## CS5830 Project 3

Kian Arnold, A02267479, Logan Liddiard, A02331163, Samuel McMillan, A02261822

## INTRODUCTION

For this project, we analyzed data involved in the first generation of Pokemon from Pokemon API. The intent behind analyzing this data was to determine important relationships between the elements of the data set. The created data set includes columns of Pokemon names, evolution levels, as well as several stats relating to Pokemon performance and viability in Pokemon combat. The results of this data may be valuable to professional gamers, video game developers by providing valuable insight into choosing the best early level Pokemon, the best overall pokemon, determining the best area for catching pokemon, and the best move sets for a variety of Pokemon types.

Slides Link
https://docs.google.com/presentation/d/1hAklT-c_khaHX8X7aTBQfKAI13zP5C3NAUj4nQyqcrc/edit?usp=sharing
Github Link
https://github.com/Logsterx7/cs5830_project3

## DATA SET

Analysis 1

For our first analysis, we used the data set present in the Github repo titled, ‘pokemon.csv’. In order to perform our analysis, we specifically looked at the columns, Order, Attack, and Health. The goal of this data set and the selected columns is to establish if the categorical variable ‘Order’, had a significant effect on an increased or decreased ratio between ‘Attack’ and ‘Health’.

Analysis 2

Similar to analysis 1, the same information was grabbed but this time habitats were appended to the proper pokemon so that we can look at types and habitats and compare them to each other.

Analysis 3

For the 3rd analysis, we used the datastip “pokemon.csv”. We looked at attack, speed, health, and defense as a combined total for determining the best pokemon.

## ANALYSIS TECHNIQUES

Analysis 1

Three analysis techniques were employed in our first analysis. The technique we

first used is a graphing of the rate of the density of the ratio of Attack over Health according to the pokemon Order. The second analysis technique used to evaluate our selected dataset is mean and standard deviation values of the ratio of Attack over Health for each Order category. The third and final analysis technique we employed is a T test. We specifically chose to perform 3 separate T Tests for this analysis, one to compare the group where Order = 1 and 2, another to compare the groups where order = 2 and 3, and another for groups 3 and 1.

Analysis 2

The main technique used was checking the standard deviation of the pokemon habitats and the count for how many appeared in each respected habitat.This allows us to check and see if pokemon are evenly distributed in each category. From there we can then proceed to check the different types of each pokemon and see if we have any kind of a correlation between types and habitats.

Analysis 3

Two techniques were used to determine the best pokemon according to combined health, speed, attack, and defense. A bar chart was used to visually compare those 4 stats averages and standard deviation was found to graph the normal distribution of the total of these stats.

## RESULTS

Analysis 1

For analysis 1, the results were surprising. The driving question toward Evaluating the data was, ‘is there a certain evolution level (Order) at which pokemon have the biggest bite per pound?’ Even though there was greater variety of weight per pound. Even with a bigger difference in values in group one (i.e., the standard deviation of .723 compared to .445 and .213 for the other groups), ‘evolution level (Order) 1 pokemon has the highest overall pound per punch (i.e., mean of 1.307 compared to 1.214 and 1.149 for the other groups).

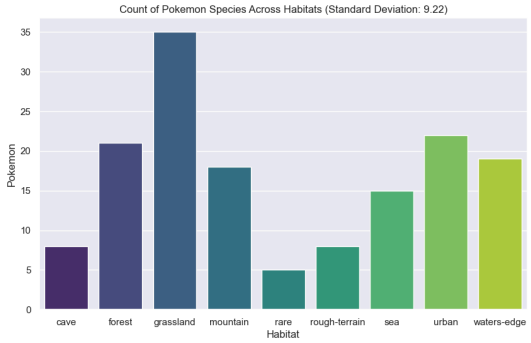
Three T tests revealed that there was no significant correlation between these three groups and their Attack/Health ratio, which leads us to reject the null hypothesis that evolution level 1 pokemon have the biggest bite per pound (Attack/Health). The p-value results for each test was greater than or equal to (0.39).

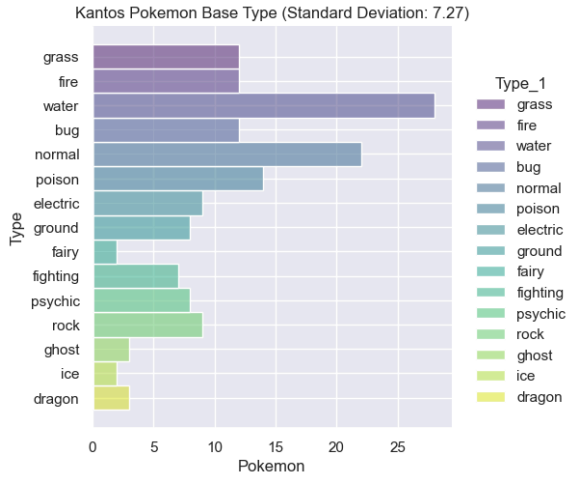
Analysis 2

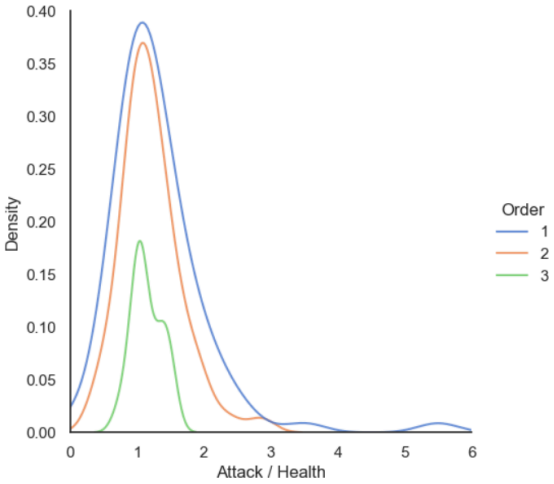
Interestingly enough the results from looking at the charts of pokemon proved to be very interesting. We wanted to look at the distribution of pokemon across different habitats. We took the standard deviation value of habitats and found that they were all pretty spread out. We then did the same with different pokemon types and found the standard deviation value was smaller but still spread out, but we found some interesting things from this. Grasslands were the most common habitat amongst pokemon but

water types were the most common base type of pokemon. When we look at the actual map of the region this makes sense since it is surrounded by water and contains rivers. On top of that this shows us that grassland habitats contain a wider variety of pokemon types since our grass types only make up less than a third of grassland populations. The key takeaway from this analysis is that if you are looking for a more diverse pokemon team you will want to search the grasslands.

Besides rare habitats, which are habitats inhabited by rare pokemon, The cave system and rough terrain were tied for having the least amount of pokemon. While they aren't great for searching for pokemon of a wide variety of types, they do serve as a place for very specific and more niche types to appear such as ground or electric types. If we refer to a map of the region it makes complete sense as to why these types and others are much more rare compared to others, there are significantly less habitable spaces for them.

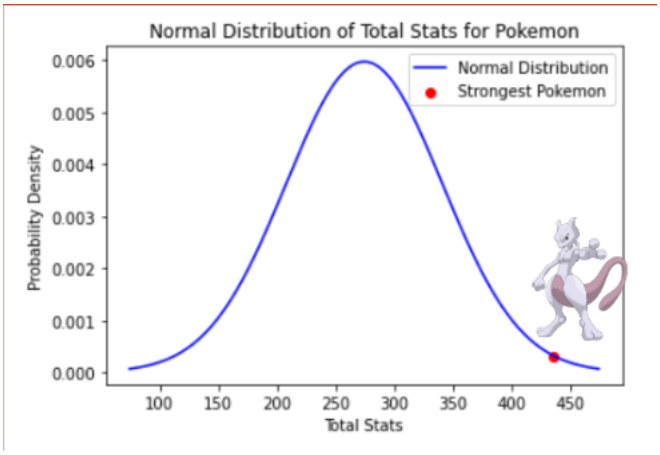



Analysis 3

The best overall pokemon was found combining the pokemons total health, defense, attack, and speed. Mewtwo was found to be the best one. Furthermore, the mean combined total of the 4 stats was around 275 for all the pokemon analyzed, so if one is worried their pokemon is generally weaker, add up the health, defense, attack, and speed to see how it compares.




## TECHNICAL

Preparation of the data for our analyses was more difficult than expected. The whole of our data was pulled from the Pokemon Api from https://pokeapi.co/. Pulling from and organizing data front this api proved troublesome as we were required to perform requests for each individual pokemon. Currently, there are a total of 1025 pokemon, but pulling the data for this many pokemon took several minutes and made it difficult to pull the exact right data that was required of our several analyses. It was divided to use the original 151 pokemon from the first generation of pokemon for sake of simplicity. However, though we decided to use a smaller number of data points for our data set, it still took several attempts to receive and format the data correctly. Eventually we were able to create two data sets, ‘pokemon.csv’, and ‘kantos-pokemon.csv’. The former included data on pokemon stats, evolution level, types, as well as pokemon name. The latter data set contained all of the same data as the previous data set but also included habitat information. Mean was used to show differences in the evolution of pokemon and total power for each evolution. The p-value and std shows if different evolution groups vary greater from one another in pound for pound strength. Standard deviation and bar charts were used for the habitat to show differences in the amount of pokemon very visually.

Preparing data for the 3rd analysis required making another column in the dataset for total (i.e., total of health, speed, attack, and defense) which could be used to graph. Each of these stats also were weighted equally important so we manipulated the higher stats on average to be lower for each pokemon and the lower stats to be higher for each pokemon (e.g., attack across all pokemon was 72 on average but was manipulated for each pokemon to be a mean of 68 to match the rest of the stats combined mean). Normal distribution was used to show the best pokemon because it shows if a pokemon is better than average in a visual way and it shows how good the best pokemon really is.