Project 3 - Pokémon Data Analysis

(Google Slides | Project Repository)

Caden Harris & Nate Taylor

Introduction

Pokemon is a popular role-playing game series involving the collection and battling of creatures called Pokemon. Because there are many factors for players to consider in choosing Pokemon lineups, movesets, and gameplay strategies, a look into the complex interplays between Pokemon stats, types, movesets, and generations has the potential to greatly increase a player's ability to strategize and efficiently progress in the games. In our analyses, we hoped to inform both experienced and novice players on choices involving which games to play, which Pokemon types to collect, and how to efficiently progress in the game.

The dataset we utilized was gathered from PokeAPI and contains Pokemon stats, types, names, generations, moves, and other related information. Our analyses involve means, Pearson correlation tests, bar charts, scatterplots, and distribution mapping to visualize important trends and correlations in the data. We found in our results the fastest Pokemon types, the inflation of stats over time, and generation-specific information on Pokemon experience yields, typing differences, and attack stat trends.

Dataset

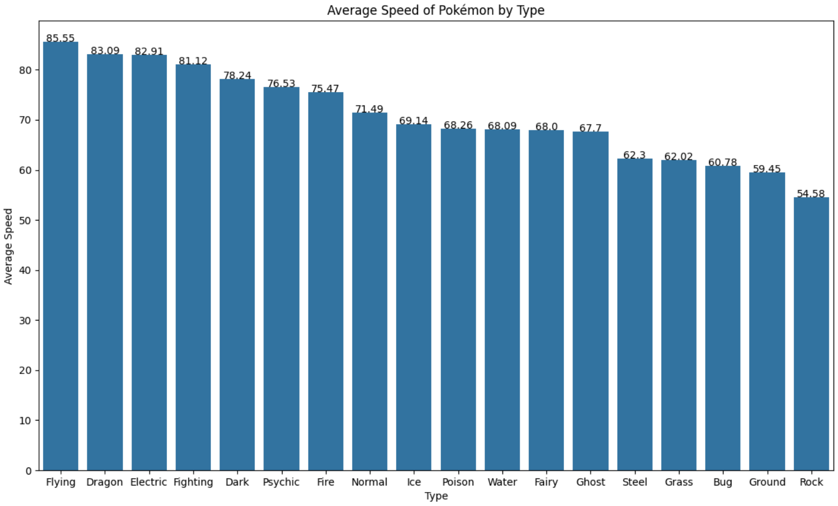
PokeAPI has a vast amount of data on all things Pokemon, especially data on Pokemon stats, generations, and movesets, which is what we focused on analyzing. It provides accurate, current information on every Pokemon and data values pertaining to them, including Pokemon name, typing, generation, base experience, base stats, and movesets. Retrieving data from PokeAPI required collecting information from endpoints specific to Pokemon game generations, as well as individual Pokemon and their respective moves. Data was returned in the form of JSON objects that were then iterated through in order to extrapolate the desired data points from each Pokemon. Checks had to be made to limit the Pokemon data and moveset data to certain generations after we collected it from PokeAPI. Because there are over a thousand Pokemon across all generations of Pokemon, retrieving the data took a considerably long time, so after we collected the data and organized it in a satisfactory manner, we saved several datasets as CSV files to reference in our analyses.

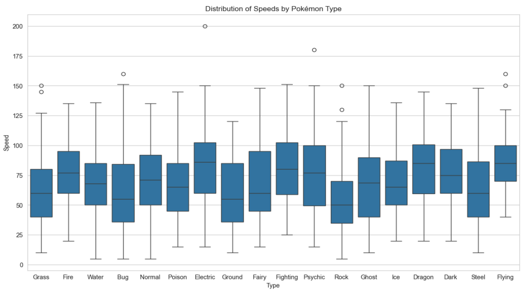
Analysis technique

We used mean analysis, Pearson correlation tests, T-tests, scatterplots, and barcharts in our analyses. Means were used to understand the average value of certain statistics like stats of specific Pokemon types to get a broad look at groups. Correlation was useful in showing trends between different stats and was particularly useful in viewing the relationship between the generation in which a Pokemon was introduced and its stats. Barcharts were useful to display easily readable stats for specific groups, like the Pokemon types. Distribution plots allowed us to visualize differences in stat distributions of different Pokemon types, while T-tests were used to confirm the significance of visual differences in data.

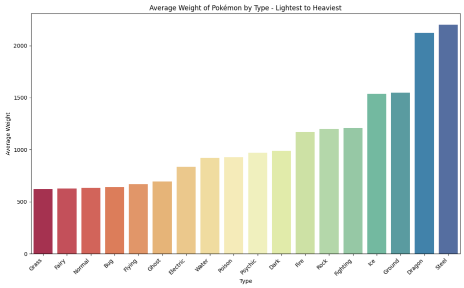
Results

For one of our analyses we wanted to look at a single Pokemon stat by type. We picked speed because it makes intuitive sense for certain types to be “faster.” The results showed that Flying types have the highest average Speed stat and Rock types have the lowest.



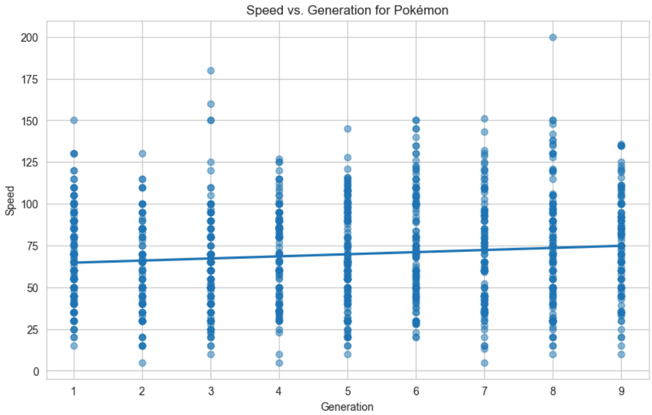


There is a clear difference between the types, and putting them in descending order shows us the trend of fastest to slowest types. Interestingly, the fastest singular Pokemon is an Electric type, Regieleki, with a value of 200. The slowest type is Rock which makes intuitive sense since rocks are slow and heavy, and flying types are going to be lighter and swifter. This is useful for showing players that if they value speed, flying types are going to be the most desirable.



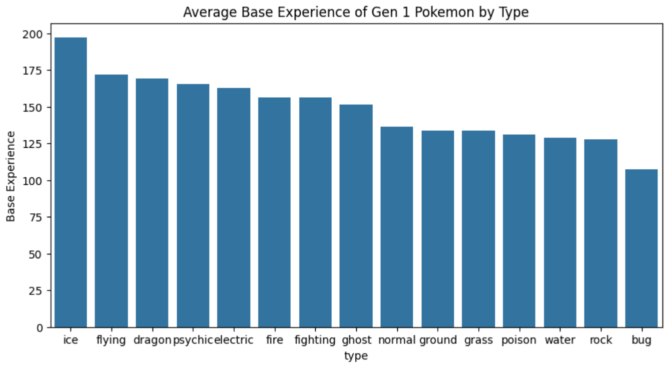
As a matter of interest, we also calculated the average weights of Pokemon by type. Flying and Rock are on opposite ends of the spectrum, but not on the polar ends as one might assume. This demonstrates that stats are more complex and many factors are involved in the composition of a single Pokemon.

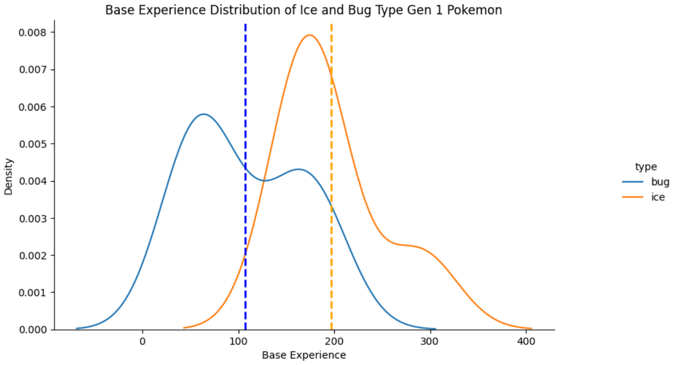
We also looked at the correlation between speed and generation (and other stats) to see if there is any connection:



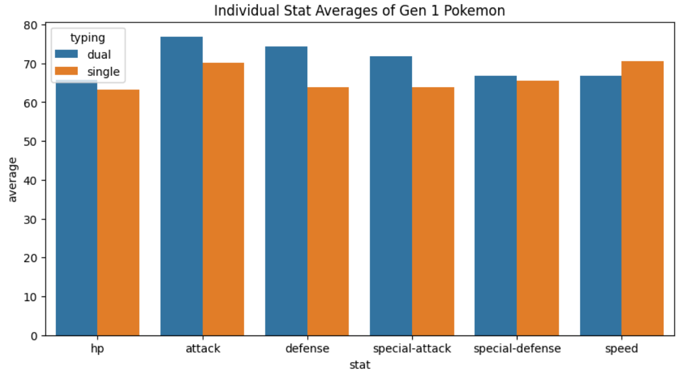
We observed a correlation of ~0.086 with a p-value of ~0.002. This suggests a slight upward increase of the stat over the generations, which shows stat inflation over time as more Pokemon are introduced. We did the same analysis for every stat in the compiled data and it yielded near-identical lines for each of them, emphasizing the trend of stat inflation. This might be useful for players because it shows that newer Pokemon have slightly higher stats and may have a competitive advantage over older Pokemon.

Our remaining analyses focused on trends seen in a single generation of Pokemon (Generation 1), in order to give players a more representative idea of gameplay in a single game. Separating Pokemon based on type shows that there are significant differences in the average experience yield of specific Pokemon types.



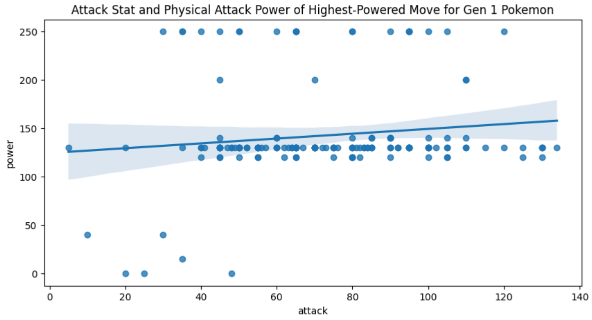


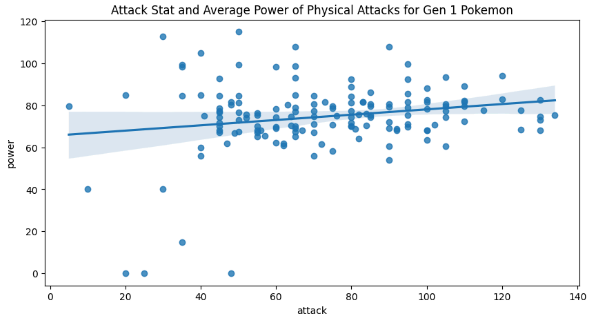
Our analysis found that Ice type Pokemon offered the highest average experience yield, while Bug had the lowest. This tells players that focusing their training/battling efforts on areas where Ice type Pokemon are abundant may allow them to gain more experience points per battle than amongst other Pokemon types (especially Bug types). The difference between Ice types and Bug types in particular was shown to be statistically significant, with a p-value of about 0.010. As expected, the Pokemon types with the lowest and highest minimum experience yield are Bug and Ice respectively; however, while Bug type has the lowest maximum experience yield value, Ice falls in the middle.



Next, Pokemon with only a single type were generally shown to be at a disadvantage when it comes to both individual and cumulative average stat values. Thus, a player can assume that any dual-typed Pokemon that they encounter may benefit from increased stats compared to single-typed. The only stat that single-typed Pokemon seemed to have advantage in was speed, implying that in addition to choosing certain types of Pokemon to prioritize Pokemon speed, limiting a player’s selection to Pokemon of only a single type might help them to narrow their selection. It should be noted that the only statistically significant difference was between the defense stats of the two groups of Pokemon. An additional fact of note is that the total difference between averages of individual stats of the type distinctions (32.630) was greater in comparison to the difference between average stat totals per Pokemon (25.126), implying that as a whole, the total of stat values per Pokemon is a little more balanced than individual stats per Pokemon.

Lastly, we found that there was a correlation between Pokemon with higher base attack stats and Pokemon with powerful moves. A weak correlation exists between base attack stat and the power of the most powerful move learnable by a Pokemon. Further, the correlation can not be shown to be statistically significant. There is, however, a slightly higher (and statistically significant) correlation between base attack stat and average power of all moves learnable by a Pokemon. These results seem to prove that although there is an intuitive relationship between attack stat and power of attacks, it is not as strong as one might expect, and many factors other than base attack stat must be considered when choosing a strong attacking Pokemon.





Technical

Preparing the data was a surprisingly complex process. It involved converting JSON objects retrieved from PokeAPI into Pandas-friendly dictionaries. Because much of the data made no distinctions between Pokemon generations, multiple layers of requests had to be made to the API and dataframe merges were required to limit data to a single generation of Pokemon. Because these API calls consumed a large amount of time, we opted to save several Pandas dataframes as CSV files to be used in performing subsequent data manipulations for analyses. This was also one of the driving reasons for limiting many of our analyses to a single generation of Pokemon. Future analyses would benefit from taking a broader look at all Pokemon and/or separating Pokemon into each generation for a more generalized look at trends.

For some of our analyses, we separated Pokemon by type and calculated mean values for statistics. Bar charts and distribution plots helped visually observe differences, and T-tests were applicable to show statistical significance of these differences because we were comparing averages of a single attribute across populations. Supplementary surface-level analyses involved sorting of data based on minimum, maximum, or true stat values. Distribution plots and Pearson correlation tests were applicable to comparisons of two different data values of a single population (all Pokemon) and helped to show and confirm trends as well. Our analyses were relatively specific and limited in scope, but future analyses would benefit from taking a look into all stat/type dynamics and the plethora of other such relationships that exist in the complex makeup of each individual Pokemon and Pokemon generation.