Pokemon Project - Gotta Catch ’Em All!

Casey Hutchinson

2023-11-28

Pokemon Project - Gotta Catch ’Em All!

Introduction:

I have been playing Pokemon since it’s rise in popularity began in the United States. I’ve collected trading cards (still have most of them), watched the anime (still do), and played just about every game on every video game platform imaginable (and will continue to). The point of this report/project is to dive deeper into the data and see what new information I can learn, but also see if the beliefs I’ve created as a Pokemon trainer over the years of playing are backed up by the data using the techniques from our previous projects.

My strongest belief as a Pokemon trainer is that you need to have Pokemon in your party that are generally from a wide variety of types, but all must posses at least an above average attack level. I prefer Pokemon with strong attacks, but I also recognize the importance of defense and speed. I’ve never paid too much attention to special attack or special defense, but I’m going to look at those stats a little closer and see if my opinions change.

We’ll also look at stats like capture rate and base total (the combined point total of attack, defense, hp, sp_attack, sp_defense, and speed stats), and also split the pokemon into smaller datasets to look at certain types of Pokemon.

In the last section I’m going to look more specifcally at the first and second generation of pokemon. These are the generations of Pokemon I am most familiar with, and I am certain beyond all doubt that the first generation is better and that the second generation can get off my lawn! I’ll look at certain subsets of the data, and take a closer look at a couple of my favorite Pokemon.

I found the original dataset on Kaggle that was put together by Rounak Banik using data scraped from serebii.net - a Pokemon database that I have used many times while playing through various games. I refined his original dataset to exclude a few variables, but I have the original included with my refined version xlsx file.

Data

The data for this project can be found here: Pokemon Data (https://www.kaggle.com/datasets/rounakbanik/pokemon/data)

Main Dataset

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.2
```

```
pokemon_refined <- read_excel("D:/School/Fall 2023/Data Exploration and Visualization/Final Proj
ect/archive/pokemon.refined.xlsx")
```

I will also add a few smaller subsets at various points throughout this project. I will define those names before I present them.

Question 1) What is the average level of attack for all Pokemon? What is the median level of attack for all Pokemon?

```
mean(pokemon_refined$attack)
```

```
## [1] 77.85768
```

```
median(pokemon_refined$attack)
```

```
## [1] 75
```

If attack is the most important factor in determining who will win a Pokemon battle, I want to know what attack values the weakest and strongest Pokemon have. Which Pokemon have these attack levels?

```
min(pokemon_refined$attack)
```

```
## [1] 5
```

```
max(pokemon_refined$attack)
```

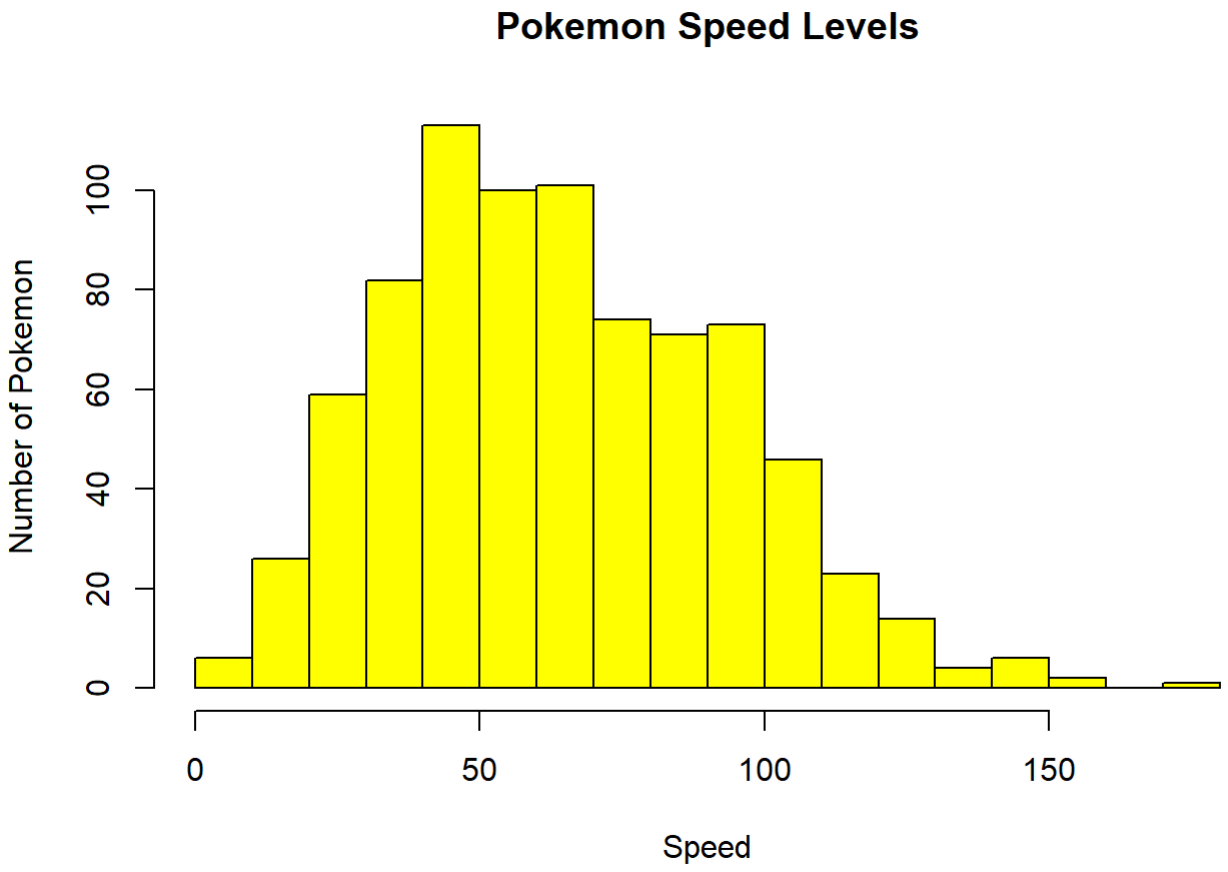
```
## [1] 185
```

The lowest attack level amongst all Pokemon is 5, shared by both Happiny and Chansey. This makes sense, as Happiny evolves into Chansey, and Chansey is renowned in the Pokemon world for their healing capabilities. They are often found working in Pokemon Centers helping heal sick and injured Pokemon, so naturally their attack would be low. The highest attack level we see is 185, and the only Pokemon with an attack this high is Heracross - a fighting AND bug type Pokemon. Fighting types typically have high attack levels, and bug Pokemon can give you useful type advantages in certain battles.

Question 2) Speed is an important factor to consider when building a Pokemon team - Pokemon with higher speed levels tend to go first in battles!

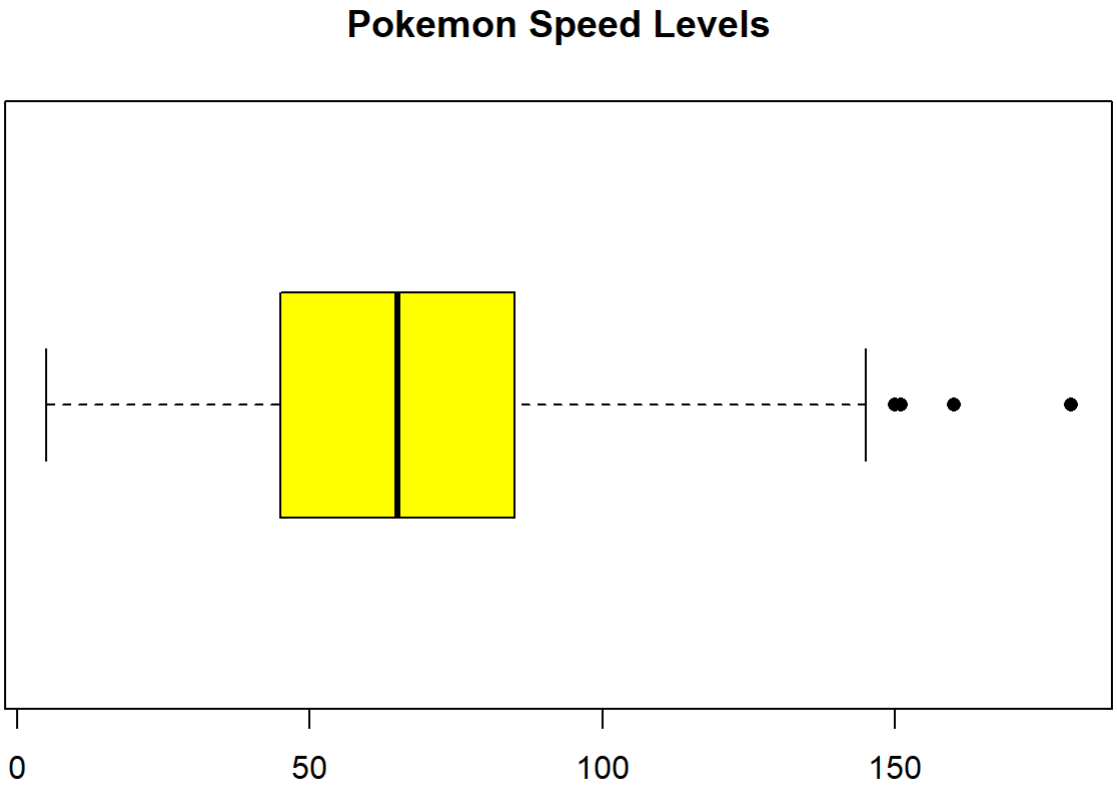
Construct a histogram and describe the shape of the Pokemons speed levels.

```
hist(pokemon_refined$speed, main = "Pokemon Speed Levels", xlab = "Speed", ylab = "Number of Pok
emon", col="yellow", breaks = 20)
```



Use a horizontal boxplot to identify whether there are any outliers present.

```
boxplot(pokemon_refined$speed, main = "Pokemon Speed Levels", horizontal = TRUE, pch = 16, col =
"yellow")
```



The histogram for Speed levels looks mostly normally distributed, but on the high end of the Speed levels you can see there are a few pokemon that skew it to the right. The boxplot for Speed levels back this up, as you can see there are at least 4 pokemon who have speed levels that are considered outliers. 3 of those Pokemon are from the first generation: Alakazam, Electrode, and Aerodactyl! Aerodactyl is a rock type Pokemon so it's very surprising to see him as an outlier for speed.

Question 3) Which is more spread out, attack stats or defense stats?

We'll start by looking at a few key data points for each: the mean, median, standard deviation, and variance.
Attack:

```
mean(pokemon_refined$attack)
```

```
## [1] 77.85768
```

```
median(pokemon_refined$attack)
```

```
## [1] 75
```

```
sd(pokemon_refined$attack)
```

```
## [1] 32.15882
```

```
var(pokemon_refined$attack)
```

```
## [1] 1034.19
```

Defense:

```
mean(pokemon_refined$defense)
```

```
## [1] 73.00874
```

```
median(pokemon_refined$defense)
```

```
## [1] 70
```

```
sd(pokemon_refined$defense)
```

```
## [1] 30.76916
```

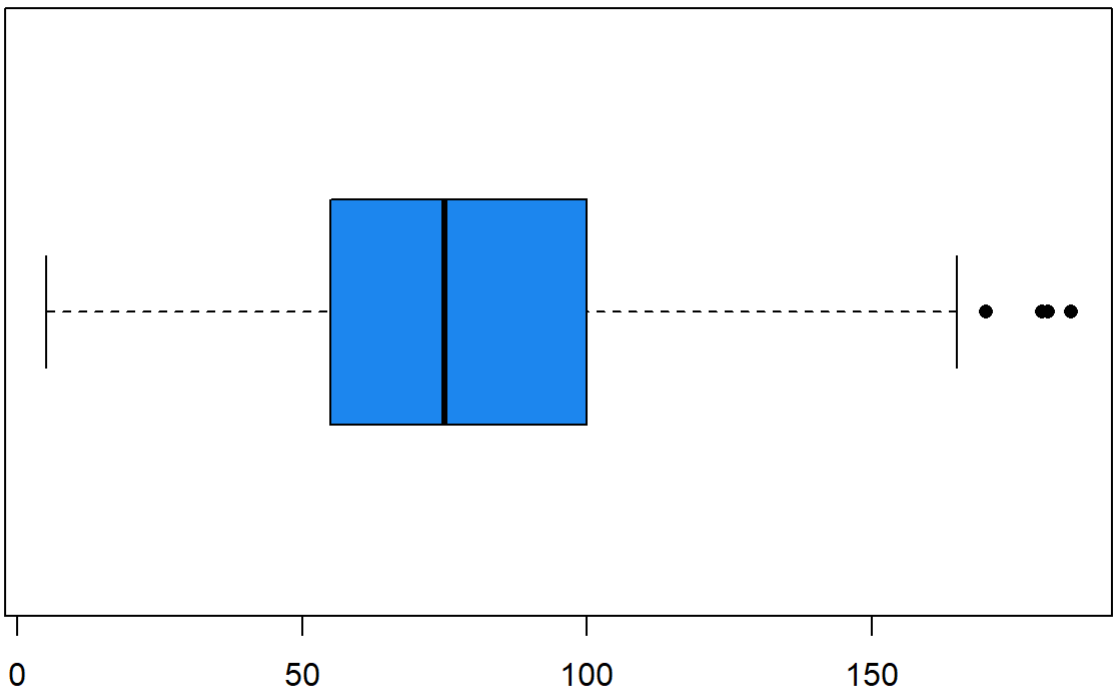
```
var(pokemon_refined$defense)
```

```
## [1] 946.7412
```

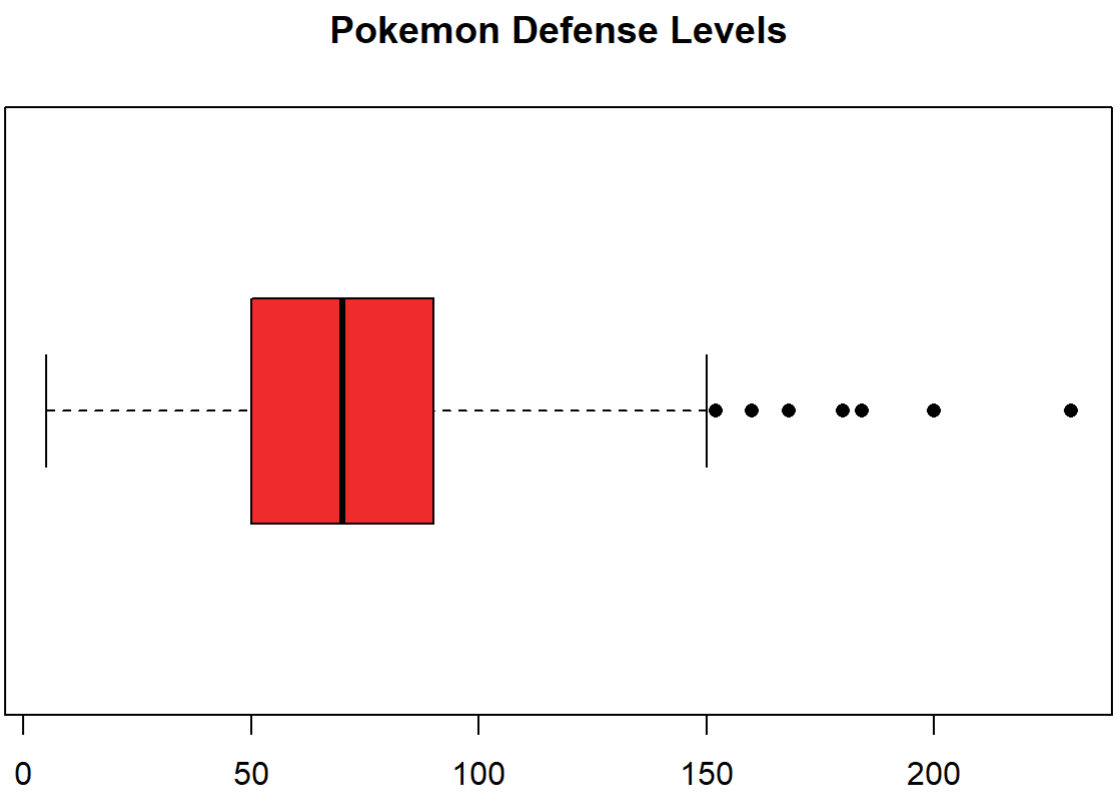
Both stats are relatively similar, and when you compare their variances both attack and defense levels among Pokemon have a wide range. This isn't surprising, as there are a ton of Pokemon types out there. I think it's fair to say that attack is a bit more spread out than defense, and I wouldn't have thought this. I figured attack stats would be much more compact, but perhaps defense is more important for Pokemon out in the wild as opposed to those training for Pokemon battles with a trainer.

Question 4) Construct horizontal boxplots for both. Which appears more symmetric?

```
boxplot(pokemon_refined$attack, main = "Pokemon Attack Levels", col="dodgerblue2", horizontal = TRUE, pch = 16)
```




```
boxplot(pokemon_refined$defense, main = "Pokemon Defense Levels", col="firebrick2", horizontal = TRUE, pch = 16)
```



Both have several outliers, but defense still has quite a few more than attack (the highest attack outlier belongs to Heracross at level 185 and the highest defense outlier belongs to 3 different Pokemon). This leads to quite a bit more skew for the defense stat, leaving the boxplot for attack looking a little more symmetric.

Question 5) If I still consider attack as the most important stat, I want to focus on raising and catching Pokemon with a strong rating in that category. What is the minimum attack stat I should look for if I only want to focus on catching Pokemon that are among top 25% in that category?

```
quantile(pokemon_refined$attack, 0.75)
```

```
## 75%
## 100
```

The minimum attack level for a Pokemon in the top 25% of attack is 100.

What if I apply the same logic to the defense stat?

```
quantile(pokemon_refined$defense, 0.75)
```

```
## 75%
##  90
```

The minimum defense level for a Pokemon in the top 25% of defense is 90.

And what if I apply it to the speed stat?

```
quantile(pokemon_refined$speed, 0.75)
```

```
## 75%
##  85
```

The minimum speed level for a Pokemon in the top 25% of speed is 85.

Question 6) This time I'm going to see what percentage of Pokemon would be eliminated from consideration if I stopped considering below level 30 for each stat. To do this, I'm going to use the pnorm command. I'm also going to find the mean and standard deviation for each stat so I can use that command.

Attack:

```
AM <-mean(pokemon_refined$attack); AM
```

```
## [1] 77.85768
```

```
ASD <-sd(pokemon_refined$attack); ASD
```

```
## [1] 32.15882
```

```
pnorm(30, AM, ASD)
```

```
## [1] 0.0683535
```

Eliminating Pokemon from consideration on my team if they have an attack level below 30 would eliminate approximately 7% of all available Pokemon. I think this is a reasonable amount, and I may even be willing to go a little higher in the future.

Defense:

```
DM <-mean(pokemon_refined$defense); DM
```

```
## [1] 73.00874
```

```
DSD <-sd(pokemon_refined$defense); DSD
```

```
## [1] 30.76916
```

```
pnorm(30, DM, DSD)
```

```
## [1] 0.08108848
```

Eliminating Pokemon from consideration on my team if they have a defense level below 30 would eliminate approximately 8% of all available Pokemon. This is also a reasonable amount, and may also warrant raising in the future.

Speed:

```
SM <-mean(pokemon_refined$speed); SM
```

```
## [1] 66.33458
```

```
SSD <-sd(pokemon_refined$speed); SSD
```

```
## [1] 28.90766
```

```
pnorm(30, SM, SSD)
```

```
## [1] 0.1043915
```

Eliminating Pokemon from consideration on my team if they have a speed level below 30 would eliminate approximately 10% of all available Pokemon. This is borderline reasonable, but I want to be careful not to eliminate too much of my available pool. I also want to factor in low-level Pokemon that may evolve and

grow their stats as they evolve - just because they are below the levels I have laid out now doesn't mean they can't be very strong down the line!

Question 7) Using the filter() command from the dplyr library, I'm going to create a new dataset that is comprised of only legendary Pokemon. I'm going to call this data set legendary. I'll also create a new dataset that is comprised of only non-legendary Pokemon. I'm going to call this data set non.legendary.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
legendary <- filter(pokemon_refined, is_legendary == 1)
non.legendary <- filter(pokemon_refined, is_legendary == 0)
```

Question 8) Compare the average attack level for legendary Pokemon to the attack level of non-legendary Pokemon. Do the same for defense. Which stat has the greater difference between legendary and non-legendary Pokemon?

Attack:

```
mean(legendary$attack)
```

```
## [1] 109.3571
```

```
mean(non.legendary$attack)
```

```
## [1] 74.84131
```

```
mean(legendary$attack) - mean(non.legendary$attack)
```

```
## [1] 34.51583
```

Defense:

```
mean(legendary$defense)
```

```
## [1] 99.4
```

```
mean(non.legendary$defense)
```

```
## [1] 70.48153
```

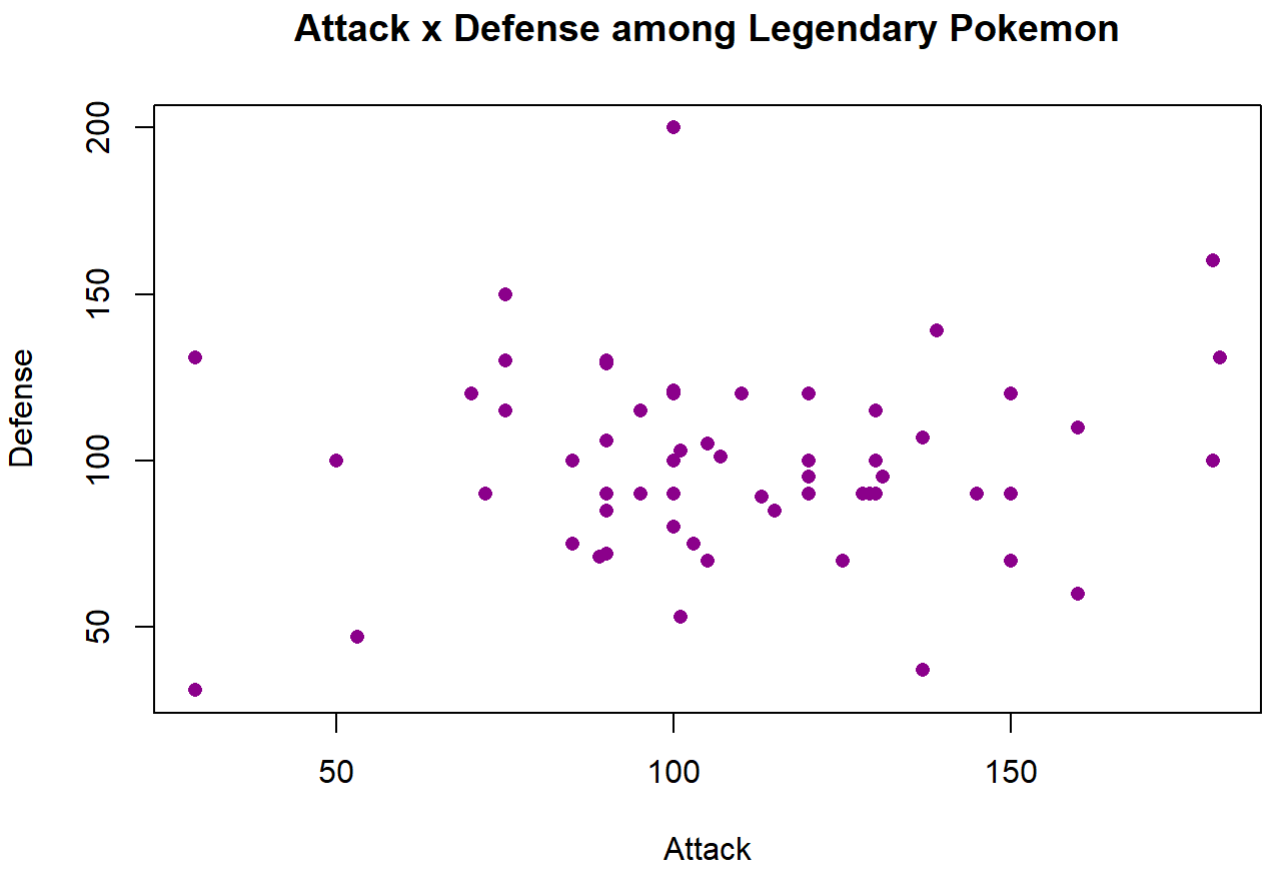
```
mean(legendary$defense) - mean(non.legendary$defense)
```

```
## [1] 28.91847
```

There is a significant difference between both average attack and average defense levels among legendary and non-legendary Pokemon. This shouldn't be too much of a surprise, it's not a prerequisite that a legendary Pokemon be physically powerful, but it seems like it's the most common trait. The difference of average attacking levels is slightly higher than average defense levels.

Question 9) Create a scatterplot that balances attack on one axis and defense on the other for legendary Pokemon. Comment on the type of association (if any).

```
plot(legendary$attack, legendary$defense, col="darkmagenta", main="Attack x Defense among Legend
ary Pokemon", xlab="Attack", ylab="Defense", pch=16)
```




```
cor(legendary$attack, legendary$defense)
```

```
## [1] 0.1012143
```

I didn't expect there to be a strong correlation between the 2 types, and the scatterplot and correlation back that up. It's interesting to see the wide variety of legendary Pokemon out there! Some have high attack and defense stats but some don't have high stats in either. It's also not surprising to not see very many below level 50 in either category.

Question 10) Legendary Pokemon are obviously not as common as non-legendary Pokemon, but are they more difficult to catch? Compare the average capture rate between the two different datasets.

```
mean(legendary$capture_rate)
```

```
## [1] 17.98571
```

```
mean(non.legendary$capture_rate)
```

```
## [1] 106.71
```

Unsurprisingly, legendary Pokemon are much harder to capture on average than non-legendary Pokemon! As a trainer, I probably won't spend too much time trying to catch any unless I need them to fill out my Pokedex.

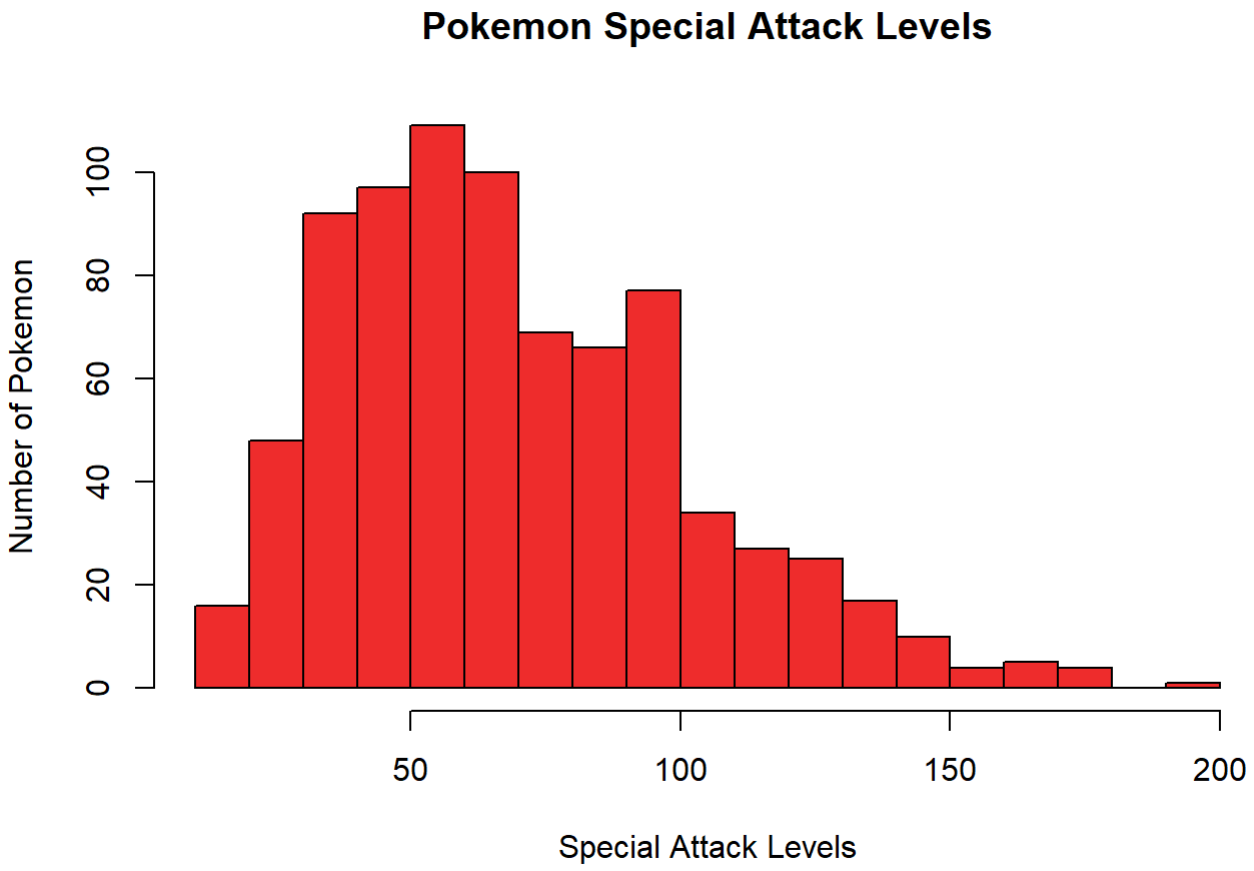
Question 11) Construct histograms of Pokemon special attack levels and Pokemon special defense levels and use them to answer the following questions:

a) Would you classify each as symmetric, skewed to the left, or skewed to the right?

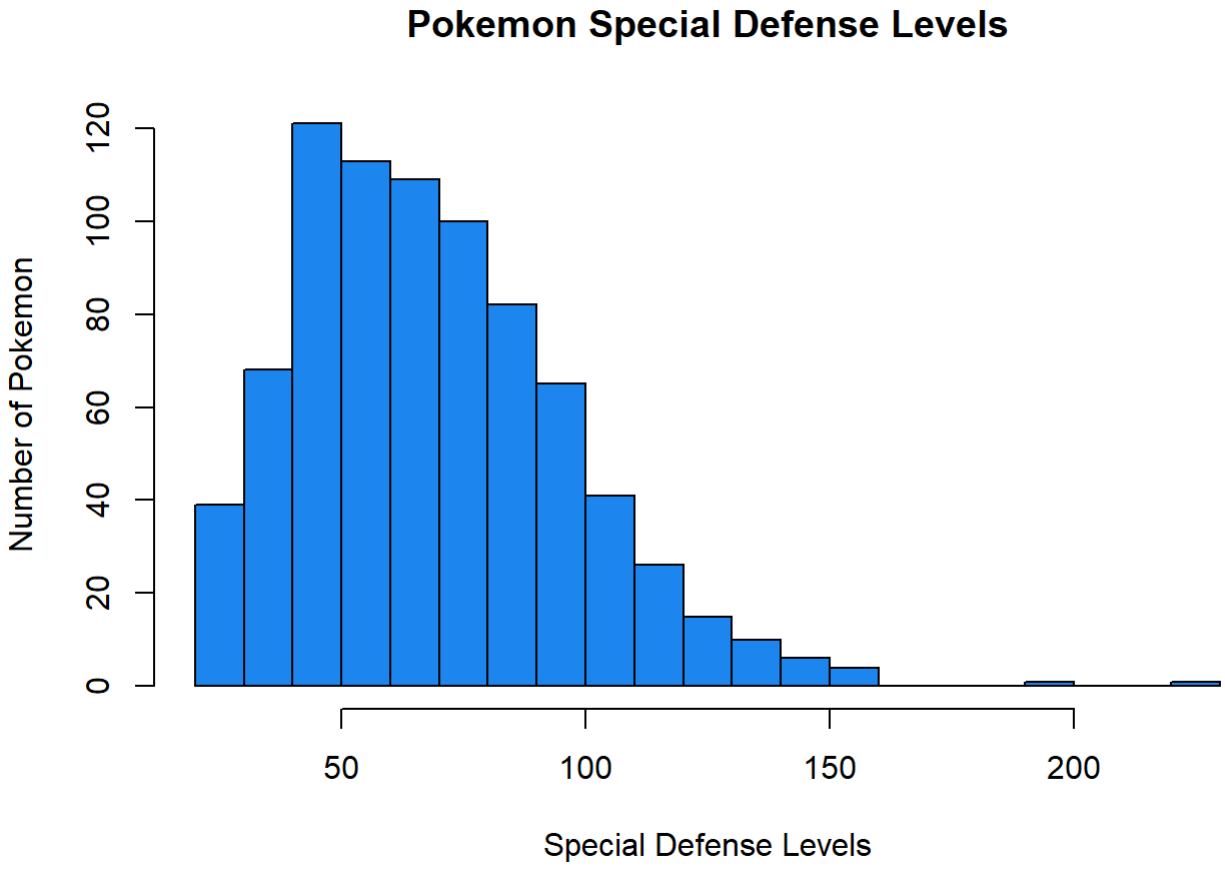
b) Would you classify each is unimodal or bimodal?

c) Does either appear to be uniform?

```
hist(pokemon_refined$sp_attack, main = "Pokemon Special Attack Levels", xlab = "Special Attack L
evels", ylab = "Number of Pokemon", col = "firebrick2", breaks = 20)
```




```
hist(pokemon_refined$sp_defense, main = "Pokemon Special Defense Levels", xlab = "Special Defens
e Levels", ylab = "Number of Pokemon", col = "dodgerblue2", breaks = 20)
```



a) I would classify both as skewed to the right. It appears to me that Special Defense has more significant outliers, so it's skewed much more than Special Attack. b) They are both also unimodal. Special Attack has a mini-peak that forms around level 100, but it doesn't reach the level that it's highest peak does so I would still consider it to be unimodal. c) Neither is uniform. This isn't much of a surprise when we consider the rest of the data we've observed.

Question 12) Now I'm going to create a new dataset comprised of only the 151 Pokemon from the first generation. I'm going to do this using the filter() command to create the OG.Pokemon dataset. I'm also going to create a second dataset comprising of only the Pokemon from the second generation called SG.Pokemon

```
OG.Pokemon <-filter(pokemon_refined, generation == "1")
SG.Pokemon <-filter(pokemon_refined, generation == "2")
```

Question 13) On average, which generations Pokemon has a higher base total level?

```
mean(OG.Pokemon$base_total)
```

```
## [1] 416.2517
```

```
mean(SG.Pokemon$base_total)
```

```
## [1] 413.18
```

The base totals for each generation are pretty close, but generation 1 has a slight edge!

On average, which generations Pokemon has a higher attack level?

```
mean(OG.Pokemon$attack)
```

```
## [1] 74.5298
```

```
mean(SG.Pokemon$attack)
```

```
## [1] 69.96
```

Again the attack averages for each generation are pretty close, but again generation 1 has a slight edge!

On average, which generations Pokemon has a higher defense level?

```
mean(OG.Pokemon$defense)
```

```
## [1] 70.07947
```

```
mean(SG.Pokemon$defense)
```

```
## [1] 71.79
```

Here we see that the second generation finally wins a category, barely beating the 1st generation with a slightly higher average defense level.

Question 14) My two favorite Pokemon types are Fire and Water. I'm going to create 2 new subsets of data for the OG dataset (Fire.OG and Water.OG) and SG dataset (Fire.SG and Water.SG) and then compare them.

```
Fire.OG <-filter(OG.Pokemon, type1 == "fire")
Water.OG <-filter(OG.Pokemon, type1 == "water")
Fire.SG <-filter(SG.Pokemon, type1 == "fire")
Water.SG <-filter(SG.Pokemon, type1 == "water")
```

On average, which generations Fire types have a higher base total?

```
mean(Fire.OG$base_total)
```

```
## [1] 463.9167
```

```
mean(Fire.SG$base_total)
```

```
## [1] 444.125
```

The first generations Fire type Pokemon have a higher base total average by almost 20 points!

On average, which generations Water types have higher base total?

```
mean(Water.OG$base_total)
```

```
## [1] 421.9286
```

```
mean(Water.SG$base_total)
```

```
## [1] 420.7778
```

The average base total between the 2 generations water type Pokemon are almost identical, but the first generation has another slight edge.

Question 15) What Pokemon among each generations Fire type Pokemon has the highest attack stat?

```
max(Fire.OG$attack)
```

```
## [1] 130
```

```
max(Fire.SG$attack)
```

```
## [1] 130
```

Each generations Fire Pokemon with the highest attack stat is at level 130. It's Flareon in the first generation and Ho-Oh in the second generation. It should be noted that Ho-oH is a legendary Pokemon (as is is second place Entei at level 115). The non-legendary Pokemon with the highest attack stat is Typhlosion at level 84.

What Pokemon among each generations Water type Pokemon has the highest defense stat?

```
max(Water.OG$defense)
```

```
## [1] 180
```

```
max(Water.SG$defense)
```

```
## [1] 115
```

The highest defense level for any Pokemon in the first generation is level 180 - and two different Pokemon actually meet this level! Both Slowbro and Cloyster have a defense level of 180. The highest defense level in the second generation belongs to another legendary Pokemon, Suicune. The highest defense level of a non-legendary Pokemon belongs to Feraligatr at level 100.

I previously mentioned Flareon in one of the answers. Flareon is the fire evolution of Eevee who also has a water evolution - Vaporeon. Flareon and

Vaporeon both have a speed level of 65. Find the z-score of each and see who is quicker relative to their type.

```
MFOG <-mean(Fire.OG$speed)
SFOG <-sd(Fire.OG$speed)
(65 - MFOG)/SFOG
```

```
## [1] -1.146494
```

```
MWOG <-mean(Water.OG$speed)
SWOG <-sd(Water.OG$speed)
(65 - MWOG)/SWOG
```

```
## [1] -0.1262005
```

By examining the z-score we can see that Vaporeon is more than one standard deviation below the mean at -1.15. Vaporeon isn't exactly speedy with a z-score of -.13, but relative to their type Vaporeon is much quicker!