Deniz Aytemiz

ECE 5984:SS: Applied Machine Learning - Homework 1

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

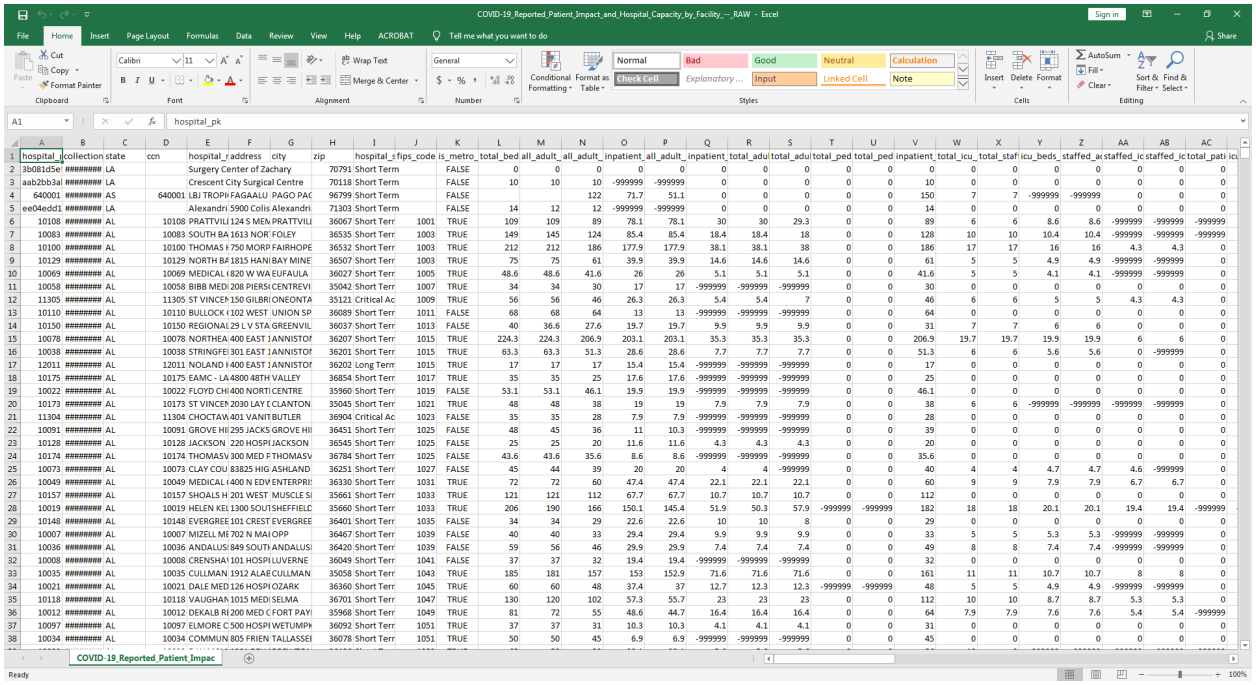


Figure 1

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.

1	hospital_pk	isnumber?
2	3b081d5ef1c552538e4af4aa593a857bb9	FALSE
3	aab2bb3ab769da90baf57242c96ec481af	FALSE
5	ee04edd185865c38c839812cb2eb5ae5d	FALSE
197	04T027	FALSE
268	04L120	FALSE
529	02d4a73785e74f88490f05c473b39d9054	FALSE
525	718456ab442d5f84683cd24a06376fcd58	FALSE
769	09005E	FALSE
963	10T290	FALSE
098	11L006	FALSE
844	764e4176df9f3b110d69f34074f855a1514	FALSE
709	28TA03	FALSE
469	a4a8d5d1a12d46859f37b88f911e03bb81	FALSE
521	719c8a84-a386-4aca-8db1-a36ddc9e050	FALSE
641	39L113	FALSE
022	44L009	FALSE
874	52C000	FALSE
003	ee04edd185865c38c839812cb2eb5ae5d	FALSE
004	aab2bb3ab769da90baf57242c96ec481af	FALSE
006	3b081d5ef1c552538e4af4aa593a857bb9	FALSE

Figure 2

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

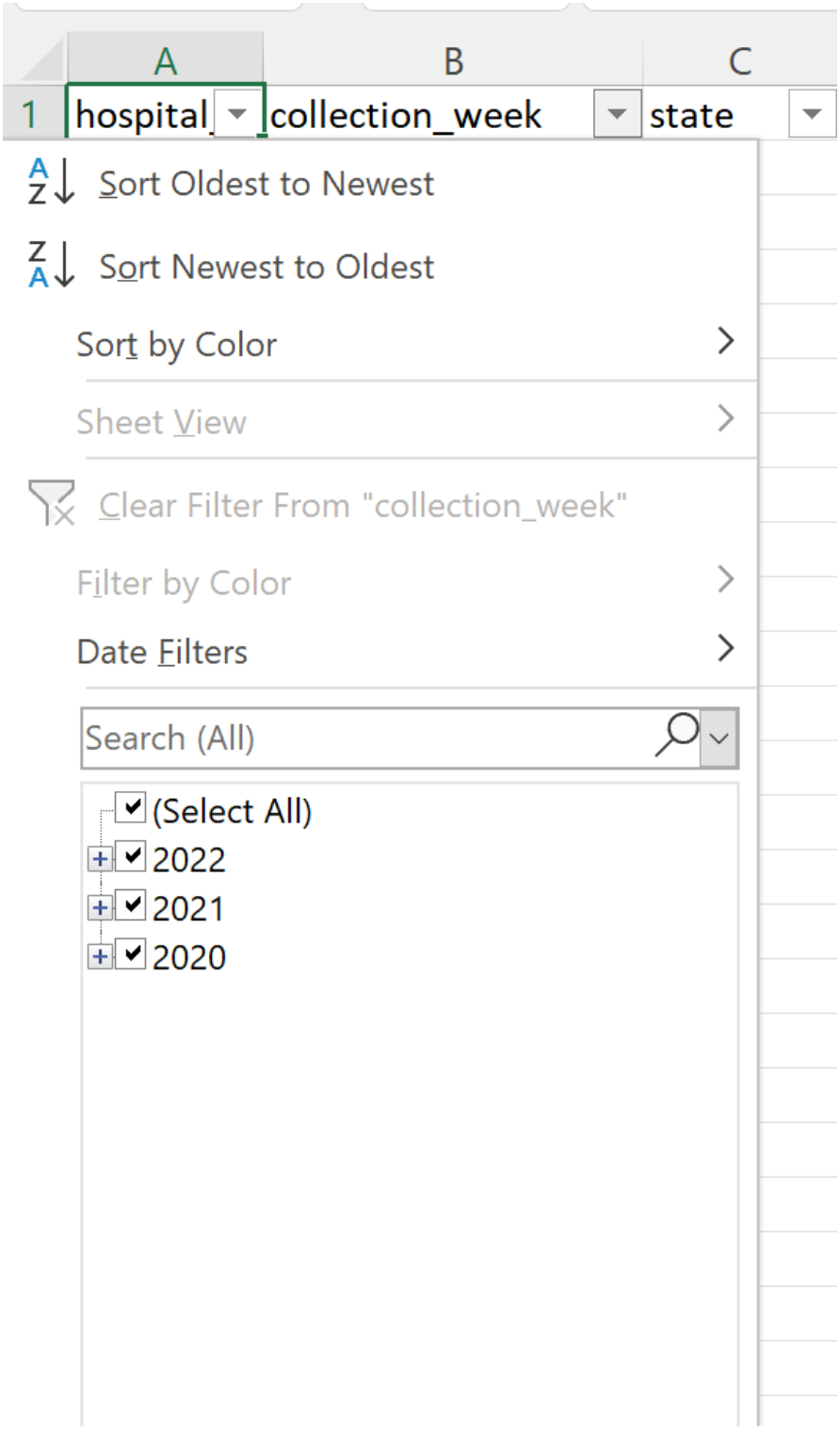
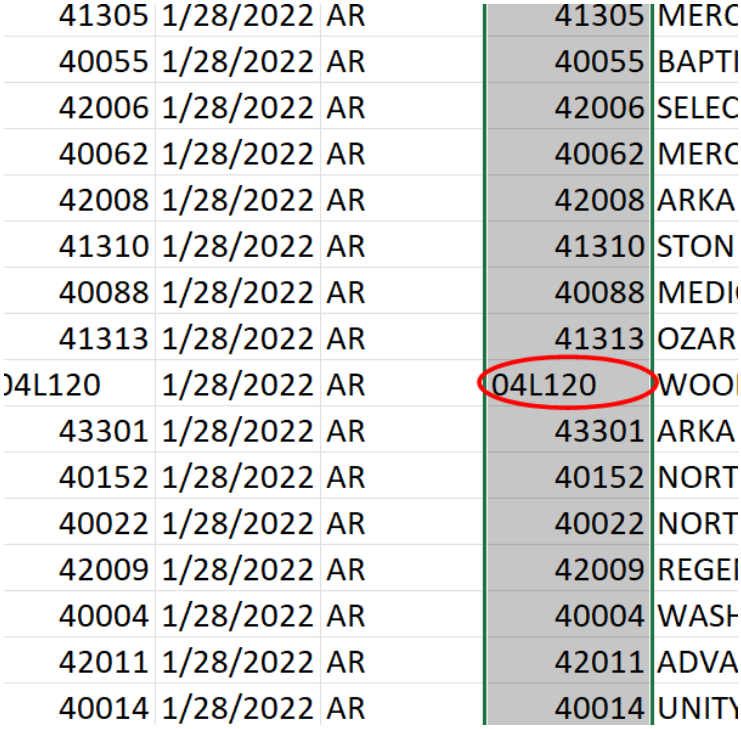


Figure 3

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.



41305	1/28/2022	AR	41305	MERC
40055	1/28/2022	AR	40055	BAPTI
42006	1/28/2022	AR	42006	SELEC
40062	1/28/2022	AR	40062	MERC
42008	1/28/2022	AR	42008	ARKA
41310	1/28/2022	AR	41310	STON
40088	1/28/2022	AR	40088	MEDI
41313	1/28/2022	AR	41313	OZAR
)4L120	1/28/2022	AR	04L120	WOO
43301	1/28/2022	AR	43301	ARKA
40152	1/28/2022	AR	40152	NORT
40022	1/28/2022	AR	40022	NORT
42009	1/28/2022	AR	42009	REGEI
40004	1/28/2022	AR	40004	WASH
42011	1/28/2022	AR	42011	ADVA
40014	1/28/2022	AR	40014	UNITY

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.

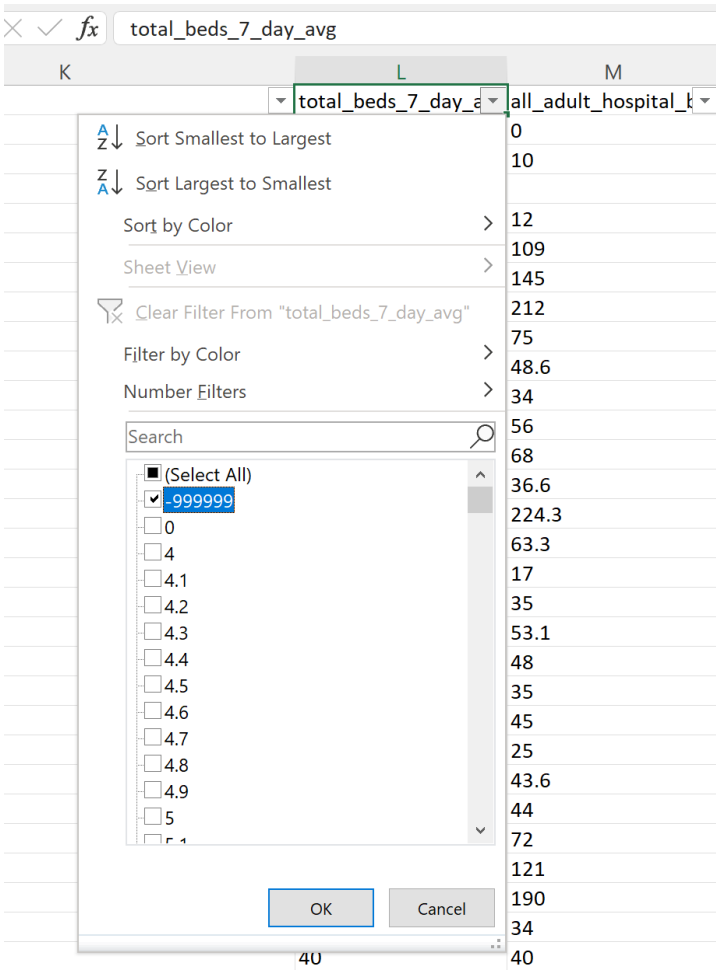

total_beds_7_day_avg
K
L
M
total_beds_7_day_a
all_adult_hospital_b
Sort Smallest to Largest
Sort Largest to Smallest
Sort by Color
Sheet View
Clear Filter From "total_beds_7_day_avg"
Filter by Color
Number Filters
Search
(Select All)
-999999
0
4
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
4.9
5
OK
Cancel
0
10
12
109
145
212
75
48.6
34
56
68
36.6
224.3
63.3
17
35
53.1
48
35
45
25
43.6
44
72
121
190
34
40
40

Figure 5

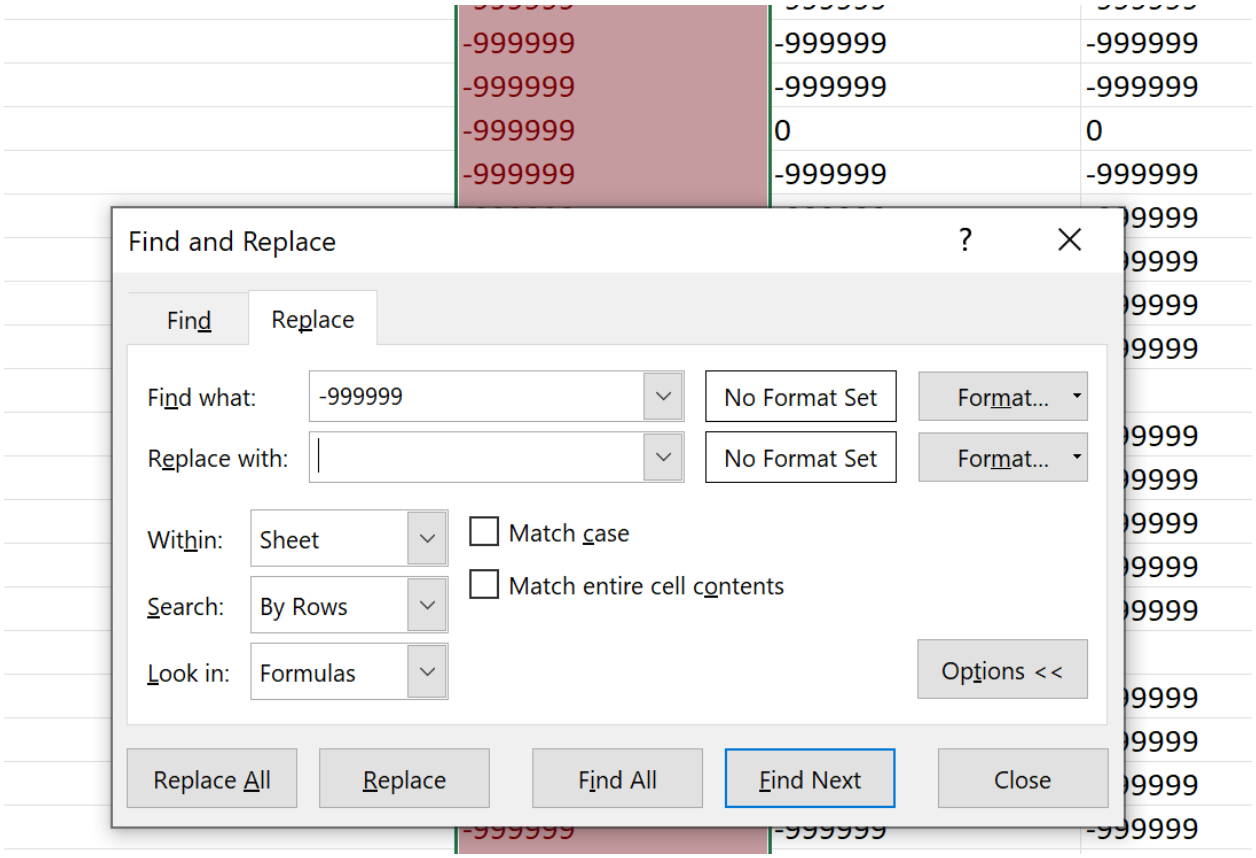


Figure 6

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.

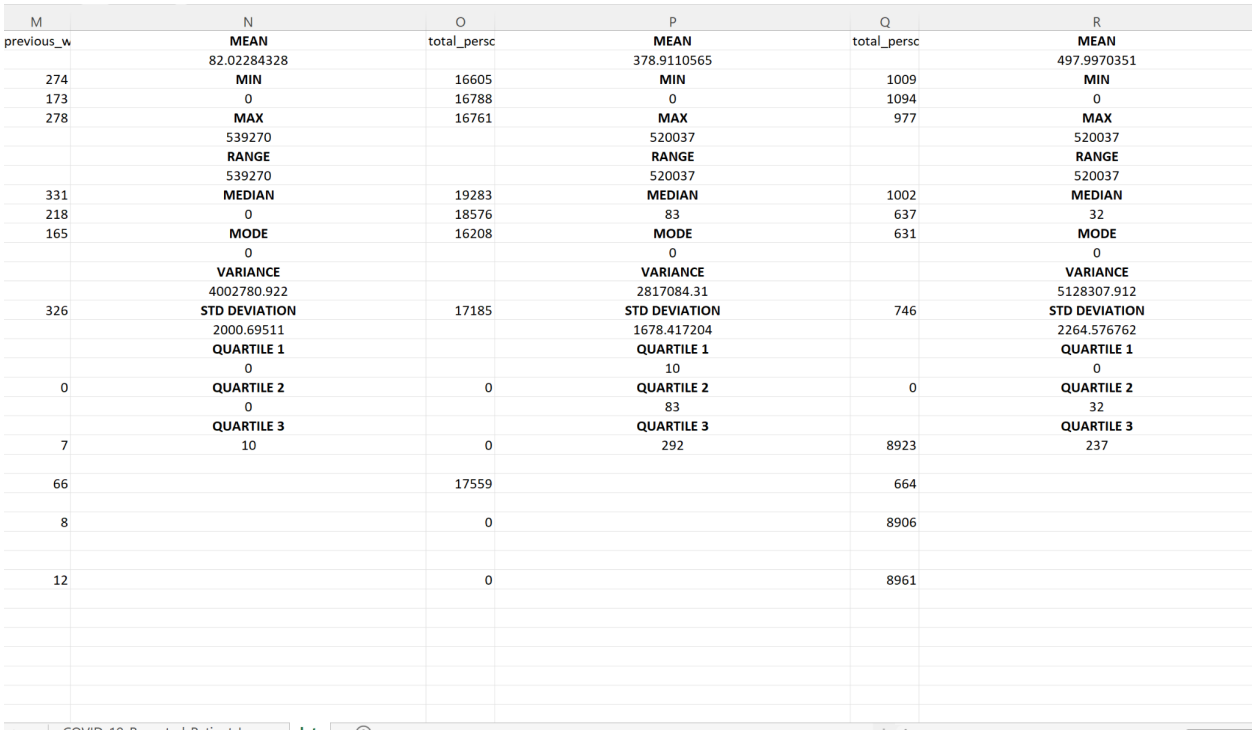
B27 fx

	A	B	C	D	E	F
1	hospital_pk	**MEAN**	collection_	**MEAN**	state	**MEAN**
2	f1c552538e4af4aa593a857bb922a4f364a412e69f912f82	267615.9428	1/28/2022	44316.64363	LA	#DIV/0!
3	b769da90baf57242c96ec481afb5ec6a2337848975e24519	**MIN**	1/28/2022	**MIN**	LA	**MIN**
4	640001	10001	1/28/2022	44043	AS	0
5	85865c38c839812cb2eb5ae5d3f8922e3b629ee98c7d9424	**MAX**	1/28/2022	**MAX**	LA	**MAX**
6	10108	677297	1/28/2022	44589	AL	0
7	10083	**RANGE**	1/28/2022	**RANGE**	AL	**RANGE**
8	10100	667296	1/28/2022	546	AL	0
9	10129	**MEDIAN**	1/28/2022	**MEDIAN**	AL	**MEDIAN**
10	10069	260009	1/28/2022	44316	AL	#NUM!
11	10058	**MODE**	1/28/2022	**MODE**	AL	**MODE**
12	11305	10108	1/28/2022	44190	AL	#N/A
13	10110	**VARIANCE**	1/28/2022	**VARIANCE**	AL	**VARIANCE**
14	10150	24513913053	1/28/2022	25375.84255	AL	#DIV/0!
15	10078	**STD DEVIATION**	1/28/2022	**STD DEVIATION**	AL	**STD DEVIATION**
16	10038	156569.1957	1/28/2022	159.2979678	AL	#DIV/0!
17	12011	**QUARTILE 1**	1/28/2022	**QUARTILE 1**	AL	**QUARTILE 1**
18	10175	140172	1/28/2022	44176	AL	#NUM!
19	10022	**QUARTILE 2**	1/28/2022	**QUARTILE 2**	AL	**QUARTILE 2**
20	10173	260009	1/28/2022	44316	AL	#NUM!
21	11304	**QUARTILE 3**	1/28/2022	**QUARTILE 3**	AL	**QUARTILE 3**
22	10091	390265	1/28/2022	44456	AL	#NUM!
23	10128		1/28/2022		AL	
24	10174		1/28/2022		AL	
25	10073		1/28/2022		AL	
26	10049		1/28/2022		AL	
27	10157		1/28/2022		AL	

Figure 7

G	H	I	J	K	L
ccn	**MEAN**	hospital_name	**MEAN**	address	**MEAN**
	267572.3225	Surgery Center of Zachary	#DIV/0!		42071
	MIN	Crescent City Surgical Centr	**MIN**		**MIN**
640001	10001	LBJ TROPICAL MEDICAL CEN	0	FAGAALU V	42071
	MAX	Alexandria Emergency Hosp	**MAX**	5900 Colise	**MAX**
10108	677297	PRATTVILLE BAPTIST HOSPI	0	124 S MEM	42071
10083	**RANGE**	SOUTH BALDWIN REGIONA	**RANGE**	1613 NORT	**RANGE**
10100	667296	THOMAS HOSPITAL	0	750 MORPI	0
10129	**MEDIAN**	NORTH BALDWIN INFIRMA	**MEDIAN**	1815 HAND	**MEDIAN**
10069	260009	MEDICAL CENTER BARBOUI	#NUM!	820 W WA	42071
10058	**MODE**	BIBB MEDICAL CENTER	**MODE**	208 PIERSC	**MODE**
11305	50515	ST VINCENTS BLOUNT	#N/A	150 GILBRE	42071
10110	**VARIANCE**	BULLOCK COUNTY HOSPITA	**VARIANCE**	102 WEST	**VARIANCE**
10150	24518455791	REGIONAL MEDICAL CENTE	#DIV/0!	29 L V STAE	0
10078	**STD DEVIATION**	NORTHEAST ALABAMA REG	**STD DEVIATION**	400 EAST 1	**STD DEVIATION**
10038	156583.7022	STRINGFELLOW CAMPUS O	#DIV/0!	301 EAST 1	0
12011	**QUARTILE 1**	NOLAND HOSPITAL ANNIST	**QUARTILE 1**	400 EAST 1	**QUARTILE 1**
10175	140167	EAMC - LANIER	#NUM!	4800 48TH	42071
10022	**QUARTILE 2**	FLOYD CHEROKEE MEDICAL	**QUARTILE 2**	400 NORTH	**QUARTILE 2**
10173	260009	ST VINCENT'S CHILTON	#NUM!	2030 LAY D	42071
11304	**QUARTILE 3**	CHOCTAW GENERAL HOSPI	**QUARTILE 3**	401 VANITY	**QUARTILE 3**
10091	390265	GROVE HILL MEMORIAL HO	#NUM!	295 JACKSO	42071
10128		JACKSON MEDICAL CENTER		220 HOSPITAL DRIVE	
10174		THOMASVILLE REGIONAL MEDICAL CENTER		300 MED PARK DRIVE	
10073		CLAY COUNTY HOSPITAL		83825 HIGHWAY 9 P O BOX 1270	
10049		MEDICAL CENTER ENTERPRISE		400 N EDWARDS STREET	
10157		SHOALS HOSPITAL		201 WEST AVALON AVENUE	
10019		HELEN KELLER MEMORIAL HOSPITAL		1300 SOUTH MONTGOMERY AVENUE	
10148		EVERGREEN MEDICAL CENTER		101 CRESTVIEW AVENUE	
10007		MIZELL MEMORIAL HOSPITAL		702 N MAIN ST	
10036		ANDALUSIA HEALTH		849 SOUTH THREE NOTCH STREET	
10008		CRENSHAW COMMUNITY HOSPITAL		101 HOSPITAL CIRCLE	
10021		DALE MEDICAL CENTER		126 HOSPITAL AVE	
10118		VAUGHAN REGIONAL MEDICAL CENTER PARKWAY CAMPUS		1015 MEDICAL CENTER PARKWAY	

Figure 8



M	N	O	P	Q	R
previous_w	**MEAN**	total_persc	**MEAN**	total_persc	**MEAN**
	82.02284328		378.9110565		497.9970351
274	**MIN**	16605	**MIN**	1009	**MIN**
173	0	16788	0	1094	0
278	**MAX**	16761	**MAX**	977	**MAX**
	539270		520037		520037
	RANGE		**RANGE**		**RANGE**
	539270		520037		520037
331	**MEDIAN**	19283	**MEDIAN**	1002	**MEDIAN**
218	0	18576	83	637	32
165	**MODE**	16208	**MODE**	631	**MODE**
	0		0		0
	VARIANCE		**VARIANCE**		**VARIANCE**
	4002780.922		2817084.31		5128307.912
326	**STD DEVIATION**	17185	**STD DEVIATION**	746	**STD DEVIATION**
	2000.69511		1678.417204		2264.576762
	QUARTILE 1		**QUARTILE 1**		**QUARTILE 1**
	0		10		0
0	**QUARTILE 2**	0	**QUARTILE 2**	0	**QUARTILE 2**
	0		83		32
	QUARTILE 3		**QUARTILE 3**		**QUARTILE 3**
7	10	0	292	8923	237
66		17559		664	
8		0		8906	
12		0		8961	

Figure 9

S	T	U	V	W	X
total_persc	**MEAN**	previous_w	**MEAN**	previous_w	**MEAN**
	795.841965		169.3544784		164.5017061
14068	**MIN**	221	**MIN**	346	**MIN**
13796	0	294	0	255	0
13711	**MAX**	462	**MAX**	149	**MAX**
	520037		520037		520037
	RANGE		**RANGE**		**RANGE**
	520037		520037		520037
15487	**MEDIAN**	216	**MEDIAN**	326	**MEDIAN**
25253	163	5	0	10	0
25404	**MODE**	13	**MODE**	13	**MODE**
	0		0		0
	VARIANCE		**VARIANCE**		**VARIANCE**
	5956082.821		4366297.79		3655142.359
13673	**STD DEVIATION**	513	**STD DEVIATION**	80	**STD DEVIATION**
	2440.508722		2089.568805		1911.842661
	QUARTILE 1		**QUARTILE 1**		**QUARTILE 1**
	32		0		0
0	**QUARTILE 2**	0	**QUARTILE 2**	0	**QUARTILE 2**
	163		0		0
	QUARTILE 3		**QUARTILE 3**		**QUARTILE 3**
8909	650	0	6	0	5
25449		7		20	
8892		0		0	
8947		0		0	

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

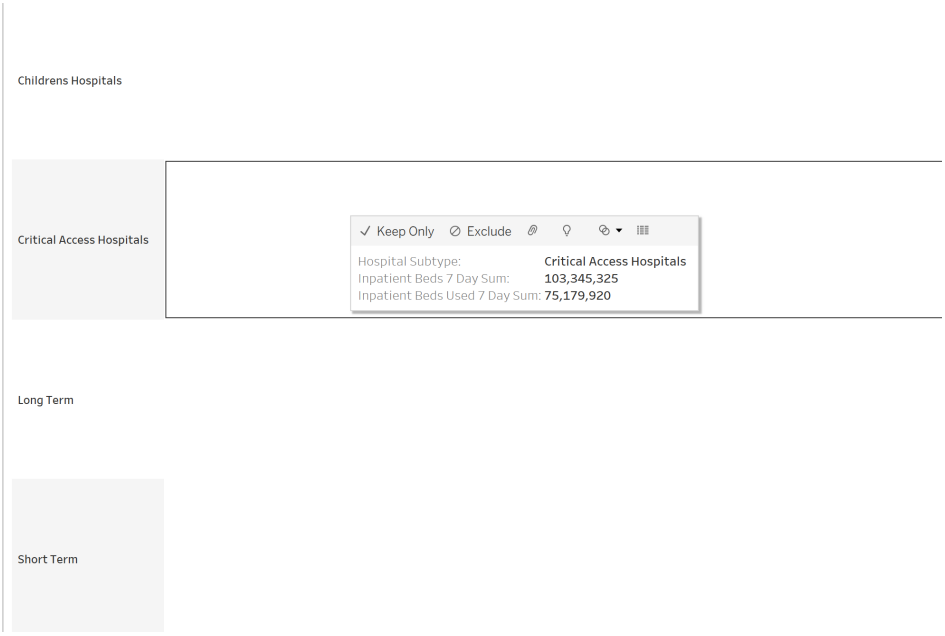


Figure 12

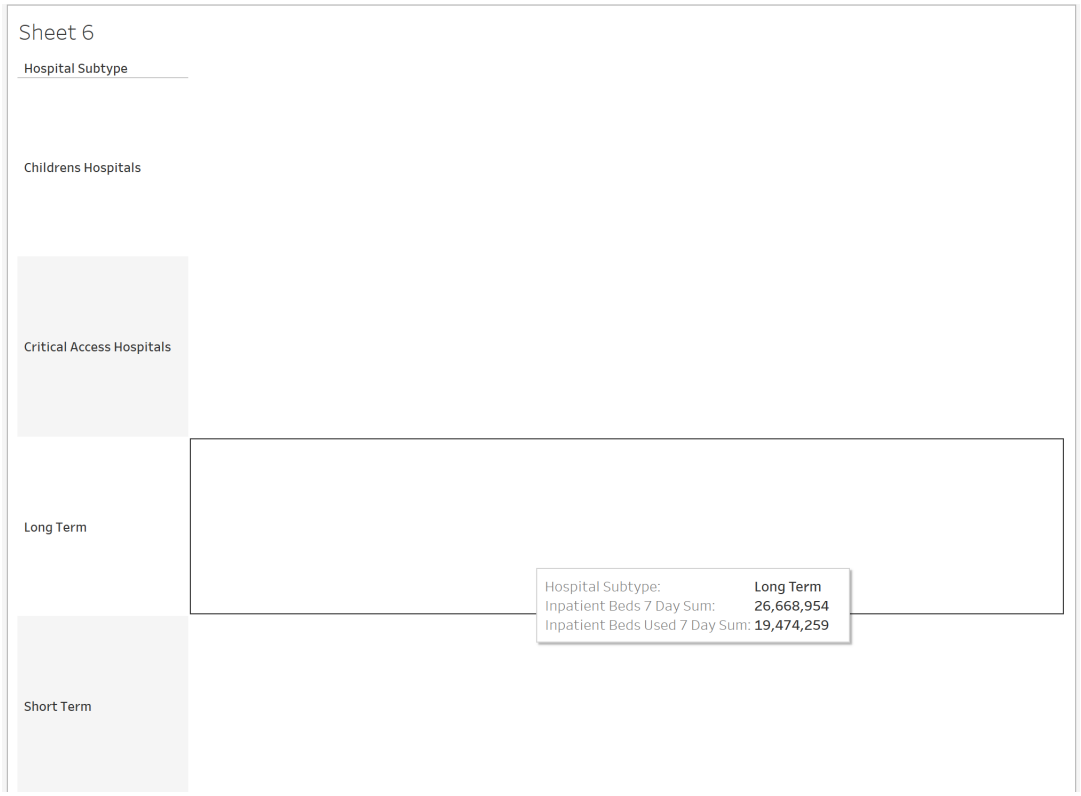


Figure 13

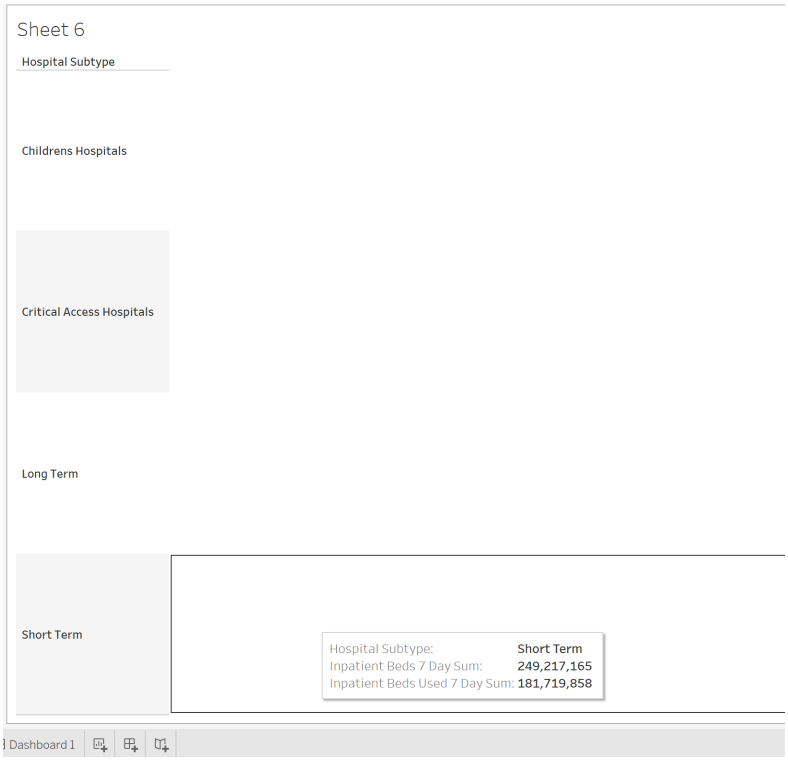


Figure 14

B. Histogram

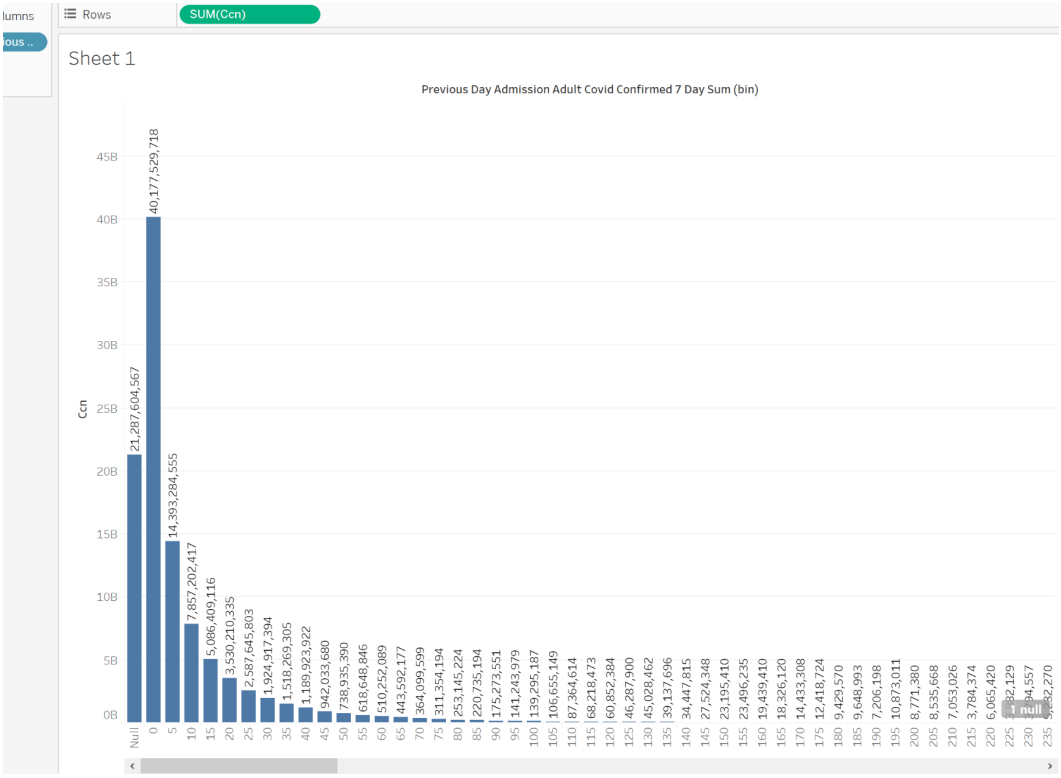


Figure 15

C. Line Graph

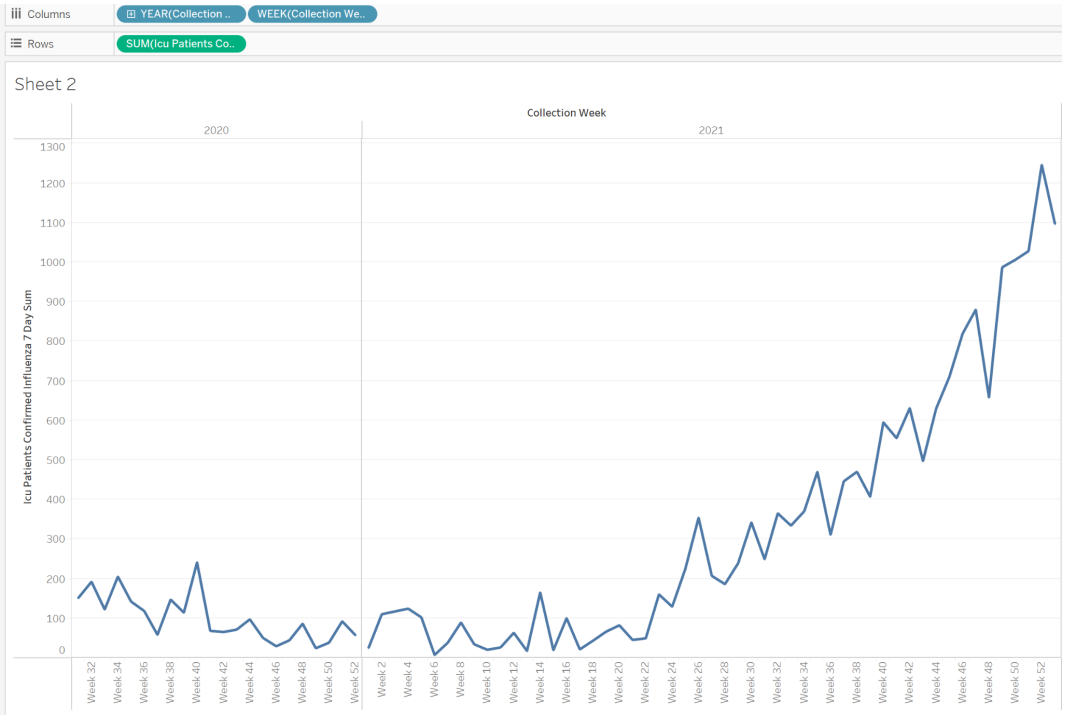


Figure 16

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

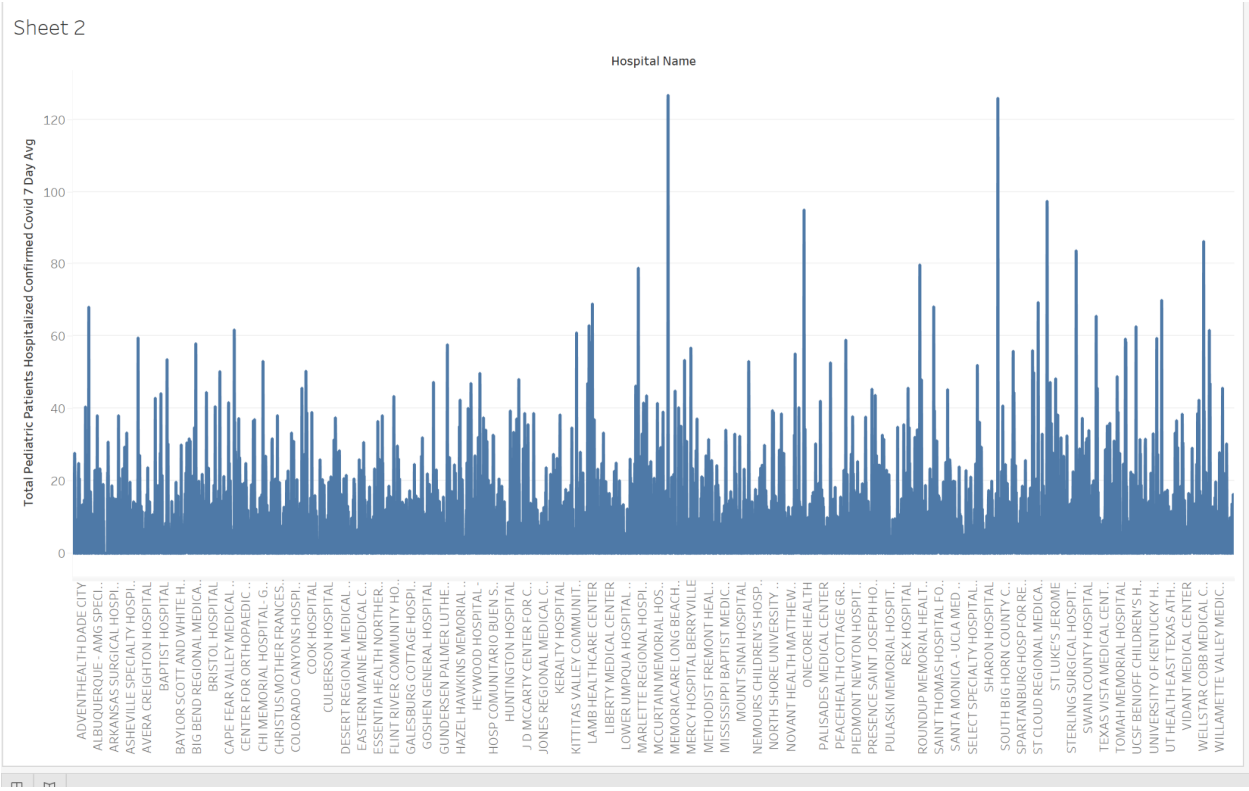


Figure 17

E.
CA with 3138
MI with 1992
FL with 1865
TX with 1699
IL with 1553

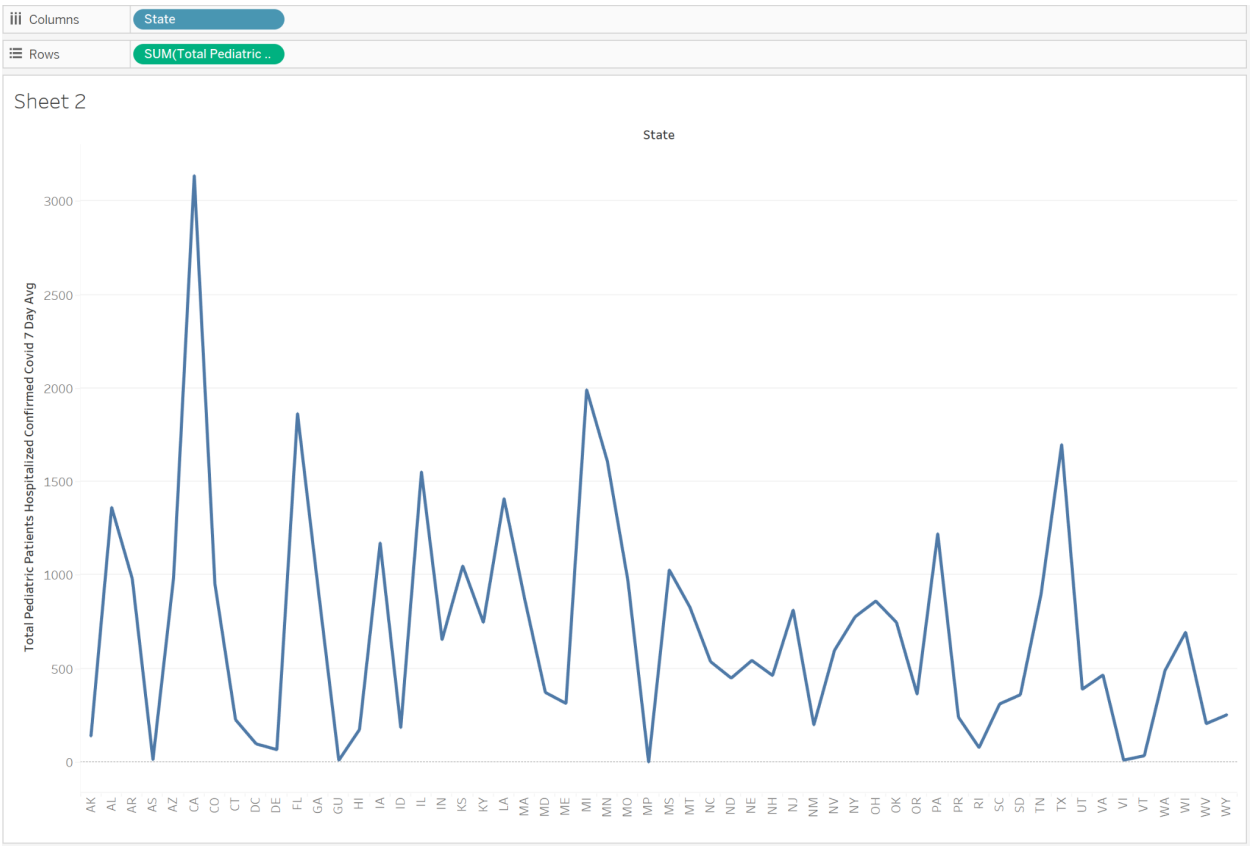

Columns
State
Rows
SUM(Total Pediatric ..
Sheet 2
State
Total Pediatric Patients Hospitalized Confirmed Covid 7 Day Avg
3000
2500
2000
1500
1000
500
0
AK AL AR AS AZ CA CO CT DC DE FL GA GU HI IA ID IL IN KS KY LA MA MD ME MI MN MO MP MS MT NC ND NE NH NJ NM NV NY OH OK OR PA PR RI SC SD TN TX UT VA VI VT WA WI WV WY

Figure 18

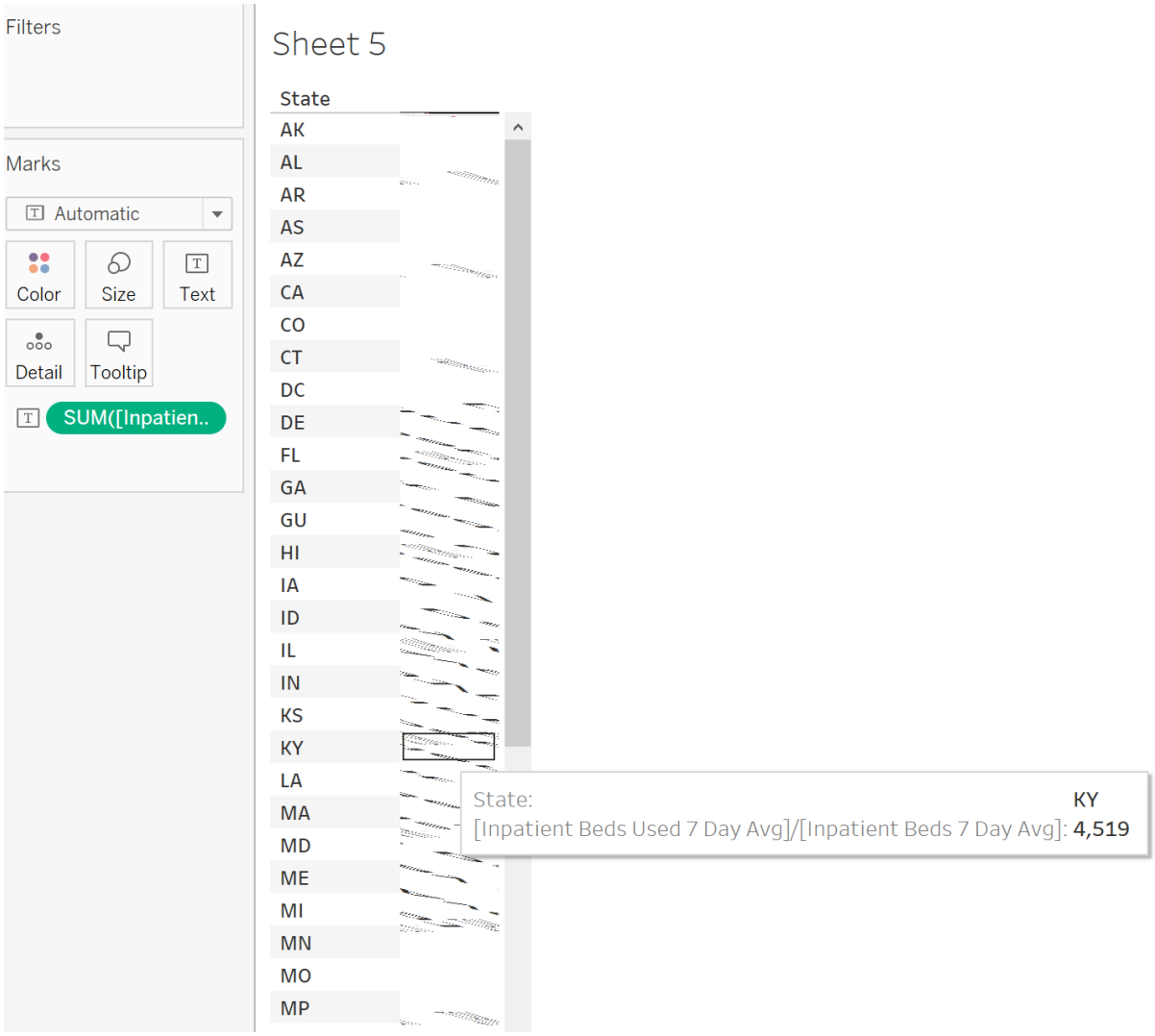


Figure 15

G.

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

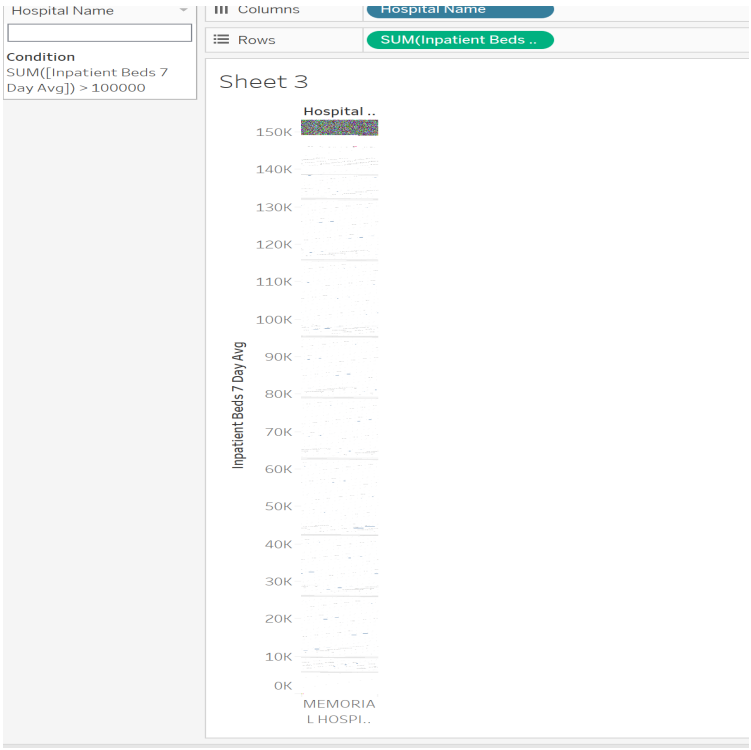


Figure 16

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

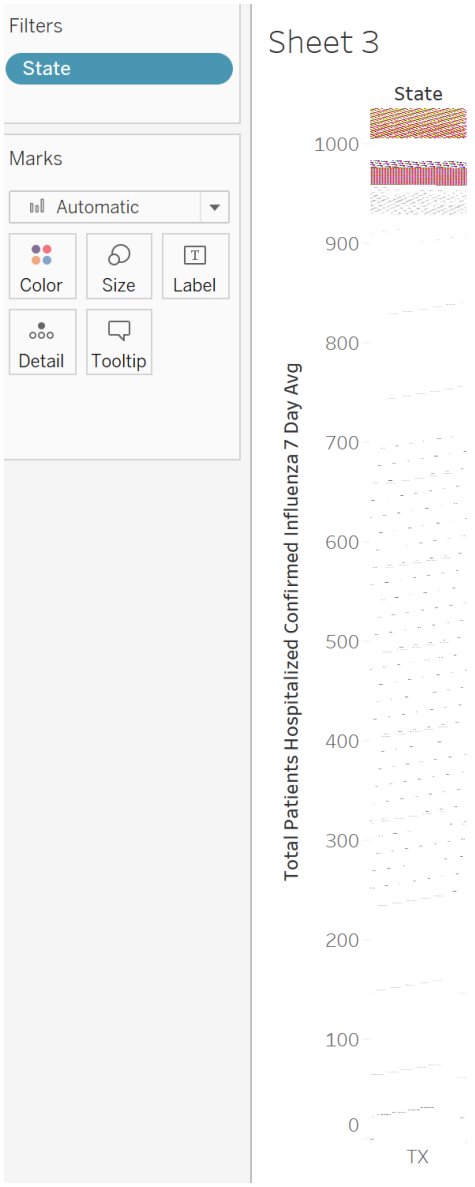

Filters
State
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Sheet 3
State
1000
900
800
700
600
500
400
300
200
100
0
Total Patients Hospitalized Confirmed Influenza 7 Day Avg
TX

Figure 16