# POKEMON DATASET

## 1. General description of dataset

The Pokemon dataset contains 3 csv files.

1. **pokemon.csv** - contains a characteristics of 800 Pokemon.
2. **combats.csv** - contains information about previous 50 000 combats. The first two columns contain the ids of the combatants and the third one the id of the winner. The Pokemon in the first columns attacks first.
3. **tests.csv** - contains two column of pokemons' ids to make 10 000 predictions.

The Pokemon characteristics come from the actual game but the battle data has been generated by a custom algorithm that doesn't take into account some of the game characteristics to simplify the dataset. The dataset does not contain any weights.

The goal of collected files is to develop a Machine Learning model able to predict the result of future pokemon combats. To do the prediction we will have the pokemon characteristics and the results of previous combats.

## 2. Description of attributes

*pokemon.csv*

| Name | Type | Meaning | Unit of measurement | Special values |
|---|---|---|---|---|
| # | Numerical(int) | ID of a single pokemon | Single pokemon | None |
| Name | Nominal(string) | Name of a pokemon | Single pokemon | There are 800 unique names for each pokemon |



*Created by: Grzegorz Meller, Mateusz Rock, Tomasz Wierciński, Robert Zwierzycki*

# POKEMON DATASET

| | | | | |
|---|---|---|---|---|
| Type 1 | Nominal(string) | First attack type | Single pokemon | Grass Fire,Water, Bug, Normal, Poison, Electric, Ground, Fairy, Fighting, Rock, Ghost ,Ice, Dragon, Dark , Steel, Flying |
| Type 2 | Nominal(string) | Second attack type | Single pokemon | Poison, Nan, Flying, Dragon, Ground, Fairy, Grass, Fighting, Psychic, Steel, Ice, Rock, Dark, Water, Electric, Fire, Ghost, Bug, Normal |
| HP | Numerical(int) | Hitpoints | Single pokemon | Scale from 1 to 255 |
| Attack | Numerical(int) | Attack force | Single pokemon | Scale from 5 to 190 |
| Defense | Numerical(int) | Defense points | Single pokemon | Scale from 5 to 230 |
| Sp.Atk | Numerical(int) | Special attack force of a pokemon | Single pokemon | Scale from 10 to 194 |
| Sp.Def | Numerical(int) | Special defense points of a pokemon | Single pokemon | Scale from 20 to 230 |
| Speed | Numerical(int) | Speed of pokemon | Single pokemon | Scale from 5 to 180 |
| Generation | Numerical(int) | Development stage of a pokemon | Single pokemon | Scale from 1 to 6 |
| Legendary | Nominal(string) | Legendary status True = legendary False = ordinary | Single pokemon | None |

*combats.csv*

| Name | Type | Meaning | Unit of measurement | Special values |
|---|---|---|---|---|
| First_pokemon | Numerical(int) | ID of the 1 pokemon in the duel | Single pokemon | None |
| Second_pokemon | Numerical(int) | ID of the 2 pokemon in the duel | Single pokemon | None |
| Winner | Numerical(int) | ID of the pokemon that won the duel | Single pokemon | None |



*Created by: Grzegorz Meller, Mateusz Rock, Tomasz Wierciński, Robert Zwierzycki*

# POKEMON DATASET

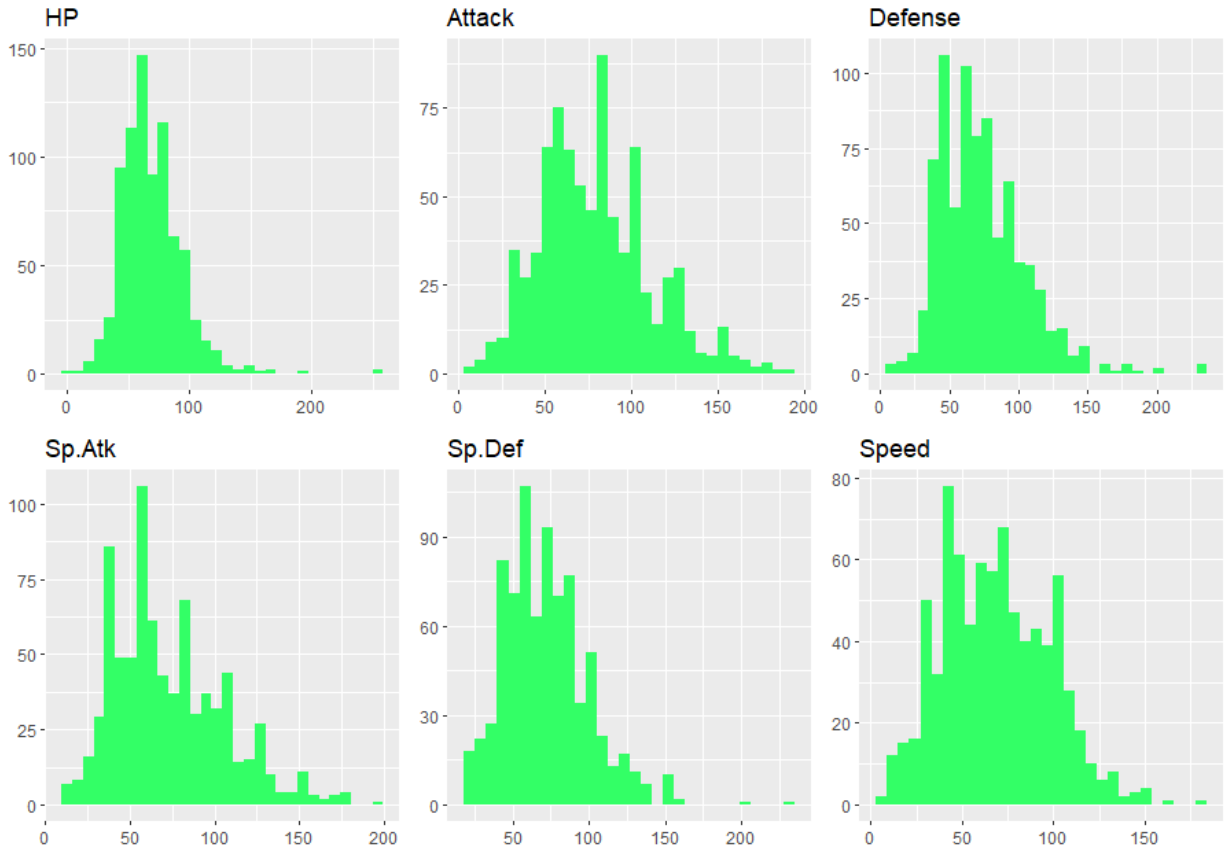
tests.csv

| Name | Type | Meaning | Unit of measurement | Special values |
|---|---|---|---|---|
| First_pokemon | Numerical(int) | ID of the 1 pokemon in the duel | Single pokemon | None |
| Second_pokemon | Numerical(int) | ID of the 2 pokemon in the duel | Single pokemon | None |

## 3. *Exploratory data analysis*

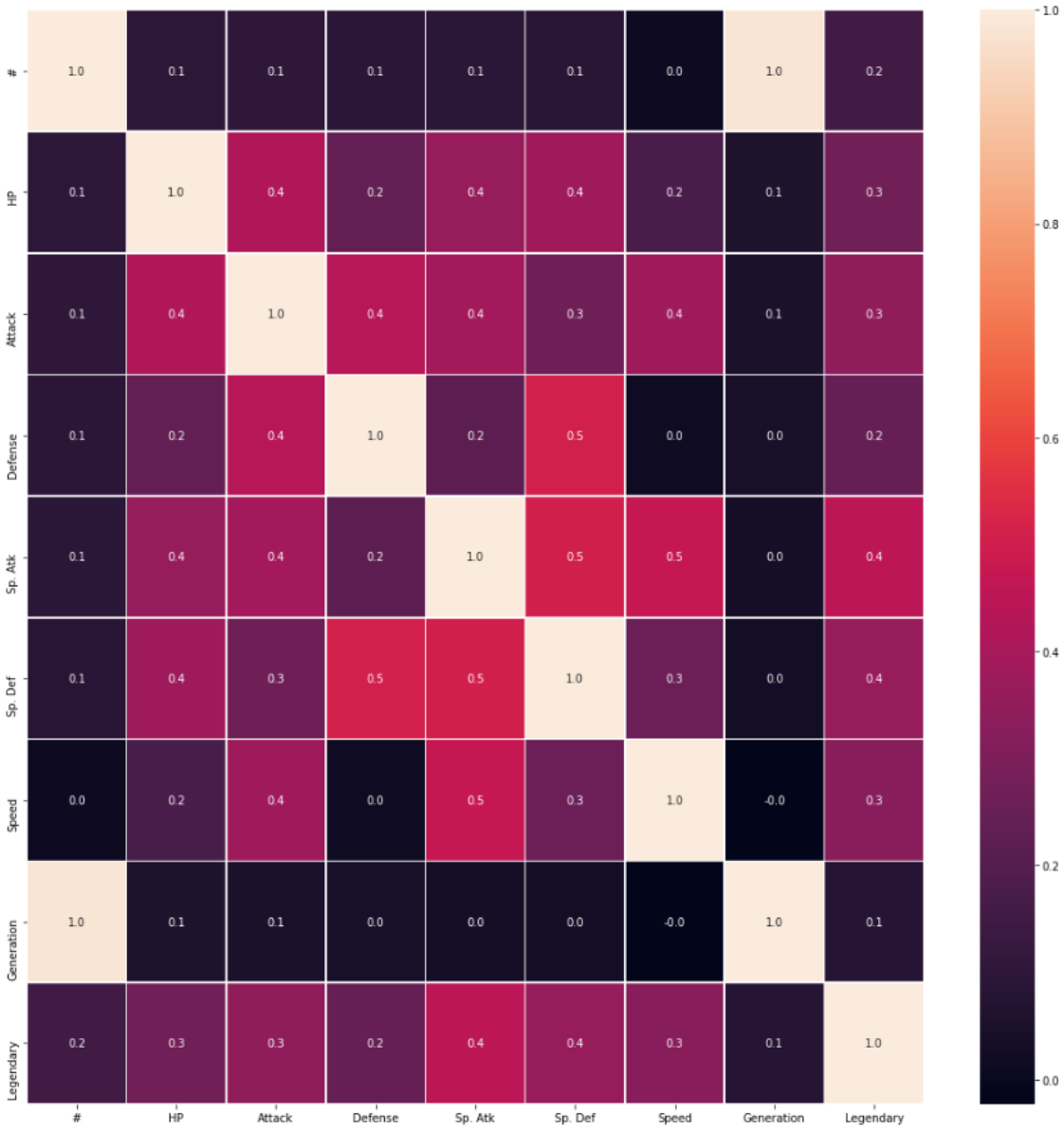
### 3.1. *Distribution of attribute values*

The distribution of numerical values from pokemon.csv



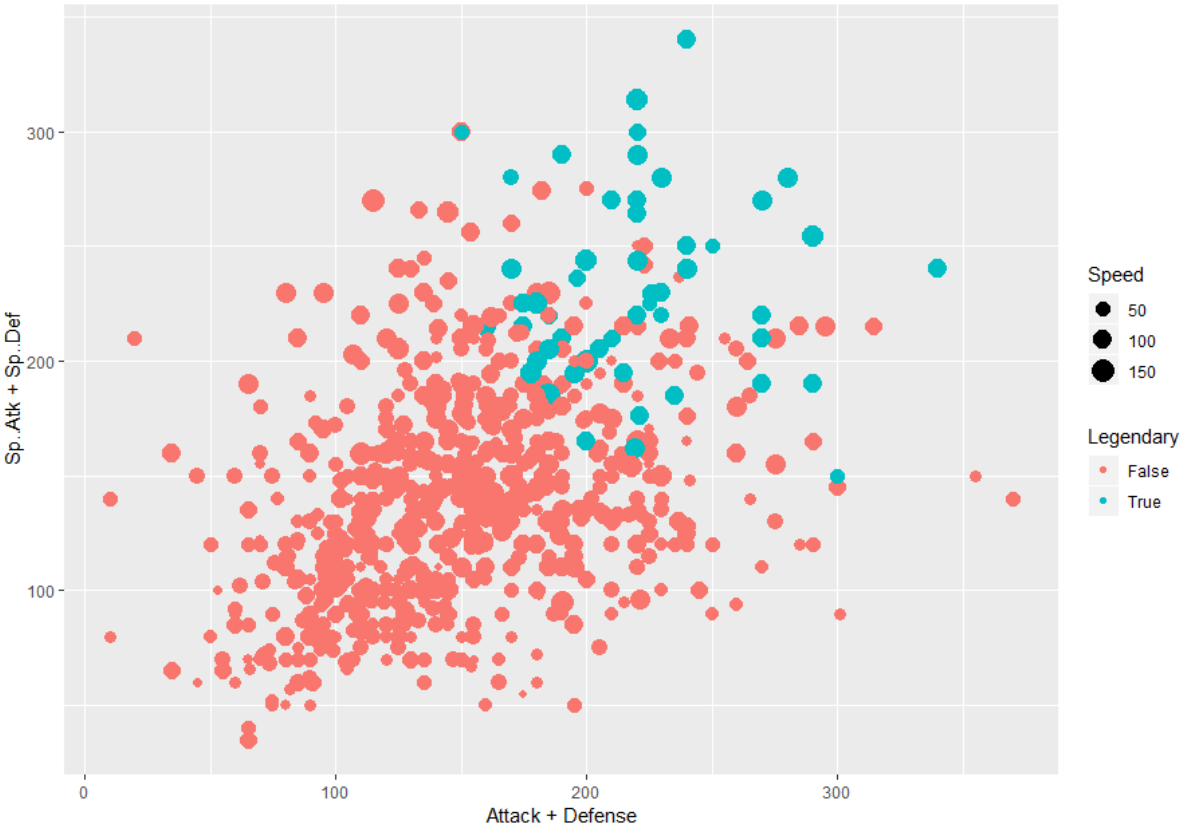
# POKEMON DATASET

### *3.2. Correlations between attribute values*




We can assume that there are no correlation.

### *3.3. Preliminary findings about the contents of the datasets*




There is a clear distinction between legendary and non-legendary Pokemon. The legendary Pokemon generally score higher across all skill points.

## *4. Data quality*

4.1. The amount of missing data – We do not have missing data in our dataset. However, there are some pokemons that do not have Second attack type.

4.2. Outliers – there are some outliers, we can see outliers on histograms

4.3. Overall data integrity – the dataset has overall data integrity

## *5. Preliminary goals for further data mining*

Pokemon holds a very special place in ours hearts as it is probably the only video game We have judiciously followed for more than 10 years. With this dataset, We wanted to be able to answer the following questions:

- Is it possible to build a classifier to identify legendary Pokemon?
- Which type is the strongest overall? Which is the weakest?
- Which type is the most likely to be a legendary Pokemon?
- Can you build a Pokemon dream team (6 pokemons)?
- Will you be able to predict the outcome of future combats?