EXECUTIVE SUMMARY

Every morning, almost half of Americans start the day with a bowl of cereal, but choosing the right healthy breakfast is not always easy. *Consumer Reports* rating is therefore calculated by an undisclosed formula based on the nutritional content. In this study, we are going to explore all the explanatory variables that have relationships with this rating and discover the formula by statistical analysis.

By performing both simple and multiple linear regression of the rating on each of the variables, it is found that there is a linear relationship between the rating and all the 9 nutritional content including calories per serving, sugars, protein, fat, sodium, dietary fiber, complex carbohydrates, potassium, and vitamins. From further analysis, a high-fiber, high-protein, low-fat, and low-sugar cereal would make a nutritious breakfast.

Although the *Consumer Reports* rating is directly calculated from nutritional content, it is interesting to find that the rating also depends on some non-nutritional information such as manufacturer and the serving size in cups and ounces, but not the supermarket display shelf location. The manufacturer *Nabisco* produces cereals with statistically higher ratings than others. This arises as there is also a linear relationship between manufacturer and the highly-correlated nutritional variables like dietary fiber, sugars, and fat.

DATA AND PROBLEM

The dataset contains per-serving nutritional information and grocery shelf location for 77 breakfast cereals from seven manufacturers. *Consumer Reports* rating (*rating*) is included in the dataset, which is calculated by an undisclosed formula presumably based on the nutritional content. There are 7 nutrients in total, which include protein (*protein*), fat (*fat*), sodium (*sodium*), dietary fiber (*fiber*), complex carbohydrates (*carbo*), potassium (*potass*), and vitamins (*vitamins*). Calories per serving (*calories*) and sugars (*sugars*) are also considered in the nutritional content. Other information we have are the manufacturer (*mfr*), supermarket display shelf location (*shelf*), type of cereal whether it is hot or cold (*type*), recommended serving size in ounces (*weight*), and recommended serving size in cups (*cups*).

As *Consumer Reports* rating is computed by an undisclosed formula based on the nutritional content, we would like to discover the formula. In other words, we would like to investigate which nutrients are essential for a nutritious breakfast cereal. Although we know that *Consumer Reports* rating is not calculated by non-nutritional information, we are interested in examining the relationship between the rating with other non-nutritional variables such as manufacturer, grocery display location, and recommended serving size.

Before performing any statistical analysis, it is noted that three cereals, *Almond Delight*, *Cream of Wheat (Quick)*, and *Quaker Oatmeal*, have missing data on the variables *potass, carbo,* and *sugars*; therefore, they are excluded in the analysis. From the remaining 74 cereals, all are served as cold, except the only one cereal, *Maypo*, so it may be a possible influential observation.

EXPLORATORY DATA ANALYSIS

As the *Consumer Reports* rating is calculated by using a formula based on the nutritional content, we first examine the relationship between the response variable *rating* and the 9 explanatory variables: *calories, sugars*, and 7 nutrients including *protein, fat, sodium, fiber, carbo, potass,* and *vitamins*.

Scatter-plots (Diagram 1) are drawn to visualize the association between *rating* and each of the explanatory variables individually. The least squares line is drawn on each plot as well. Since the data of the three variables, *protein, fat,* and *vitamin*s, are discrete with several distinct values, box-plots (Diagram 1) are more relevant to represent the relationship between *rating* and these variables. These box-plots are drawn with widths proportional to the square roots of the number of observations in the groups. A notch is also drawn in each side of the boxes. If the notches of two plots do not overlap, there is strong evidence that the two medians differ. All the six scatter-plots and the three box-plots are graphed geometrically in Diagram 1.

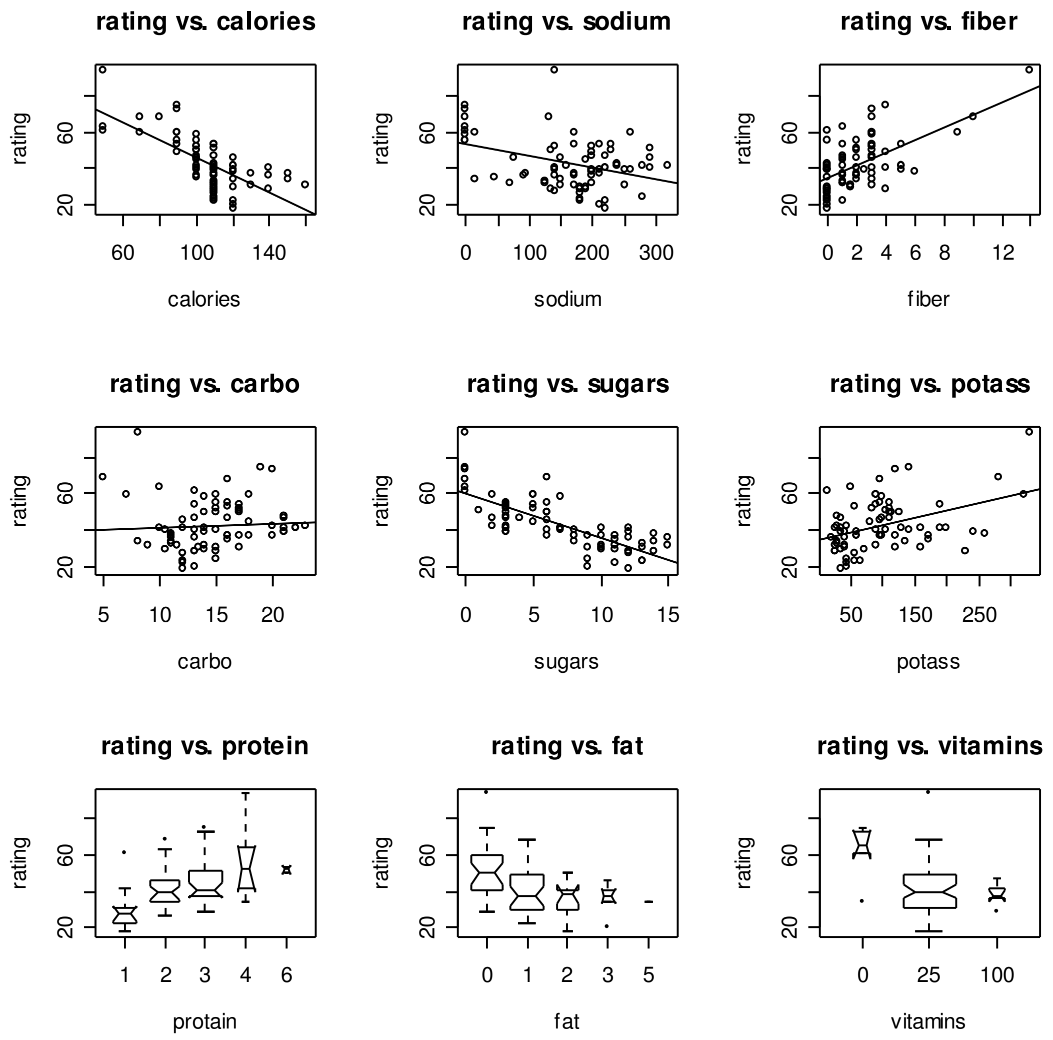

Diagram 1 – Plots between *Consumer Reports* Rating and the 9 Explanatory Variables.

From Diagram 1, it is clear that the three variables, *fiber*, *potass,* and *protein* are positively correlated with *rating*. It means that *rating* increases as these variables increase. *Fiber* has the most positive correlation with a value of 0.60. It implies that *rating* seems to depend strongly on *fiber*. On the other hand, the variables *carlories, sodium, sugars, fat,* and *vitamins* are all negatively correlated with *rating*. To have a high rating, these variables have to be as low as possible. Among these, *sugars* have the most negative correlation of -0.76 while *calories* also have a high correlation of -0.69. It shows that *rating* is dependent strongly on both *sugars* and *calories*. However, the correlation between *rating* and *carbo* is only 0.056, which is very close to 0. The rating seems to depend very weakly on carbohydrates. Diagram 1 also suggests that some outliers or influential points are present in the data, which may require special attention in building a sound model.

Simple Linear Regression

After the examination of these plots, it is reasonable to fit a simple linear regression of *rating* on each of the 9 explanatory variables. All the correlations, the R^2 values (proportion of variability explained by the model), the regression coefficients, and the p-values of testing the significance of the coefficients are summarized in Table 1. Note that it is sorted in an ascending order of R^2 and a descending order of p-values.

Table 1 – Simple Linear Regression of Rating on each of the 9 Explanatory Variables

Explanatory Variables	**Correlation**	**p-value**	**R^2**	**Model**
carbo	0.056	0.636	0.003129	39.40 + 0.20 carbo
vitamins	-0.21	0.0665	0.046	46.29 - 0.14 vitamins
sodium	-0.38	0.000757	0.1467	52.92 - 0.06 sodium
fat	-0.41	0.000344	0.1641	48.02 - 5.65 fat
potass	0.42	0.000230	0.1729	34.26 + 0.08 potass
protein	0.47	2.72e-05	0.2182	27.05 + 6.09 protein
fiber	0.60	1.27e-08	0.3641	34.77 + 3.49 fiber
calories	-0.69	7.25e-12	0.4813	94.88 - 0.49 calories
sugars	-0.76	6.92e-15	0.5715	59.67 - 2.43 sugars

Fitting simple linear regression confirms that the correlation between *rating* and *sugars* is the strongest. Its model gives the largest R^2 and the smallest p-values to reject that the regression coefficients are zero. Table 1 also provides evidences that the regression coefficients are all statistically significant in the model regressing *rating* with all the other variables except *carbo*. Fitting *rating* on *carbo* gives a p-value of only 0.636, which is even larger than 0.1. Again, it confirms the week linear relationship between *rating* and *carbo.*

Multiple Linear Regression

We are now fitting a multiple linear regression of *rating* on all the 9 explanatory variables: *rating ~ calories + protein + fat + sodium + fiber + carbo + sugars + potass + vitamins.* We first examine the plausibility of the basic assumptions of this multiple linear regression model by some diagnostic plots.

To examine the assumption of linearity and the pattern of dispersion, we plot the residuals vales versus the fitted values in Diagram 2. The scale-location plot (Diagram 2) provides the plot of the square root of the standardized residuals versus the fitted values. Both plots show no curvature, indicating the assumption of linearity holds. The pattern of dispersion about the x-axis is also uniform, which confirms the assumption of homogeneous dispersion. There are a few points of leverages as shown in the plots. The cereals 16 *(Corn Flakes)*, 44 (*Muesli Raisins, Peaches, & Pecans)*, and 67 (*Total Corn Flakes)* are possible influential values. Note that the observation number is not matched with the name in the original dataset since the missing data is removed before analysis.

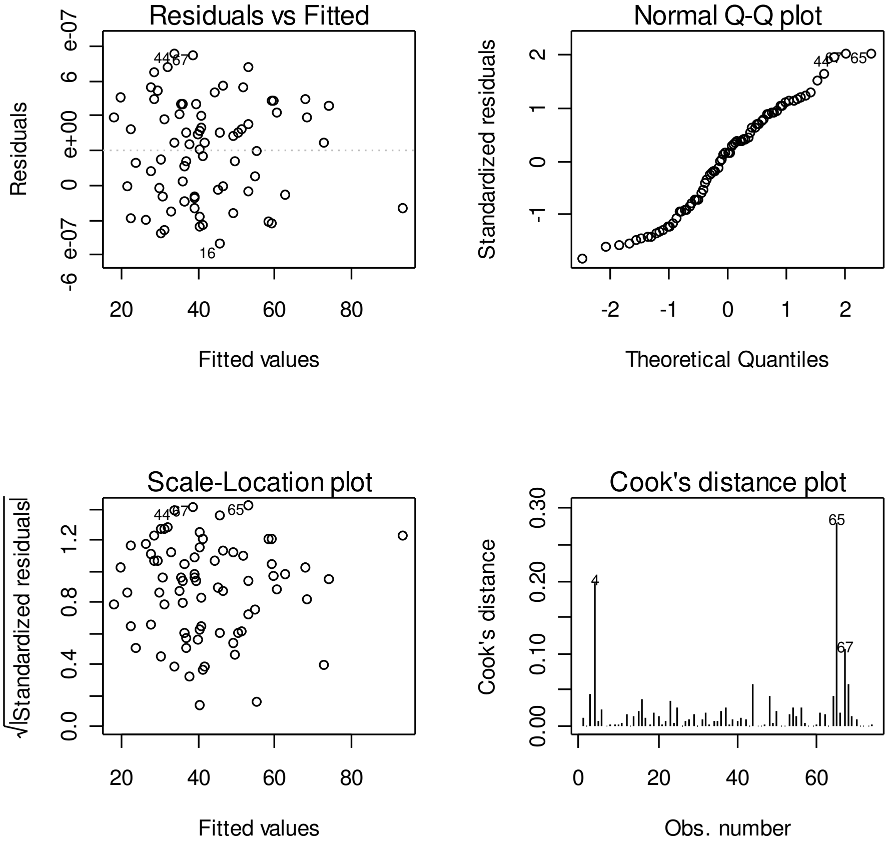

Diagram 2 – Diagnostic Plots of Multiple Linear Regression of Rating on 9 Variables

To examine normality, normal quantile plot is graphed above. The linear shape of the Q-Q plot implies that the assumption of normality holds. It also highlights that the cereals 44 (*Muesli Raisins, Peaches, & Pecans)*, 65 (*Special K*), and 67 (*Total Corn Flakes)* are deviant observations.

From the Cook's distance plot, it is clear that cereal 65 (*Special K*) is the most influential data with the largest Cook's distance, but it is not close to 1 at all. Cereals 4 (*All-Bran with Extra Fiber*) and 67 (*Total Corn Flakes)* are also possible influential observations. Note that *All-Bran with Extra Fiber* is the one with the highest rating. Removing these observations do not make a difference as their Cook's distance is much smaller than 1.

After examining model assumptions, we are ready to have the analysis of variance of this multiple regression model:

```
lm(formula = rating ~ calories + protein + fat + sodium + fiber +
    carbo + sugars + potass + vitamins)

Residuals:
       Min         1Q     Median         3Q        Max
-5.343e-07 -2.537e-07  3.961e-08  2.424e-07  5.513e-07

Coefficients:
              Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.493e+01  2.794e-07  196559702   <2e-16 ***
calories    -2.227e-01  7.501e-09  -29694282   <2e-16 ***
protein      3.273e+00  5.551e-08   58964906   <2e-16 ***
fat         -1.691e+00  8.101e-08  -20877762   <2e-16 ***
sodium      -5.449e-02  4.910e-10 -110974232   <2e-16 ***
fiber        3.443e+00  4.756e-08   72399805   <2e-16 ***
carbo        1.092e+00  3.492e-08   31287364   <2e-16 ***
sugars      -7.249e-01  3.311e-08  -21895192   <2e-16 ***
potass      -3.399e-02  1.601e-09  -21228850   <2e-16 ***
vitamins    -5.121e-02  1.779e-09  -28778552   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 3.069e-07 on 64 degrees of freedom
Multiple R-Squared:      1,      Adjusted R-squared:      1
F-statistic: 1.696e+16 on 9 and 64 DF,  p-value: < 2.2e-16
```

This model gives a perfect fit with R^2 strictly equals to 1. The p-values of t-tests on each variable are all strictly less than 2.2e-16, which implies that all the 9 variables are statistically significant at the level 0.001. Although the variable *carbo* is individually non-significant in simple linear regression model, it is significant when it is combined with other variables in the multiple linear regression model. As $R^2 = 1$, 100% of variability is explained by this model. Further transformation or reduction of any variables will not improve the model at all. Therefore, this model is the final model:

rating = 54.93 -0.22 calories + 3.27 protein -1.69 fat -0.054 sodium + 3.44 fiber + 1.09 carbo -0.72 sugars -0.034 potass -0.051 vitamins

This is also the formula of how *Consumer Reports* rating is computed. The result is consistent with the scatter-plots and box-plots in Diagram 1. The rating increases when the nutrients *fiber* and *protein* increase. Decreasing *fat* and *sugars* will also increase the *rating*. Other variables have less effect on *rating* as they have coefficients close to 0.

INVESTIGATION

Although *Consumer Reports* rating is not calculated by non-nutritional information, we are interested in examining the relationship between *rating* and other non-nutritional variables including manufacturer, supermarket display shelf location, and serving size in cups and ounces. The relationship between *rating* and type of cereal is not examined as there is only one hot cereal in the dataset. Similar to Diagram 1, the scatter-plots and box-plots of their relationships are graphed in Diagram 3 on the next page.

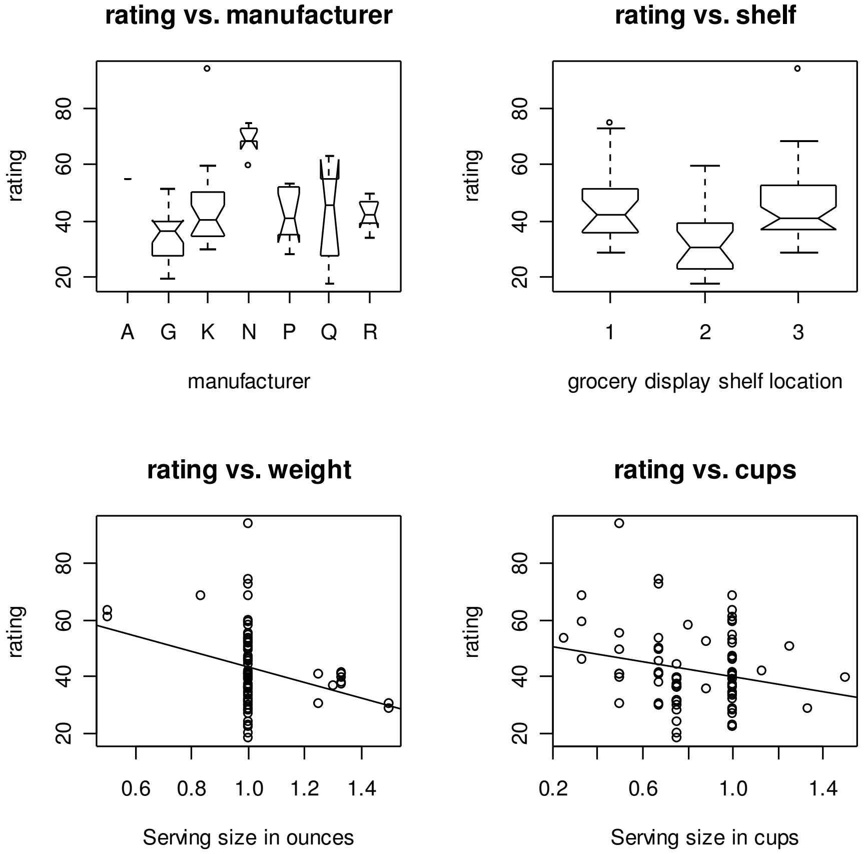

Diagram 3 - Plots between *Consumer Reports* Rating and the 4 Non-nutritional Variables

From the above plots, *rating* seems to be dependent on manufacturers. *Nabisco* tends to produce cereals with higher ratings than the other 5 manufacturers, *General Mills, Kelloggs, Post, Quaker Oats,* and *Ralston Purina.* The mean rating produced from *General Mills* is the lowest. Note that there is only one cereal produced by *American Home Food Products* in the dataset.

The second box-plots suggest that there seems to be non-linear relationships between *rating* and *shelf.* The other two variables *cups* and *weight* are negatively correlated with *rating*, but the plots are not roughly elliptical. Logarithmic transformation of these two non-nutritional variables may give better fits. By fitting simple linear regression on each variable, Table 2 summarizes all the correlations, the R^2 values, and the p-values of F-test. Again, it is sorted in the order of R^2 and p-values.

Table 2 – Simple Linear Regression of Rating on each of the 4 Non-nutritional Variables

Non-nutritional Variables	**Correlation**	**p-value**	**R^2**
Supermarket display shelf *(shelf)*	0.0510	0.666	0.002605
Serving size in cups *(cups)*	-0.223	0.0567	0.04951
Serving size in ounces *(weight)*	-0.300	0.0093	0.09028
Manufacturer *(mfr)*	N/A	4.276e-05	0.3512

The variable manufacturer is individually significant as its p-value is 4.276e-05 which is less than 0.001. The variables *cups* and *weight* are statistically significant at the level of 0.1 and 0.01 respectively. However, *shelf* is non-significant as it has large p-value. If we take logarithms of *cups* and *weight,* the situation is slightly improved:

Table 3 – Simple Linear Regression of Rating on Log Transformation of Variables

Non-nutritional Variables	**Correlation**	**p-value**	**R^2**
Log *(cups)*	-0.262	0.0242	0.06854,
Log*(weight)*	-0.313	0.00655	0.09821

Fitting multiple linear regression of *rating* on these 4 variables generates a similar result. The model *rating ~ mfr + shelf + log(weigh) + log(cups)* gives a R^2 of 0.5069. Manufacturer is significant at the level of 0.001 while log(*cups)* and log(*weight)* are statistically significant at the level of 0.05 and 0.001 respectively. The variable *shelf* is non-significant at any level.

This provides evidences that although the rating is not calculated from non-nutritional information, it is dependent on manufacturer and the serving size in cups and ounces. This may be due to a linear relationship between these non-nutritional variables and the 9 nutritional variables. We may further examine the association between manufacturer and the nutritional variables. Here, we choose to produce box-plots of manufacturer and 4 most highly-correlated nutritional variables in the previous analysis.

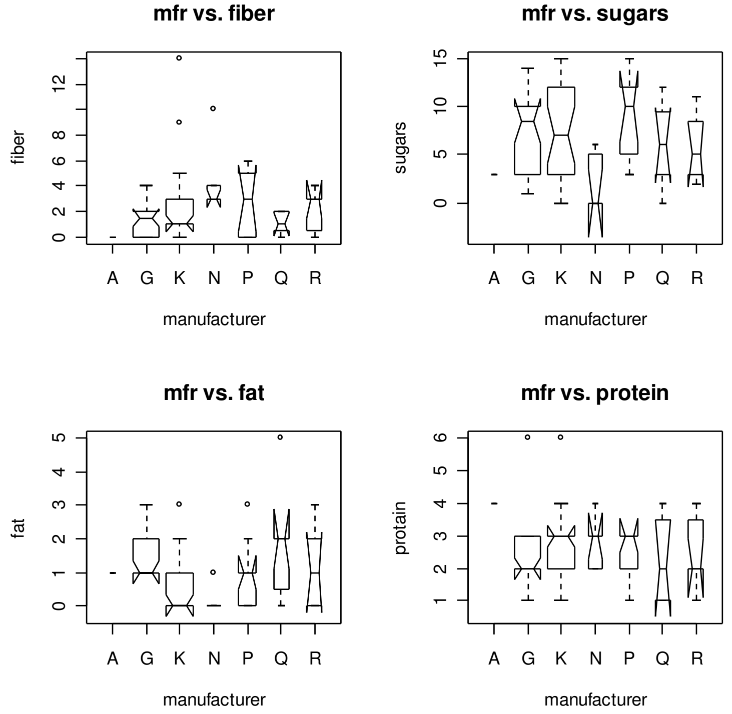

Diagram 4 – Box-plots between Manufacturer and 4 Nutritional Variables

The manufacturer *Nabisco* produces cereals with higher ratings than others. From the above four plots, it also gives significantly different results. *Nabisco* tends to produce high-fiber, low-fat, and low-sugar cereals as well. Therefore, the reason why *rating* depends on *mfr* is due to the linear relationship between *mfr* and the nutritional content.

CONCLUSIONS

Consumer Reports rating of breakfast cereals is calculated by using a formula based on 9 nutritional content including calories per serving, sugars, protein, fat, sodium, dietary fiber, complex carbohydrates, potassium, and vitamins. The multiple linear regression model suggests that there is a linear relationship between the rating and these 9 nutritional variables.

Among all these variables, the rating is highly and positively correlated with dietary fiber and protein, while it is highly and negatively correlated with fat and sugars. Other variables have effect on the rating as well but with weaker relationships. In other words, *Consumer Reports* give cereals high ratings for ones with high dietary fiber, high protein, low fat, and low sugars. It can be supported by the cereals with the highest and lowest ratings. For instance, *All-Bran with Extra Fibe*r, having the highest rating of 93.70, has the highest dietary fiber of 14.0 grams and high protein of 4.0 grams with neither fat nor sugars. On the other hand, *Cap'n'Crunch*, having the lowest rating of 18.04, has high sugars of 12.0 grams, high fat of 2.0 grams, and low protein of 1.0 gram with no dietary fiber at all.

From this 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. It is no wonder why most manufacturers nowadays would emphasis by adding the words “Low Fat” or “With Extra Fiber” to promote their “nutritious” products. Some brands with healthy sounding names, however, do not make the grade. For example, *100% Natural Bran* has the highest fat of 5.0 grams and high sugars of 8.0 grams with only 2.0 grams of dietary fiber, so it is given a low rating of 33.98. As we would not know the *Consumer Reports* rating of each breakfast cereal when we purchase in supermarket or grocery, it is a good idea to choose those with high fiber, high protein, low fat, and low sugar.

Although the *Consumer Reports* rating is not calculated directly from non-nutritional information, it is interesting to find that it also depends on manufacturer and serving size in cups and ounces. The manufacturer *Nabisco* produces cereals with significantly higher ratings while *General Mills* produces the ones with lower ratings. It is consistent with the previous conclusion that most *Nabisco*’s products are high-fiber with no fat and sugars. If the rating is unknown when we purchase cereals, choosing *Nabisco* might be a good choice for heavy breakfast.

Serving size is also another non-nutritional variable that correlates negatively with the rating. If more serving size is required for each breakfast, the rating becomes lower. It can be explained by most high-fiber cereals requires smaller serving size. The serving size in ounces is negatively correlated with fiber by -0.51. Furthermore, it makes sense that the rating does not depend on the grocery display shelf location, that is, cereals on the top floor do not mean that they are more nutritious!

There are some limitations in this study. There are only 9 nutritional variables available in the dataset. The formula of *Consumer Reports* rating may be dependent on other nutrients or other non-nutritional information such as price, total size, design of package, etc. The way the data was collected might affect the scope of our conclusions.