Exploring Nutritional Data

Kristie Kooken

January 15, 2023

Project 3: Exploring Nutritional Data

The purpose of this project is to explore nutritional information and compare calories, fat and sugar (and potentially other indicators) of foods that are prepared in restaurants to foods one could prepare at home and how each compares to Reference Daily Intake (RDI) information.

Sources:

The data sources that I will use for this project are listed below:

- Flat file: nutrition.csv
    - Nutritional values for common foods and products, approximately 8,800 types of food
        - This data source has 77 columns which includes descriptions of the food item, serving size, calories and other nutritional components including fats, minerals, vitamins, amino acids and water.
    - Website: https://www.kaggle.com/datasets/trolukovich/nutritional-values-for-common-foods-and-products
- API: CalorieKing is a site that offers a calorie counting database/app, weight loss assistance and recipes. This site also has API of nutritional information for favorite brands and fast-food restaurants.
    - API is JSON format
        - Mass and volume are in grams and milliliters
        - Nutrient information is presented (22 columns)
        - Brand, name and classification are present (considered Summary information) (4 columns)

        - Specific details about food item (considered Detailed information) (14 columns)
    - https://www.calorieking.com/us/en/developers/food-api/documentation/
- Website: The reference dietary intake (RDI) tables for daily values, vitamins and minerals from Wikipedia
    - Dietary Reference Intakes (DRIs) are a set of reference values used to plan and assess nutrient intakes of healthy people as regulated by the FDA in the US and Health Canada in Canada. On the Wikipedia website, there are three tables of dietary reference intake information
        - 1 Daily Values Table with 2 columns
            - Nutrient and Value
        - 2 Dietary Reference Intake Tables with 3 columns each
            - Nutrient, Old RDI and New RDI
        - I plan to use Daily Values Table with 2 columns for this project
            - Creating a total daily calorie variable from information on this page (it is not included in the tables)
    - Site documentation includes examples of calls
    - Website: https://en.wikipedia.org/wiki/Reference_Daily_Intake

Relationships:

For relationships between the three data sources, I have spent time considering how I could merge these data to use in a meaningful way for exploratory data analysis. The flat file of nutrition has food item names but no formal coding system of the food items. The API from

CalorieKing has an ID field that is a UUID identifying the brand name of a food item which seems very useful but there is no matching field for the flat file. However, the API data also has a category column. At this stage, I am hoping this category column could match merge to food item names in the flat file however, I will have to see if this is feasible. If it is not feasible, I plan to create a grouping variable in the API, for example, all hamburgers would go under Hamburger as a grouping variable then I would merge/join the nutritional data for hamburgers to this group variable (in this example, I would have a one-to-many merge). For the website information, since this information for a daily intake of food, I will merge this first to the flat file onto every record and then derive variables such as a percentage of daily calorie variable for the food items. To do this merge, I would transpose the RDI data and create columns instead of rows, create a key variable like key = 1 in both the flat file as well as the website data and join them together.

This is an example for a food item for my proposed relationships between the data sources, the number of columns is truncated in this example:

Item (.csv)	Calories (.csv)	Item (API)	Calories (API)	Fat (API)	RPI (wiki) Total fat	% fat RDI (derived)
Hamburger	100	McDonalds Big Mac	250	65g	72g	97%
Hamburger	100	Wendy Kid's burger	125	22g	72g	30%
Hamburger	100	Red Robin Hamburger	450	125g	72g	170%

Approach to Project:

The approach I will take for this project is to bring each data source and do as much initial data cleaning and processing as possible. Because there are many variables between the flat file and API source – along with many overlapping variables (for example, both data sources have calories of food items), I will rename variables to unique names and discard variables that I will not use in analysis. Once each data source has been prepped, I will then merge this data as outlined in the previous section and using the tools described in Milestone 5 and then derive at least two variables (percentage of fat from RDI and percentage of calories from RDI). I will then plot the visualizations as described and complete the write-up. The purpose of the write-up is to better understand eating environments (home versus restaurant or potential fast food versus chain sit down restaurant) and how that can change our caloric intake. In terms of challenges, I am a little apprehensive about the API data. It seems like it is larger and I am inexperienced with parsing JSON except in a rudimentary way. I am worried I will either make a very small request and need to a few requests or I will request the full data available and be overwhelmed on how to process the data. Likewise, I have concerns around my proposed method for merging the flat file and API – however I will need to be creative and flexible if my initial proposal does not go as planned. This project will test my python skills which can seem shaky. At the same time, I am interested to improve my skills and do more complicated merging and data transformations. I am also apprehensive that I will need to create many variables, though I have a general idea of what is needed, I am guessing there will be complications that are yet to be discovered.