**Capstone Project 1 Data Wrangling**

**Data Exploration:**

The dataset I will be using comes from Kaggle. Initially, the csv file provided in the download seemed appropriate for use with the project idea I had in mind. However, upon further examination, the csv turns out to be a reformatted version of the original json file storing the scraped data. The 600+ columns in the csv are a result of parsing unique tags and ingredients from the categories and ingredients data from the json file, respectively. Since the provided json file also includes the recipe directions and measurements, I will be using it as the main source of data which would be supplemented should the csv’s formatting prove useful.

**Data Wrangling:**

Having been originally scraped from Epicurious.com two years ago, this recipes dataset inherits some of the problems present at the time on the website itself. One such issue is the missing nutritional data on roughly 4000 out of the 20,000 total recipes. This presented an opportunity to do further web scraping as a part of data acquisition and wrangling. To this end, I began scraping additional recipes from Epicurious and keeping only ones with all the nutritional data included. Furthermore, carbohydrate seems to have been completely omitted which can be obtained for the entire dataset. This will give the full picture of the macronutrients for each recipe. There is also a very small number of rows with missing values for all columns which were removed altogether.

In preparation for scraping Epicurious again, the name of the recipes need to undergo some simple cleaning. The text are stripped of trailing whitespace and changed into the proper format to produce the search url. This step is necessary since some recipe pages have unique ids. As such, the url for the actual recipe will need to be obtained from the search page. The dataset also has duplicated recipes, some of which are actually duplicated data while others are different versions. Duplicated entries are sorted by rating first, then the top entry with the most info available will be kept and all other duplicates are dropped.

In terms of outliers, the numeric columns calories, fat, protein, and sodium all contain quite a few. These are all contained in the nutritional info section so they naturally paint the same picture when looking for erroneous data. It appears that some recipes provide nutritional info on a per serving basis while others will list contents for the recipe in its entirety. This is shown most prominently by cake/pie recipes as well as stuffing, dressing and various other types of side recipes. This discrepancy will need to be considered since these will most likely be classified as outliers. Otherwise, there are a handful of obvious errors such as calories being in the hundred thousand/ millions range. If possible, these will be replaced by the proper values and otherwise removed altogether. It appears that nutritional data errors are still present in recently scrape recipes. Obviously erroneous values such as infinity for any nutritional data point will be omitted.