Data Visualization — Bubble Plots in R with ggplot2

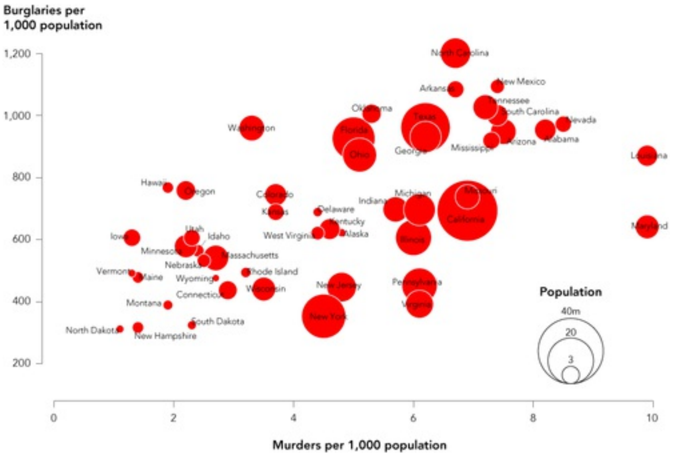
Introduction

Have you ever wanted to plot some data in R, but realize that there are three variables to account for? You find yourself saying these statements to yourself:

“How am I gonna choose which two variables to use?” “Maybe I should just use colors for the third variable… But my third variable has too many options for a color to be feasible…”

You try turning to the default `plot` function that R provides, but find that it’s limited in its usage, and it doesn’t really look that good. What do you do?

Well there’s a tool out there for your problems! The `ggplot2` package in R has a type of plot available that’s great for this type of situation. It’s called the **bubble plot**, and it’s one of the better ways to plot out data with three dimensions.



Motivation

My motivation towards this topic was to learn more about ggplot, as I think it’s a fascinating tool for plotting out data. Coming from a background of Python and Pandas for data science, I’ve been curious about `ggplot2` ever since I learned about it, and wanted to do more research on the tool itself.

To do that, I looked more closely at the different types of plots that can be done with ggplot, and found the bubble plot to be interesting both in its uses and in how aestheically pleasing it is. It is with that motivation that I continue to learn more about the topic and present that information in this post.

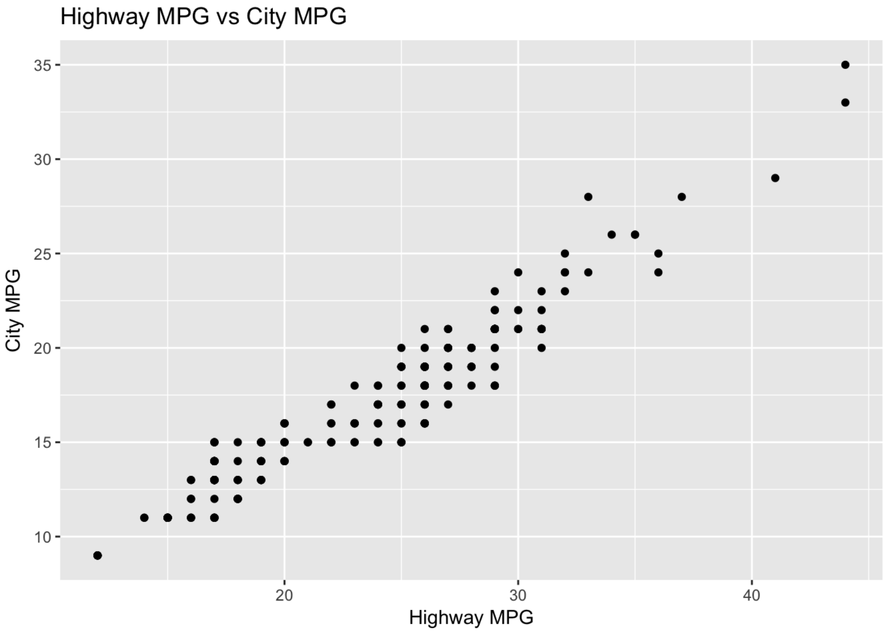
Background – An Introduction to `ggplot2`

`ggplot2`, for those of you who don’t know, is “a plotting system for R based on the grammar of graphics” according to the ggplot2 website (ggplot2.org). It’s very easy to use, as shown below with this simple scatterplot:

```r
library(ggplot2)
```

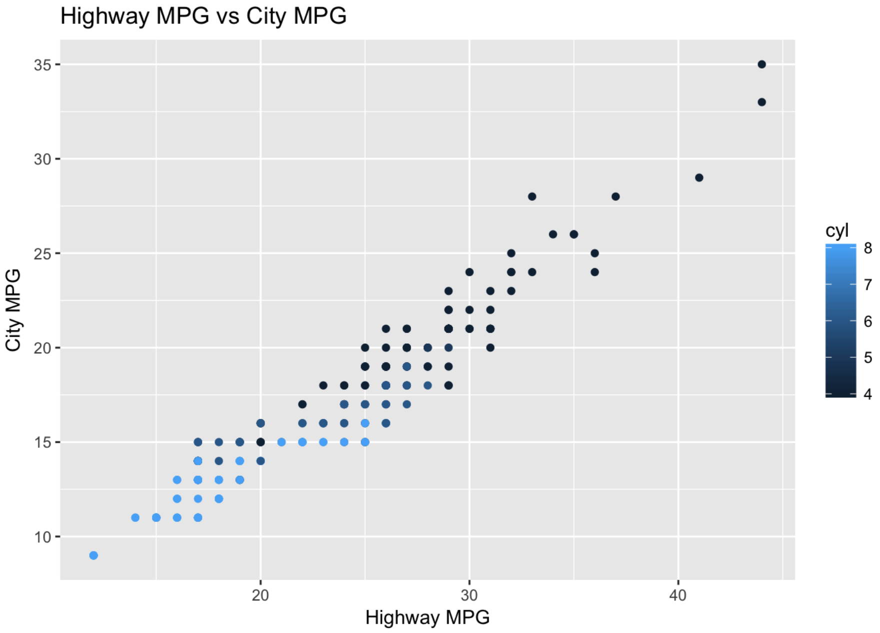
```
## Warning: package 'ggplot2' was built under R version 3.3.2
```

```r
# Creating a simple scatterplot
ggplot(mpg, aes(hwy, cty)) +
  geom_point() +
  labs(title = "Highway MPG vs City MPG",
       x = "Highway MPG", y = "City MPG")
```



You may be thinking: why can't I just use the default `plot` function that R has? Well, `ggplot2` has many simple functions to create some really cool graphs, particularly the above scatterplot with color added:

```
# A more colorful scatterplot
ggplot(mpg, aes(hwy, cty)) +
  geom_point(aes(color = cyl)) +
  labs(title = "Highway MPG vs City MPG",
       x = "Highway MPG", y = "City MPG")
```



For this plot, the colors represent the number of cylinders the engine has. It's pretty simple to add to the function: I only put an `aes(color = cyl)` inside the `geom_point()` function, which is basically saying to add a color scale according to the variable `cyl`.

I hope you agree that it looks pretty good aesthetically, and it was pretty simple to do. There are a whole bunch of other types of graphs that can be created (line, bar, histogram, etc.), but for the sake of clarity I won't go too deep into those. Just know that `ggplot2` is great for plotting all sorts data in R, both qualitative and quantitative.

Discussion

What is a bubble plot?

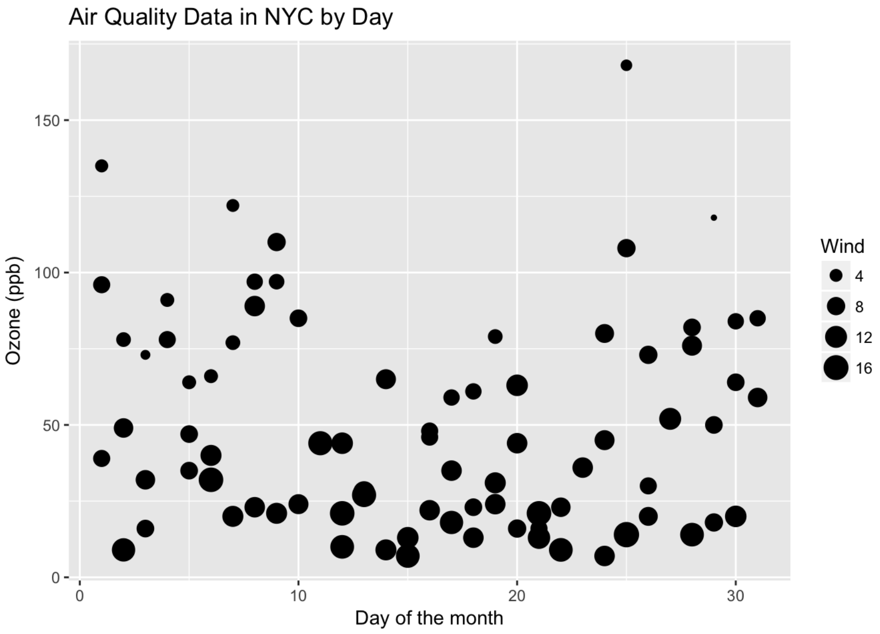
So with that little introduction to `ggplot2`, we now go into the bubble plot. The bubble plot is a variant of the scatterplot – in other words, it is created similarly to the scatterplot by calling `geom_point()` on the `ggplot()` function.

As stated above, a bubble plot is best used when there are three variables in the data. Like the scatterplot, a bubble plot is best used for quantitative (numerical) data, although qualitative data can be used to some degree (which will be shown below).

Here's a simple example of a bubble plot:

```
# Credit for data trimming goes to:
# http://t-redactyl.io/blog/2016/02/creating-plots-in-r-using-ggplot2-part-6-weighted-scatterplots.html
data(airquality)
aq_trim <- airquality[which(airquality$Month == 7 |
                            airquality$Month == 8 |
                            airquality$Month == 9), ]
aq_trim$Month <- factor(aq_trim$Month,
                        labels = c("July", "August", "September"))

# A simple bubble plot
ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(aes(size = Wind)) +
  labs(title = "Air Quality Data in NYC by Day",
       x = "Day of the month", y = "Ozone (ppb)")
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```

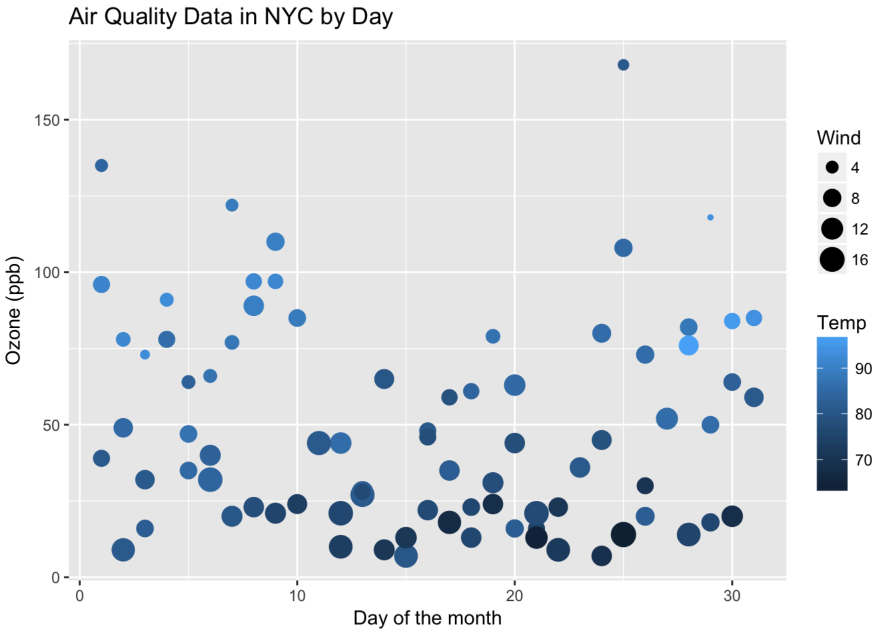


Cool things you can do with the bubble plot

One nice thing about the bubble plot is that we can plot data of more than two variables. As mentioned above, three variables are usually the norm for this type of plot. However, we can add color as a fourth variable:

```
# The same plot, but with color
ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(aes(size = Wind, col = Temp)) +
  labs(title = "Air Quality Data in NYC by Day",
       x = "Day of the month", y = "Ozone (ppb)")
```

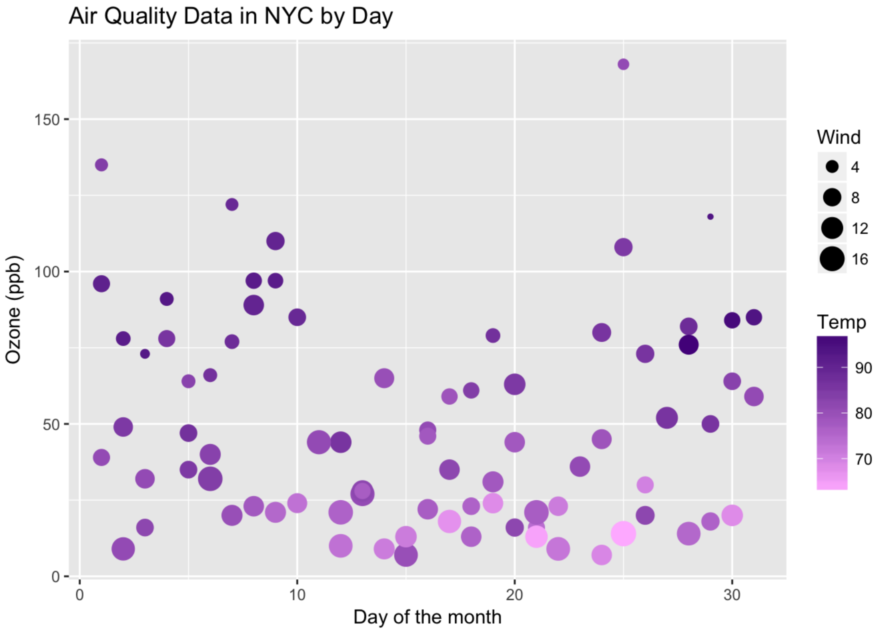
```
## Warning: Removed 11 rows containing missing values (geom_point).
```



By adding `col` to the `geom_point()` variable, we can add another variable to the mix. You can also change the color of the plot to match your tastes:

```
# Plot with a different color
ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(aes(size = Wind, col = Temp)) +
  labs(title = "Air Quality Data in NYC by Day",
       x = "Day of the month", y = "Ozone (ppb)") +
  scale_color_continuous(low = "plum1", high = "purple4")
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```

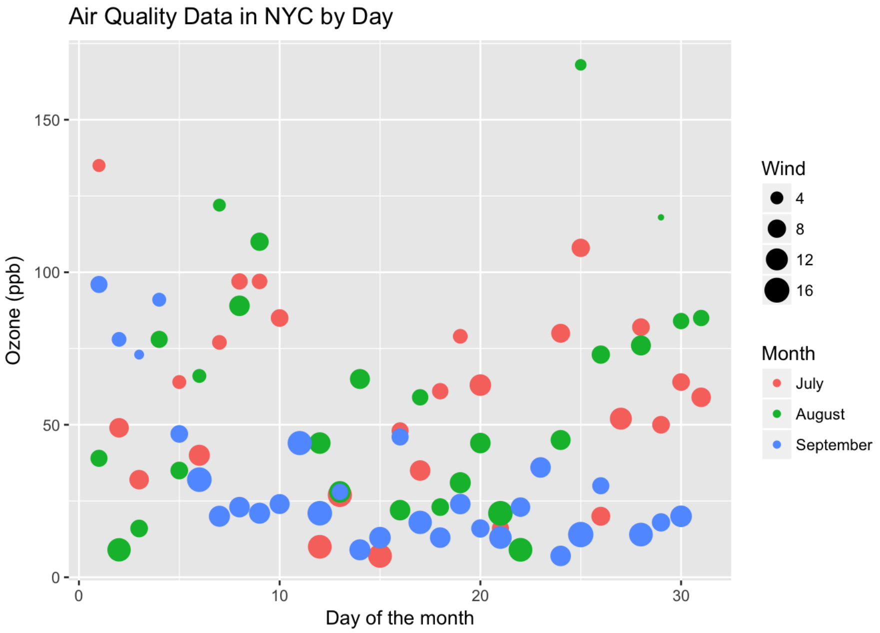


To do this, add `scale_color_continuous()` to the `ggplot()` function as above.

Colors aren't limited to quantitative variables, however. This time qualitative variables can join in on the fun, like in the plot below with the 'Month' variable:

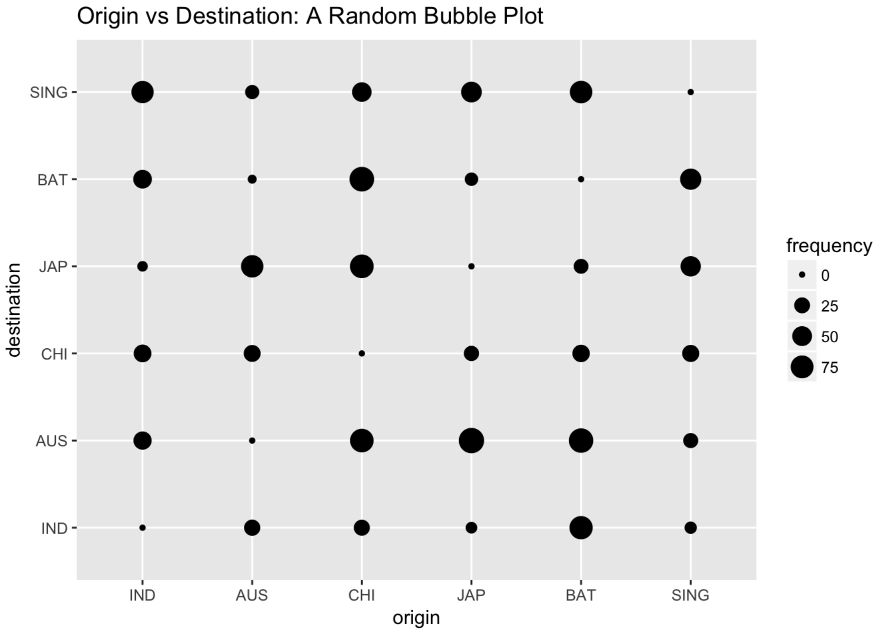
```
# Plot with color as qualitative variable
ggplot(aq_trim, aes(x = Day, y = Ozone)) +
  geom_point(aes(size = Wind, col = Month)) +
  labs(title = "Air Quality Data in NYC by Day",
       x = "Day of the month", y = "Ozone (ppb)")
```

```
## Warning: Removed 11 rows containing missing values (geom_point).
```



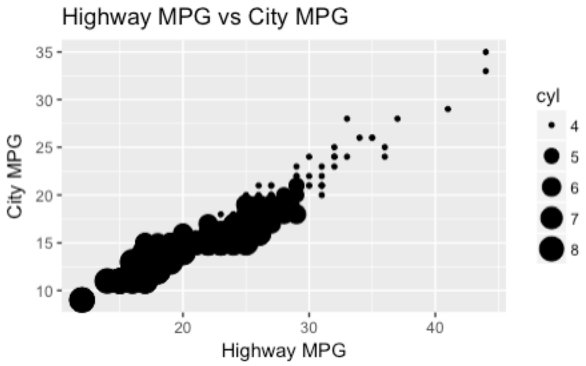
As stated above, quantitative data is usually the best type of data for the bubble plot. However, there are cases where qualitative data can be used to create a bubble plot. Take, for instance, this example from this Stack Overflow post here:

```
# Generating random data
countries = c('IND', 'AUS', 'CHI', 'JAP', 'BAT', 'SING')
frequencies = matrix(sample(1:100, 36), 6, 6, dimnames = list(countries, countries))
diag(frequencies) = 0

# Casting the matrix data to a suitable format for plotting
library(reshape2)
frequencies_df = melt(frequencies)
names(frequencies_df) = c('origin', 'destination', 'frequency')

# Using bubble plot to plot data
ggplot(frequencies_df, aes(x = origin, y = destination, size = frequency)) +
  geom_point() +
  labs(title = "Origin vs Destination: A Random Bubble Plot")
```



Limitations of the Bubble Plot

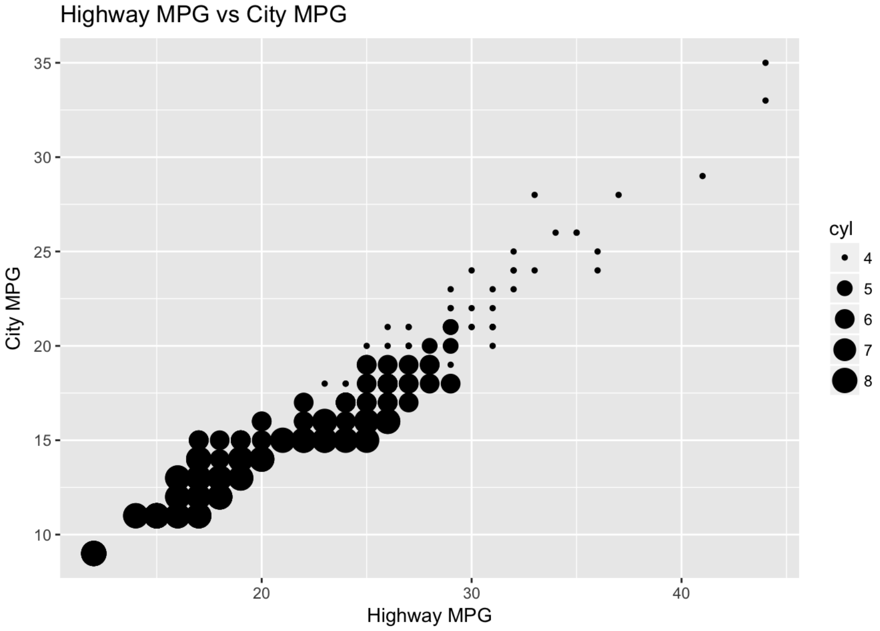
The bubble plot is nice visually, but it has its downsides. Take, for example, the plot below:



As you can see, if you're not careful there can be overlap between points, making it hard to find any sort of meaningful patterns in the plot. There are a few ways to fix this, however.

One solution is to just increase the length and width of the plot. By doing so, the plot is more spaced out and points aren't as bunched up, allowing you to be able to see patterns within the data. Here's the same plot, but spaced out more:

```
# A bubble plot with... something wrong
ggplot(mpg, aes(hwy, cty)) +
  geom_point(aes(size = cyl)) +
  labs(title = "Highway MPG vs City MPG",
       x = "Highway MPG", y = "City MPG")
```

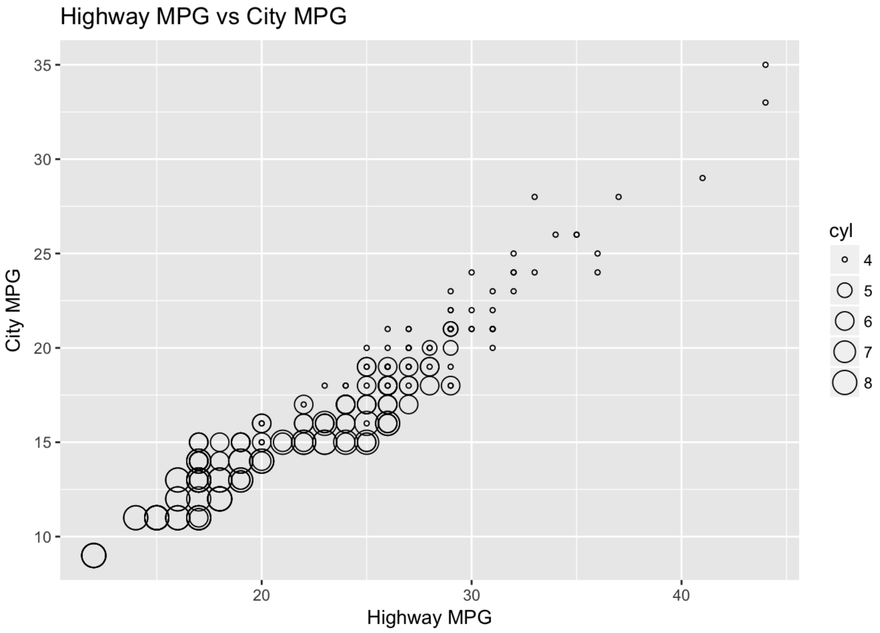


You may be thinking: "problem solved, right? Looks good to me."

Well what if I told you there were points completely overlapping each other? That's pretty bad: it leads to a misleading plot, and patterns derived from this plot wouldn't truly reflect the data.

There are a few ways to fix this, however. One is to change the shape of the points, as shown below:

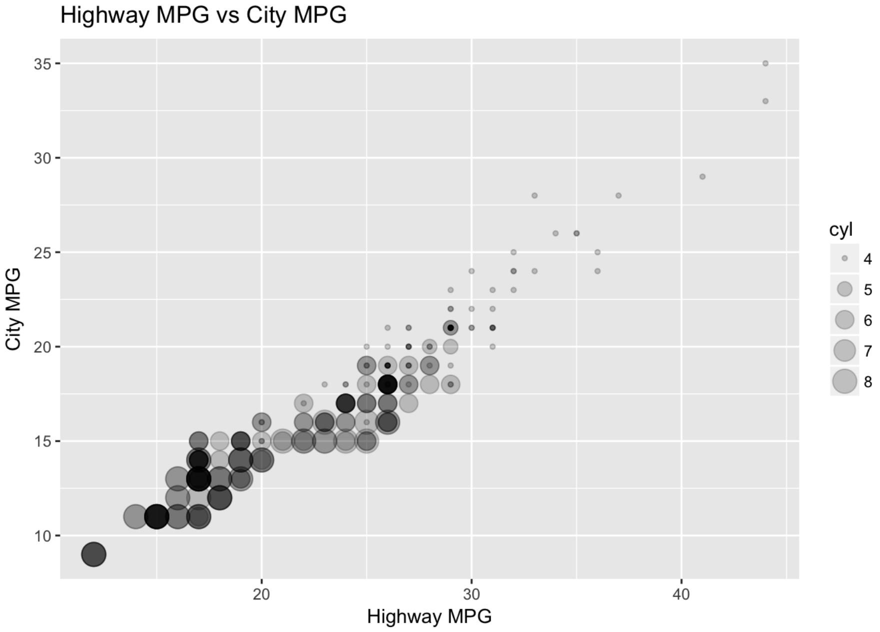
```
# A bubble plot with rings instead of points
ggplot(mpg, aes(hwy, cty)) +
  geom_point(aes(size = cyl), shape = 21) +
  labs(title = "Highway MPG vs City MPG",
       x = "Highway MPG", y = "City MPG")
```



By setting the `shape` variable in `geom_point()`, we can see the points that are completely overlapped, although it is still somewhat hard to see any patterns

Another method is to adjust the transparency of the points so that the overlapped points can be seen:

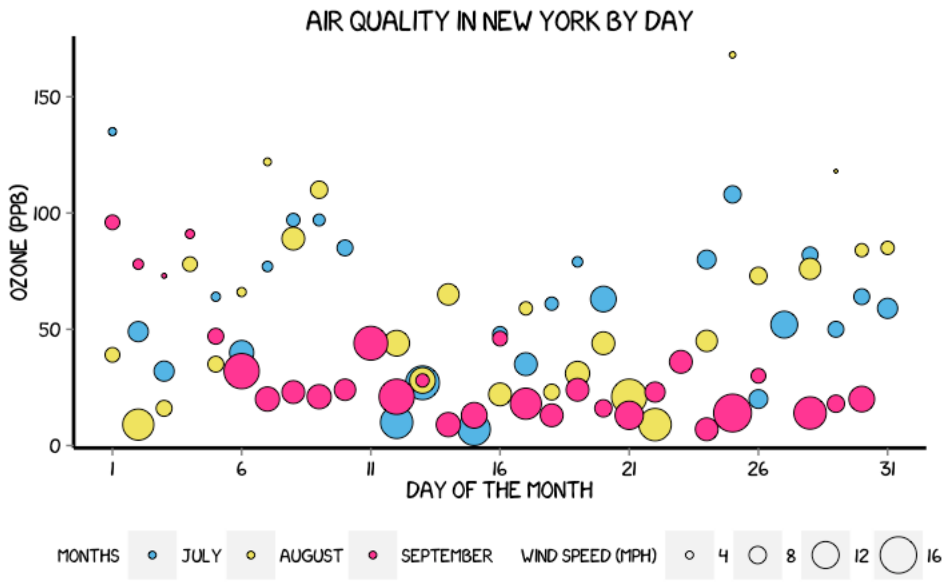
```
# A bubble plot with transparency added
ggplot(mpg, aes(hwy, cty)) +
  geom_point(aes(size = cyl), alpha = 0.2) +
  labs(title = "Highway MPG vs City MPG",
       x = "Highway MPG", y = "City MPG")
```



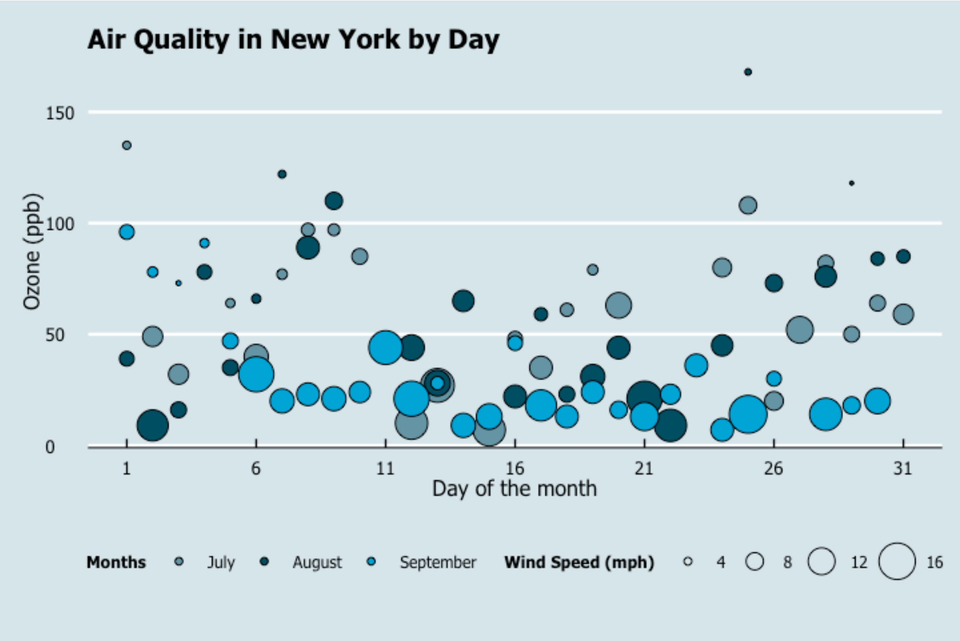
By changing the `alpha` value inside `geom_point()`, we can adjust the transparency of the points, allowing us to see just how many points are overlapping in a particular area.

An aside: Themes in `ggplot2`

One thing I found in my research that I thought was pretty dang awesome was that you can use themes in your plots. For example, you can make a plot in the style of XKCD (thanks to Mauricio Vargas Sepúlveda from http://t-redactyl.io/blog/2016/02/creating-plots-in-r-using-ggplot2-part-6-weighted-scatterplots.html for creating the theme). Unfortunately, I couldn't get the code to work correctly, so here's an image example of a plot in the theme:



And how about a graph in the style of The Economist (thanks again to the same person as above):



You can even create your own theme, as long as you have the patience and creative eye. Just one of the many benefits of using the `ggplot2` library.

Conclusions

So in conclusion, we saw the best ways to use the bubble plot as well as different ways to manipulate it to create new interesting twists such as color scheme, shapes, and more.

If you have more than two variables in your data and want to create a nice-looking plot for them, why not use a bubble plot?

References

- http://t-redactyl.io/blog/2016/02/creating-plots-in-r-using-ggplot2-part-6-weighted-scatterplots.html
- http://blog.revolutionanalytics.com/2010/11/how-to-make-beautiful-bubble-charts-with-r.html
- http://ggplot2.org
- http://sharpsightlabs.com/blog/bubble-chart-in-r-basic/
- http://r-statistics.co/Top50-Ggplot2-Visualizations-MasterList-R-Code.html
- http://dni-institute.in/blogs/bubble-chart-using-ggplot/
- https://stackoverflow.com/questions/26292484/programming-in-r-bubble-chart-visualization
- https://stackoverflow.com/questions/21313905/how-to-set-ggplot-alpha-transparency-value-for-all-points-at-once
- http://ggplot2.tidyverse.org/reference/scale_gradient.html