Pokémon Dataset Transformations Term Project Milestone 1

David D Berberena

Bellevue University

DSC 540 Data Preparation

Catherine Williams

April 7, 2024

Pokémon Dataset Transformations Term Project Milestone 1

**Introduction: Project Subject Area**

As I have been tasked with researching datasets that share a commonality amongst each other to perform various methods of data preparation, I wanted to work on data that I was either interested in seeing or on data that I was already familiar with to work with on a much grander scale. The parameters of this assignment immediately had me thinking about video games, as they are packed with a variety of data on specific characters, stats, and other qualities. Following this train of thought led me to think of one of the most popular video games in today’s society that seems to have a never-ending roster of characters: Pokémon. With the many different Pokémon featured in their video games, there is a plethora of data out there relating to each specific creature noted in the franchise. The subject of this term project will be centered around these Pokémon and their attributes.

**Data Sources**

Now that I have realized what it is that I’d like to do with this project, I managed to find the datasets needed for the project and its guidelines. I found a flat file (CSV dataset) from Kaggle.com, as well as a website that data present can scraped from, and an API URL that allows me to pull data from the Pokémon Company. Below is a small breakdown of the file formats I have retrieved for this data preparation project:

1. Flat file: The CSV document I have chosen outlines almost every single Pokémon in existence up to this point in time along with many of the important video game stats attached to each one. The link to the flat file dataset that I am using is here: <https://www.kaggle.com/datasets/rohanpatil63/pokemon-dataset>
2. Website: The data contained on the website I found matched up very nicely with the dataset from the CSV file, as its purpose is the same as the previous file format. The data featured here on the website identifies the Pokémon (this time up to a certain number), the video game stats, and a couple of other metrics that shed some more light on the Pokémon that are featured. The link to this website is as follows: <https://pokemon.fandom.com/wiki/List_of_Pok%C3%A9mon_according_to_base_statistics>
3. API: This API deals with everything related to the Pokémon games that occurs or is present in-game. Many different API URLs on the API website can request different types of information, yet the API URL I will be employing in this project specifically relates to the retrieval of data pertaining to the Pokémon themselves. The links for the API documentation and the API URL I will be using are below:

API Documentation Link: https://pokeapi.co/docs/v2

API URL Link: https://pokeapi.co/api/v2/pokemon/{id or name}/

**Relationships**

After looking at these forms of data, they are very closely related regarding the information held within. The CSV flat file holds information such as Pokémon name, ID (PokéDex) number, and stat information like Attack, Defense, Speed, and others. The website data also holds this exact same information, showing that these two have many commonalities and relationships. The API data requested from the API URL gets us data including the Pokémon name, the ID (PokéDex) number, and stats like Speed. The connection I wish to emphasize in my chosen datasets is the Pokémon names, as those names are consistent throughout each dataset and have the least chance of being input correctly, containing missing or NaN values, etc. For the API data to be compiled similarly to the structure of the other datasets while containing the same Pokémon names as the other datasets, several API calls would have to be made using the names featured in the other two datasets, as the names of the Pokémon are unchanging and can be called from multiple datasets with the same output.

**Project Plan, Challenges, and Ethical Implications**

Upon looking at the guidelines for the rest of the project, not much data transformation is happening intending to understand what the data means concerning certain variables. It seems as if we simply will be making the datasets readable to humans and combining them to store the finished dataset in a database. Data visualization is where we can start to see where patterns and trends arise in our relational datasets. As we are using data manipulation libraries like Pandas and data visualization libraries like Matplotlib, I wish to be able to glean answers from the data, such as which Pokémon has the highest HP, or which Pokémon is the slowest, or how many Pokémon have a stat total of 500 or more. These answers can be found with what I currently know about transforming the datasets and do not present a challenge. Now working with the API calls to generate a human-readable dataset may present a challenge though. I plan to loop through each ID number and make an API call for each ID number, then extract from the JSON dictionaries the specific key-value pairs I am interested in and which also relate to the other two datasets. This may sound easy enough, but I do foresee this step being cumbersome to manage at this point in the project. As always, data ethics comes into question when a data-centric project is underway. People ask, “Is the data safe”, or, “How did you get that data to begin with?” In the case of the various Pokémon datasets I have acquired, everything regarding this type of data has been provided by the Pokémon Company, which allows this data to be public knowledge. The nature of the data is innocent and does not contain any inappropriate, illegal, or defaming information, among other things. The Pokémon data that is being worked with is derived directly from the Pokémon games, which are rated E for everyone to enjoy, with no questionable ethics behind the message the data can send. I am looking forward to crafting these datasets as I am passionate about the nature of the data as well as I am eager to understand how these three datasets can come together to form one all-encompassing spectrum of data.