**CS 4460 ASSIGNMENT 1**

**Name: CHANG, Yingshan**

**Student ID: 903457645**

1. Five questions

* What is the meaning of “-1” in some of the columns such as “cup”, “sugar” and “carbohydrates”? Does it mean “data not available”?
* What should be the unit for each numeric column? Do they represent “density” or “total amount”?
* In the “manufacture” column, there is only one instance with manufacture “A”. This might not be enough to well-represent the feature of all cereal produced by A.
* For all columns filled by numbers, which ones should be considered as “numeric” and which ones should be considered as “categorical”, since numbers can represent not only the amount of something, but also the index or class of something?
* Judging from the first column, some food other than cereal, such as fruits, raisin and nuts, are added to some cereal products while other cereal products are whole-grain. Thus, the data in other columns may not represent the property of the cereal but a combination of cereal and other food. Would this lead to inconsistency across different rows?

1. Five insights

* Previously I thought “calorie” might be positively associated with “carbohydrates” and “sugar”. While it is quite notable that “high calorie” is probably associated with “high sugar”, the association between calorie and carbohydrates is not that clear.
* Since carbohydrates and sugar always come into our mind together, they might represent similar properties of the cereal. However, it seems that cereal products containing more carbohydrates tend to have less sugar.
* Products with higher purity of “bran” or “wheat” might consist of more fiber.
* At first sight of “shelf”, I guess it might be the index of shelf in the supermarkets that a certain cereal product was placed. Then there might be some reasons for the arrangement. i.e., cereal products on the same shelf may share some similarities. However, I did not discover any convincible reason from the given dataset. It is also possible that they are simply placed onto different shelves according to their brands.
* When it comes to the first column, it is interesting to note that some entries basically illustrate the components of the cereal product, while other entries might simply represent the brand and give no ideas about what the product contains. Further, it cannot be decided yet the source of the first column of data. There might be some inconsistency if some entries in the first column are brand names while others are product names.

1. My procedure of data exploration and analysis

I loaded the data into Excel where the dataset can be viewed as a table, with the first row indicating the name of each column. After recognizing what each column represents, I began to find ways of exploring characteristics which cannot be discovered intuitively. Basically, I used two functions provided by Excel: average and sort. Hoping to understand the difference between different manufactures and types, I used AVERAGE function to compute the average of some values corresponding to a particular manufacture or type. After that, I used SORT to detect possible association between two columns. For example, I sorted “carbohydrates” in ascending order and looked at “sugar” in order to decide whether or not there is an approximately ascending or descending trend. During this process, I also took notes about aspects of the dataset that might lead to confusions and difficulties of analyzing data in this way

1. Challenges of doing data analysis in this way

In the process of doing this assignment, I found three main challenges to analyze data without visualization tools. First of all, while there are 78 rows and 15 columns, I can only focus locally on a small region of 3-4 entries at the same time when I was scanning though the dataset only by my eyes. It was really hard to capture a global picture of this dataset because as my visual focus changing, I tend to forget what I previously saw and had to go back again to check. Secondly, so many numbers are very annoying because they are not arranged in order. Even though the sorting function of Excel helped me a lot, it can only keep one column sorted at a time while all others are still in a mess. Moreover, this method of data analysis cannot scale up when it comes to huge datasets. Many problems have already emerged when I was only asked to explore a table with 78 rows. In real-world problems, we may encounter gigabytes of data. I cannot imagine what can we do without visualization tools. Overall, without visualization tools, data analysis could become a really arduous, time-consuming, unstraightforward and inefficient task.