Case Study 1

BSDS Spring 2021

Instructions

- You must work in your assigned group on Monday, 2/22 in a breakout room during the live lecture (unless you have made prior arrangements).
- On Friday, 2/26 each group must present one or two of the open-ended problems in the assignment.
- Each group will submit a Jupyter Notebook file (.ipynb) and an associated .html file via Canvas. Each group member will submit the same two files, except for the "equal work pledge".
- The submitted notebook must start with a Markdown cell header indicating the Case Study number, dataset, group members, and any external sources. Everyone will additionally include an "equal work pledge" which states that they understand all code/answers in the assignment and that all group members contributing equally. If this is not the case please indicate.
- Each question must be annotated appropriately with Markdown cells. The Notebook file should written in a way that a third party with no knowledge of the questions can read it.
- All group members will receive the same grade unless the "equal work pledge" is violated.

Import the `pokemon.csv` dataset available on Canvas. Here is a description for each variable:

- `name`: The English name of the Pokemon
- `japanese_name`: The Original Japanese name of the Pokemon
- `pokedex_number`: The entry number of the Pokemon in the National Pokedex
- `percentage_male`: The percentage of the species that are male. Blank if the Pokemon is genderless.
- `type1`: The Primary Type of the Pokemon
- `type2`: The Secondary Type of the Pokemon
- `classification`: The Classification of the Pokemon as described by the Sun and Moon Pokedex
- `height_m`: Height of the Pokemon in metres
- `weight_kg`: The Weight of the Pokemon in kilograms
- `capture_rate`: Capture Rate of the Pokemon

- abilities: A stringified list of abilities that the Pokemon is capable of having
- experience_growth: The Experience Growth of the Pokemon
- base_happiness: Base Happiness of the Pokemon
- hp: The Base HP of the Pokemon
- attack: The Base Attack of the Pokemon
- defense: The Base Defense of the Pokemon
- sp_attack: The Base Special Attack of the Pokemon
- sp_defense: The Base Special Defense of the Pokemon
- speed: The Base Speed of the Pokemon
- generation: The numbered generation which the Pokemon was first introduced
- is_legendary: 1 if the Pokemon is legendary and 0 otherwise.

1. Create the following visualizations.

a. Make a scatter plot of different Pokemon's attack vs. their defense stat.

b. Create a histogram of Pokemon's heights.

2. Add a trend line using the geom_smooth layer to the scatter plot from question one. Does this trend line make sense? Now color the data points by type1 without altering this trend line.

3. Create a dataset of just Pokemon with type1 as grass in generation 3.

4. List the mean defense for each type1 variables.

5. Create a dataset of just legendary Pokemon. Add columns for combined attack + defense and sp_attack + sp_defense. Visualize these two new variables.

Note: In the following questions (6. - 10.) include summary statistics and/or visualizations to support your claims!

6. Is there a relationship between attack and sp_attack? If so, what? What is the best way to visualize this relation by type1?

7. Count the number of missing values for each variable. Why do you think these variables have missing values?

Open-ended Questions

8. Create a dataset of `non-legendary` Pokemon and devise a metric from the Pokemons' stats to rank them. Based on this metric, which "type" of Pokemon is the best? "Type" can be interpreted broadly as any of the categorical variables or something like "the tallest Pokemon are the best."

9. Compare/contrast `legendary` vs. `non-legendary` Pokemon using both numeric and categorical variables.

10. Develop and answer you own hypothesis; get creative!