Phuong Ngo

INF550

Homework 1

Insights:

- Most brands are high (>100) in calories
- The relationship between calories to cups and calories to brands. I want to know if there's any correlation between those.
- I also want to know the calories/sodium rates per cup or weight. If I am able to extract analysis from that, I might be able to find or predict which brand comes with the most calories/sodium or nutrition when compared to overall size of the same pack. For example: the for 1 cup of two different brands, one might be more packed with calories and sodium, hence not as healthy as the other brand
- Another question would be whether hot cereal is healthier or worse than cold cereal
- There is one class that isn't being fully utilized, which is the shelf class. I wonder if that has anything with the shelf position of each products at supermarket? If it is, then with the nutritional data and all, will I be able to find the correlation between shelf position and brand or nutrition values.

I used the dataset from the excel sheet to work on data exploration and analysis. For most of the numerical data, I tried to find min and max for each variable (calories, sodium, and so on). What I was trying to do was to plot some kind of histogram to show the frequency of all these values. I was able to plot a few histograms based on the frequencies that I found for each

nutritional variables. I also used Excel to help me do the math. The histograms help me with some information analysis of course. But not to the extent that I was expecting.

I was expecting it to be easier, but having to do it without the help of any analytical tool was hard, frustrating and time consuming. I am not used to using Excel to do manipulate and explore data. I am also a bit clueless since I missed two weeks of class due to registration problems. So despite reading all the class materials prior to doing this, I still found it quite taxing to do the homework even though I know it is not hard. To be honest, the process is too long. Between trying to slice the data by myself, calculating the numbers (even with the help of Excel) and trying to create some types of graphs so that I can see the data more clearly, it was clearly too much to do it manually. Moreover, doing this has its own limitations just as I stated about about finding the correlations or making assumptions and predictions about the dataset.