Kevin Paulovici

Bellevue University

DSC 540 Data Preparation

Mid-term Project

For this mid-term project, I decided to use a Pokémon dataset. I chose this dataset because I was fairly familiar with the columns and would be able to relate the data to the required task for this assignment. However, one thing I struggled with while completing the tasks was generally how I wanted to complete it. Since this project was focused on demonstrated our competence we were not required to come up with general questions while completing the tasks. This allowed a lot of variability in the way tasks can be performed.

The following is a summary of how each task was completed.

Replace headers: Since the headers we generally understandable, I decided to replace a commonly occurring header. “against\_” was replaced with “facing\_” which applied to multiple columns. This was done by finding the columns “against\_” occurred for, then switching is out with new string text. This was a little tricky because I created the headers on in-line with replacement vice having the headers set ahead of time.

Format data: pprint is able to provide a reasonably clean view of the data. However, there were just way too many columns and it is not clear to see some of the more important columns. I selected a few and looped through my data and only printed the selected columns. This was again limited to a few rows for simplicity.

Identify outlier: This one can depend on the analysis performed The data itself was expected to range greatly, so you’d have to be careful not to assume a high/low value was an outlier. I decided to use the “is\_legendary” column to identify what may be considered an outlier and reasonably removed.

Find duplicates: Duplicates were determined by the “name” column and counted. This dataset is an inclusive list of all Pokémon, there should only be unique Pokémon. While there is probably multiple ways to handle this, the best option I saw was to exclude them when creating the data set during the replace headers part.

Conduct fuzzy matching: This was probably the most challenging part of the assignment. Since my data set did not have an obvious example, I used the type2 column to mimic this task. I used a random variation of “fire” and then compared the variations to “fire”. After comparing various ratios, I was surprised by the variation of results. I assumed with a simple comparison they would result in high ratios. However after seeing the results I realize fuzzy matching requires some thought behind before just taking a randomly high ratio value.

Through this project I learned a little more and strengthened what I we have been practicing up to this point. For the final project, I plan to spend more finding and researching my dataset. Determining good data analysis questions will help my focus on the different approaches to data wrangling. Additionally, I’d like to deviate from our text book [1] a little more and practice more generic ways to solve problems; that way they can be applied without specific requirements. Additionally, I’d like to spend more time on condensing code that way I am not duplicating code unnecessarily.