FINAL Project –The Battle of the Neighborhoods

for Applied Data Science Capstone Course

Karl D Huber
April 2020

INTRODUCTION

This project strives to **identify the Southern California neighborhoods that are at high risk for severe impact from COVID-19**, the disease caused by the novel coronavirus (SARS-cov-2).

Over 2.5 million cases of COVID-19 have been reported worldwide (as of mid-April 2020) with over 190,000 deaths. This is the biggest global health event in the 21st century and the spread of the coronavirus and its potential impact upon specific communities has become a paramount concern to cities and governments everywhere, including Southern California (SoCal).

Two key factors that have been cited as being correlated to the likelihood of contracting COVID-19 are *population density and availability of hospital beds*. We will look at these factors and use data science techniques to combine relevant data with geographical locations to indicate on a **SoCal map those areas that the data show have highest risk for COVID-19**.

This map and the data behind it will allow civic and other leaders in Southern California to clearly see where they need to take mitigating steps (like arranging for additional hospital beds or imposing social distancing measures) that could save lives.

DATA

Here is the key data gathered for this project:

- population by Southern California (SoCal) zip code downloaded from **US Census using their public API**
- SoCal hospital name and latitude/longitude gathered using **Foursquare API**
- SoCal hospital name, city and number of beds scraped from **American Hospital Directory website** https://www.ahd.com/states/hospital_CA.html
- COVID-19 overall hospitalization rate calculated from data downloaded from **NYC Health website** https://www1.nyc.gov/site/doh/covid/covid-19-data.page (New York City is similar in size to the SoCal region and has the most COVID data)
- geocoordinates from **geopy and Nominatum** for Los Angeles, CA as an appropriate centerpoint to map SoCal using **Folium**

NOTE: zip code was chosen as the core measure (rather than city or county) to provide the most granular information. This will allow Los Angeles to be split into its far-ranging geographic communities (e.g. Century City & Westwood & Downtown Los Angeles) rather than just lumping all LA population data into one wide spread blob called Los Angeles.

Also obtained this supplemental data to aid presentation and ease comprehension:
cross-reference of SoCal zipcodes to their corresponding city using scraped & filtered data from https://www.laalmanac.com/communications/cm02_communities.php

Here's how the data was used once gathered:

1. Calculate population and number of beds for each city/community based on zip codes
2. Convert population to max # of beds needed for each city/community using the overall hospitalization rate from NYC
3. Calculate the 'bed supply ratio' for each city/community = max # of beds/number of beds (a value < 1 indicates there are enough hospital beds to cover the max # of beds needed; a value > 1 indicates a possible shortage of beds)
4. Assign each zip code in a city or community the 'bed supply ratio' for that city or community
5. Plot the 'bed supply ratio' for each zip code on a thematic map of SoCal visually highlighting the high risk areas
6. Plot each hospital on that SoCal map with its name and number of beds appearing when clicked
7. Tabulate the 10 least at risk communities

Data Gathering, Cleaning and Wrangling

As expected, this stage consumed the largest amount of time during the project. Certain data (e.g. population by zip codes) was relatively clean and required little manipulation—in that instance, just needed to remove an unwanted character from the population values. Other data (e.g. hospital beds) was incomplete and required manual lookup and value setting. Details of how each dataset was put together follow.

Population by Zip Code: easily downloaded latest available American Community Survey population data by zip code for 2018 using **read_html** (as a browser) **from the US Census site**. Kept only two columns zip and population. Needed to remove an unwanted trailing] from the population values.

Hospital Geocoordinates (Lat/Long): got downtown LA lat/long coordinates from **geopy** via **nominatum**. Put those geocodes into Foursquare API to obtain just the hospital venues around and associated latitude/longitude coordinates surrounding Los Angeles. That only returned 25 hospitals, as apparently not many Foursquare users think of reviewing hospitals (or maybe they're just healthy?). Looking at the hospital beds data, saw there should be more than 100 hospitals in the SoCal area.

Was able to use **OpenCageGeo API** to gather lat/long for the over 100 hospitals listed in American Hospital Directory data.

Hospital Beds: scraped hospital name, city and number of beds from AHD website using **read_html** in python. Certain hospitals had zero beds listed; for those, found figures on internet and plugged those values back into dataframe using .iat method in pandas.

Hospitals and ZipCodes: Unfortunately, the OpenCageGeo API could only find zip code for about half the hospitals. Looked up the remaining zip codes one by one using latlong.net, combined those into an Excel sheet then imported that using read_excel into jupyter notebook as a dataframe.

Extra hospitals found in Foursquare data: The Foursquare hospital data, although limited, did identify a handful of hospitals that were not included in the AHD hospital beds list. Beds and zip code data for these hospitals was gathered online. Lat/long coordinates were obtained using OpenCage. These hospitals were then merged back into the data.

COVID-19 statistics: found overall infection rates in this online article https://nypost.com/2020/04/17/more-have-been-infected-with-coronavirus-than-believed-study/ ; used the latest NYC data at their website for cases/hospitalizations to calculate max hospitalization rate.

Final cleaning & prep: As our approach involved combining data from several datasets, needed to ensure key columns would 'match' from set to set. Made sure all columns were similarly named (e.g. City) across datasets. Also stripped leading & trailing spaces from the City and HospitalName columns. Set certain values for beds and hospital name directly in the dataframes to make sure values matched and data was in place.

METHODOLOGY

From the data gathering, cleaning & wrangling above, we have the following data in these dataframes:

1. **Communities and their corresponding zip codes** in *df_zipz2*
2. **Hospital names and their lat/long coordinates** in *df_hosp_coords*
3. **Hospitals and their corresponding zip codes** in *df_hosp_zip*
4. **Hospitals and number of beds** in *df_beds1*
5. **Population by zipcode** in *df_zip*

We also have two COVID-19 values to use in our calculations:

infection rate = 4%
hospitalization rate = 26%

Here's how we'll use this data to move forward on our task of plotting those SoCal communities at highest risk from COVID-19

1. Combine all hospital dataframes into a single df (*df_hosp*) with all the hospital related data
2. Add the population data for each zip code to the communities and zip codes dataframe and name that *df_zipz_pop*
3. Use groupby with sum to create a new df (*df_comm_pop*) that contains the total population for each city
4. Use groupby with sum to create a new df (*df_comm_bed*) that contains the total beds for each city, name that column TotalBeds
5. Add the hospital beds totals for each city to *df_comm_pop* name that new dataframe df_CALC
6. Add a new calculated column to df_CALC = pop total X infection rate X hospitalization rate = MaxBeds
7. Add a second calculated column in df_CALC BSR = BedSupplyRatio = TotalBeds/MaxBeds (a value > 1 indicates enough beds)
8. "Lookup" and assign the BSR to each zip code in df_zipz2 using the City as the crossreference (every zip in the city gets the same BSR).
9. Assign a BSR of 0 to any city that doesn't have any hospitals directly in their zip codes

Then we'll plot the hospitals onto a SoCal map with markers including bed info.
Next, we'll create a choroplot using the BedSupplyRatio for each zip code and add that to the SoCal map.
We'll then use the BSR data to tabulate the 15 SoCal cities least at risk for COVID-19

Here are some snippets of the results of the data manipulation as noted above. The first table shows the summary data including the calculated columns MaxBeds and BedSupplyRatio at the community level.

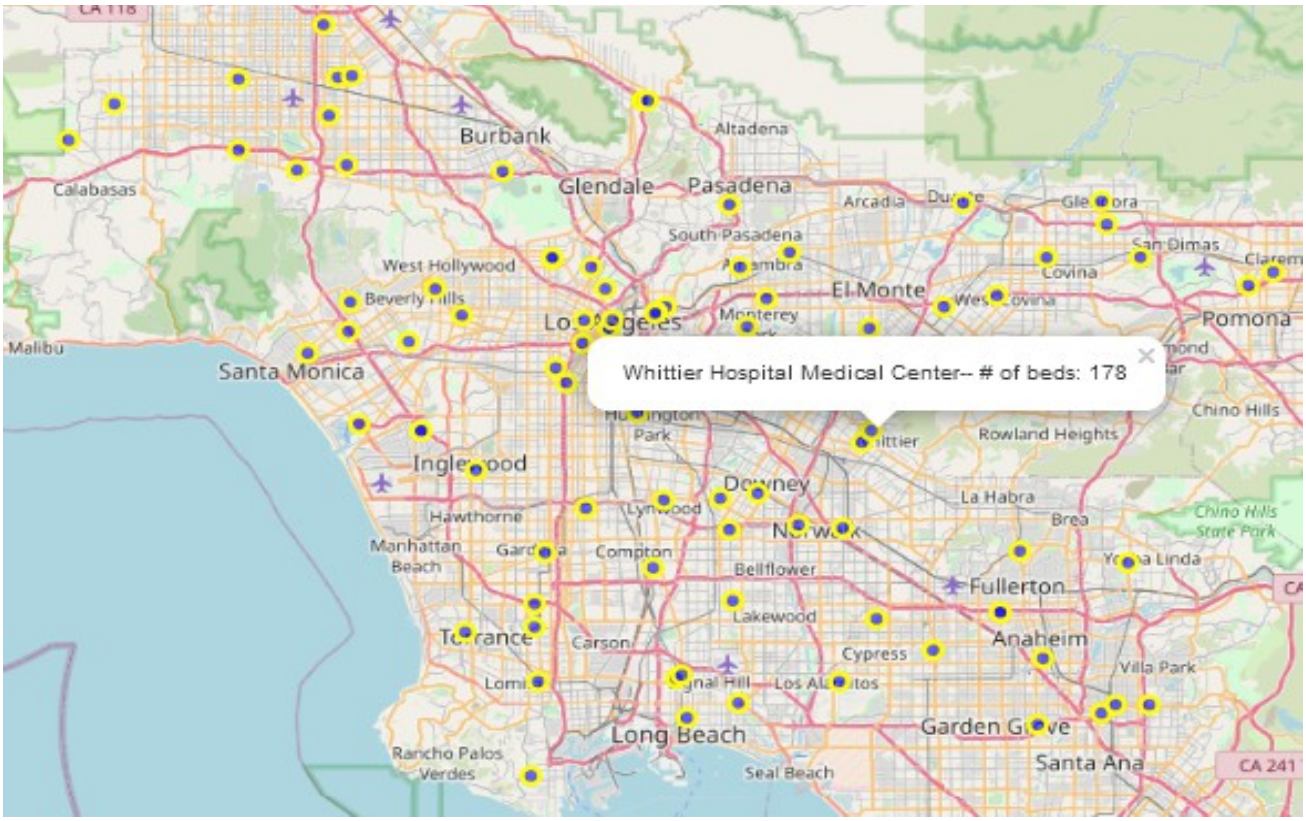
City	TotalPopulation	TotalBeds	MaxBeds	BedSupplyRatio
West Carson, Torrance	18188.0	NaN	189.0	NaN
West Covina	109408.0	371.0	1138.0	0.326011
West Fairfax	28085.0	NaN	292.0	NaN
West Hills	80356.0	228.0	836.0	0.272727
West Hollywood	93914.0	880.0	977.0	0.900716
West Los Angeles	47967.0	265.0	499.0	0.531062

This second table shows that same data assigned back to the individual zip codes associated with each community. This is the data we need to map BedSupplyRatio on top of zip code outlines.

	City	ZipCode	TotalPopulation	TotalBeds	MaxBeds	BedSupplyRatio
0	Acton	93510	7626.0	NaN	79.0	NaN
1	Agoura Hills	91301	25631.0	NaN	267.0	NaN
2	Agoura Hills	91376	25631.0	NaN	267.0	NaN
3	Agua Dulce	91390	19053.0	NaN	198.0	NaN
4	Alhambra	91801	84864.0	144.0	883.0	0.16308

Using that summary data, **Folium maps** were created to visualize this data all at once for the SoCal region.

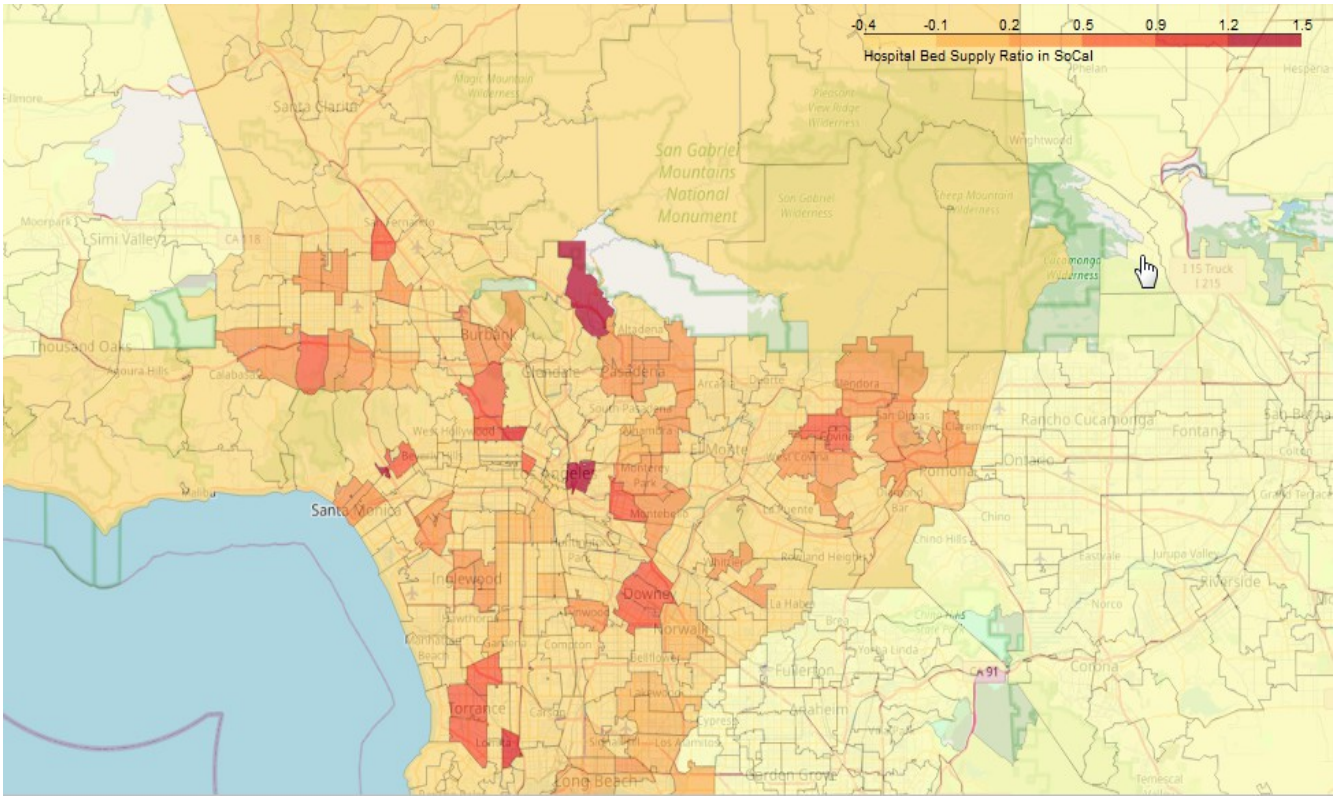
First, hospitals were plotted using their geocodes. When an individual marker is clicked on the map, the hospital name and number of beds are displayed.



MAP OF SOUTHERN CALIFORNIA HOSPITALS

Next, a thematic map showing relative BedSupplyRatio (BSR) by zipcode was created. Key here was obtaining an LA area zipcode geojson file from the LA Times.

A BSR of 1 or greater indicates that community has enough hospital beds to cover the likely hospitalization required by its population. The darker colors are for higher BSR.



HOSPITAL BED SUPPLY RATIO IN SOUTHERN CALIFORNIA

ANALYSIS

First, let's perform some high-level investigation of our data and determine the overall SoCal BSR. This will give us an idea of how ready the entire region is to deal with COVID-19.

Recall from data gathering that our infection rate is 4% and the hospitalization rate is 26%. We need to get the SoCalTotalPop & SoCalTotalBeds figures from our data. Then we can calculate SoCalMaxBeds and SoCalBSR. Here's how that went:

```
#get total population in the dataset from df_CALC
TotalSoCalPop= df_CALC['TotalPopulation'].sum()
print(TotalSoCalPop)
```

```
18441887.0
```

```
#apply infection and hospitalization rates to get SoCalMaxBeds
SoCalMaxBeds= TotalSoCalPop*inf_rate*hosp_rate
print(SoCalMaxBeds)
```

```
191795.6248
```

```
#get total beds in the dataset from df_CALC
TotalSoCalBeds= df_CALC['TotalBeds'].sum()
print(TotalSoCalBeds)
```

```
22879.0
```

```
#calculate overall SoCal BSR = SoCalMaxBeds/SoCalTotalBeds
SoCalBSR=TotalSoCalBeds/SoCalMaxBeds
print(SoCalBSR)
```

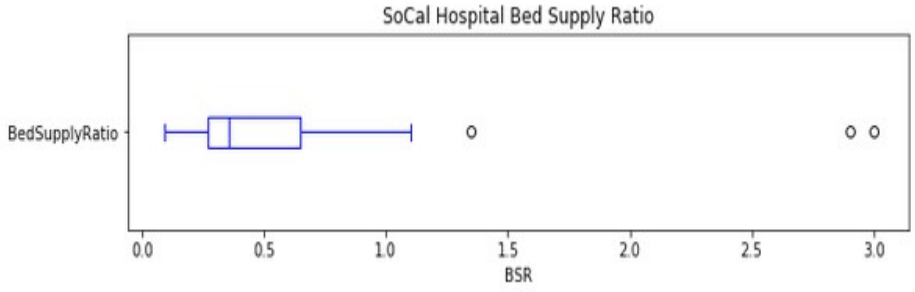
```
0.11928843540543581
```

So, overall the **SoCalBSR is 11.9%.** This means that SoCal *as a region* does not have enough beds to cover its maximum COVID-19 risk. Now, let's dive more deeply into the data.

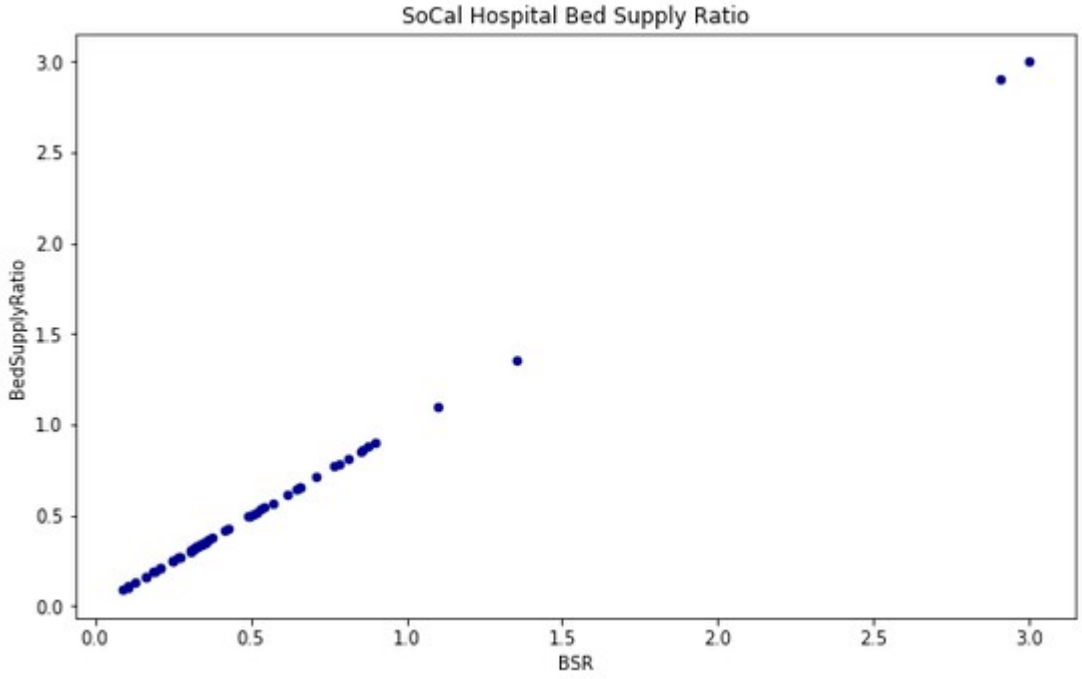
Looking at the statistics for the BSR data, one community immediately stood out, Veterans Administration (VA) with a BSR of 94.5! The VA is actually a large hospital that occupies a significant enough land area to warrant its own zip code. It has very few residents and thus an unusually high BSR. To make our data fit and graphs have meaning, set the VA BSR = 3.0 in most cases.

Here are the BSR stats with that adjustment. We can see the two outliers near 3.0 and that most communities have a BSR between 0.10 and 1.1-- with the average of 0.53. This means the average community with hospital beds can cover 53% of its maximum COVID-19 hospitalization.

	BedSupplyRatio
count	56.000000
mean	0.537853
std	0.538922
min	0.088703
25%	0.269313
50%	0.356400
75%	0.649787
max	3.000000

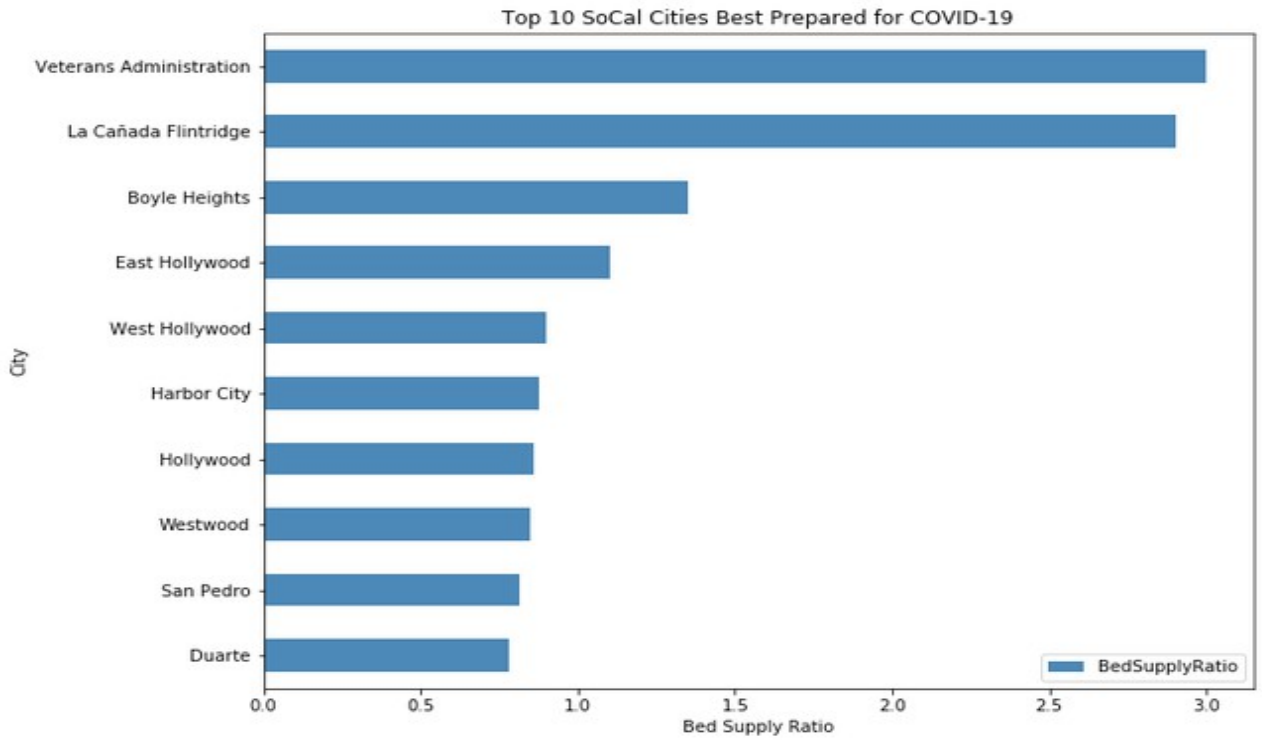


Let's scatter plot this data to see if we can identify some groupings...



Again, see the outliers and a couple likely groupings-- i.e. one above 1, one from 0.7 to 1.0, another from 0.5 to 0.7. We'll use kmeans clustering to determine and map these clusters shortly.

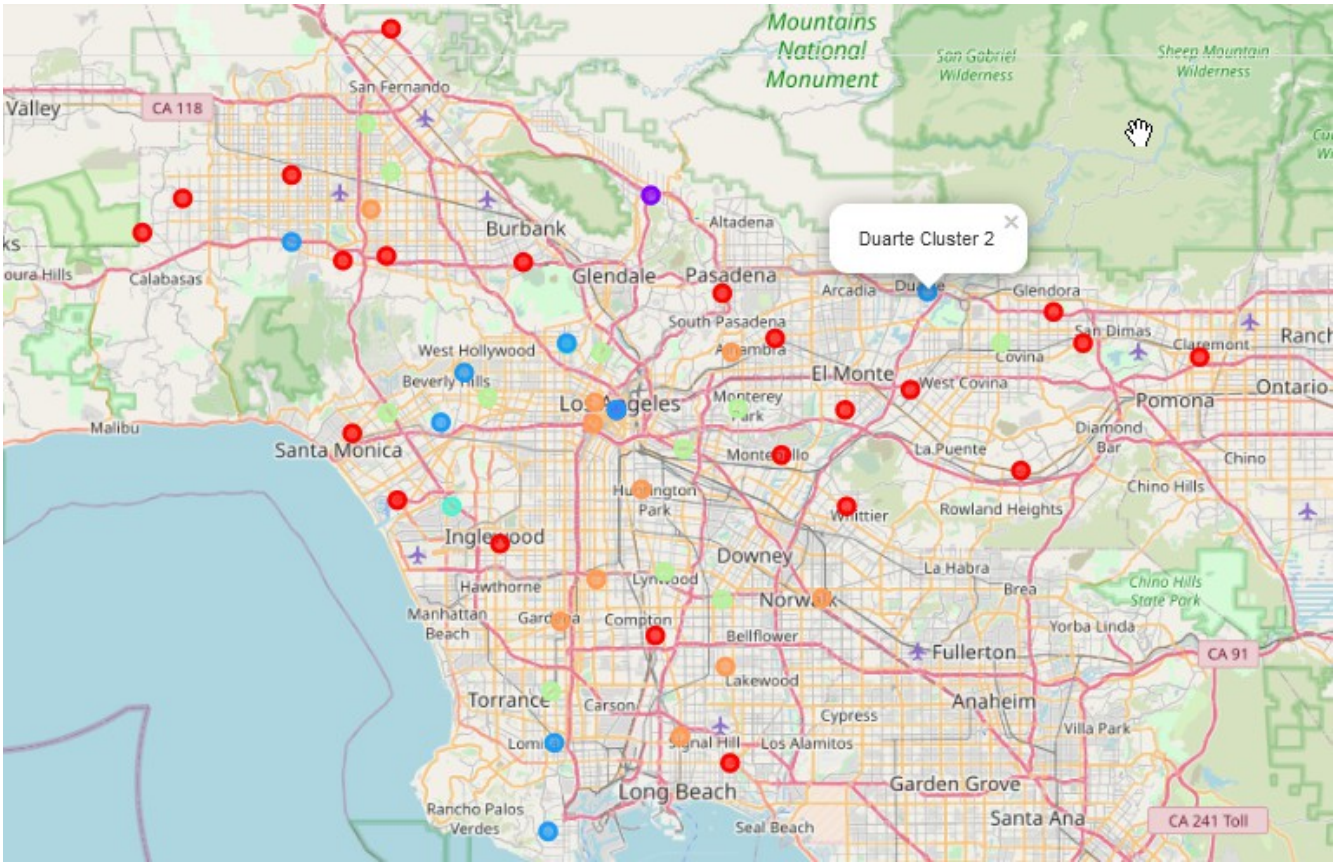
Over 80% of the communities did not have any hospitals in any of their zip codes. Thus, all 250+ of those communities tied for the title of community at the 'highest-risk'. Here's a bar graph showing the ten communities that are the best-prepared and 'least-risk' .



Here's just a portion of the complete table showing all communities, their BSR score and underlying data like TotalPopulation, TotalBeds and MaxBeds (the likely number of beds that population will require).

City	TotalPopulation	TotalBeds	MaxBeds	BedSupplyRatio
Veterans Administration	916.0	945.0	10.0	94.500000
La Cañada Flintridge	20423.0	616.0	212.0	2.905660
Boyle Heights	100534.0	1413.0	1046.0	1.350860
East Hollywood	37898.0	434.0	394.0	1.101523
West Hollywood	93914.0	880.0	977.0	0.900716
Harbor City	28185.0	257.0	293.0	0.877133
Hollywood	195084.0	1743.0	2029.0	0.859044
Westwood	50288.0	445.0	523.0	0.850860
San Pedro	84123.0	712.0	875.0	0.813714
Duarte	26601.0	217.0	277.0	0.783394
Tarzana	30804.0	246.0	320.0	0.768750
Westlake	49492.0	366.0	515.0	0.710680

Finally, used kmeans = 6 to cluster the BSR data. Here's the map showing that.



BSR ~3.0 BSR ~1 BSR >0.7 BSR 0.5-0.7 BSR 0.2-0.4 BSR <0.2

BED SUPPLY RATIO CLUSTERS IN SOUTHERN CALIFORNIA

RESULTS AND DISCUSSION

Overall, the Southern California region is not well prepared for COVID-19. *It only has 12% of the maximum hospital beds it likely will need for hospitalized COVID-19 patients.* Thus, additional beds must be created and/or other steps taken (e.g. social distancing) to ensure the actual level of hospitalization remains below the number of available beds.

Looking at specific communities in the region, turns out over 80% (roughly 250 out of 300) do not have any hospitals at all and thus must rely on nearby hospitals to serve their population should they become ill. Almost every SoCal community is at risk for COVID-19, making a Top 10 neighborhoods at risk list moot.

Instead, a Top 10 graph was created showing those communities who are the most prepared to handle COVID-19 for their populations. One 'community' (the Veterans Administration) actually had more hospital beds than residents. This 'community' is really a hospital complex that is large enough in area to have its own zip code. The Top 10 cities can provide at least 75% of the hospital beds required for their population and so do not need to take immediate additional measures.

On average, the 56 SoCal communities that do have hospitals can provide 53% of the hospital beds their population requires.

Looking at the map of hospitals in SoCal, we see a fairly even distribution of hospitals in the region. The shaded map showing calculated BedSupplyRatio (TotalBeds/MaxBeds) for each zip code also indicates areas that can cover their hospitalization needs are spread out geographically. The cluster map confirms the reasonable spacing of relatively well-prepared communities.

CONCLUSION

This project set out to identify where in the SoCal region leaders might need to take extra action to ensure their communities could deal well with COVID-19. Using population and hospital bed information by zip code, it has provided maps and tables to allow such decisions to be made.

Almost all communities in SoCal, need to take extra steps--as they do not have enough hospital beds to cover the maximum hospitalization from COVID-19.

This project only looked at one factor (hospital beds) that has been identified as key in combating COVID-19. Further analysis of other factors such as age, health of population, population density should be considered along with hospital beds to determine a more complete assessment of neighborhood risk.