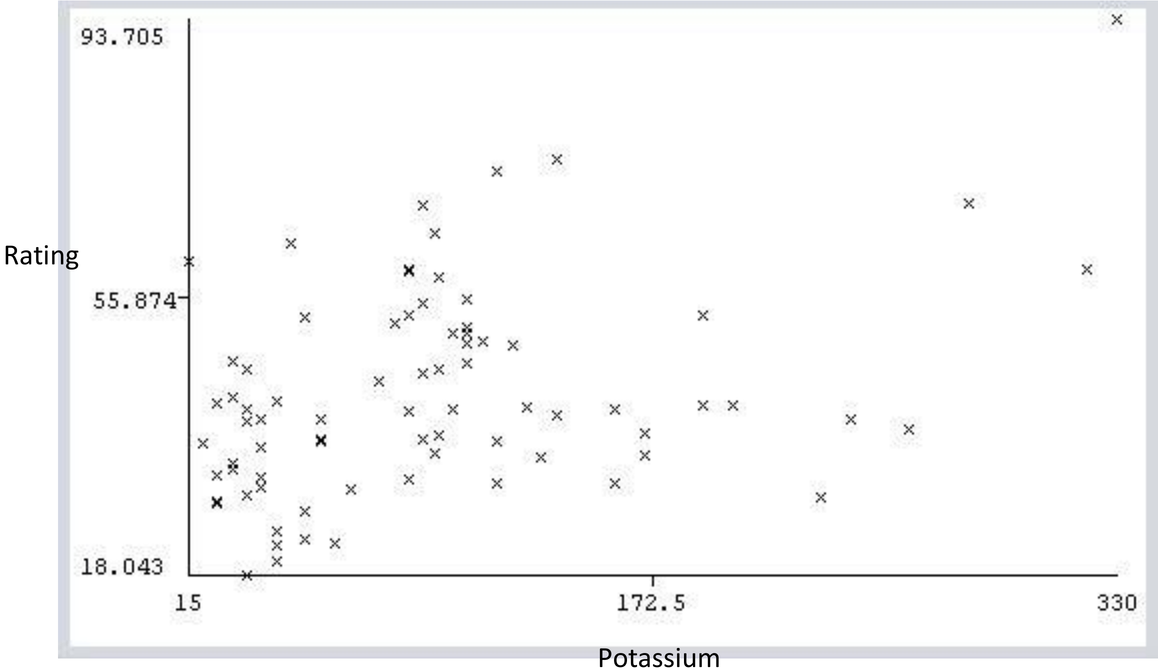
Fiber_Rating correlation – From data we calculated correlation coefficient r= 0.584. So we can say that Fiber and Rating are positively correlated. Cereals having High fiber are popular.



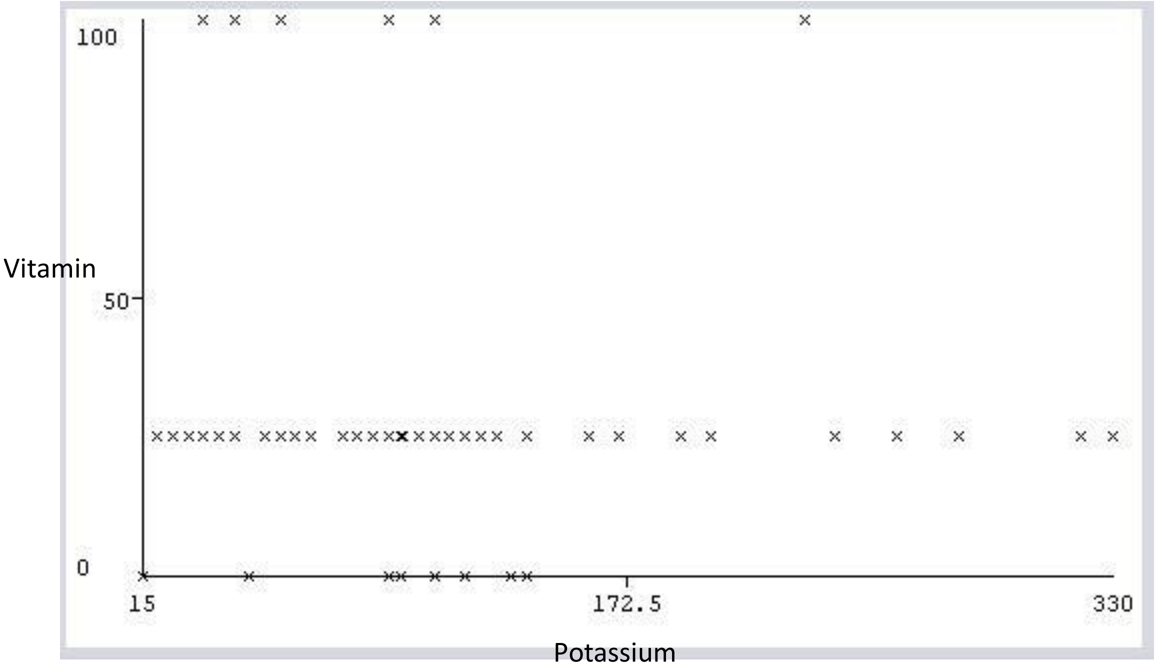
Fiber_Sugar correlation – From data we calculated correlation coefficient r= -0.141. So we can say that Fiber and Sugar are negatively correlated.



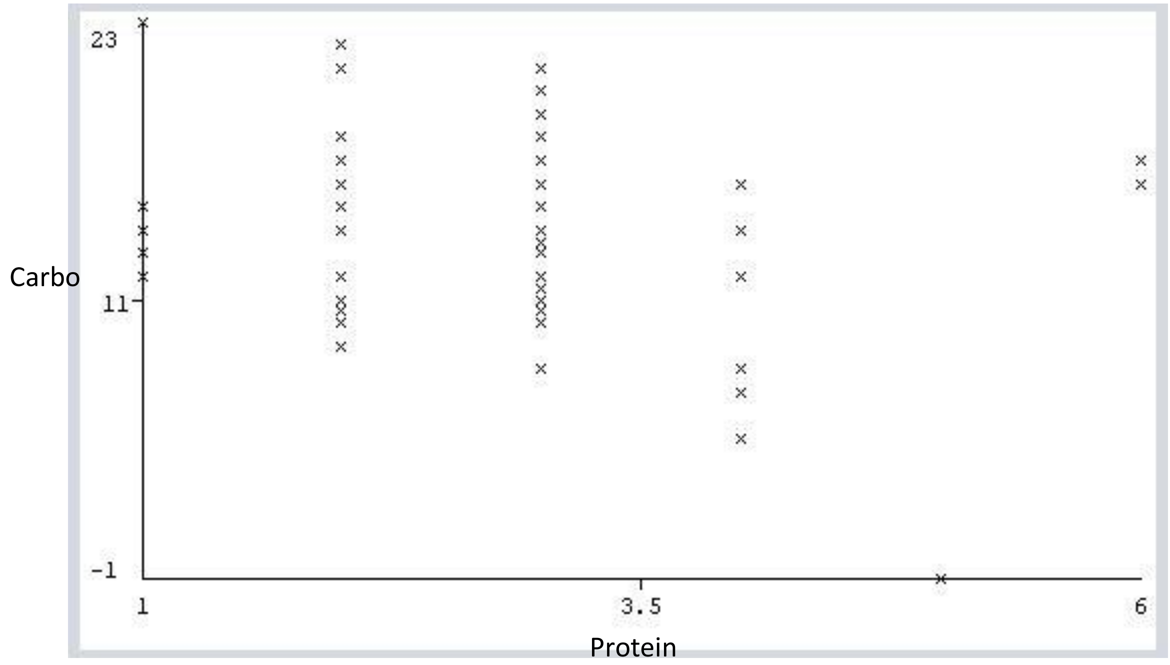
Fiber_Vitamin correlation - From data we calculated correlation coefficient r= -0.032 which is close to zero. So we can say that Fiber and Vitamin are not correlated.



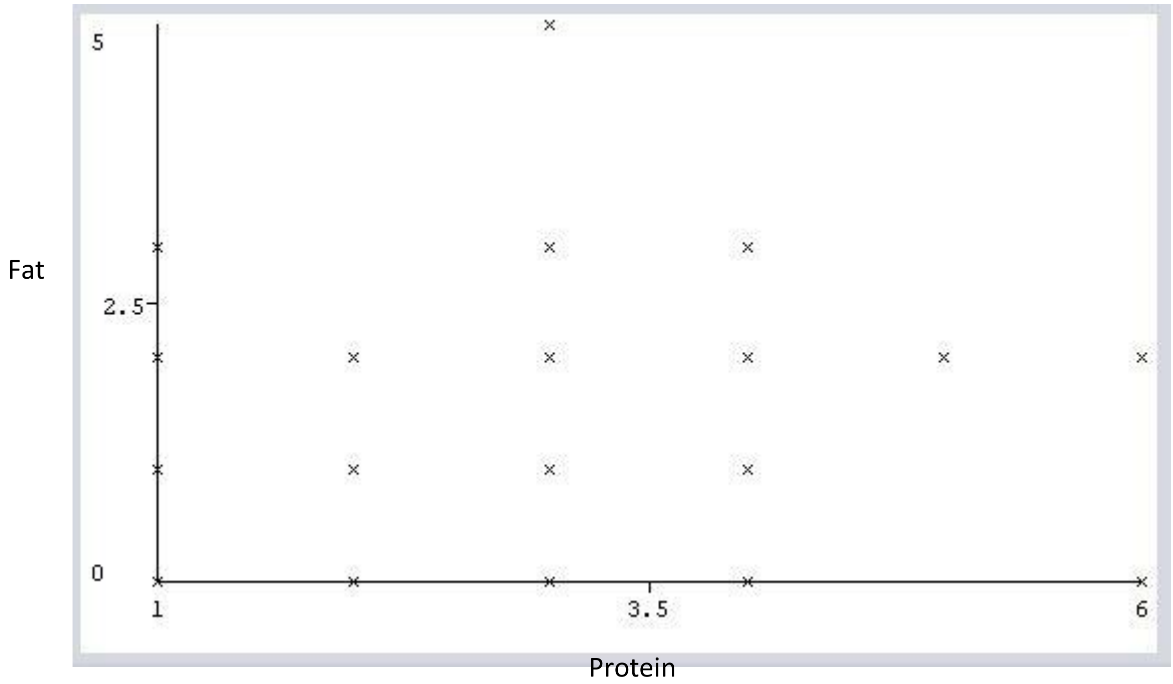
Potassium_Rating correlation – From data we calculated correlation coefficient r= 0.380. So we can say that Potassium and Rating are positively correlated. Cereals high in potassium are rated high.



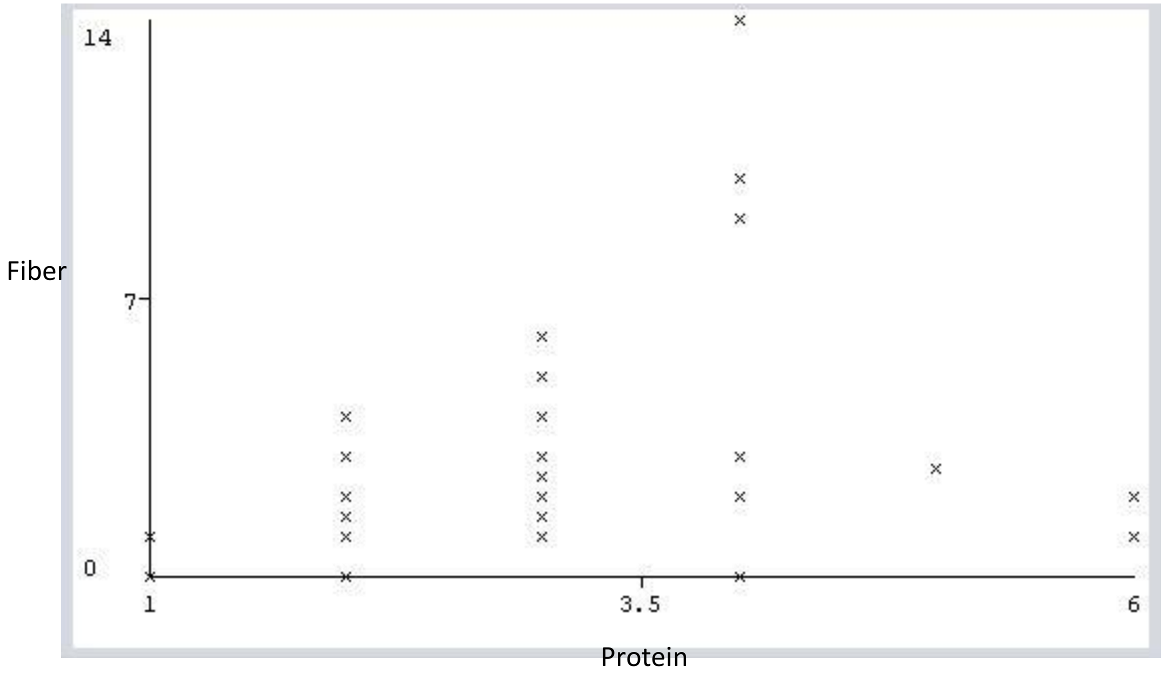
Potassium_Vitamin correlation – From data we calculated correlation coefficient r= 0.021 which is close to zero. So we can say that Potassium and Vitamin are not correlated.



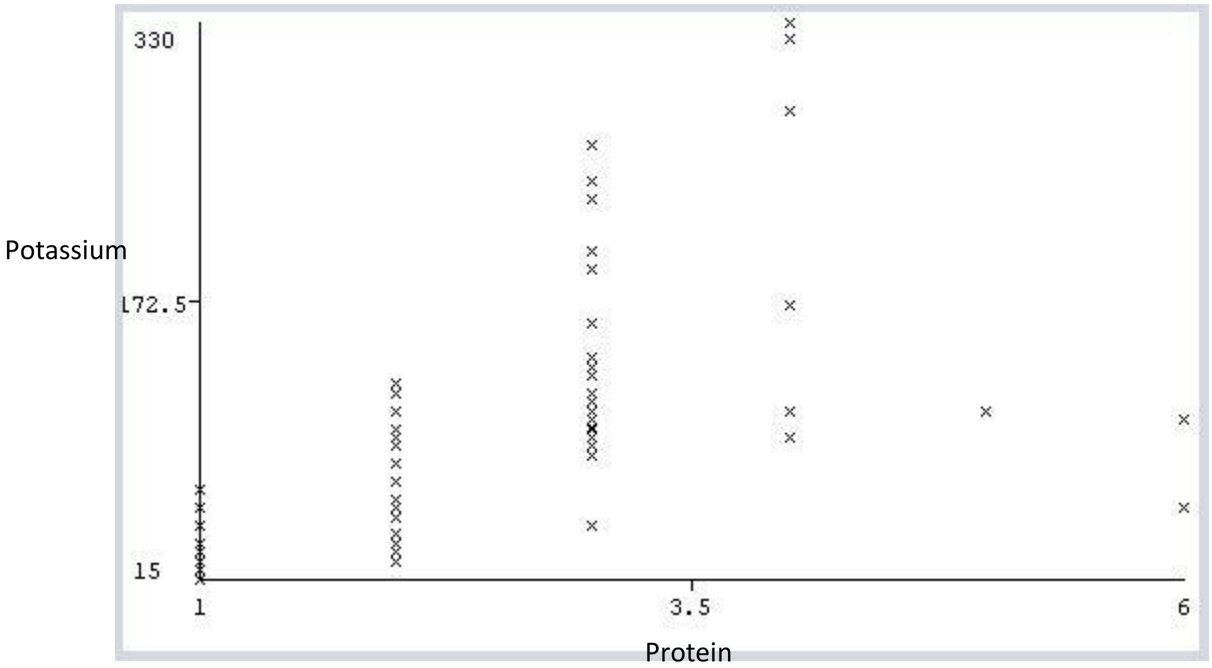
Protein_Carbo correlation – From data we calculated correlation coefficient r= -0.131. So we can say that Protein and Carbohydrate are negatively correlated.



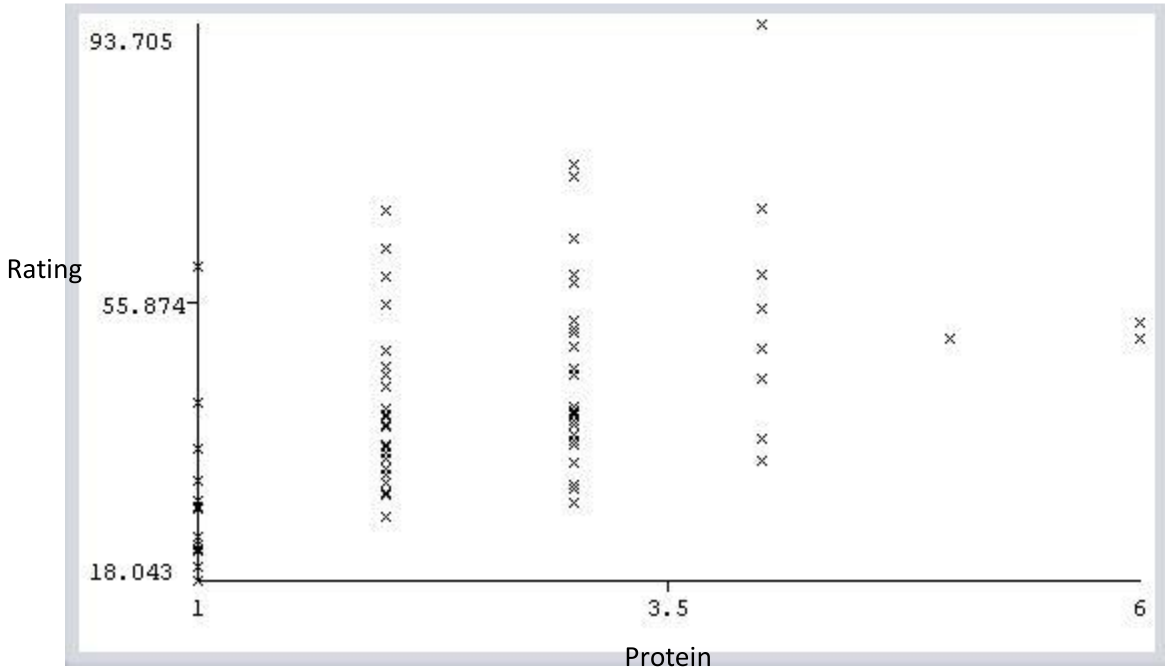
Protein_Fat correlation – From data we calculated correlation coefficient r= 0.208. So we can say that Protein and Fat are positively correlated.



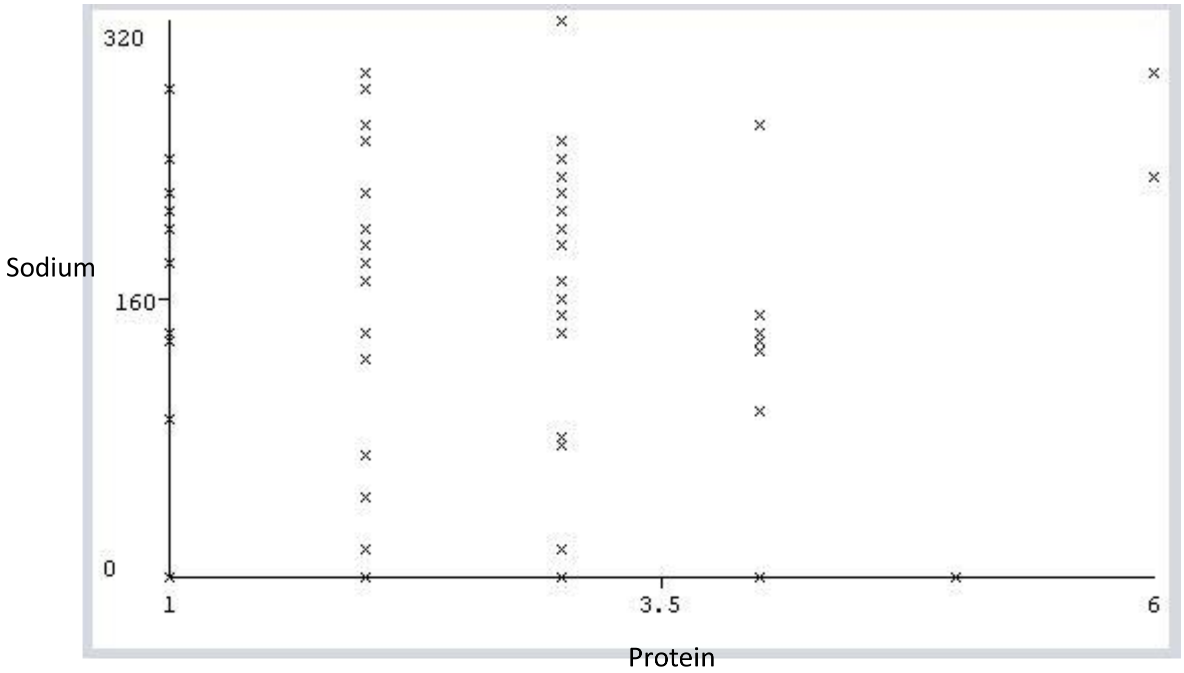
Protein_Fiber correlation – From data we calculated correlation coefficient r= 0.500. So we can say that Protein and Fiber are positively correlated. Cereal having high protein also have high fiber.



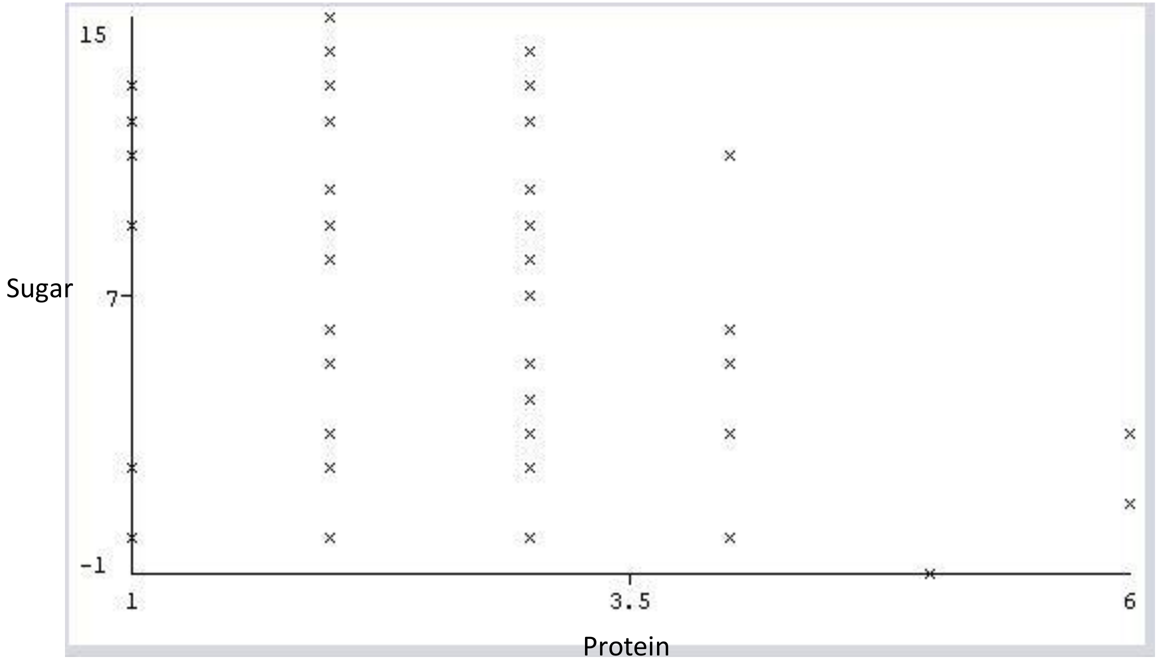
Protein_Potassium correlation – From data we calculated correlation coefficient r= 0.549. So we can say that Protein and Potassium are positively correlated. Cereal having high protein also have high Potassium.



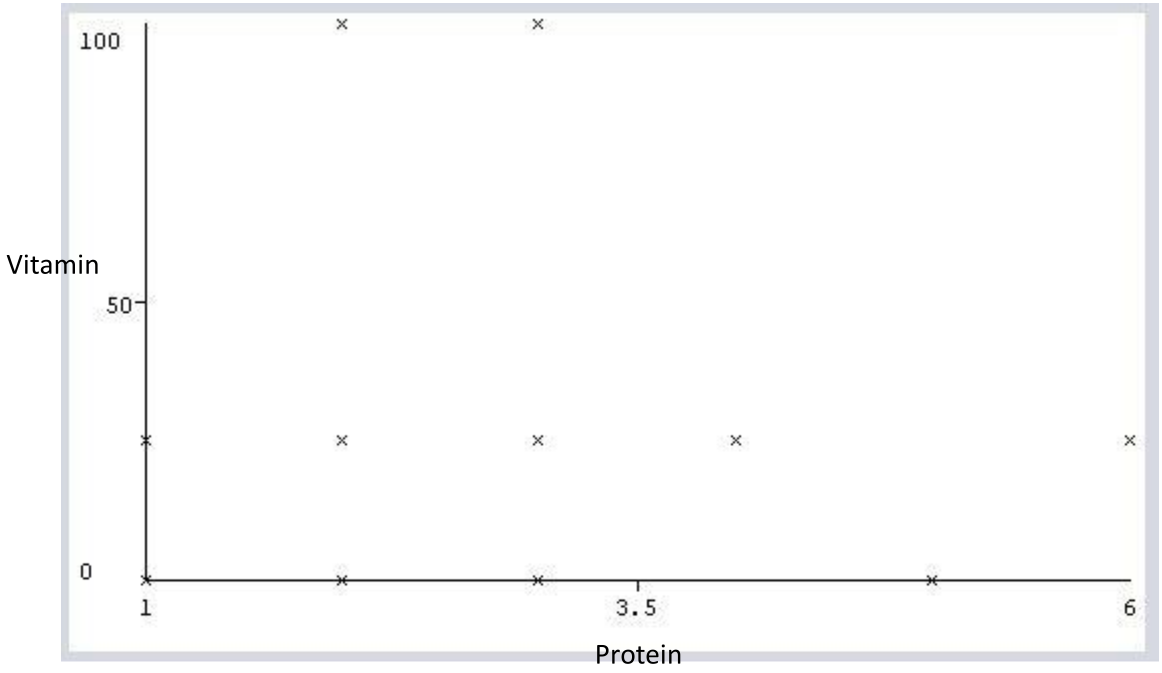
Protein_Rating correlation – From data we calculated correlation coefficient r= 0.471. So we can say that Protein and Rating are positively correlated. Cereal having high protein also have high Rating.



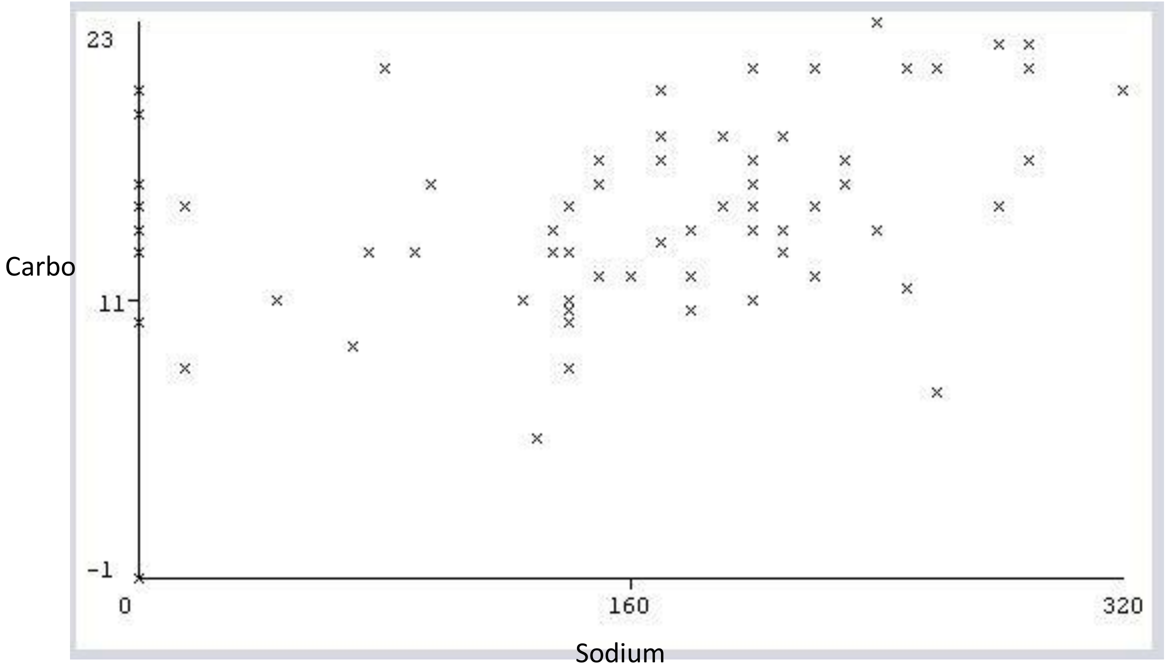
Protein_Sodium correlation – From data we calculated correlation coefficient r= -0.054 which is close to zero. So we can say that Protein and Sodium are not correlated.



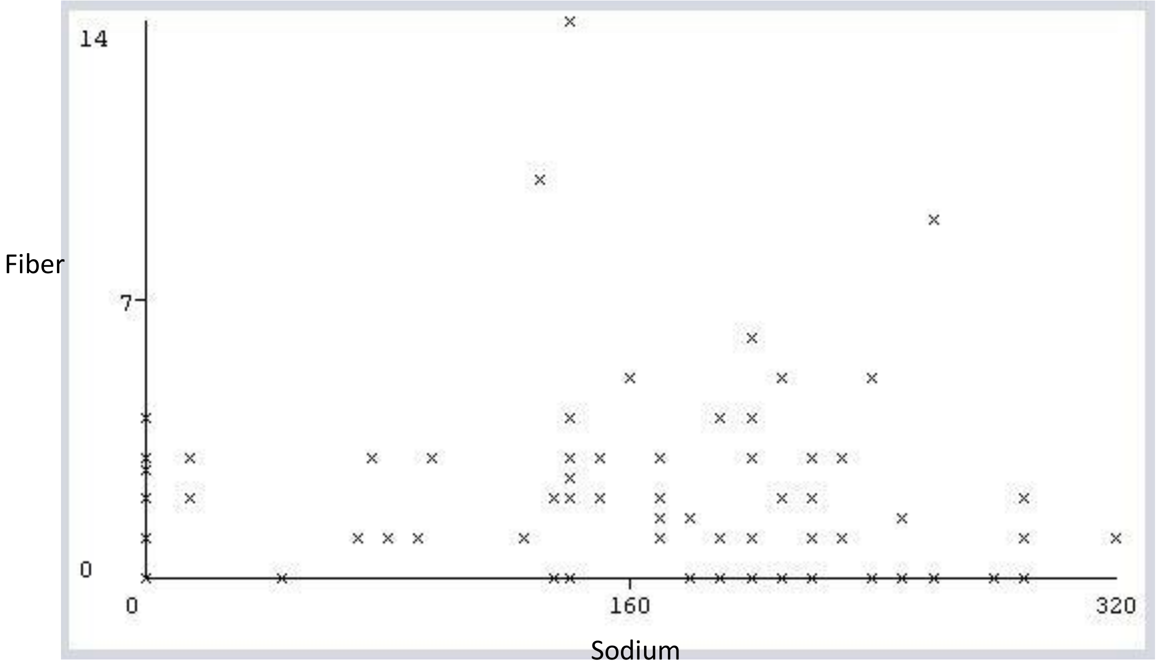
Protein_Sugar correlation – From data we calculated correlation coefficient r= -0.329. So we can say that Protein and Sugar are negatively correlated.



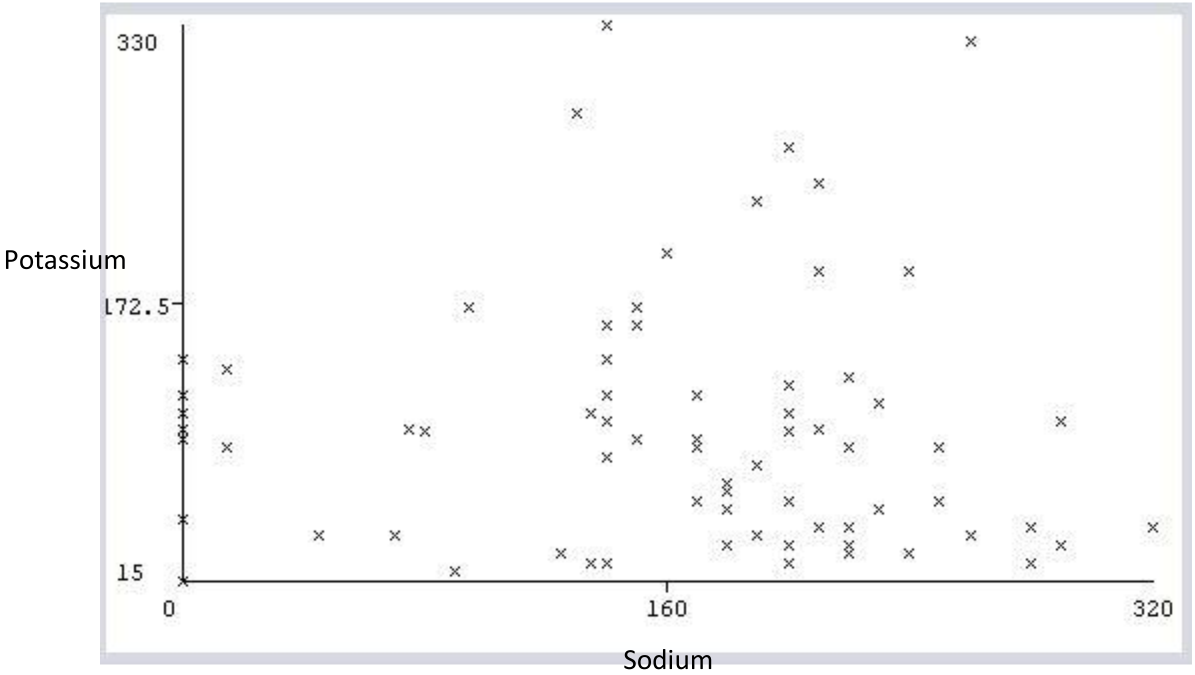
Protein_Vitamin correlation – From data we calculated correlation coefficient r= 0.007 which is close to zero. So we can say that Protein and Vitamin are not correlated.



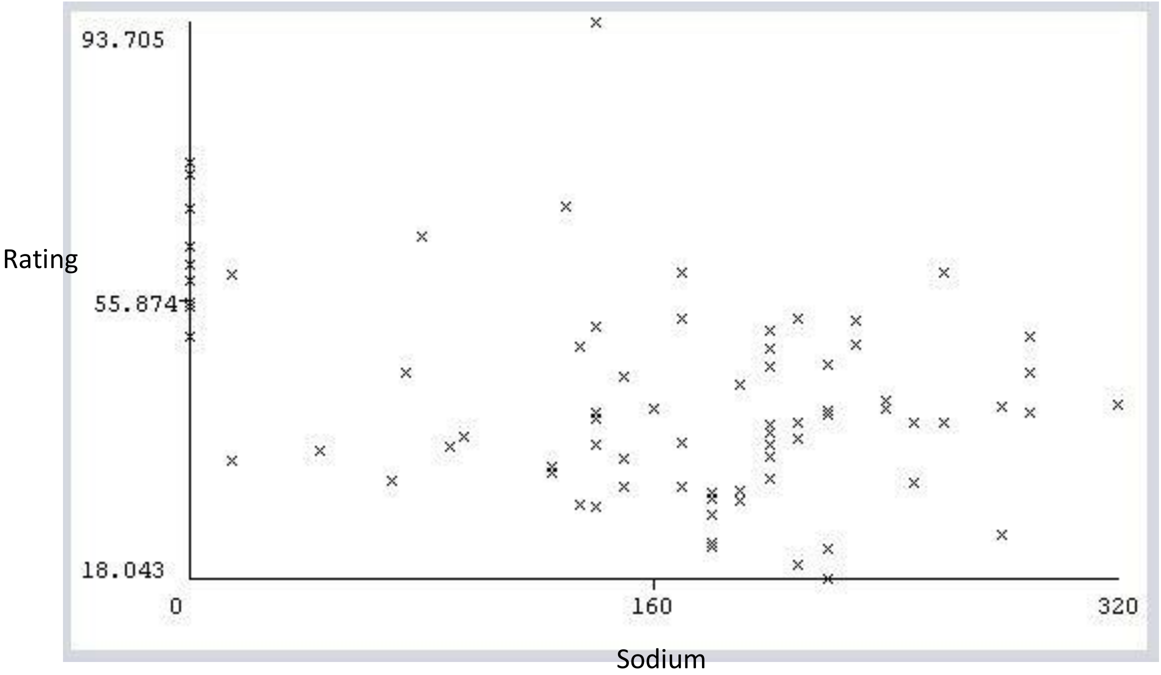
Sodium_Carbo correlation – From data we calculated correlation coefficient r= 0.356. So we can say that Sodium and Carbohydrate are positively correlated.



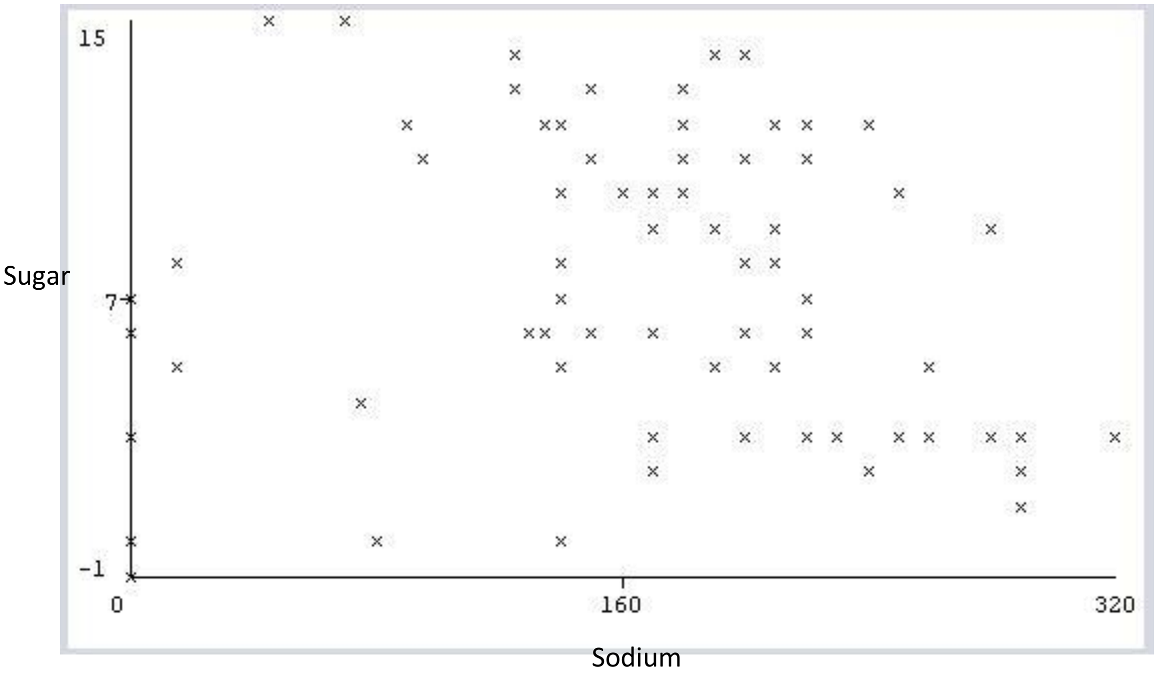
Sodium_Fiber correlation – From data we calculated correlation coefficient r= -0.071 which is closer to zero. So we can say that Sodium and Fiber are not correlated.



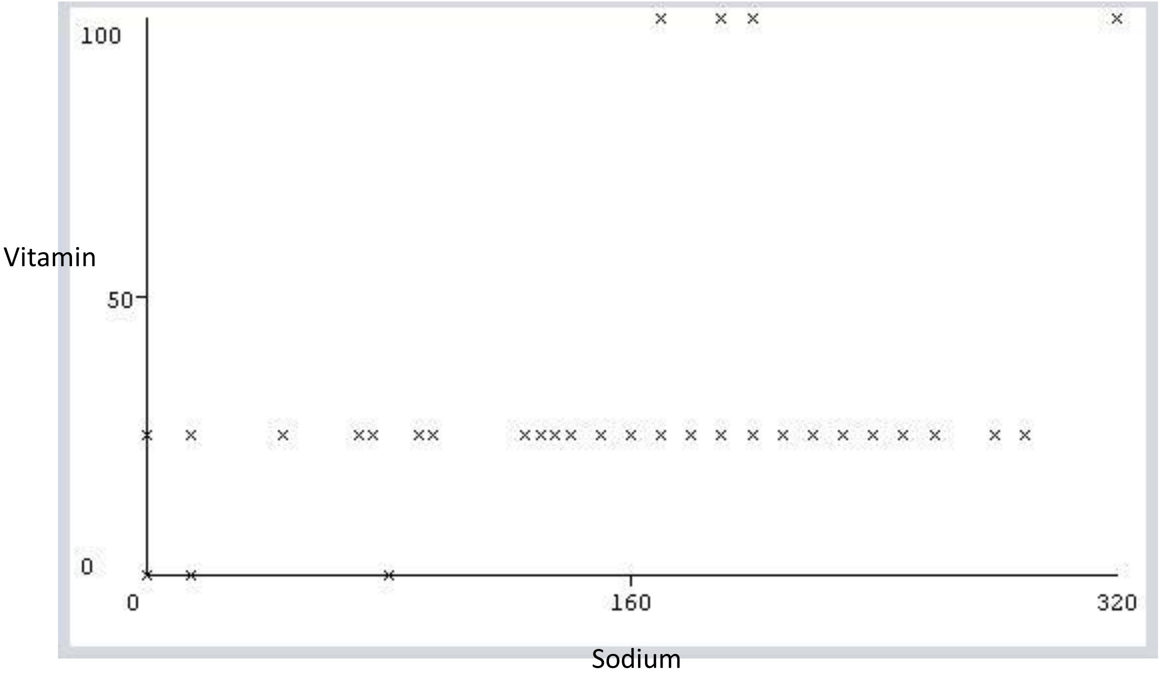
Sodium_Potassium correlation – From data we calculated correlation coefficient r= -0.032 which is closer to zero. So we can say that Sodium and Potassium are not correlated.



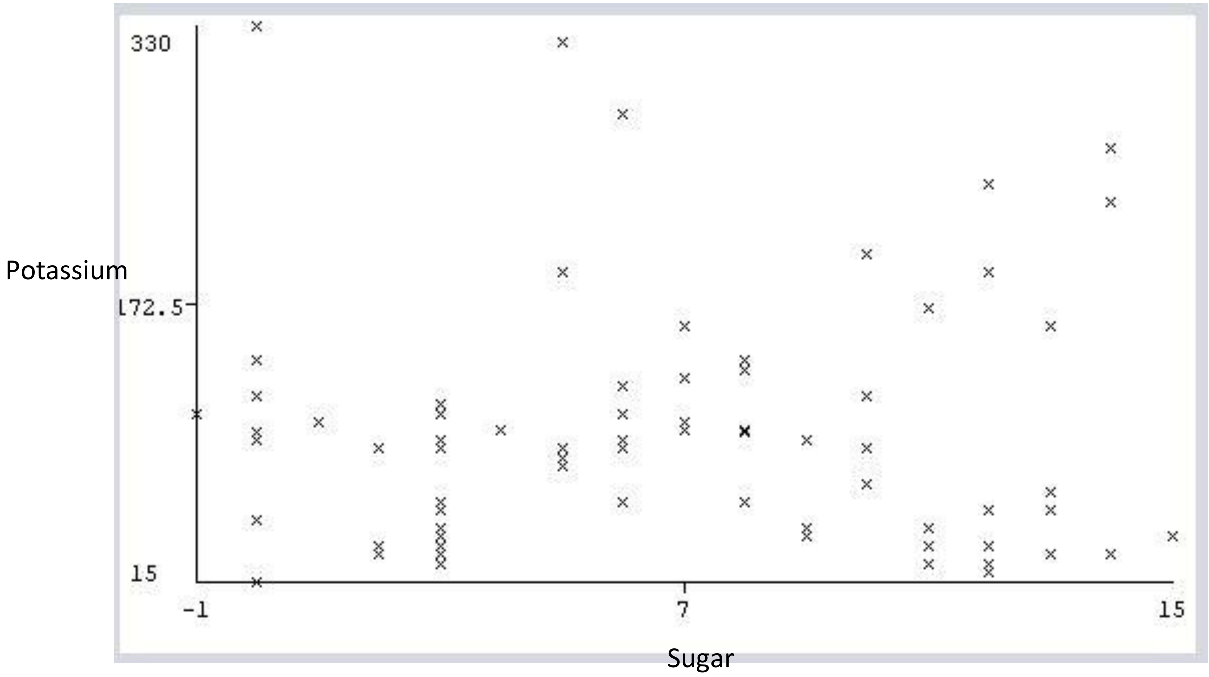
Sodium_Rating correlation – From data we calculated correlation coefficient r= -0.401. So we can say that Sodium and Rating are negatively correlated.. High sodium cereals are low rated.



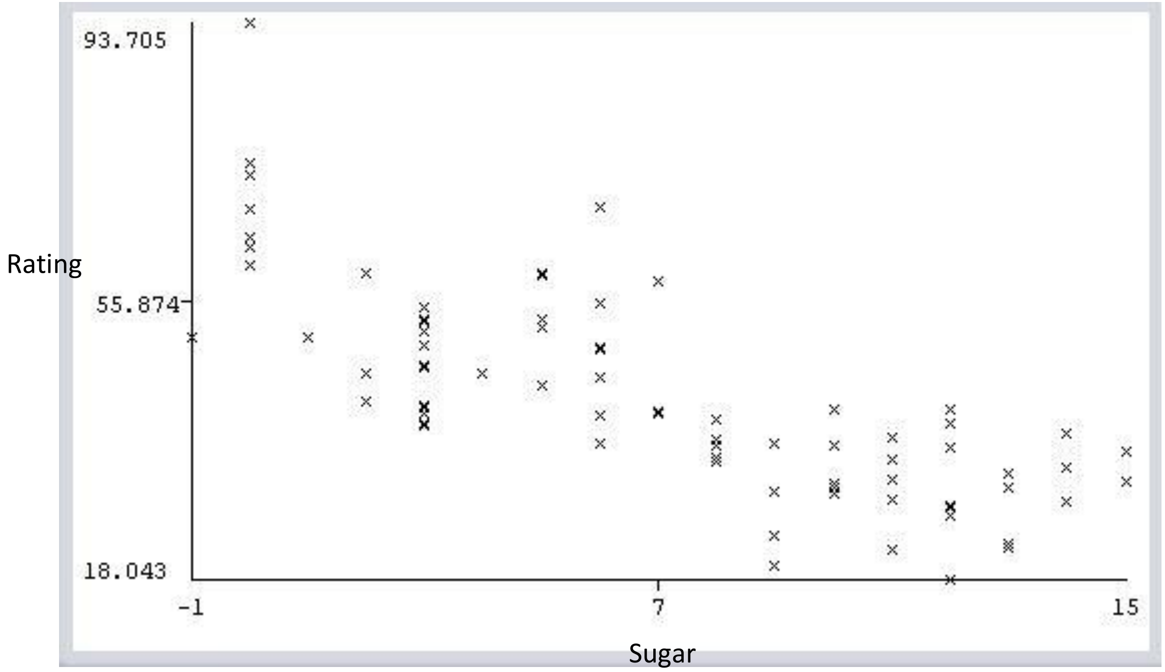
Sodium_Sugar correlation – From data we calculated correlation coefficient r= 0.101. So we can say that Sodium and Sugar are positively correlated.



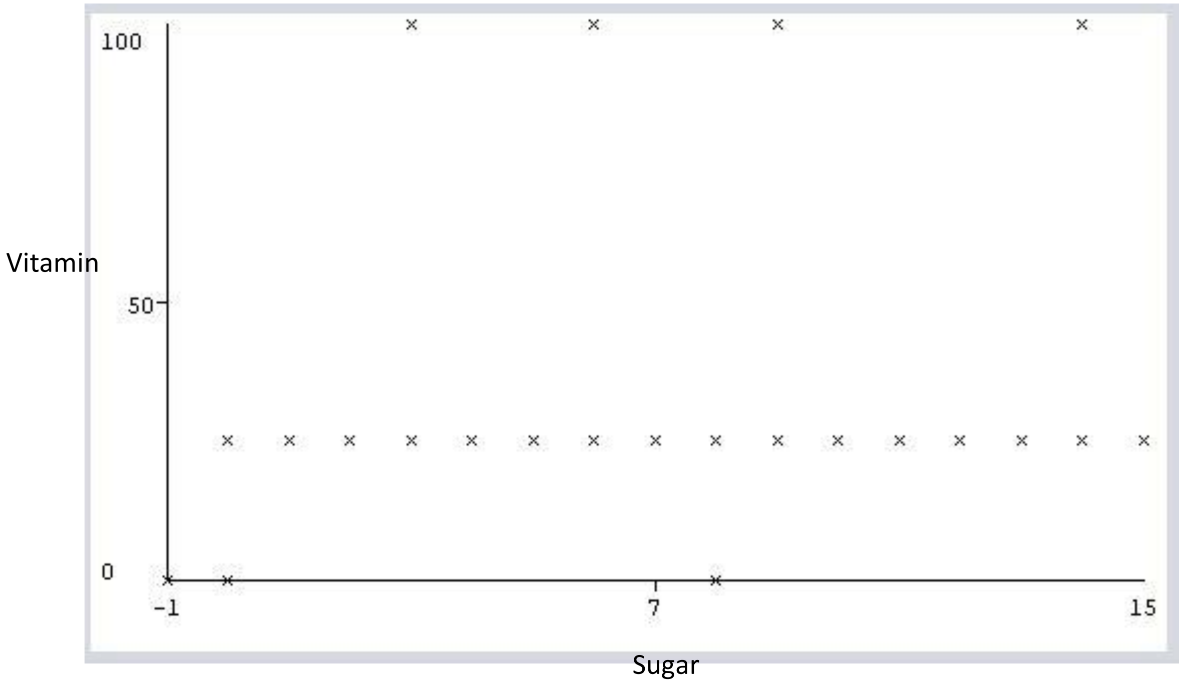
Sodium_Vitamin correlation – From data we calculated correlation coefficient r= 0.361. So we can say that Sodium and Vitamin are positively correlated.



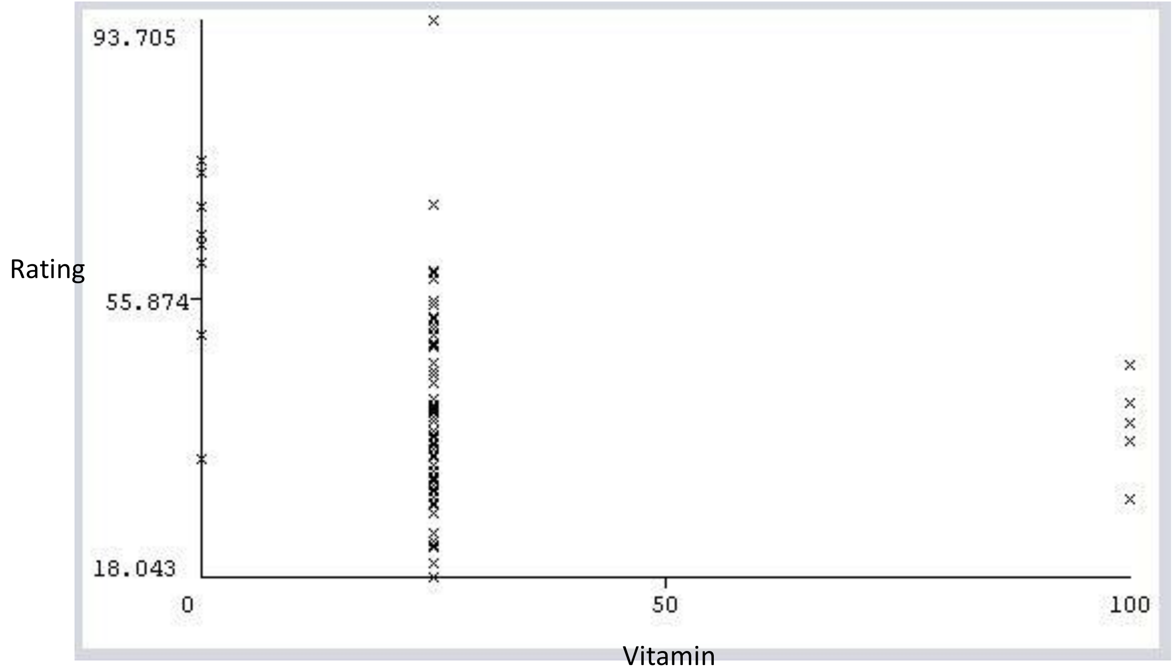
Sugar_Potassium correlation – From data we calculated correlation coefficient r= 0.022 which is closer to zero. So we can say that Sugar and Potassium are not correlated.



Sugar_Rating correlation – From data we calculated correlation coefficient r= -0.759. So we can say that Sugar and Rating are strong and negatively correlated. High in sugar cereals are less popular.



Sugar_Vitamin correlation – From data we calculated correlation coefficient r= 0.125. So we can say that Sugar and Vitamin are positively correlated.



Vitamin_Rating correlation – From data we calculated correlation coefficient r= -0.241. So we can say that Vitamin and Rating are negatively correlated. People do not really look for vitamins in cereals.