Notes On Data

Derek R Neilson

September 25, 2024

Abstract

This document contains notes on the data. The notes are intended to demonstrate how I filter and manipulate the data, and are purely for my instructor to review.

1 Introduction

In data analysis, the ability to effectively filter and manipulate data is crucial for extracting meaningful insights. This document outlines the methodologies and tools I employ to pre-process and analyze the dataset. The primary focus is on cleaning the data, handling missing values, and transforming data to suit the analytical objectives. These notes serve as a comprehensive guide for understanding my data processing workflow.

2 Data Collection

The data was collected from `https://fdc.nal.usda.gov/`. The dataset is 2.9GB and is labeled `Branded` and is in JSON format. I chose this dataset because it is large and one can assume that it has the most rows because it is so large.

To download the data, I used the following commands:

baeginlstlisting[caption=Download, Extract, and Remove Dataset File, label=lst:download$_e$*xtract*]*wgethtt*p
//fdc.nal.usda.gov/fdc − *datasets/FoodData*$_C$*entral*$_b$*randed*$_f$*ood*$_j$*son*$_2$024 −
04−18.*zipunzipFoodData*$_C$*entral*$_b$*randed*$_f$*ood*$_j$*son*$_2$024−04−18.*ziprmFoodData*$_C$*entral*$_b$*randed*$_f$*ood*$_j$*son*$_2$02
04 − 18.*zip*

As shown in Listing red1Download, Extract, and Remove Dataset Filelstlisting.1, the commands download, extract, and remove the dataset file.

It is worth noteing that I am using git to track changes in the code and data. The git commands will not be shown in this documen for brevity sake.

3 Data Inspection

I receved the following files after extracting:

- `brandedDownload.json` I am assuming that this is the main file
- `foundationDownload.json` I am assuming that this is a supporting file