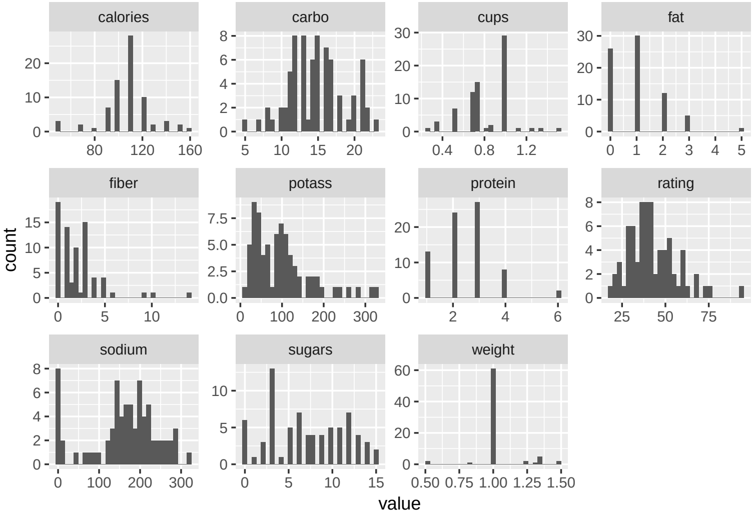
Healthy Breakfast EDA

Gustavo Esparza, Lindsay Brock, Brian Schetzsle

11/7/2019

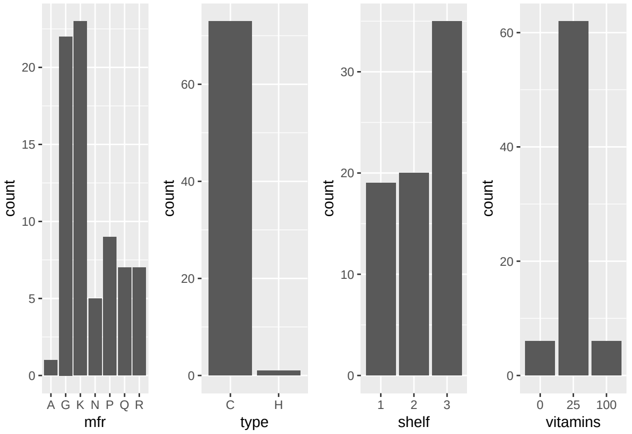
1 Do you agree with Consumer Reports grading of the cereals?

We will first observe the distributions for each individual variable in the data set:



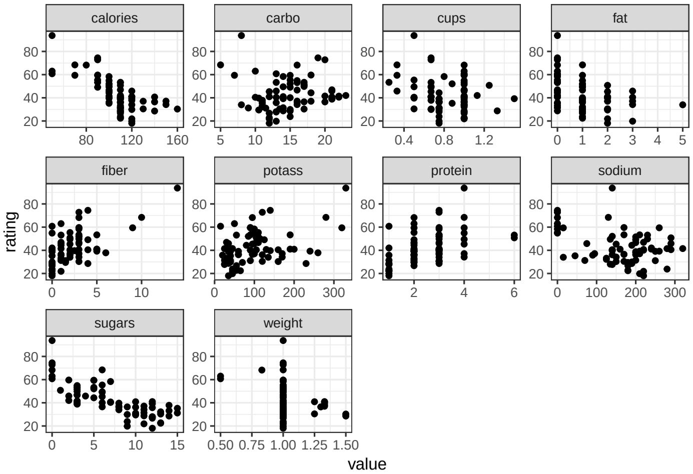
Initially, we can observe the following variables to have minimal variance: **Cups, Weight**. These two variables pertain to industry standards for the amount of cereal in a single serving, and as long as they are observed and followed by the consumer they should have no impact on the Consumer Reports rating. This was confirmed during original exploration of the data, when an attempt was made to modify the variables so that all serving sizes were equivalent to one cup. It was observed that there was no significant impact made to the distribution of the individual variables and their relationship to rating, so they cup sizes were left to their original observations for this analysis. The variables **Calories**, and **Carbohydrates** appear to have a bell curve shape while other variables such as **Fat, Fiber, Potassium,** and **Protein** are skewed right.

Now, we will inspect the categorical variables in order to understand their distributions:



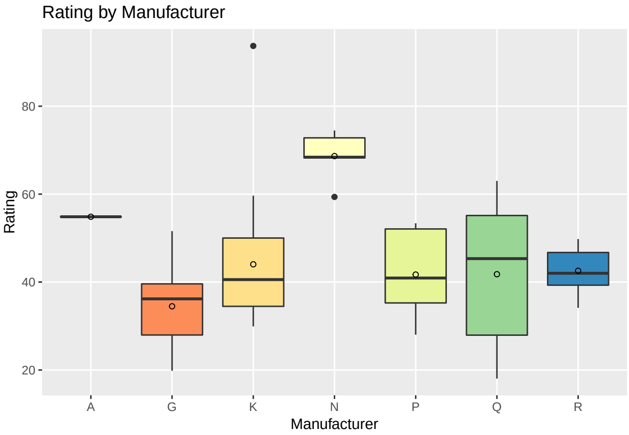
It is immediately evident that the variables **Type** and **Vitamins** have minimal variation and can be assumed to be uniform across most manufacturers of cereal. Looking closer at the observed values, **Type** (hot/cold cereals) contained so few hot cereals it is safe to say almost all of cereals are served cold and very few cereals contained 0 or 100 percent **Vitamins**, leading to their uniformity. **Shelf location** does seem to have more cereals located on the third shelf, and manufacturers in this data set appear to primarily come from manufacturers **General Mill** and **Kelloggs**.

Now, we will take a look at our primary point of interest, the relationship between the response of *Rating* against all other predictor variables:

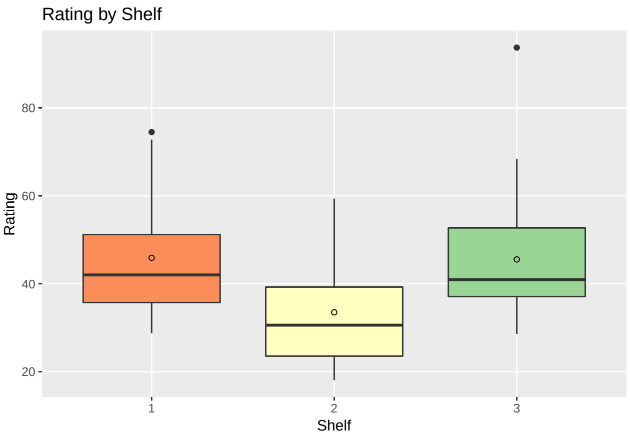


By a simple inspection of our scatter plots, we can observe the following variables to have a clear positive relationship with Ratings: **Fiber, Potassium,** and **Protein** and the following variables to have a clear negative relationship with Ratings: **Calories, Fat,** and **Sugars**. It is clear that none of these relationships are perfectly linear.

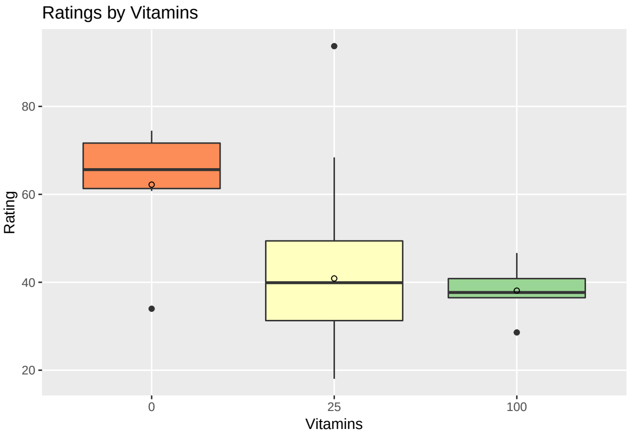
In addition, we will display the relationship between our categorical variables and Rating:



We can observe that most Manufacturers have a similar distribution of ratings, where **Nabisco** appears to have the highest ratings overall. In addition, we can observe one cereal from **Kelloggs** having the highest rating overall and therefore boosting the average ratings for the entire group. Based on this finding, this observation could be considered an outlier and will be discussed later in this analysis.

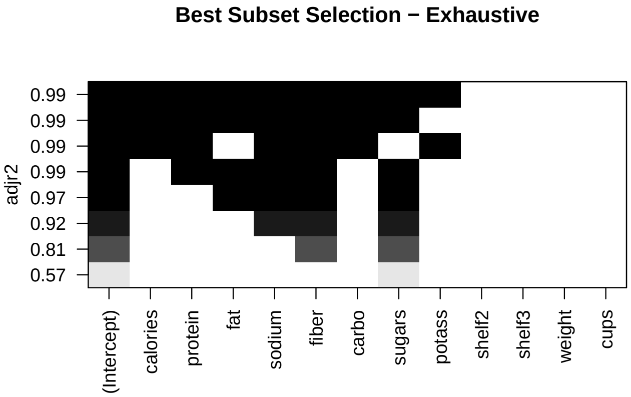


By inspection of the shelf location and its impact on ratings, we can observe that all three shelves are mostly uniform in distribution with the exception of shelf 2 having lower ratings.



By inspection of vitamins and their relationship to rating, it can be observed that there is a negative trend with rating and the percentage of vitamins in a cereal. It should be noted that most cereals in the data set have 25 percent of the vitamins present, which may contribute to the visual skewing observed.

By observing the relationship between our predictors and the response of Consumer rating, it does seem that there is a general consensus of “healthier” variables found within the cereal to correlate to higher consumer ratings. To further establish this conclusion, we will perform a best subsets regression on our original data:

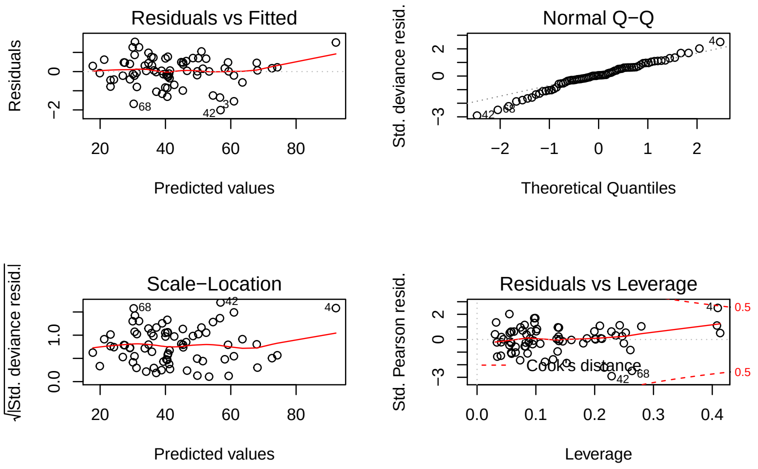


Using the model with the highest adjusted r^2 from the plot above we find that the variables *calories, protein, fat, sodium, fiber, carbo, sugars,* and *potass* were the most likely to have been included as they account for the highest amount of variance within the data set.

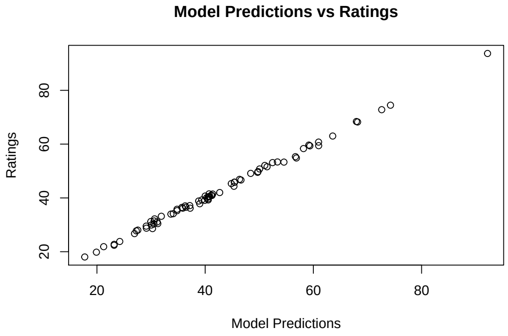
Now, we will fit a model using the optimal predictors:

$$ratings = \hat{\beta}_0 + \hat{\beta}_1 calories + \hat{\beta}_2 protein + \hat{\beta}_3 fat + \hat{\beta}_4 sodium + \hat{\beta}_5 fiber + \hat{\beta}_6 carbo + \hat{\beta}_7 sugars + \hat{\beta}_8 potass$$

To ensure the model is fit appropriately we must not forget to confirm it upholds the all important four assumptions using the plots below. The Residual vs Fitted plot which we use for the linearity assumption, seems straight enough when considering the red dotted line with variances that seem evenly spaced although slightly bunched to the left. This could be from the single point up high and to the right, an outlier. The Scale-Location plot which tells us about constant variance looks okay but could be better. It should follow a straight red line and the points should be more consistently spread out instead of clustered to the left, yet again. The Normal QQ Plot which represents the normality of the standard residuals seems straight enough along the dotted line besides the few trailing outliers at the bottom left. As for the assumption of independence, we would hope that the data was collected independently.



Finally, looking at the plot of the observed ratings compared to those predicted using the model we can see it is almost perfectly linear. The fit of the model seems to be appropriate.



We will now compute the Mean Squared Error of the Consumer Ratings model to get an estimate of how accurate it is in predicting the original observed ratings. It is extremely small supporting our conclusion that the model fit is appropriate.

```
## [1] 0.5367158
```

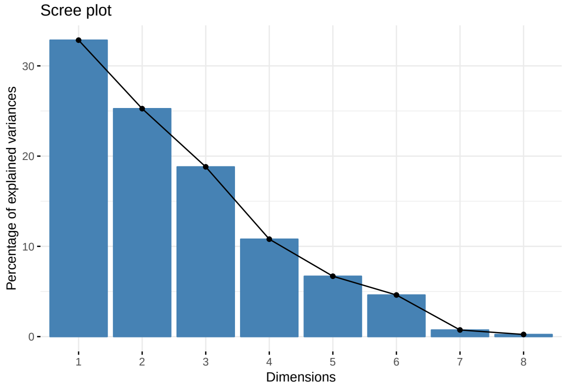
Overall, after analysis of the individual variables and evaluation of the variables Consumer Reports actually used to create the ratings we would agree with their conclusions. Although we agree, we also believe these ratings could be created using fewer variables as described in the analysis below.

2 BUILD YOUR OWN NUTRITIONAL MODEL

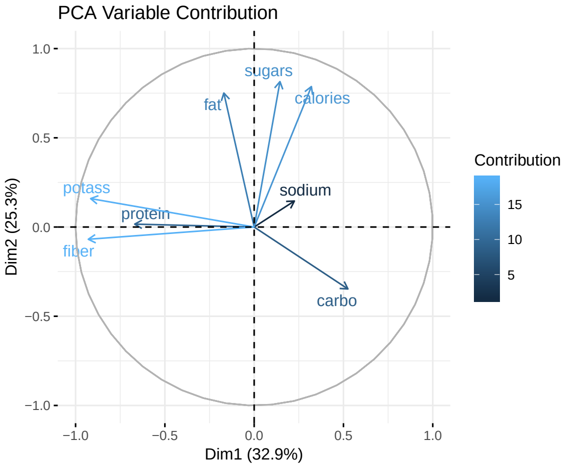
We wish to build a model that uses a minimal number of features to determine the nutritional value of a cereal. We have just analyzed our given data set and determined several nutrition-related variables to have a clear relationship with the response of **Rating** which are **Calories, Protein, Fat, Sodium, Fiber, Carbohydrates, Sugars,** and, **Potassium**. The objective here is to construct a model from these predictors that is minimal in dimensionality, but optimal in inferential and predictive efficiency.

PCA

We can begin our model development by performing a Principal Component Analysis on the Health related variables. Here is a scree plot of our computed principal components along with their percentage of explained variance:



We can observe that the first three components explain a majority of the variance (around 80%) within the data. We would like to analyze which variables contribute to which components, and to which capacity they contribute. We can display such a plot in the following manner:

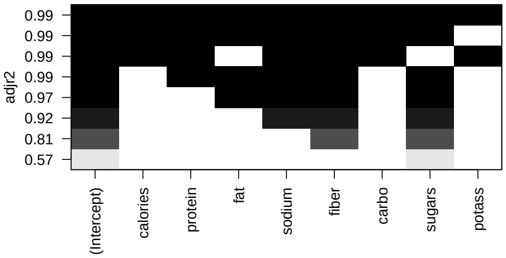


Using the first two components, we are able to see a distinct grouping of “Healthy” variables **Potassium, Protein,** and **Fiber**, with **Potassium** and **Fiber** having the most contribution. There is also another visible grouping of “Unhealthy” variables **Fat, Sugars,** and **Calories** which all see to have the same amount of contribution to the principal components. **Carbohydrates** and **Sodium** seem to be the only ungrouped variables that have the smallest contribution, overall.

BEST SUBSETS

PCA was useful in observing which variables collectively contributed to the first two components. Keeping this information in mind, we should now investigate whether a subset of these predictors can construct a regression model that retains inferential and predictive power. For this purpose, we will perform a best subsets analysis on the variables used in the original Consumers Rating model constructed above, using R^2 as a reference point for the validity of this new model.

Here is the adjusted R^2 plot for our subsets, which will be used to select a best subset model:

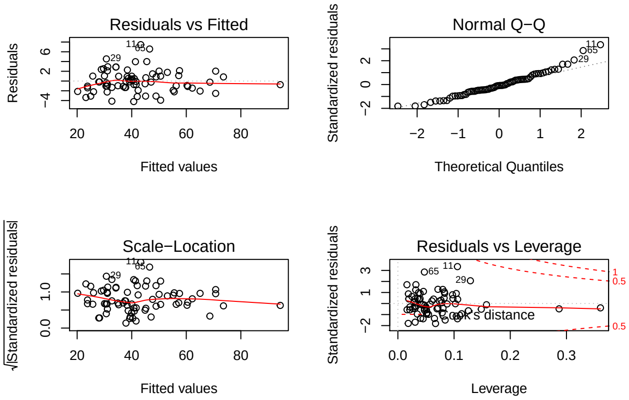


By inspection of R^2, we can observe that a model including four predictors reaches a correlation coefficient of .97. Although adding more predictors does increase the overall R^2 value, it is not explaining a great deal of additional variation in the overall data and would seem like an appropriate choice for our new model.

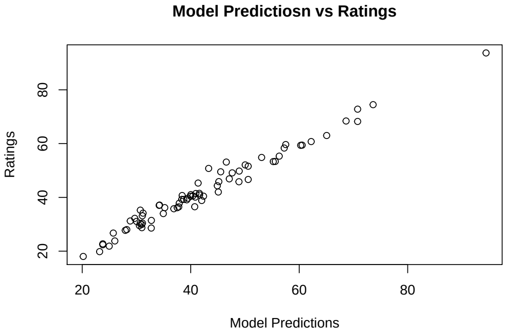
Now, we will fit a model using the optimal predictors: **Fat, Sodium, Fiber,** and **Sugars**. Here is the provided summary for the model:

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 62.20743569 0.83559381  74.44698 1.205077e-67
## fat         -3.33393693 0.28682484 -11.62360 6.570200e-18
## sodium      -0.05521875 0.00334548 -16.50548 1.174305e-25
## fiber        2.85519324 0.11571494  24.67437 5.892658e-36
## sugars      -1.93414352 0.06703325 -28.85350 3.099557e-40
```

To ensure this new model is fit appropriately we must confirm it upholds the all important four assumptions, yet again, using the plots below. The Residual vs Fitted plot used for the linearity assumption, seems straight enough when considering the red dotted line with variances that seem evenly spaced although slightly bunched to the left. This could be from the single point to the right, an outlier. The Scale-Location plot used to describe constant variance looks okay but could be better. It should follow a straight red line and the points should be more consistently spread out instead of clustered to the left, yet again. The Normal QQ Plot representing the normality of the standard residuals seems straight enough along the dotted line besides the few outliers now at the top right. Again, we would hope that the data was collected independently for the assumption of independence.



Finally, looking at the plot of the observed ratings compared to those predicted using the model we can see it is almost perfectly linear. The fit of the model seems to be just as appropriate as the original model previously fit.



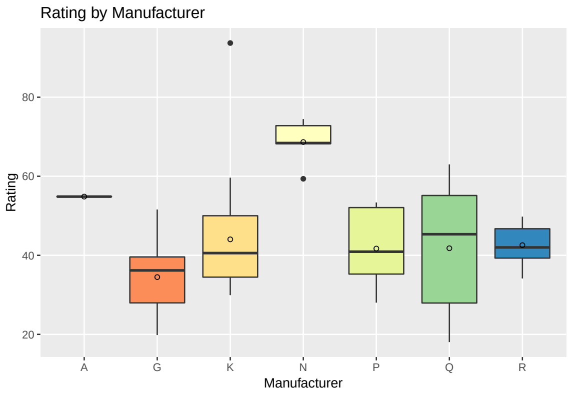
We will now compute the Mean Squared Error of our model to get an estimate of how accurate our model is in predicting the original ratings observed.

```
## [1] 5.189127
```

From the results of our reduced “Nutritional” model, we can observe that there is a general agreement between our model and the original Consumer Ratings.

3 MANUFACTURERS

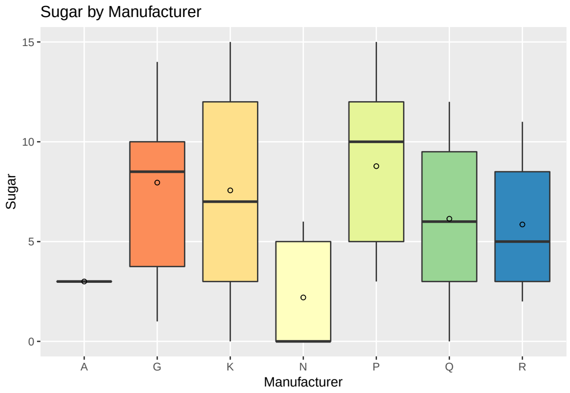
If we were to decide which manufacturers make the healthiest cereals based on the ratings of the Consumer Reports, we want to look at the average rating of each manufacturer as in the bar plot below. It is abundantly clear that Nabisco wins with an average rating of 67.97. The second runner up would be American Home Food Products with an average rating of 54.85.



Instead of relying simply on the rating we might also want to take a look at which variables are most important to the Consumer Reports model and see which manufacturers are “healthiest” with respect to these variables. Using the bagging approach on the regression tree fit with the variables from the original Consumer Reports model, we find of the variables listed below the most important to be *sugars* with *calories* following close behind. We will now begin to group the cereals by manufacturer and observe their interaction with variables associated with nutritional value.

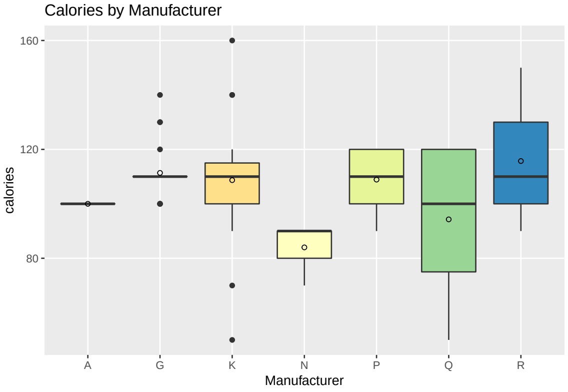
```
##             %IncMSE IncNodePurity
## calories   7.997040     3058.7965
## protein    6.422024     1007.0005
## fat        4.759244      791.4768
## sodium     8.360199     1754.1890
## fiber      6.425356     1812.9336
## carbo      5.019082     1037.3926
## sugars    10.568401     3901.5199
```

Sugars



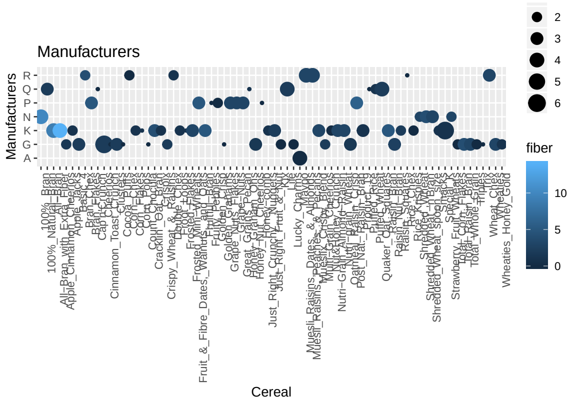
Plotting the average amount of sugar by manufacturer tells a different story of “health” compared to the rating. We can see from the plot above that Nabisco could be considered healthiest as it has the least amount of average sugar with American Home Food Products following close behind, yet again.

Calories



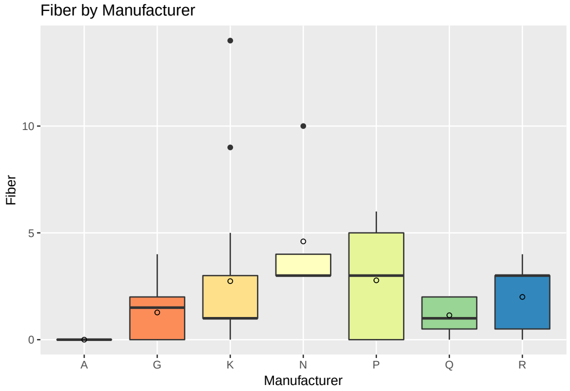
Plotting the average amount of calories by manufacturer (above) we can conclude again the “healthiest” cereals come from Nabisco based on its average calories. This seems to fall right in line with the conclusion solely relying on rating, that the “healthiest” manufacturer of the group seems to be Nabisco with American Home Food Products following close behind.

Fiber and Protein:



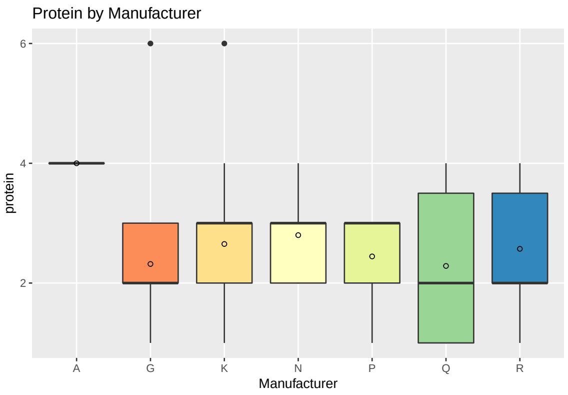
In this plot, Lighter shaded and larger sized circles indicate a cereal with a higher Fiber and Protein count. We can observe these cereals mainly belong to the **Kelloggs** group. Groups **Post** and **Quaker** contain cereals that have some of the lowest counts for these two nutritional parameters.

Fiber



When comparing Fiber across all manufacturers, we once again observe that **Nabisco** contains the cereals with the highest counts.

Protein



For protein counts, we see a fairly even distribution of protein values for all manufacturers. We can still observe that groups **General Mills** and **Kelloggs** each contain a single cereal with a high Protein count.

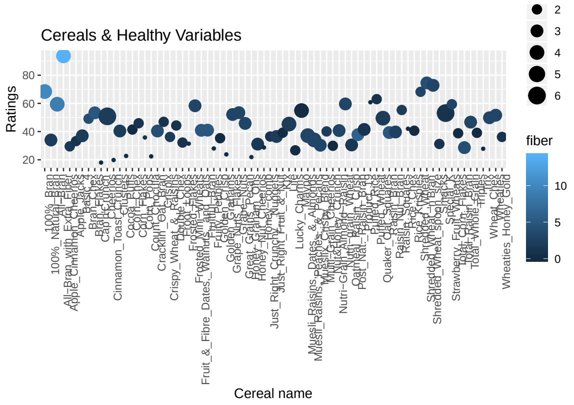
4 CONTRADICTIONS

One particular pair of cereals that seemed to have contradicting feature trends were *Cheerios* and *Special K*. Both of these cereals had relatively high Protein levels (6) yet had strikingly low ratings (around 50 for both). Initially, this seems quite strange when considering the Protein value but it is important to consider the high sodium and carbohydrate levels found in both observations.

name	mfr	type	calories	protein	fat	sodium	fiber	carbo	sugars	potass	rating
Cheerios	G	C	110	6	2	290	2	17	1	105	50.76500
Special_K	K	C	110	6	0	230	1	16	3	55	53.13132

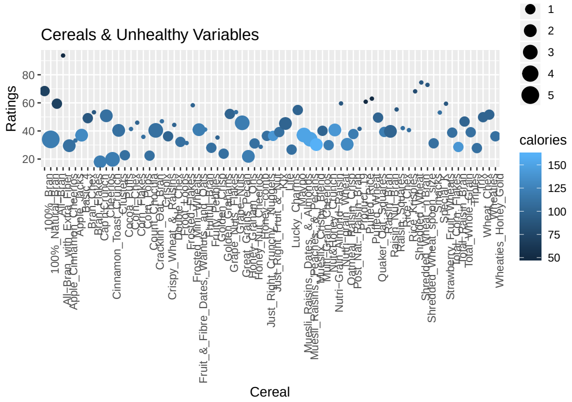
5 CONTINUOUS IMPACT ON RATING

First, we will observe a graph that plots Rating against healthy variables Fiber and Protein:



We can observe that the highest rated cereals generally contain high amounts of protein and fiber (large circles than are generally lighter). There is a single observation that is the highest rated cereal, which clearly has a higher fiber count than any other cereal in the data set.

Next, we will plot Ratings against some of the unhealthy variables: Calories and Fat



We can observe that the highest rated cereals are those that have minimal fat and calories (small dark circles).

```
summary
```

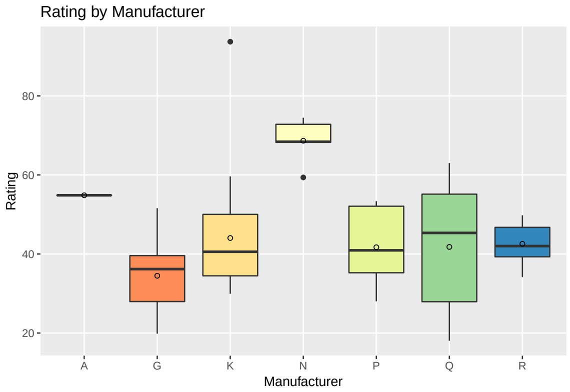
```
##
## Call:
## lm(formula = rating ~ ., data = cereal3)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.213 -1.331 -0.200  1.049  7.463
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 62.207436   0.835594   74.45   <2e-16 ***
## fat         -3.333937   0.286825  -11.62   <2e-16 ***
## sodium      -0.055219   0.003345  -16.50   <2e-16 ***
## fiber        2.855193   0.115715   24.67   <2e-16 ***
## sugars      -1.934144   0.067033  -28.85   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.359 on 69 degrees of freedom
## Multiple R-squared:  0.9733, Adjusted R-squared:  0.9717
## F-statistic: 628.6 on 4 and 69 DF,  p-value: < 2.2e-16
```

We can also understand the relationship between continuous variables and the response using the coefficient estimates produced by the model we fit using a subset of only continuous variables. From the given summary above, the intercept tells us that if all of the other variables in this model (fat, sodium, fibers, and sugar) were zero, the average cereal rating would be around 62.5. A higher average rating seems appropriate as all of our unhealthy variables are zero but the average rating could also be limited because there is no healthy protein. It would be quite impossible to have zero calories or zero of the other variables for that matter, so lets take a look at the individual variables.

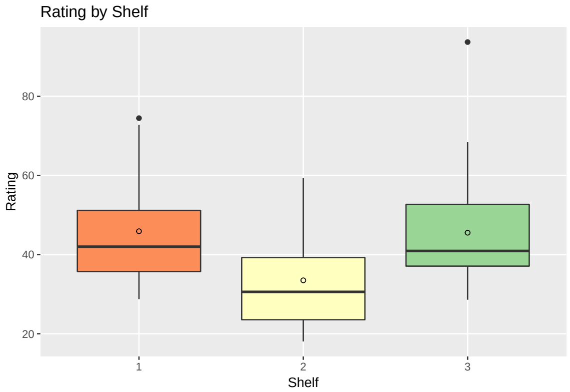
The estimated coefficient of -3.33 for fat tells us that while holding all other variables constant, for each increase in unit of fat the rating would decrease by -3.33. As the estimate is a random variable itself we must also take into consideration its standard error of 0.29. Using this standard error we can be confident that 95% of the time the average decrease in rating for a unit increase in fat will be between -3.89 and -2.77. As each unit of sugar increases, we also find the rating decreases by -1.93 [95% CI: (-2.06, -1.8)], holding all other variables constant. Every unit increase in sodium also produces a decrease in rating of -0.06 [95% CI: (-0.07, -0.05)], holding all other variables constant, which seems rather small when compared to fat and sugar. Finally we see that for every increase in unit of our healthy variable fiber, increases the rating by 2.86 [95%

CI: (2.63, 3.09)] when holding all else constant.

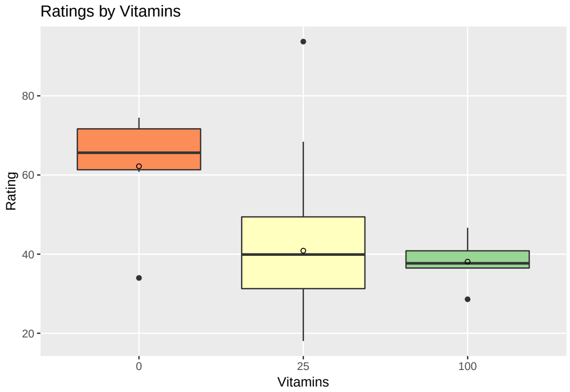
5B CATEGORICAL IMPACT ON RATING



Using the plot above we can observe that most Manufacturers have a similar distribution of ratings, but the **Nabisco** group appears to have the highest ratings overall. In addition, we can observe one cereal from the **Kelloggs** group having the highest rating overall and therefore boosting the average ratings for the entire group.

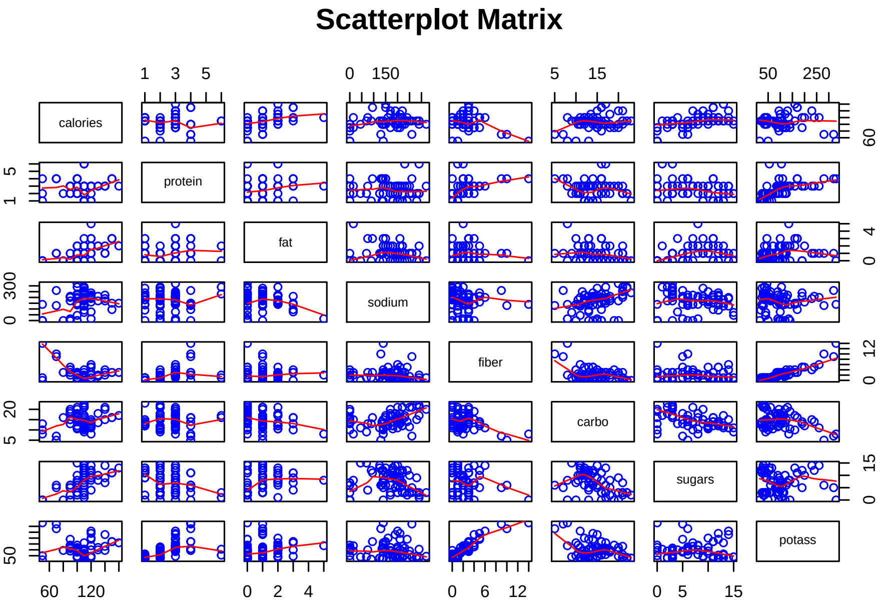


By inspection of the shelf location and their impact on ratings, we can observe that they are all mostly uniform in distribution with the exception of shelf 2 having lower ratings.

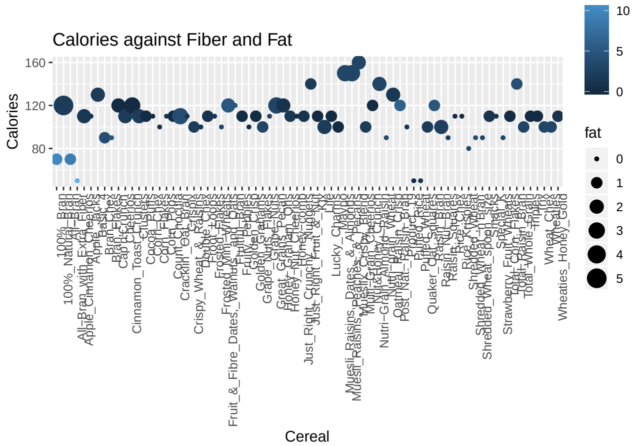


By inspection of vitamins and their relationship to rating, it can be observed that there is a negative trend with rating and the percentage of vitamins in a cereal. It should be noted that most cereals in the data set have 25 percent of the vitamins present, which may contribute to the visual skewing observed.

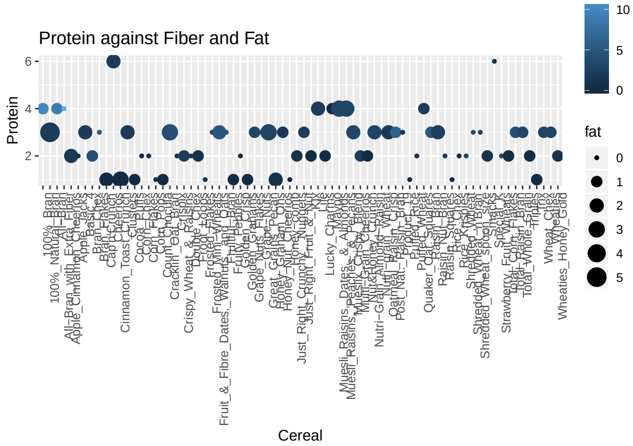
6 RELATIONSHIPS BETWEEN CONTINUOUS



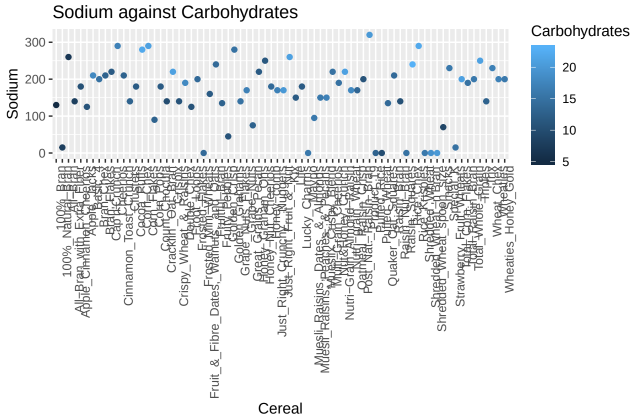
Above is a simple scatter plot matrix that displays the relationship between our continuous variables. Initially, we can observe that the variable Calories has a strong positive relationship with Fat and a strong negative relationship with Fiber. Potassium and fiber also seem to have a strong positive relationship, where one variable increases seemingly at the same rate as the other does. There seems to be an interesting relationship between calories and sugar as it seems that as sugar increases, calories do not necessarily increase. This seems almost counter intuitive. There also seems to be a negative relationship between carbohydrates and sodium.



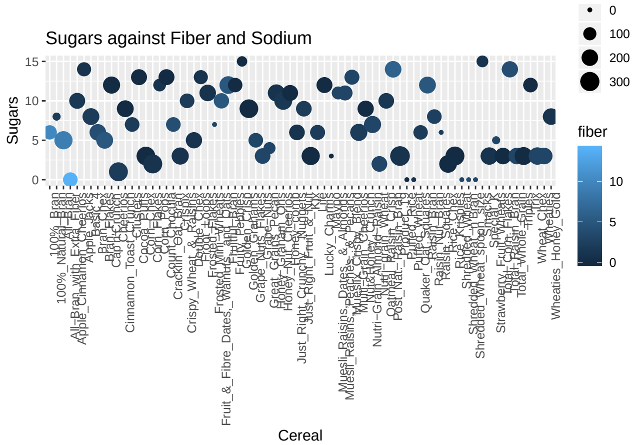
In this plot, We can observe the cereals with the highest calorie count also having a large amount of fat and minimal Fiber. These are designated by the larger dark circles.



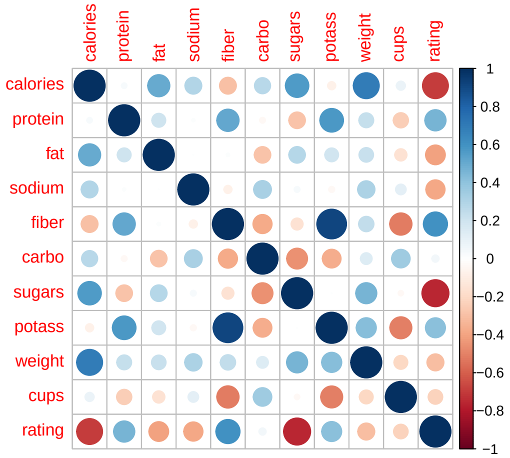
Protein appears to have the strongest relationships with Fat and Fiber. Above is a more detailed plot to display this relationship. Here, The highest protein value is designated by a large dark circle on the left and a smaller dark circle on the right. These indicate both cereals have minimal fiber, which is surprising. However, reducing fiber to the 4 count tier produces several cereals with high fiber and low fat.



Sodium appears to have a strong relationship with Carbohydrates, which we can examine in the plot above. We can clearly observe the cereals with a higher sodium count also have a higher carbohydrate count, establishing a relationship between "Unhealthy" variables.



Sugars appear to have a strongly negative relationship with Fiber and Sodium in the plot above. We can see the highest sugar content cereals do indeed have minimal fiber and sodium. A zero-sugar cereal located in the lower left has a large amount of fiber, indicating a possible outlier in the data set.

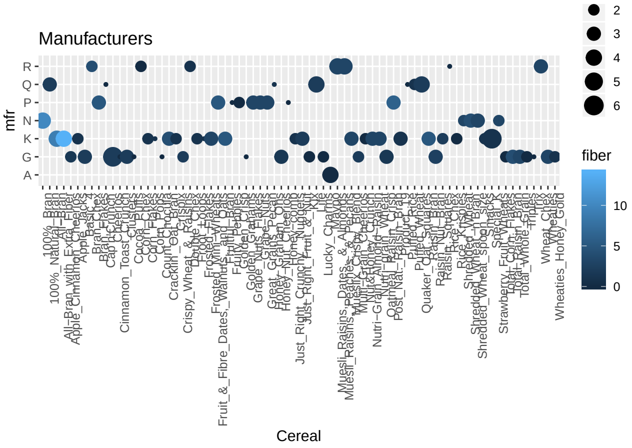


Finally, we will present a correlation matrix above that gives a further emphasis on the relationship between our continuous variables. From this correlation matrix, we can observe the previous claims found in the plots to be reestablished in the simple correlation matrix. Rating has a strong negative correlation with calories, as calories decrease so does the rating. Potassium has a large positive correlation with fiber. The relationships between sugars and calories, potassium and protein, and rating and fiber all have relatively larger positive correlation with one another compared to the rest of the correlations represented in this plot. Similarly, sugars, fiber, and potassium all have a relatively strong negative correlation with carbohydrates compared to the other variable correlations represented above.

6B RELATIONSHIPS BETWEEN CONTINUOUS AND CATEGORICAL

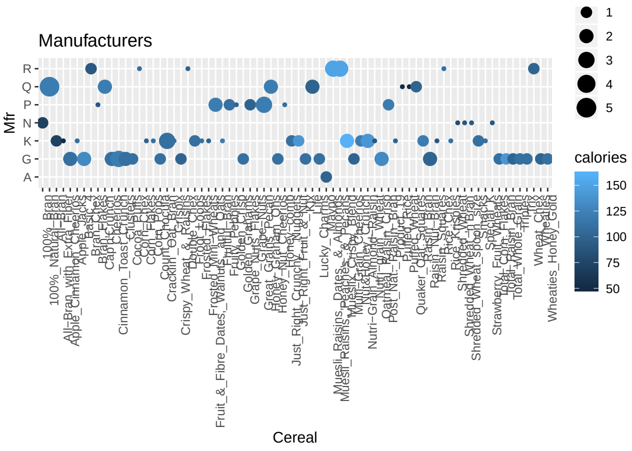
Manufacturer

Healthy Variables and Manufacturer



As previously seen, we can observe that **Kelloggs** and **Nabisco** contain cereals that have high Protein and Fiber counts.

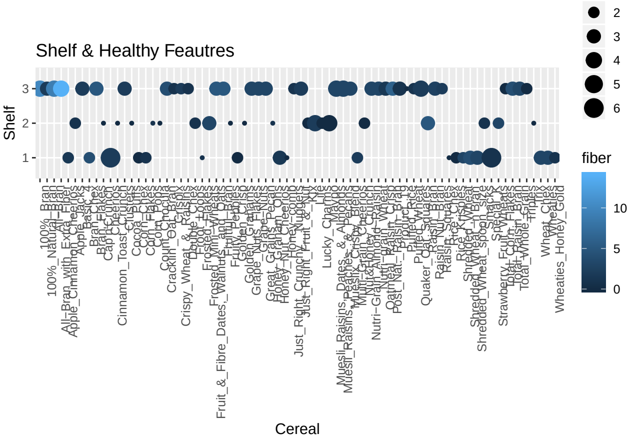
Unhealthy Variables and Manufacturers



Now focusing on unhealthy variables Calories and Fat, we can observe unhealthy cereals as the larger light circles. These designated cereals mainly belong to group **Ralston Purina**, but **Kellogs** also contains a few cereals with high calorie counts.

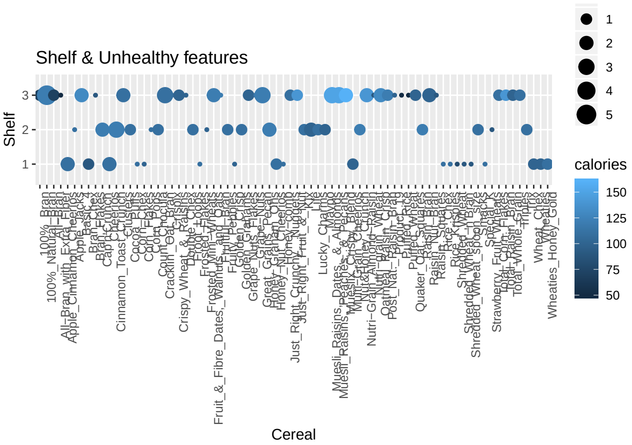
Shelf

Healthy Variables and Shelf



Here, the healthier (large and light) cereals can be located on Shelf 3, while the most unhealthy cereals are typically found on Shelf 2.

Unhealthy Variables and Shelf



It is interesting to note that Shelf 3 once again appears to contain many cereals with high calorie and fat count. Do not select cereals from shelf 3 blindly!

7 ADDITIONAL VARIABLES

When considering the health rating for a cereal, it is becoming and more common for **Gluten Free** to be a desired option. Although this would simply be a binary variable, we could still be able to observe a trend of gluten free cereals having a higher consumer rating.

It is also common for healthier food options to also be more expensive. While we can consider the healthiest foods in a store to be the most expensive, there may be a optimal price point that allows for healthy options to be available for a larger demographic of consumers.

Another variable that might be interesting to take into consideration is the number of ingredients. It has been said that food items that contain fewer ingredients tend to be healthier as they most likely contain fewer preservatives, artificial flavoring, colors, and other chemicals. It would be interesting to see if the number of ingredients is correlated with any of our predictor variables and if it had a relationship with ratings.

8 IRRELEVANT VARIABLES

As previously observed, the variable *TYPE* consisted of nearly all cold cereals with barely any hot cereals.This variable merely seemed to establish the point that most cereals are cold, but did not provide any insight into how the ratings were developed.

Continuing with this discussion, we also previously observed the variables **Cups and Weight** to have very little variation and merely represent the Industry standards for cereal production and manufacturer suggested serving size.

9 OUTLIERS

Continuing on with the previous discussion in problem **4**, we have discussed the anomaly pair of cereals with high protein yet low ratings. It was discussed that the high sodium levels contributed to the low rating, which caused the protein level to be disregarded in total rating.

One other outlier to take into consideration is the only cereal to have a rating above 80: All Bran With Extra Fiber. We will take a look at the features for this cereal in order to understand how such a high rating was determined:

name	mfr	calories	protein	fat	sodium	fiber	carbo	sugars	potass	rating
All-Bran_with_Extra_Fiber	K	50	4	0	140	14	8	0	330	93.70491

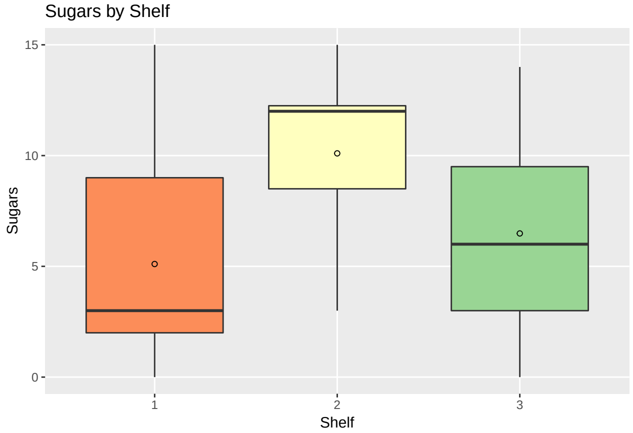
This cereal clearly has a high amount of Fiber combined with 0 fat and sugar. By inspection of other cereals with 0 fat and sugar but lower fiber, we observe ratings ranging from 60 -75. This is still a quality rating, but it is clear that the extremely high Fiber count contributes to the higher rating for this cereal.

Next we can look at observational model outliers by using Cook's Distance on the original model for Consumer Ratings. Cook's distance computes how far on average predicted values change if a specific observation is dropped from the data set. The cooks.distance function used here specifically uses an algorithm that captures information about predicted values using a distance measure that is a combination of leverage and each value of the data set. Higher cook's distances indicate outliers. Using the criteria mentioned, Cheerios, Golden Crisp, and Special K seem to be the outliers of the original model.

```
## [1] "Cheerios"     "Golden_Crisp" "Special_K"
```

10 Do you have any suggestions for the consumers of cereals that might have been missing from the report?

My first and foremost piece of advice for the consumers of cereals that are missing from our data set is to avoid the cereals on the second shelf if they are trying to avoid sugar and stay healthy! As we can see from the plot below, those cereals that reside on the second shelf tend to have a higher sugar content than the rest and therefore a lower rating (as seen in the ratings vs shelf plot from 5B).



If we could set up a consulting booth in the cereal isle to help these consumers in person, as we are now a consulting agency we could use the reduced model for the prediction of Consumer Ratings to help determine the rating of their selected cereal as it still explains a majority of the variance. For example, a "healthy" shopper wants advice on purchasing a cereal that contained 1 gram of fat, 15 milligrams of sodium, 10 grams of fiber, and 2 grams of sugar. With a rating of 82.7288628 we would say it might be a little bland but head to the check out line! Now let's say a mom wants to know if a certain cereal her child picks out at least has a score of 60 or greater. With 2 gram of fat, 200 milligrams of sodium, 3 grams of fiber, and 10 grams of sugar it ends up with a score of 33.7199561, unfortunately it does not make the cut. Maybe her next plan of action is to tell her child they can pick a cereal from any shelf besides the second.

The final most general piece of advice would be how to select a "healthy" cereal if the store wasn't lucky enough to have a cereal consulting booth. Based on our model the four variables to keep in mind when selecting a cereal are fat, sodium, fiber, and sugars. Looking back at our original analysis of the variables, it would be important to select a cereal with a fat content below 2 grams, a sodium count close to zero, a fiber count above 7, and a sugar count as close to zero as possible in order for the cereal to have a high rating and therefore be relatively “healthy”. Remembering our analysis of rating and other variables grouped by manufacturer, selecting a cereal made by Nabisco might also aid in the selection of a healthy cereal.

12 SAMPLING SCHEME

After researching some of the cereals in our data set, it is clear not all of them are still offered in most supermarkets. Our first step in a sampling scheme to verify the information in the data set would be to compare this list to what is currently available at the local supermarket. If 75% of the cereals are still in production that would leave us with approximately 57 cereals to sample from. Randomly sampling 50 of these remaining cereals would provide us with 74% of the original data, which seems like it would provide a sufficient enough representation of our original data set. We could then verify the information from the data set against the cereals from our sample and maybe even update the ratings from Consumer Reports to see if they are still the same. The major issue with the example is sitting in the cereal isle with a pile of cereals and trying to record all of the pertinent information.