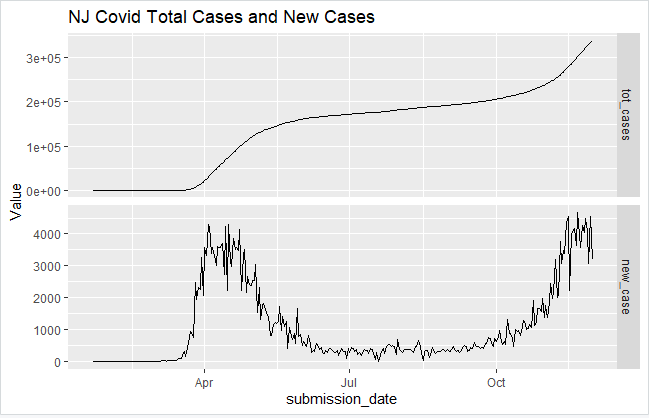
502

Use the file

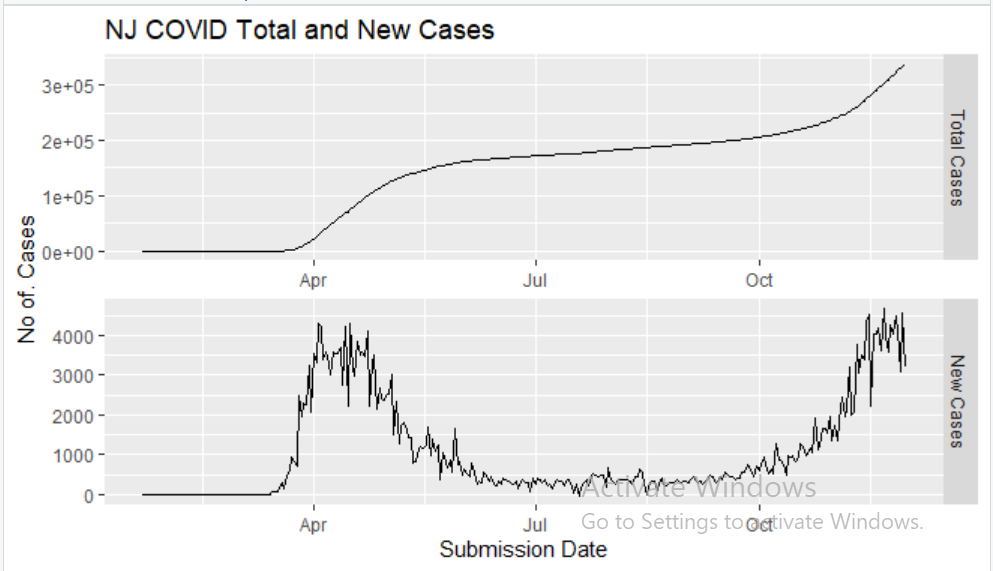
United\_States\_COVID-19\_Cases\_and\_Deaths\_by\_State\_over\_Time:

1. You will bring the data into R and manipulate as needed to create

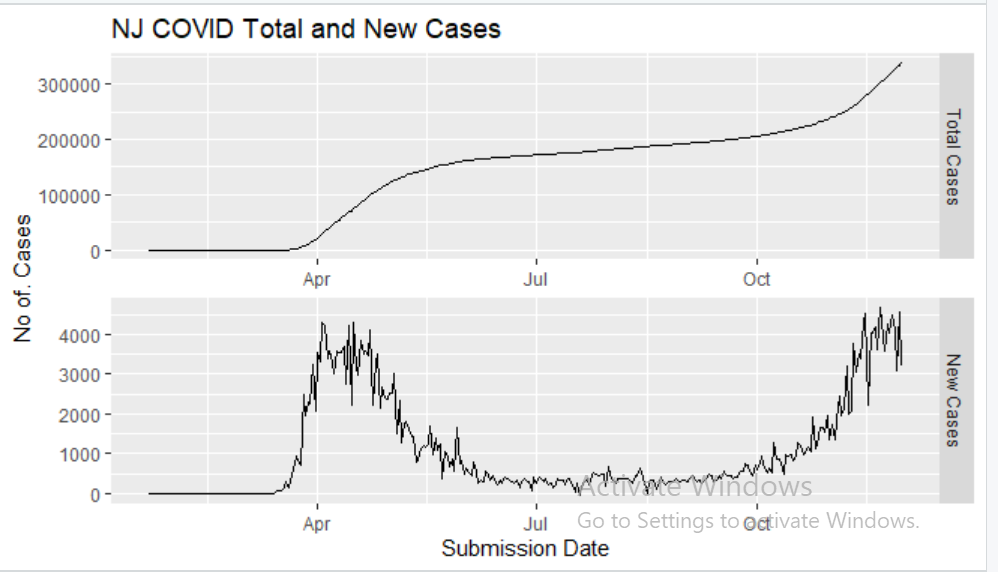
The plot below using ggplot and facet\_grid functions. Obviously your graph will have better labeling!



**My Visuals:**

****

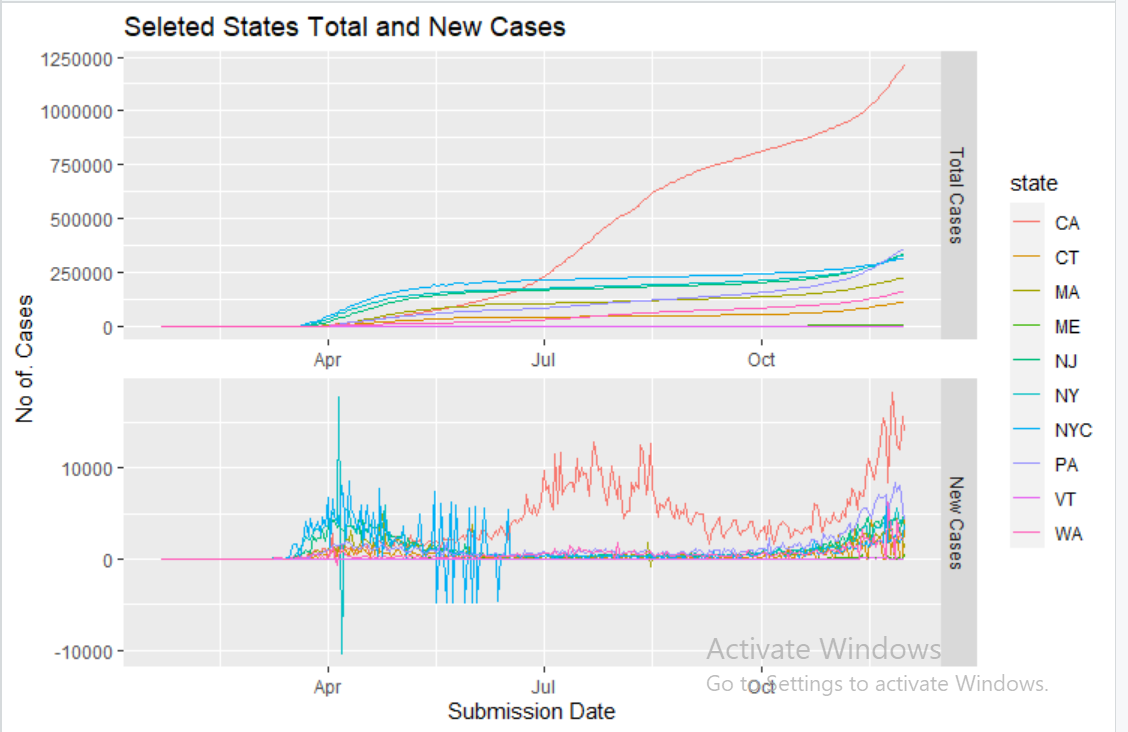
I have also modified the scientific notion to regular for total cases:



1. Do the same as above, except you will isolate more states from the original dataset and each graph will have these extra states in addition to New Jersey: New York, New York City, Pennsylvania, Connecticut, Massachusetts, Vermont, Maine, Washington, and California.

What channel is being used in this second graph to distinguish between the states? What mark is being used for this visual? What unusual behavior is observed for some states and why do you think that is (don’t fret too much on the why, just provide your best insights)?

States Added: ('NJ', "NY" ,'NYC','PA', 'CT', 'MA', 'VT', 'ME', 'WA', 'CA')



Marks: Lines

Channel: Color Hue (to distinguish between states)

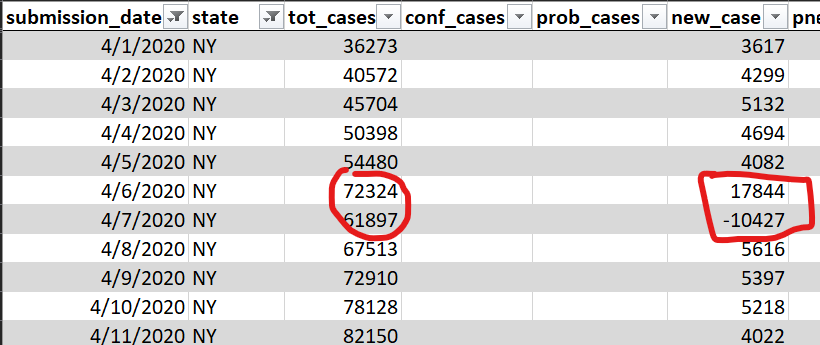
Unusual behavior: Some states has negative numbers of new Cases.

Why is that?

1st: Maybe Because of the people has recovered.

But another thing I found weird is that before going negative there is a positive spike in the plot (all most every time)

2nd: It could be people getting tested and those results of negative our weight the number of positive tests done that day.

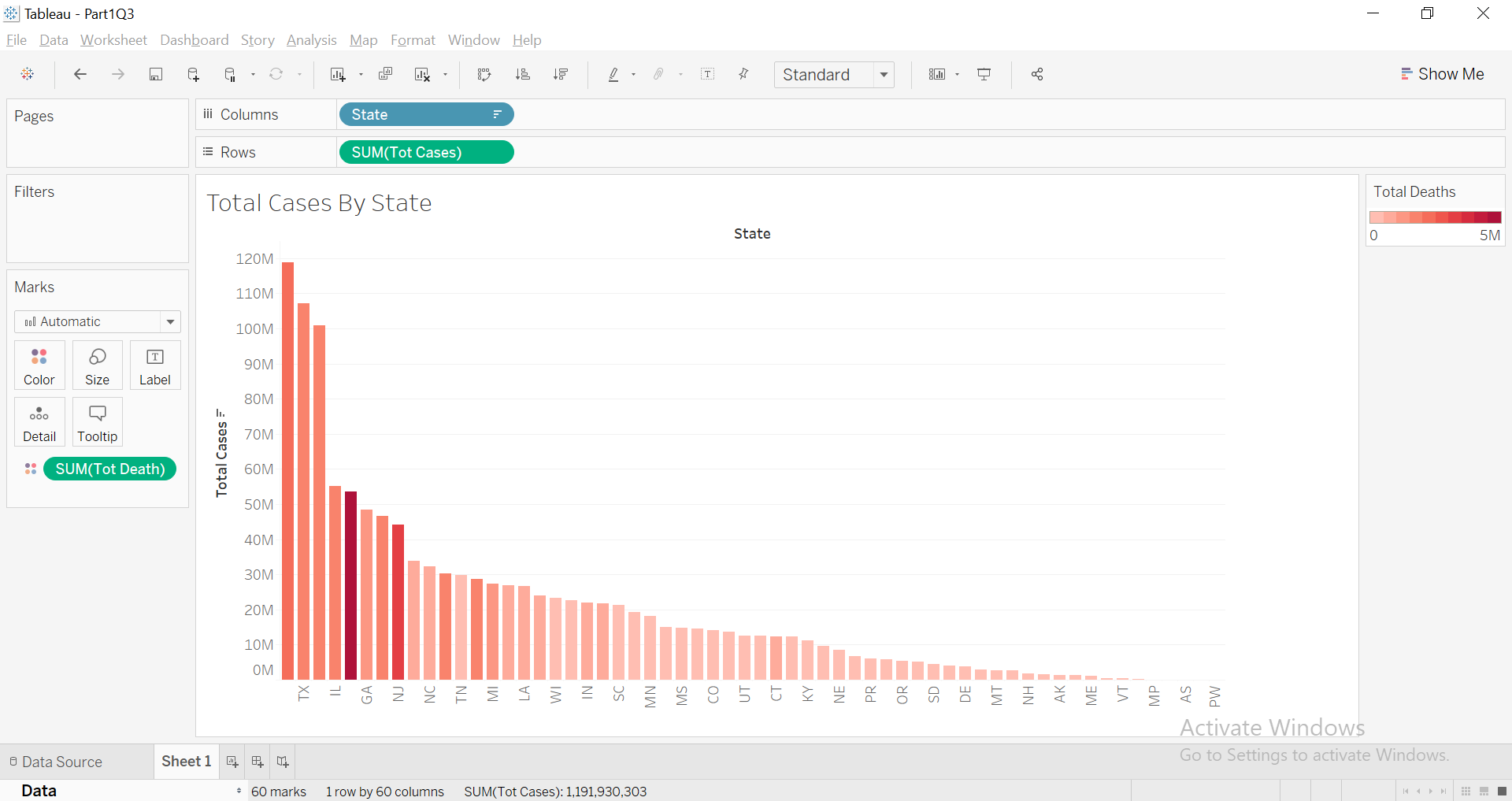


3rd: the negative no. of new cases has subtracted from total cases. I think there is miss calculation of the new cases at previous date that has been fixed in subsequent dates.

That could be the reasons we have seen sudden increase and then later instant decrease at various data points in the plot.

I will go with my third hypothesis.

1. You will use the same original dataset and create one bar graph in tableau that represents the total cases for every state in the US, and also uses the appropriate color channel to encode the total deaths. Provide a brief explanation of what you observe.



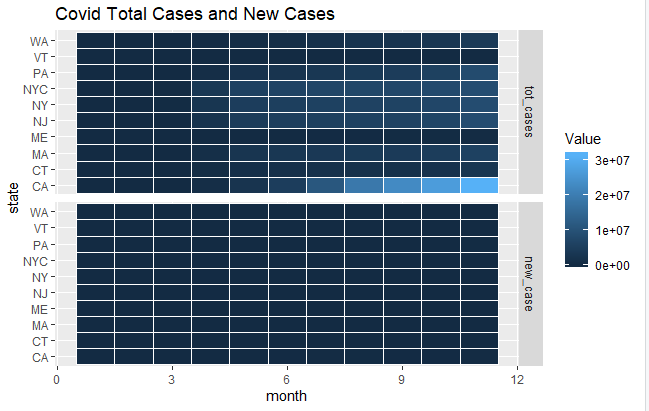
Observation: Even though some states has less number of total cases as compared to other sates yet they have high death rates.

For example: NJ and NYC

1. Lastly, you will use this second dataset:

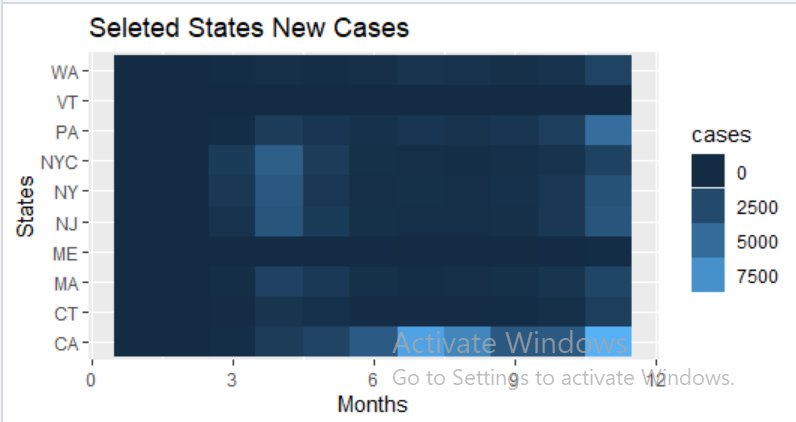
StatesCasesCovid\_Grad.csv

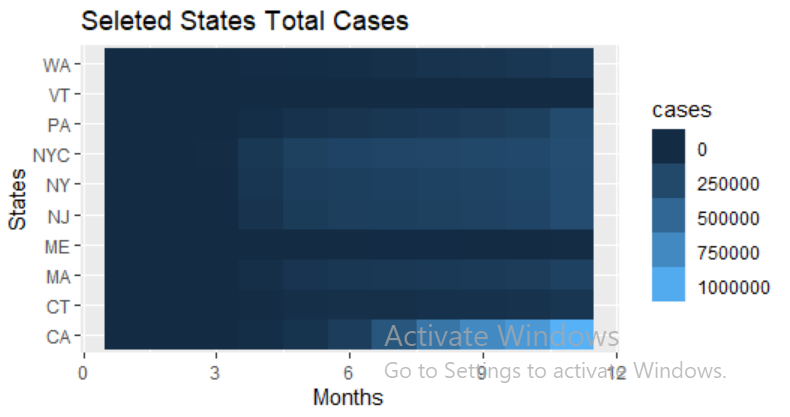
to help you create a couple visuals and answer the question: why is the below heatmap not the best representation of the data?



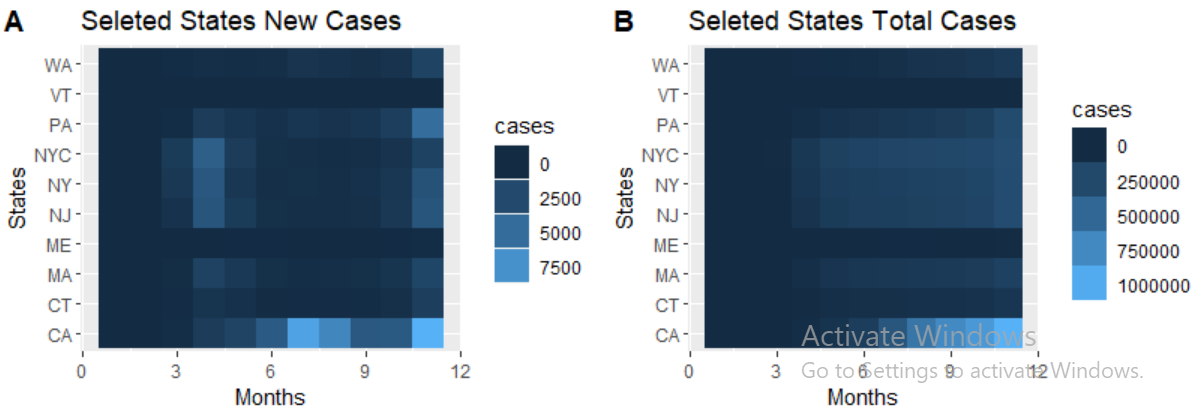
The visuals you will create are two separate heatmaps, one for total cases and one for new\_cases, that will be a part of this exam but should also help you answer the question above. You will have to aggregate the values to one sum total for each month. The dataset is prepped for you to accomplish this, but you will have to remember how to aggregate/summarize and combine the two graphs into one visual reflecting both.

**Note:** I am using the original dataset.





**Both Together:**

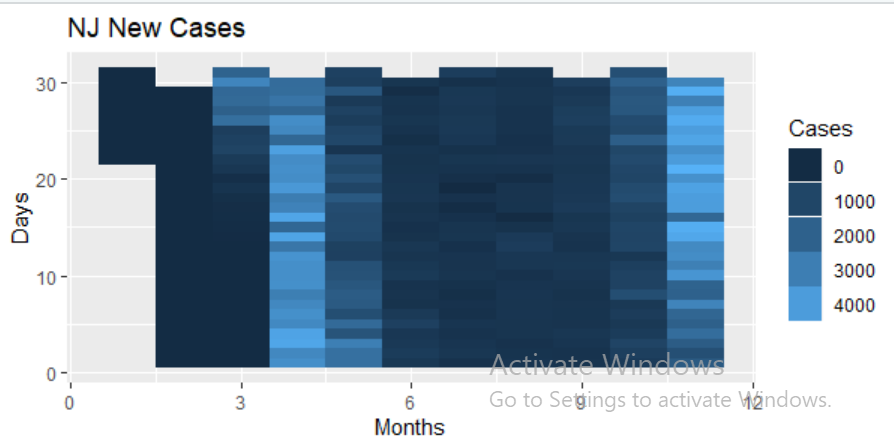
****

**Problem with the plot in question:**

Both plots have a huge difference in values. plotted on same scale (i.e. the legend on right is same for both). New cases are comparatively far less than the total cases. Which makes the new cases graph to much skewed. Hence, not representing the true insight of the data.

Plotting both separate and comparing them has solved the issue.

1. You will use the same dataset as above in 4 and create a heatmap for the state of NJ with month on the x-axis and day on the y-axis to see the trends of new cases from day to day across months. What do you observe?



**Observation:** Same as the line plot in question 1.

No number of cases in up to the 3rd week of January. Then a few cases started slowly. By the end of march cases began rising and in April it reached the peak then gradually getting down in May. Very few cases from July to Oct. Again, cases start rising in October up until November.

In short, April, May, Oct and Nov was hit badly.

1. EXTRA CREDIT\*\*\* 5 points extra if you are able to use the original dataset for are the visuals without having to pull in the second dataset. To accomplish this you will need to map values (ie replace month names with month numbers).

**I used original data set.**

month(submission\_date, label=F)

**label = F** gives the month number instead of word. Hence, no need to replace month names in that cases.

Hint\* Along with all the work you have done this semester, I think homework 7 will also help you. Remember that you have a time based column to handle, don’t forget to look at the structure of the variables if you are having trouble (ie timestamp, factor, integer, numeric, etc etc).

Submission instructions:

Please submit code along with a snapshot of your visuals in a word document. Also, please provide your R code file and tableau file. I should be able to run through your code to recreate your visuals. You can also comment in your answers in the rcode itself. I plan to go through the code, but in case there are any issues, I would like the word document as a backup of your work and visuals.