Stat 445 Report

Executive Summary:

Given the dataset for 77 breakfast cereals from 7 manufacturers, we plan to analyze the cereals with interesting research questions. Before we begin, we include a background information sheet for those who may be familiar with the terms like calories, vitamins, protein, and fibers, but couldn't really tell the difference between them. It is essential to understand some background materials before conducting data analysis. Without background information, variables might get mixed up, with the resulting analysis being less useful.

The study addresses the following four research questions:

1) Does calories = 4 protein + 4 (carbohydrates + sugars) + 9 fat?

With the nutritional formula for calculating food calories, we would like to answer if the dataset sufficiently fit this formula. The conclusion is that calories do fit the formula while there may be some questioning rounding errors from the manufacturers.

2) How do we choose a healthy, nutritious cereal?

First, we would like to understand how ratings are related to each variable by performing a multiple regression. Next, we choose cereals that are both low in fat and high in fiber and try to compare them to the population. We also concluded that Nabisco is geared towards healthy minded consumers.

3) Describe the relationship between display shelf and the cereals.

We use some graphics to show the differences between the 3 shelves. Noting that fiber is an indicator of healthy cereals, using boxplots, we discovered that shelf 3 contains the highest level of fiber. Doing a regression with dummy variables, we find the difference between cereals placed on shelf 3 and those placed on shelf 1 and 2 are significant.

4) Are there some cereals that are essentially the same?

We represent the dataset in multidimensional, and try to visually distinguish between similar features for our cereals.

Background Information:

Protein is essential for building and repairing muscles, red blood cells, hairs and other tissues. Protein is a source of calories and can be used for energy if inadequate levels of carbohydrates are available.

Carbohydrates are a source of calories from sugars and starches that fuel your muscles and brain. The liver breaks down carbohydrates into glucose (blood sugar), which is used for energy by the body. Carbohydrates are classified as simple or complex. Simple carbohydrates refer to sugars, while complex carbohydrates refer to starches.

Fat is a source of stored energy (calories) that is burned mostly during low-level activities (e.g., reading and sleeping) and long-term activity (e.g., long distance running). When energy from food can't be used by our body, it may be stored as fat.

Calories are units of energy. People tend to associate calories with food, but they can be applied to anything that contains energy. To replenish energy, human beings need to eat. Specifically, calories measure the potential energy for each food item. Now, food calories come from proteins, carbohydrates and fats. *A gram of carbohydrate contains 4 calories. A gram of protein also contains 4 calories. A gram of fat contains 9 calories.* Therefore, given information on how many proteins, carbohydrates and fats are in each food item, we can calculate the amount of calories that item contains.

For example, if we look at the nutrition level for 100% Bran located on the first row of the dataset, we find that it contains 1 gram of fat, 4 grams of protein and 11 grams of carbohydrates and sugars. After multiplying and summing, the total is (9+16+44) 69 calories. We see that calories provided, 70 calories, is rounded by the manufacturer.

Vitamins are metabolic catalysts that regulate chemical reactions within the body. Most vitamins are chemical substances that the body does not manufacture, so we must obtain them through diet. Vitamins are not a source of energy.

Minerals are elements obtained from foods that help regulate body processes. Minerals such as iron, magnesium, **sodium** and **potassium** do not provide energy.

Fiber includes the part of plant cells that humans can't digest. High-fiber foods such as wheat bran contain mainly insoluble fiber. This type of fiber passes through the digestive tract more or less intact, and it helps keep digestion running smoothly. Fibers do not contribute to calories, since it cannot be digested.

References:

Howstuffworks, "How Calories Work", http://health.howstuffworks.com/calorie.htm

Sports Nutrition Guidebook, 2ed Nancy Clark

Research Question **#1**:

In our dataset, does calories = 4 protein + 4 (carbohydrates + sugars) + 9 fat? Since there are many variables, we want to figure out as many relationships as possible before carrying forward.

To answer, we can try do a multiple regression for calories by formulating that calories = B0 + B1(protein) + B2(carbo) + B3(sugars) + B4(fat). Now, if the coefficients B0, B1, B2, B3, and B4 end up as 0, 4, 4, 4, 9 respectively, we can conclude that calories do follow the formula. Notice that 58th cereal in our dataset has both carbo and sugars missing, so we omitted it in the regression.

Table 1:

Call: lm(formula = calories ~ protein + carbo + sugars + fat)				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.04175	3.9273	0.011	0.992
protein	4.15389	0.59962	6.928	1.59E-09
carbo	4.06656	0.17194	23.651	< 2e-16
sugars	3.94232	0.16534	23.844	< 2e-16
fat	8.59671	0.63562	13.525	< 2e-16

The results, in Table 1, shows that the coefficients are not the perfect whole numbers that we expected, but these estimates are extremely close. The p-values for protein, carbo, sugars and fat are significant, while the intercept is not significant, meaning that B0 is not significantly different from zero. When re-examining the calories in the dataset, it shows that all the values are rounded to the nearest 10th digit. This might explain the slight deviations from 0, 4, 4, 4, and 9.

Figure 1.

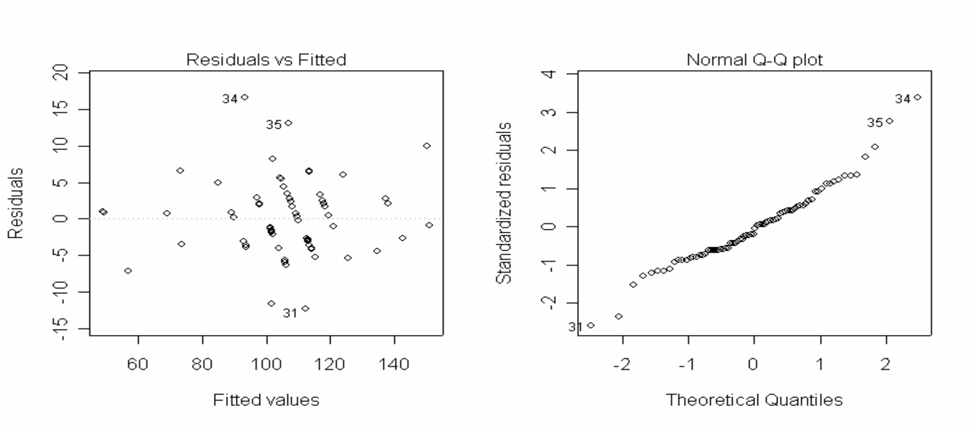


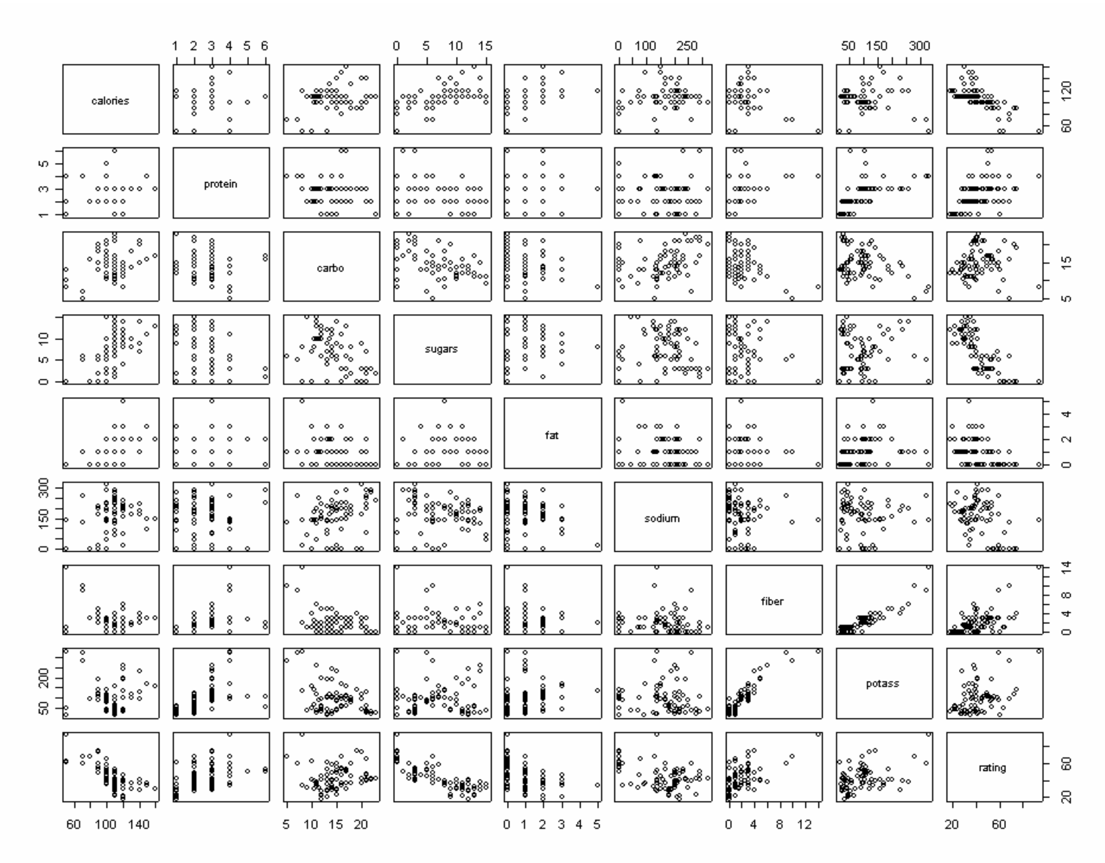
Fig. 1 helps us to check the assumptions for our regression. The residuals vs. fitted plot shows a randomly distributed scatter. The Q-Q plot shows a linear relation. Therefore, the diagnostics plots confirm the assumption of normally, independently distributed residuals, saying this is a good model.

We conclude that calories do follow: 4 proteins + 4 (carbohydrates + sugars) + 9 fats, with some rounding errors on the manufacturer's part. This relationship can save us a lot trouble when interpreting regression on the dataset. Since calories are linear combinations of protein, carbo, sugars and fat, any regression done to investigate the relationship between calories with any of its components would always show significant linear relationships.

Research Question #2: How do we choose a healthy, nutritious cereal?

First we would like to construct a model that predicts “consumer reports” rating. Ratings are an indication of nutritional content and understanding it can provide useful insights. Before doing any regression, it is best to draw a scatter plot matrix. Fig. 2 contains the scatter plot for the continuous variables in our dataset. Carefully viewing the dataset, we can immediately spot the negative correlation between calories and ratings (1st row, 9th col), saying that the more calories the cereal, the less the rating. There is also a strong correlation between fiber and potassium (7th row, 8th col).

Fig 2.



To construct a regression for ratings, a first approach would be to use calories as a predictor. Since we know calories is the sum of the parts: protein, carbo, sugar, fat, using these four as predictors would explain more variability than calories itself. If we add both predictor variables fiber and potassium to the model, the information would be redundant due to collinearity. After dropping carbo as the insignificant term by adding sodium, our regression model is shown in the following table.

Table 2:

Call: lm(formula = rating ~ protein + sugars + fat + fiber + sodium)					
Coefficients					
(Intercept)	**protein**	**sugars**	**fat**	**fiber**	**sodium**
58.151	1.978	-1.803	-3.898	2.422	-0.057

Sine rating measures how nutritious a cereal is, to gain insight from Table 2 , we infer that fat is twice as bad for you when comparing it to sugars. Also, fiber and protein both are healthy indicators of a cereal.

Let us turn our attention to cereals that are both low in fat and high in fiber. Summarizing both fat and fiber variable using stem and leaf plot, we get the following:

Cereals by Fat

```
 0 | 000000000000000000000000000
 1 | 000000000000000000000000000000
 2 | 00000000000000
 3 | 00000
 4 |
 5 | 0
```

Cereals by Fibers

```
 0 | 00000000000000000000000000000000000555
 2 | 000000000057000000000000000
 4 | 00000000
 6 | 0
 8 | 0
10 | 0
12 |
14 | 0
```

The stem and leaf plot for both variables are in single digit. Each digit in the plot represents a single observation. Both distributions are positively skewed, saying that a lot of cereals contain 0 fat and/or 0 fibers. The median for fat = 1, for fiber = 2. We choose our nutritious cereals as those having 0 fat and fiber greater than 2.

Table 3: subset of data [fat==0 & fiber>=3,]

	name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	shelf	rating
4	All-Bran_with_Extra_Fiber	K	50	4	0	140	14	8	0	3	94
10	Bran_Flakes	P	90	3	0	210	5	13	5	3	53
27	Frosted_Mini-Wheats	K	100	3	0	0	3	14	7	2	58
29	Fruitful_Bran	K	120	3	0	240	5	14	12	3	41
34	Grape-Nuts	P	110	3	0	170	3	17	3	3	53
51	Nutri-grain_Wheat	K	90	3	0	170	3	18	2	3	60
64	Shredded_Wheat	N	80	2	0	0	3	16	0	1	68
65	Shredded_Wheat_'n'Bran	N	90	3	0	0	4	19	0	1	74
66	Shredded_Wheat_spoon_size	N	90	3	0	0	3	20	0	1	73
69	Strawberry_Fruit_Wheats	N	90	2	0	15	3	15	5	2	59

Table 3 contains 10 cereals that are considered nutritious. The mean rating for this table is 63.3, while the mean rating for the entire population is 42.6. This huge increase in rating indicates this group being more significantly nutritious.

Using this subset of healthy cereals, another question arises: Can we tell which manufacturers specialize in healthy cereals? By looking at the healthy cereals in Table 3, we see that Nabisco, Kellogg's, and Post are major players. But when comparing to the total cereals seen in Table 4., Nabisco has 4 out of 6, 67% of their total product aiming to be healthy cereal. We can conclude that Nabisco is a niche product.

Table 4.

mfr (total population)						
American H	**General M**	**Kellogg's**	**Nabisco**	**Post**	**Quaker**	**Ralston**
1	22	23	6	9	8	8
mfr (healthy subset)		17%	67%	22%		
American H	**General M**	**Kellogg's**	**Nabisco**	**Post**	**Quaker**	**Ralston**
0	0	4	4	2	0	0

Fig 3.

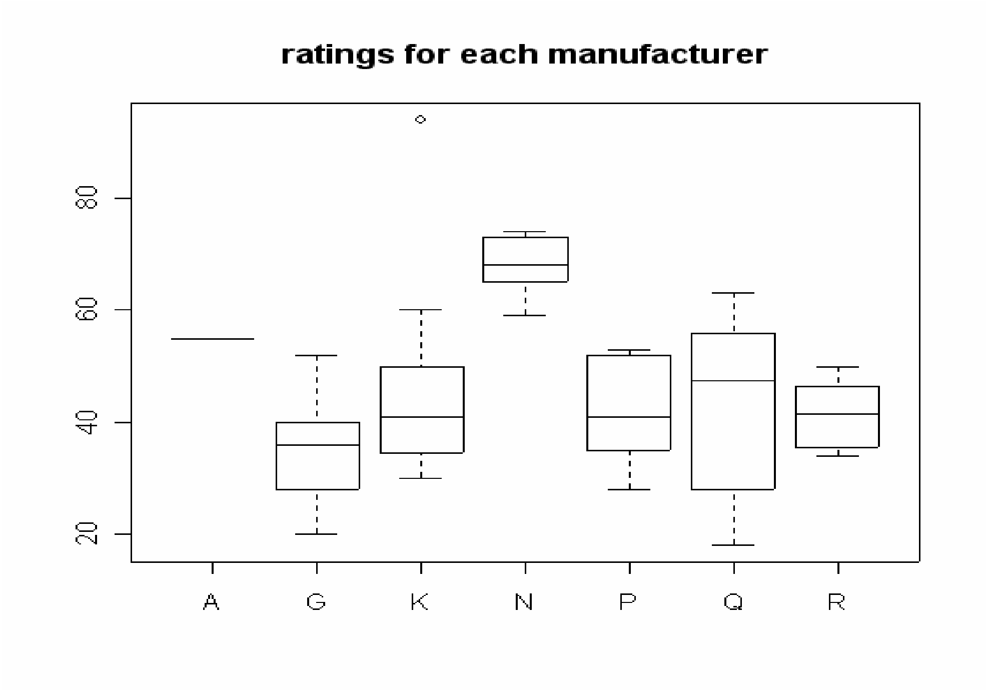


Fig 3. is the boxplot for ratings according to each manufacturer for 77 cereals, it also supports our claim that Nabisco is highly specialized toward healthy minded consumers.

<u>Research Question #3</u>: Describe the relationship between display shelf and the cereals

Focusing on the shelf variable in the dataset, we can ask questions such as, "What kinds of cereals are located on what shelves". In order to describe any meaningful relationship, it's always useful present graphical plots.

Fig. 4

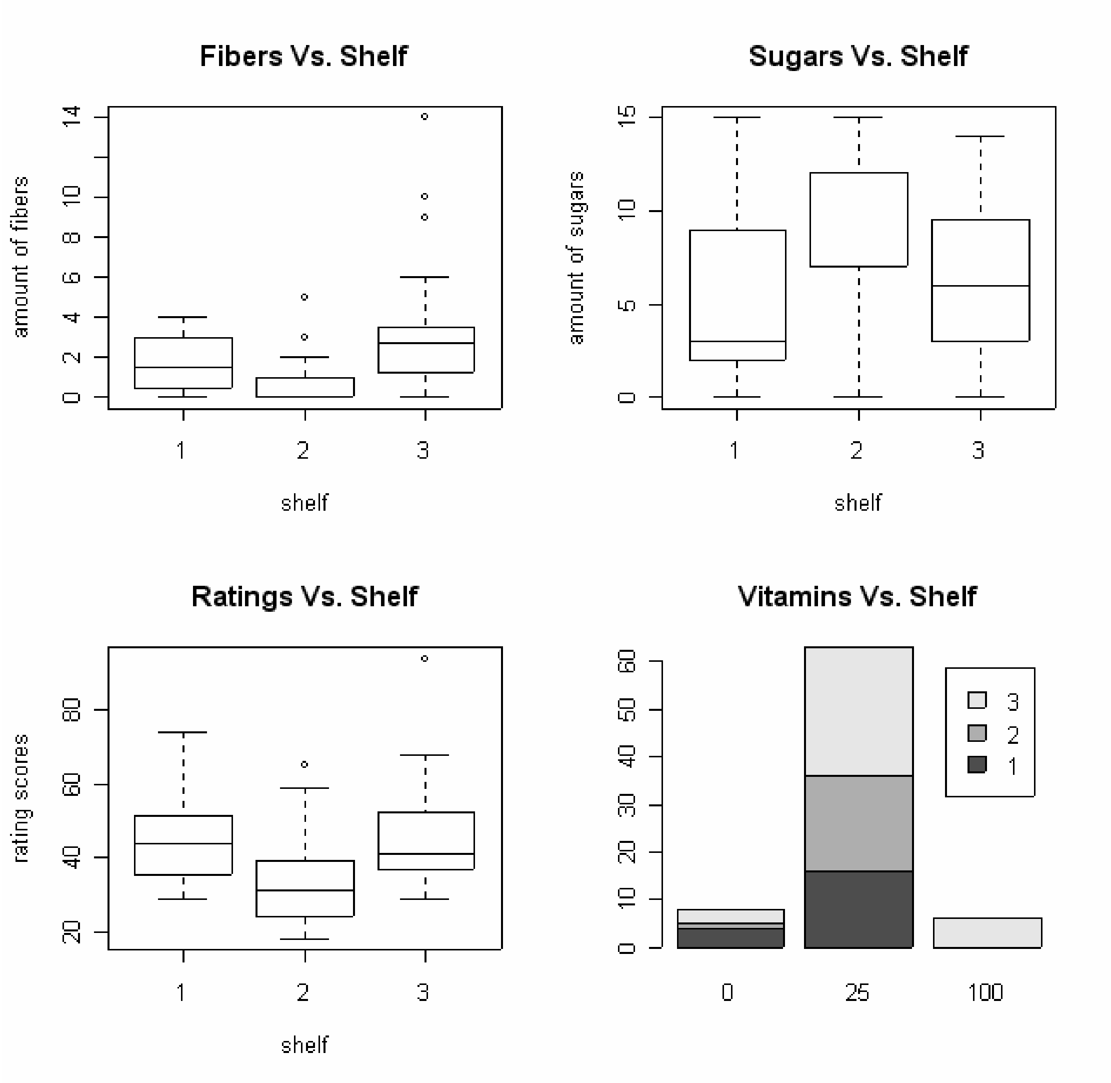


Figure 4 shows cereals on shelf 3 have the highest fiber, while cereals on shelf 2 contain the most sugars. It is also interesting to see that the overall rating is best for shelf 1. The last barplot in fig. 4 is a plot of 2 factors, since there are 3 levels of vitamins, 0, 25, 100. We note that the majority of cereals are 25% vitamins enriched, and distributed equally. There are so some cereals with no vitamins that can also be found on each shelf. However, if we want to find cereals that are 100% enriched in vitamins, we can only find it on shelf 3.

From the above, we may conclude, cereals placed on shelf 3 are aimed at healthy consumers. Since we cannot tell multiple values apart from our boxplots, we try another graphic by breaking the data into 3 distinct shelves as follows in Figure 5.

Fig 5.

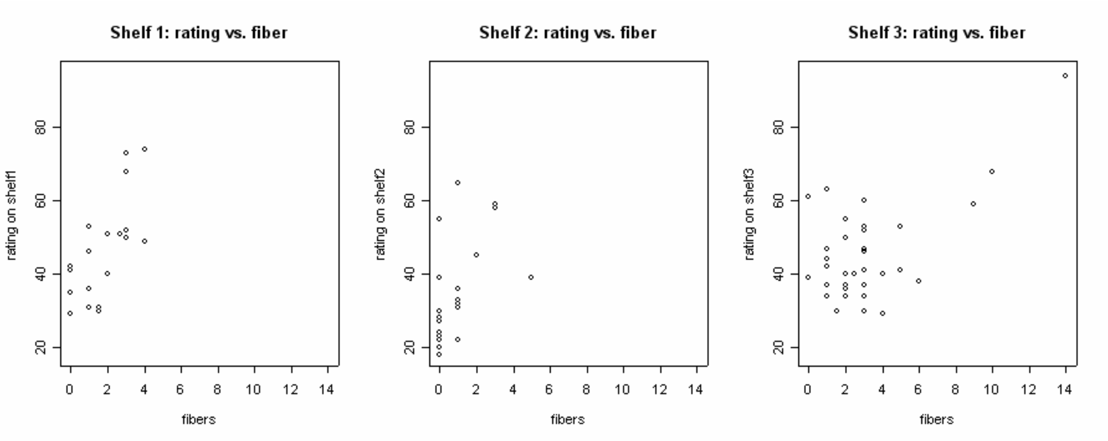


Figure 5 shows us that there are definitely more linear relationship between fibers and ratings for shelf 3 only. It is also evident, that on shelf 3, fibers have more variability per cereal than on shelves 1 and 2. Notice some observations are squeezed together in shelf 2 and shelf 1, transformation on the fibers variable may help spread out the values.

A statistical question would be to treat the shelf as a dummy variable and see if there is significant difference between average fibers by shelf, refer back to first graph in Fig. 4. Since we want to see if the differences between shelves are significant, we set Z1 as the indicator for shelf 1, and Z2 as the indicator for shelf2. Regress fiber rating by B0 + B1 (Z1) + B2 (Z2), we get the following:

Table 5:

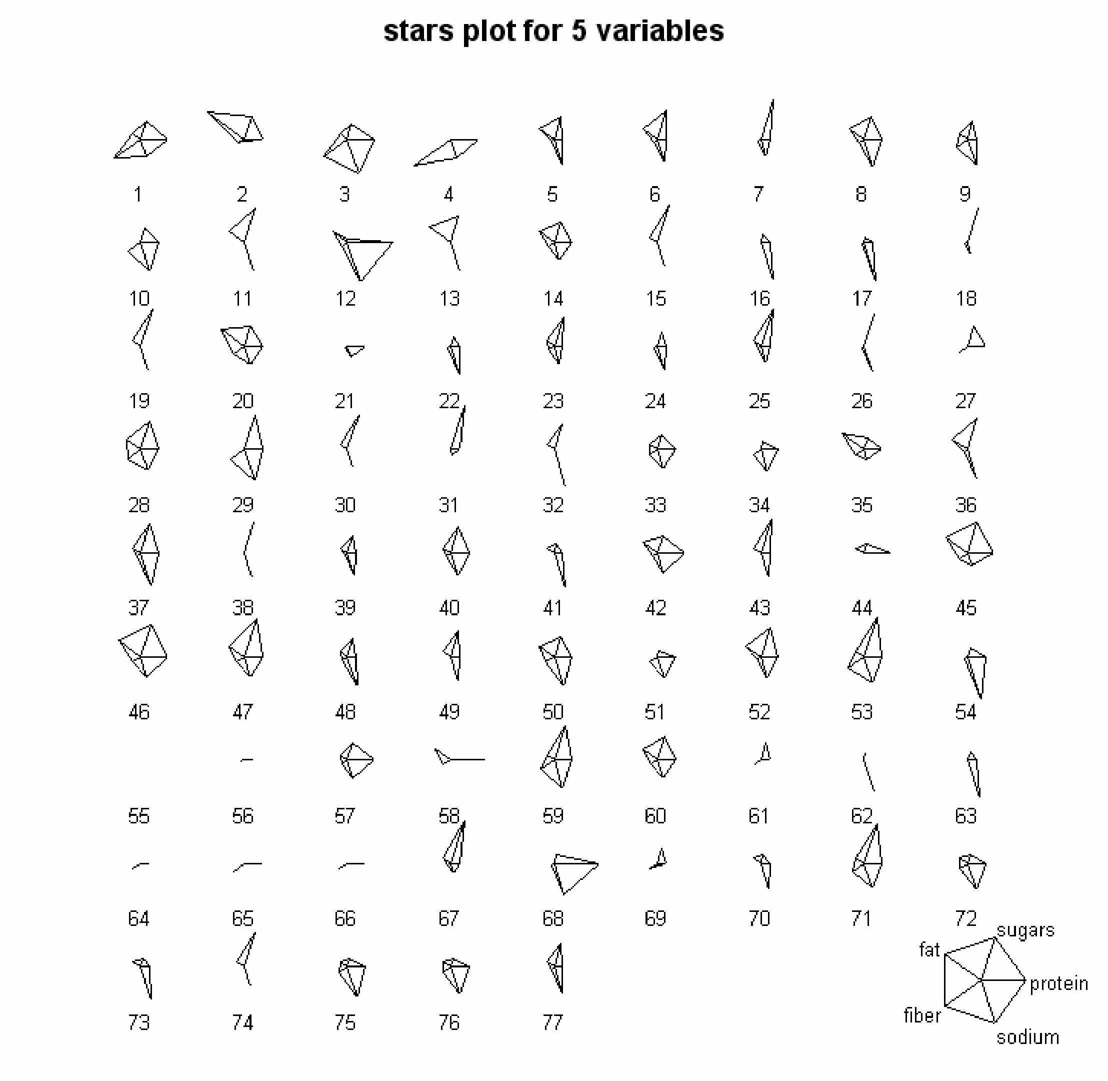
Call: lm(formula = fiber ~ Z1 + Z2)				
	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	3.1389	0.3674	8.543	1.22E-12
Z1	-1.4539	0.6148	-2.365	0.020663
Z2	-2.2341	0.6053	-3.691	0.000425

By looking at the p-values in the last column of Table 5, all less than alpha = 0.05, meaning they are significantly different from the null hypotheses of coefficients being zero. Since our base group is shelf 3, the intercept, B0 indicates the mean fiber level of 3.13. We can conclude the difference between shelf 2 and shelf 3, shelf 1 and shelf 3 are significant. This may support our assumption that cereals placed on shelf 3 are generally healthier, since they have more fiber.

<u>Research Question #4:</u> Are there some cereals that are essentially the same?

If we can somehow represent variables in a multi-dimensional way, we can visually describe and group these observations. We can use the stars command in R. Let us choose the 5 variables used in the regression in *research question #2*, they are protein, sugars, fat, fiber and sodium

Fig. 7



There are a lot of possibilities for grouping. For example, we can see 75 and 76 as being in similar shape, implying similar cereal. Also, we can group 8, 50, and 52 together as a radom sample:

	name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	vitamins	shelf	rating
8	Basic_4	G	130	3	2	210	2	18	8	100	25	3	37
50	Nutri-Grain_Almond-Raisin	K	140	3	2	220	3	21	7	130	25	3	41
52	Oatmeal_Raisin_Crisp	G	130	3	2	170	1.5	14	10	120	25	3	30

A closer look shows these three are pretty similar in every category. The drawback for stars plot is that too many dimensions would result in blurry perception, and the plot wouldn't be useful anymore. Stars plot can help us discover whether cereals having similar attributes would have the same ratings. Due to page restrictions, grouping the stars plot is left as a fun exercise for readers.