Group 18 – Milestone 2

Group Name/ID: Group 18

Team Members: Ariel Guerrero (aes604)

<Captain> Hunter Long (mia014)

Joshua Cherry (qvm945)

Data Analysis Goal (0)

The goal of our data analysis is to find correlations between Pokémon types/generations and their various base stats. We will create a good number of descriptive/summary statistics and display them using the plotting techniques we learned in class. We will also use linear regression to train a machine learning model to predict the stats of future Pokémon from each type.

Data Description (1)

1. Our data was found on kaggle.com - link: <https://www.kaggle.com/terminus7/pokemon-challenge>
2. Our data is clean and organized already, so almost no preprocessing is required. The only “preprocessing” we did was to remove and add a few columns to the data.
3. Our data is in a csv file.
4. Our data is 41 columns long, so it isn’t practical to show an entire row here. Here are the first 10 columns of the first 5 rows to give an idea of the layout and data points:

Table

Description automatically generated

1. Number of rows: 1028, Number of columns: 41
2. We have all the data that we need.

Preliminary Analysis (2)

1. The first 16 of our data points are informational stats about the Pokémon, such as name, type, height, and abilities. The next 7 data points are the Pokémon’s battle stats like health points, attack, and speed. The final 18 data points are the Pokémon’s damage multipliers against all types of Pokémon. Again, because there are 41 data points, it would be excessive to show the distributions of them all. Below is the distribution of the attack and height stat:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

1. We used standard normal distribution to print a report including probability estimations for all Pokémon heights(m) and weights(kg):

﻿How many Pokémon heights are less than 7 meters?

# Expected: 1026 / 1027

# Observed: 1014 / 1027

How many heights are between 2 and 15 meters?

# Expected: 23 / 1027

# Observed: 110 / 1027

How many heights are greater than 2 meters?

# Expected: 23 / 1027

# Observed: 111 / 1027

What percentage of the Pokémon population is between 1 and 2 meters?

Answer: 27.64 %

What percentage of the Pokémon population is heavier than 50kg?

Answer: 56.09 %

What percentage of the Pokémon population is heavier than 100kg?

Answer: 40.77 %

What percentage of Pokémon are taller than 2.66807m? (mean+(1 std.dev.))

Actual: 65/1027 = 6.329%

Theory: 15.866%

What percentage of Pokémon are heavier than 75kg?

Actual: 246/1027 = 23.953%

Theory: 0.003%

1. We did not have a hypothesis.
2. First, we used some basic statistical methods, like the attack stat distribution above, the attack sample mean, and the attack stat Z score:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

We plotted the raw attack values against the Z-score:

Chart, line chart

Description automatically generated

We also found the SEM and the CI for the attack stat mean:

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

We found the Pearson, Spearman, and Kendall correlation coefficients for the Pokémon’s weight versus health points:

Chart, scatter chart

Description automatically generated

We found the height distribution of the Pokémon (shown above) and the PMF for these heights:

Chart, histogram

Description automatically generated

We graphed the PMF for the heights of dragon type Pokémon versus not dragon types:

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Finally, we graphed the CDF of the heights of dragon type Pokémon versus not dragon types and the CDF versus the PMF for the heights of dragon type Pokémon:

Chart

Description automatically generatedChart, histogram

Description automatically generated

Chart

Description automatically generated

1. We plan to continue implementing the statistical methods we learned in class, as well as use linear regression to train a machine learning model to predict the stats of future Pokémon from each type. We feel this is a good direction to continue the project in because after, we will be able to use the correlations we find to draw conclusions about the most powerful types of Pokémon and have a potentially working machine learning model.
2. We will measure the success of our simple data analysis by the quality of the conclusions we can draw based on our conclusions and test the performance of our model by running it testing it with a few Pokémon that we don’t use to train it.