For this term’s project. I have selected Nutrition data from 3 different sources which are listed below.

1. Excel file – This file gives the Nutrition details such as calories, carbs, proteins, fiber, net carbs, sugars, cholesterol, and many more of the food items and also the food group that it belongs to.

<https://tools.myfooddata.com/nutrition-facts-database-spreadsheet.php>

1. Table from a website – This table primarily contains the calories, proteins, carbs, fiber, fat and saturated fats.

<https://en.wikipedia.org/wiki/Table_of_food_nutrients>

1. API – This will be used to return results of items based on a searched item and the ingredients in them.

<https://fdc.nal.usda.gov/api-guide.html#bkmk-6>

All the above data sources are related by the item name. The Excel file has one-one relation with the data from website The data from API has one-many relationships with the data from Excel file.

In order to accomplish the milestones, I have to fetch the data from all these sources clean the headers, make sure that they are of the same name across the sources if they are related, join the data between the sources once they are loaded into the tables in sql lite and get meaningful insights from the data. One thing I need to consider is that the serving size in the items from the excel are of 100 gms where as in the table from website and the api , the serving size is different. I have to make sure to convert them before doing a lookup. I have to make sure that there are no duplicates in the data. As the key is the item names, I have to ensure that there are no duplicate items. I need to make sure that the data is valid and that there’s no bad data such as characters in number fields. I have to work on identifying any anomalies in the nutrition information. If inconsistent values are found, they need to be fixed. Fuzzy matching needs to be done to identify similar items. Also, these steps need to be performed across the three data sets as part of the different milestones. In the milestone 4, I have to create a public key and pull data from the api source. The data will be returned in json format and I have to convert that into tabular format.

In the final milestone, the data from these data sets needs to be loaded into the SQL Lite database as different tables. 5 visualizations need to be created using either python libraries or the power BI and must span between at least two tables.