Team 20: Final Project Proposal

Minh Tran Quoc, Melody Yu, Eric Salguero, Danh Nguyen

Overview

For our project, we decided to see if we could determine if a pokemon's special defense attribute was based on other factors, namely its health points (HP), speed and defense. We wanted to focus on special defense rather than its attack parameters since we believe defense to be underappreciated when it comes to determining a pokemon's battle capacity. To do so, we used multivariate linear regression and found that special defense was indeed influenced by some of the other factors, but not all. To specify, our analysis showed that there was a positive relationship between special defense, defense and HP; however, speed was found to have little to no effect.

Background

Data

Our data analysis centers around the popular video game and media franchise, Pokemon. In the span of two decades, eight generations of Pokemon have been released, bringing the Pokemon population from its original population of 151 to nearly 900 today.

Our dataset is a subset of Pokemon from the first six generations (link: https://www.kaggle.com/abcsds/pokemon). It consists of 721 samples and contains 13 attributes, which are described below:

- **ID** - pokemon's identification number
- **Name** - pokemon's name
- **Type1** - pokemon's base type, which determines weakness or resistance to attacks
- **Type2** - pokemon's second type; some pokemon have this
- **Total** - the sum of all the following statistics and can be used generally to determine a pokemon's fighting capabilities
- **HP** - health points; defines how much damage a pokemon may withstand before fainting
- **Attack** - the base modifier for normal attacks
- **Defense** - the base damage resistance against normal attacks
- **SP Atk** - special attack, the base modifier for special attacks
- **SP Def** - special defense, the base damage resistance against special attacks
- **Speed** - determines which pokemon attacks first each round; a higher speed indicates it will attack first
- **Generation** - the generation that the pokemon belongs to
- **Legendary** - whether a pokemon is a legendary pokemon or not

Our only label is the special defense attribute.

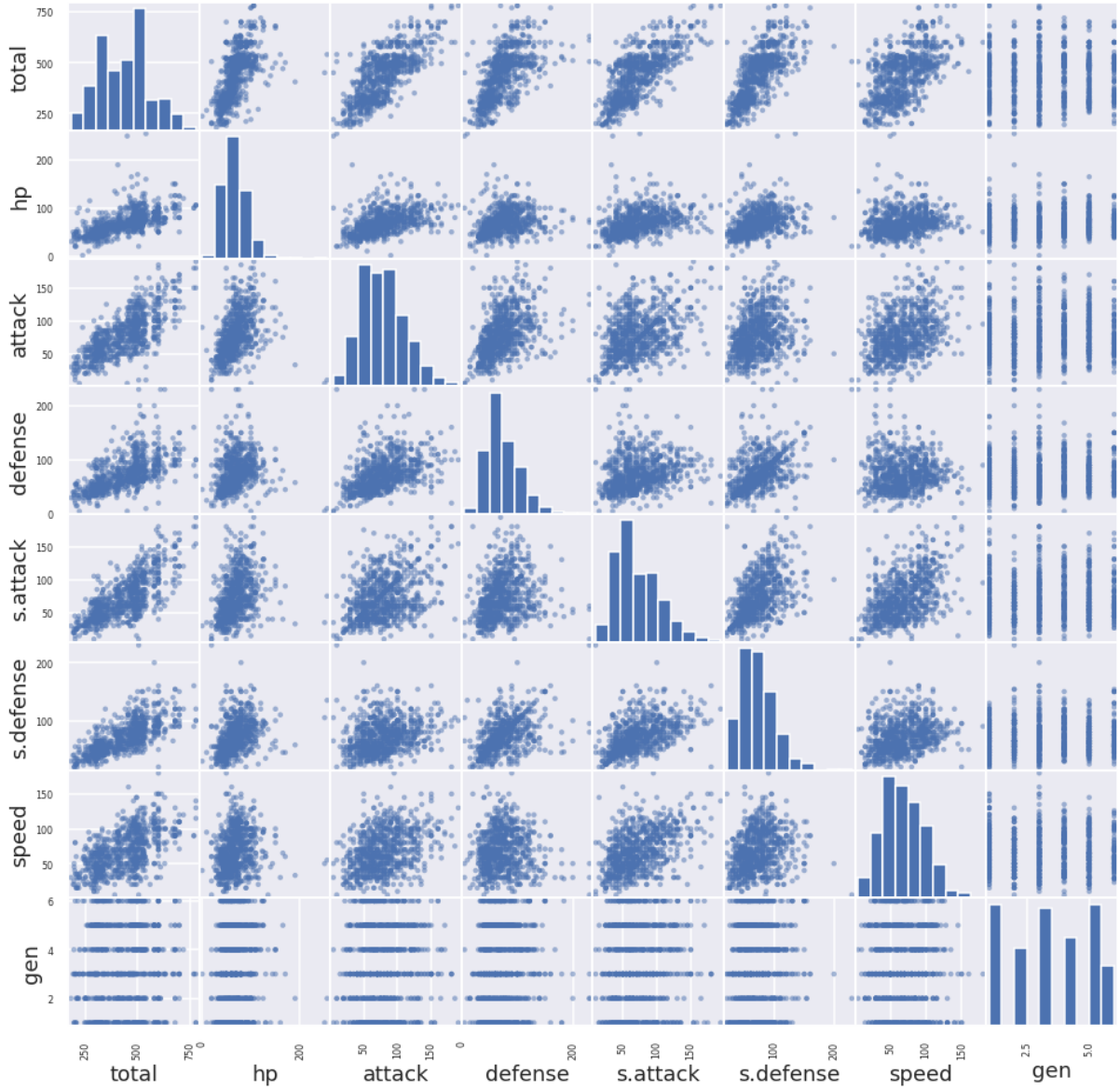
Research Question:

Does a pokemon's special defense statistic get influenced by their other defensive statistics (i.e. health points, speed and defense)?

Methods

Data Visualization:

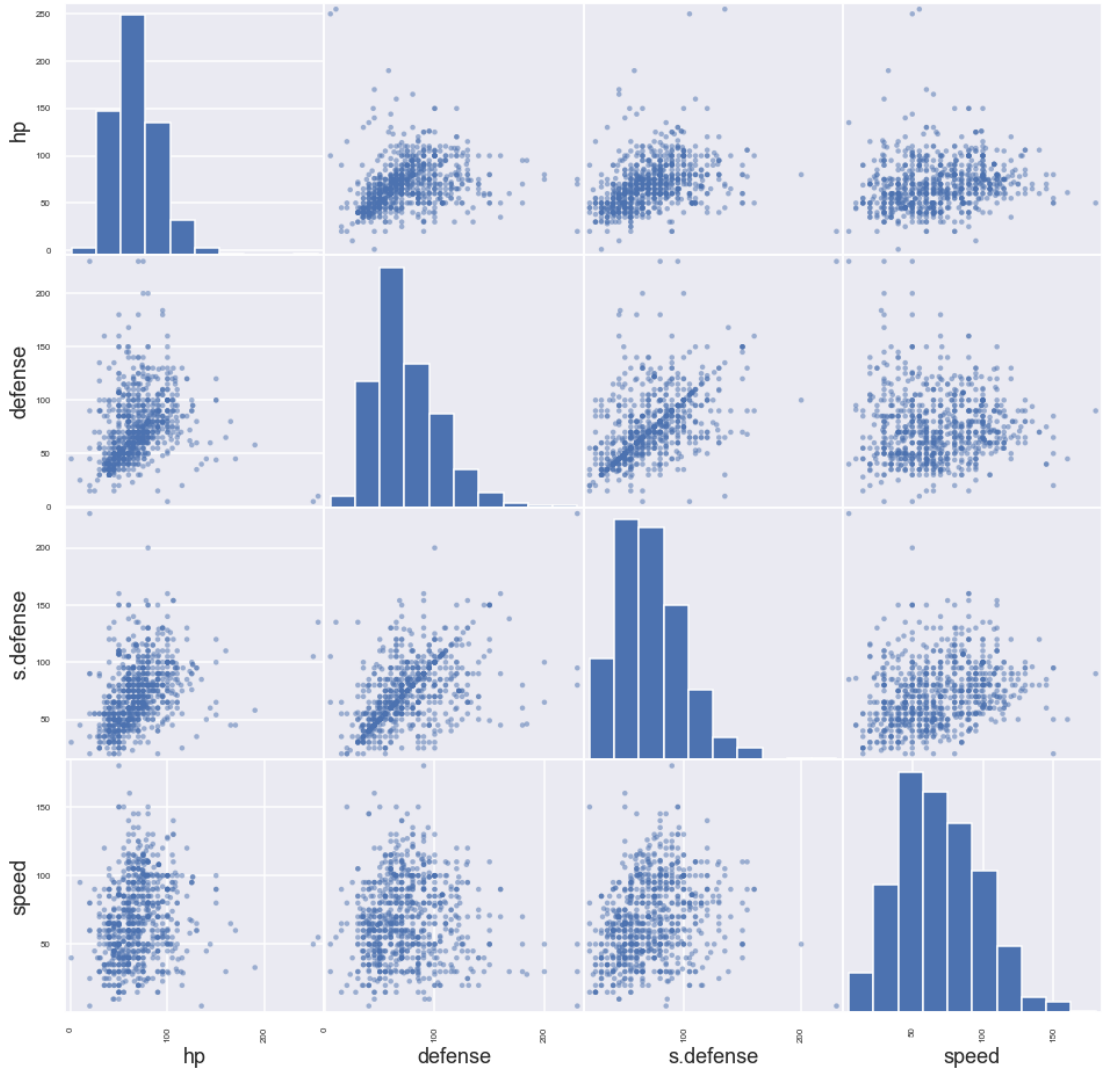
Prior to performing analysis on our dataset, we observed the data with a heat map and scatter matrices to see how other statistics correlated with special defense. We first noticed that the Total was positively correlated with the other statistics, which was understandable given that Total is the sum of each statistic. Then, we saw that most other attributes were positively correlated with each other, which hinted that statistics were correlated in some way with others.



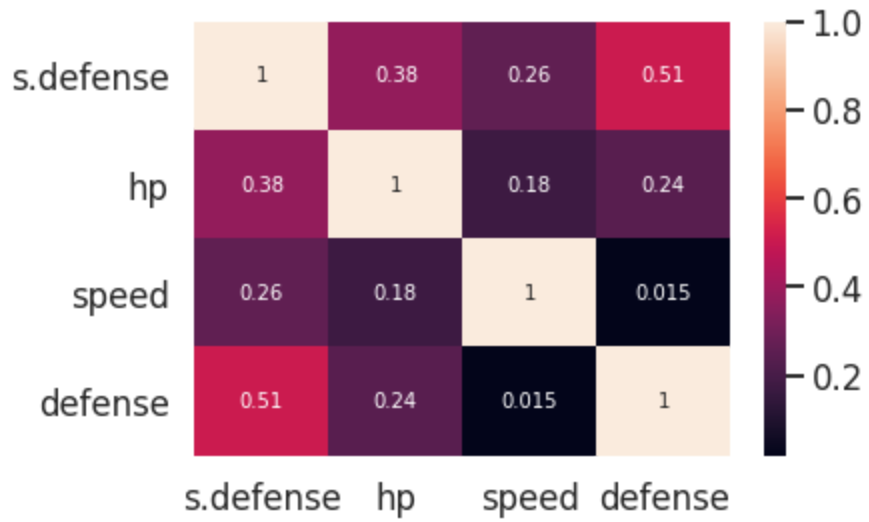
Data Cleaning:

Afterwards, we cleaned the data. In particular, we determined whether the dataset contained any rows that were duplicates or null values. If they did, we would delete those rows, as they could potentially skew our results. Ultimately, we found that none of the rows contained duplicate data or were null. Thus, our sample size remained unchanged, with 721 observations.

Following that, we dropped unnecessary data variables and their corresponding columns from our data. This included: ID, Name, Type1, Type2, Total, Attack, SP attack, generation and legendary. This left four columns of data. Since we dropped columns, we decided to look at correlation again to see whether the data would be more specific.



To further analyze the data, we used a heatmap to see correlation scores. We did this to fully understand our above scatter matrix and saw that there was some correlation between the variables.



Primary Analysis:

With those four columns, we conducted our primary analysis, multivariate linear regression. First, we randomly split the data into 80/20 sets for training and testing, respectively. For each model, we used three different splits and ensured that the split generation was based on the same seed so that our notebook ran the same split combinations. This was done so that we could later use cross validation.

Then, we created and fit three models on each training set to decide which would suit our dataset best:

- Model 1: s.defense = w0 + w1*hp + w2*defense + w3*speed
- Model 2: s.defense = w0 + w1*hp + w2*hp^2 + w3*defense + w4*$defense^2$ + w5*speed + w6*$speed^2$
- Model 3: s.defense = w0 + w1*hp + w2*defense + w3*speed + w4*hp*defense + w5*hp*speed + w6*defense*speed

For the three models, we chose these three variables (HP, defense and speed) because we believed them to be in some way, an influence on special defense. Choosing HP and defense seemed obvious, but we included speed as a variable because we thought the pokemon with the higher speed would likely have a higher defense as well, since they would be the initial attacker.

Therefore, Model 1 investigates whether those three variables independently influence special defense. Model 2 works similarly, except it involves second order terms to capture a more complex relationship. Model 3 differs from the previous two as we included interaction terms to see whether it was the work of two variables in tandem that influenced special defense.

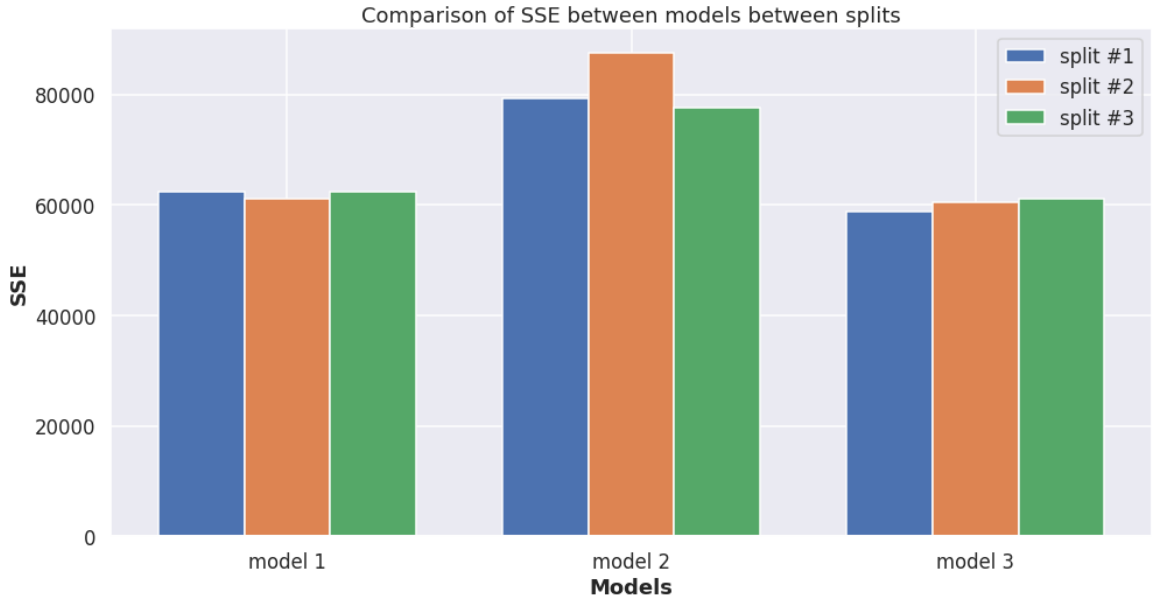
Using these models, we calculated the SSE for the testing set to evaluate which model generalizes the best.

Results

With our methods, we determined three model equations for each split:

Split 1	Model 1	s.defense = 8.953148 + 0.276588*hp + 0.422832*defense + 0.189525*speed
	Model 2	s.defense = 1.663093 + 0.160942*hp + 0.000552*hp^2 + 0.632726*defense + -0.001062*$defense^2$ + 0.308183*speed + -0.000821*$speed^2$
	Model 3	s.defense = -7.940202 + 0.474775*hp + 0.721609*defense + 0.201399*speed + -0.003825*hp*defense + 0.000974*hp*speed + -0.000841*defense*speed
Split 2	Model 1	s.defense = 9.238354 + 0.281296*hp + 0.389170*defense + 0.204420*speed
	Model 2	s.defense = -6.895276 + 0.118630*hp + 0.000760*hp^2 + 0.777981*defense + -0.002073*$defense^2$ + 0.467941*speed + -0.001835*$speed^2$
	Model 3	s.defense = 5.424525 + 0.418885*hp + 0.500522*defense + 0.081257*speed + -0.002755*hp*defense + 0.000272*hp*speed + 0.001544*defense*speed
Split 3	Model 1	s.defense = 9.608260 + 0.270239*hp + 0.409876*defense + 0.199401*speed
	Model 2	s.defense = 2.112712 + 0.048542*hp + 0.001072*hp^2 + 0.670893*defense + -0.001273*$defense^2$ + 0.392654*speed + -0.001316*$speed^2$
	Model 3	s.defense = 3.659380 + 0.329839*hp + 0.699190*defense + 0.036856*speed + -0.004023*hp*defense + 0.003059*hp*speed + -0.000399*defense*speed

Following this, we looked at a bar graph containing the SSEs of each split and their models on the test set.



From our graph, we saw that Model 3 generalized the best and had the lowest SSE. This is likely due to the relationships between the interaction terms, which would be difficult to capture with a model like Model 1, since it focuses only on individual attributes. While Model 1 isn't necessarily bad, it is outperformed by Model 3. We also saw that Model 2 was heavily overfitted and was therefore a bad generalization of the dataset. This is understandable as Model 2 involves second order terms, which overtly capture the variance in the training set.

After determining that Model 3 was the best model for our data, we looked back at the coefficients of Model 3 within the splits. Across all the splits, we noted that special defense was affected most by HP and defense (though not necessarily to a large degree) and that speed had little to no influence. Additionally, the correlation between special defense and speed was actually reliant on the interaction between HP and defense, as was shown by our regression analysis.

Discussion

Our results are not surprising for the Pokemon world. In real life, the animals that Pokemon are based on depend on speed for survival. In the Pokemon world, however, this is not the case, as speed only impacts who attacks first. We estimate that with our results, Pokemon trainers can focus on improving their HP and defense to ensure battle survivability.

We recognize that there are limitations to our results, such as the dataset being too limited, the result potentially being due to overfitting and/or being restricted by the types of features we use. There are also external factors that affect our results, such as in-game items that scale each statistic a different amount or the potential inaccuracy of these statistics. Additionally, these statistics only reflect the first six generations and there's no indication that the following generations will not differ.

For future analyses of this question, we believe that this question could be better answered if we had more samples to work with and/or if we included special attack to see how it affects special defense.