WDABC Project 2 Alex Miller

Report:

The project was inspired by an article on influenster.com listing most preferred Halloween candy of each state and D.C. in the U.S.

Data Extraction:

Data set 1: Preference List from influenster

<https://www.influenster.com/article/americas-favorite-halloween-candy-state-by-state>

After inspecting the website the list were paragraphs <p>. I considered building a text scraping but because it was a simple list I copied (from the site) and pasted to excel. The list was one column and state and candy were separated by a hyphen/dash. I created a second column using the data separator in excel then converted it to a csv.

Data set 2: Candy Nutrition from popsugar.com

This data is a table <Table> in the html itself. After unsuccessful attempts to use jupyter notebook to scrape the table from the website I simply copy pasted and converted to csv using excel.

<https://www.popsugar.com/fitness/Calories-Halloween-Candy-Fun-Size-Treats-5452936>

Data set 3: Oral Hygiene by state

This site posted a table with data from a cdc survey. I copied and pasted to excel and concerted to csv.

<https://www.toothbrush.org/us-oral-health/>

Data Transformation:

The plan was to read, merge and clean in pandas. I tried many times and many different techniques, and I was unsuccessful. Since I was already behind and catching up on other projects I decided create the final data set using excel. Had this project used a lot more data I would have figured it out with the aid of TAs/ classmates.

In excel I sorted joined state preference with candy nutrition to create st\_can\_nut.csv then I filled in nutritional that didn’t match up due to not being on the candy preference list, with data found online. I then joined the oral health by state and state/candy nutrition set.

Data Loading:

I merged the 2 created set into one combine and then dropped unnecessary columns. To create “combined final.csv” It isn’t an amazing data set nor did I execute in the preferred programs(pandas/postgres) but, it is not an incomplete.