# Project Proposal

STA210: Brian Daniel and Nate - Daniel Cohen, Nate Krall, Brian Kim Team

```
library(tidyverse)
library(tidymodels)
library(ggplot2)
library(dplyr)

# add code to load data
pokemon <- read.csv("data/pokemon.csv")
```

## Introduction

The expansive world of Pokémon, at its core, is a children's game. However, a deeper look at the numbers and statistics the game is built from reveals several intricate relationships among the different pokémons' characteristics. Each creature has its own specific set of base statistics, colloquially referred to as "base stats," including attack, special attack, defense, special defense, and speed, which indicate that pokémon's battle prowess. Summing these stats yields a pokémon's total base stats, which is the best measure of a pokémon's overall strength when all pokémon are put at an even playing field – common knowledge for any pokémon fan. We are interested in measuring a pokémon's strength without using base stats as predictors, giving us insight on how strong the relationships among pokémon's different characteristics actually are. Thus, we are looking to answer the following research question: **Can we predict a Pokémon's Base Stat Total from other variables?** In other words, we are analyzing how well variables such as the pokémon's type, capture rate, growth rate, generation, height, weight, base happiness, weaknesses, and if the pokemon is legendary or not can predict a pokémon's total base stats. We hypothesize that a multiple linear regression model including some formation of these predictor variables will be a somewhat strong predictor base_total – thinking about the game, stronger pokémon would seem to have certain values for these predictor variables when compared to weaker ones: for example, legendary pokémon tend to be stronger in battle than non legendary pokémon, so we might expect is_legendary to be a useful predictor for base_total, for example.

## Data description

We retrieved the dataset from kaggle.com, a large data science online community, and the dataset is called "The Complete Pokemon Dataset" created by Rounak Banik in 2017. The dataset was retrieved via web scraper from the website serebii.net, an all-in-one, reliable data hub for all things pokemon in 2017. Since it was formed in 2017, the dataset does not include pokémon from more recent games, but still includes a whopping 801 pokémon, meaning the dataset has 801 observations, one for each pokémon. The dataset contains the following variables, taken directly from the kaggle website for the dataset:

- name: The English name of the Pokemon
- japanese_name: The Original Japanese name of the Pokemon
- pokedex_number: The entry number of the Pokemon in the National Pokedex
- percentage_male: The percentage of the species that are male. Blank if the Pokemon is genderless.
- type1: The Primary Type of the Pokemon (every pokémon has this)
- type2: The Secondary Type of the Pokemon (not all pokémon have this)
- classification: The Classification of the Pokemon as described by the Sun and Moon Pokedex
- height_m: Height of the Pokemon in meters
- weight_kg: The Weight of the Pokemon in kilograms
- capture_rate: Capture Rate of the Pokemon
- base_egg_steps: The number of steps required to hatch an egg of the Pokemon
- abilities: A stringified list of abilities that the Pokemon is capable of having
- experience_growth: The Experience Growth of the Pokemon
- base_happiness: Base Happiness of the Pokemon
- against_?: Eighteen features that denote the amount of damage taken against an attack of a particular type (18 of these, one for each type)
- hp: The Base HP of the Pokemon
- attack: The Base Attack of the Pokemon
- defense: The Base Defense of the Pokemon
- sp_attack: The Base Special Attack of the Pokemon
- sp_defense: The Base Special Defense of the Pokemon

- speed: The Base Speed of the Pokemon
- generation: The numbered generation which the Pokemon was first introduced
- is_legendary: Denotes if the Pokemon is legendary.

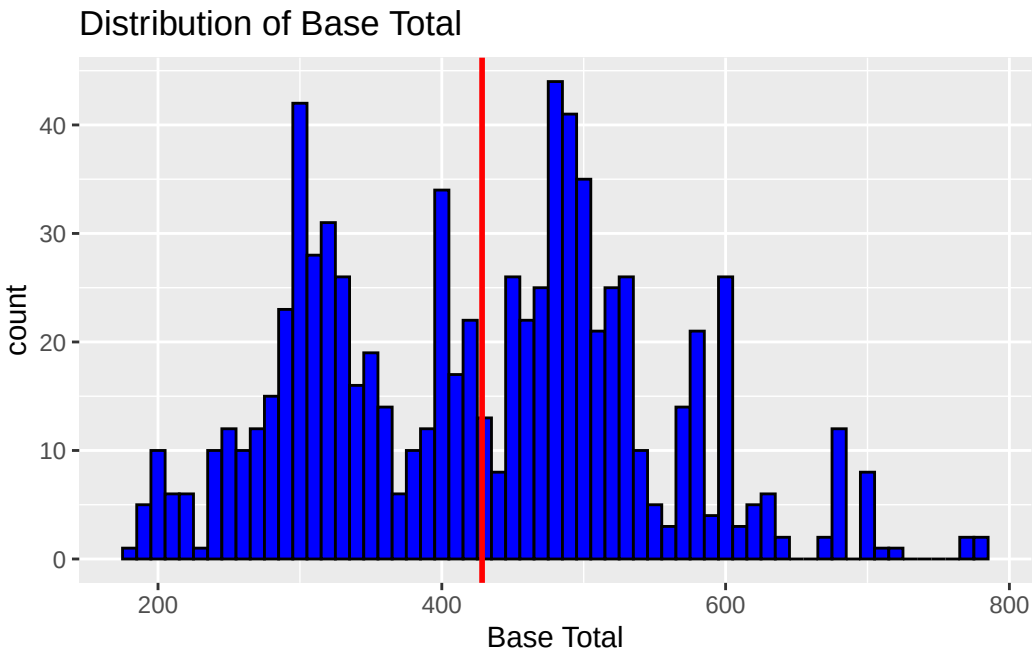
Source: (https://www.kaggle.com/datasets/rounakbanik/pokemon/)

As you'll notice, the dataset splits the base stats of each pokemon into the individual stats, but we only need to know about the base_total variable, which is included in the csv file. Each variable describes the pokémon at hand in a different way. Some, like is_legendary, may prove to be extremely important in our regression model, while with others, like name and Japanese name, we can remove them from consideration as they are simply unique identifiers.

## Initial exploratory data analysis

### Distribution of the response variable (base total):

```
ggplot(pokemon, aes(x = base_total)) +
  geom_histogram(binwidth = 10, fill = "blue", color = "black") +
  geom_vline(xintercept = mean(pokemon$base_total), color = "red", size = 1) +
  labs(title = "Distribution of Base Total", x = "Base Total")
```



```r
tidy(summary(pokemon$base_total))
```

```
# A tibble: 1 x 6
  minimum    q1 median  mean    q3 maximum
    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl>
1     180   320    435  428.   505     780
```

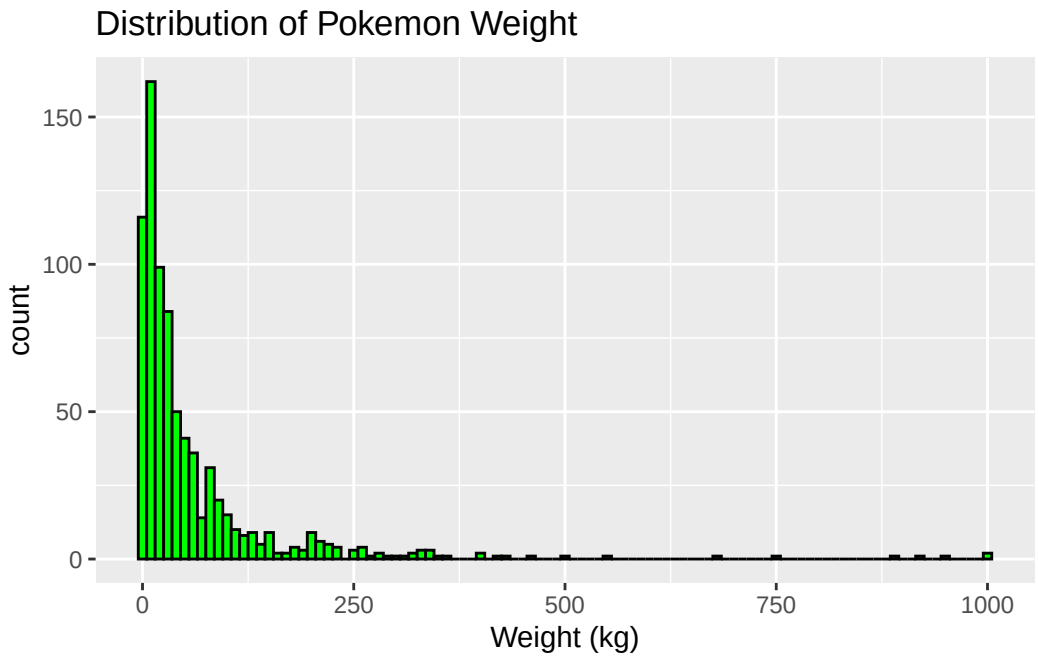
Description:

We can see from the distribution of the base total, that is seems to be roughly trimodal with three peaks (one around 300, 400, and 475). The base total points vary from about 180 to 780 with a mean at around 428.4. The median is 435 which is higher than the mean by a little bit which could mean that there are more extreme low base total Pokemon bringing the mean down a little bit. The IQR is 185 (505 - 320) and looking at the data there might be a potential outlier around 780 base total.

**Distributions of one potential quantitative predictor variable and one potential categorical predictor variable:**

- **Weight(quantitative predictor):**

```r
ggplot(pokemon, aes(x = weight_kg)) +
  geom_histogram(binwidth = 10, fill = "green", color = "black") +
  labs(title = "Distribution of Pokemon Weight", x = "Weight (kg)")
```




```
tidy(summary(pokemon$weight_kg))
```

```
# A tibble: 1 x 7
  minimum    q1 median  mean    q3 maximum    na
    <dbl> <dbl>  <dbl> <dbl> <dbl>   <dbl> <dbl>
1     0.1     9   27.3  61.4  64.8   1000.    20
```

Description:

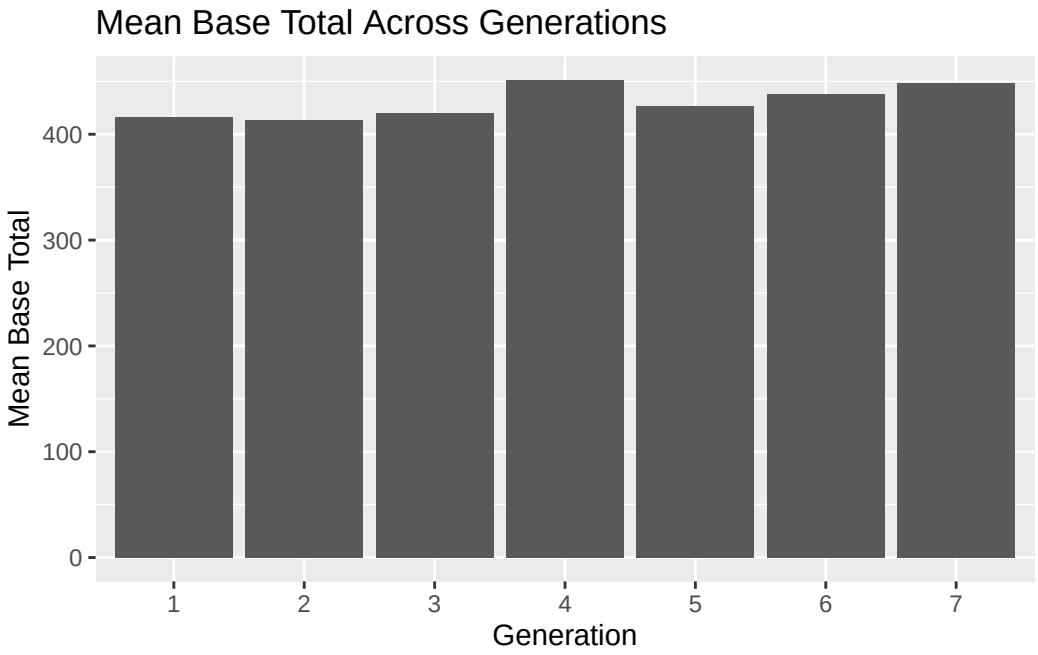
The distribution of the weights of the pokemons show a heavily right skewed distribution. The mean is around 61.38 while the median is only at around 27.3 which suggests the extremely heavy Pokemon are bringing the mean weight up a very significant amount whch leads to the mean being much higher than the median. The range of the weights of Pokemon show that the lightest one is around .1 kg while the heaviest is around 999.9 kg. Since there is a heavy right skew the median will be a better predictor for the average Pokemon weight so we can say the median weight of the Pokemon is around 27.3kg

- **Generation(qualitative predictor):**

```
generation_means <- pokemon |>
  group_by(generation) |>
  summarise(mean_base_total = mean(base_total)) |>
```

```
  arrange(desc(mean_base_total))

ggplot(generation_means, aes(x = as.factor(generation), y = mean_base_total)) +
  geom_bar(stat = "identity", show.legend = FALSE) +
  labs(title = "Mean Base Total Across Generations",
       x = "Generation",
       y = "Mean Base Total")
```



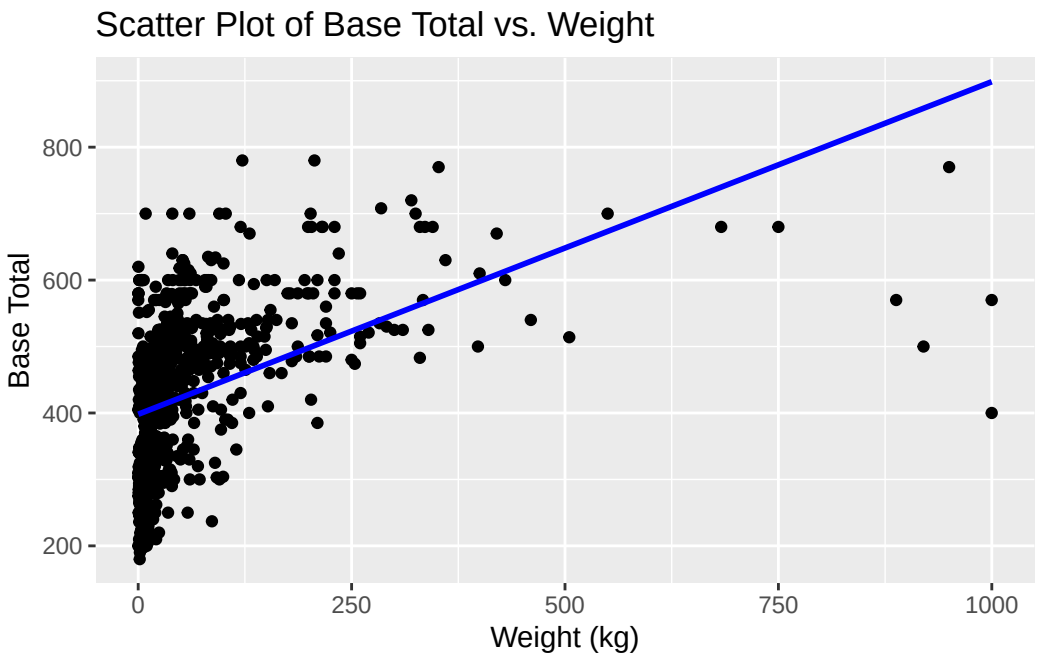

Description:

This is the graph showing the relationship between different generation Pokemon and their mean base total. We can see from this graph that the mean base total of generation 4 Pokemon were the highest but overall we can't see any distinct relationships between the generation of the Pokemon and their mean base total. One can maybe say there could be a very slight positive linear relationship between the generation of the Pokemon and their mean base total as we obsere that as generation increases the mean base total tends to increase slightly.

**Relationships between response and predictor from above:**

- **Weight and Base Total:**

```
ggplot(pokemon, aes(x = weight_kg, y = base_total)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, col = "blue") +
  labs(title = "Scatter Plot of Base Total vs. Weight", x = "Weight (kg)", y = "Base Total
```

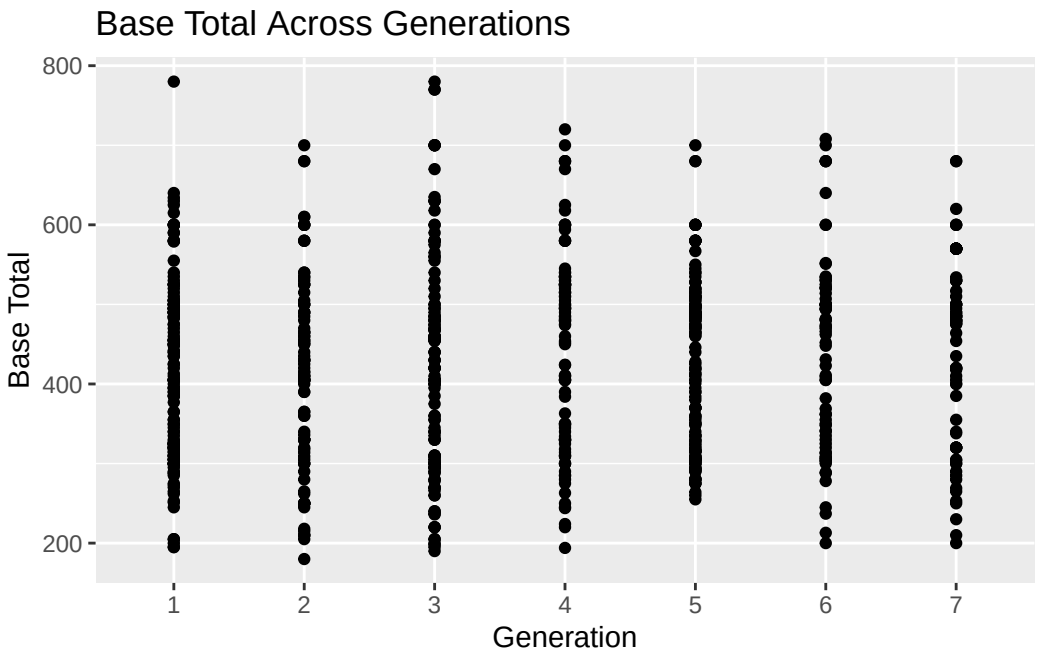



Description:

From the scatterplot I can see that there is no apparent correlation between the weight and the base total. There may be a very low positive correlation between these two variables but most of the points are focused around when the weight of the Pokemon is less than 100kg so it's hard to tell the relationship. There seems to be a few outliers near 900-1000kg.

- **Base Total and Generation:**

```
ggplot(pokemon, aes(x = as.factor(generation), y = base_total)) +
  geom_point() +
  labs(title = "Base Total Across Generations",
       x = "Generation",
       y = "Base Total")
```

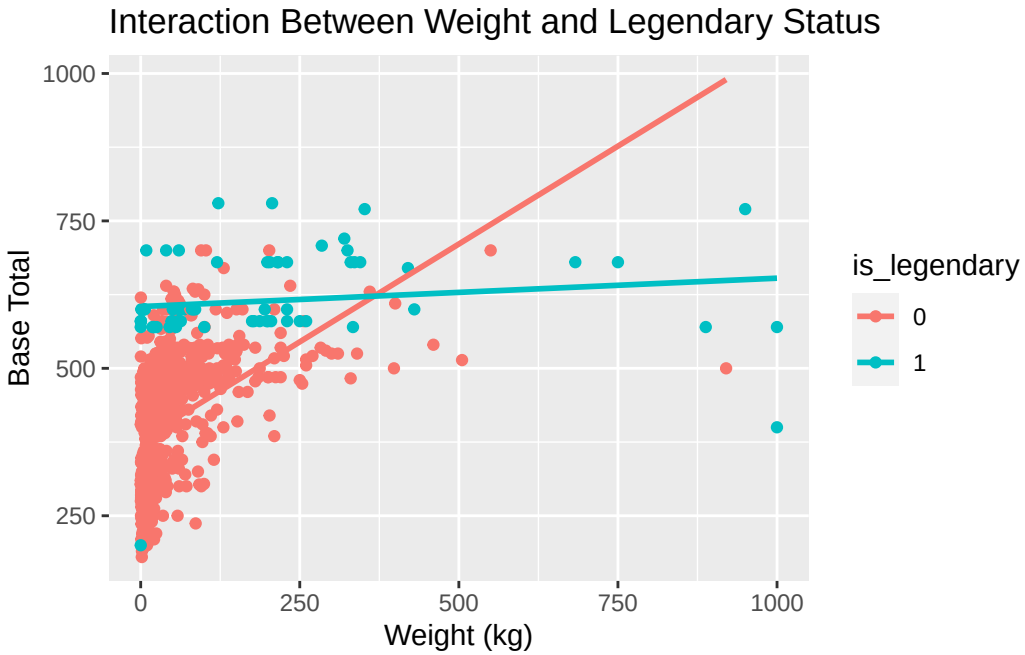



Description:

From the graph I can see that for each generation of Pokemon there seems to be a few Pokemon that have a base total that is comparably higher than the others but most of the Pokemon seems to be around 400 base total points for all generations and we can see that generation 1 had the highest base total Pokemon out of the 7 generations of Pokemon. We can also see that from the top base total Pokemon in each generation, generation 7 had the lowest rated one.

Potential Interaction Effect (Base total and is legendary):

```
pokemon$is_legendary <- as.factor(pokemon$is_legendary)
ggplot(pokemon, aes(x = weight_kg, y = base_total, color = is_legendary)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Interaction Between Weight and Legendary Status", x = "Weight (kg)", y = "
```




```
levels(pokemon$is_legendary)
```

```
[1] "0" "1"
```

Description:

This graph is showing the relationship between the weight of the Pokemon and the base total for legendary and non legendary Pokemon. The red line is for non legendary Pokemons (is_legendary = 0) and the blue line is for legendary Pokemon (is_legendary = 1). Something interesting we can note is that legendary Pokemon all tend to have a much higher and consistent base total for all weights. Across all weights most base total for legendary Pokemons stay between 550- 700. Among the legendary Pokemon there doesn't seem to be any correlation between the weight and the base total. Among the non-legendary Pokemon however, there seems to be a very slight positive correlation between the weight of the Pokemon and the base total. As the weight of the non-legendary Pokemon goes up, the Base total seems to go up slightly although with the cluster of points under 100 kg Pokemon, it is hard to conclude, and regardless, the correlation does not appear to be strong between these variables.

Note that from this graph we can tell that Legendary pokemon appear to tend to have much higher base stat totals than non-legendary pokemon on average, which is something that will definitely impact our multiple linear regression model.

## Analysis approach

The response variable is **base_total**, which is the total of all base stats (attack, special attack, defense, special defense, speed, and hitpoints) for a given pokémon. It is a generally great measure of a pokémon's strength.

The potential predictor variables include:

- percentage_male
- type1
- type2
- height_m
- weight_kg
- capture_rate
- base_egg_steps
- abilities
- experience_growth
- base_happiness
- against_? (18 of these, one for each type)
- generation
- is_legendary

As our output variable is quantitative and we have several predictor variables, the regression model we are using is **multiple linear regression**.

## Data dictionary

The data dictionary can be found here