COVID-19: best equipped San Francisco neighborhoods

21 June 2020 by Zharas Aitmambet
https://github.com/zaitmambet/covid19-best-equipped-sanfrancisco-neighborhoods

Introduction: Business Problem

The COVID-19 pandemic has hit the large urban areas in the U.S. the hardest. New York has had the largest outbreak in the country, although the number of new cases has declined over the recent weeks. Detroit's Wayne County and Chicago's Cook County are not far behind the epicenter in the number of coronavirus-related deaths.

San Francisco is the **second most densely populated** large U.S. city, behind New York City, and the fifth most densely populated U.S. county, behind only four of the five New York City boroughs. High population density is one of the factors that could potentially make San Francisco one of the U.S. cities that are most vulnerable to the novel coronavirus.

The objective of this project is to provide a helicopter view on the COVID-19 situation in San Francisco by clustering neighborhoods based on **population density** and either **(i) number of hospitals** or **(ii) number of hospital beds** within the neighborhood borders that are used as indicators of neighborhood's adaptation and mitigation capacity against the pandemic. I will use a clustering algorithm and feed into it four variables for each neighborhood: neighborhood population density, number of hospitals per thousand people (fetched from Foursquare API), number of staffed hospital beds per thousand people, and number of intensive care unit (ICU) beds per thousand people. The results of clustering could be used as dummy variables for further more comprehensive analysis. The maps can provide a rough overview of concentration of hospital beds to general public living in San Francisco area.

Data sources

I will collect data from several sources:

List of 41 San Francisco neighborhoods created by grouping 2010 Census tracts:

- Data content: Neighborhood, Geographic coordinates (multipolygon)
- Data format: GeoJSON
- Data source: DataSF

San Francisco population data:

- Data content: Neighborhood, Population as of 2016
- Data format: PDF
- Data source: City and County of San Francisco

USA Hospital Beds:

- Data content: Hospital name, Geographic coordinates (point), Hospital type, Address, Number of licensed beds, Number of staffed beds, Number of ICU beds
- Data format: CSV
- Data source: COVID-19 GIS Hub

Location of venues with the category = "hospitals" in San Francisco:

- Data content: Name, Category, Geographic coordinates (point)
- Data format: JSON
- Data source: Foursquare API

Data collection and cleaning

Official San Francisco neighborhoods data

The official website of San Francisco city has quite a few versions of neighborhood division. I decided to use the one with 41 neighborhoods, because it it the only one for which reliable population information is available. Now, this dataset contains geodata for every neighborhood's borders in the multipolygon format, which I will later use to assign hospitals to neighborhoods. Here I add centroid points such that we can more easily see concentration of neighborhoods on a map below.

In [2]:
```
url_n41 = r"https://data.sfgov.org/api/geospatial/p5b7-5n3h?method=export&format=GeoJSON"
feat = requests.get(url_n41).json()
neighborhoods_sf_raw = gpd.read_file(url_n41)
neighborhoods_sf_41 = neighborhoods_sf_raw
neighborhoods_sf_41['Latitude'] = neighborhoods_sf_41["geometry"].centroid.y
neighborhoods_sf_41['Longitude'] = neighborhoods_sf_41["geometry"].centroid.x
neighborhoods_sf_41.rename(columns={"nhood" : "Neighborhood"}, inplace = True)
neighborhoods_sf_41["Population"] = 'NA'
```

In [3]:
```
neighborhoods_sf_41.head()
```

Out[3]:

	Neighborhood	geometry	Latitude	Longitude	Population
0	Bayview Hunters Point	MULTIPOLYGON (((-122.38158 37.75307, -122.3815...	37.730889	-122.386016	NA
1	Bernal Heights	MULTIPOLYGON (((-122.40361 37.74934, -122.4037...	37.740364	-122.415664	NA
2	Castro/Upper Market	MULTIPOLYGON (((-122.42656 37.76948, -122.4269...	37.762319	-122.435217	NA
3	Chinatown	MULTIPOLYGON (((-122.40623 37.79756, -122.4055...	37.796140	-122.407081	NA
4	Excelsior	MULTIPOLYGON (((-122.42398 37.73155, -122.4239...	37.718562	-122.431807	NA

Extracting population data from official report on San Francisco neighborhoods

There is not much relevant information on neighborhood population available online. In fact, there is no clearly defined list of neighborhoods in San Francisco. For that reason, I decided to rely on most reliable data available online: official reports. I use **pdfminer** and **PyPDF4** packages to extract text from PDF.

```
url = 'http://default.sfplanning.org/publications_reports/SF_NGBD_SocioEconomic_Profiles\
/2012-2016_ACS_Profile_Neighborhoods_Final.pdf'

r = requests.get(url, stream=True)
r.raw.decode_content = True

open('SF_neighborhoods_official_41.pdf', 'wb').write(r.content)
```

```
4392742
```

Next, I extract text from the report, and do a bit of text cleaning with **ftfy** package such as replacing ligatures with separate letters and removing line breaks and special symbols.

```
pdfConverter = PdfConverter(file_path="SF_neighborhoods_official_41.pdf")
text = pdfConverter.convert_pdf_to_txt()
text_clean = text.replace('\n', ' ')
text_clean = text_clean.replace('\x0c', ' ')
text_clean = text_clean.replace('- ', '')
text_clean = ftfy.fix_text(text_clean)
ft.append(text_clean)
```

Official US hospital beds data (incl. ICU beds) ¶

I download official data on hospital beds in the U.S. and filter out all non-relevant information and rename some of the variables.

```
us_hospitals = "https://opendata.arcgis.com/datasets/1044bb19da8d4dbfb6a96eb1b4ebf629_0.csv"
us_hospitals_data = requests.get(us_hospitals).content
hosp_csv = open('us_hospitals_bed_capacity_2020.csv', 'wb')
hosp_csv.write(us_hospitals_data)
hosp_csv.close()
print('Data downloaded!')
```

```
hosp_beds_overall = pd.read_csv(
    "us_hospitals_bed_capacity_2020.csv",
    encoding = 'cp850')
hosp_beds_overall.head(2)
```

	X	Y	FID	HOSPITAL_NAME	HOSPITAL_TYPE	HQ_ADDRESS	HQ_ADDRESS1	HQ_CITY	HQ_STATE	HQ_Z
0	-112.066157	33.495498	1	Phoenix VA Health Care System (AKA Carl T Hayd...	VA Hospital	650 E Indian School Rd	NaN	Phoenix	AZ	
1	-110.965885	32.181263	2	Southern Arizona VA Health Care System	VA Hospital	3601 S 6th Ave	NaN	Tucson	AZ	

The dataframe of 17 hospitals in San Francisco from the official dataset of US hospital beds looks like this:

HOSPITAL_NAME	HOSPITAL_TYPE	LATITUDE	LONGITUDE	NUM_STAFFED_BEDS	NUM_ICU_BEDS	ADULT_ICU_BEDS
California Pacific Medical Center - Mission Bernal Campus (FKA California Pacific Medical Center - St Lukes Campus)	Short Term Acute Care Hospital	37.747703	-122.420667	120.0	10	10
California Pacific Medical Center - Davies Campus	Short Term Acute Care Hospital	37.768193	-122.435709	137.0	8	8
Zuckerberg San Francisco General Hospital (FKA San Francisco General Hospital)	Short Term Acute Care Hospital	37.755727	-122.404738	284.0	58	58

Forsquare API: retrieving data on San Francisco hospitals

We will use Foursquare data to create an alternative simplified clusterization of neighborhoods based on hospital locations instead of hospital beds data. First, I enter all necessary prerequisite inputs (client_id, client_secret, version etc.). After that, using the same address as was used to create the maps, we will now retrieve venues (**limit** = 100) that are situated within San Francisco city borders (**near** = address) and have category key that corresponds to **categoryId**= "4bf58dd8d48988d196941735".

The resulting dataset of 35 hospitals fetched from Fousquare API looks like this:

	name	categories	lat	lng	address	postalCode
0	Kaiser N.I.C.U.	Hospital	37.782523	-122.442392	2425 Geary Blvd	94115
1	Zuckerberg San Francisco General hospital and ...	Hospital	37.755659	-122.404956	1001 Potrero Ave	94110
2	Adult Infusion Center	Hospital	37.766612	-122.390380	1825 4th St	94158
3	Kaiser Permanente Pediatrics	Hospital	37.783285	-122.440394	2238 Geary - 5th Floor	NaN
4	SFGH Cafeteria	Hospital	37.754976	-122.404324	1001 Potrero Ave	94110
5	UCSF Bakar Precision Cancer Medicine Building	Hospital	37.766648	-122.389704	1825 4th St	94143
6	Seton Medical Center - Cardiac & Pulmonary Reh...	Hospital	37.679535	-122.474396	1900 Sullivan Ave	94015
7	Saint Francis Memorial Hospital Physical Therapy	Hospital	37.777448	-122.389567	24 Willie Mays Plz	94107
8	Bldg 90	Hospital Ward	37.756950	-122.406035	NaN	94110

Data Preparation

Neighborhoods dataset

Adding population data

In previous section we collected data from various sources. Now that we have all necessary data, we extract relevant population data using **re** package, and then merge it with the neighbourhood location dataset.

	Neighborhood	geometry	Latitude	Longitude	Population
0	Bayview Hunters Point	MULTIPOLYGON (((-122.38158 37.75307, -122.3815...	37.730889	-122.386016	37600
1	Bernal Heights	MULTIPOLYGON (((-122.40361 37.74934, -122.4037...	37.740364	-122.415664	26140
2	Castro/Upper Market	MULTIPOLYGON (((-122.42656 37.76948, -122.4269...	37.762319	-122.435217	21090
3	Chinatown	MULTIPOLYGON (((-122.40623 37.79756, -122.4055...	37.796140	-122.407081	14820
4	Excelsior	MULTIPOLYGON (((-122.42398 37.73155, -122.4239...	37.718562	-122.431807	39340
5	Financial District/South Beach	MULTIPOLYGON (((-122.38753 37.78280, -122.3875...	37.790418	-122.397053	17460
6	Glen Park	MULTIPOLYGON (((-122.44738 37.74648, -122.4472...	37.739605	-122.436326	8210
7	Inner Richmond	MULTIPOLYGON (((-122.45932 37.78752, -122.4592...	37.780950	-122.465434	22500
8	Golden Gate Park	MULTIPOLYGON (((-122.44092 37.77363, -122.4407...	37.769029	-122.481922	90
9	Haight Ashbury	MULTIPOLYGON (((-122.43200 37.77143, -122.4319...	37.768374	-122.444467	18050
10	Hayes Valley	MULTIPOLYGON (((-122.42081 37.77400, -122.4210...	37.774787	-122.429044	18250

Adding area

Further, we convert degrees into meters and then use **.area** method to compute area, not forgetting to divide by 1000^2 to get kilometer values. As a result, we now have a dataset containing area of each neighbourhood which will be necessary for computing population densities.

	Neighborhood	geometry	Latitude	Longitude	Population	Area, sq.km
0	Bayview Hunters Point	MULTIPOLYGON (((-122.38158 37.75307, -122.3815...	37.730889	-122.386016	37600	21.369042
1	Bernal Heights	MULTIPOLYGON (((-122.40361 37.74934, -122.4037...	37.740364	-122.415664	26140	4.453833
2	Castro/Upper Market	MULTIPOLYGON (((-122.42656 37.76948, -122.4269...	37.762319	-122.435217	21090	3.543839
3	Chinatown	MULTIPOLYGON (((-122.40623 37.79756, -122.4055...	37.796140	-122.407081	14820	0.929571
4	Excelsior	MULTIPOLYGON (((-122.42398 37.73155, -122.4239...	37.718562	-122.431807	39340	5.748898

Foursquare API data

We continue with merging Foursquare hospital location data with neighborhood boundaries, that will show which neighborhood each of 36 hospitals belongs to.

	name	categories	lat	lng	address	postalCode	geometry	Neighborhood
30	UCSF Medical Center Building 1	Hospital	37.764049	-122.457281	400 Parnassus Ave	94143	POINT (37.764 -122.457)	Inner Sunset
31	Gateway Medical Building	Hospital	37.766169	-122.390217	1825 4th St	94158	POINT (37.766 -122.390)	Potrero Hill
32	The Jewish Home	Hospital	37.727730	-122.430836	302 Silver Ave	94112	POINT (37.728 -122.431)	Excelsior
33	Chinese Hospital 東華醫院	Hospital	37.795672	-122.409201	845 Jackson St	94133	POINT (37.796 -122.409)	Chinatown
34	UCSF Medical Center At Mount Zion	Hospital	37.784893	-122.439164	1600 Divisadero St Fl 2	94143	POINT (37.785 -122.439)	Japantown

In the next step, applying the .**groupby** method, we count the number of hospitals located within neighborhoods' boundaries.

Neighborhood	Number of hospitals
Bernal Heights	1
Castro/Upper Market	1
Chinatown	1
Excelsior	1
Inner Sunset	6
Japantown	2
Lincoln Park	1
Lone Mountain/USF	2
Mission	6
Mission Bay	1
NA	2
Nob Hill	1
Pacific Heights	1
Portola	1
Potrero Hill	5
Presidio Heights	1
Western Addition	2

Now it's the time to merge this count data with Foursquare hospitals dataset and derive a couple of useful coefficients.

```
hospitals_fsq_demog["Population Density"] = hospitals_fsq_demog.apply(
    lambda row: (row["Population"]/row["Area, sq.km"]), axis=1)

hospitals_fsq_demog["Hospitals per Thousand People"] = hospitals_fsq_demog.apply(
    lambda row: (row["Number of hospitals"]/(row["Population"]/1000)), axis=1)
```

```
hospitals_fsq_demog.head()
```

	Neighborhood	Number of hospitals	Latitude	Longitude	Population	Area, sq.km	Population Density	Hospitals per Thousand People
0	Bernal Heights	1.0	37.740364	-122.415664	26140	4.453833	5869.101582	0.038256
1	Castro/Upper Market	1.0	37.762319	-122.435217	21090	3.543839	5951.173482	0.047416
2	Chinatown	1.0	37.796140	-122.407081	14820	0.929571	15942.838445	0.067476
3	Excelsior	1.0	37.718562	-122.431807	39340	5.748898	6843.050818	0.025419
4	Inner Sunset	6.0	37.758473	-122.464732	29120	5.885239	4947.972448	0.206044

Official US hospitals dataset

Similarly, we merge official data on US hospital beds with neighborhood boundaries data to compute remaining relevant variables. Aggregating the merged data by neighborhoods we can see how many staffed hospital beds are there within neigborhood boundaries.

Neighborhood	NUM_STAFFED_BEDS	NUM_ICU_BEDS
Bernal Heights	120.0	10
Castro/Upper Market	137.0	8
Chinatown	65.0	6
Excelsior	13.0	0
Inner Sunset	911.0	134
Lincoln Park	124.0	1
Lone Mountain/USF	392.0	57
Mission	284.0	58
Nob Hill	156.0	34
Potrero Hill	289.0	51
Twin Peaks	6.0	1
West of Twin Peaks	13.0	0
Western Addition	274.0	44

After merging the data on staffed and ICU beds with the official neighbohoods data, we get a new dataset which will let us compute additional coefficients: **Beds Per Thousand People** and **ICU Beds Per Thousand People**.

	Neighborhood	NUM_STAFFED_BEDS	NUM_ICU_BEDS	Latitude	Longitude	Population	Area, sq.km	Beds Per Thousand People	ICU Beds Per Thousand People
36	Portola	0.0	0.0	37.726792	-122.408993	16410	3.410645	0.0	0.0
37	Presidio	0.0	0.0	37.797377	-122.466370	3830	9.782601	0.0	0.0
38	Presidio Heights	0.0	0.0	37.786313	-122.451658	10720	2.078046	0.0	0.0
39	Treasure Island	0.0	0.0	37.820655	-122.369540	3090	3.680632	0.0	0.0
40	Visitacion Valley	0.0	0.0	37.712867	-122.410104	18570	2.526024	0.0	0.0

Methodology

In this project we will direct our efforts on identifying San Francisco neighborhoods that have the highest concentration of hospitals, in particular hospitals with the largest number of staffed beds and ICU beds. In addition, we will take into account population density of neighborhoods, because it is believed to be an important factor in the spread of contagious infectional diseases.

In first step we have collected the required data: border coordinates and population of neighborhoods, location and category of every hospital in San Francisco *(according to Foursquare categorization)*, and the numbers of hospital beds from the public database.

In the data preparation step of our project, we calculated several variables such as '**population density**', '**number of hospitals per thousand people**', '**number of staffed hospital beds per thousand people**', '**number of ICU hospital beds per thousand people**' across different neighborhoods of San Francisco - these coefficients will be used to identify neighborhoods with the highest capacity to fight COVID-19.

In third and final step we will use **k-means clustering** to create **clusters of neighborhoods** that have common characteristics in terms of two sets od variables derived at the previous step:

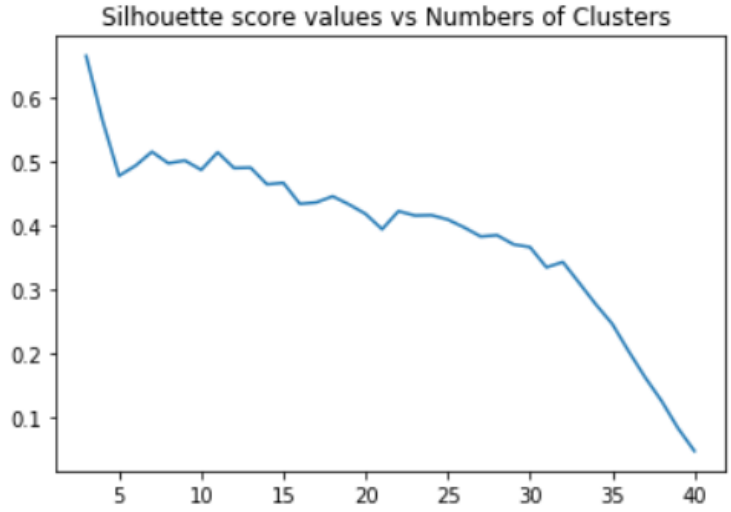
- number of hospitals per thousand people and population density
- number of staffed hospital beds and ICU beds per thousand people and population density

We will present and interpret maps of neighborhood clusters created using alternative approaches, and discuss their similarities and differences.

First, we start with defining a function that will identify the optimum number of clusters and visualize the simulation results.

Clustering based on Foursquare data

To create clusters from Foursquare data, we select and keep two relevant variables which we will focus on: **Population Density** and **Hospitals per Thousand People**. After that, we normalize data and determine the optimal number of clusters before proceeding with k-means clustering.

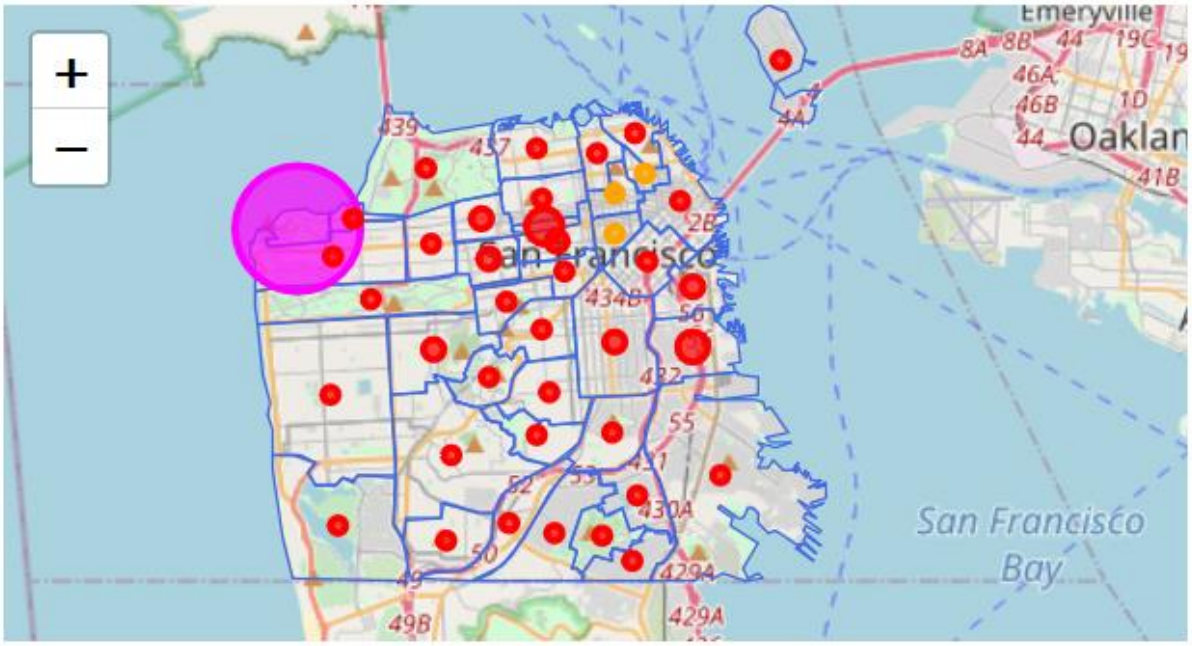



```
Optimal number of components is:    3
```

Once we added the clusters to our Foursquare dataset, we can visualize them on a map to see how the clusters look like. We have three clusters:

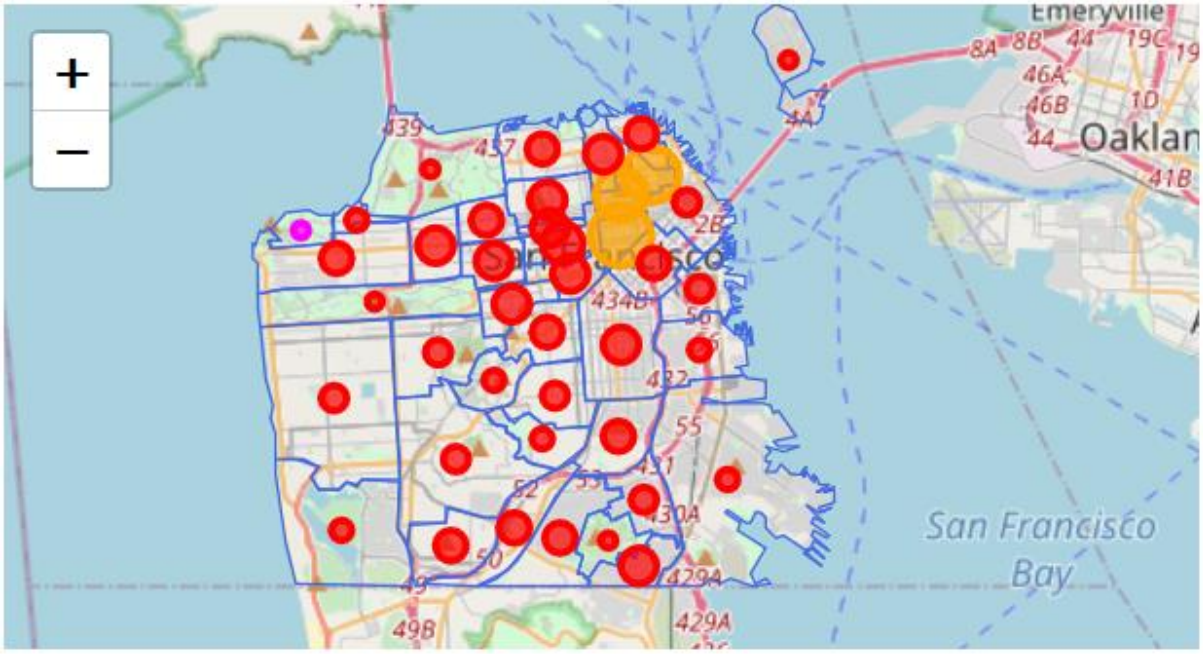
- the most populated cluster 0 is characterized by a low number of hospitals per thousand people and small to medium population density
- next, cluster 1 is actually an outlier that has a very small population of 320 people coupled with one hospital within its borders, which results in an extremely high hospital coverage
- finally, we have cluster 2 with three neighborhoods that have the highest population density

Map of San Francisco neighborhood clusters



Bubble size corresponds to Hospitals per Thousand People

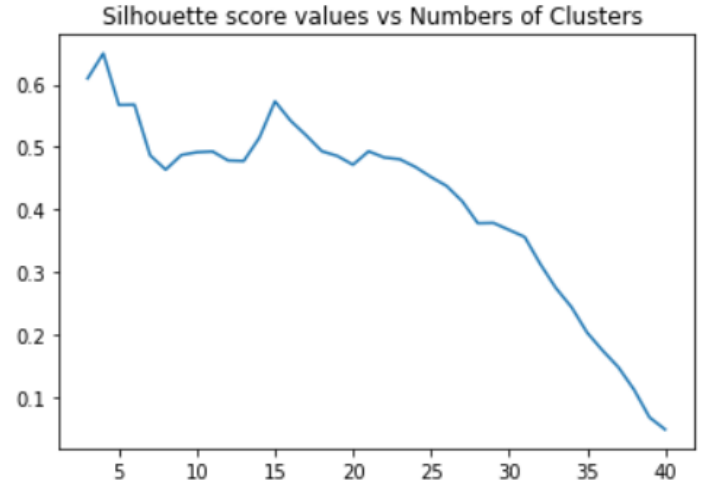
Map of San Francisco neighborhood clusters



Bubble size corresponds to Population Density

Clustering based on US hospital beds data

In order to identify clusters of neighborgoods from US hospital beds data, we focus on three variables: **Beds Per Thousand People, ICU Beds Per Thousand People, and Population Density**. As before, we normalize data and apply the function to determine the optimal number of clusters. After that, we proceed with applying the k-means clustering algorithm and visualizing the results.

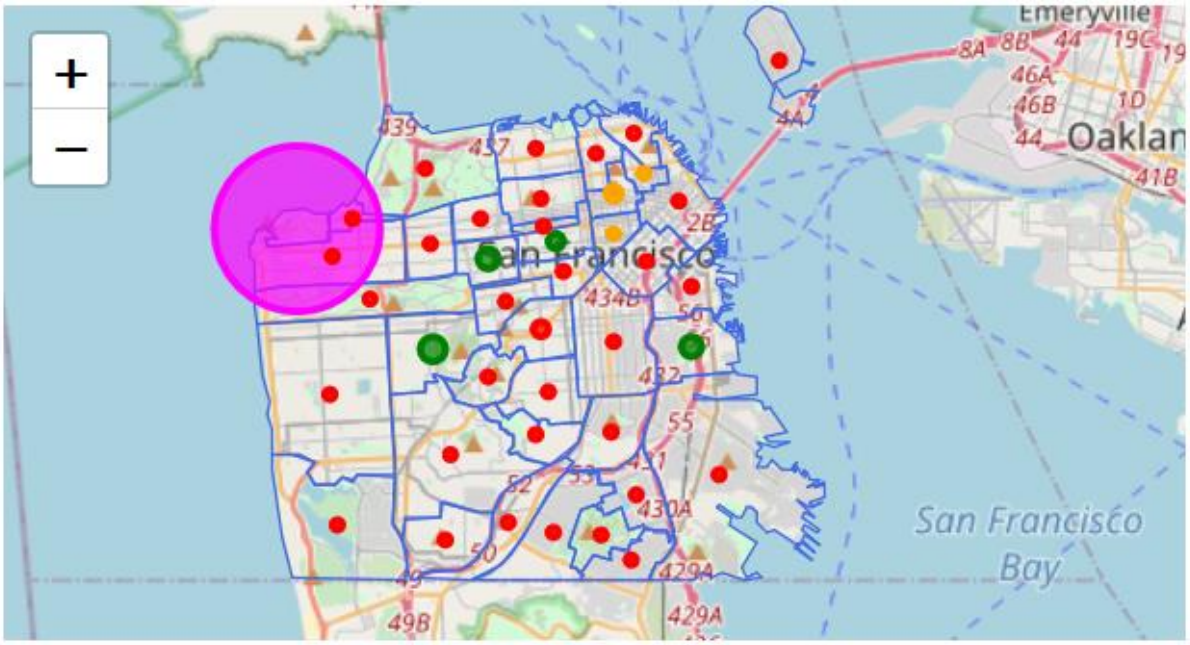



```
Optimal number of components is:    4
```

In this case, we have three independent variables, and as a result, the optimal number of clusters is four. Let's have a look at the visualization and try to interpret how the clusterization algorithm worked. We have four clusters:

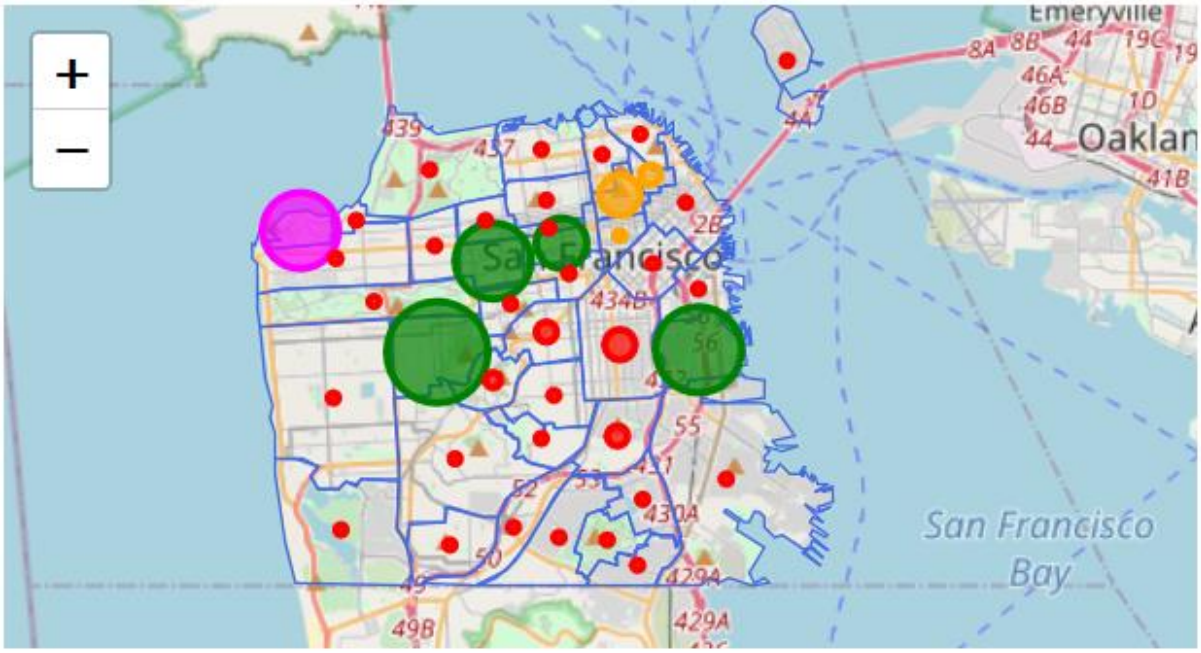
- the most populated cluster 0 is characterized by a very low number of both types of hospitals beds per thousand people and small to medium population density
- cluster 1 with four neighborhoods also has a low ratio of staffed hospital beds per thousand people, but the highest ratio of ICU beds per thousand people despite being similar to cluster 0 in terms of population density
- next, cluster 2 is the same neighborhood that was an outlier in the previous clustering. Its extremely small population density and large number of both types of hospital beds makes it the best equipped neighborhood in terms of hospital beds coverage
- finally, we have cluster 2 with three neighborhoods that have the highest population density, and low to medium beds to thousand people ratios

Map of San Francisco neighborhood clusters



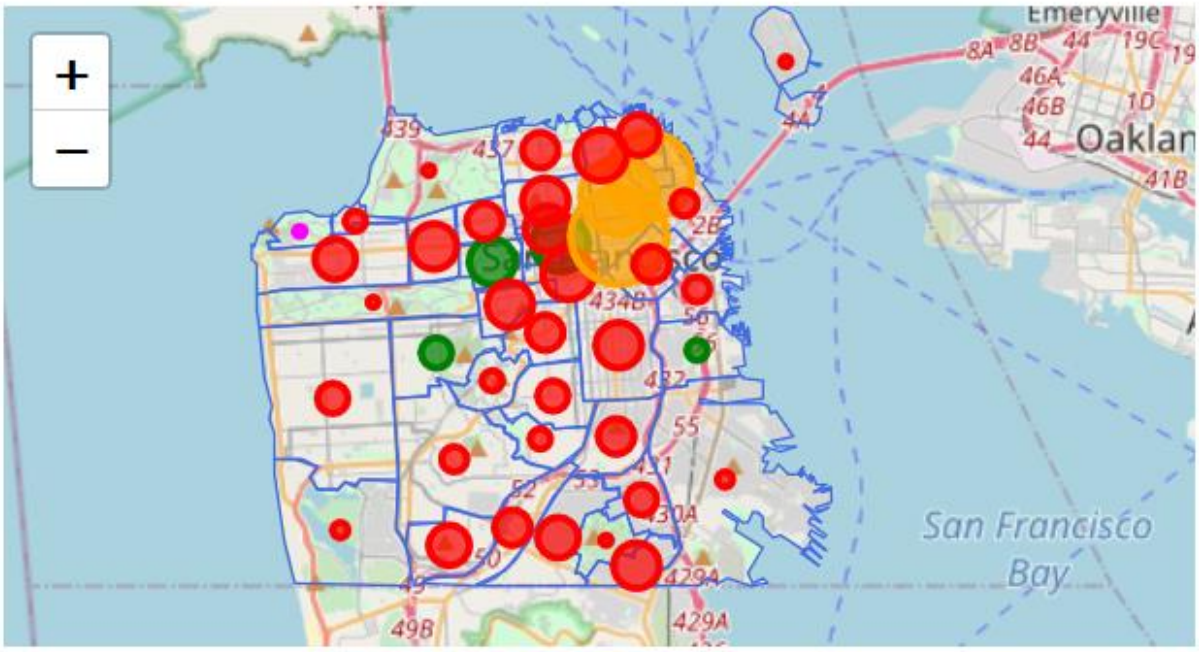
Bubble size corresponds to Beds Per Thousand People

Map of San Francisco neighborhood clusters



Bubble size corresponds to ICU Beds Per Thousand People

Map of San Francisco neighborhood clusters



Bubble size corresponds to Population Density

Results and Discussion

Our analysis shows that San Francisco neighborhoods can be divided into 3 to 4 clusters that have similar features when it comes to availability of hospitals, hospital beds and population density. Using Foursquare data, we detected the concentration of health facilities is skewed towards northern and eastern parts of the city, which corresponds to uneven distribution of population density. However, using data on exact number of staffed hospital beds, as well as ICU beds, it is possible to identify a few neighborhoods that have the highest beds per capita ratios and, at the same time, low to medium density of population. These include:

- Lincoln Park
- Inner Sunset
- Lone Mountain
- Potrero Hill
- Western Addition

The first of these, a clear outlier, could be identified on the basis of hospitals per capita data only. The other four could not be distinguished without more detailed hospital beds data.

The next cluster was correctly defined by both approaches - with Foursquare hospital data and with official US hospital beds data. This should not be surprising, as the distinctive feature in this case is that these neighborhoods are characterized by the highest population densities:

- Chinatown
- Nob Hill
- Tenderloin

Finally, the least favorable cluster of 33 neighborhoods that have either no hospitals and hospital beds, or the corresponding per capita ratios are substantially lower than in the other clusters. Subsequently, this cluster 0 with the lowest number of staffed hospital beds and ICU beds per capita should be considered as potential points of focus for future health policy interventions.

We do not suggest that these areas within cluster 0 are indeed least prepared to fight the pandemic. First of all, we do not suggest that there is a causal relationship between a hospital's proximity (in particular, its location within neighborhood's borders), and availability of health services to people living in the respective neighborhood. Second, apart from population density and hospital proximity, there are many more factors that could potentially have stronger impact on population's vulnerability to contagious diseases such as COVID-19 in terms of adaptation and mitigation capacity.

Conclusion

Purpose of this project was to apply a simple k-means clustering algorithm to identify San Francisco neighborhoods that are most (and least) prepared to counter the COVID-19 pandemic, as inferred from a small selection of potentially relevant variables.

We used visualization tools to highlight distinctive features of neighborhoods, and discussed the differences between the clusters. According to this clustering analysis, the most well-positioned neighborhood in terms of hospital beds per capita ratio and population density is **Lincoln Park**.