8.3 Final Project Step 2

Harlan Wittlieff

11/7/2021

Problem statement

Do physical characteristics (height, weight, and gender) impact Pokemon base stats?

How to import and clean my data.

Data Import

```
Pokemon_df <- read.csv("Data/pokemon.csv")
```

Data Cleaning

- The weight_kg field has missing values.
- The height_m field has missing values.
- The % male field contains NA values for certain Pokemon. This is not missing data, but instead an "Unknown" designation used in the Pokemon universe for Pokemon that do not have a set gender.
- The original data set contains 41 variables. This has been reduced to 10 for this analysis.
  - For example, the data set contains a field for japenese_name. This variable is not needed for this analysis and has been removed.

For the missing values, these are manually looked up from the source data on serebii.net and added to the data set. An example of the code to correct missing values is listed below.

```
Pokemon_df[19, "height_m"] = 0.7
Pokemon_df[19, "weight_kg"] = 3.5
```

What does the final data set look like?

- The final data set contains base stats and physical characteristics for 801 different Pokemon.
  - Physical characteristics
    * height
    * percentage_male
    * weight_kg

  - Base stats
    * attack
    * defense
    * hp (hit points)
    * sp_attack
    * sp_defense
    * speed
  - Pokemon Name
- A subset of the data used is included below.

```
head(pokemon_df_reduced)
```

```
##         name attack defense height_m hp percentage_male sp_attack sp_defense
## 1  Bulbasaur     49      49      0.7 45            88.1        65         65
## 2    Ivysaur     62      63      1.0 60            88.1        80         80
## 3   Venusaur    100     123      2.0 80            88.1       122        120
## 4 Charmander     52      43      0.6 39            88.1        60         50
## 5 Charmeleon     64      58      1.1 58            88.1        80         65
## 6  Charizard    104      78      1.7 78            88.1       159        115
##   speed weight_kg
## 1    45       6.9
## 2    60      13.0
## 3    80     100.0
## 4    65       8.5
## 5    80      19.0
## 6   100      90.5
```

What information is not self-evident?

- To discover new information that is not self-evident, correlations between variables will be investigated. Additionally, a regression model will be built and statistical significance determined.

What are different ways you could look at this data? / How do you plan to slice and dice the data?

- Physical characteristics can be compared to base stats individually.
  - For example, a regression model could be built that attempts to predict attack based on height, percentage_male, and weight_kg.
- A new variable could be created from the base stats that summarizes these variables into one metric. This metric will then be used and the dependent variable to be predicted from the physical characteristics.
  - Option A will build a ranking system based on a specific Pokemon's percentile for each individual metric and then combine these rankings into one metric.
  - Option B will combine these fields into one metric based on the total sum.

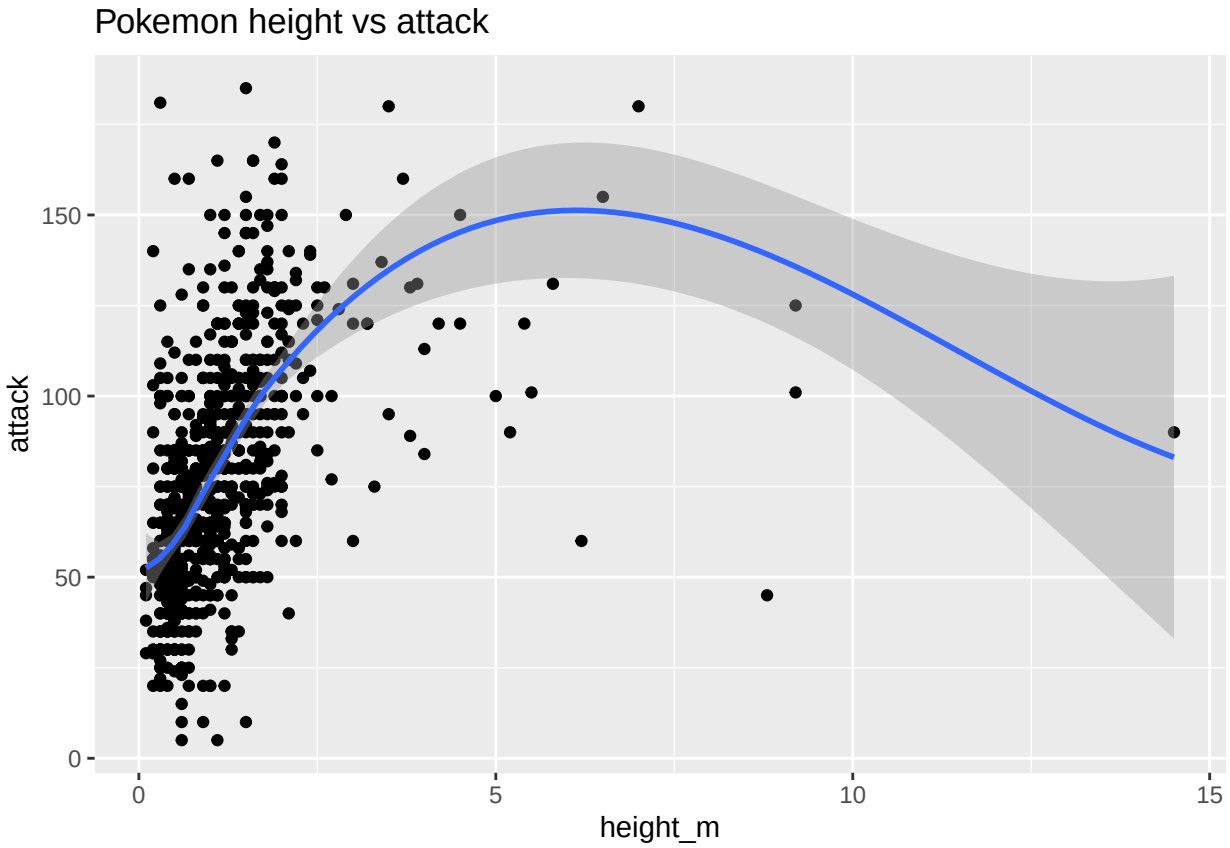
How could you summarize your data to answer key questions?

Results from the regression model will be reported.

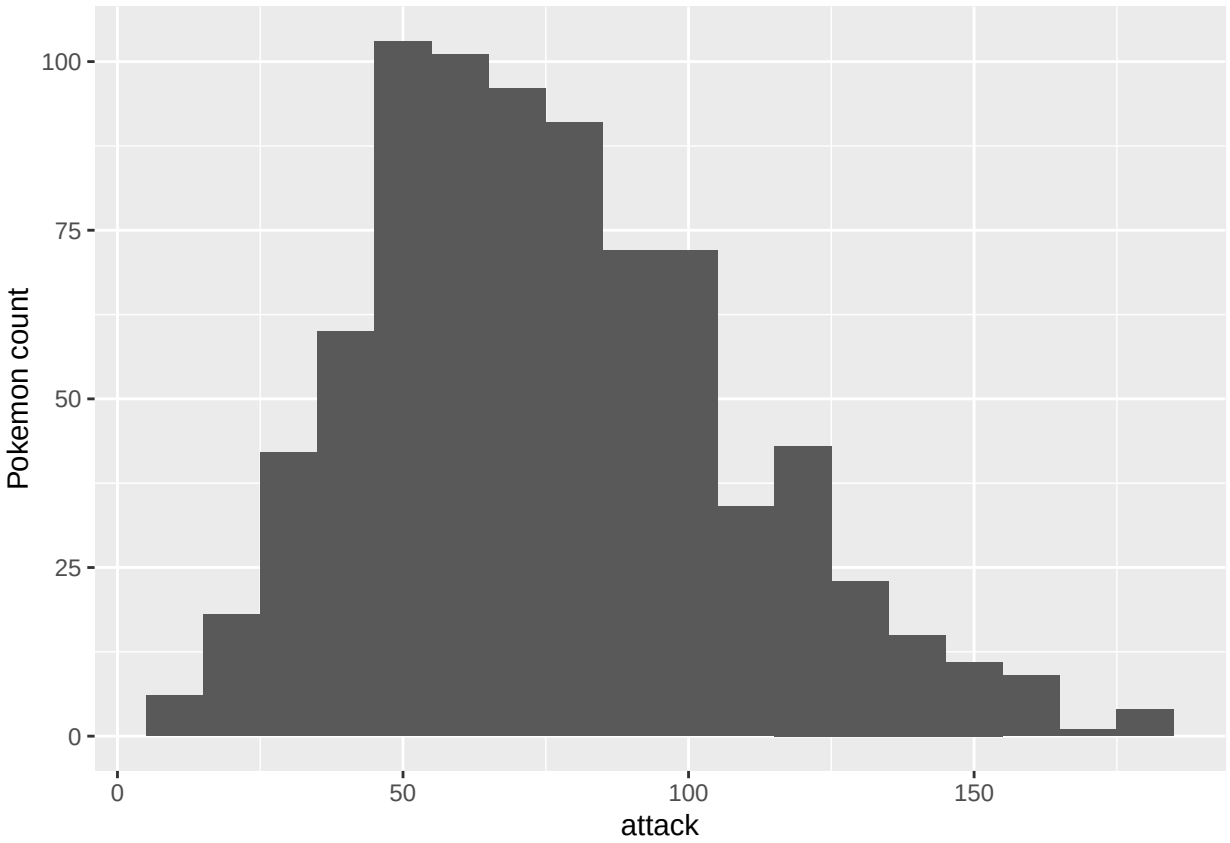
What types of plots and tables will help you illustrate the findings to your questions?

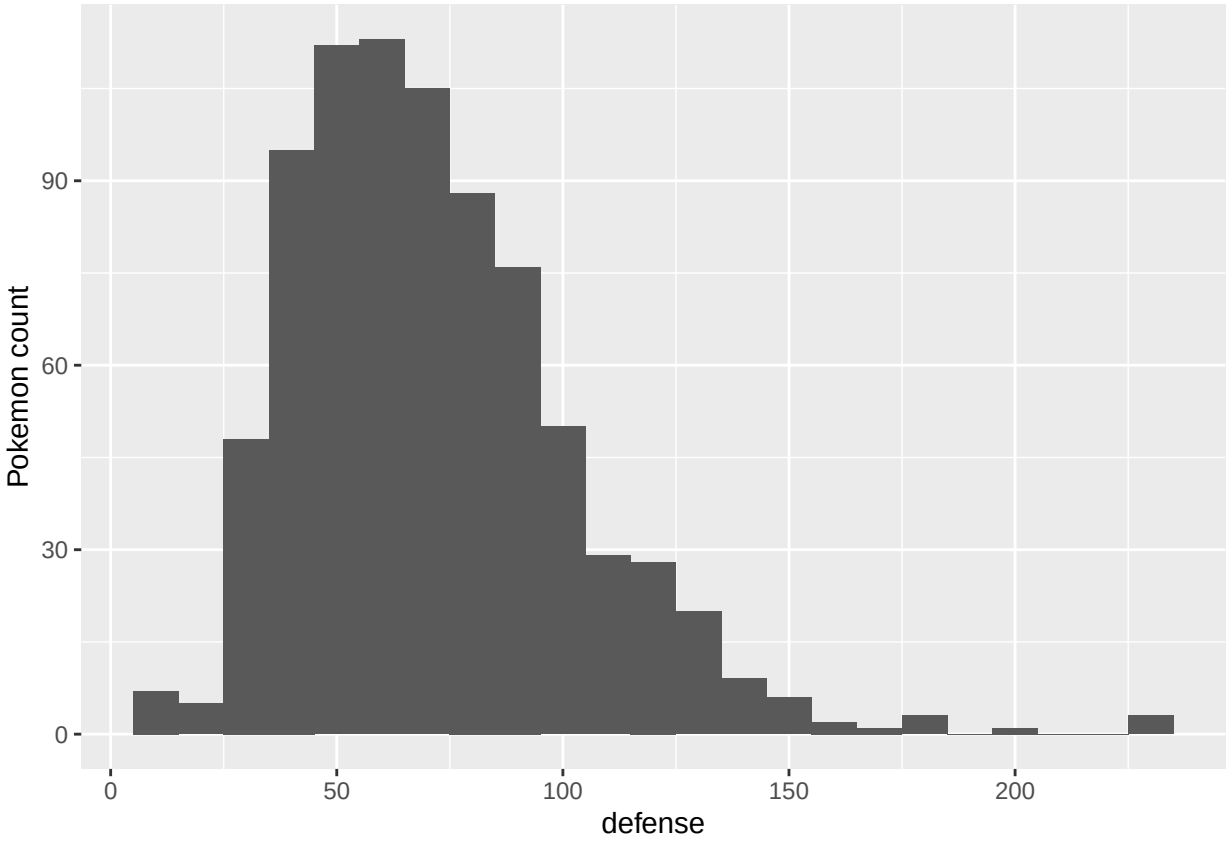
- To gain a visual understanding of the impacts, various physical characteristics can be charted against base stats in a scatter plot.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



To communicate the distribution of base stats across all Pokemon, histograms will be used.


100
75
50
25
0
Pokemon count
0
50
100
150
attack



Do you plan on incorporating any machine learning techniques to answer your research questions?

- Yes, a regression model will be built. Statistical significance of physical characteristic's impact on base stats will be investigated.

Questions for future steps.

1. Additional research on how to handle the "Unknown" gender characteristic is needed.
2. Assumptions for a regression model will have to be validated.