CS234 Mini-project: Cleaning Food Data Observations

Overview

In the first week of classes, students in CS 234 collected data about their eating habits by taking photo of their meals before and after eating. By discussing together the theory of "tidy data" we came up with a list of variables to represent the observations. A spreadsheet with the variable names was created and shared with the class. Every student filled out the spreadsheet with their meal information.

These spreadsheets were collected and put together to create a single CSV containing all observations. Every observation starts with the student ID, a string that was assigned randomly to every student, "st1", "st2", "st3", etc. There are in total 13 variables for each observation, here is the list of all of them:

```
SID, date, time, duration, beverage?, caffeine, bev-type-1,
bev-type-2, location, hasveggie, isvegetarian, waste?, nrplates.
```

Despite our efforts for normalization (make the labeling uniform), different errors were introduced in the data. One is the milligram vs. gram in caffeine amount, another is the addition of % for the amount of waste, different date formatting string, etc. In this collaborative class task, you'll write code to clean the data to the best of your abilities and create a new CSV file that can be loaded into pandas to do data exploration in our next step.

Class Activity on Friday 09/29/2017 -- Complete by 10/03/2017

1. Download the CSV file which is part of the foodcleaning.zip folder in the schedule.
2. Create a group of 3 students to work together.
3. For every variable in the CSV look at the data and try to find out inconsistencies in values. Make a list of all the fixes one needs to do for each variable [if necessary].
4. Open the notebook that imports pandas and loads the CSV as a dataframe. Proceed into writing code for cleaning the data and fixing the issues. Try to write code that doesn't use FOR loops, but **uses vectorized operations in pandas**. However, if you cannot, you can also use Python with for loops.
5. Keep working on this over the weekend. For Tuesday, we want to have at least one completely cleaned CSV file to continue with the rest of exploration. Students who want to challenge themselves should use pandas for cleaning.