Cereal Report

Cody Frisby

3/18/2017

I have been asked to report my findings on the 50 cereals and the variables that the agency has analyzed: *calories*, *protein*, *fat*, *sodium*, *fiber*, *carbohydrates*, *sugar*, and *potassium*.

I start by looking at summary statistics of the data. Additionaly, a method we can apply that can help in identifying varibles who have considerable correlation and also in identifying observations that are possible extremes, so as to further our exploratory investigation of the data, is the use of the bivariate box plot. Like the univariate box plot, this method helps us identify potential "outliers", or observations that are extremes. For our data we have $\binom{8}{2} = 28$ plots since there are 8 variables. Using this method and identifying the outliers on each of those 28 plots the most frequent cereals taht are "outliers" are summarized in the table below.

Table 1: Outliers

Brand	frequency
AllBran	17
QUAKER2	13
PuffedRice	10
PuffedWheat	8
RaisinBran	8
Cheerios	7
FruitfulBran	7

This table includes the name of the cereal and the number of times it was an outlier. Figure 1 shows one of the plots out of 28 to illustrate the correlation of two of the variables, those with the largest correlation, $cor(fiber, potassium) = 0.8671198$. We can clearly see that **ALLBRAN** is way out to the top and right. If we are to move forward with further analysis we may want to consider excluding this ceral and perhaps one or two others that are requent outliers. It can be clearly seen from figure 2 that **ALLBRAN** is an extreme for the variable *fiber* AND *potassium*, **PuffedRice** and **PuffedWheat** are extremes for the variable *calories* (perhaps this makes sense since those cerals are mostly air?), and **QUAKER2** appears to be an extreme for the variable *Carbohydrates*.

Figure 1 also illustrates the possibility of some clustering of the cereals around certain values. Investigating further it appears that almost half, $\frac{22}{50}$, of the cereals have a reported 110 calories. Is this due to the manufacturers controlling serving size so that each serving hits a certain amount? Or could it be an issue with our measurement? Perhaps this warrants further investigation from the board.

What kind of questions do we have about the data? Do we wish to know which of the observed variables are the most influential in predicting a cereals calories? Are there certain natural groupings of the variables that we wish to explore the relationships among? **Are there cereals that are alike?** What kind of assumptions must we make about the data in order to use statistical methods like regression or principal components analysis? We need to keep these things in mind as we proceed with an analysis of the data.

One sanity check on the data that could be performed is to see if the *Calories* reported makes sense **given** the reported grams of protein, fat, and carbs. For example, it is well known that there are 9 calories per gram of fat and 4 calories per gram of protein and carbs. Assuming that *fat*, *protein*, and *carbohydrates* are in grams per serving units I proceed with the following calculation.

$$\hat{calories} = 4Protein + 9Fat + 4Carbohydrates$$

Table 2 displays the discrepencies between our calculated calories and the reported ones ($delta = Calories - \hat{cals}$) ordered by largest *delta* on top and excluding those that don't have a *delta* greater than 10. As you can see from table 2, we have quite a few cereals whose *delta* is greater than 10. We may need to go back and check our measurements. Either our *Calorie* value is incorrect or our values for grams of *Protein*, *Fat*, and/or *Carbohydrates*are wrong. Those with the largest negative deltas are of most concern to me. This would mean that either the company is under-reporting the calories or there are concerning issues with our data gathering (measuring or other) procedures. Either way, I might consider excluding these observations from further analysis while we check on them.

Table 2: Discrepancies

	Calories	cals_hat	delta
CerealCheerios	120	46.12	73.88
USCommodityCornRice	110	169.83	-59.83
QuakerOatmealraisens	137	187.42	-50.42
QUAKER2	162	211.19	-49.19
QuakerOatmealfrut	135	183.00	-48.00
RALSTONEnrichedWheat	130	171.80	-41.80
CreamOfWheat	130	149.56	-19.56
CerealsCornYellow	107	122.20	-15.20
SpecialK	110	98.20	11.80

Above I posed the question "**Are there cereals that are alike?**". I explore that question using principal component analysis. For the uninitiated, principal component analysis is a technique that attempts to summarise and/or visualize the given data into fewer dimensions, or variables. Each principal component is a linear combination of the included variables, for our case there are 8. Using this technique we would in turn derive 8 principal components, but not all need be considered important or significant. If there are few that are important than principal component analysis is more useful that simply using the original data since we can then summarize relationships and/or groups of the variables with fewer dimensions.

Figure 3 displays our derived principal components. The plot on the left we would like to see a sudden bend or elbow, which we really are not seeing. The right plot shows the accumulated variance as a function of the number of principal components. Ideally, you'd like to hit 80-90% by component 2 or 3. Here we hit 91.7 by component 5. Perhaps we can reduce the dimensions of our data from 8 down to 5.

Our first principal component might be thought of as the most "healthy" cereals considering cereals that are high in both protein and potassium (see figure 3). These are those that fall to the right and outside the inner ellipse. **PuffedWheat** and **PuffedRice** appear to be grouped high and to the left. This can be thought of as those cereals that are't high in *fiber* or *potassium* and low, or have negative association, with the other variables, particularly low *calories* and *sugar*. If we desired a ceral that had low sugar content and lots of fiber and potassium (perhaps a more healthy choice?) then we may be able to look to the right and up.

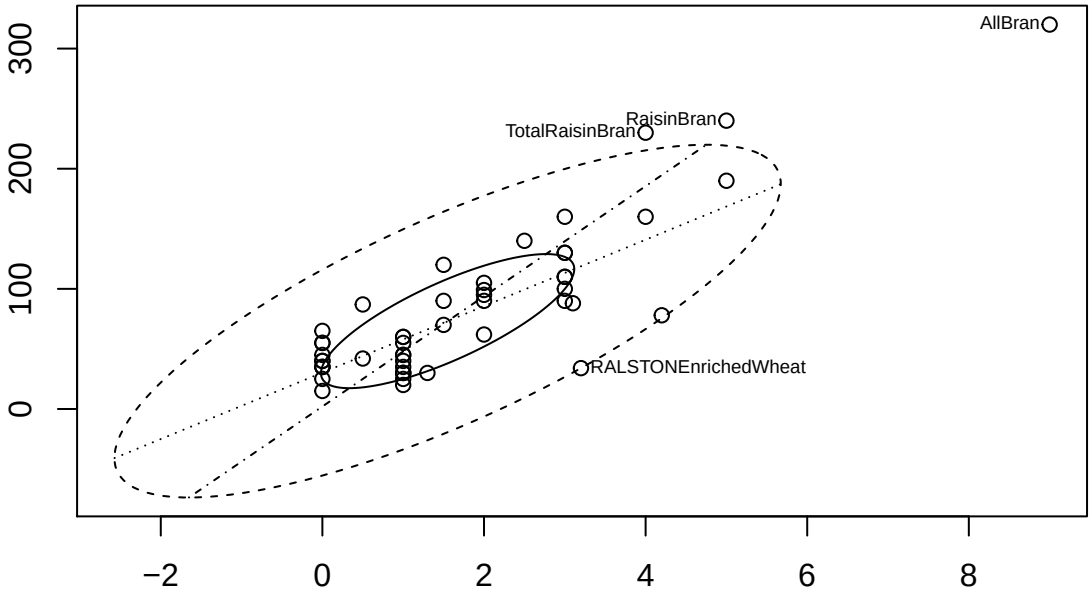

Figure 1: Fiber vs. Potassium

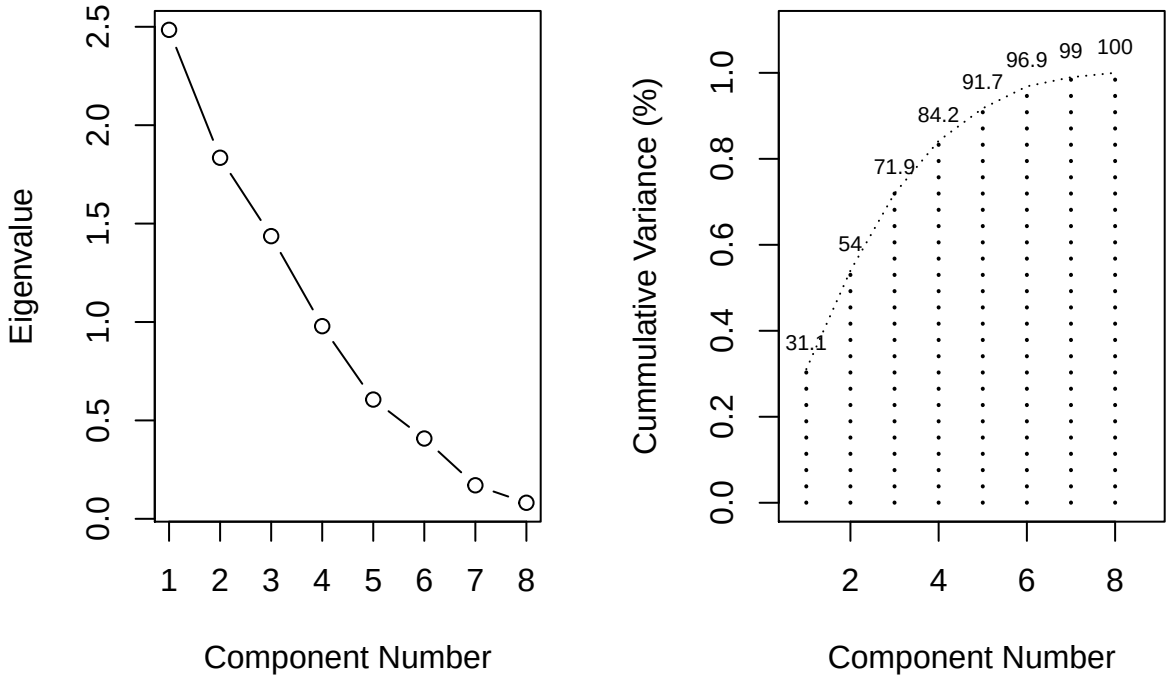

Figure 2: Principal Component Summaries

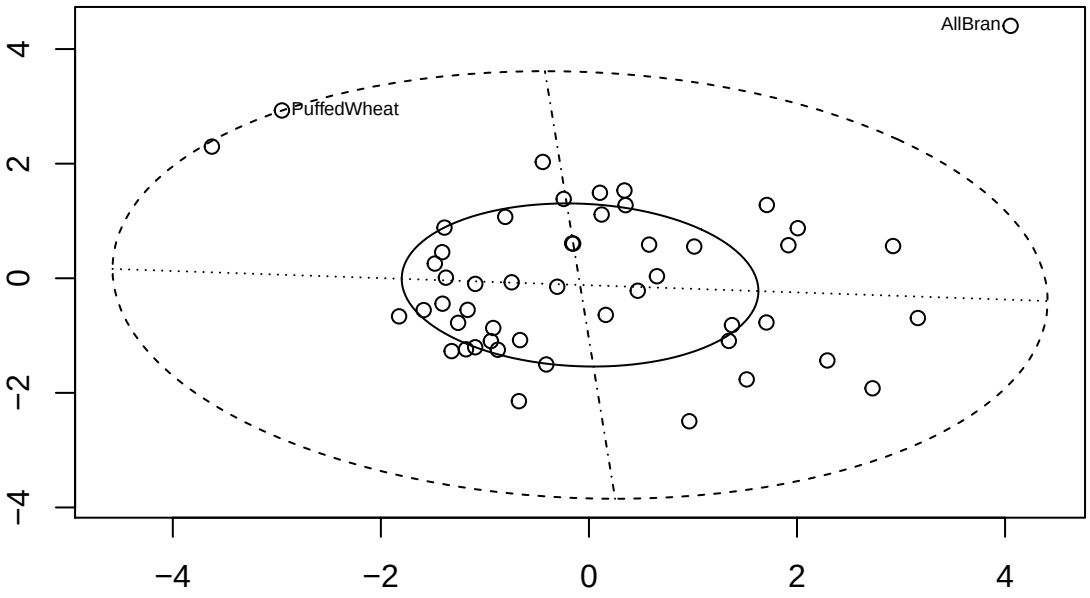

Figure 3: PC1 vs. PC2