Analyzing Categorical Datasets using Cleveland Dot Plots

STAT133 Fall 2017 Post #1

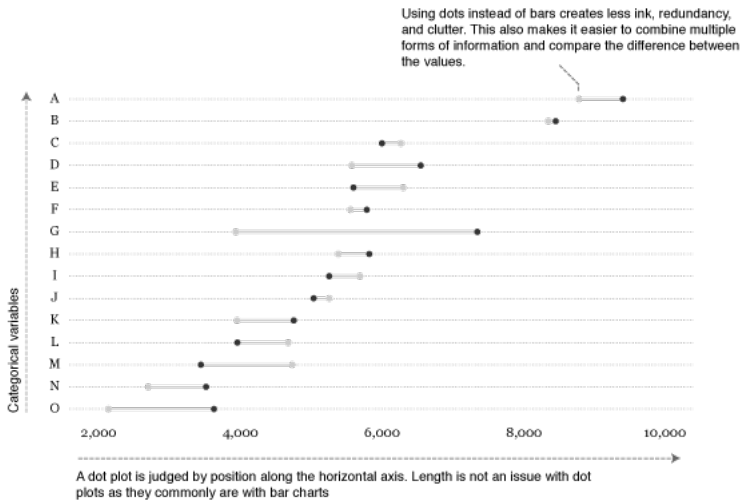
Alan Chuang

10/31/2017

Let's analyze some data!

Introduction to Cleveland Dot Plots

In the world where big data has become so mainstream, being able to process and understand data is increasingly important. Throughout the data analysis cycle, data must be prepared, analyzed, and finally reported. However, if results are not properly and effectively reported, the gargantuan efforts spent preparing and analyzing data are wasted. One vital aspect of reporting data involves data visualizations. There are many different types of ways to visualize data: graphs, charts, tables, and more. A common, basic data visualization is the bar chart. The simplicity of the bar chart makes it very easily understandable by readers. However, when comparing bar charts, sometimes it is difficult to visually identify all the minute differences between charts. The Cleveland Dot Plot, designed to take advantage of the judgements readers make when looking at data visualizations, solves this problem. In this post, we will illustrate the usefulness of the Cleveland Dot Plot by analyzing some data about characters from the classic, popular game Pokemon!



An example of a Cleveland Dot Plot.

Set-Up

To create Cleveland Dot Plots, we will need to load some packages in R.

```
library(readr) #for importing the data
library(dplyr) #for data manipulation and wrangling
library(ggplot2) #for creating the data visualizations
```

In this post, I will be working with some data involving Pokemon. All the data comes from this source. If you're following along, try to work with a data set that involves some sort of categorical data. In general, we want to use the plots we're about to create when working with categorical data.

```
data = read_csv(file = '../Data/pokemon.csv')
```

Comparing Attack Points by Type

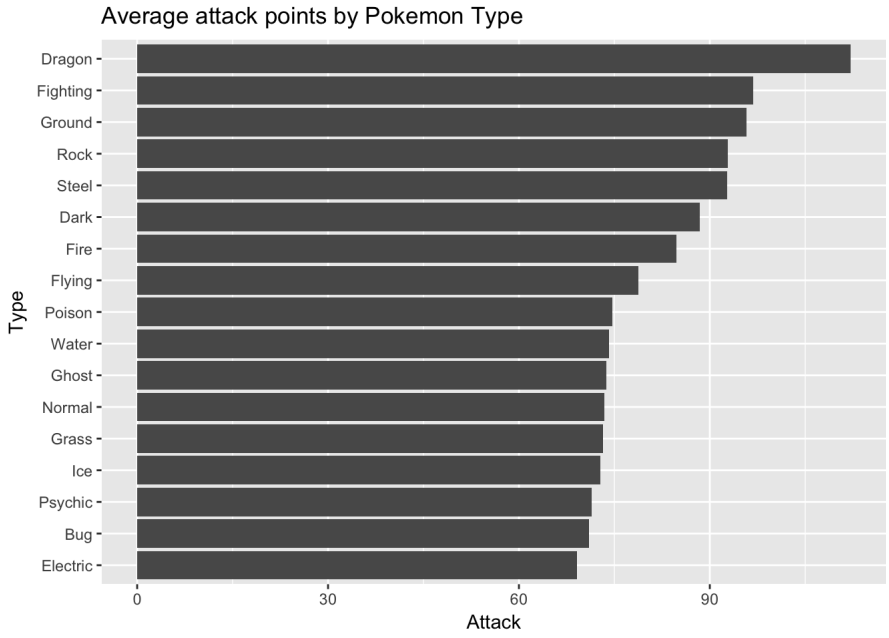
This post assumes that the reader has some basic knowledge of how work with R and the packages dplyr and ggplot2. For a quick review of how to create dotplots using ggplot2, see this tutorial. In addition, a dplyr tutorial can be found here.

Let's compare the average Attack Points based on the primary type of a Pokemon. We'll first need to do some data preparation.

```
atk_points = data %>%
  group_by(Type) %>%
  summarise(Attack = mean(Attack))
```

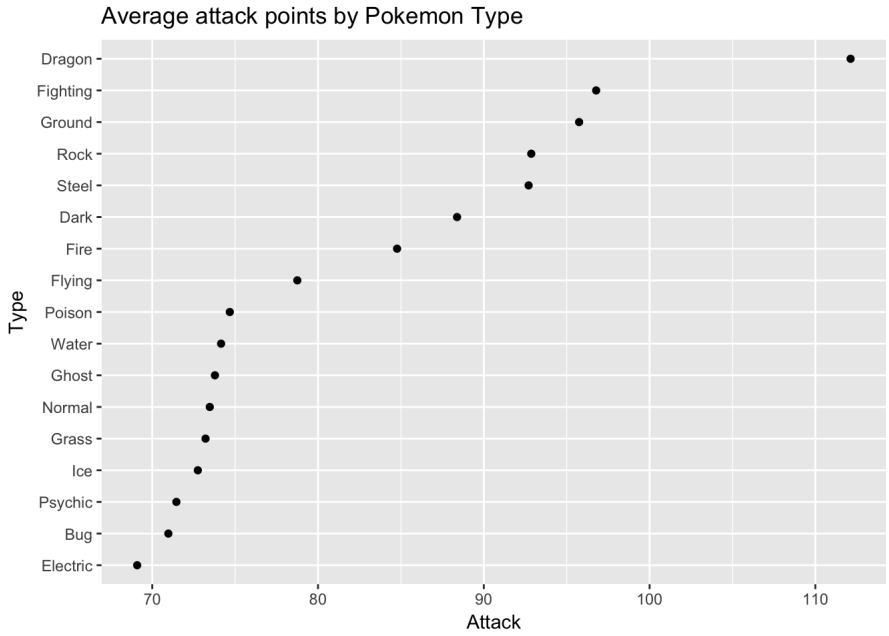
Let's visualize the data using a barchart and a dotplot. Here's the barchart.

```
ggplot(atk_points, aes(reorder(Type, Attack), Attack)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  ggtitle("Average attack points by Pokemon Type") +
  xlab("Type")
```



Here's the dotplot.

```
ggplot(atk_points, aes(reorder(Type, Attack), Attack)) +
  geom_point() +
  coord_flip() +
  ggtitle("Average attack points by Pokemon Type") +
  xlab("Type")
```

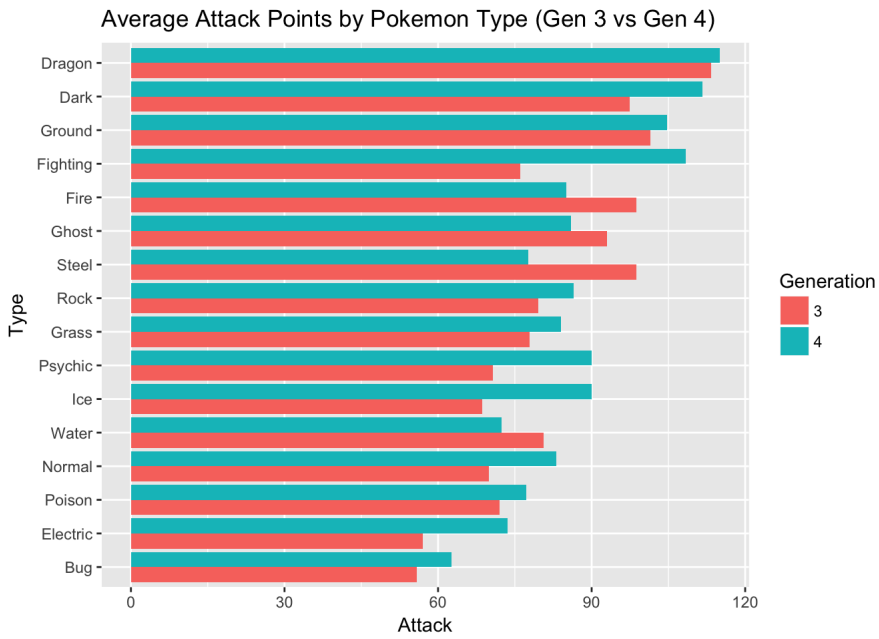


Now, say we wanted to compare the average Attack Points for Generation 3 Pokemon and Generation 4 Pokemon by their type. Let's first create a data frame to do this.

```
atk_gen3_gen4 = data %>%
  filter(Generation == "3" | Generation == "4") %>%
  group_by(Type, Generation) %>%
  summarise(Attack = mean(Attack))
```

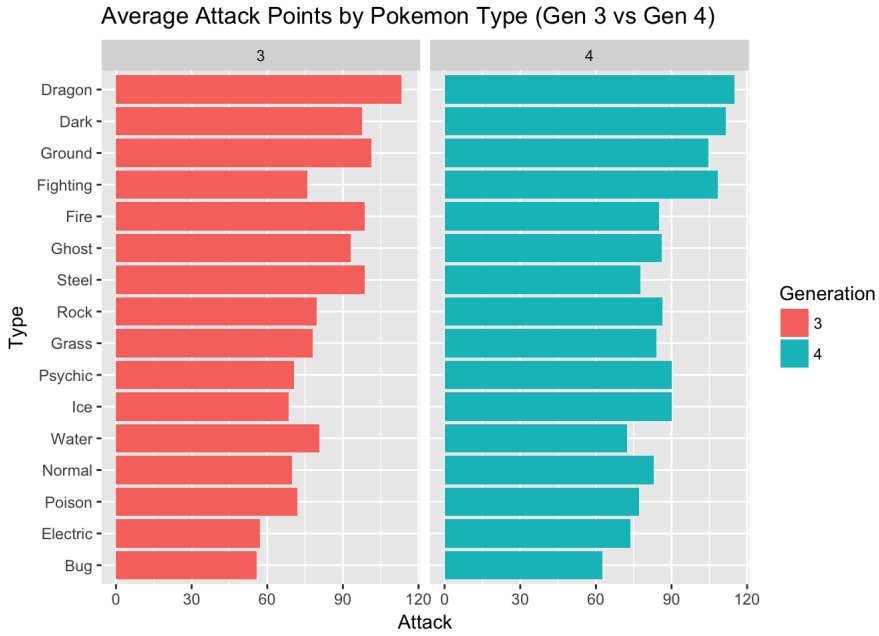
Now, we will try to visualize this using a barchart.

```
ggplot(atk_gen3_gen4, aes(reorder(Type, Attack), Attack, fill = Generation)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  ggtitle("Average Attack Points by Pokemon Type (Gen 3 vs Gen 4)") +
  xlab("Type")
```



Or, alternatively, we can make two separate barcharts.

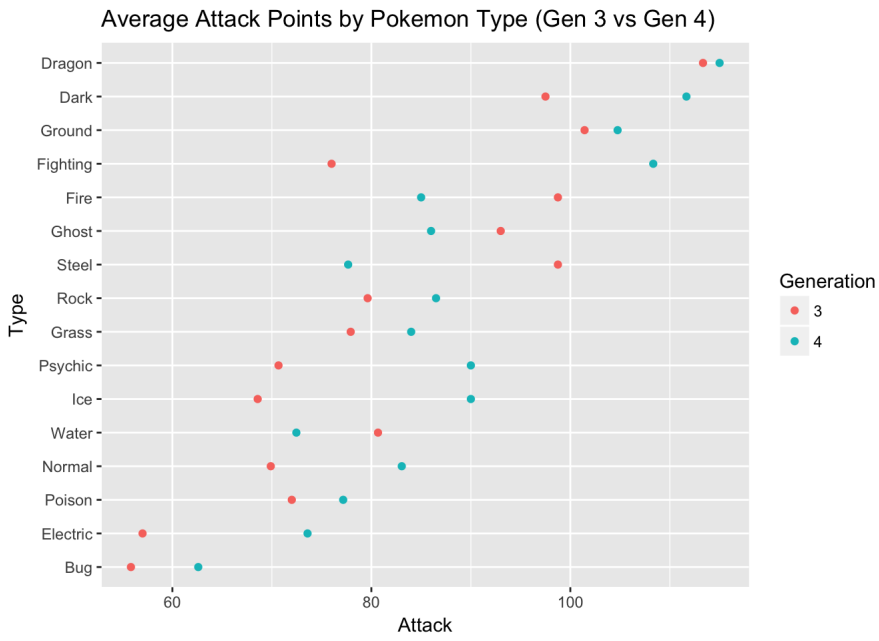
```
ggplot(atk_gen3_gen4, aes(reorder(Type, Attack), Attack, fill = Generation)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  facet_wrap(~ Generation) +
  ggtitle("Average Attack Points by Pokemon Type (Gen 3 vs Gen 4)") +
  xlab("Type")
```



While these bar charts are nice, it's hard to really understand and draw conclusions from them without taking a long time looking at them and comparing them.

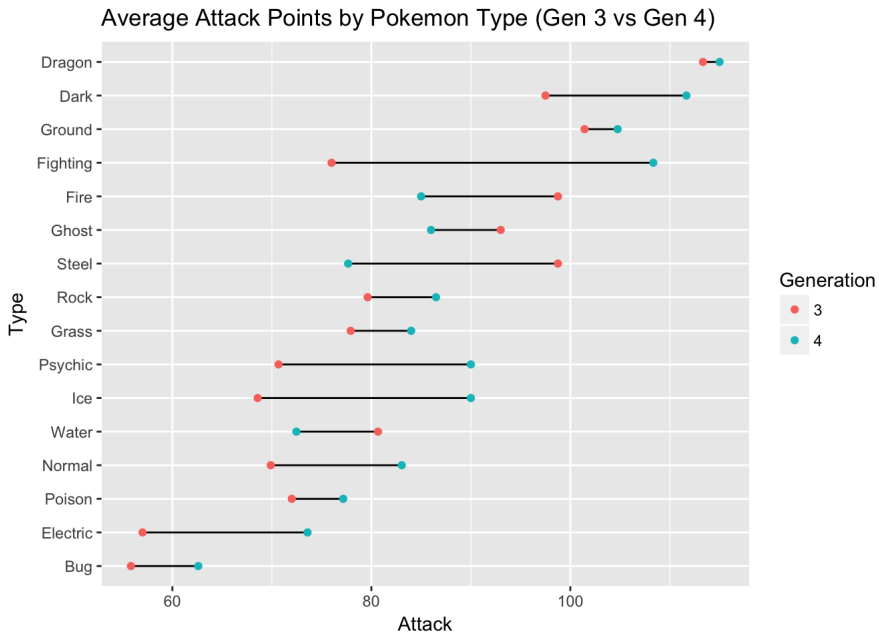
Let's try a dot plot!

```
ggplot(atk_gen3_gen4, aes(reorder(Type, Attack), Attack)) +
  geom_point(aes(color = Generation)) +
  coord_flip() +
  ggtitle("Average Attack Points by Pokemon Type (Gen 3 vs Gen 4)") +
  xlab("Type")
```



This is an improvement on the barcharts, but now let's make one final touch: connecting the dots for each type. This allows us to more concretely see the difference between average attack points in Generation 3 versus Genration 4 for each type. We can do this with the `geom_line` function, grouping by Type.

```
ggplot(atk_gen3_gen4, aes(reorder(Type, Attack), Attack)) +
  geom_line(aes(group = Type)) +
  geom_point(aes(color = Generation)) +
  coord_flip() +
  ggtitle("Average Attack Points by Pokemon Type (Gen 3 vs Gen 4)") +
  xlab("Type")
```



Now, we can very clearly see the differences. It is very clear that Dragon types are by far the strongest, but seem to very slight changes in attack points, whereas Fighting types got a lot stronger from Generation 3 to 4. We can also see that Steel Types got significantly weaker. The main point to take away here is that this dot plot with connected lines, called a **Cleveland Dot Plot**, is much easier to read and gives us more useful information more quickly than the bar charts or tradtional dot plot.

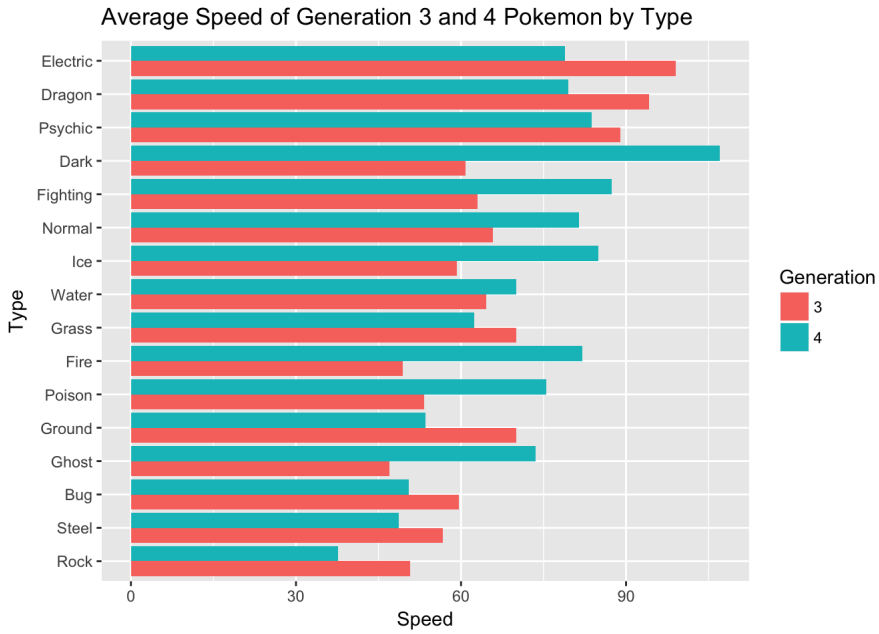
A Second Example Comparing Speed

Now, let's do a second example to really hammer the point home. Let's compare the average speed per type of the two same generations, Generation 3 and Generation 4.

```
speed_gen3_gen4 = data %>%
  filter(Generation == "3" | Generation == "4") %>%
  group_by(Type, Generation) %>%
  summarise(Speed = mean(Speed))
```

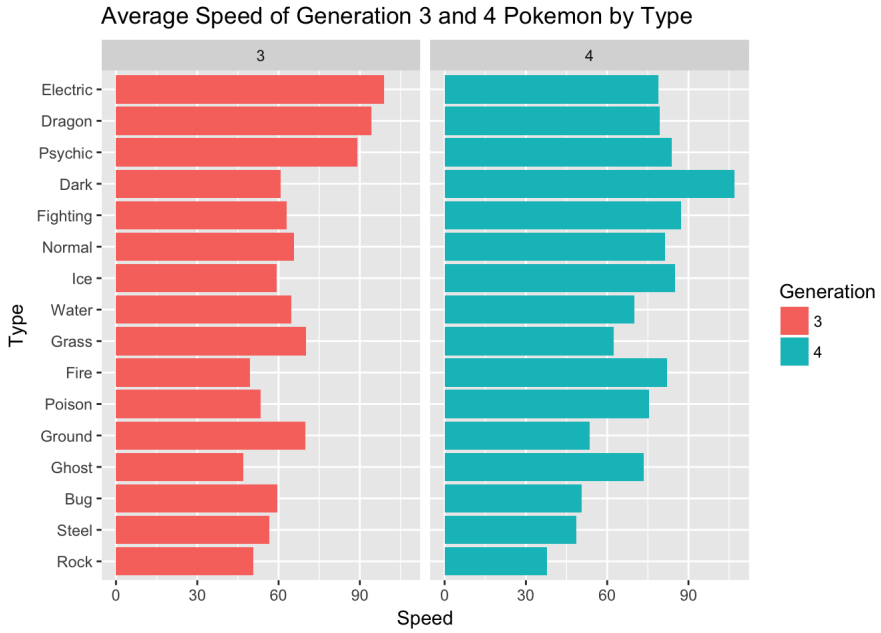
Again, let's compare bar charts with our Cleveland Dot Plot. First, the bar charts.

```
ggplot(speed_gen3_gen4, aes(reorder(Type, Speed), Speed, fill = Generation)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  ggtitle("Average Speed of Generation 3 and 4 Pokemon by Type") +
  xlab("Type")
```



Again, we have the possibility of making two separate barcharts side-by-side, but this is still difficult to read.

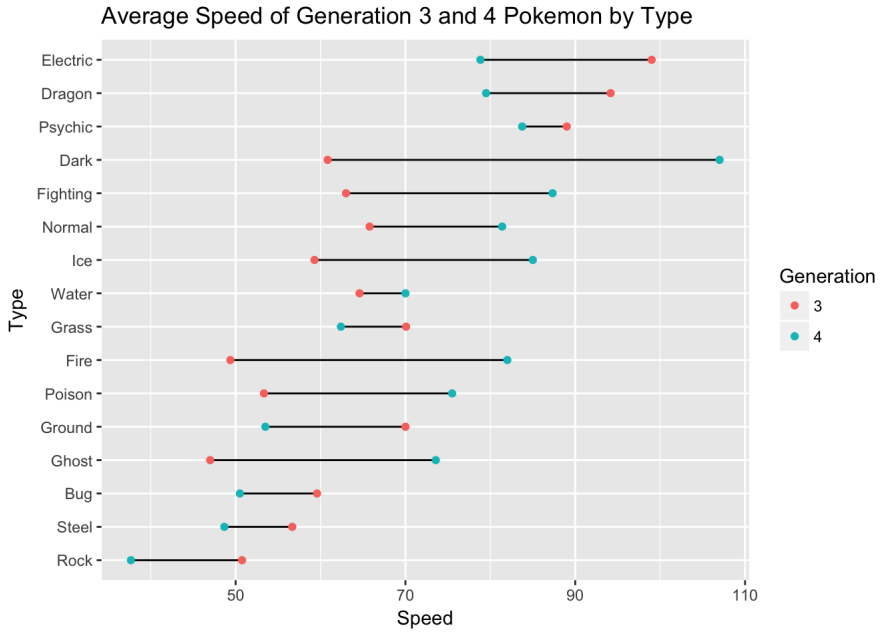
```
ggplot(speed_gen3_gen4, aes(reorder(Type, Speed), Speed, fill = Generation)) +
  geom_bar(stat = "identity", position = "dodge") +
  coord_flip() +
  facet_wrap(~Generation) +
  ggtitle("Average Speed of Generation 3 and 4 Pokemon by Type") +
  xlab("Type")
```



From the above two barcharts, try to deduce which types had an increase in speed, and which types had a decrease in speed, as well as how big those changes were. Time yourself!

Now, let's take a look at the Cleveland Dot Plot.

```
ggplot(speed_gen3_gen4, aes(reorder(Type, Speed), Speed)) +
  geom_line(aes(group = Type)) +
  geom_point(aes(color = Generation)) +
  coord_flip() +
  ggtitle("Average Speed of Generation 3 and 4 Pokemon by Type") +
  xlab("Type")
```



Now try to do the same exercise as previously mentioned. Deduce which types had an increase and which types had an decrease, as well as how big those changes were. Was it faster this time?

Again, we can see that the Cleveland Dot Plot is *much* easier to read. We can easily see the changes in average Speed for each type from Generation 3 to Generation 4. Dark type Pokemon got a massive buff in speed, while Electric types and Dragon types got slower. The types with a red dot on the left hand side had increases in speed, while the types with a blue dot on the left had decreases in speed. We can look at the length of the line to determine whether the change was significant.

It is clear this type of plot allows us to make these kinds of observations and conclusions much quicker than a double barchart would.

Take Home Message

- Cleveland Dot Plots are much easier to read and understand when comparing data based on categories.
- Barcharts are nice, but they can take a lot of time to analyze.
- We can create Cleveland Dot Plots pretty easily using the R package ggplot2. It takes very little extra effort to make a much more readable and understandable graphic.
- It is important to really think about what results we are trying to portray when reporting data. Certain graphs or charts may be much better visual aides for specific conclusions.
- Using a Cleveland Dot Plot is just one example of where choosing a certain plot can make a big difference. In general, taking the extra time to consider all the different types of graphs (and even creating them and analyzing them) is well worth it to make sure the big idea gets across.
- We can make more improvements! Take a look at some of these resources to learn about some more graphs and when to use them! Cool graphs to learn how to make: Box and Scatter Plot, Stacked Area Chart, Hourly Heat Map.

Sources:

- https://greatbrook.com/wp-content/uploads/2015/05/data-analysis-charts.png
- http://uc-r.github.io/cleveland-dot-plots
- https://www.kaggle.com/abcsds/pokemon
- http://www.joyce-robbins.com/blog/2016/06/02/datavis-with-rdrawing-a-cleveland-dot-plot-with-ggplot2/
- http://genomicsclass.github.io/book/pages/dplyr_tutorial.html
- http://www.win-vector.com/blog/2013/02/revisiting-clevelands-the-elements-of-graphing-data-in-ggplot2/
- http://www.wekaleamstudios.co.uk/wp-content/uploads/2010/03/graphs-dotplot.pdf
- http://www.r-graph-gallery.com/273-custom-your-scatterplot-ggplot2/
- https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf