COVID-19-Chill

Jeremy Harris

March 23, 2020

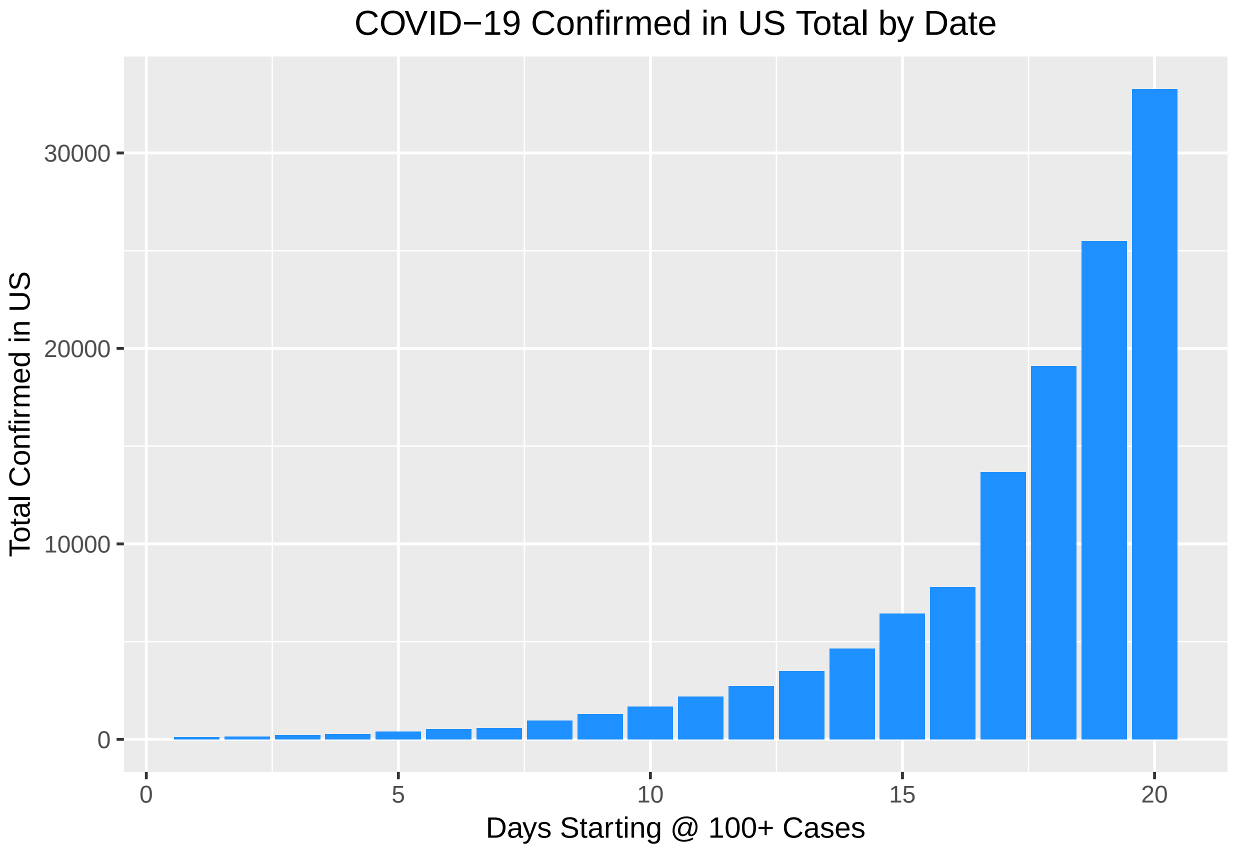
COVID-19 Data & Real-World Statistics

COVID-19 has taken our world over and while it does seem to be highly contagious, I want to look further into the data and provide information that is purely driven by statistics. The data I'm using is from Johns Hopkins GitHub account. The primary purpose of this project is to tell the story behind the numbers and hopefully reduce the potential of over-hyped decisions and further economic disaster if it is not warranted.

Load & Prep The Data

I'll be pulling the data from the Johns Hopkins' time series reports. The reports are broken down into three different csv files that are updated daily: **Confirmed, Recovered, Death**. I'll be consolidating these files so that I can pull out the US data for analysis. I'll be comparing this to Italy as it is used most often in the media for various statistics.

Confirmed Cases Analysis



It's pretty clear to see that there is a dramatic increase in the total confirmed cases per day in the US. However, keep in mind a few things:

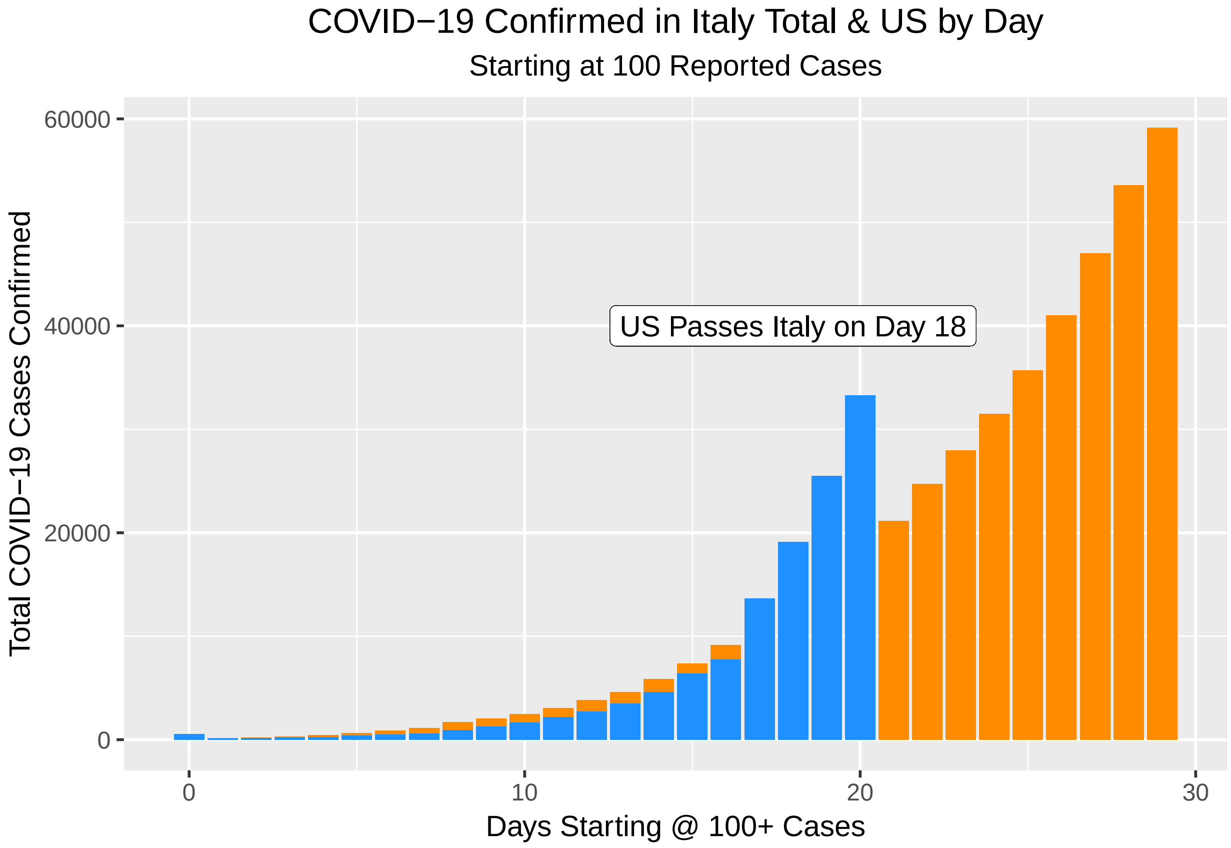
- We are only testing (for the most part) sick or highly likely sick people.
- We don't really know how many cases we have as a starting point to compare growth.

Let me explain further. . .

To get a true understanding of the infection rate we would need to test a random sample of the population and keep testing that same sample over set time periods to determine the infection rate inside of our sample. Then, we would have an established baseline and we could use math/statistics to provide realistic probability outcomes. If we are only testing a biased group then our results will be biased. Furthermore, if we are not continuing to take a sample population and re-test them then we don't really know if any of the measures we are putting in place are working properly. Instead, you see a **shiny graph** like the one above and immediately think: *"Oh my God, we're all going to die!"*. Well, not so fast.

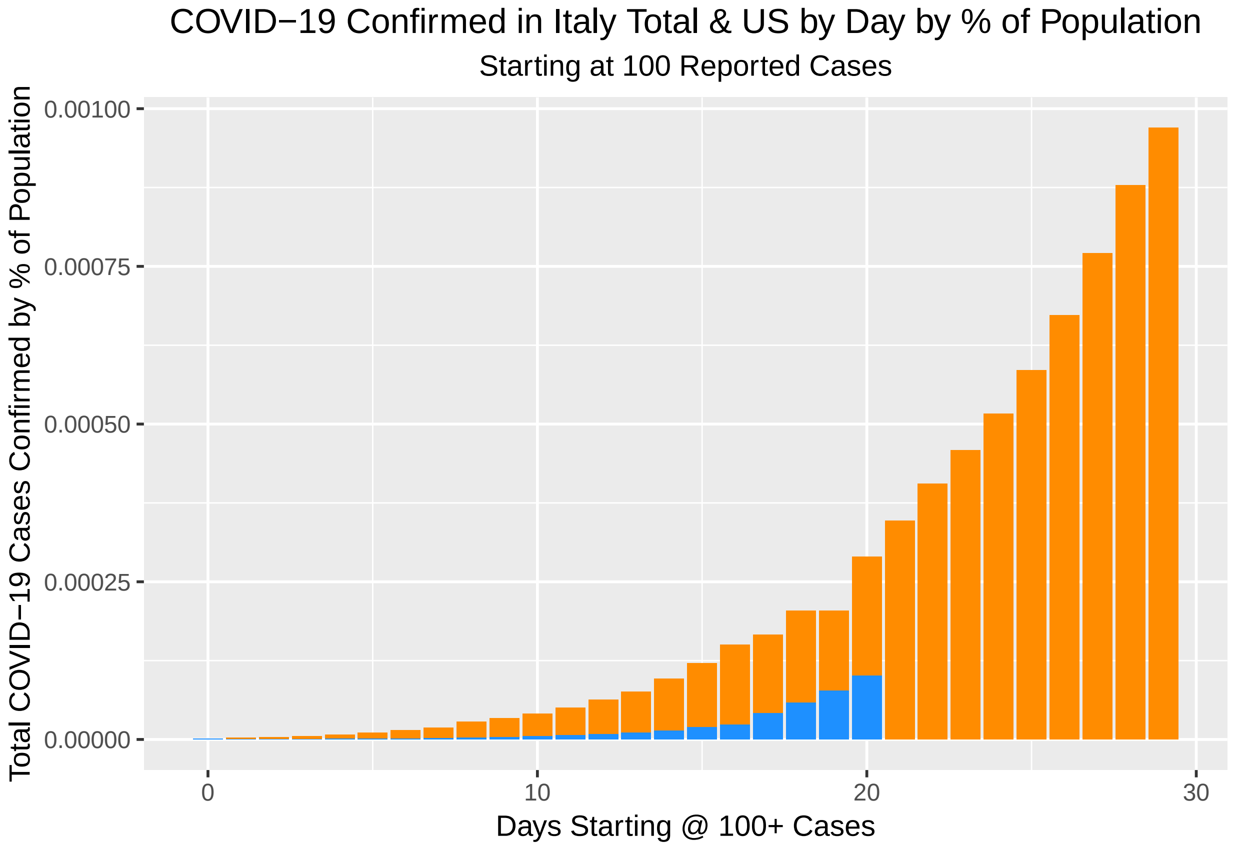
Let's Compare Our Growth Rate Based on Population Percentage to Italy's

Since Italy has become the "worst case scenario" for this virus, then let's see what it looks like if we compare our current infection rate based on the number of cases per day to that of Italy when we factor in our the population of the two countries.



This is a popular graphic that is going around currently. At first glance, it shows that the US is outpacing Italy if you measure from the first day that each country reported at least 100 confirmed cases. However, we haven't taken population or population density into account yet.

US Compared to Italy as a Percentage of Population



```
## geom_label: parse = FALSE, label.padding = 0.25, label.r = 0.15, label.size = 0.1, na.rm = FALSE
## stat_identity: na.rm = FALSE
## position_identity
```

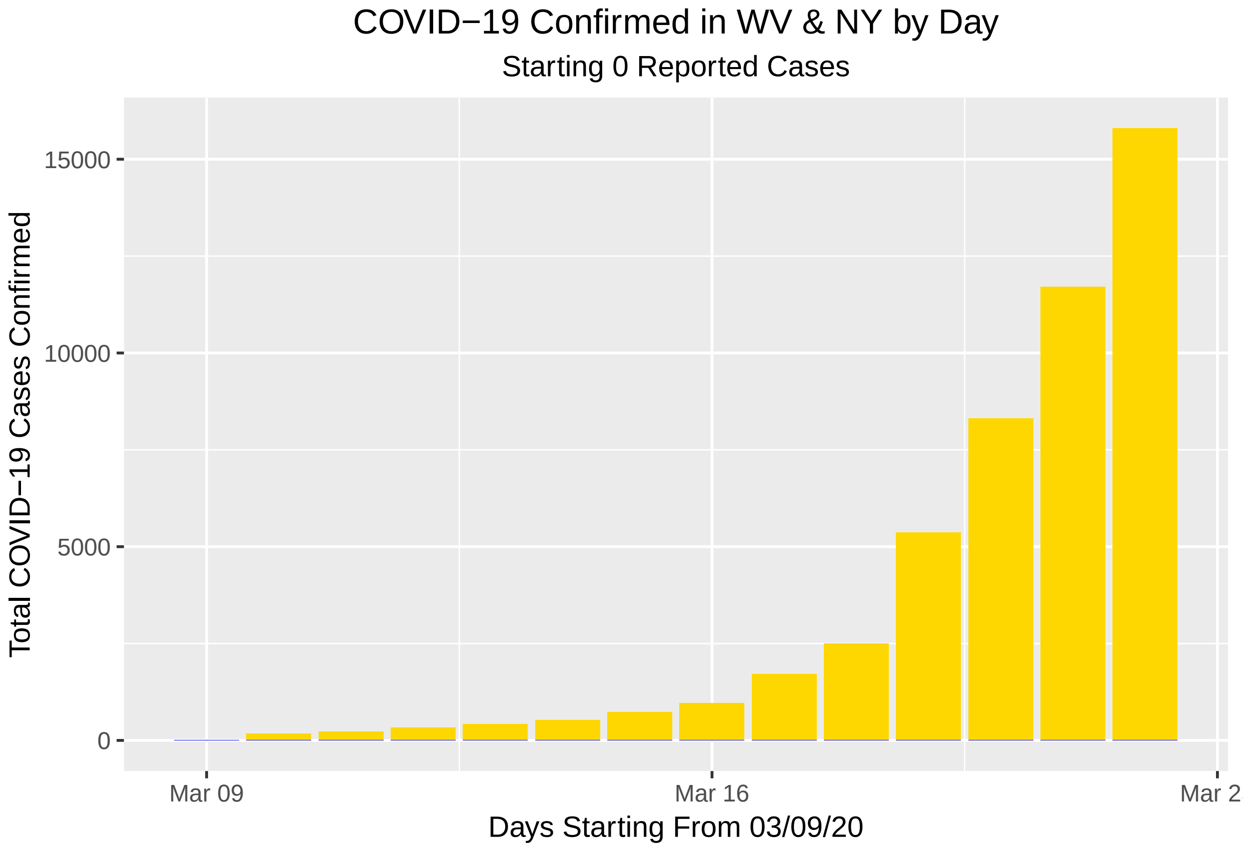
Now this is a little bit better of a picture of how the virus is spreading give the vast population difference. It is clear that there is still an increase in cases but again, we don't have a way to truly measure the spread unless we have a set sample group that is completely unbiased and tested frequently.

What about Population Density Differences?

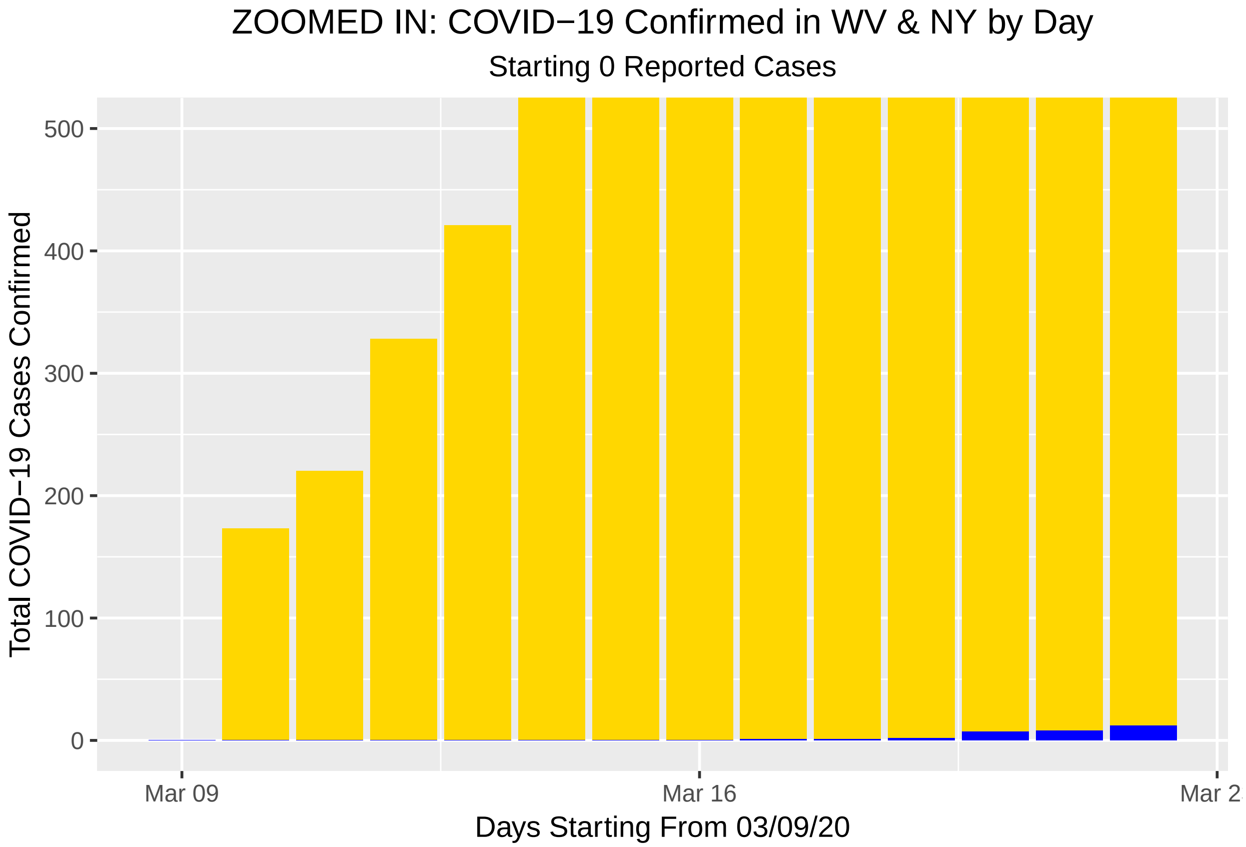
In the US, West Virginia was the last state to report a COVID case. That is partially because testing was delayed and also because there isn't a high population density. West Virginia has a total population of 1.8M. Let's contrast that to New York which seems to be really struggling with COVID. The entire state population of New York is 20M with New York City alone making up a staggering 18.8M of that number.

Let's Compare New York City to West Virginia

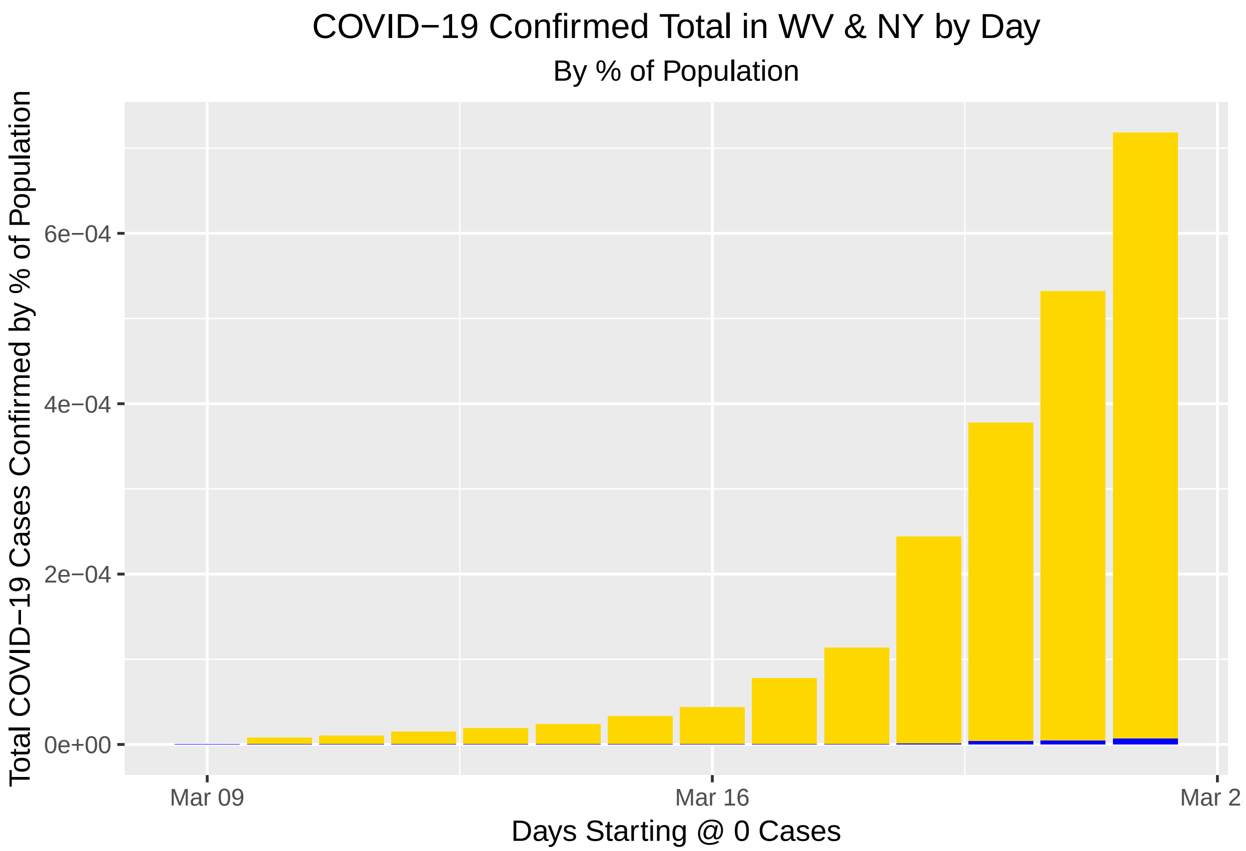
Using the same data, let's see what the spread looks like for a densely populated area as opposed to a more rural area.



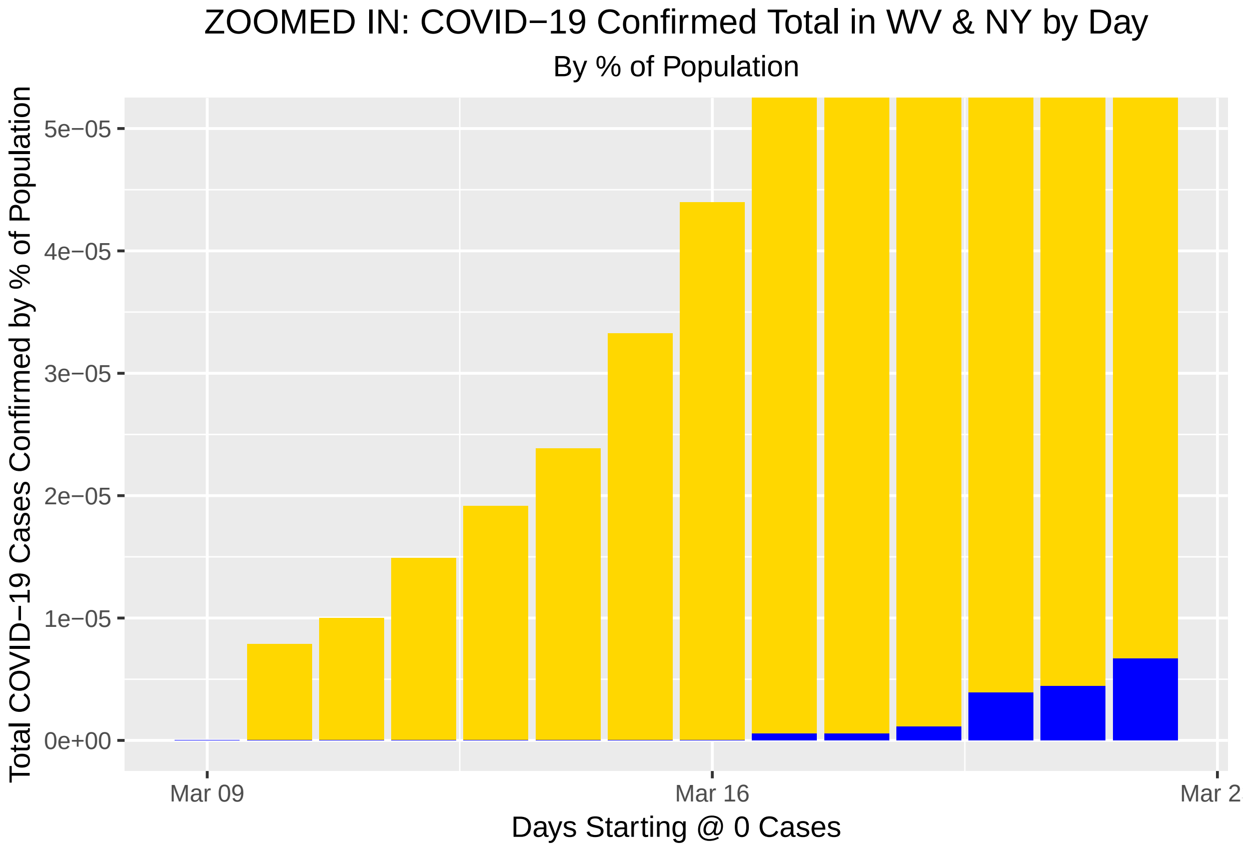
It looks like the data from WV isn't showing up, but in reality, the data is present but there are so few cases in WV that we can't see it at the current scale that the NY data is show. Here I zoom in to a scale capped at just 500 cases so that the WV data actually shows up.



Now, let's take a look again this time considering the % of population which is more relevant in terms of determining if population density has any correlation to the spread of COVID.



I can see the WV data showing up but again it is so few cases that I need to zoom in. I'll cap the percentage to .0004% so we can get a better picture. But notice what I said. . . I'll cap the percentage to .00005% of the population. Even in NY, we are at a very small percentage of the population at this point.



I think we can at least see that population density might have something to do with the spread of this virus which makes obvious sense. With more people packed in a smaller space we would expect a higher transmission of a highly contagious virus.

What About Confirmed Cases By Day by Square Mile

Now this should be even more telling. Let's compare the WV data to the NY data and instead of dividing the confirmed cases by the population, let's divide the cases by the area of the state in square miles. Area of WV is: 24,000 and the area of NY is: 302 and the area of Italy is 116,000 per Google.

To do this, I will create a small dataset that contains each location's area, population & confirmed cases. The confirmed cases will be the first 10 days of the virus being reported in each location. From there, I'll calculate the confirmed infection per population % and per square mile.