**Doug Marcum - DSC 550 Original Analysis Case Study Final Documentation**

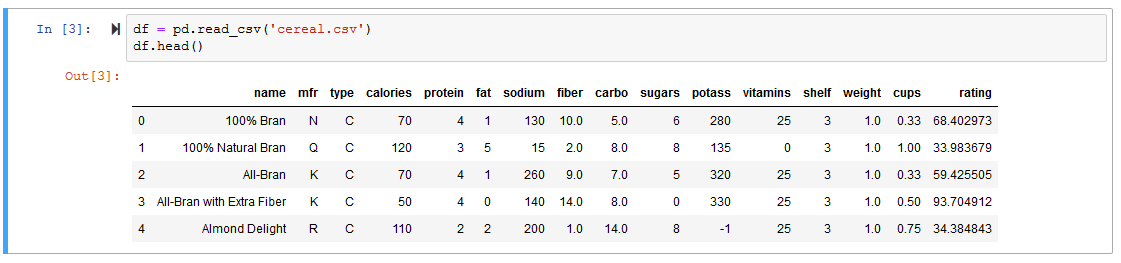
**Case Study: Analyze data to evaluate cereal information to predict consumer rating**

This course has occurred at a time of global high stress, so this project is to be something lighthearted. I reviewed many medical, political, and sports focused datasets that all just seemed to serious. Luckily, I stumbled on the topic of cereal, an interesting and somewhat silly topic. The data for this case study is a list of numerous brands of cereal by multiple manufactures. While not all encompassing, the data is comprehensive enough to predict how a cereal might score with a consumer based on several variables/features.

**Step by Step Graph Analysis:**

## **Part 1**

**Data Source:** <https://www.kaggle.com/crawford/80-cereals>

**1. The originally data can be found on Kaggle.com. Load downloaded CSV file. Display first five rows to make certain data loaded properly and explore columns.**

**2. Variables\*** (See below for complete list)

**3. Display dimensions and information of data frame**

Shape: (77, 16)

Info:

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 77 entries, 0 to 76

Data columns (total 16 columns):

name 77 non-null object

mfr 77 non-null object

type 77 non-null object

calories 77 non-null int64

protein 77 non-null int64

fat 77 non-null int64

sodium 77 non-null int64

fiber 77 non-null float64

carbo 77 non-null float64

sugars 77 non-null int64

potass 77 non-null int64

vitamins 77 non-null int64

shelf 77 non-null int64

weight 77 non-null float64

cups 77 non-null float64

rating 77 non-null float64

dtypes: float64(5), int64(8), object(3)

memory usage: 9.8+ KB

None

**4. Check for missing values**

0 missing values for name

0 missing values for mfr

0 missing values for type

0 missing values for calories

0 missing values for protein

0 missing values for fat

0 missing values for sodium

0 missing values for fiber

0 missing values for carbo

0 missing values for sugars

0 missing values for potass

0 missing values for vitamins

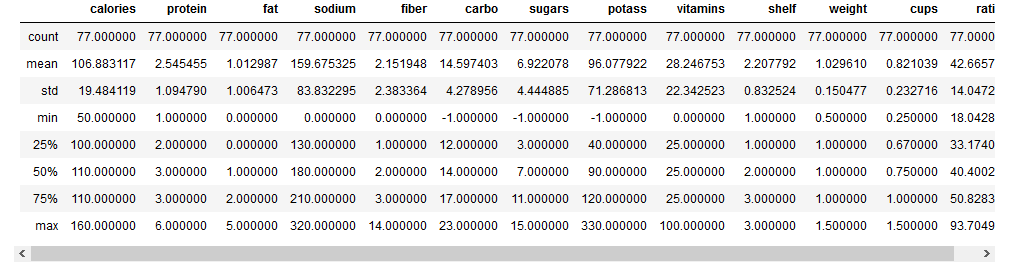
0 missing values for shelf

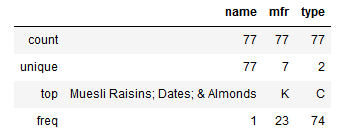
0 missing values for weight

0 missing values for cups

0 missing values for rating

**5. Run summary information on data (total, mean, min, max, freq, unique, etc.)**

****

****

**6. Fields showing -1 are not possible values for sugar, carbohydrates, and potassium. It is believed that these values may have been entered to replace previous NaN values. Since 0 is an acceptable value for these columns, cereals with -1 values are dropped.**

7. Create 'score' variable, grouping cereals into 3 scoring possibilities

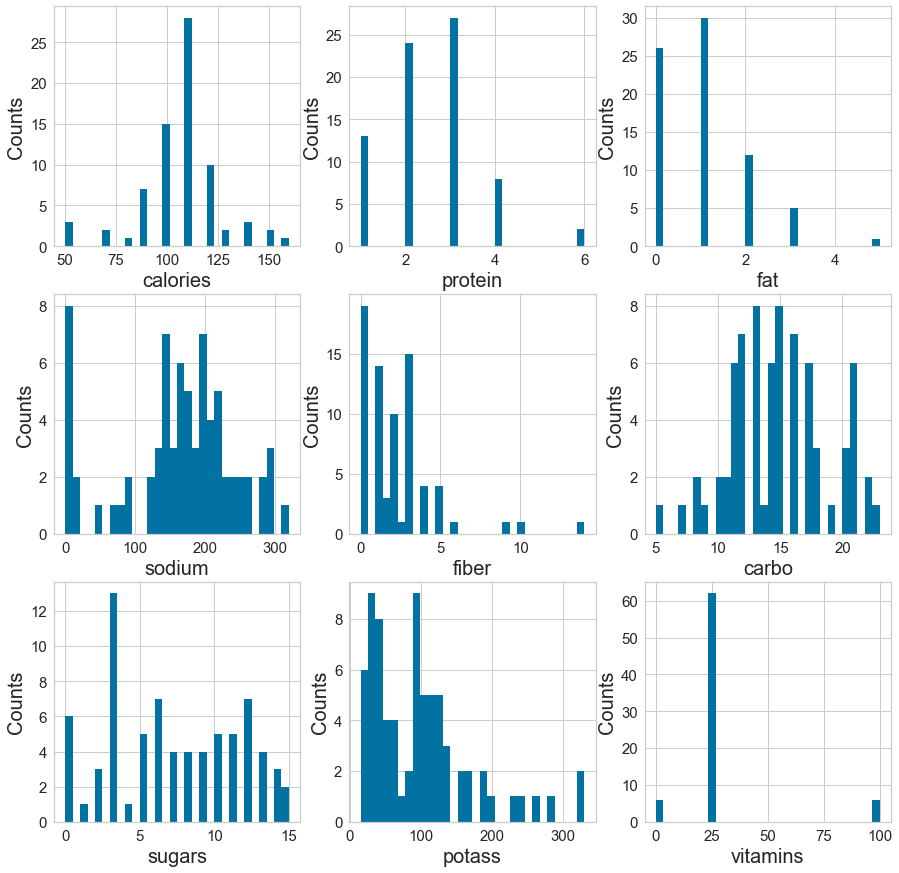
2 = Cereals with a consumer rating in the 75th or higher quantile

1 = Cereals with a consumer rating in the 25th - 75th quantile

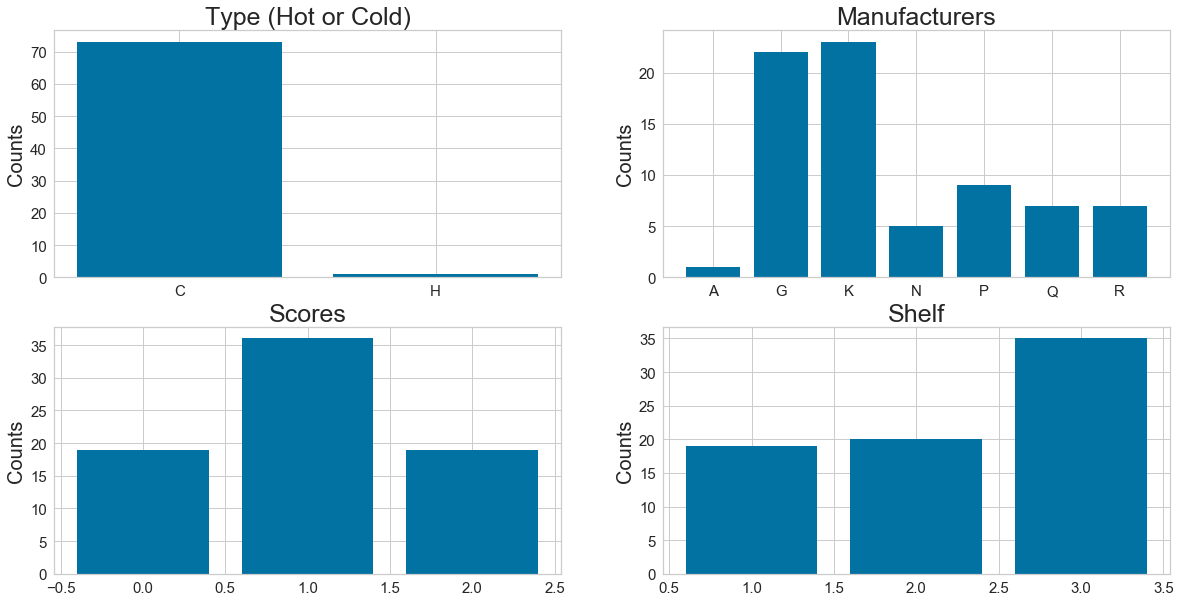
0 = Cereals with a consumer rating in the 25th or lower quantile

From this, 'score' is the target and the other variables will be the features.

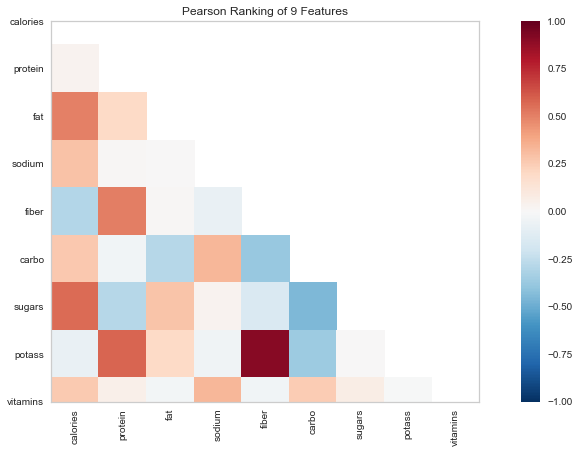
**8. Plot histograms to display and explore data**



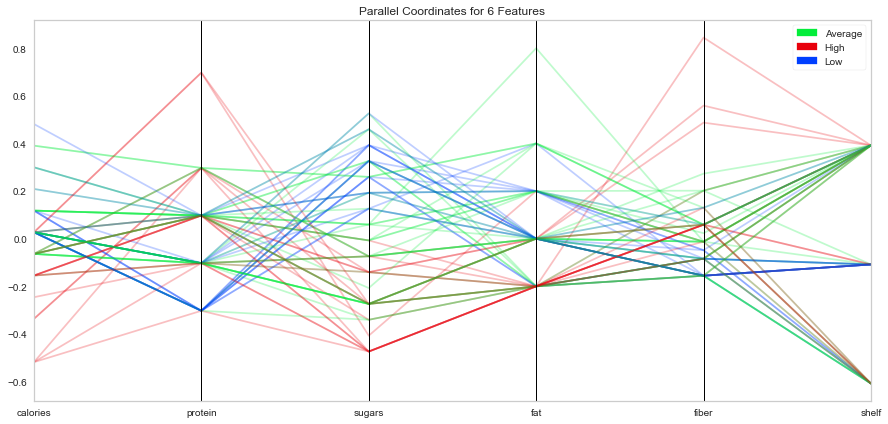
**9. Plot box charts for data with fewer variables**



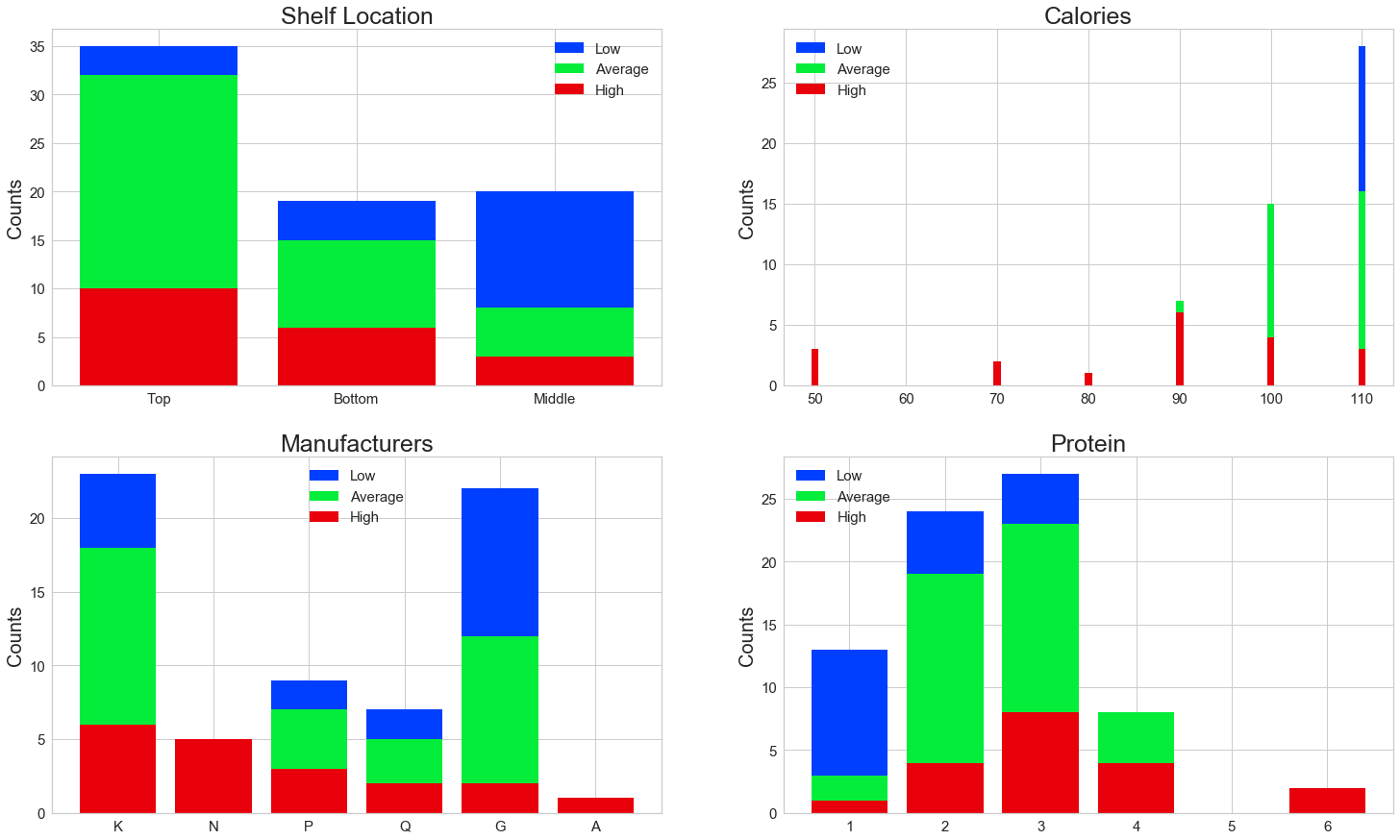
**10. Use Pearson Ranking charts to see what data is correlated**

****

**11. Parallel Coordinates visualization is used to compare the distributions of numerical variables between cereals with high, average, and low scores.**

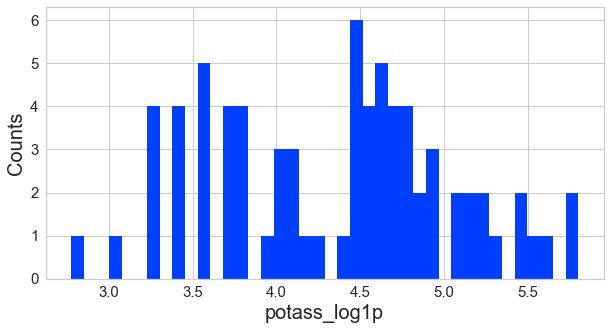
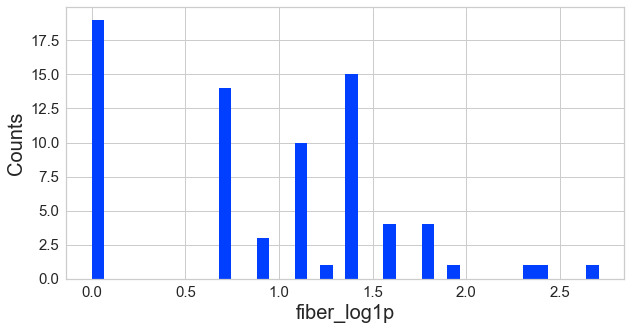
****

**12. Stack Bar Charts are used to compare cereals of which scored high, average, and low based on the other variables.**

****

## **Part 2**

**13. Transform 'fiber' and 'potass’ since both were skewed heavily in previous observation.**

****

**14. Perform feature reduction using LassoCV.**

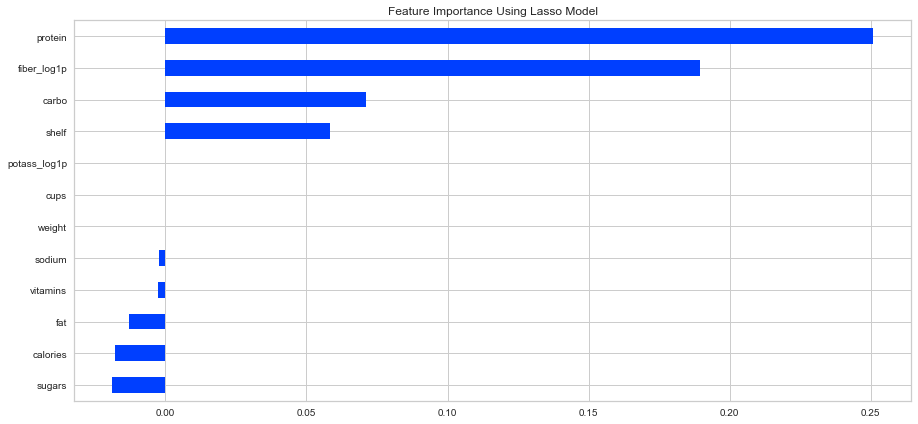
Best alpha using LassoCV: 0.020203

Best score using LassoCV: 0.825267

Total number of features before elimination: 12

Number of features eliminated: 3

Number of features remaining: 9

****

**15. Convert categorical variables ('mfr', 'type') to numeric.**

mfr\_A mfr\_G mfr\_K mfr\_N mfr\_P mfr\_Q mfr\_R type\_C type\_H

0 0 0 0 1 0 0 0 1 0

1 0 0 0 0 0 1 0 1 0

2 0 0 1 0 0 0 0 1 0

3 0 0 1 0 0 0 0 1 0

5 0 1 0 0 0 0 0 1 0

## **Part 3**

**16. Training - Data is split into two sets: Training and Testing. (Testing Size of 30%)**

Samples in training set: 51

Samples in testing set: 23

Counts of High, Average, and Low scores in the training set:

Average 26

Low 13

High 12

Name: scores, dtype: int64

Counts of High, Average, and Low scores in the validation set:

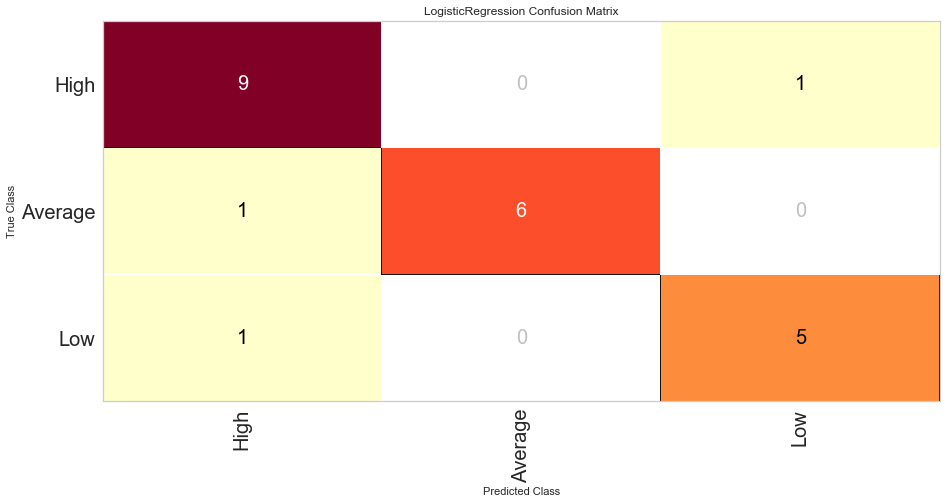
Average 10

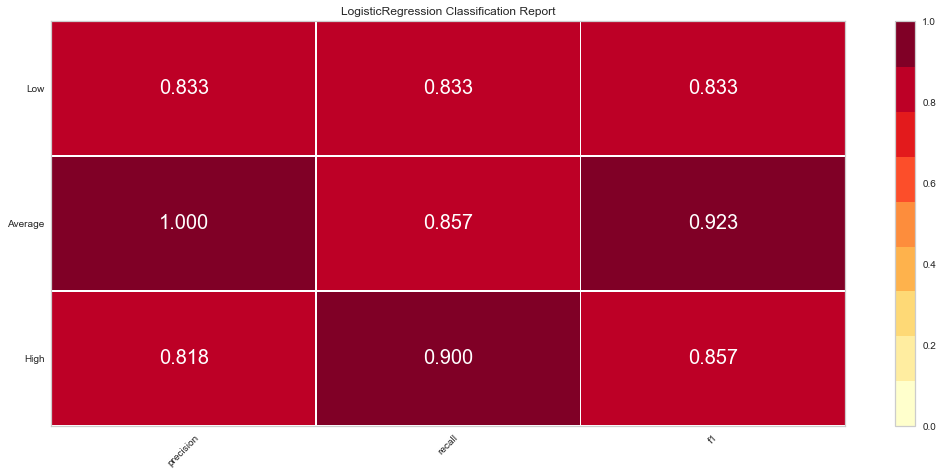
High 7

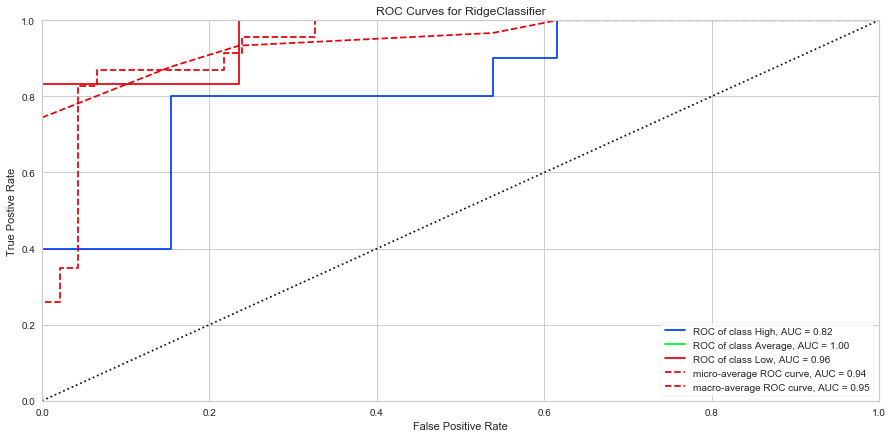
Low 6

Name: scores, dtype: int64

**17. Evaluation of logistic regression model using Confusion Matrix, Precisions, Recall, and F1 Score. ROC Curve evaluation was performed using Ridge regression. This classifier first converts the target values into {-1, 1} and then treats the problem as a regression task (multi-output regression in the multiclass case).**

****

****

****

**Conclusion**

**While the data set was limited in size, some interesting insights were still obtained through the analysis.**

1. **Cereal with the highest caloric values tend to score low, while those with higher scores tend to have lowest caloric values.**
2. **Cereals with high in grams of protein per serving scored higher than those with low grams of protein per serving.**
3. **In terms of product placement on grocery shelves, cereals with the lowest scores tend to be placed in the middle shelves. These shelves are typically at eye level and also easily seen by children.**

**In constructing the model and the reviewing the results, the features determined to have the most importance allowed for successful model. For each category, precision, recall, and F1 scores exceeded 0.81 and scored higher than 0.9 in a number of categories.**

**\*Variables**

name: name of cereal

mfr: manufacturer of cereal

A = American Home Food Products

G = General Mills

K = Kellogg’s

N = Nabisco

P = Post

Q = Quaker Oats

R = Ralston Purina

type:

C = cold

H = hot

calories: calories per serving

protein: grams of protein

fat: grams of fat

sodium: milligrams of sodium

fiber: grams of dietary fiber

carbo: grams of complex carbohydrates

sugars: grams of sugars

potass: milligrams of potassium

vitamins: vitamins and minerals - 0, 25, or 100, indicating the typical percentage of FDA recommended

shelf: display shelf (1 = bottom, 2 = middle, or 3 = top, counting from the floor)

weight: weight in ounces of one serving

cups: number of cups in one serving

rating: rating of cereals from Consumer Reports