Gotta Graph 'em All

Caroline Huang, Sean Jin, David Kang, Matt Liu

Introduction

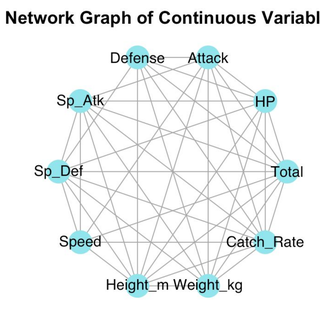
Pokemon is a popular franchise many of us have grown up with. Along with Ash Ketchum, our love for understanding these creatures has resulted in us yearning to “catch ‘em all”. The franchise has grown from a show to a video game, and now, it is easier than ever to immerse ourselves into this world with Pokemon Go. For this reason, our group is analyzing a Pokemon dataset from Kaggle. This dataset focuses on the stats and features of the Pokemon in the first six generations. There are 721 samples and 21 variables, plus Pokemon ID and name. **By looking into the data more closely, we wanted to know if certain types were stronger than others, and how different Pokemon stats were related or not related to each other. Additionally, “legendary” Pokemon seemed to be stronger or harder to catch than others so we were curious if the Pokemon stats backed up the “legendary” status of certain Pokemon.**

Findings and Conclusions

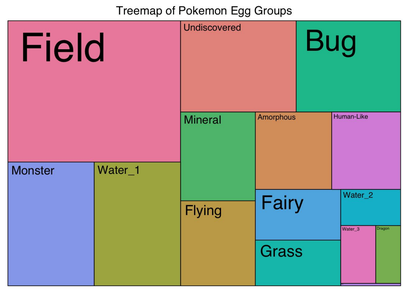
We can conclude that there is no conclusive evidence whether certain types were stronger than others as we can see on the box plot as there is no clear difference between each of the types, the “legendary” status of pokemon does play a role when it comes to their stats as seen through the principal component analysis, and the pokemon’s stats have some sort of relationship to each other and they share a similar distribution with one another. We also found about the frequencies of primary Pokemon Types as Water and Normal are the most frequent and Flying and Fairy are the least frequent, weight and height have a positive correlation with each other as they are similar to real animals, “Field” being the most apparent egg group out of all of them, and how the pokemon’s color and type are related to each other.

Figures and Analysis

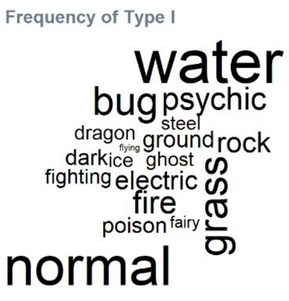
As a way to observe the overall relationships that the variables have, we plotted this dependence graph of the continuous variables in the Pokemon dataset. Also, we chose to plot this network graph in the shape of a circle so that it is easier for readers to observe which variables are related. From this graph, we can conclude that most of the variables have relationships with one another since they are connected with lines.



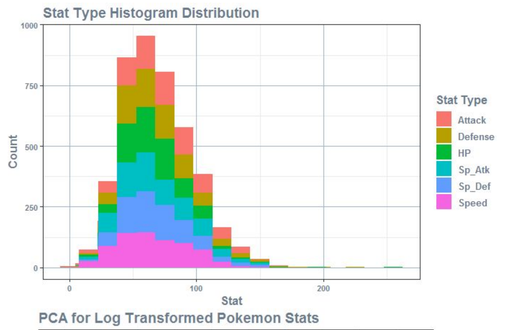
While analyzing the distribution of Pokemon Egg Groups, we used a treemap to compare the occurrences or frequencies of the egg groups. The groups are all scaled based on total frequency and labelled accordingly, and we can se that “Field” dominates most of the egg groups, as it dominates the majority of the “whole” rectangle. It seems to make sense that “Fairy” or “Dragon” egg types are less frequent since as mentioned before, those Pokemon were released much later compared to others.



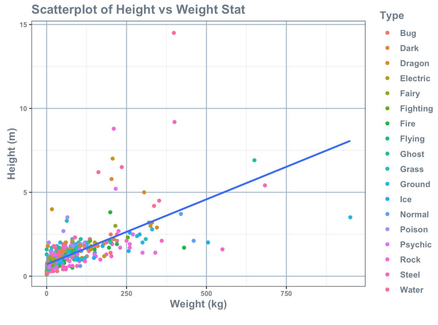
In order to assess the frequency of the primary Pokemon types, we used a word cloud to visually determine the occurances. From this graph, we are able to conclude that Water and Normal type Pokemon are the most common, followed by Bug, Grass, Fire, Psychic, and Rock. Flying was by far the least common, followed closely by Fairy. It makes sense that Fairy is the least common because the Fairy type was released years later than the rest of the types.



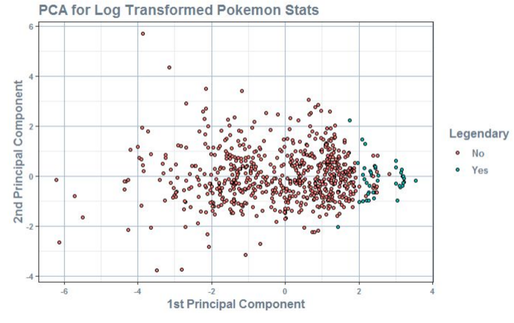
After examining the distribution of total stats based on Type_1, we were curious to see what the distribution of each individual stat type looked like. We created a stacked histogram and separated the types by color coding them according to the legend. Based on the graph we can see that the distribution of each stat appears to be about the same with a few outliers on the higher end of the spectrum, however the data seems to mainly be centered around a stat of 60. Each of the Stats seemed to be similarly distributed.



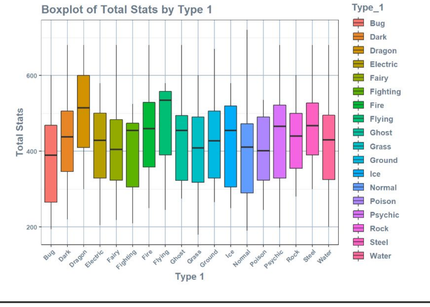
First, to see what variables change based on the type of a Pokemon, we created a scatterplot of the height and weight of the Pokemon. Overall, there appears to be a positive correlation between the height (in meters) and weight (in kg) of Pokemon. This is corroborated by the linear trendline that shows a positive correlation between the two variables. Each point is also colored by the type of the Pokemon. While observing the colors of the points, there does not seem to be a trend between the type of a Pokemon and the corresponding height or weight.



We then wanted to see the correlation between whether Legendary Pokemon and their log transformed stats (excluding total). By doing a principal component analysis, we can see whether there is a correlation between a Pokemon being a legendary or not and their log transformed stats. So, a principal component analysis graph was made in order to visually see the correlation between the Pokemon’s legendary status and their log transformed stats. We can see that there is some correlation between Pokemon’s stats and whether a Pokemon is a legendary or not as there is a clear distinction where the distribution of legendary Pokemon lie compared to non-legendary Pokemon.



Next, we used a box plot to compare the total stats of each primary Pokemon type. This reveals the min, max, Q1, Q3, and median total stat of each Pokemon type. Because every box plot overlaps with every other box plot, we cannot say that any two types have significantly different total stats. Normal type Pokemon have the largest range of total stats and poison have the smallest. Flying type Pokemon have the largest mean, followed closely by Dragon, while Bug type Pokemon have the smallest mean.



Finally, we examined the relationship between Pokemon type and the color of the Pokemon. Our initial thoughts that there is a higher frequency of color that is associated with the type was correct, in that water Pokemon were most frequently blue and electric Pokemon were most frequently yellow, and so on. However we can still see a large variety and distribution of colors among Pokemon regardless of type, and found it interesting that some types shared certain colors (Ground and Normal both majority Brown, Rock and Steel both majority Grey).