Brandon Palonis

Dr. Garrison

CS1699

4/20/2019

Project 3

**W0:** The dataset I chose was one of Pokémon, obtainable from <https://www.kaggle.com/rounakbanik/pokemon> . This dataset contains information on all 802 Pokémon that currently exist, including their typing, all six of their stats and the total, and their possible abilities. To make this dataset more akin to the kind that are necessary for this project, I had to remove several columns from the original dataset that would not work for a public dataset. This was done to treat each Pokémon like a person and make this dataset innocuous enough to be released, but not so innocuous as to not be able to be further anonymized for this project. I removed columns containing: a Pokémon’s name in English, a Pokémon’s Pokédex number, A Pokémon’s name in Japanese, and a Pokémon’s classification. My assumption is that if this dataset were to be publicly released by a company, these would be direct identifiers as to the occupants of the dataset. A Pokémon’s English and Japanese name would not be released in a public dataset, just as a person’s name would not be released in a public dataset. The Pokédex number is also unique to each Pokémon, so it can be immediately linked to a Pokémon and therefore reveals too much information to resemble a publicly available dataset. Lastly, while there is some overlap on Pokémon classification, for instance there are four Pokémon classified as the Seed Pokémon, but there is only one Pokémon classified as the Moonlight Pokémon. Due to this, I felt classification would still be too direct of an identifier for some Pokémon, and therefore it was removed from this dataset. The removal of these fields does not remove utility from the set, there is still useful information to be gained by what is left, this data can still be used to analyze Pokémon as a whole. For instance, this still contains information on the typing of Pokémon and their stats, and one could use this dataset to determine the percentage of fire type Pokémon that have less than 50 base speed. The original dataset is included here as pokemon\_clean.csv, and the dataset with removed fields used for anonymization is pokemon.csv.