Ihor Kovalyshyn

26 Feb 2017 • on R visualization scatterplot Big Data ggplot2

When Scatter Plot doesn't work

In this blog post I'm going to demonstrate a case when you shouldn't use scatter plots for visualizing a relationship between two variables.

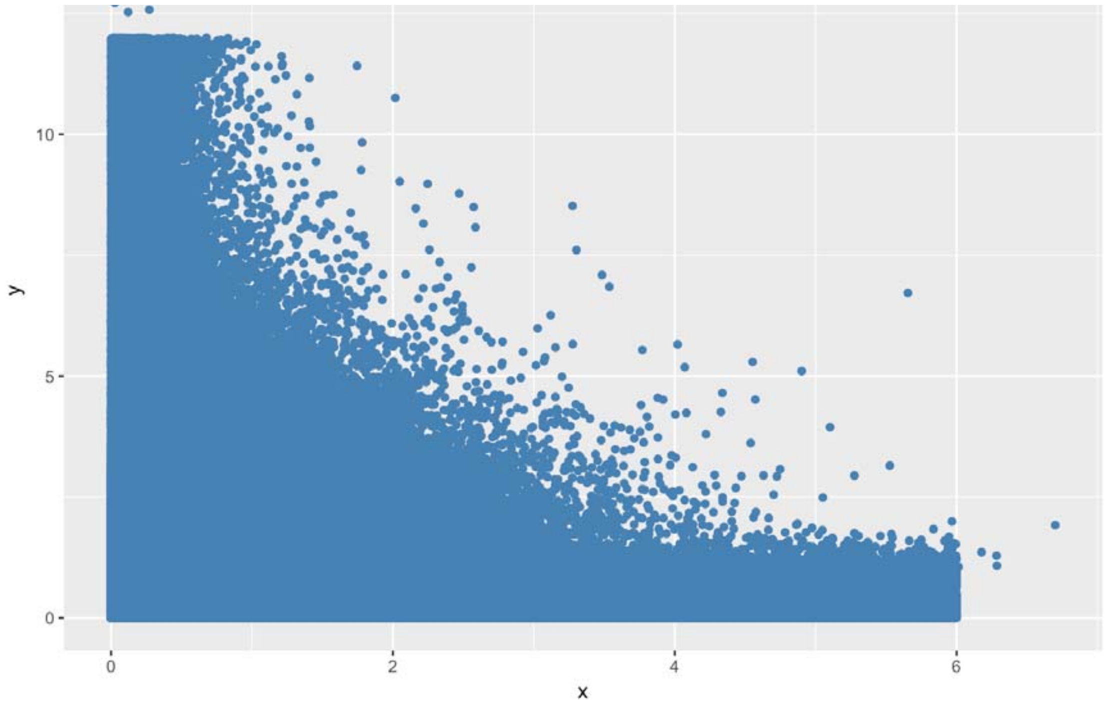
While working with data set with more than 4.5 million samples, I encountered a problem with visualizing a relationship between two variables. I noticed that scatter plot doesn't show correct patterns because of enormous overplotting. And that can be very misleading even for experienced people. Besides that, ggplot2 takes crazy amount of time to plot that number of data samples as a scatter plot.

So, here's an example. I generated a data set with 3.3 million samples and added some patterns. Let's see if we can see those patterns on scatter plot.

So, here's a scatter plot of the generated two variables:

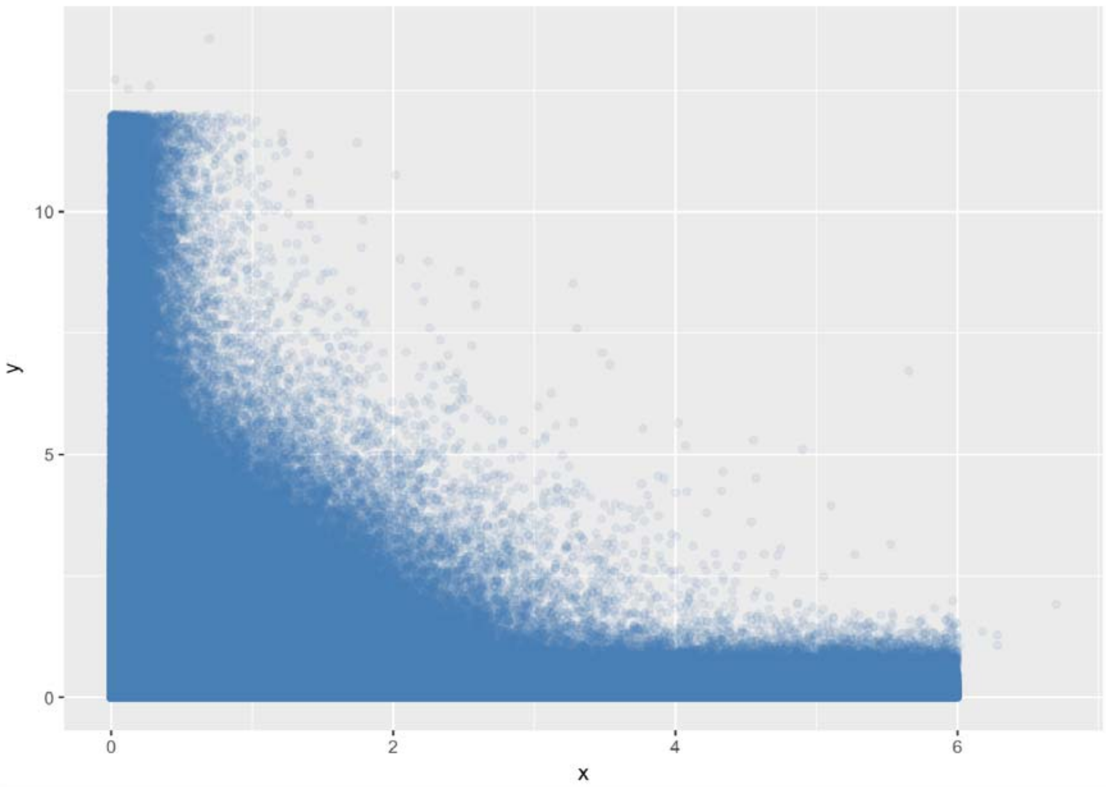
```
library(ggplot2)
library(viridis)

ggplot() +
    geom_point(aes(x = x, y = y),
               colour = 'steelblue')
```



We can see that overplotting issue. Let's try to change alpha to 0.1:

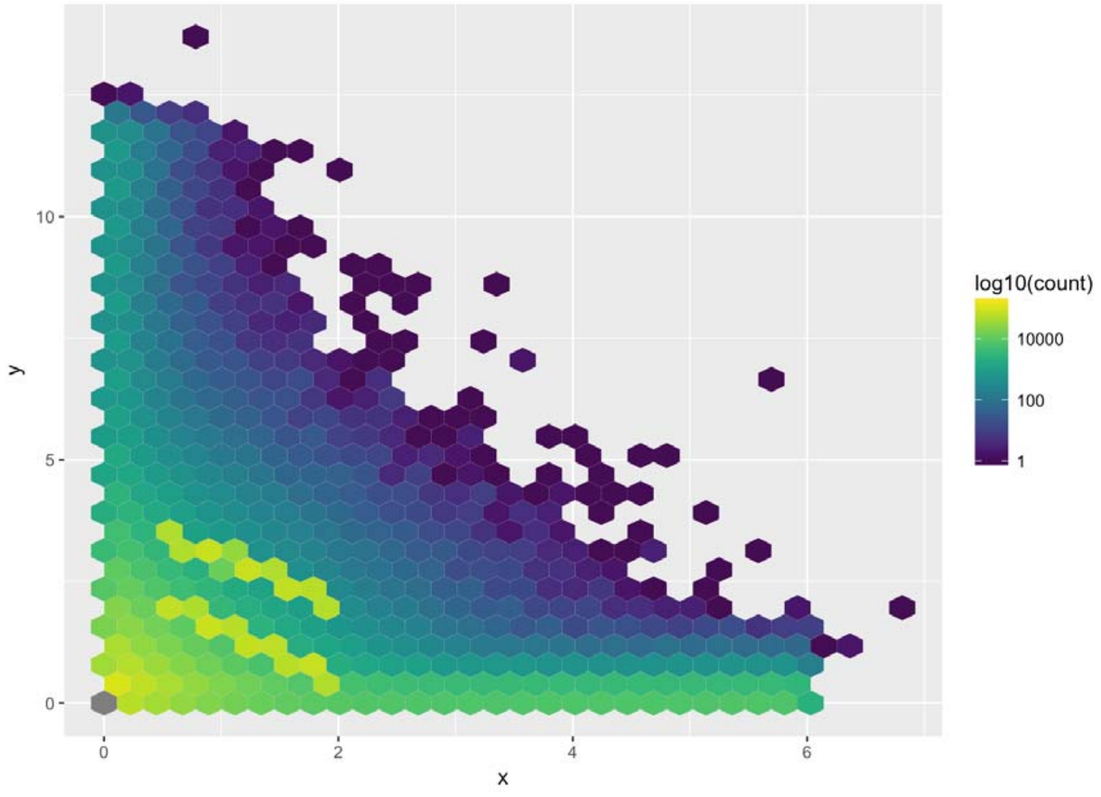
```
ggplot() +
    geom_point(aes(x = x, y = y),
               colour = 'steelblue',
               alpha = 0.1)
```



wrong decisions about the relationship between the variables from these plots. However, let’s continue experimenting with the visualization.

Here’s an alternative to scatter plots. Let’s try to use Hex Bins.

```
ggplot() +
    stat_binhex(aes(x = x, y = y)) +
    scale_fill_gradientn(colours = viridis(256),
                         trans = 'log10',
                         limits=c(1, 200000),
                         name = 'log10(count)'
                         )
```



The picture has changed a lot. Now we can see that there are a lot of (0, 0) samples (if to be precise - 30% of all samples) and two lines. Obviously we couldn’t see those things on the scatter plot.

You can check the code for generating the data set on my github repo.

comments powered by Disqus