Finding the Best County in California State to Fight the Pandemic

Jun 1, 2021

BACKGROUND:

COVID-19 is an infectious disease caused by a newly discovered coronavirus. Most people infected with the COVID-19 virus will experience mild to moderate respiratory illness and recover without requiring special treatment. Older people, and those with underlying medical problems like cardiovascular disease, diabetes, chronic respiratory disease, and cancer are more likely to develop serious illness. The COVID-19 virus spreads primarily through droplets of saliva or discharge from the nose when an infected person coughs or sneezes.

Therefore, it is advantageous to know which counties in a particular state (in our case CALIFORNIA state) are well equipped with hospitals & beds. This can be done by calculating the bed per person ratio. For example, this analysis can be useful to improve those counties whose bed per person ratio is less.

PROBLEM:

The pandemic situation has turned so worse that as of June 1, 2021, the state with highest number of COVID-19 cases in the United States was California. And over 3.5 million cases have been reported across the States of California.

My main motive behind this is to create some useful insights on this situation. And in this project, we will determine which county is best prepared for this pandemic, by finding out the **best ratio of hospital beds per person** and **ICU beds per person** for each county in California. We will also cluster the counties based on the above ratios.

DATA ACQUISITION and CLEANING:

We will be collecting data from the following sources:

Data Sources:

1. California State Dataset with Counties, Number of Hospital Beds and ICU Beds.
   - source: CA data set
2. California Counties Population data
   - source: worldpopulationreview.com
3. Hospitals Information (Names, Coordinates)
   - source: **FOURSQUARE API**
4. Counties coordinates
   - source: simplemaps
   - Alternative source: **FOURSQUARE API** (county key: 5345731ebcbc57f1066c39b2)

Data Cleaning:

Data downloaded or leveraged from multiple sources were combined into one table. And we sort the table alphabetically based on County names. **Alpine & Sierra counties** are excluded based on low population. So only **56 counties** are considered out of 58.

California hospital beds dataset contained data dated from **march 29, 2020**, I've cleaned most of the data and considered data only dated as of **June 1, 2021.** After merging different tables, we include two columns namely **Bed_per_100_people** and **ICU_Bed_per_100_people** and it resulted in the following final data frame:

```
3 cal_df.head()
```

county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
Alameda	2493.0	110.0	1680480	37.6469	-121.8889	0.148350	0.006546
Amador	53.0	0.0	40446	38.4464	-120.6511	0.131039	0.000000
Butte	451.0	7.0	196880	39.6669	-121.6007	0.229074	0.003555
Calaveras	33.0	8.0	46319	38.2046	-120.5541	0.071245	0.017272
Colusa	48.0	5.0	21805	39.1775	-122.2370	0.220133	0.022931

METHODOLOGY:

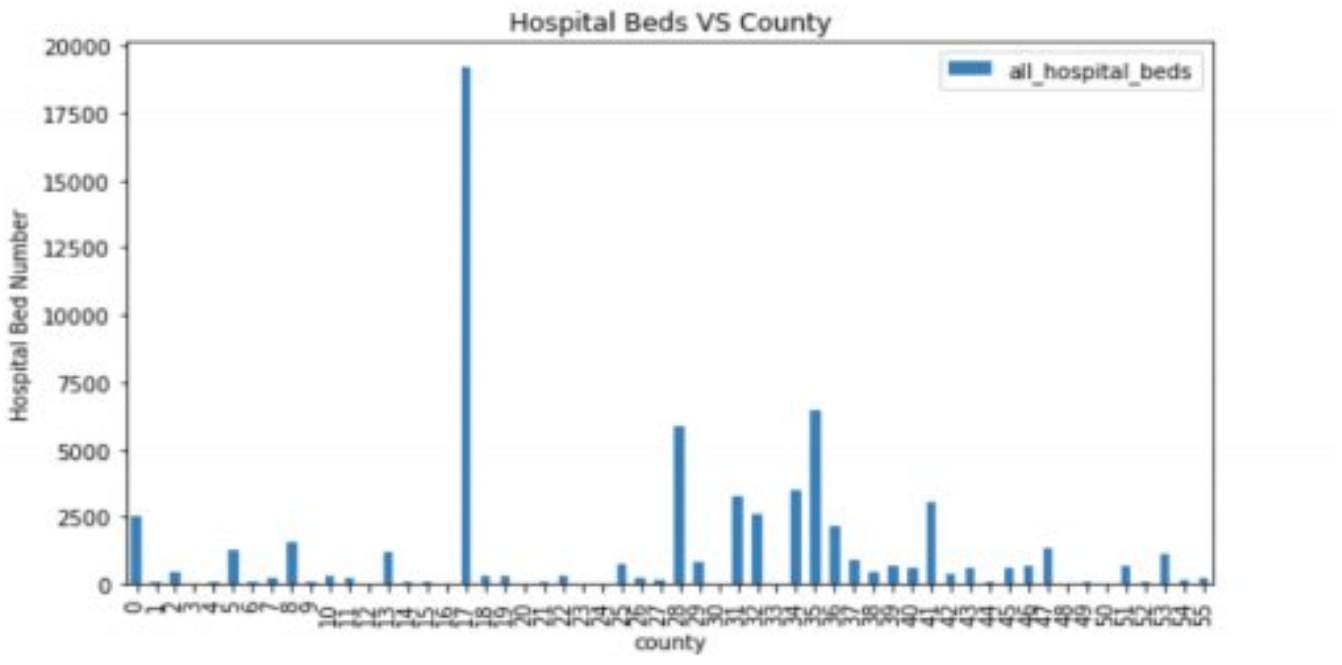
★ **Step 1: California hospitals and beds data from** [CA data set]

county	todays_date	hospitalized_covid_confirmed_patients	hospitalized_suspected_covid_patients	hospitalized_covid_patients	all_hospital_beds	icu_covid
Lassen	2020-03-29	0.0	2.0	NaN	NaN	
Yolo	2020-03-29	2.0	3.0	NaN	NaN	
San Francisco	2020-03-29	50.0	73.0	NaN	NaN	
Los Angeles	2020-03-29	489.0	1132.0	NaN	NaN	
San Diego	2020-03-29	121.0	211.0	NaN	NaN	

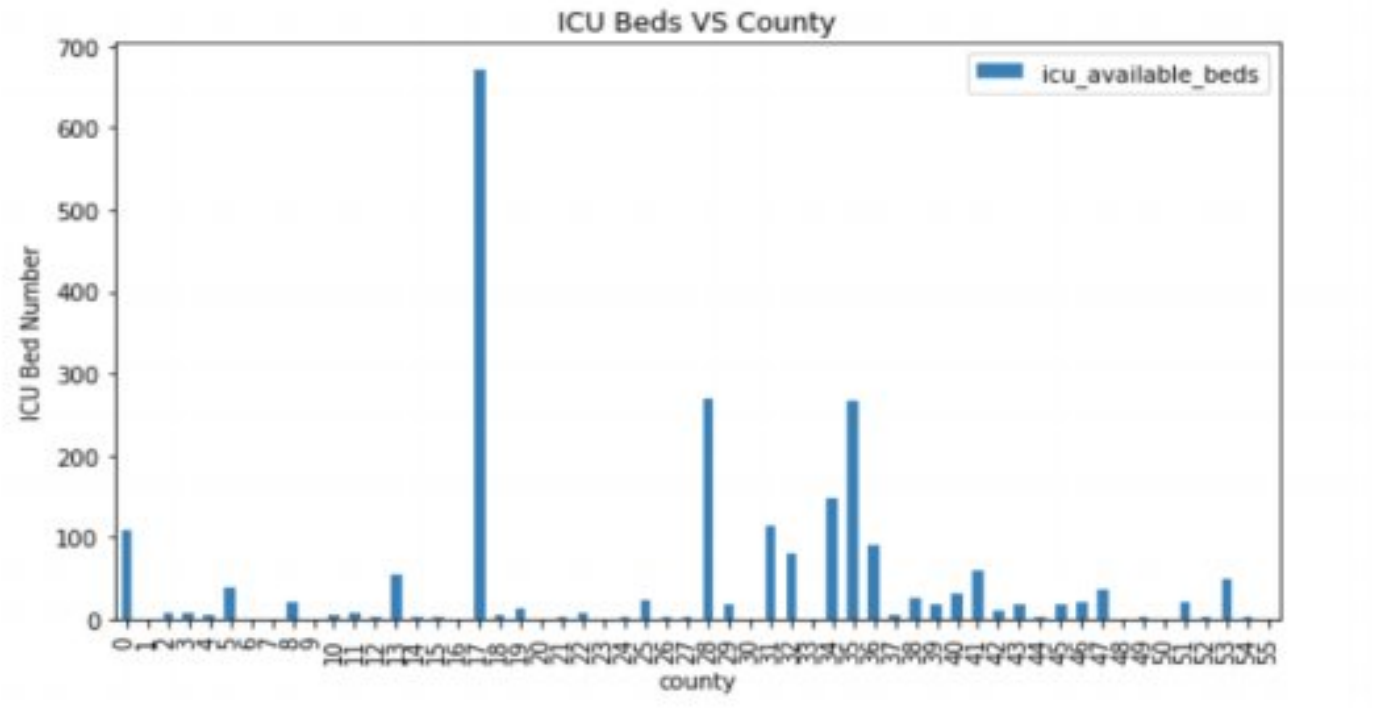
As we can see the data is dated from 2020-03-29 to 2021-01-06. By using the tail method, we get the data as of June 1, 2021. Removing columns related to patients, and sorting data frame alphabetically based on county names, the resulting data frame (**total 56 counties**) looks like the following:

	county	all_hospital_beds	icu_available_beds
0	Alameda	2493.0	110.0
1	Amador	53.0	0.0
2	Butte	451.0	7.0
3	Calaveras	33.0	8.0
4	Colusa	48.0	5.0

★ **Step 2:** Plotting bar charts for **Hospital Beds VS County** and **ICU Beds VS County**



Here the **county with max number of hospital beds** is present in county **index-17** that is **Los Angeles County**



Again, **Los Angeles** with county index-17 has the **max number of ICU Beds**

★ Step 3: Finding County with 0 ICU Beds

```
# counties with 0 icu_available_beds
no_beds_df = cal_df[cal_df['icu_available_beds']==0]
no_beds_df
```

	county	all_hospital_beds	icu_available_beds
1	Amador	53.0	0.0
6	Del Norte	53.0	0.0
7	El Dorado	190.0	0.0
9	Glenn	47.0	0.0
16	Lassen	25.0	0.0
20	Mariposa	14.0	0.0
23	Modoc	12.0	0.0
30	Plumas	35.0	0.0
33	San Benito	25.0	0.0
48	Sutter	14.0	0.0
50	Trinity	25.0	0.0

In total the above **11** counties have **0** ICU available Beds

So, the above 11 counties don't have ICU Beds.

★ Step 4: Collecting and Cleaning Population Data

Collecting the population data from the worldpopulationreview.com and reading it into a pandas data frame. It looks like the following:

```
pop_df = pd.read_csv('cal_pop_data.csv')
for county in pop_df['CTYNAME']:
    pop_df.replace(to_replace=county,value=county.rstrip('County'),inplace=True)

pop_df.sort_values('CTYNAME',inplace=True)
pop_df.reset_index(drop=True,inplace=True)
pop_df.head()
```

	CTYNAME	pop2021	GrowthRate	popDensity
0	Alameda	1680480	11.0701	2273.7070
1	Alpine	1209	4.1344	1.6358
2	Amador	40446	6.7571	54.7239
3	Butte	196880	-10.4883	266.3807
4	Calaveras	46319	1.8716	62.6701

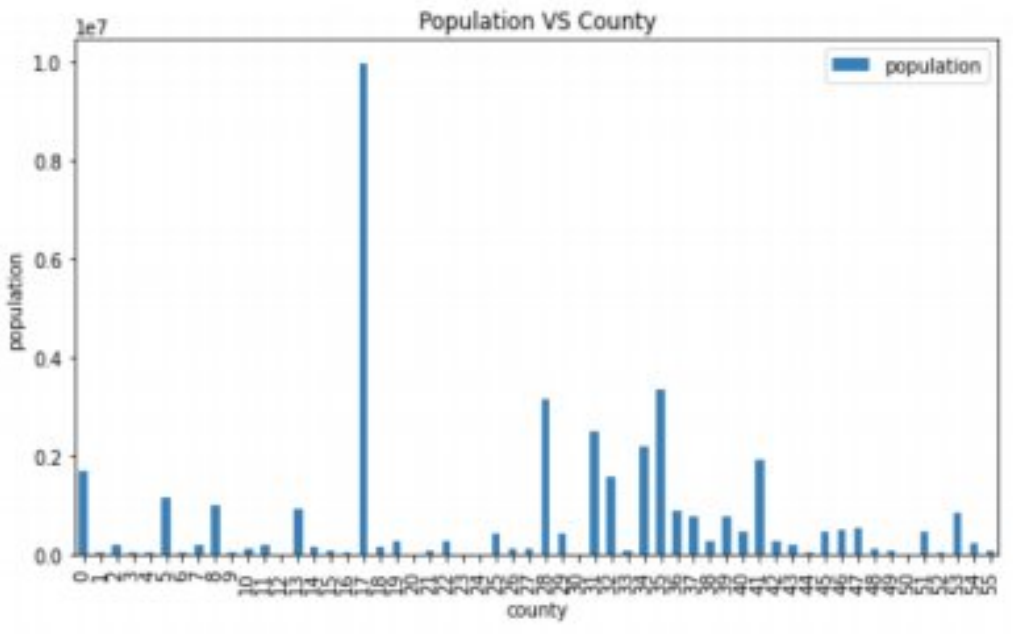
Remove **Alpine (index-1)** and **Sierra (index-45)** counties as discussed earlier

```
pop_df.drop([1,45],axis=0,inplace=True)
pop_df.drop(columns=['GrowthRate','popDensity'],inplace=True)
pop_df.reset_index(drop=True,inplace=True)
```

```
pop_df.rename(columns={'pop2021':'population'},inplace=True)
pop_df.head()
```

	CTYNAME	population
0	Alameda	1680480
1	Amador	40446
2	Butte	196880
3	Calaveras	46319
4	Colusa	21805

★ Step 5: Plotting Population VS County



Los Angeles (index-17) is the most populated county in California State and is the reason for having more hospital and ICU Beds

★ Step 6: Collecting and Cleaning Counties Coordinates data set

Collecting the Counties Coordinates data from simplemaps website and reading it into a data frame. It results in the following data frame:

```
us_counties_df = pd.read_csv('uscounties.csv')
us_counties_df.head()
```

	county	county_ascii	county_fips	state_id	state_name	lat	lng	population
0	Los Angeles	Los Angeles	6037	CA	California	34.3207	-118.2248	10081570
1	Cook	Cook	17031	IL	Illinois	41.8401	-87.8168	5198275
2	Harris	Harris	48201	TX	Texas	29.8577	-95.3936	4646630
3	Maricopa	Maricopa	4013	AZ	Arizona	33.3490	-112.4915	4328810
4	San Diego	San Diego	6073	CA	California	33.0341	-116.7353	3316073

The above df contains all the counties present in the United States Country. We need only those in the California State and clean the df by keeping only **county, lat** and **lng** columns, it results in the following df:

	county	lat	lng
0	Alameda	37.6469	-121.8889
1	Amador	38.4464	-120.6511
2	Butte	39.6669	-121.6007
3	Calaveras	38.2046	-120.5541
4	Colusa	39.1775	-122.2370

★ Step 7: Merging all the df's to get the final df

We will be merging all the till known df's (cal_df, pop_df, us_counties_df) and add 2 more columns **Bed_per_100_people** & **ICU_Bed_per_100_people.** The resulting df looks like the following:

Final DataFrame

Including **Bed per 100 people** and **ICU Beds per 100 people**, we get our final dataframe

```
cal_df['Bed_per_100_people'] = (cal_df['all_hospital_beds']/cal_df['population'])*100
cal_df['ICU_Bed_per_100_people'] = (cal_df['icu_available_beds']/cal_df['population'])*100
cal_df.head()
```

	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
0	Alameda	2493.0	110.0	1680480	37.6469	-121.8889	0.148350	0.006546
1	Amador	53.0	0.0	40446	38.4464	-120.6511	0.131039	0.000000
2	Butte	451.0	7.0	196880	39.6669	-121.6007	0.229074	0.003555
3	Calaveras	33.0	8.0	46319	38.2046	-120.5541	0.071245	0.017272
4	Colusa	48.0	5.0	21805	39.1775	-122.2370	0.220133	0.022931

★ Step 8: Finding counties with the best bed_per_person ratio's

Finding the counties with best beds per 100 people ratio

```
cal_df['Bed_per_100_people'].max()
```

```
0.3226604030164421
```

```
cal_df[cal_df['Bed_per_100_people']==0.3226604030164421]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
55	5	Yuba	261.0	1.0	80890	39.269	-121.3513	0.32266	0.001236

So **Yuba County** with **0.32266** bpp ratio has the best beds per 100 people ratio

```
cal_df['ICU_Bed_per_100_people'].max()
```

```
0.022930520522815866
```

```
cal_df[cal_df['ICU_Bed_per_100_people']==0.022930520522815866]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
4	0	Colusa	48.0	5.0	21805	39.1775	-122.237	0.220133	0.022931

And **Colusa County** with **0.0293** ibp ratio has the best ICU beds per 100 people

From the above image, we can see that **Yuba County** has the **best Bed_per_100_people ratio (0.32266)** and **Colusa County** has the **best ICU_Bed_per_100_people ratio (0.02293).**

★ Step 9: Leveraging and Cleaning Data from FOURSQUARE API

By passing the hospital key (4bf58dd8d48988d196941735) in the URL, we get the hospital's names and coordinates data as a JSON file from FOURSQUARE's API. Normalizing and leveraging the required data, we create a pandas data frame namely hospitals_df and use this data to map the locations of the hospitals on a folium California map. Actually, there isn't much need of this data, as collecting data from FOURSQUARE API is an important criterion, we will be using this data to map the hospital's location. Instead of collecting the counties coordinates data from simplemaps we can get the similar data from FOURSQUARE API using the county key (5345731ebcbc57f1066c39b2). The hospital_df looks like the following:

```
1 hospital_df.head()
```

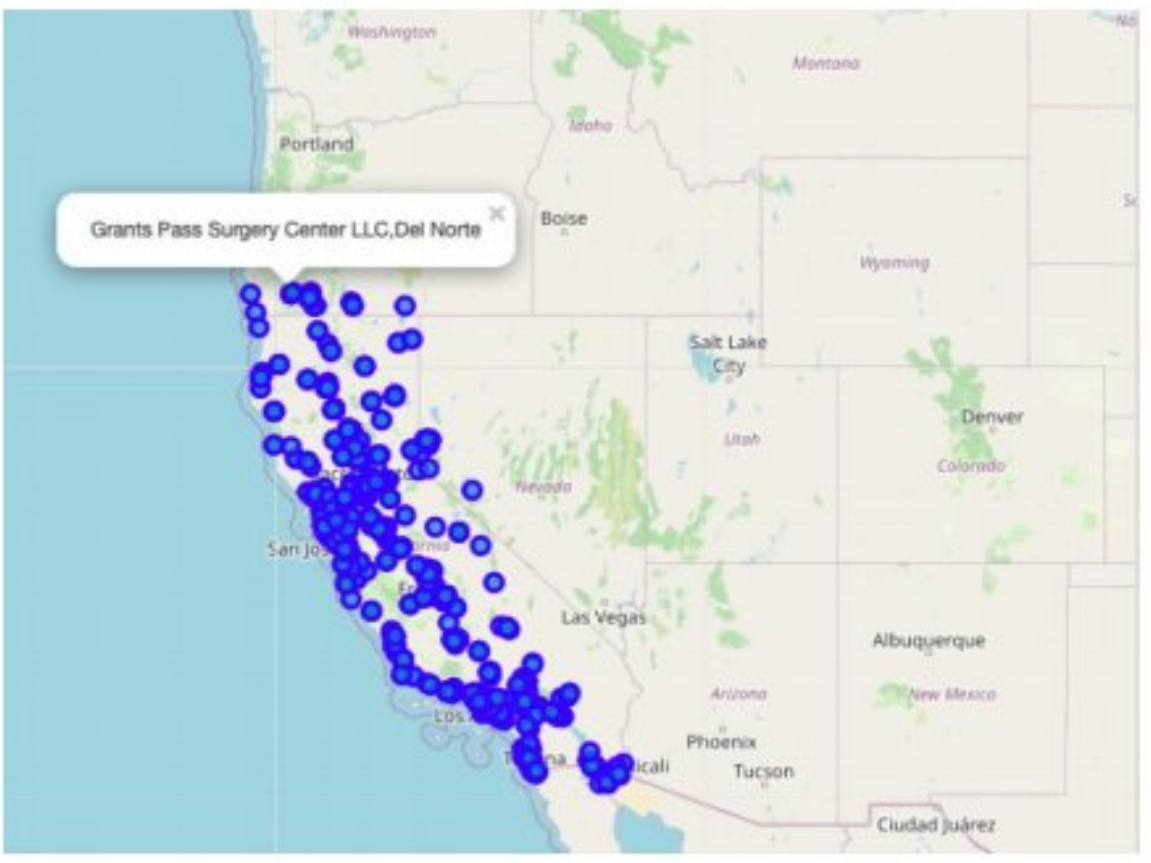
	ID	Name	Latitude	Longitude	county
0	5d3d9ca51e4a070007882bc2	Zuckerberg San Francisco General Hospital and ...	37.755659	-122.404956	Alameda
1	4a73f4d8f964a520a2dd1fe3	Palo Alto Medical Foundation	37.548328	-121.973723	Alameda
2	4a1dc9f8f964a520967b1fe3	Lucile Packard Children's Hospital (LPCH)	37.435998	-122.175331	Alameda
3	4a8f5a59f964a520091520e3	El Camino Hospital	37.369134	-122.079735	Alameda
4	52e694c611d265590dffd4e9	One Medical	37.773986	-122.422218	Alameda

After cleaning by removing ID column, it results in:

	Name	Latitude	Longitude	county
0	Zuckerberg San Francisco General Hospital and ...	37.755659	-122.404956	Alameda
1	Palo Alto Medical Foundation	37.548328	-121.973723	Alameda
2	Lucile Packard Children's Hospital (LPCH)	37.435998	-122.175331	Alameda
3	El Camino Hospital	37.369134	-122.079735	Alameda
4	One Medical	37.773986	-122.422218	Alameda

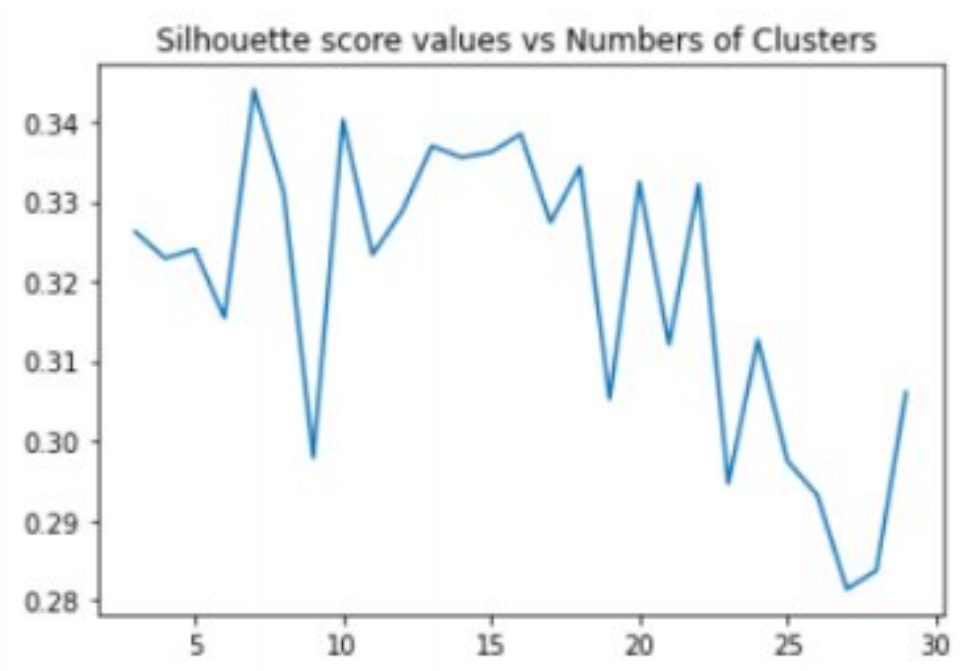
In total I managed to fetch 2245 hospitals from FOURSQUARE API.

Now plotting the data on map:



★ Step 10: Clustering by K-MEANS

We are going to cluster the data based on **population**, **Bed_per_100_people** and **ICU_Bed_per_100_people** values. The data is first normalized using the **StandardScaler.** We will be using the **silhouette score** to find the optimum number of clusters; it is a good indication that the underlying model fits best at that point. In the visualizer, value of k turned out to be 7.




```
Optimal number of components is:  7
```

The score suggests us to have **7** clusters

Merging the cluster labels into the cal_df. The data frame looks like this:

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
0	4	Alameda	2493.0	110.0	1680480	37.6469	-121.8889	0.148350	0.006546
1	1	Amador	53.0	0.0	40446	38.4464	-120.6511	0.131039	0.000000
2	5	Butte	451.0	7.0	196880	39.6669	-121.6007	0.229074	0.003555
3	0	Calaveras	33.0	8.0	46319	38.2046	-120.5541	0.071245	0.017272
4	0	Colusa	48.0	5.0	21805	39.1775	-122.2370	0.220133	0.022931

★ Step 11: See Which County goes to Which Cluster

Let's see which county goes to which cluster

Dataset for **cluster 1**:

Cluster 1

```
1 cal_df[cal_df['Cluster_labels']==0]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
3	0	Calaveras	33.0	8.0	46319	38.2046	-120.5541	0.071245	0.017272
4	0	Colusa	48.0	5.0	21805	39.1775	-122.2370	0.220133	0.022931
12	0	Inyo	29.0	2.0	18225	36.5111	-117.4107	0.159122	0.010974
24	0	Mono	17.0	2.0	14526	37.9391	-118.8868	0.117032	0.013768

Dataset for **cluster 2**:

Cluster 2

```
1 cal_df[cal_df['Cluster_labels']==1]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
1	1	Amador	53.0	0.0	40446	38.4464	-120.6511	0.131039	0.000000
6	1	Del Norte	53.0	0.0	27956	41.7431	-123.8972	0.189584	0.000000
8	1	Fresno	1554.0	22.0	1013400	36.7582	-119.6493	0.153345	0.002171
9	1	Glenn	47.0	0.0	29245	39.5982	-122.3920	0.160711	0.000000
10	1	Humboldt	274.0	6.0	134186	40.6993	-123.8756	0.204194	0.004471
18	1	Madera	279.0	5.0	158217	37.2180	-119.7627	0.176340	0.003160
23	1	Modoc	12.0	0.0	8923	41.5898	-120.7250	0.134484	0.000000
26	1	Napa	206.0	3.0	135654	38.5065	-122.3305	0.151857	0.002212
29	1	Placer	799.0	18.0	410327	39.0635	-120.7175	0.194723	0.004387
30	1	Plumas	35.0	0.0	18939	40.0046	-120.8385	0.184804	0.000000
50	1	Trinity	25.0	0.0	11721	40.6507	-123.1126	0.213292	0.000000

Dataset for **Cluster 3**:

Cluster 3

```
1 cal_df[cal_df['Cluster_labels']==2]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
17	2	Los Angeles	19186.0	670.0	9969510	34.3207	-118.2248	0.192447	0.00672

Dataset for **Cluster 4**:

Cluster 4

```
1 cal_df[cal_df['Cluster_labels']==3]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
7	3	El Dorado	190.0	0.0	197037	38.7787	-120.5247	0.096429	0.000000
14	3	Kings	98.0	3.0	156056	36.0753	-119.8155	0.062798	0.001922
16	3	Lassen	25.0	0.0	30483	40.6736	-120.5943	0.082013	0.000000
20	3	Mariposa	14.0	0.0	16799	37.5815	-119.9054	0.083338	0.000000
21	3	Mendocino	93.0	2.0	85445	39.4402	-123.3915	0.108842	0.002341
22	3	Merced	271.0	7.0	284738	37.1919	-120.7177	0.095175	0.002458
33	3	San Benito	25.0	0.0	65490	36.6057	-121.0750	0.038174	0.000000
37	3	San Joaquin	889.0	5.0	781462	37.9348	-121.2714	0.113761	0.000640
39	3	San Mateo	704.0	19.0	762357	37.4229	-122.3290	0.092345	0.002492
48	3	Sutter	14.0	0.0	98217	39.0346	-121.6948	0.014254	0.000000
49	3	Tehama	59.0	2.0	67216	40.1256	-122.2341	0.087777	0.002975
54	3	Yolo	128.0	2.0	221264	38.6866	-121.9016	0.057849	0.000904

Dataset for **Cluster 5**:

Cluster 5

```
1 cal_df[cal_df['Cluster_labels']==4]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
0	4	Alameda	2493.0	110.0	1680480	37.6469	-121.8889	0.148350	0.006546
28	4	Orange	5828.0	270.0	3175130	33.7030	-117.7611	0.183552	0.008504
31	4	Riverside	3256.0	114.0	2520060	33.7437	-115.9938	0.129203	0.004524
32	4	Sacramento	2584.0	81.0	1578680	38.4493	-121.3443	0.163681	0.005131
34	4	San Bernardino	3507.0	147.0	2206750	34.8414	-116.1784	0.158921	0.006661
35	4	San Diego	6486.0	267.0	3347270	33.0341	-116.7353	0.193770	0.007977
41	4	Santa Clara	3057.0	60.0	1918880	37.2318	-121.6951	0.159312	0.003127

Dataset for **Cluster 6**:

Cluster 6

```
1 cal_df[cal_df['Cluster_labels']==5]
```

	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
2	5	Butte	451.0	7.0	196880	39.6669	-121.6007	0.229074	0.003555
36	5	San Francisco	2162.0	90.0	883255	37.7562	-122.4430	0.244776	0.010190
43	5	Shasta	567.0	17.0	180822	40.7637	-122.0405	0.313568	0.009402
47	5	Stanislaus	1305.0	36.0	555728	37.5591	-120.9977	0.234827	0.006478
55	5	Yuba	261.0	1.0	80890	39.2690	-121.3513	0.322660	0.001236

Dataset for **Cluster 7**:

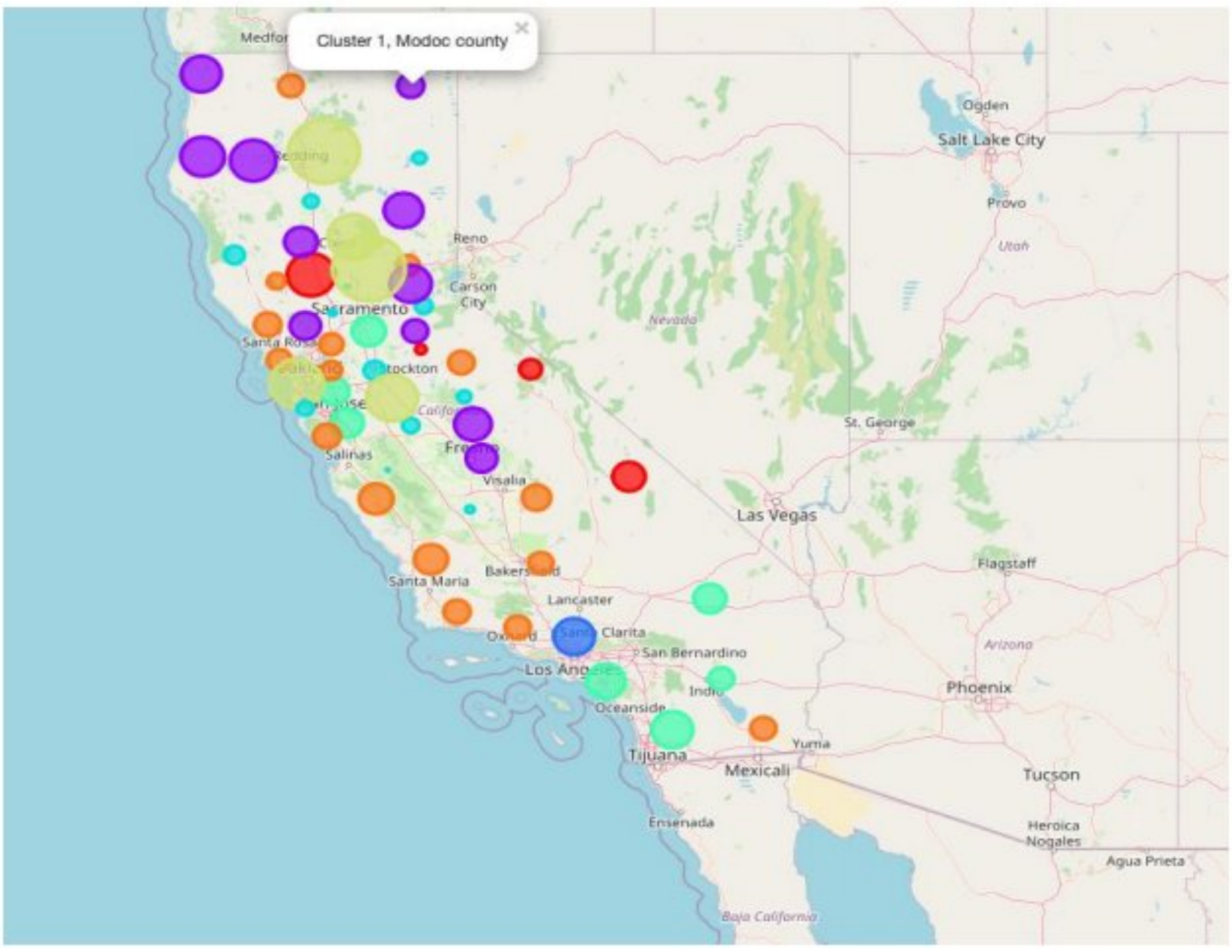
Cluster 7

```
1 cal_df[cal_df['Cluster_labels']==6]
```

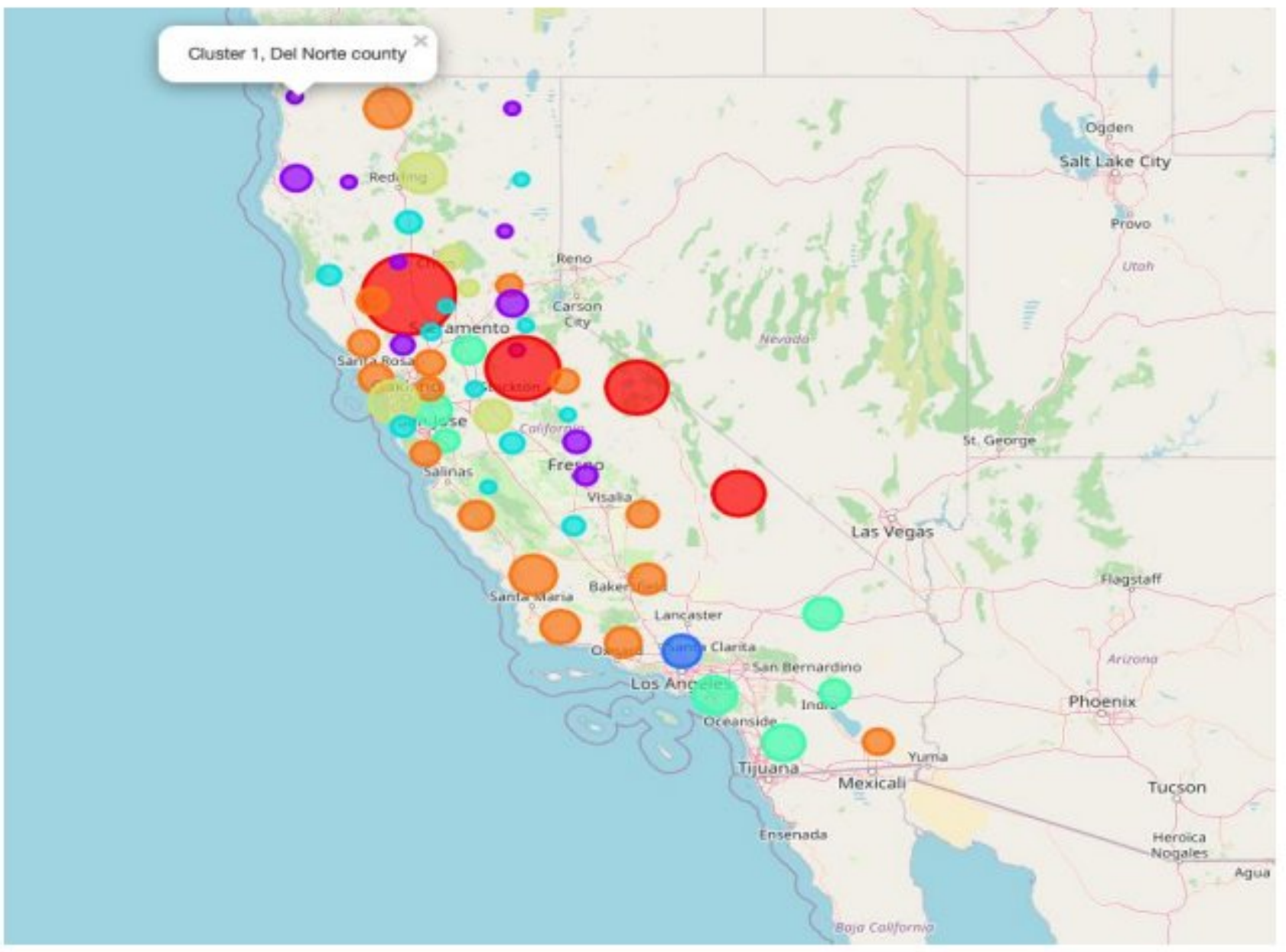
	Cluster_labels	county	all_hospital_beds	icu_available_beds	population	lat	lng	Bed_per_100_people	ICU_Bed_per_100_people
5	6	Contra Costa	1293.0	39.0	1159540	37.9191	-121.9278	0.111510	0.003363
11	6	Imperial	236.0	8.0	180599	33.0395	-115.3654	0.130676	0.004430
13	6	Kern	1163.0	55.0	913090	35.3429	-118.7299	0.127370	0.006024
15	6	Lake	64.0	3.0	64524	39.0996	-122.7532	0.099188	0.004649
19	6	Marin	320.0	14.0	257154	38.0734	-122.7234	0.124439	0.005444
25	6	Monterey	716.0	24.0	434283	36.2172	-121.2392	0.164869	0.005526
27	6	Nevada	121.0	3.0	100249	39.3014	-120.7685	0.120699	0.002993
38	6	San Luis Obispo	460.0	26.0	282625	35.3871	-120.4045	0.162760	0.009199
40	6	Santa Barbara	603.0	31.0	447937	34.6729	-120.0165	0.134617	0.006921
42	6	Santa Cruz	372.0	11.0	271957	37.0562	-122.0018	0.136786	0.004045
44	6	Siskiyou	56.0	4.0	43517	41.5927	-122.5404	0.128685	0.009192
45	6	Solano	560.0	19.0	451479	38.2700	-121.9329	0.124037	0.004208
46	6	Sonoma	655.0	22.0	485722	38.5283	-122.8874	0.134851	0.004529
51	6	Tulare	673.0	22.0	469407	36.2201	-118.8005	0.143372	0.004687
52	6	Tuolumne	72.0	2.0	54660	38.0276	-119.9548	0.131723	0.003659
53	6	Ventura	1106.0	50.0	841734	34.4565	-119.0836	0.131395	0.005940

VISUALIZING FOLIUM MAPS

The first map illustrates the clusters where the radius of the circle marker is proportional to hospital beds per 100 people



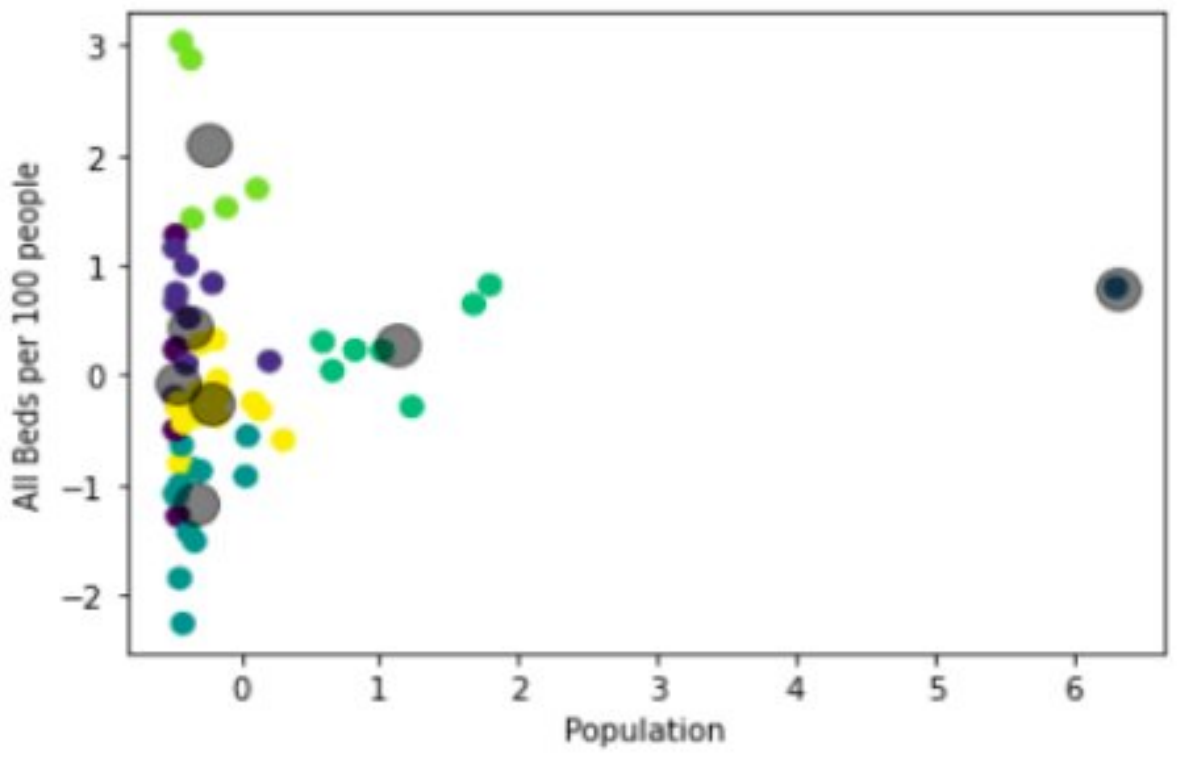
The second map illustrates the clusters where radius of the circle marker is proportional to ICU beds per 100 people

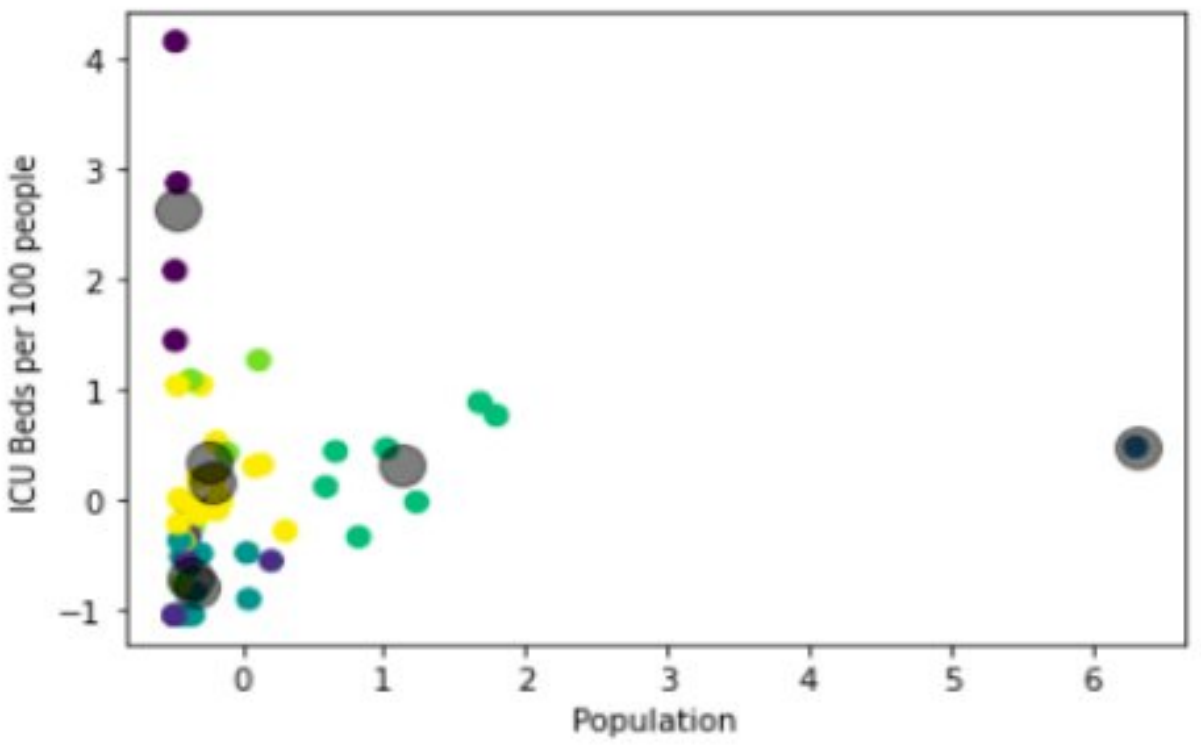


We can see the **only 1 county** with **steel blue circle marker** which is **Los Angeles County** is a cluster.

SCATTER PLOTS:

Let's look at the scatter plots of our data and define the clusters with different colors.





We can observe the outliers here. In the first plot, we see the **top green circle outlier** which is **Yuba County** with **best bed per 100 people ratio** and the next green marker below Yuba County is **Shasta County** which has the **next best bed per 100 people ratio.** The other **black circle outlier** in first scatter plot is **Los Angeles County** because of its **high population.**

And coming to the second scatter plot the **purple circle outlier** is **Colusa County** which has the **best ICU bed per 100 people ratio** and the **black circle outlier** is **Los Angeles County** due to its **high population.**

RESULTS AND DISCUSSION:

During the analysis, a total of **7 clusters** were defined. **Single cluster** (cluster 3), which is **Los Angeles County**, is an **outlier** because of its **huge population** compared to others. **Yuba County** and **Colusa County** have the **best bed per 100 people** and **ICU bed per 100 people ratios respectively**. And they too are **also outlier**s in our scatter plots for the above reason. There are **11 counties** with **no ICU beds** and they are from **clusters 2 and 4**, which means they need to concentrate on providing more emergency treatment (like ICU Beds) and also 10 out of 11 falls in the category with the least beds per 100 people (the exception being El Dorado County with 190 beds). **Los Angeles** is also an outlier because of its high population but has only **0.192447 bed per 100 people** and **0.00672 ICU bed per 100 people**, which means they need to improve a lot to provide more beds to fight the pandemic.

CONCLUSION:

Finally, to conclude, basic exploratory data analysis was performed to identify the well-equipped county in California State. During the analysis, important features like **counties**, their **population** and **total beds available** and **ICU beds available** as of **June 1, 2021** were considered. And were clustered based on the above features. **Yuba County** has the best bed per 100 people ratio and **Colusa County** has the best ICU bed per 100 people ratio. **Los Angeles**, an outlier due to its high population has lot of scope for providing more beds.

REFERENCES:

- California Health and Human Services Open Data Portal
- us counties
- simplemaps
- FOURSQUARE API
- project idea (Special mention for giving me a similar idea about this project)