Coursera-IBM Data Science Capstone Project

---

(1) Introduction

(1-1) Background

In difficult public health circumstances, supplies may become scarce, and communities may not be affected equally. Looking at a range of demographic, public health, and hospital data may allow decision-makers to create informed plans about how to mobilize medical resources in a targeted fashion.

In more regular times, certain regions may still be underserved with regards to their healthcare needs. Geospatial analysis can also help determine which areas may benefit most from a newly constructed hospital or increased medical staffing.

(1-2) Problem

This capstone project for the Coursera-IBM Data Science Certificate will explore hospital data by cities in Los Angeles County, as well as COVID case data in the region between mid-December 2020 to mid-January 2021.

More specifically, this assignment tries to analyze data in two areas.

1. Which zip codes in LAC are located furthest away from the nearest medical center, as found by Foursquare, and would therefore perhaps benefit most from a new hospital.
2. Which cities in LAC have the highest ratios of COVID cases to hospital beds, and could require increased medical assistance in the near future; as well as which cities in LAC have the highest ratios of people to hospital beds, and may benefit from increased hospital capacities in the longer term.

(2) Data

In addition to the Foursquare venue data required for this project, several other data sources contributed to this assignment.

(2-1) Cities In Los Angeles County And Their Populations

A list of cities that make up LAC, and their populations, was provided by Wikipedia.

List of cities in Los Angeles County, California - Wikipedia

Features include

- **City Name**
- Data City Was Incorporated Into County
- **City Population As Of 2010 US Census**

This data spans 88 cities in LAC.

(2-2) United States Zip Codes

A list of zip codes that make up the US, including a handful of other features, was compiled by Schuyler Erle at Geocoder, and provided by Civic Space Labs.

download – CivicSpace Labs

Features include

- **Zip Code**
- **City**
- State
- **Longitude**
- **Latitude**
- Time Zone
- Observance Of Daylight Savings Time

Zip codes were filtered so all matched up to a city in LAC. This filtered data spans 404 zip codes in 75 cities in LAC.

(2-3) Los Angeles County Healthcare Facility Beds

A list of California state healthcare facilities and number of beds, among several other features, was provided by the California Health And Human Services Open Data Portal.

Licensed Healthcare Facility Listing - Licensed Healthcare Facility Listing, December 31, 2020 - California Health and Human Services Open Data Portal

Features include

- Facility ID
- **Facility Name**
- License ID
- Facility Level
- Address
- **City**

- **Zip Code**
- County Code
- County Name
- Emergency Services Provided
- **Number Of Beds**
- Open Or Not
- Opening Data
- License Type
- License Category
- **Latitude**
- **Longitude**

Facilities were filtered to include only those that were located in LAC, are currently open, and have a specified number of beds. This filtered data spans 625 healthcare facilities across LAC.

(2-4) Los Angeles County COVID Cases

A list of daily-updated cumulative COVID case counts for California state since 20200228 was provided by the Los Angeles Times.

california-coronavirus-data/latimes-place-totals.csv at master · datadesk/california-coronavirus-data (github.com)

Features include

- **Date**
- County
- Federal Information Processing Standards (FIPS) Code Assigned To County
- **City**
- **Confirmed Cases So Far**
- Notes
- Latitude
- Longitude
- Population

Data points from LAC on 20201215 and 20210115 were taken to calculate the number of new cases in the past 30 days in the county as of 20210115.

Cities not in Wikipedia's LAC city list were assumed to be neighborhoods in Los Angeles City, and their cases were summed.

This final data spanned 85 LAC cities.

(2-5) Foursquare Los Angeles County Medical Center Venue Location Data

Foursquare venue data was accessed to retrieve the 50 closest hospitals, hospital wards, emergency rooms, and urgent care centers within 100,000 meters of the centers of each zip code in LAC. Duplicate venues were removed.

Features extracted from the JSON data were

- **Venue Name**
- **Zip Code**
- **City**
- **Latitude**

- **Longitude**
- **Venue Type**

Data points that did not include zip codes and/or cities were labelled to the closest zip codes based on latitude and longitude, and mapped to the corresponding cities.

This data only spanned 209 healthcare facilities, compared to 625 from the earlier bed data set.

(2-6) Los Angeles County GeoJSON Zip Code Boundaries

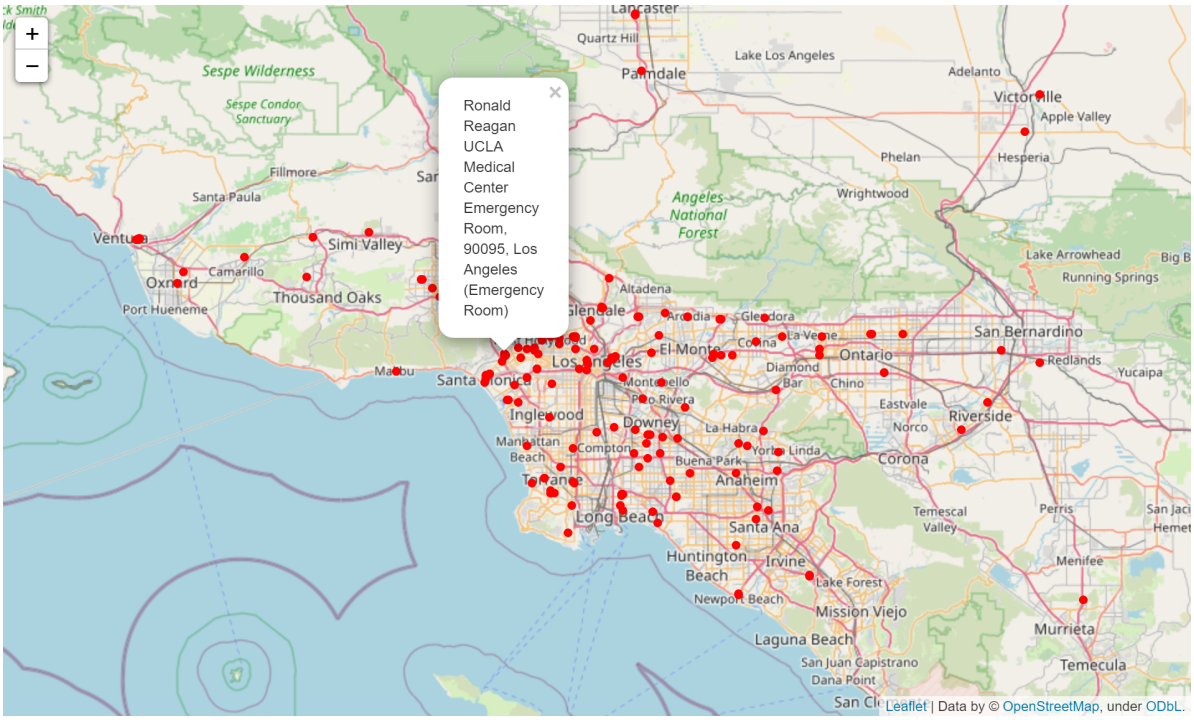
Los Angeles County zip code boundaries were provided, again, by the Los Angeles Times. These were used when visualizing data in Folium maps.

[Mapping L.A. Boundaries API - Data Desk - Los Angeles Times (latimes.com)]

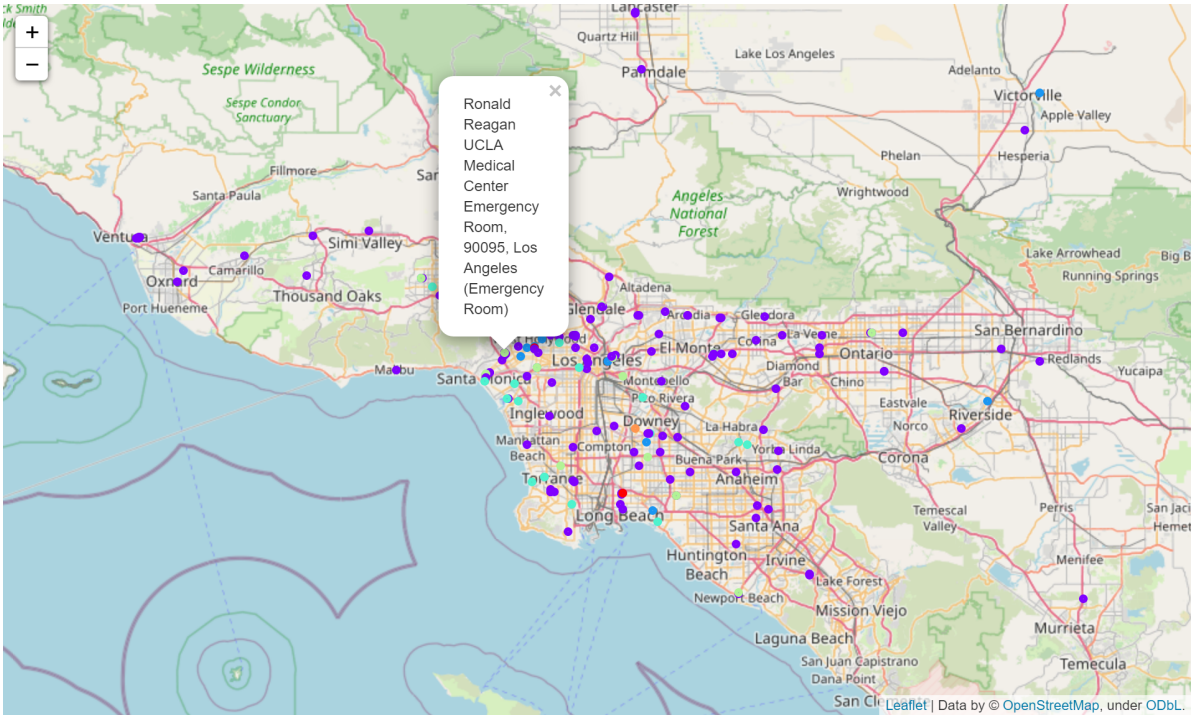
(3) Methodology

(3-1) Foursquare - Medically Remote Zip Codes

Foursquare medical center location data was mapped in Folium.



A similar map to the one above, but with location dots color-coded by medical center type was also created.



In the above color-coded map, purple dots indicate hospitals, blue dots medical centers (general), light green dots emergency rooms, and turquoise dots urgent care centers. The red dot is a hospital ward and the orange dot an office.

A simple algorithm was used to calculate the the distances between zip codes and the medical facilities identified by Foursquare.

(3-2) City Beds

The LAC COVID data was merged with the LAC medical facility bed data to determine the cities in the region with the highest ratios of coronavirus cases per hospital bed. A choropleth map was created in Folium to visualize the data.

The LAC city populations data was merged with the LAC medical facility bed data to determine the cities in the region with the highest ratios of people per hospital bed. Again, a choropleth map was created in Folium to visualize the data.

(4) Results

(4-1) Foursquare - Medically Remote Zip Codes

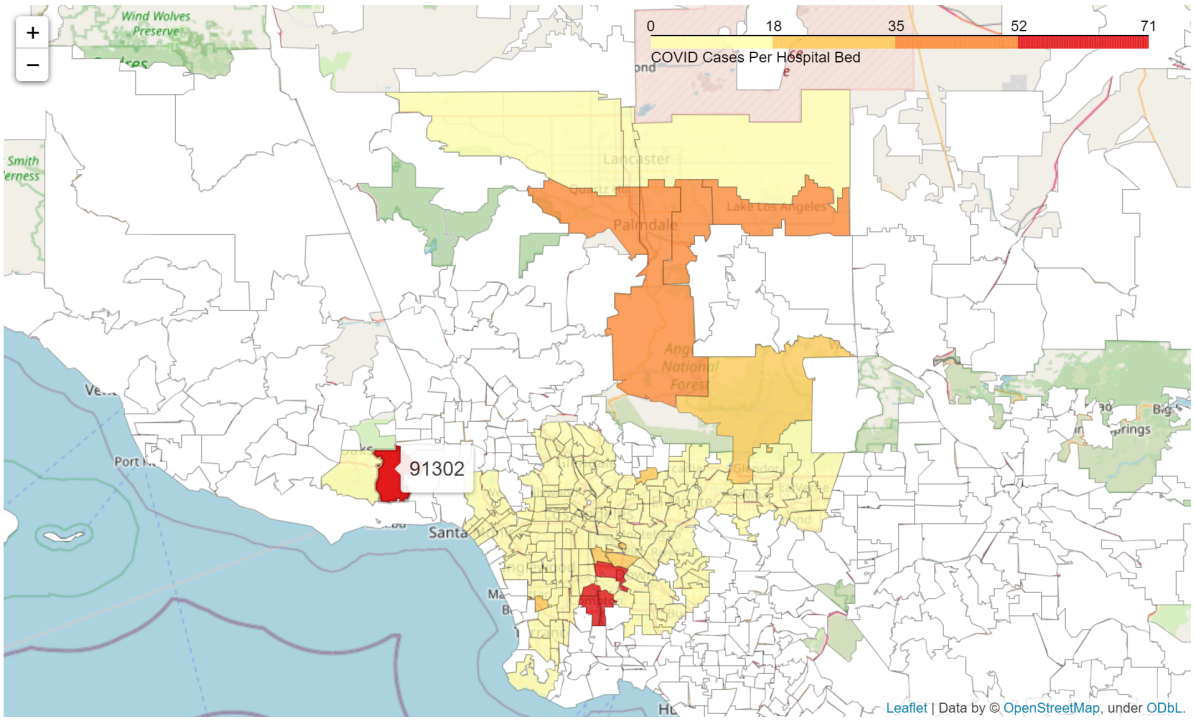
From the two maps created, it appears that the majority of health care facilities in LAC are situated near the center of the region. Areas in the more rural, outer reaches of the county, with smaller populations and fewer roads, may benefit from increased healthcare infrastructure.

Zip Codes Furthest From Nearest Medical Center (Foursquare)
91359
91361
90704
90265
91301
93536
90263
91302
91711
91767

(4-2) City Beds

The following table shows the top 5 cities in LAC in terms of new COVID cases between 20201215-20210115 per hospital bed.

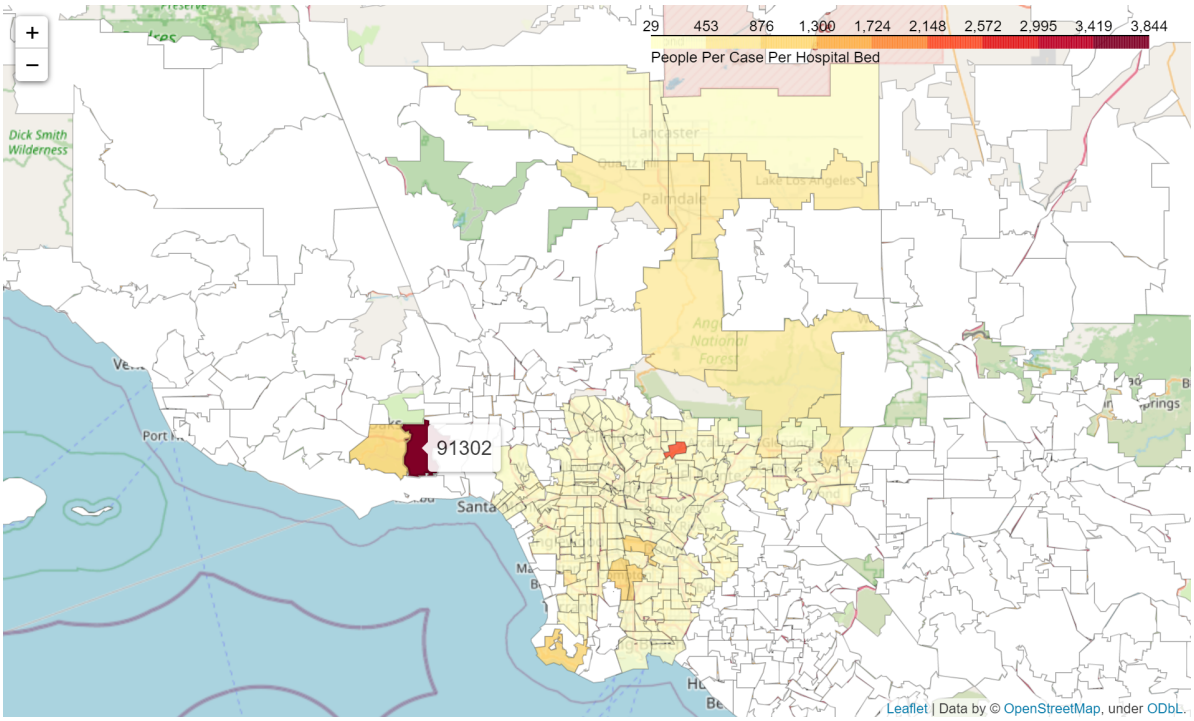
City	COVID Cases	Medical Facility Beds	Cases/Bed
South Gate	6916	99	69.858586
Compton	6174	99	62.363636
Calabasas	336	6	56.000000
Palmdale	8468	214	39.570093
Azusa	2194	65	33.753846



Zip codes 90220-90223, 90280, and 91302 have the most COVID cases per hospital bed. These correspond to the cities of Compton, South Gate, and Calabasas.

The following table shows the top 5 cities in LAC in terms of people per hospital bed.

City	2010 Population	Medical Facility Beds	People/Bed
Calabasas	23058	6	3843.000000
San Marino	13147	6	2191.166667
Rancho Palos Verdes	41643	34	1224.794118
Agoura Hills	20330	20	1016.500000
Compton	96455	99	974.292929



Zip codes 91302, 91108, and 90275 have the highest ratios of people in the general population for each hospital bed. These correspond to the cities of Calabasas (again), San Marino, and Rancho Palos Verdes.

---

(5) Discussion

A major obstacle, and a huge liability in the project results, is the lack of consistency across data from each of the sources used. Between the medical bed data and the Foursquare venue data, facility names were formatted differently, with bed data being in all-caps and occasionally including city names at the end. Latitude and longitude coordinates between the two data sets often had differing levels of precision, and sometimes pointed to separate areas. Foursquare venue data was also unable to capture the complete set of facilities included in the bed data provided by the California government.

The numbers of beds counted for certain communities were suspicious, like having only 6 beds for the top two cities in terms of ratio of people per bed, Calabasas and San Marino.

In addition, different data sets separated LAC in different ways. Some split Los Angeles city into separate neighborhoods of more comparable population sizes to the other cities in the county; others didn't. Furthermore, zip code boundaries sometimes don't match up with city boundaries, furthering complicating matters.

Lastly, the GeoJSON file included a number of cities in Orange County to the southeast of LAC, potentially among other areas not within LAC.

---

(6) Conclusion

Geospatial data analysis can be useful in issues of public concern, in addition to solving business problems. Combined with resource data, location info can be employed to help more efficiently provide access to government, or other, supplies.

As a general result reached by this project, many remote and underserved communities may require better access to healthcare resources to bolster their responses to public health crises and maintain higher qualities of care for their populations even in the absence of severe outbreaks.

In the future, more well-documented data sets across entities could help provide better analyses. This could include standardized naming conventions, standardized county sub-divisions, and greater precision in geospatial coordinates, among other features. Organizations that collaborate and coordinate their data collection and formatting can create higher quality data that is more comprehensive and can yield more useful results.