Pokemon Exploration

1. Introduction

This project explores the pokemon dataset found at:
https://www.kaggle.com/rounakbanik/pokemon

It contains various information on pokemon from each generation such as name, various stats like hp and attack, and the pokemon's type. In this project I explore some differences between pokemon in each generation. This is of particular interest because one may wonder if the popularity of pokemon has influenced the creation of new pokemon in newer generations. Perhaps the Pokemon company has taken notice of popular pokemon or pokemon types and decided to create new ones that are stronger than previous generations.

2. Data Processing

The data contains 801 pokemon and 41 columns. Looking through the data, there appears to be no null or missing values. The next step was to only select the columns needed for the purpose of this project. I also grouped the pokemon together by generation that they were released. At the time of the data being released, there were seven pokemon generations.

3. Data Analysis

3.1 Egg Step and Stat Differences Per Generation

In this section we will look at the average egg steps required to hatch the pokemon in each generation and the average stat per generation. Some background first: each pokemon can be obtained via breeding in-game in egg form and the player must walk around with the egg for a certain number of steps before it hatches. Each pokemon also has certain stat values such as health, attack, defense, and more. For this project we will look at the average total stat of each pokemon instead of individual stats.

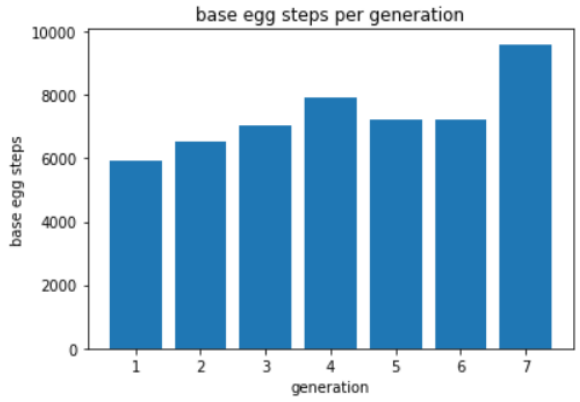

Fig 1 Egg steps per generation

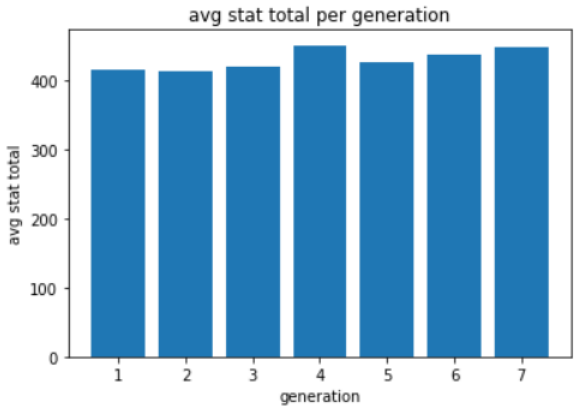

Fig 2 Avg stat total per generation

Fig1 shows the average number of egg steps per generation. At first glance, we can see a clear increase of egg steps in successive generations. This enforces the idea that there are some differences between pokemon in each generation but what lead to this increase of egg steps?

First we look at the average stat total in each generation since stronger pokemon may mean more effort to obtain. In Fig2 we can see a slight increase in average stat total per generation but nothing too alarming to warrant a massive spike in the number of egg steps.

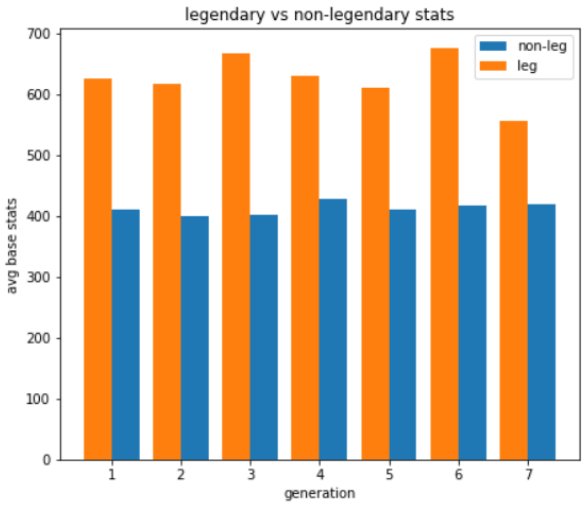

Fig 3 leg vs non-leg stats

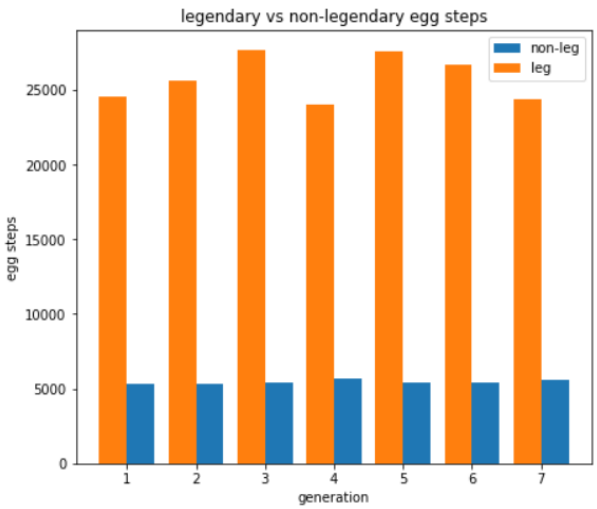

Fig 4 leg vs non-leg egg steps

In Fig4, we can see that eggs of legendary pokemon make up most of the average egg steps. If we look at how many legendaries are in each generation (from 1 to 7): [5, 6, 10, 13, 13, 6, 17], we can see that the amount of legendaries increased each generation which could be a potential answer for the increase in egg steps that we observed.

Looking further into the stat split between legendary and non-legendary pokemon in Fig3, we can see that it is also roughly the same per generation. However, one would expect more legendaries = higher stats yet from Fig2 and 3 we can see that while gen7 had the most legendaries, there is not a huge increase in stat. In fact gen7 legendaries had the lowest average legendary stat which leads one to wonder if the Pokemon company intentionally reduced the stats. Besides this finding, there does not seem to be other differences in stats.

	pokedex_number	generation	base_egg_steps
pokedex_number	1.000000	0.985904	0.197278
generation	0.985904	1.000000	0.128941
base_egg_steps	0.197278	0.128941	1.000000
base_happiness	-0.154776	-0.115974	-0.524551
base_total	0.140363	0.084358	0.496466
defense	0.102412	0.057603	0.265230
experience_growth	0.084687	0.052530	0.373837
height_m	0.008580	-0.025525	0.381455
attack	0.137633	0.101375	0.331088
hp	0.108079	0.074065	0.359218
sp_attack	0.109285	0.070146	0.404477
sp_defense	0.093848	0.048711	0.335404
speed	0.004294	-0.019914	0.289447
weight_kg	0.123567	0.085767	0.443286
is_legendary	0.196785	0.139029	0.873488

In Fig5, we can see a correlation matrix and the main focus of Fig5 is on the correlation of "base_egg_steps" with the other variables used. There weren't any high correlations besides "is_legendary" but that can be explained since "is_legendary" only takes on values 0 or 1 (for false or true respectively) and legendaries have higher egg steps. "base_total" shows moderately high positive correlation. "Base_happiness" has negative correlation but that is because pokemon tend to have the same base happiness and pokemon with high egg steps tend to have a lower happiness value.

Fig 5 Correlation Matrix for base_egg_steps

3.2 Type Differences

In this section we will look at how the stats of pokemon types changed per generation. There are 18 different pokemon types which represent some type of "element" that the pokemon specializes in. At most a pokemon can have two different types but some may only have one. The goal of this section is to determine if there is an increase in stats for pokemon of certain types per generation. If a pokemon was a water and ground type then it would be counted for both water and ground calculations since each type on a pokemon is equally important regardless of whether or not it is the first or second type listed. For this project I only had time to look into the fairy type.

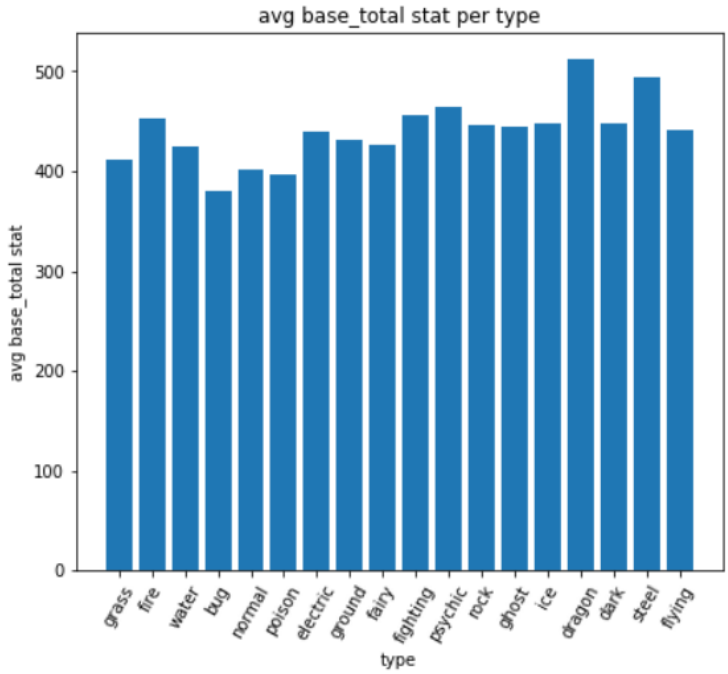

Fig 6 avg stat total per type

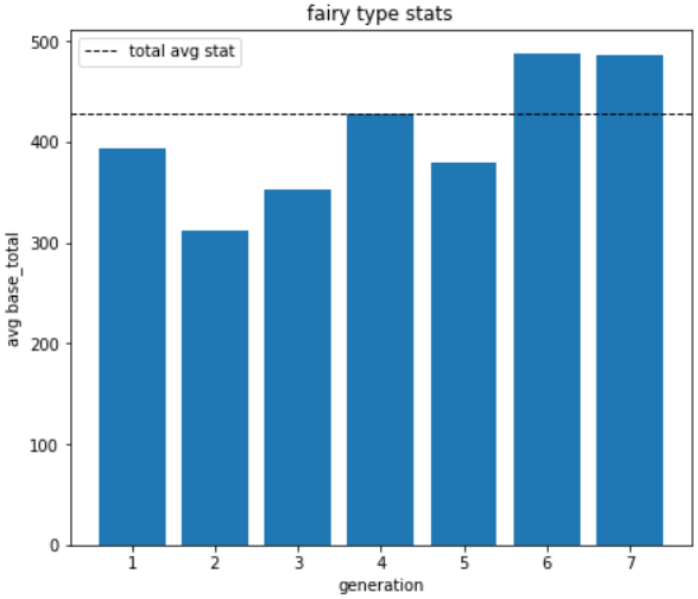

Fig 7 avg stat for fairy type per gen

Fig6 shows the average stat total of each type and we can see that fairies are about average. In Fig7, I show the average fairy stat per generation and compare it to the overall average stat of all pokemon. Gen6/7 fairies clearly have above average stats while previous gens are below average. Some background: fairy types were introduced in gen6 and the Pokemon company added the fairy typing to some previous pokemons. Knowing this, the question arises: are the stats of gen6/7 fairies statistically significant/larger than the previous generations of fairies? In

other words, did the Pokemon company create new gen6/7 fairies stronger than their predecessors?

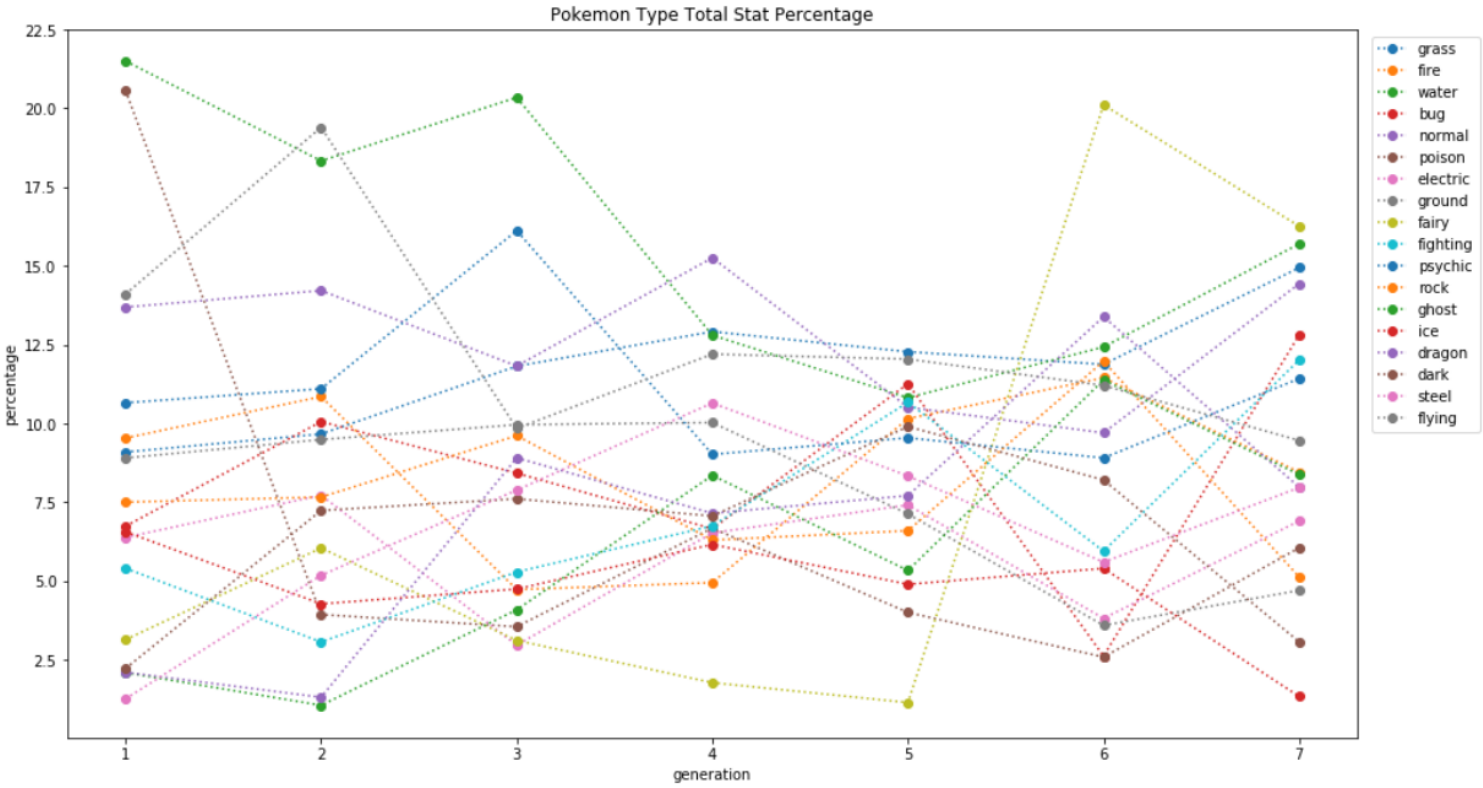

Fig 8 stat percentage per type per generation

In Fig8, it shows the percentage of a certain type's total stats over total stats for the specified generation. In other words, if we were looking at dragon types for gen1, I calculated the sum of all dragon type stats in gen1 and divided it by the sum of all pokemon stats in gen1. This allows me to see how much of gen1's stats were allocated to dragon types. We can see that fairy types hold most of the stats for gen6/7.

In order to test for significance, I used a one sided t-test where the null hypothesis is the average fairy stat for the respective generation = average fairy stat for gen1 fairies. The alternative hypothesis is that the average fairy stat for the respective generation is larger than the gen1 average fairy stat. First, I tested to see if the samples were gaussian with the shapiro function from scipy and it most likely is. I then proceeded to do a t-test using scipy.stats.ttest_ind() with alpha = 0.08. Since ttest_ind() is a 2-sided test, I divided the p-value returned from the function by 2 and checked if it was less than alpha. The result was a rejected null in support of the alternative for gen6/7.

In Fig9 below, I plotted the p-value of each gen for fairy types compared against gen1 fairies. Only gen6/7 had p-value < alpha.

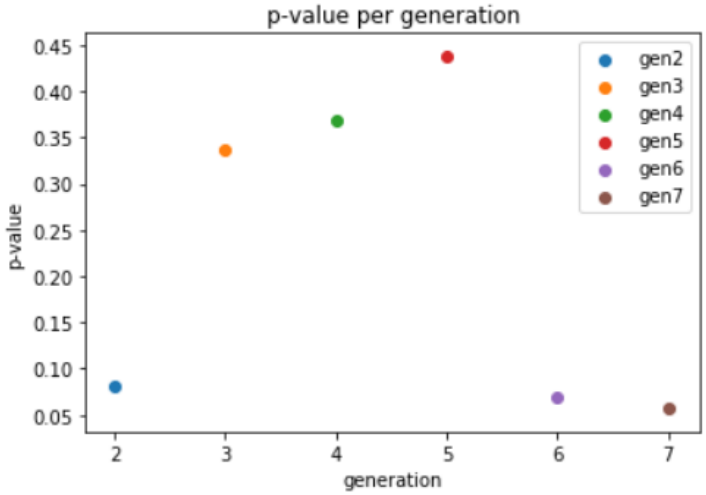

Fig 9 p-value per gen for fairy types

4. Conclusion

In conclusion, there are some signs of generational differences between pokemon. For example, even though gen7 has the most legendaries, the average legendary stat seems to be lower than previous generations. Egg steps also increased every generation, possibly due to the fact that the amount of legendaries increased per generation. In addition, we looked at differences between types, specifically fairy types, and found that in gen6/7 the average fairy stat is statistically larger than the average gen1 fairy stat. This suggests that the Pokemon company created gen6/7 fairy types to be stronger than previous generation fairy types.