CS7450: Information Visualization **(Due: 09/14/2020)**

Homework Assignment #1

Instructor: Alex Endert *Name:* Chen-Yang(Jim), Liu *GTID:* 90345****

Question 1:

List (bullet list of items) five "insights", chunks of knowledge, or deeper questions that you either encountered or gained while exploring the data. An insight could be some understanding of the data and its characteristics that is not relatively obvious or intuitive. It is something that most people might not realize initially. Note that an insight or knowledge chunk simply may be a deeper question that arose in your mind while exploring the data, and your analysis may not have been sufficient to answer the question.

- Should we use abbreviated names like Unknown Manufacturers and Types for visualization? This dataset might be from workers who've been familiar with each variable. However, for viewers, they are just a bunch of one characters that are meaningless. Therefore, while visualizing data, we should keep an eye on viewers' background. If we plot just one character, probably people could not understand this visualization. In other words, the goal of data visualization is to let people clearly and intuitively understand what this is for.

- How do we measure cups?
  Normally, we think a cup as a certain amount or use it as know how much we eat. But in the dataset, we could see some cups with non-integer. This is a little confusing because we unlikely know how to calculate 0.67 cups. What's more, we never know how much a cup is from different manufacturers. Is a cup from Quaker same as Rice Chex? We do not know.

- How do we calculate calories?
  Each cereal product somehow overestimate calories. Usually $1g$ of protein, fat and carbonhydrate represents 4, 9 and 4, respectively. However, in each product, it compensate some calories in each product. From the standpoint of statisticians, is this from random error? Accurate calculation seems not true in everyday goods. What's more, if we know total calories, protein and fat in Quaker Oatmeal, we should know how much carbonhydate is in this product. However, in this dataset, it just shows carbonhydate as -1

- Statistics for cereal products
  From the information provided in cereal products, in calorie column, the mean is 106 and the standard deviation is 19. Mean and standard deviation show that products are in certain range and give us a general idea on how data spread.

- Vitamins
  From our experience, each product would list different kinds of Vitamins. But, in this dataset,

we could not compare each product correctly because maybe some products include Vitamin A but some products contain Vitamin B. Those two things are different. If one customer wants to eat something helpful for his eyes, he might choose a cereal product with Vitamin A. Thus, categorized Vitamins help customers better understand what they need.

Question 2:

Write one paragraph about the process you used to do the exploration and analysis. Did you load the data into Excel, work manually, or do both? What did you do in Excel? Did you draw pictures? Just tell me (briefly) what you did.

Since my major is statistics, I just used Excel to derive some useful statistics to catch a blueprint of data. For statisticians, we usually use mean, variance, standard deviation, confidence intervals, quartiles and correlation. I could know central part lies around 106. And, calories of most products are 110. To my surprise, originally, I expected calories would have strong linear relationships among carbonhydrate, fat and protein. Those values are not above 0.5. Therefore, I am a little suspect how they calculate calories for each product.

Question 3:

Write one paragraph about challenges or problems that you encountered in doing the analysis this way. Did anything limit or frustrate you? If nothing did, perhaps there was something that was more difficult than you thought it should be. Nothing is perfect, so you should be able to list some potential issues here.

As an statistician, I usually use data visualization to get a glimpse of how data locate and further determine which model is suitable for this dataset.

- Visualization helps people know more
  I could not use visualization methods to see whether there are outliers which affect model accuracy. I do not have much time to remove outliers manually.

- Where is product difference?
  If we are given a dataset, we might try to know each example has what kind of information. Are they similar or distinct? Each product only list their own values, but it is hard to discern difference and similarity.

- What is scales for each column?
  Only a number is represented in each column. We could not know that it is gram, mm or lbs. Easily, we will think sodium has a large amount in each product, but actually it just a scale of mg. We inclined to assume each column is used in the same scale, but actually it is not.