Table of Contents

Introduction:

"You are what you eat", is what a well-known saying advises us. Breakfast cereals are eaten in many households, perhaps due to the millions of advertisements touting healthiness, tastiness, kid-friendliness, and a host of other good qualities. We are interested in how the actual nutritional value of a cereal as printed on the box and marketing techniques influence consumer ratings of the product. There may also be other interesting trends involving different manufacturers, or types of cereals, or other qualities.

Executive Summary

We have a data set containing manufacturer name, nutritional information, consumer ratings, and the supermarket shelf location for 77 different cereals made by several different manufacturers, and want to determine a possible relationship between these.

The Exploratory Data Analysis section investigates a few possible associations between different variables, using correlation coefficients to decide which variables are most strongly related. Two interesting relationships are between rating and calories, and rating and sugar. We also observe some trends when we separate the data by manufacturer name.

The Confirmatory Data Analysis section goes into more detail, where we try to construct a suitable model that describes rating accurately, without being too complex. Our final model is

rating = 58.4 + 2.11*protein - 3.99*fat - 0.053*sodium + 2.40* fiber - 1.77*sugars - 0.048*vitamins

This expression tells us that rating increases with higher protein and fiber content in the cereal, and decreases with higher sodium, sugar content, and also vitamins. This model is questionable for several reasons, but the r-squared value is 0.99, which indicates that it explains 99% of the variation in rating! This makes it an acceptable formula for our sample data, although it should be tested on more breakfast cereals.

Exploratory Data Analysis

We are given a data set with entries for 77 different cereals detailing the name, the manufacturer *mfr*, the *type* (hot or cold), *calories* per serving, *protein* (grams), *fat* (grams), *sodium* (milligrams), dietary *fiber* (grams), complex *carbo*hydrates (grams), simple *sugars* (grams), *potass*ium (milligrams), the percentage of the FDA recommended daily amount of *vitamins* and minerals, recommended serving sizes in ounces *weigh*, the recommended serving sizes in *cups*, and the *rating* given to the cereals. There are only 4 missing values, so these will be ignored for now. We begin with a summary of some of the more interesting variables using histograms:

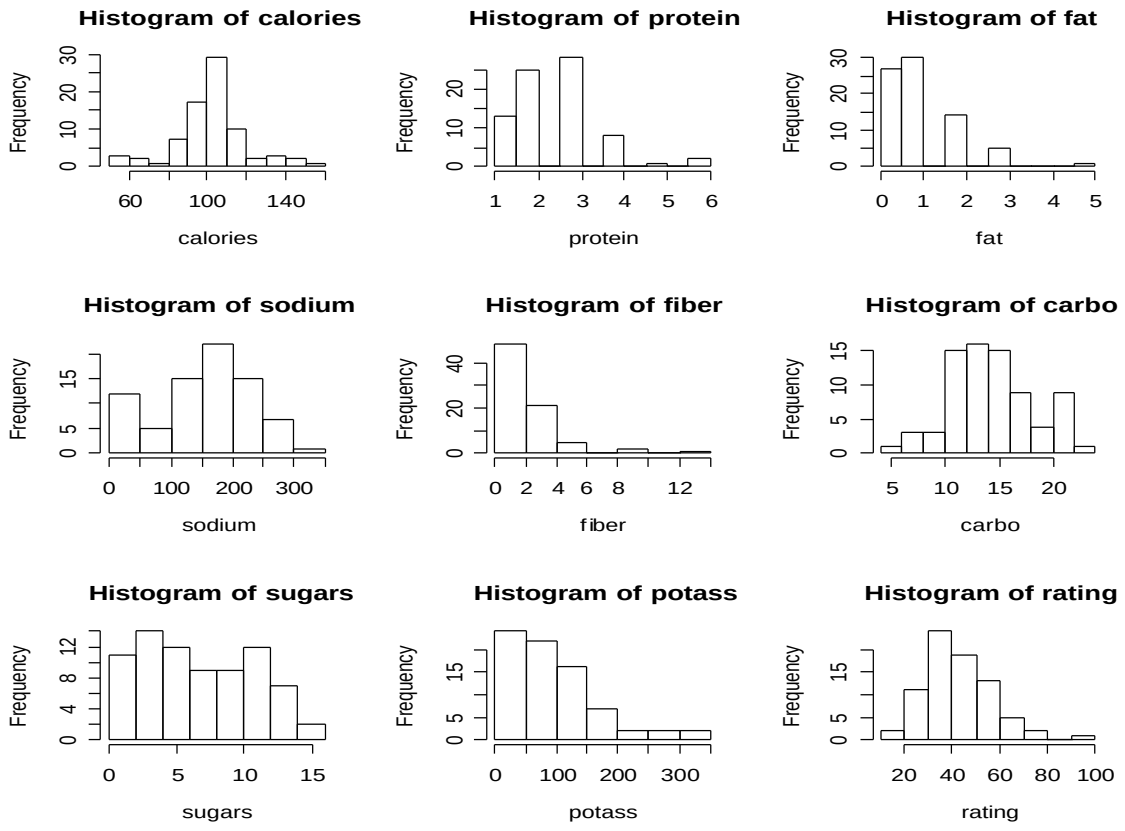


Figure 1.

These are just for the reader to obtain a better sense for the data, and cannot really tell us what is going on. But it is interesting to observe the way some variables such as fiber are skewed (the majority of the cereals in the sample have low fiber content) and some are more uniformly distributed, such as sugars. We may also suspect outliers in some variables, such as the right-most bar in the histogram of fat.

The largest sample correlation found is actually between potassium and fiber (0.91) – but this is a detail related to cereal manufacturing and is not of interest to us. We are looking for a connection between the nutritional information variables (e.g. calories, protein, fat), the shelf on which the cereal is displayed, and the rating for the cereal.

The next pair of variables with the highest absolute sample correlation r=-0.76 is *sugars* and *rating*, which suggests that the more sugar a cereal contains, the lower its rating is. We can also see if the manufacturers also tie into this somehow; for example, a certain manufacturer makes low sugar cereals. Figure 2 below does not suggest any significant relationships except that all the Nabisco cereals appear to be low in sugar and with high ratings.

Figure 2.

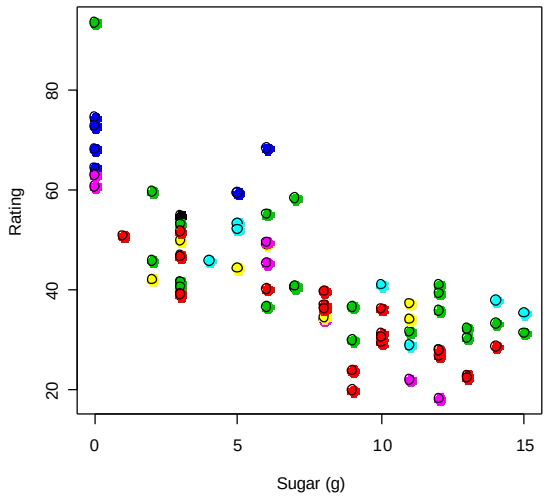
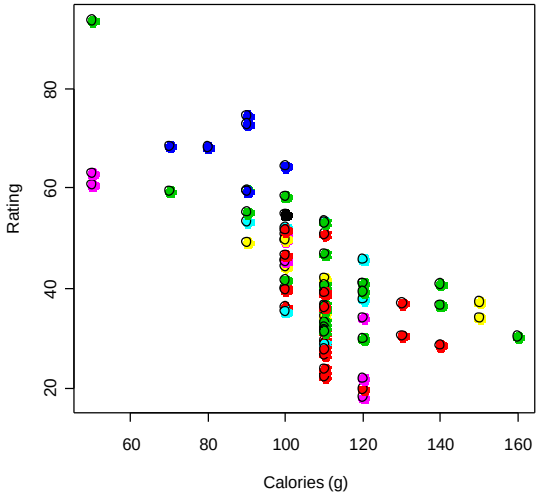


Figure 3.



There is also a significant sample correlation between *rating* and *calories* per serving, which are plotted in Figure 3. We can see a strong downward trend in rating as the amount of calories per serving increases. Again, Nabisco stands out with generally higher ratings, but the rest of the manufacturers are clumped together. Let us then explore the question, “is there a difference between cereals made by different manufacturers?” Of course, we are not done with rating yet – there are many influential factors beyond sugars and calories – but it suffices to show a relationship exists, and it will be examined in the next section.

When we separate the data by manufacturer name and construct some pair-wise plots, some interesting trends appear with regards to rating (Figure 4), sugars (Figure 5), and sodium (Figure 6). By interesting, we mean that there is a strong possibility that the sample median for a variable differs from one manufacturer to the next. This is indicated in our boxplots when the range covered by the notches of one manufacturer's boxplot does not overlap that of another manufacturer.

Figure 4.

A American Home Food Products
G General Mills
K Kelloggs
N Nabisco
P Post
Q Quaker Oats
R Ralston Purina

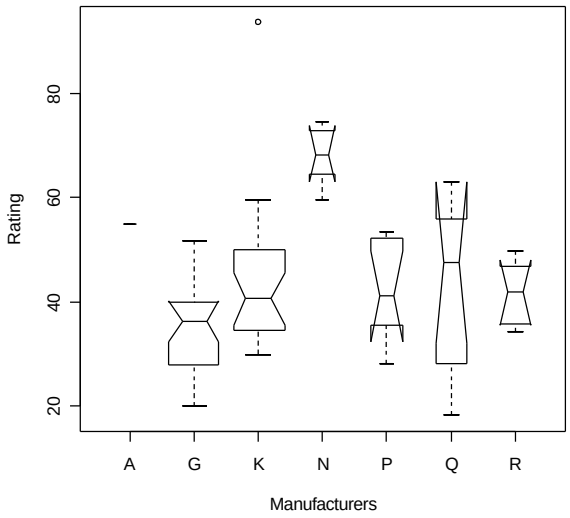


Figure 5.

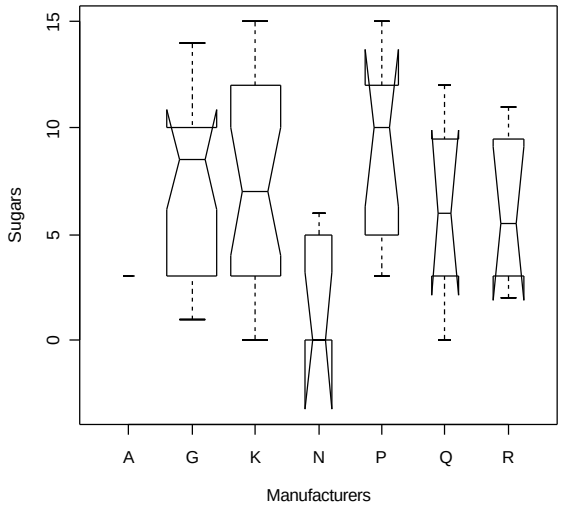
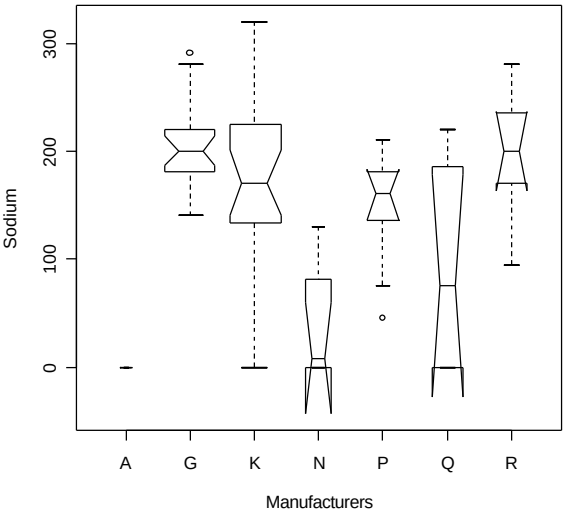


Figure 6.



There appear to be some real differences between manufacturers. For example, the sugar content of Nabisco cereals from the sample is significantly lower than Kelloggs or Post, which would certainly interest anyone on a diet! Nabisco cereals also seem to have lower sodium content than any of the other brands except Quaker Oats. Figure 4 simply confirms what we suspected from earlier, than Nabisco cereals in the sample had relatively higher ratings.

Confirmatory Data Analysis

We would like to know how the consumer rating is calculated, and suspect it largely depends on the nutritional information on the cereal box (after all, besides taste, there is no other good basis for a rating). Thus we can try to express rating as a formula involving the other variables.

Attempt 1: We must begin from a realistic model, first taking into consideration all other variables. We know that there should be no bias and that the manufacturer should not in any way influence the rating, so we may ignore *mfr*. Also, assuming hot and cold cereals have no intrinsic "healthy" or "unhealthy" qualities, the *type* variable can be discarded (especially since there are only 3 hot cereals in our sample). Lastly, the supermarket display *shelf* for the cereal should not influence the rating unless young children not tall enough to see the top shelves assigned the ratings. Thus, we try to fit

$$\text{rating} = \bullet_0 + \bullet_1\text{*calories} + \bullet_2\text{*protein} + \bullet_3\text{*fat} + \bullet_4\text{*sodium} + \bullet_5\text{* fiber} + \bullet_6\text{*carbo} + \bullet_7\text{*sugars} + \bullet_8\text{*potass} + \bullet_9\text{*vitamins} + \bullet_{10}\text{*weigh} + \bullet_{11}\text{*cups}$$

where the $\bullet_i$ are coefficients (numbers). The result is pleasant: the regression has r-squared value 1! This means that rating can be 100% explained by only the variables *calories, protein, fat, sodium, fiber, carbo, sugars, potass, vitamins, weigh,* and *cups* (our assumptions are okay).

Attempt 2: Now, let us try to simplify the model as much as possible.. Since the caloric value of food can be calculated from the protein, fat, carbohydrates and sugars content using a mathematical formula, *calories* is redundant and we will leave it out. In our previous attempt, both *weigh* and *cups* had p-values > 0.1, indicating they have a weak relationship with rating, so we leave them out too. The resulting fit should be nearly as good, and indeed, gives us a r-squared value of 0.9935. The model still explains over 99% of the variation in rating.

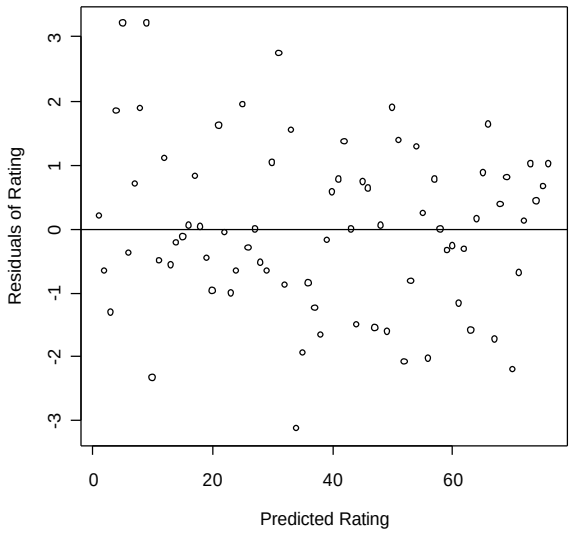
Attempt 3: The least significant coefficient from our second regression is for *carbo*, with a p-value of 0.00696. Removing this variable and fitting the resulting model gives us a r-squared value of 0.9928, which is still good!

Attempt 4: From our third regression attempt, we notice that *potass* coefficient has a p-value of 0.00051, so we will try to fit the data without this variable. The r-squared value in this case is 0.9909, suggesting *potass* can be removed from the model since it is not significant. Thus we have a regression of *rating* on *protein, fat, sodium, fiber, sugars,* and *vitamins* with p-values for all the coefficients much smaller than 0.0001, and we can be satisfied with this model:

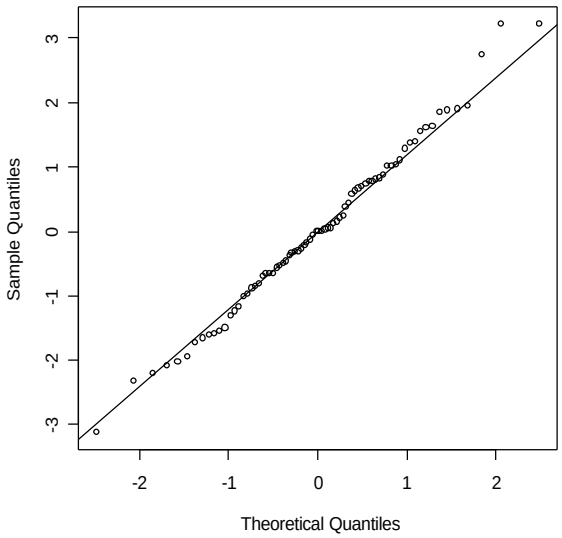
$$\text{rating} = 58.4 + 2.11\text{*protein} - 3.99\text{*fat} - 0.053\text{*sodium} + 2.40\text{* fiber} - 1.77\text{*sugars} - 0.048\text{*vitamins}$$

The appropriateness of the model and the t-tests used to determine the significant of coefficients assumes that the residuals are normally distributed. The residuals, or the differences between the value predicted by the model compared to the actual data, should have uniform variance in the range of values for *rating.* We examine this assumption in Figure 7:

Residual Plot (Figure 7.)



Normal Q-Q Plot for the Residuals (Figure 8.)



The residuals are well-behaved, with no obvious trends. However, the variance seems to be slightly larger for small values of rating. In Figure 8, we have a quantile-quantile plot of the residuals. If they are normally distributed, then they should follow a straight line – and they do, except for 3 odd points near the top right-hand corner. This is an indicator for "too many" large positive residuals and in fact, these correspond to the 3 points in Figure 7 at the top left-hand side which give us the impression of larger variance.

Attempt 5: We can choose to ignore these 3 possible outliers, given that there are 77 data points in total. But might we obtain a better fit by removing these points? The resulting model is almost identical, with an r-squared value of 0.9928 (slightly larger than 0.9909, but not significantly), and is given by

rating = 58.2 + 2.15*protein - 3.98*fat - 0.052*sodium + 2.40* fiber - 1.79*sugars - 0.047*vitamins

The difference in both r-squared value and coefficients is negligible, so we will keep the 3 data points and our old model.

Conclusions & Open Issues

The *Consumer Reports* rating can be calculated 99% from just a few nutritional variables: protein, fat, sodium, fiber, sugars, and vitamins. We do not know if this is the actual formula used, but we do know it is very accurate for all practical purposes. Our final model is

rating = 58.4 + 2.11*protein - 3.99*fat - 0.053*sodium + 2.40* fiber - 1.77*sugars - 0.048*vitamins

This tells us that according to our model, the "default" rating for all cereals is about 58. For every gram of protein per serving, the rating goes up by 2.11 points, and for every gram of fat per serving, the rating goes down by about 4 points! Clearly, *Consumer Reports* does not like fat in cereals. The rating also decreases with increases in sodium and vitamins, which seems to contradict common sense since more vitamins and minerals is beneficial. At the risk of a worse fit, we may choose to ignore vitamins in further analysis. Fiber is good, so we are not surprised to learn rating increases by 2.4 for every gram of dietary fiber per serving in the cereal. Simple sugars are regarded as unhealthy, and accordingly, rating decreases by 1.77 for every gram of sugar per serving. In summary, protein and fiber are "good" nutrients that increase rating, while fat, sodium, sugar, and possibly vitamins are "bad" and decrease rating.

Since the data set contains missing a few missing values, some of our results may be slightly off, but we hope the effect is small since most of the data is complete. It is also interesting how the complex carbohydrates content does not have a significant effect on rating, according to our model. We suspect that recommended serving sizes may come into play as well, because all of the other variables (such as protein, fat, etc.) depend on the serving size, which we have ignored as insignificant. If *Consumer Reports* calculates rating based on nutrients per serving however, then our model is still valid.

Manufacturers never came into our model, since we assumed unbiased ratings. It seems like a safe assumption, given that our first model without manufacturers was able to explain 100% of the *rating* variable. However, we did notice trends from one manufacturer to another. Nabisco cereals seem to be low sugar, low calorie and low sodium – and also highly rated by *Consumer Reports*, so they are likely high in fiber and protein as well.

The "rating formula" derived correctly estimates the given ratings in the sample by 99%, and satisfies normality requirements. Our next step might be to collect a larger sample of breakfast cereals and test to see if our model fits this new data set accurately.