Learning more about the uses and importance of ggplot2

Introduction

We have had a brief introduction into ggplot2 by creating some basic graphs like scatterplots, histograms, and barcharts. But the importance of this package cannot be overstated in the modern statistics community. Many things we learn in our academic curriculum are often meant to teach us conceptual fundamentals and theory behind certain applications, but in this class we actually learn many hands-on, practical skills in the very relevant statistical programming language R.

This post will revolve around showing how specifically the ggplot2 package is a prime example of R's highly functional and easy-to-use nature. I was motivated to do my post on this topic because I myself was exposed to just how important R is in the corporate world in helping to work through actual data companies look at every day. In two of my internship experiences I have seen ggplot2 been referenced and used to show different visualizations in decks that are actually presented in meetings. When I saw just how easily readable and digestable these graphs were in displaying complex information on real company metrics and data, I was intrigued to learn more.

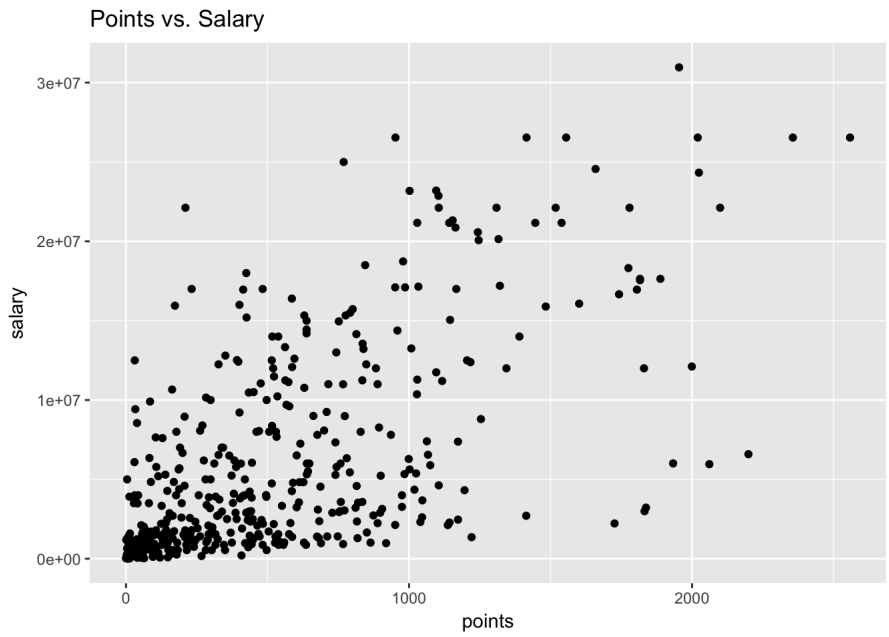
This class has been a great initial exposure to ggplot2 and how the basics of it work. In my post I hope to review some of the things we have already learned, but also introduce some new aspects or features of ggplot2 that we have not learned about, and finally I want to make sure to highlight how important ggplot2 is in the statistics world.

Part 1: Reviewing classwork

Let's begin by reviewing some of the stuff we have already learned.

Our most basic form of graph was the scatterplot. Using our regular NBA dataset we get:

```
# Generate scatterplot of points vs salary
ggplot(data = dat, aes(x = points, y = salary)) +
  geom_point() +
  ggtitle("Points vs. Salary")
```

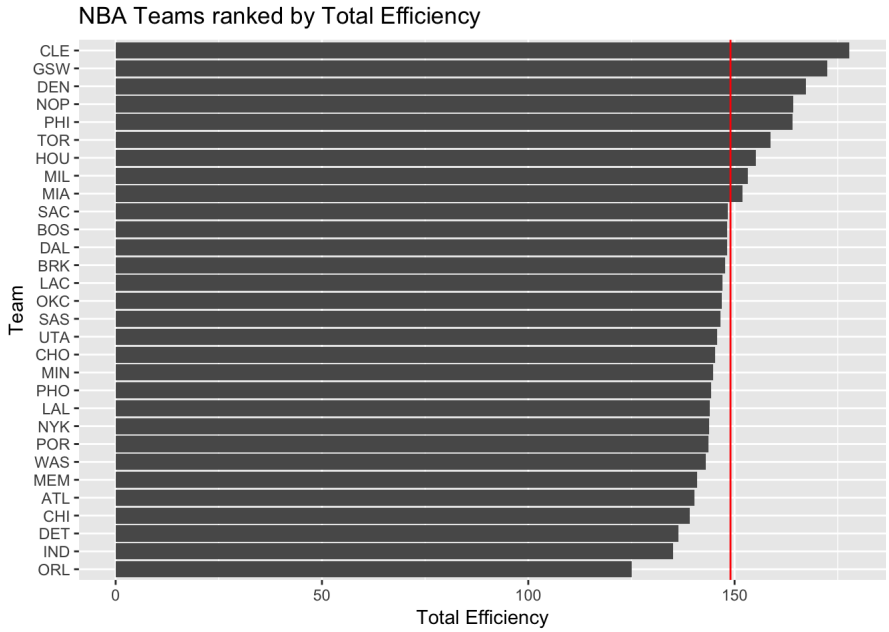


This is a very simple ggplot use but it is important to actually understand what is going on. The first argument, "data", is the whole data set that contains the variables you want to graph. The second parameter, "aes", stands for aesthetics and takes in two additional parameters. The first parameter is the variable we want on the x-axis, and the second parameter is the label we want on the y axis. Lastly we add "geom_point()" to tell ggplot2 what it is we actually want on our graph. In this case we want a plot of points. These are the fundamental building blocks in our ggplot call. We have only used a few in this example, but there are many more like scales, coordinate systems, position adjustments, and even faceting.

```
# Reference: http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html
```

By understanding this concept of building blocks, we have actually enabled ourselves to skip into something more complex.

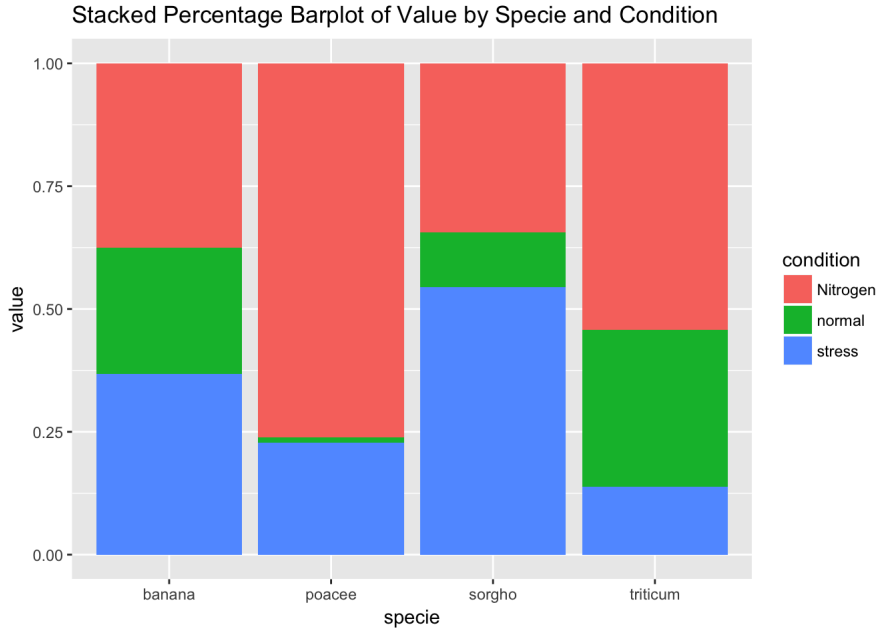
```
# Sorted bar chart of NBA teams ranked by total efficiency
ggplot(teams, aes(reorder(team, efficiency), efficiency)) +
  geom_bar(stat = 'identity') +
  coord_flip() +
  geom_hline(yintercept = mean(teams$efficiency),color = "red") +
  labs(y = "Total Efficiency") +
  labs(x = "Team") +
  labs(title = "NBA Teams ranked by Total Efficiency")
```



This is evidently more complicated, but still fairly easy to breakdown if we think of each addition as another building block. We follow the same setup, but we order the y-axis, flip the coordinates, add a line, and add some labels. Clearly going from something very simple to something more complex in ggplot2 is not very difficult as it simply requires understanding the right blocks to add to the same basic code. In this way, we can easily play around with different features of the graph by just adding and removing as we go. I think this serves as a good review; now we can move on to looking at some newer applications of ggplot2 we haven't used before.

Part 2: Some new material

```
# Create a stacked percentage barplot from new data set
ggplot(dat2, aes(fill=condition, y=value, x=specie)) +
    geom_bar( stat="identity", position="fill") +
    ggtitle("Stacked Percentage Barplot of Value by Specie and Condition")
```




```
# Reference: http://www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2/
```

In this graph, we have not only made a barchart that is stacked, but we have calculated each individual bar's percentage of the total bar. This is very useful because we can see directly how all the bars relate to one another instead of just roughly comparing heights. Let's examine the new building blocks we used here. We used a new parameter called fill that would tell us how to fill each bar (by condition). Secondly we used a position parameter that tells us to scale the bars heights - effectively creating the percentages instead of just raw values.

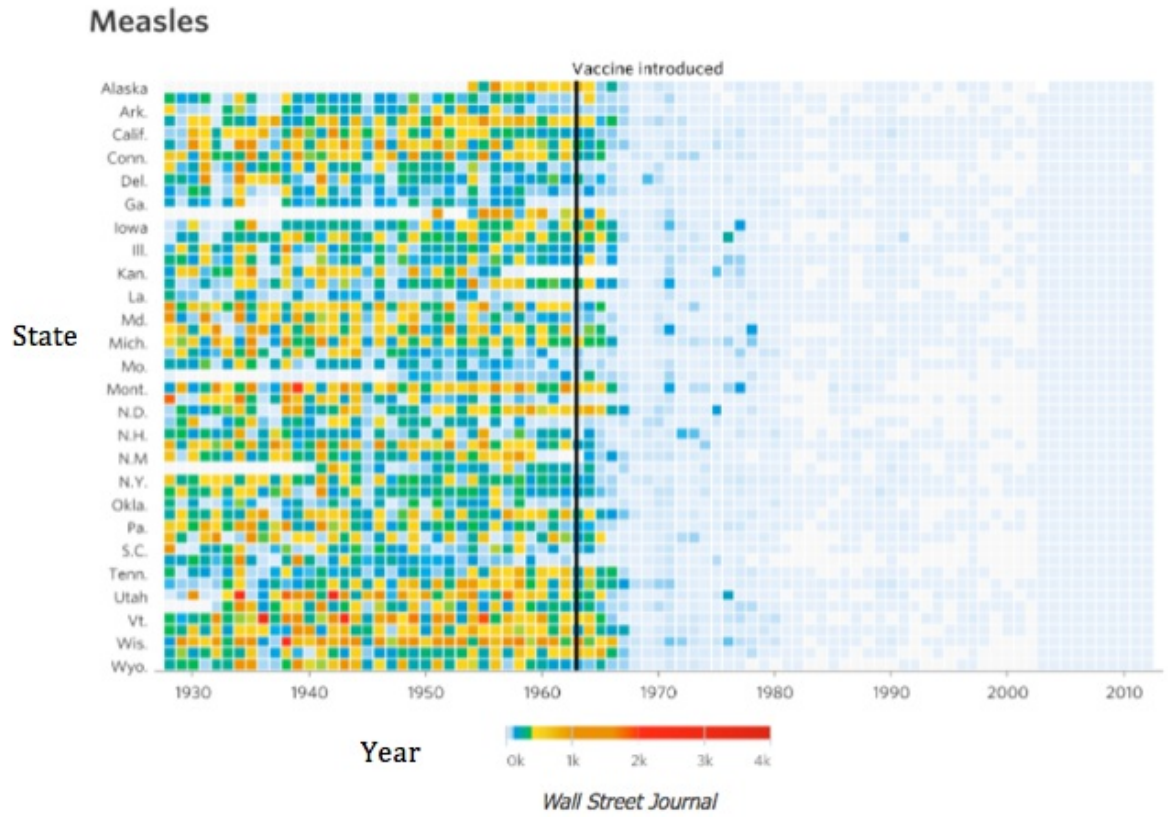
Another cool thing we can do with our ggplot graphs is add themes. Themes allow us to better create the setting and surroundings of a graph so that it makes more sense. By adding this necessary context and design, themes can make graphs much more visually appealing as well.

```
# Reference: https://cran.r-project.org/web/packages/ggthemes/vignettes/ggthemes.html
```

Some examples of themes we can use are theme_base, theme_calc, and theme_excel. My favorite kinds of themes are those that we see in publications like theme_wsj for The Wall Street Journal, and even theme_economist for The Economist. This is an example of a graphic that was created using ggplot2 that was actually used in a Wall Street Journal article. It is describing the number of people with measles in different states over time, but pay more attention to the slightly different fonts and colors being used in the theme. Ignore the axes titles' placement and font

because I had to add those in.

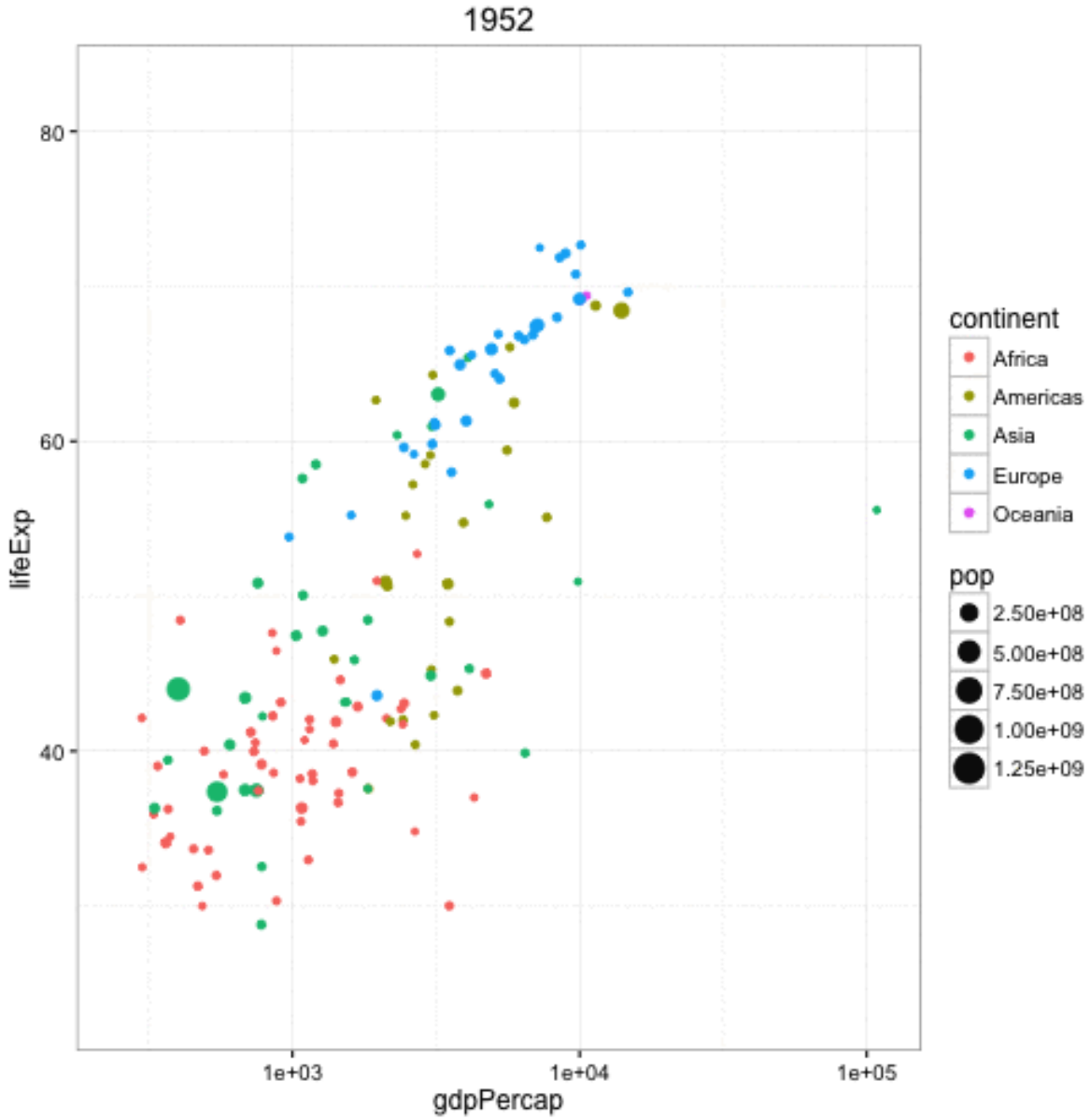
Measles Count by State and Year




```
# Reference: http://graphics.wsj.com/infectious-diseases-and-vaccines/
```

This graph tells a very clear story, and it is in large part due to the great design and theme it is using. While adding themes is a useful thing to do, the last new material I want to introduce is more on the functional side. Something that I have found very interesting in ggplot2 is the animation. Using animations in graphs can be useful for a number of reasons. Most commonly, they are used to show change over time. Static graphs can only be so useful, and that is why an introduction to using animation in ggplot2 is necessary. Let's take a look at an animation in ggplot2.

GDP Per Capita vs Life Expectancy from 1952 - 2007



Because of animations we are able to visualize how the correlation between GDP and life expectancy in different continents has changed over time. Had this graph just been a snapshot of one year (static), it would not be nearly as interesting or informative.

Let's take a look at the code used to build it.

```
# Basic scatter plot
mapping <- aes(x = gdpPercap, y = lifeExp,
               size = pop, color = continent,
               frame = year)
p <- ggplot(gapminder, mapping = mapping) +
  geom_point() +
  scale_x_log10()
# Animate
gganimate(p)
```

```
# Reference: http://www.sthda.com/english/articles/16-r-packages/58-gganimate-create-animations-with-ggplot2/
```

What we see are mostly our regular building blocks like geom_point or scale. But one very important addition to make the animation work is what we actually want to see change. This is what the parameter "frame" is for. In this case, we want to see the year change. That is why we set "frame" equal to the variable "year". The last step is simply to run our plot through the function gganimate. Now that we have been introduced to some new aspects of ggplot2, let's talk about its importance.

Part 3: Importance of ggplot2

Since we are in a statistics class after all, one of the best testaments to the importance of this package is its number of downloads. ggplot2 is the 3rd most downloaded package out of any in R.

Most downloaded packages

Name	Direct downloads	Indirect downloads	Total
1. viridisLite	151,069	9,184	160,253
2. R6	76,754	130,090	206,844
3. ggplot2	67,371	148,045	215,416
4. readr	67,337	51,023	118,360
5. dplyr	62,270	124,468	186,738

```
# Reference: https://www.rdocumentation.org/trends
```

Its creator, Hadley Wickham, notes that ggplot2 was so successful because previous visualization packages were not appealing from a theoretical point of view. Wickham was able to fix this by creating a “grammar of graphics”. He also notes that his effort in making the defaults aesthetically pleasing was a key part of ggplot2’s success. In our earlier parts, we have seen firsthand how much easier both these things made our plotting.

```
# Reference: https://qz.com/1007328/all-hail-ggplot2-the-code-powering-all-those-excellent-charts-is-10-years-old/
```

Lastly, I wanted to touch on ggplot2’s use going forward. Statistics is a constantly evolving field, but ggplot2 will remain relevant. Machine learning often uses ggplot2 in clustering algorithms as well as for other predictive analytical tools. Data science is one of the newest, hottest professions and understanding how to use ggplot2 is still something that data scientists recommend as a skill in your toolbelt.

```
# Reference: http://data-informed.com/the-skills-you-need-to-become-a-data-scientist/
```

Conclusion

The message of my post is clear - learning about how to use ggplot2 in depth is an important skill to have for any statistician. I hope that I was able to break down more precisely what we did in class with ggplot2, show you some new things that can be done in the package, and finally convince you of its importance. Evidently there is a lot more to learn and discuss about ggplot2, but this should serve as an inspiration to go out and learn more about what visualiations actually interest you.

References

http://tutorials.iq.harvard.edu/R/Rgraphics/Rgraphics.html

http://www.r-graph-gallery.com/48-grouped-barplot-with-ggplot2/

https://cran.r-project.org/web/packages/ggthemes/vignettes/ggthemes.html http://graphics.wsj.com/infectious-diseases-and-vaccines/

http://www.sthda.com/english/articles/16-r-packages/58-gganimate-create-animations-with-ggplot2/

https://www.rdocumentation.org/trends

https://qz.com/1007328/all-hail-ggplot2-the-code-powering-all-those-excellent-charts-is-10-years-old/

http://data-informed.com/the-skills-you-need-to-become-a-data-scientist/