EXECUTIVE SUMMARY

This report investigates what factors are important in achieving a high rating in breakfast cereals. A dataset of 77 different breakfast cereals is used in this research, and the rating is calculated by *Consumer Reports*. This dataset contains per-serving nutritional information and grocery shelf locations from seven manufacturers. Table 1.1 displays the variables and descriptions in the dataset used in this research. Three data points contained missing data item, so they are ignored and not taken into account.

Variable	Description
name	name of cereal
mfr	manufacturer[1]
type	C = cold or H = hot
calories	calories per serving
protein	protein (grams)
fat	fat (grams)
sodium	sodium (milligrams)
fiber	dietary fiber (grams)
carbo	complex carbohydrates (grams)
sugars	sugars (grams)
potass	Potassium
vitamins	typical percentage of Food and Drug Administration (FDA) recommended daily amount of vitamins and minerals
shelf	supermarket display shelf (counting 1 = floor)
weight	recommended serving size (ounces)
cups	recommended serving size (cups)
rating	*Consumer Reports* rating

Table 1.1 Variables and description of the dataset used.

It is of interest to investigate which nutritional ingredient affect the *rating* by the consumers. In addition, an extra factor, *manufacturer*, is also considered when determining the *rating*. The investigation shows that the *manufacturer* and ingredients play a significant role in determining the *rating*.

SCREENING OF VARIABLES

Not all the variables from the dataset presented in Table 1.1 are used in the investigation. For instance, the *type* variable is not of interest since by common sense, most cereals are consumed cold. Also, after having deleted the three missing data points

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

mentioned earlier, the dataset now contains only 1 data point of type “hot”, which is not very meaningful if it were included in the following analysis. The variable *vitamins* is also not included due to the distribution of this variable. By not including *vitamins* in the analysis, it does not imply that this variable is not important in determining the *rating*. It is just the case that there is not enough data in this variable to address this investigation precisely. The *vitamins* variable contains data of three different percentage values (0%, 25%,100%) of which 61 out of 73 data points (83.5%) are under the 25% category, 6 out of 73 under the 0% and 6 out of 73 under the 100%. Most of the data points in the *vitamins* variable are clustered around one value, which does not have much variability for analysis.

DATA ANALYSIS

First of all, *rating* between different manufacturers is compared and is shown in Figure A. Most manufacturers have ratings that are quite similar to each other, as shown by the overlapping of boxplots. However, only manufacturer *Nabisco* (N) is relatively higher compared to the rest. Manufacturer *American Home food Products* (A) contains only one data point; hence, no variability can be assessed and as a result, manufacturer A is ignored in the analysis. A potential outlier is detected from this boxplot, which has the highest rating value. This point corresponds to *Kelloggs’* (K) cereal with a rating of 93.7. By comparing this value with the other manufacturers’ cereals in Table 1.2, this data point is suspicious of being a potential outlier. However, this point is not ignored, because it might be the fact that this particular cereal is rated high by the consumers.

Table 1.2 further explains the summary of the *rating* values for each of the manufacturer. The range of ratings in this dataset is 18.04 to 93.70. The mean of the *rating* ranges from 34.49 to 68.66, with the lowest standard deviation (sd) 5.78 to the highest of 14.46. A high sd implies that the estimate of the average rating for that particular manufacturer is not as reliable as compared to one with a lower sd. Manufacturer Q for example has the highest sd, and this is shown graphically in Figure A by the long boxplot which shows a great amount of variability. On the other hand, manufacturer R has the lowest sd, and also the shortest boxplot in Figure A. Graphically, the length of the boxplot signifies the reliability of the estimate values of *rating*. The shorter the boxplot, the smaller the sd; hence, the more reliable the estimate of *rating*.

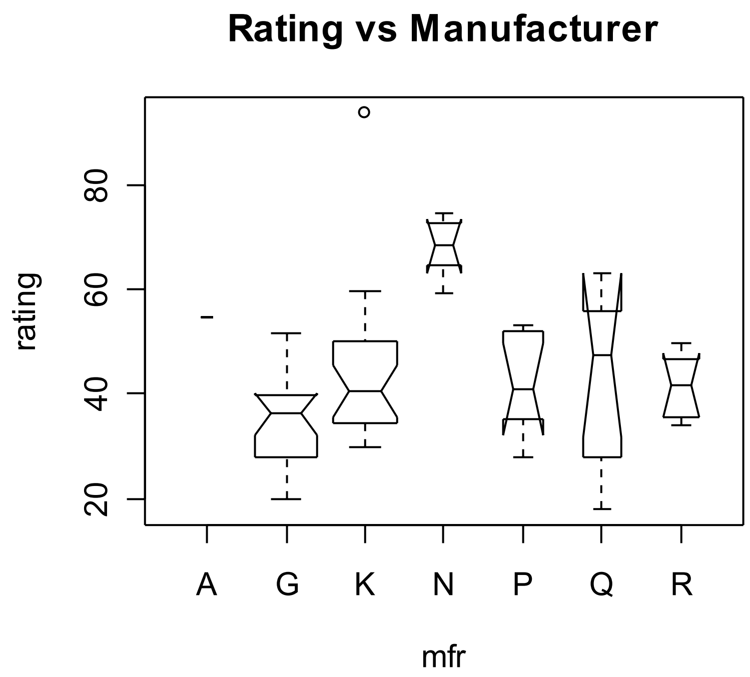

Figure A Boxplot of rating versus manufacturer.

Manufacturer/ Statistics	G	K	N	P	Q	R
samplesize	22	23	5	9	7	7
min	19.82	29.92	59.36	28.03	18.04	34.14
mean	34.49	44.04	68.66	41.71	41.79	42.57
max	51.59	93.70	74.47	53.37	69.01	49.79
standard deviation	8.95	14.46	5.87	10.05	17.82	5.78

Table 1.2 Summary of ratings grouped by the manufacturer.

Next, it is observed how ingredients and manufacturer affect *rating* with the following regression model constructed,

$$\text{rating} \sim \beta_0 + \beta_1 \textit{mfr} + \beta_2 \textit{calories} + \beta_3 \textit{protein} + \beta_4 \textit{fat} + \beta_5 \textit{sodium} + \beta_6 \textit{fiber} + \beta_7 \textit{carbo} + B_8 \textit{sugars} + B_9 \textit{potass} \qquad \textbf{(Model I)}$$

Figure B displays a pairs-plot which illustrates how two variables are related with each other, amongst the ingredients, manufacturer and rating variables. Along with the plots, a histogram is shown on the diagonals to illustrate the distribution for each variable. The bottom left of the graph displays scatter plots of two variables with a red smooth curve that helps to capture the relationship.

The top right of the graph shows the correlation coefficient between the two corresponding variables. A correlation coefficient ranges from 0 to 1, and the higher the value, the higher the collinearity[2] between the two variables. The existence of collinearity means that the variables are "moving with each other" in the dataset, which makes it more difficult to sort out the impact of each variable on *rating*. From the graph, it is very obvious that there is a linear trend between *fiber* and *potassium*, yielding a correlation coefficient of 0.91. This collinearity is further investigated by the circle plots in Figure C. The circle plot is similar to a scatter plot of two variables plotted against each

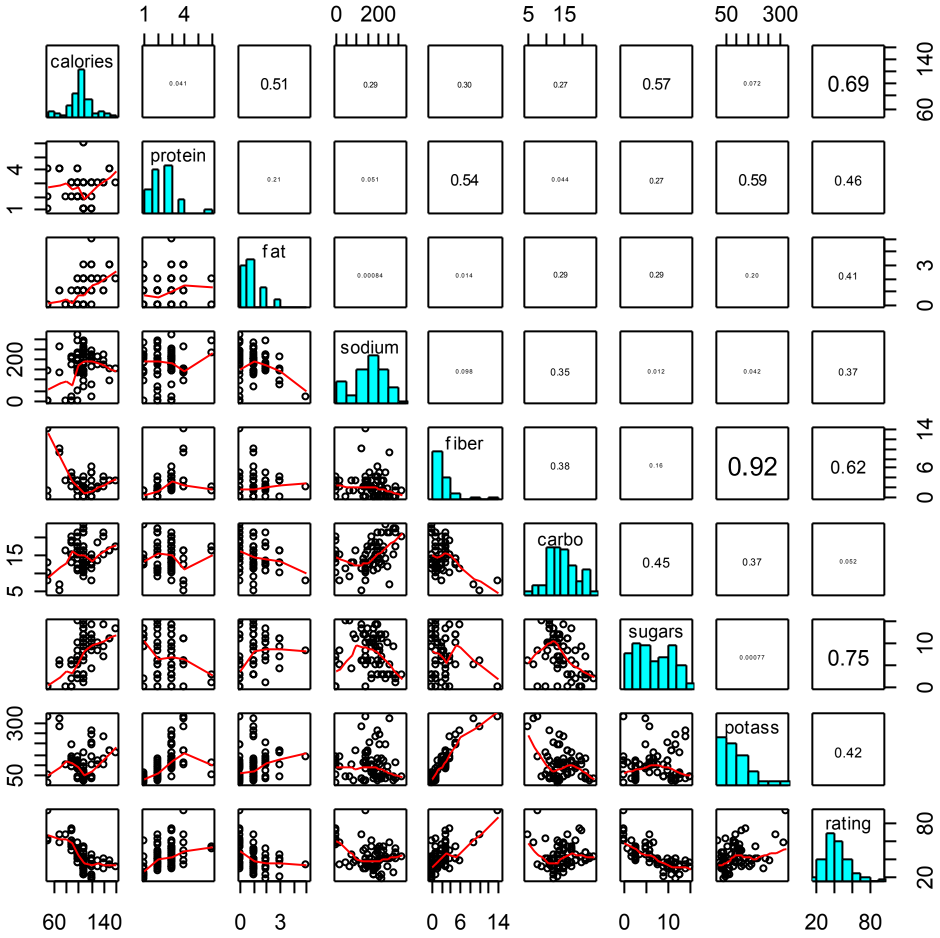

Figure B Pairs-plot for the ingredients, manufacturer and rating variables.

[2] "Collinearity causes variances of the coefficients of the regressors in the model to be inflated due to collinearity…This inflation of variance underscores the damage due to the condition of collinearity." (page 125 of Raymond H. Myers', Classical and modern regression with applications, 1990.

other. In this case, the two variables are *rating* versus *potassium*. A third variable is also added into the plot, which is *fiber.* This variable is shown by the size of the circles which is proportional to the size of the amount of fiber. In Figure C, each color corresponds to the manufacturer, and a dotted smooth curve is drawn to emphasize the relationship between *rating* and *potassium*. A pattern is observed from this plot which shows that as potassium increases, so do the sizes of the circles (ie. fiber). Similarly, as rating increases, both fiber and potassium also increase. This plot also illustrates the variability of potassium, fiber and rating for the different manufacturers. For example, Kelloggs (K), indicated by the green circles, has a variety of different cereals ranging from the lowest potassium values in this dataset to the highest potassium values. Other manufacturers, G, for instance, have most of its red circles clustered in a shorter range of potassium values. These red circles have a lesser variability (shorter range of values) in potassium and rating with the sizes of the circles being roughly the same size. This implies that the *fiber*

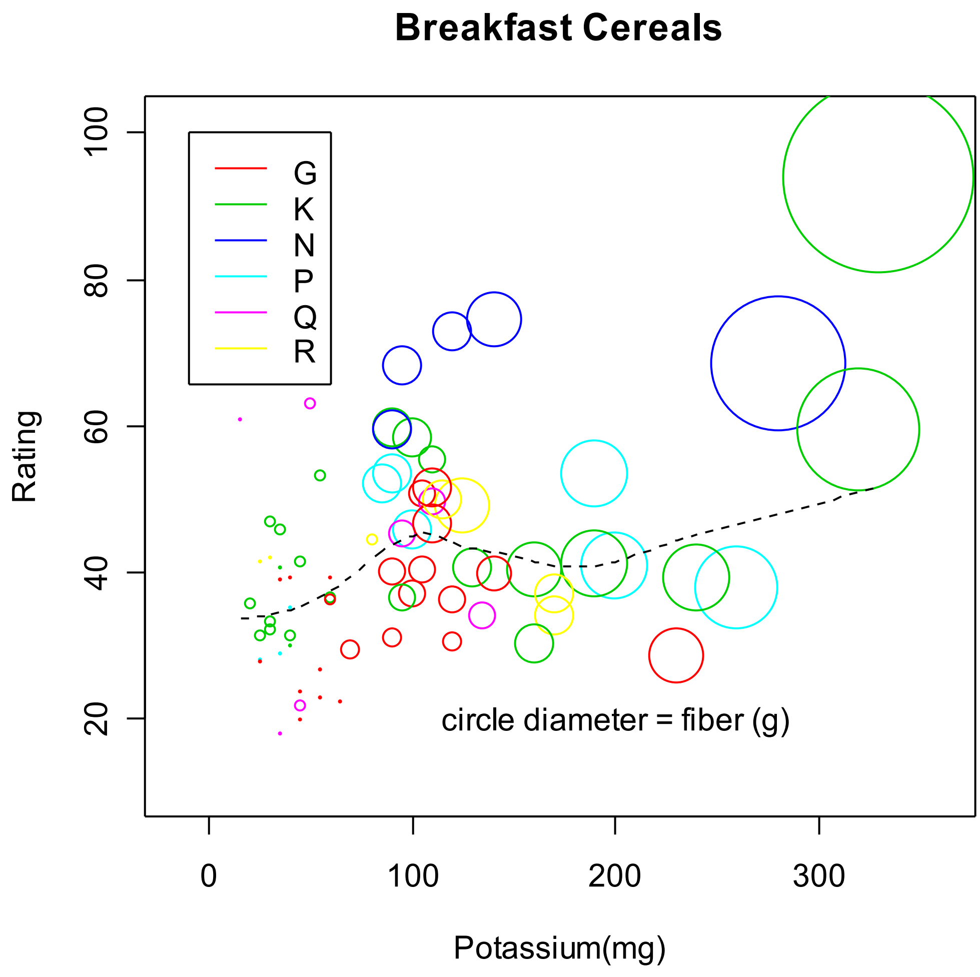

Figure C Circle-plot of rating vs. potassium, with the size of the radius proportional to the Amount of fiber.

values are somewhat constant throughout the different cereal brands for manufacturer G.

A strong collinearity exists between the two variables, *fiber* and *potassium,* and one of them is taken out. Two statistical partial F-tests[3] were conducted and it revealed that with *fiber* in the model, *manufacturer* becomes an insignificant variable. However, without *fiber* in the model, *manufacturer* becomes a significant variable in predicting the rating. By commons sense, *manufacturer* should be a factor that would affect the *rating*. Since *fiber* and *potassium* are collinear with each other, no information is lost if fiber is taken out. Having *fiber* removed from Model I, significant information from the *manufacturer* variable contributes to *rating.*

Model I is constructed without the *fiber* variable and the following table displays values of the estimated betas.

Variable	Estimated Beta Value	Standard Error
(intercept)	64.71	2.43
mfrK	3.30	0.87
mfrN	4.50	1.70
mfrP	2.42	1.11
mfrQ	-0.46	1.24
mfrR	2.62	1.16
calories	-0.19	0.06
protein	2.74	0.46
fat	-2.40	0.72
sodium	-0.05	0.01
carbo	0.29	0.28
sugars	-1.40	0.26
potass	0.06	0.01

Table 1.3 A table of the estimated beta values and standard errors from Model I.

The preceding estimated beta values is interpreted as follows. The rating will tend to increase by 1% if the following is increased:

- 2.74 (g) of protein; 0.29 (g) of carbohydrates; 0.06 (mg) of potassium

and if the following is decreased:

- 0.19 calories per serving; 2.40 (g) of fat; 0.05 (mg) of sodium; 1.40 (g) of sugars

The *manufacturer* beta values suggest how much higher the ratings are compared to the

[3] Partial F-tests give information regarding the importance of a single variable in the model involving all other variables.

other remaining manufacturers that are held fixed. For instance, Kelloggs (mfrK) has a rating that is 3.30 units higher than other manufacturers when all the ingredients variables are held fixed for all the other manufacturers.

Replacing the estimated beta values in Table 1.3 with the betas in Model I (without *fiber*), a prediction equation is constructed,

$$\text{rating} = 64.71 + 3.30\textit{mfrK} + 4.50\textit{mfrN} + 2.42\textit{mfrP} - 0.46\textit{mfrQ} + 2.62\textit{mfrR} - 0.19\textit{calories} + 2.74\textit{protein} - 2.40\textit{fat} - 0.05\textit{sodium} + 0.29\textit{carbo} - 1.40\textit{sugars} + 0.06\textit{potass} \qquad \textbf{(Eq II)}$$

Manufacturers can now use the above equation to predict what the rating value would be.

CONCLUSION

The dataset used in this investigation had some missing data items, and variables that contained too little information for analysis. Hence, these data points and variables were taken out in the analysis. The topic of interest is to observe how the factors, ingredients and manufacturer, affect *rating*. Even though *vitamins* may also be an important factor, but it was not included in the analysis because the data did not contain enough variability to be assessed. Furthermore, the two collinear variables, *fiber* and *potassium*, are remedied by removing *fiber* out of the analysis. Having done so, *manufacturer* becomes a significant factor, which makes sense, since ratings by consumer should somehow be affected by what manufacturer the cereal is produced from. Again, not including the *fiber* variable does not mean that *fiber* has no effect on *rating*, it is the fact that *fiber* and *potassium* are collinear, which means that *potassium* captures that information about *fiber*. As seen with the pairs-plot and circle-plots, an increase in fiber shows a significant increase in potassium. All in all, variables in Eq. II are all important in predicting the *rating* values.

LIMITATIONS

In the section “Screening of Variables”, *vitamins*, could not be used since most of the data were of value 25 with not a lot of variability to be assessed. This dataset is obtained from an observational study, hence data cannot be controlled to produce a desired result. Therefore, no matter what, no data can be obtained so that there is a range of different percentage of Food and Drug Administration recommended daily amount of vitamins and minerals. Another limitation is the restriction on the prediction equation (Eq II). This equation is only good for values of each variable that are within the range of this dataset. That is, if values for the variables are outside the range (min to max values in Table 1.2), the prediction equation fails.