ECE 5984:SS: Applied Machine Learning - Homework 1

Raw table is uploaded to excel as shown in Figure 1.

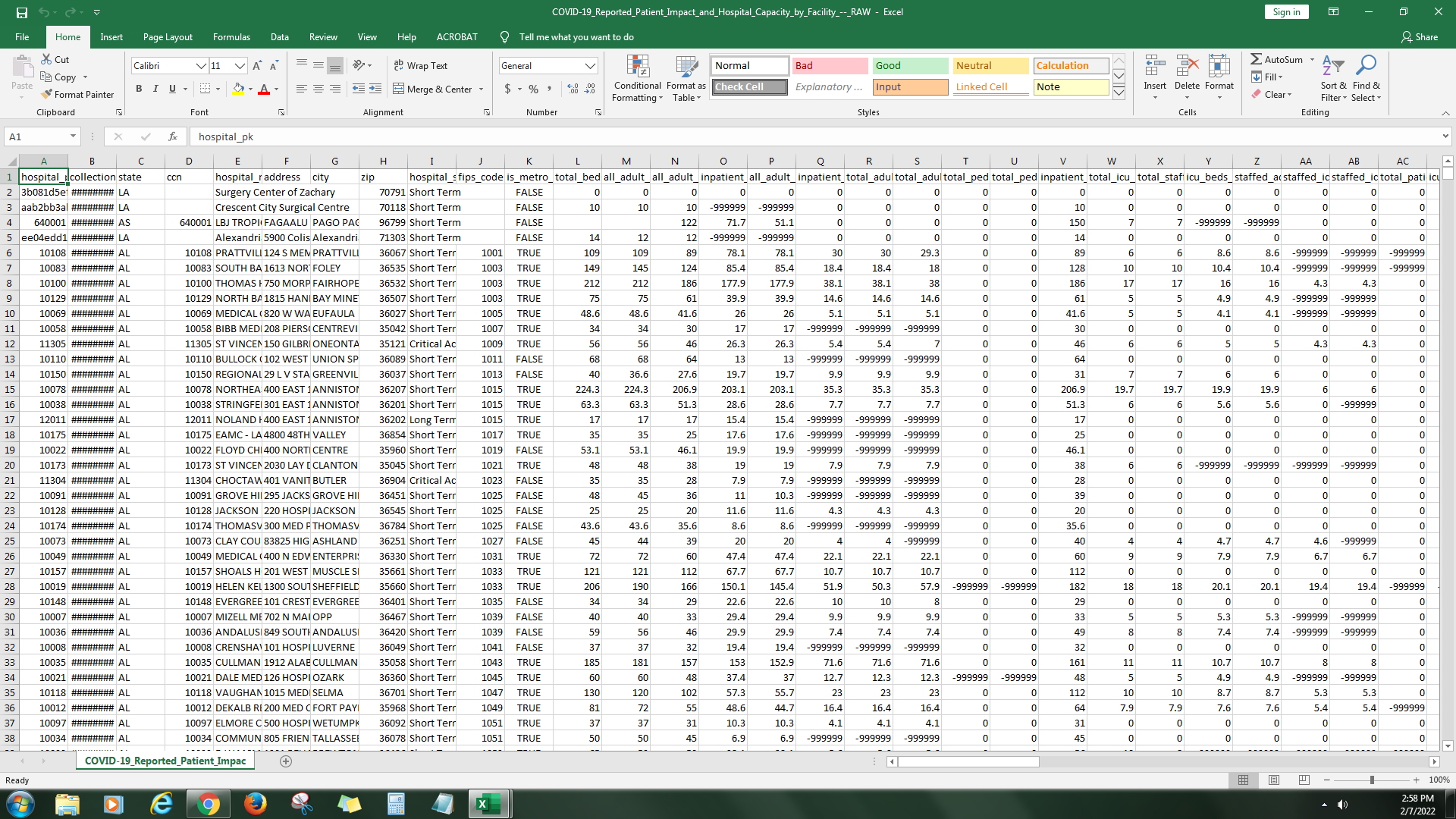


Figure 1

For the first column I just checked whether there is completely numeric entries but there is non as shown in Figure 2. They all accepted as valid since they are ids.

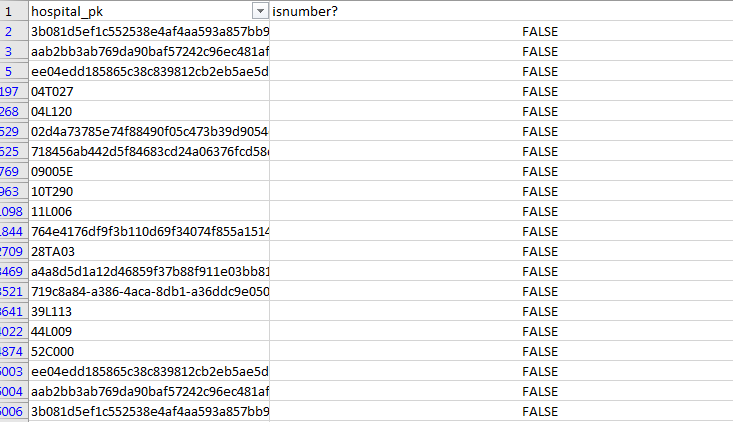


Figure 2

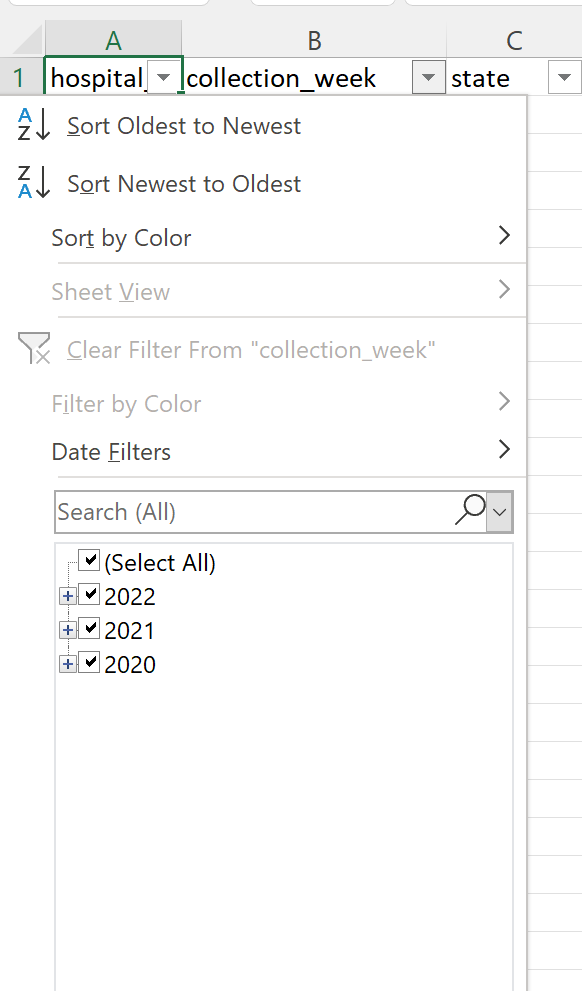
For the second column named collection week using filter tool we can observe the dates entered are valid.

Figure 3

Moving on to the 4th column, there are some crn numbers that are not numbers and they are circled again with data validation as shown in Figure 4. Data validation is done with the whole number selected in the range 10000-700000. But those outliers are the same with the id so I consider them as valid entries.

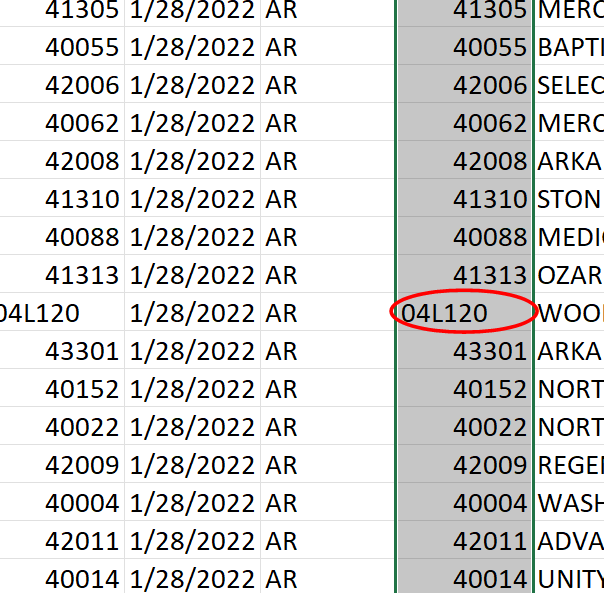


Figure 4

Moving to the 5th column,since this is the address column I just checked if there is completely numerical entry. If not, I accepted them as valid.

For address columns I use the same method and look for pure numerical entries, since there was none, nothing is replaced until numerical columns starting with total\_beds\_7\_day\_avg. As shown in Figure 5 there is invalid entry of -999999 and it can be observed throughout the data table. Therefore we remove it from the whole workbook since there is no column that data is valid. This is shown in Figure 6, the within part is changed from sheet to workbook.

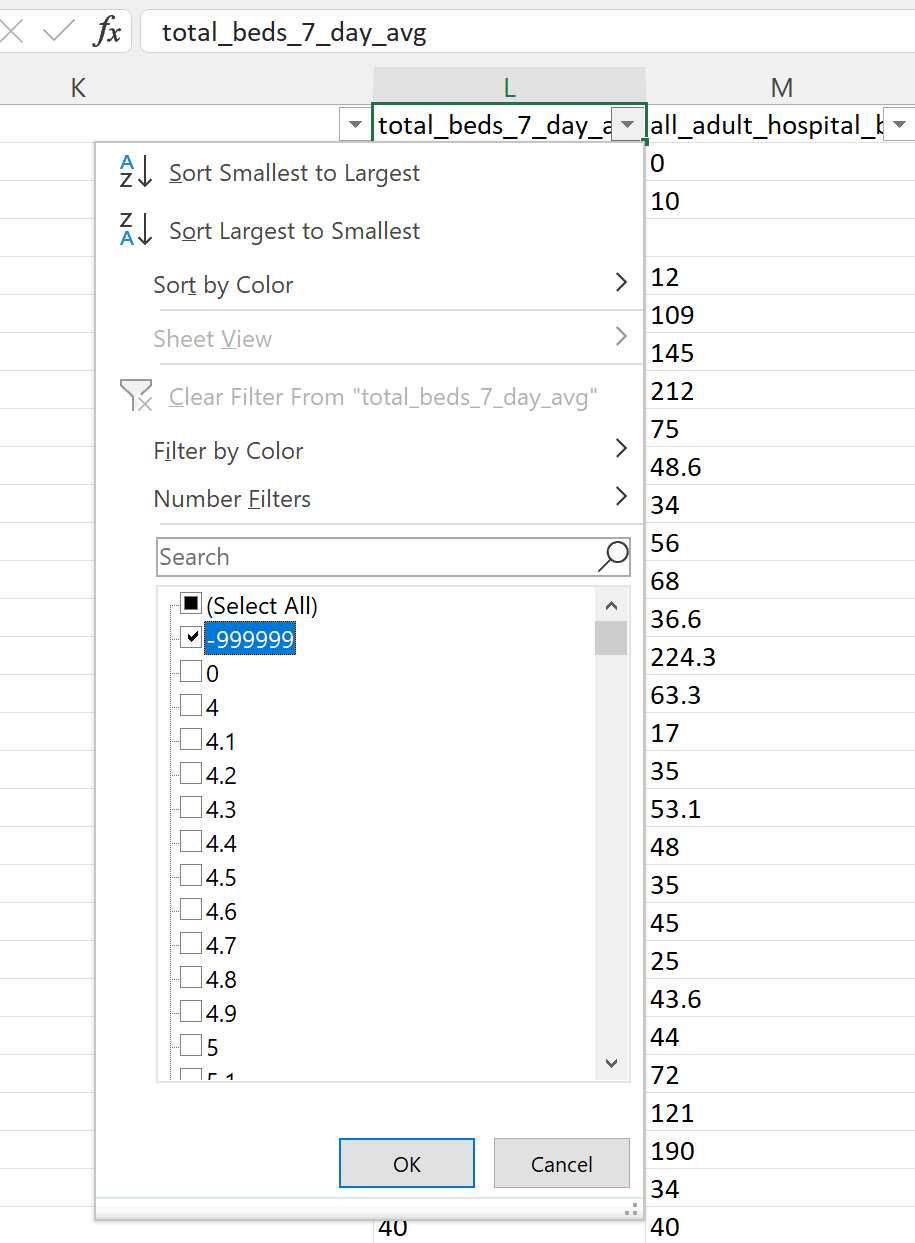


Figure 5

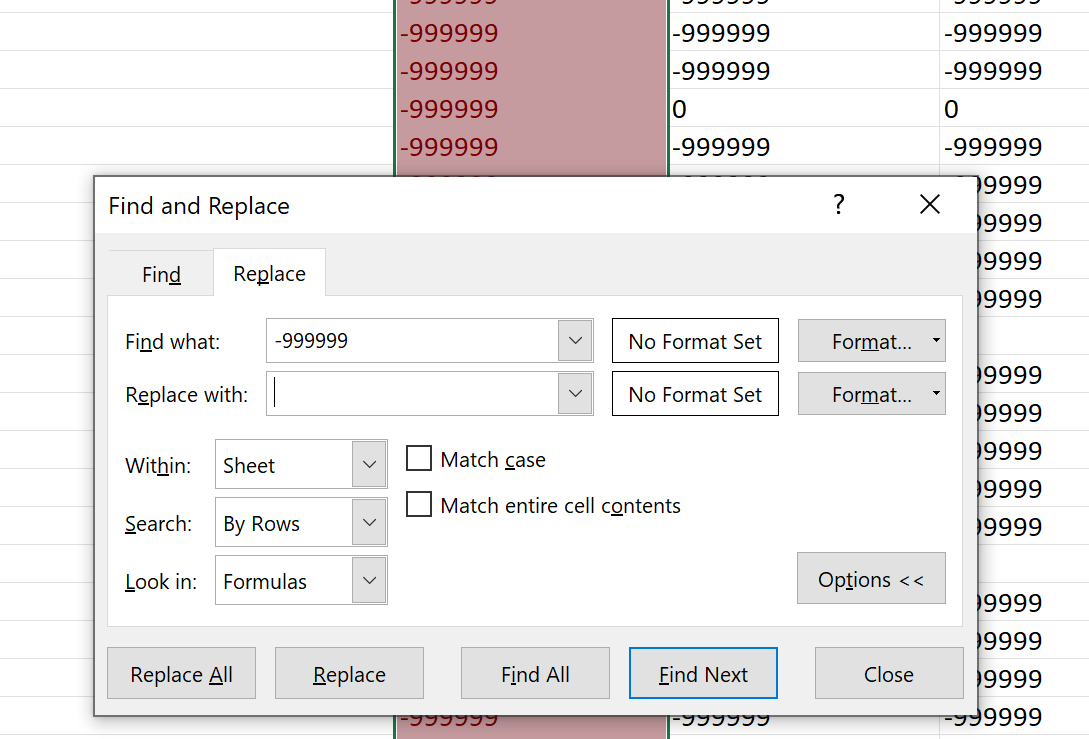


Figure 6

For the rest of the numerical columns, using the find and replace tool, the most observed invalid/negative values are removed from the worksheet. No systematic way is used in the process because there is too much data. The values observed and removed are;

| Number of entries replaced | 1504698 | 119730 | 57311 | 1813 | 169 | 345 | 513 | 378 | 84550 | 18 | 2 | 9 | 4 | 4 | 30 | 1 | 8 | 2 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| The invalid data | -999999 | -8 | -7 | -6 | -5 | -4 | -3 | -2 | -1 | -0.1 | -0.2 | -0.4 | -0.6 | -0.7 | -9 | -3.8 | -2.3 | -2.6 |

In Figure 7,8 and 9 the descriptive stats of the columns are displayed.

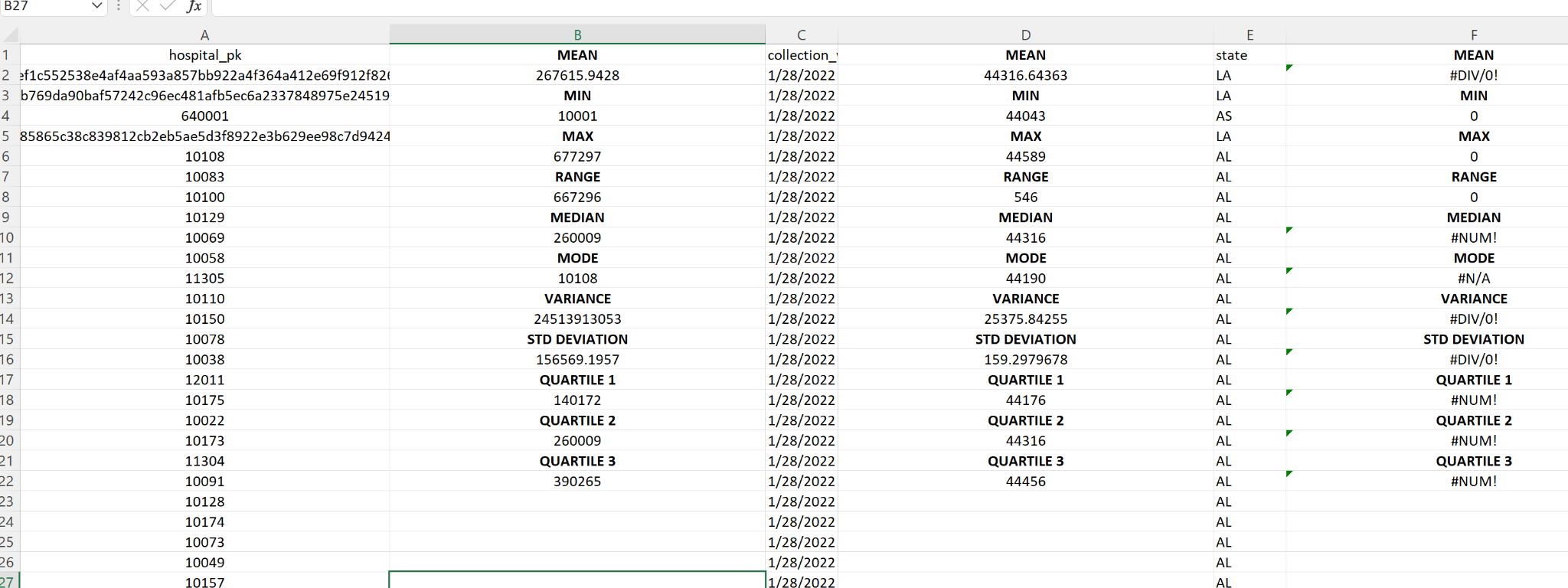


Figure 7

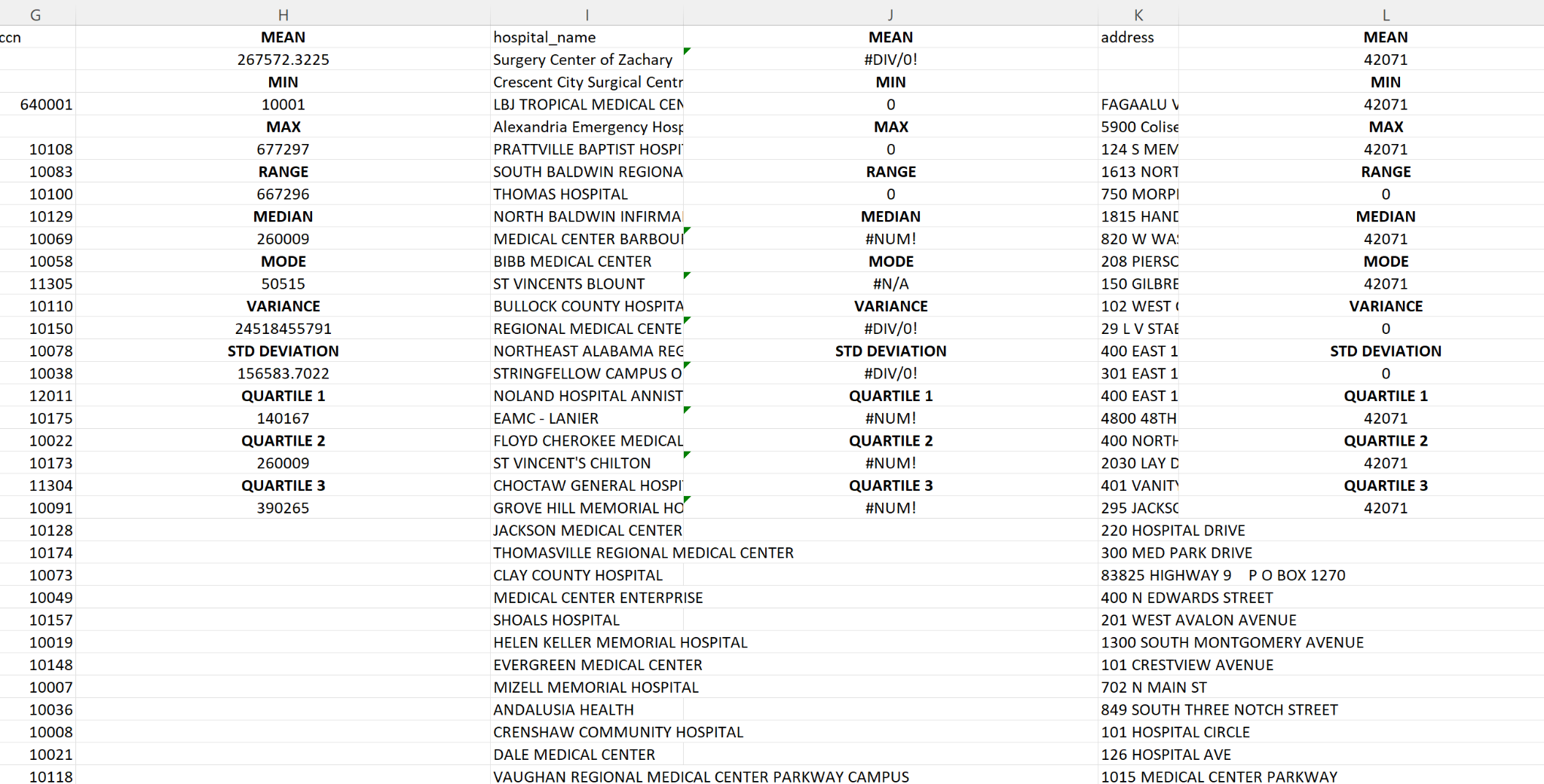


Figure 8

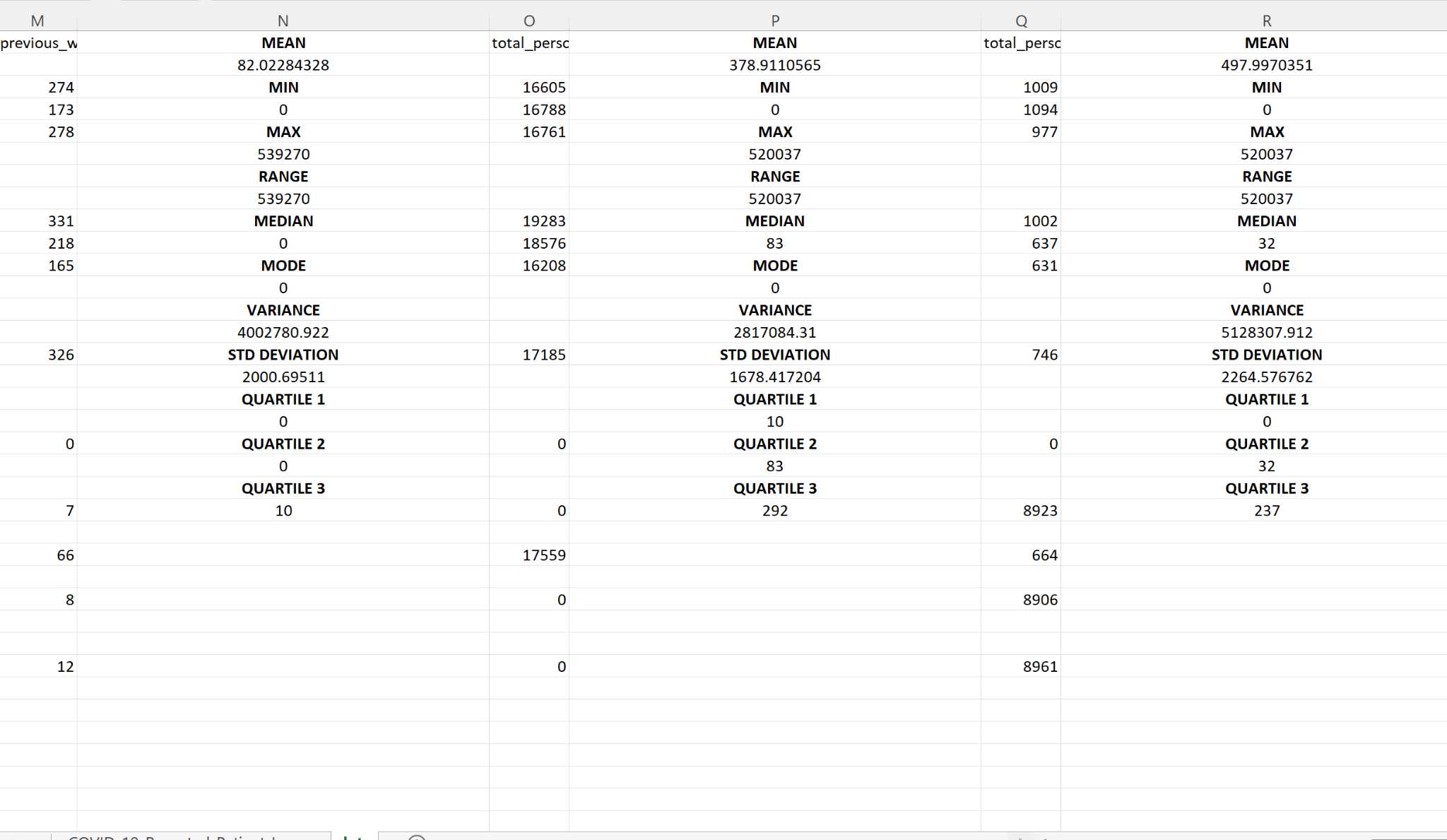


Figure 9

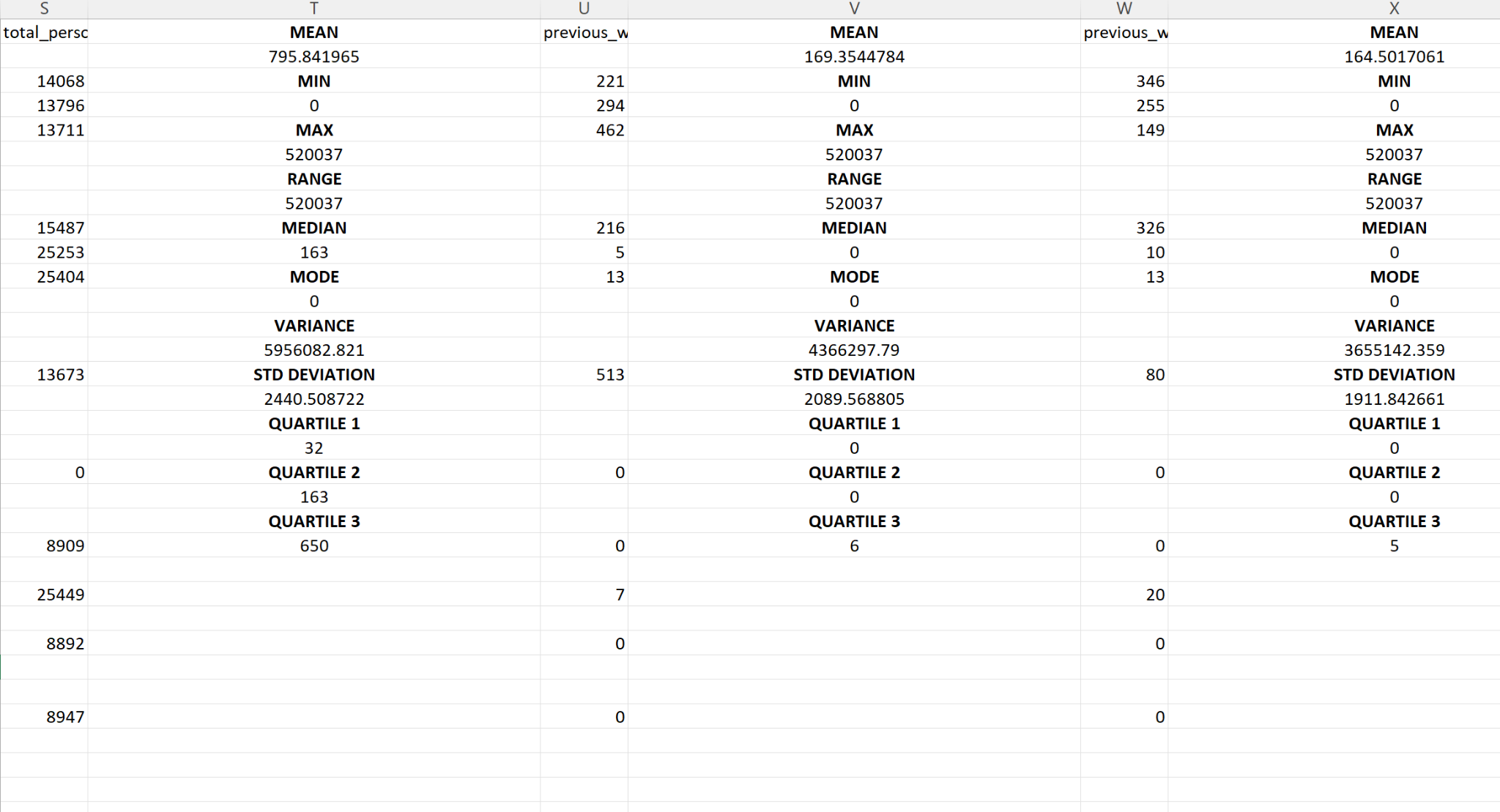


Figure 10

Question 6.

A.Table

The table created is given in Figure 11,12,13,14. The table is made but the bars are invisible for some problem unsolved therefore their values are displayed individually as follows,



Figure 11

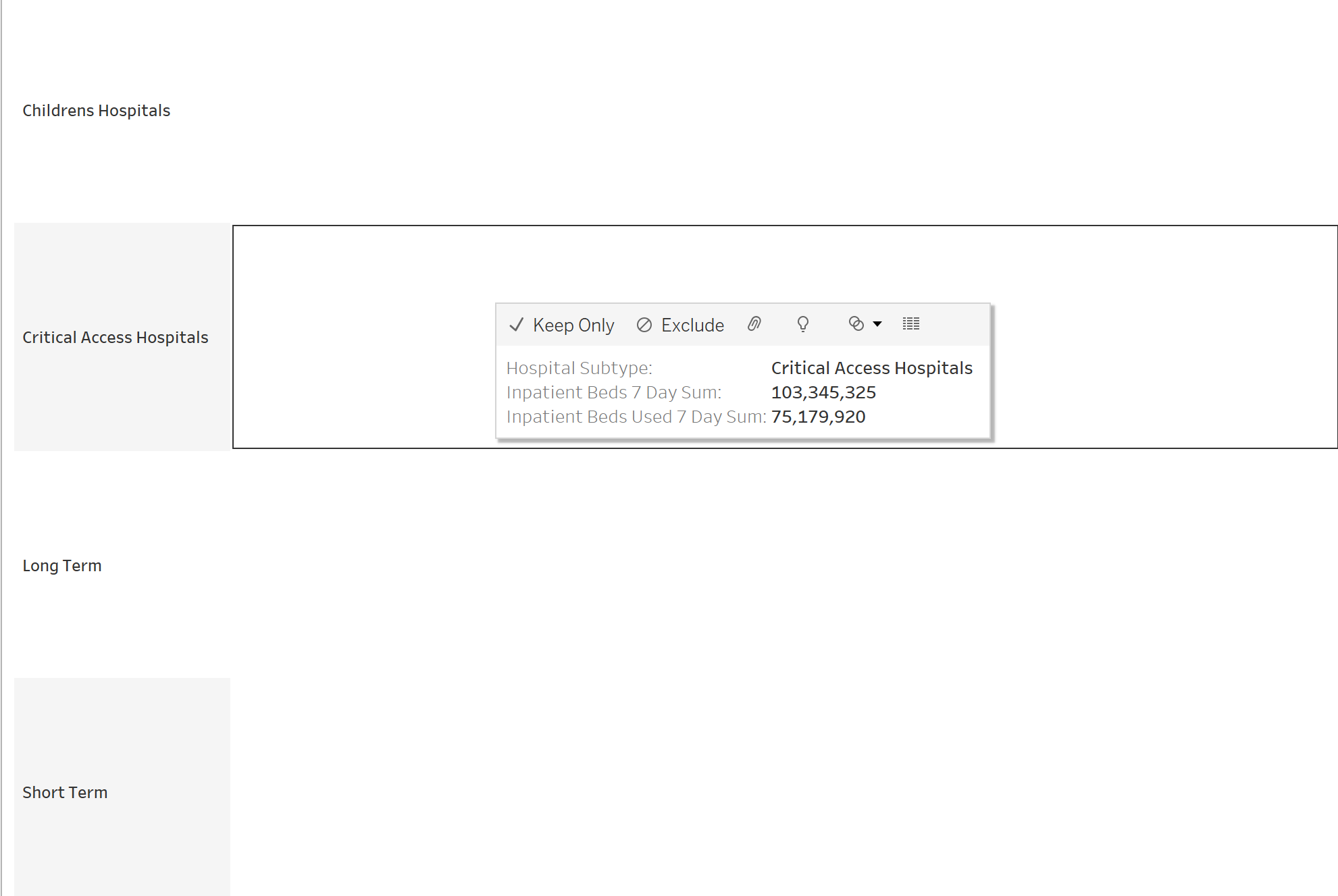


Figure 12



Figure 13

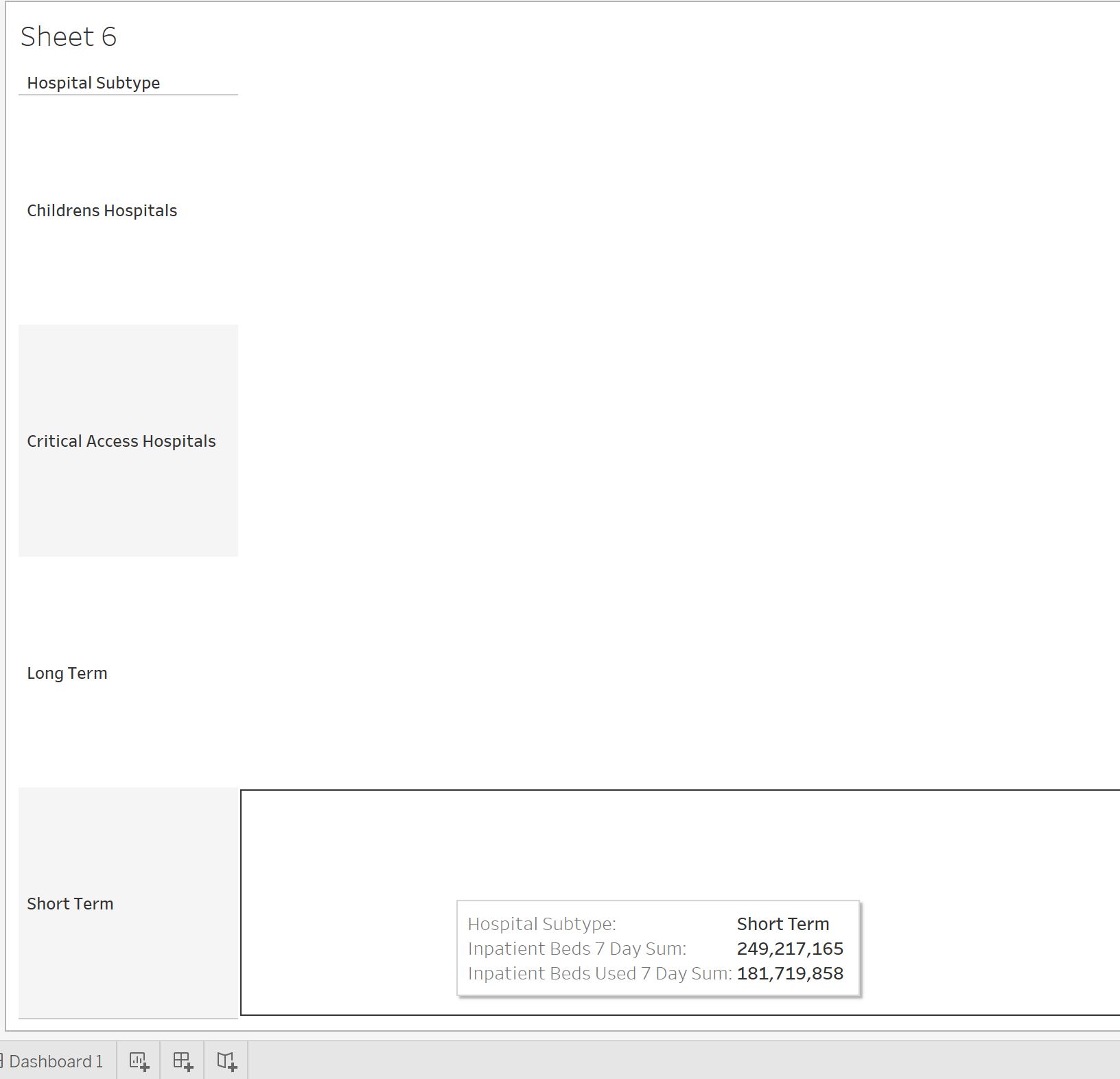


Figure 14

B. Histogram

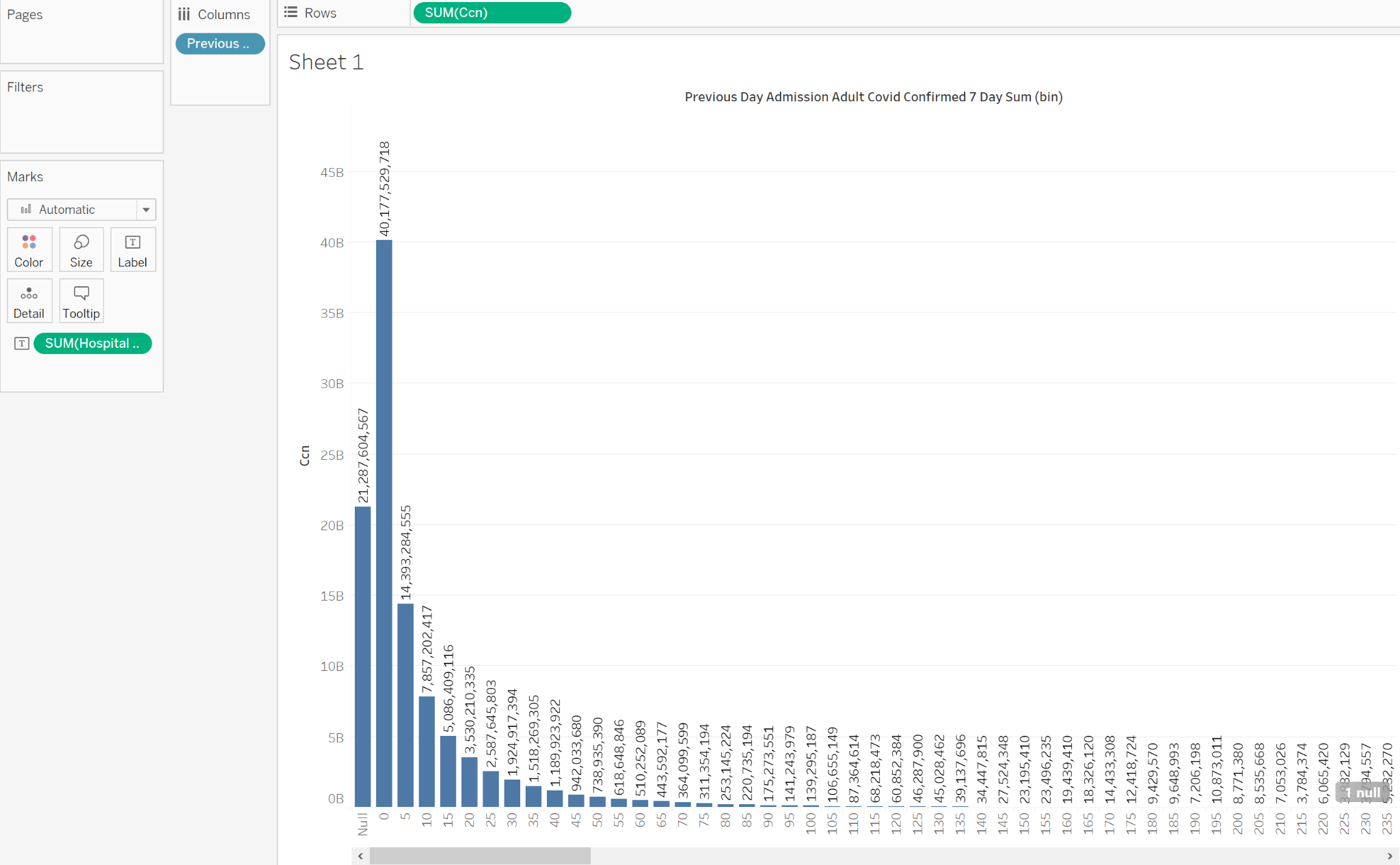


Figure 15

C. Line Graph

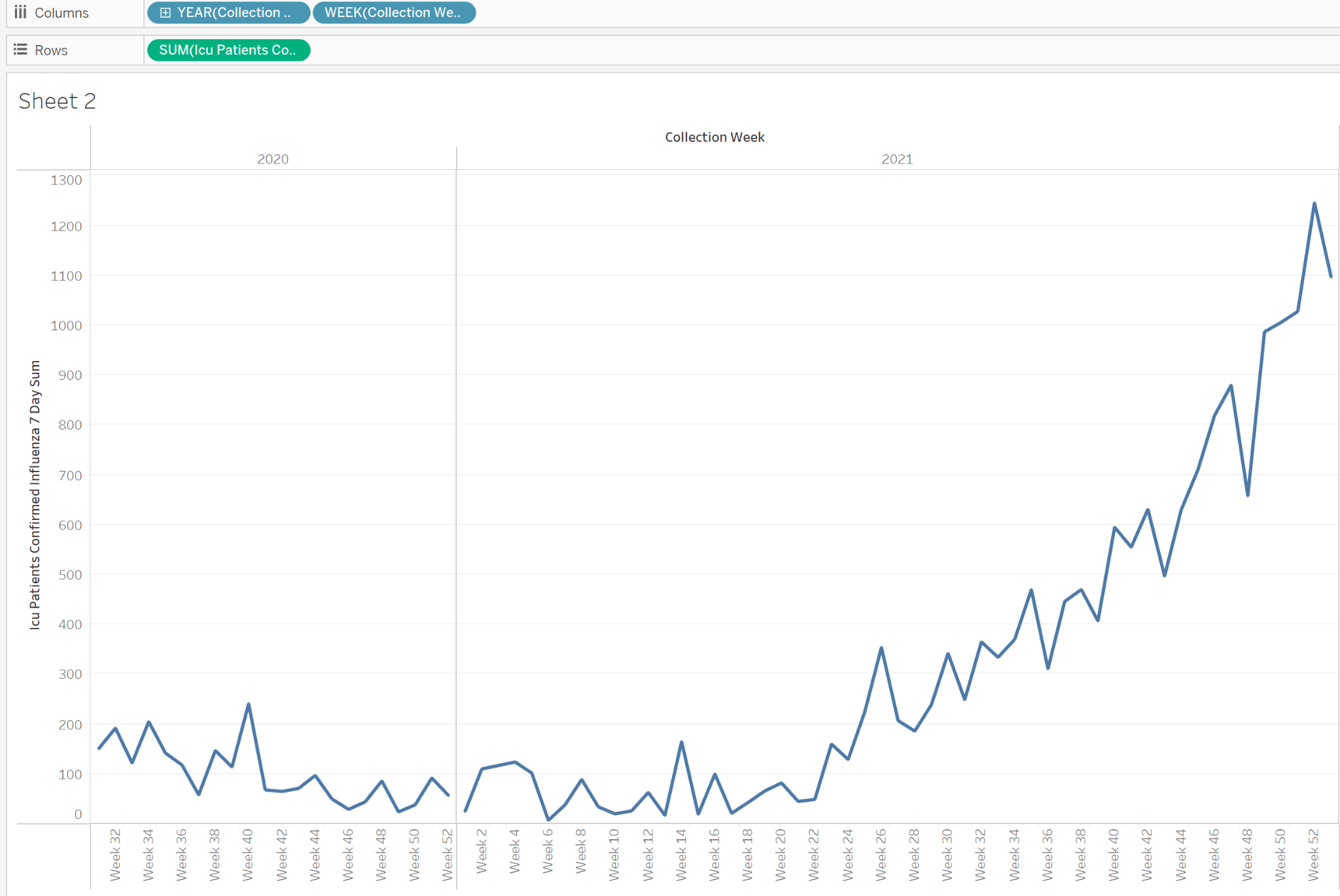


Figure 16

D.

Medical center of the rockies with 126.7

Signature Healthcare Brockton Hospital with 125.9

ST Joseph Medical Center with 97.4

Oklahoma Center for Orthopaedic&Multi SP with 95.0

Wellstar Douglas Hospital with 86.3

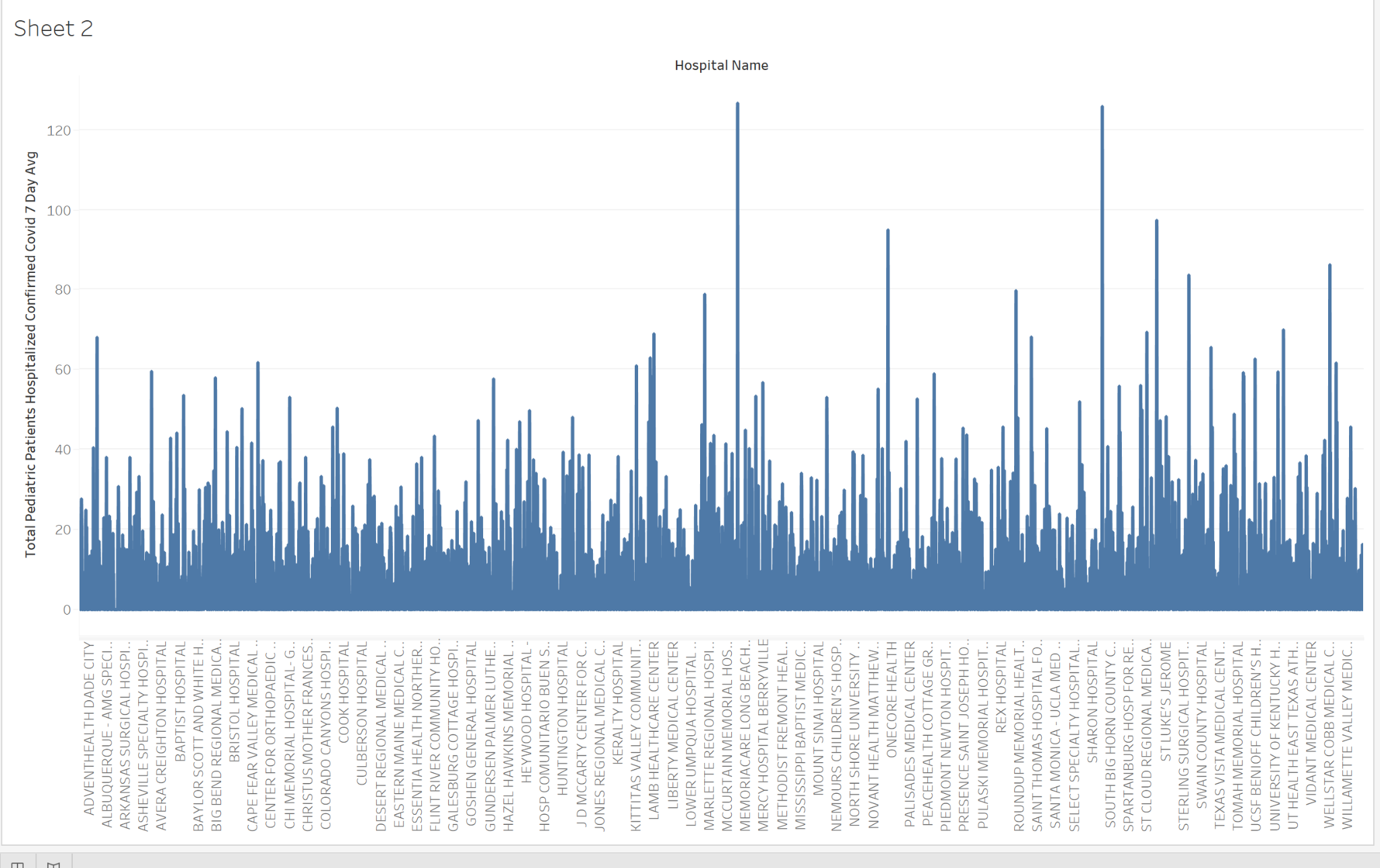


Figure 17

E.

CA with 3138

MI with 1992

FL with 1865

TX with 1699

IL with 1553

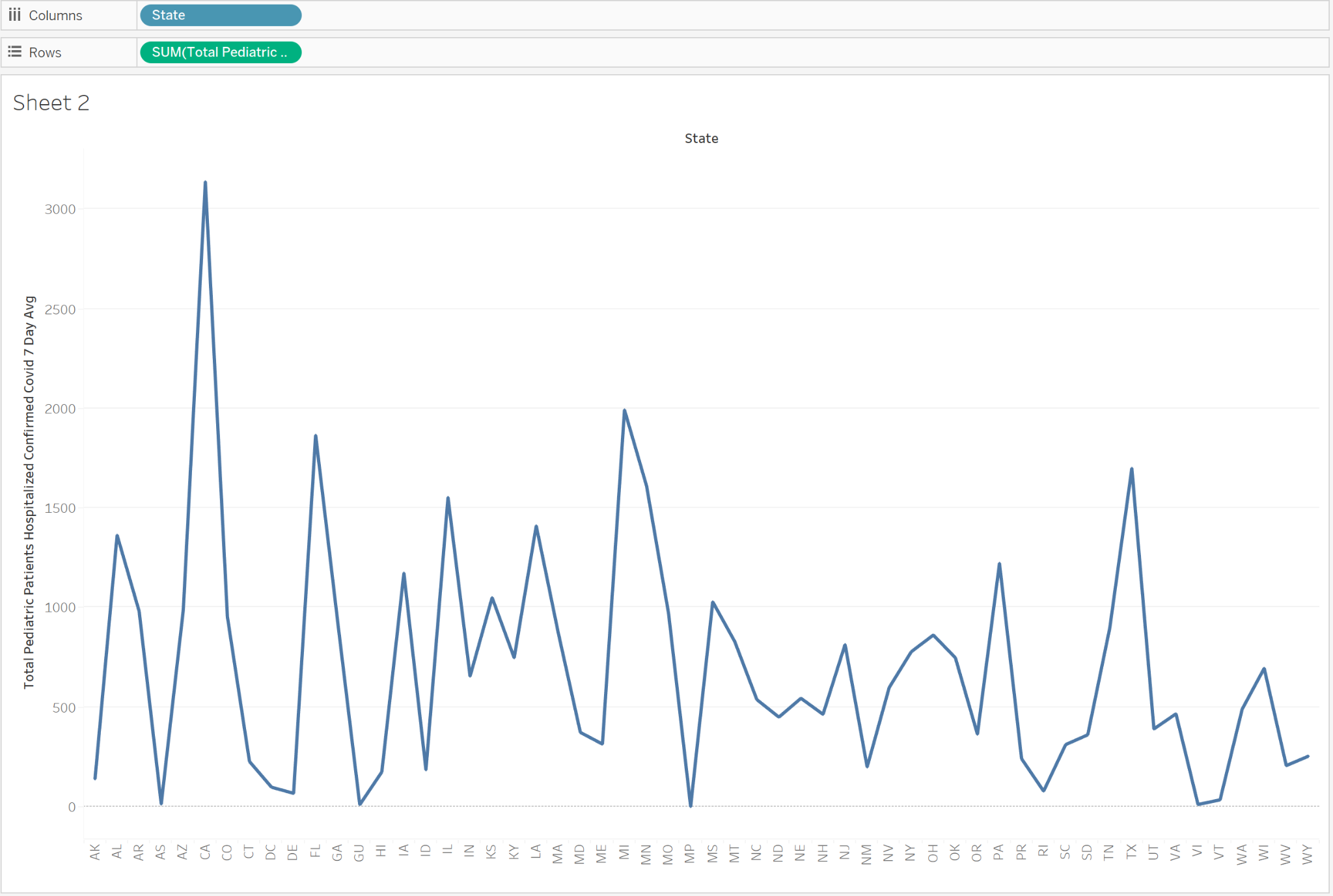


Figure 18

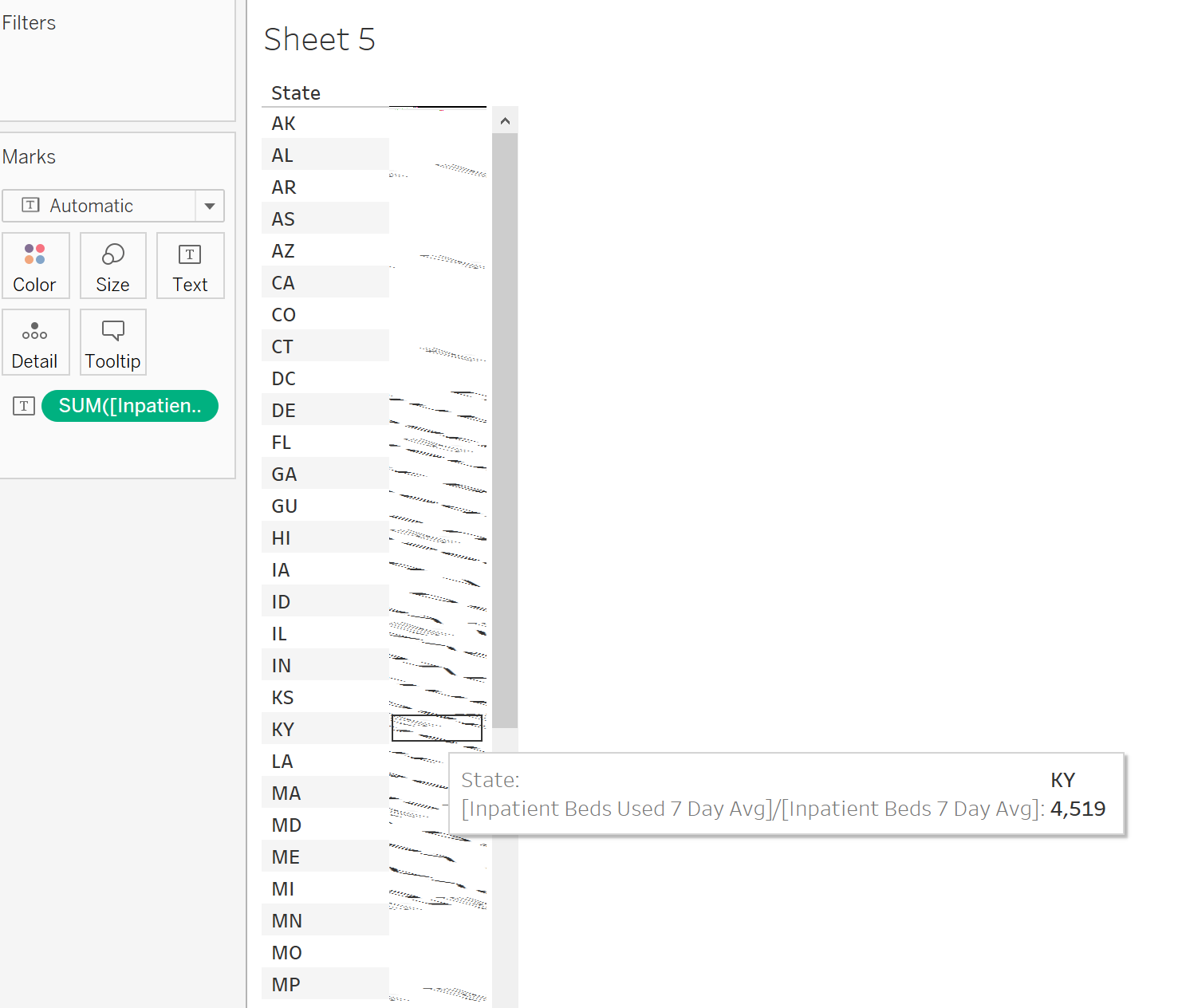


Figure 15

G.

Memorial Hospital with 145841. This is found with the filter as shown in Figure 16.

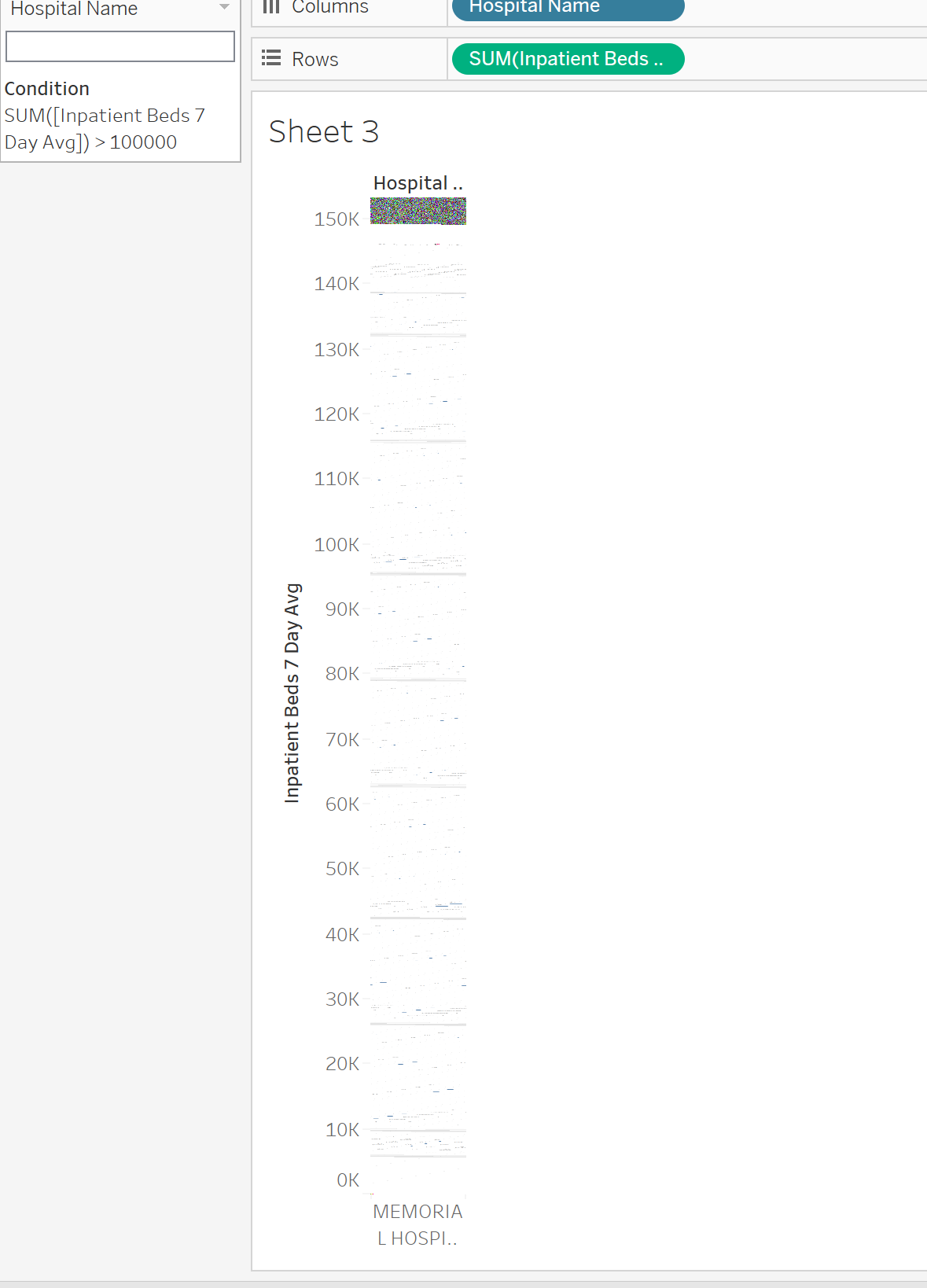


Figure 16

H. TX with 985.2. It is displayed in Figure 16.

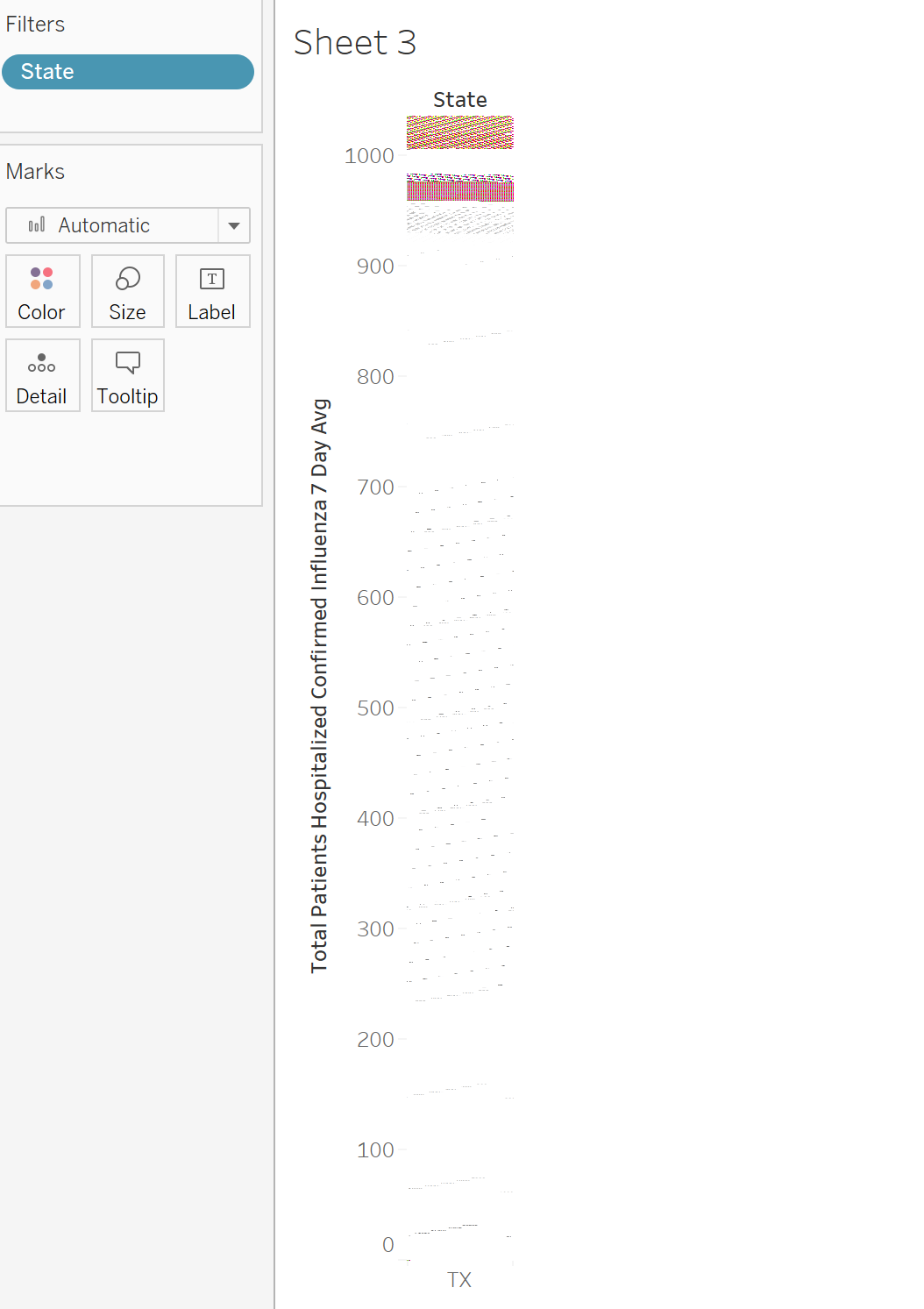


Figure 16