

San Francisco Bay Area Hospitals and Food Venues Study

IBM Capstone Project using Foursquare



Table of Contents

Table of Contents	2
1. Introduction	2
2. Data Acquisition and Cleaning	3
3. Methodology	6
4. Analysis	10
5. Results and Discussion	10
6. Conclusion	11

1. Introduction

1.1 BACKGROUND

The COVID-19 novel coronavirus hit regions of the United States in different ways. Shelter in Place orders have quarantine people in their home, except for essential workers, many of whom are healthcare workers. This project seeks to find top food venues near hospitals in the San Francisco Bay area. The Bay Area is a region comprised of nine counties. Those counties are Alameda, Contra Costa, Marin, Napa, San Francisco, .San Mateo, Santa Clara, Solano, Sonoma. These counties contain approximately 101 cities, 7.4 million inhabitants, and cover around 7,000 sq miles.¹ The region had some of the first cases of COVID and the Mayor of San Francisco issued one of the earliest shelter-in-place orders² ³. in hopes of not only keeping restaurants afloat, but also to potential connecting healthcare workers with generous folks who would give gift certificates or other gifts to healthcare workers in

¹ <https://mtc.ca.gov/about-mtc/what-mtc/nine-bay-area-counties>

² <https://www.latimes.com/california/story/2020-04-11/bay-area-coronavirus-deaths-signs-of-earlier-spread-california>

³ <https://www.sfchronicle.com/local-politics/article/Bay-Area-must-shelter-in-place-Only-15135014.php>

order to keep up morale. A few volunteer organizations in NY have done similar things for healthcare workers in New York⁴.

Kmeans clustering will also be used to cluster food venues and see if there 'at-risk' food clusters patterns—or any patterns really. With shelter-in-place many restaurants are losing income and work. A secondary analysis will see if there are food venues that are at risk of closing and recommend that the community leverage social good and support both the local economy and healthcare workers. (Note: the most reliable source of 'at-risk' came from the Yelp-Scrape. Yelp declined to give ratings or reviews for food venues that had closed. Therefore, the clusters found give a different pattern of hospitals. Please see the IBM Capstone Project notebook for more information on venues that may have not survived the past month of SIP).

1.2 Problem

This project seeks to find clusters of food venues near hospitals in the Bay Area.

1.3 Interest

This project seeks to do social and community analysis during a time of crisis.

2. Data Acquisition and Cleaning

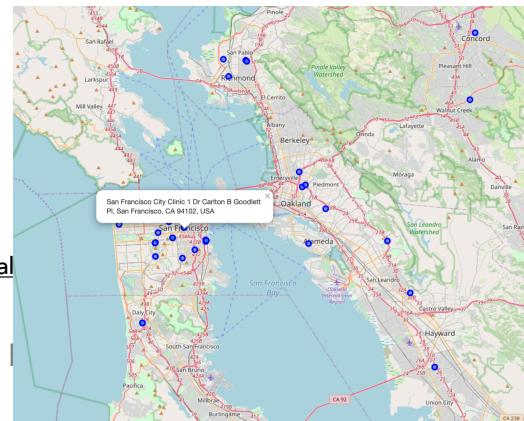
2.1 Data Sources

This project will use hospital geolocation data, Foursquare API food venue information as well as United States Census Data (for healthcare and food service workers) from the 2018 American Community Survey. Another source of data will be from the Yelp website.

DATA SOURCE A : BAY AREA HOSPITALS

For more code, please look at the "Hospital Name and Location Scrape Notebooks" on the GitHub repository. A scrape of the wikipedia on Bay Area hospitals in conjunction with

⁴ <https://nypost.com/2020/04/08/how-to-say-thank-you-to-essential/>



Google's Geocoder retrieved addresses and latitude and longitude coordinates and saved in a CSV file. Both the notebook and csv file is available on the GitHub repo.

There were redundant data points and some data exploration revealed duplicated geolocation information, even with different Hospital names. Errors were also found not the wikipedia site as hospitals from Long Beach (outside of Los Angeles, CA, over 8 hours away), Modesto, and Fresno, California were included on the scraped list. Around 48 hospitals in the bay area were left after removing duplicate address and coordinates. Another interesting note in the dataset: No hospitals in Marin County were scraped. I chose not to add them as 48 hospitals seemed a large enough number for analysis. The above folium map shows the geolocations of the original 54 hospitals webscraped. Below is sample code for the webscrape.

CODE TO GET HOSPITAL NAMES AND GEOLOCATIONS

```
2]: 1 # wget the wiki page of hospitals in the Bay area
2 !wget https://en.wikipedia.org/wiki/Category:Hospitals_in_the_San_Francisco_Bay_Area

]: 1 # Run soup and other scrape functions to get a list of hospital names in the Bay Area
2 soup = ''
3 with open('Category:Hospitals_in_the_San_Francisco_Bay_Area', 'r') as data_file:
4     for x in data_file.readlines():
5         soup += x
6
7 soup_p = bs4.BeautifulSoup(soup)
8
9 list_of_hospitals = []
10 category = []
11
12 # gets links to Oakland and SF City hospitals
13 for x in soup_p.find_all('a'):
14     if 'Categories:' in str(x.get('href')):
15         category.append(str(x.get('href')))
16     if ':' in str(x.get('href')):
17         pass
18     elif '.org' in str(x.get('href')):
19         pass
20     elif 'Main' in str(x.get('href')):
21         pass
22     else:
23         list_of_hospitals.append(x.get('href'))
24
25 # creates soup links for oakland and SFCity hospitals
26 hospst_ = []
27 for x in category[1:3]:
28     print(x)
29     new_link = 'https://en.wikipedia.org/' + x
30     hospst_.append(new_link)
31
```

DATA SOURCE B: FOUR SQUARE VENUE LOCATION DATA

Please see the notebook “IBM Coursera Capstone” for in-depth data acquisition and cleaning steps.

Overall, the code and data collected is very similar to the Manhattan and Toronto projects previously completed in this repo. More Foursquare information was explored and the url used to request information was changed. First, the radius of venues was increased to 1600 meters to account for more rural locations and second, the section parameter was set to food.

Also, the city information was collected in order to build a yelp scrape. Also, the ‘url’ category was ‘returned’ from the requests json at one point to determine delivery options, but only 11 out of 94 venues for the first hospital included grub hub delivery urls and therefore, I ended up not including it and instead, collected city data in order to build a link to collect yelp data for each venue.

```
In [213]: 1
Out[213]: 1
          name      categories    lat     lng   city      id      url
0  Trabocco Kitchen and Cocktails  Italian Restaurant  37.757138 -122.251750 Alameda  52a8c5e911d28b6a4122b614 NaN
1  Sidestreet Pho  Vietnamese Restaurant  37.762867 -122.245128 Alameda  5127d17de4b02a1718eb694 NaN
2  Julie's Coffee & Tea Shop        Café  37.761632 -122.245141 Alameda  4ae61a25f964a520b9a421e3 NaN
3  Doggy-Style Hot Dogs  Hot Dog Joint  37.761648 -122.244870 Alameda  4e6acf1e18a833989eca2e245 NaN
4  Burma Superstar  Burmese Restaurant  37.763652 -122.243411 Alameda  4b3fcf3cf964a32058a25e3 NaN

In [108]: 1 # 22 out of 94 had grubhub urls...is that enough... ack. I'll keep it.
2 nearby_venues[nearby_venues['url'].notnull()].shape
In [216]: 1 nearby_venues['categories'] = nearby_venues['categories'].str.lower()
In [217]: 1 nearby_venues['categories'].unique()
Out[217]: array(['italian restaurant', 'vietnamesee restaurant', 'café',
       'hot dog joint', 'burmese restaurant', 'middle eastern restaurant',
       'pizza place', 'american restaurant',
       'eastern european restaurant', 'mexican restaurant',
       'sushi restaurant', 'diner', 'afghan restaurant',
       'new american restaurant', 'iranian restaurant',
       'bakery', 'burger joint', 'sandwich place', 'noodle house', 'burrito place',
       'asian restaurant', 'polk place', 'salad place', 'thai restaurant',
       'chinese restaurant', 'taco place', 'cuban restaurant',
       'ethiopian restaurant', 'breakfast spot', 'bagel shop',
       'korean restaurant', 'japanese restaurant', 'fast food restaurant',
       'bbq joint', 'food truck', 'food', 'japanese curry restaurant'],
      dtype=object)

In [103]: 1 nearby_venues.shape
Out[103]: (94, 6)

In [90]: 1 nearby_venues.iloc[[41, 69, 71]]
Out[90]: 1
          name      categories    lat     lng   city      id
41  Noah's Bagels  bagel shop  37.757246 -122.253706 4f32113919833175d60d433e
69  House of Bagels  bagel shop  37.764104 -122.242950 4aabdf77964a520975a20e3
71  Bagel Street Cafe  bagel shop  37.756982 -122.252732 54c82f5a498e05697a006089
```

Data was stored in pandas data frames that made it easy to merge for analysis purposes later on.

DATA SOURCE C: YELP WEB-SCRAPE DATA

Using venue data from FOURSQUARE, I made a yelp scrape function that scraped ‘ratings’ and number of ‘reviews’ for each venue off the corresponding yelp site. Some regex was needed to get good data. If ‘nan’ values were returned, there was a high correlation to that venue being closed.

DATA SOURCE D: IPUMS 2018 CALIFORNIA American COMMUNITY SURVEY DATA

A 1% sample of US Census data for the State of California was retrieved from IPUMS USA, University of Minnesota, www.ipums.org. The percentage of food service workers, healthcare workers and top jobs for each of the bay counties came from the 2018 ACS survey.

DATA SOURCE E: COVID-19 Data For the final folium cholorpleth map, a repo was forked from the NY TIMES Covid-19 repository [repository on Github](#).

3. Methodology

This project had three parts to it. First, the hospital scrape. Second, the food venues and K-Means clustering analysis; third, Venue and Hospital Analysis using Covid Data, food venue ratings and categories, and the K-Means clustering data.

3.1 The Hospital Scrape

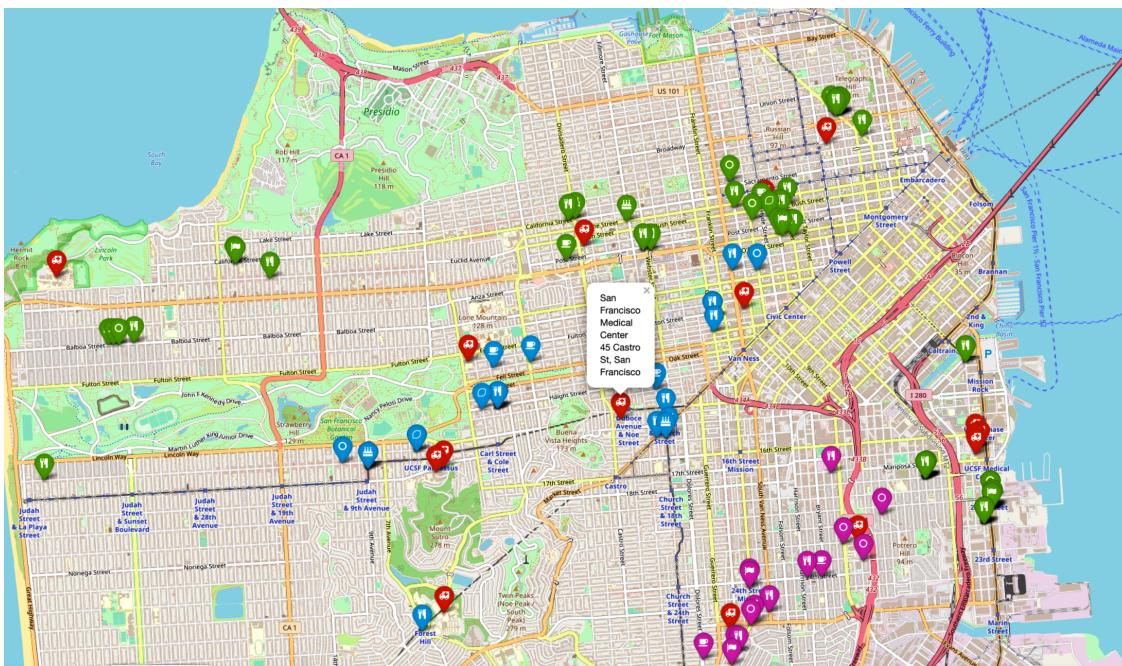
Google geolocation data was used to get the addressed and coordinates of 48 hospitals in the Bay Area.

3.2 Foursquare API and Food Venues

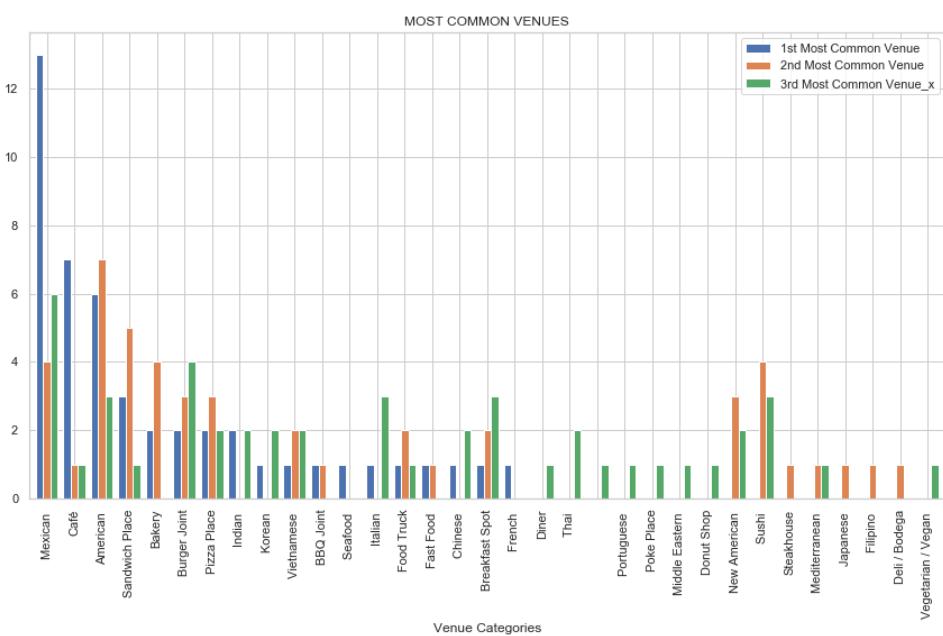
The ultimate goal of this project was to find food venues near major hospitals in the Bay Area. A radius of 1 mile was used (1600 meters) to account for more rural counties in the area. Unfortunately, in the densely packed counties in the area, this meant many duplicate venues and information. A secondary attribute was then considered : yelp ratings and the number of yelp reviews. Using the information from Foursquare API data, yelp links were created and then used to scrape this data off of each food venues' website.

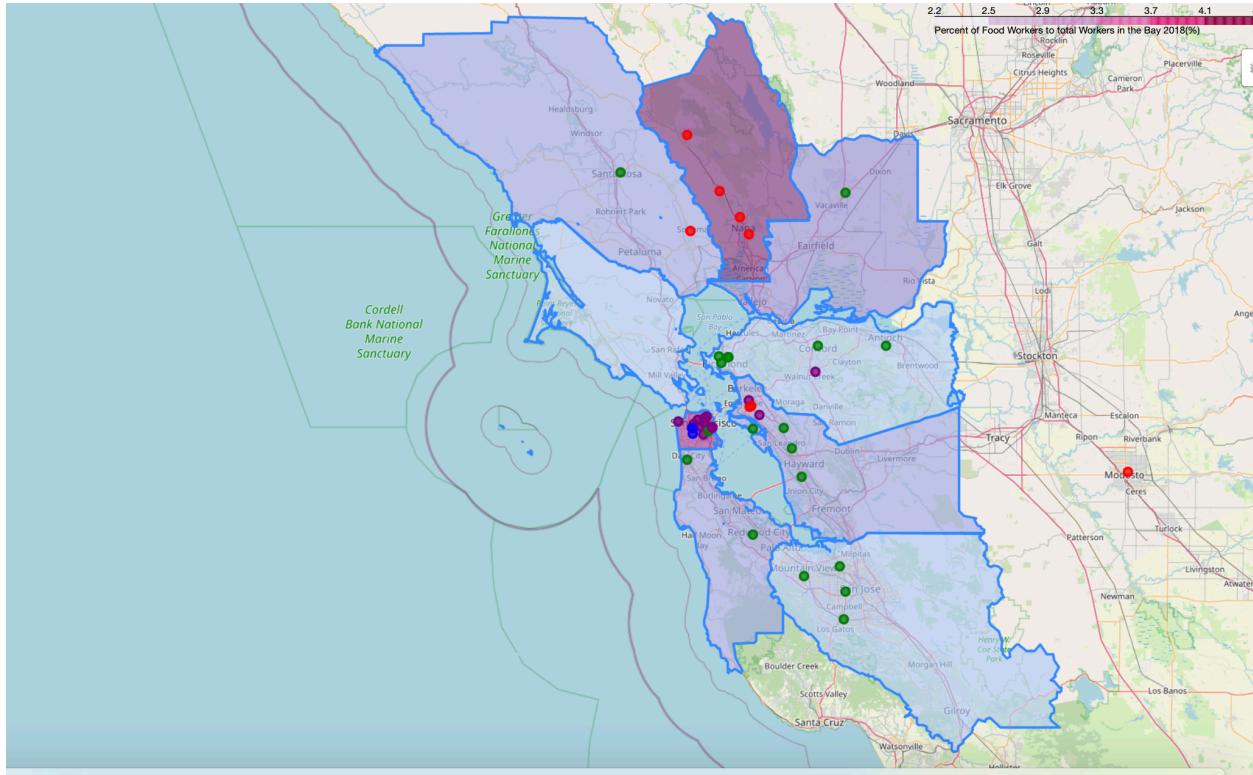
As I mentioned above, if the scrape did not work—even after minor regex workarounds to double check names of venues and locations, a random sample of venues returned as CLOSED and were dropped from the data set. Approximately 1200 venues remained after dropping venues without yelp ratings and reviews. These

venues were then used to create clusters based on food venue types. Labels for each hospital were then added to the data frame. Further analysis on the top three venues was completed (see charts below).

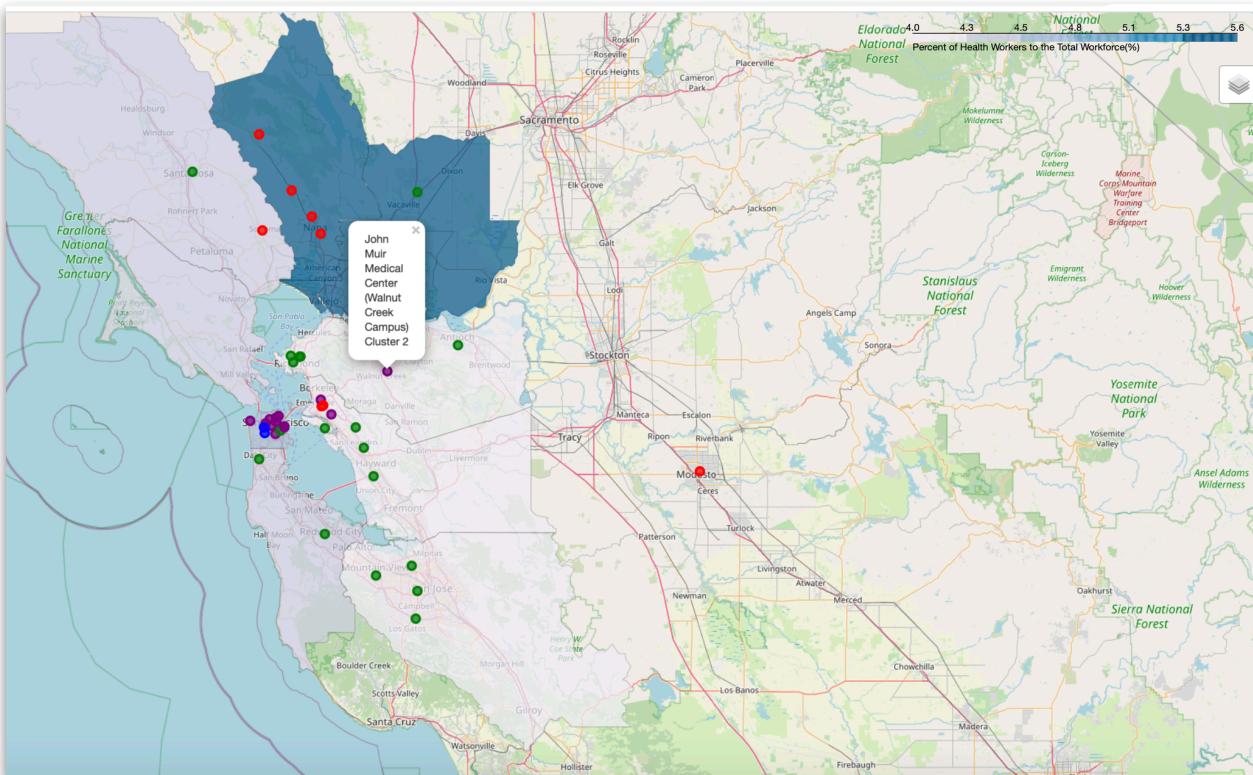


K-Means Clustering of Food Venues around Hospitals in San Francisco





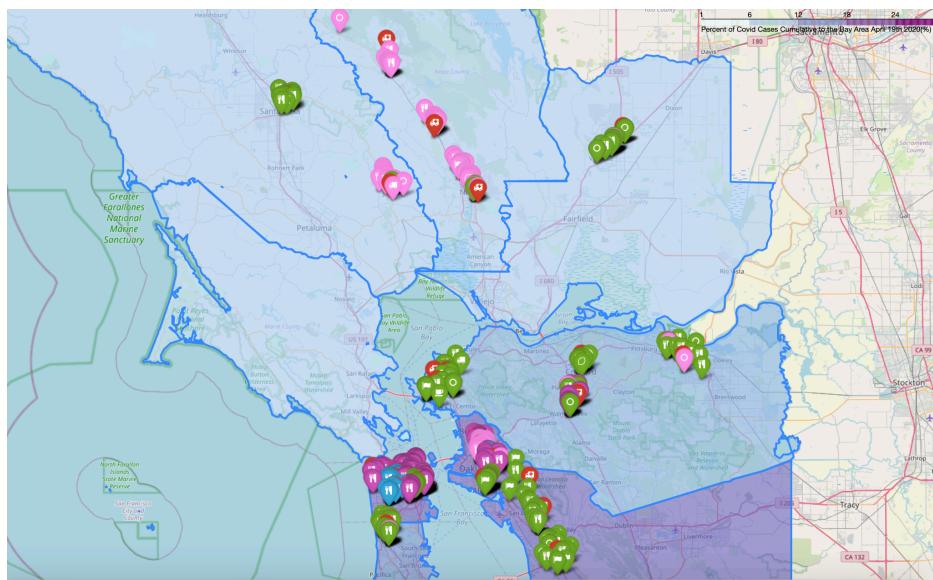
Choropleth of Food Workers (as a percentage of the total workforce) in Bay Area



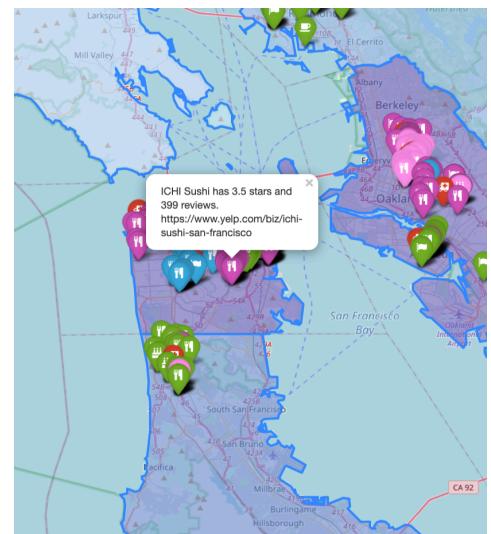
Choropleth of HealthWorkers (as a percentage of the total workforce) in the Bay Area

3.3 Gift Card Recommendations

The third notebook contains more detailed work on determining what food venues should be included in the final Choropleth. The data was filtered for Yelp reviews over 3.5 and customized icons were made for loose categories of food options. For example, a birthday cake symbol represents nearby bakeries. Utensils represent restaurants; a circle: burger, hot-dog, and pizza places.



The choropleth map contains the yelp information as well as the yelp url in case a user would like to order from that restaurant. Also, hovering over each county gives more information on the county and number of COVID cases. No corollary analysis has been done, it is more for a representation of the number of cases in the area.



4. Analysis

There are four primary clusters. The Blue cluster (cluster label '0') contains a lot of sandwich and 'sushi restaurants, and there are three hospitals in this cluster. The Green cluster (cluster label '1') that primarily has ethnic restaurants in its vicinity, from Vietnamese, Israeli, to Mediterranean and Filipino. There are also a large amount of 'fast food' restaurants in this cluster. The Purple/dark Pink cluster (cluster label '2') is mostly cafes, bakeries, delis, and seafood places while the light pink cluster contains 'New American' restaurants and more food venues labeled 'restaurants'. Overall, Mexican food venues were in all the clusters as shown above, was a top venue choice overall.

There did not seem to be much correlation to where food service workers lived in the Bay Area and particular food venues. Same issue with health workers. Most food service workers work in the county of Napa and the highest percentage of health workers live in Solano. Neither of those counties have the highest rates of COVID-19 infections. However, further analysis would confirm an assumption that two counties may unique adverse effects by economic shutdown and COVID-19.

Given the final choropleth, there are ample choices of quality food venues in the Bay Area.

5. Results and Discussion

This project was broad in scope and is best viewed as a preliminary analysis that points to further research to determine more economic and social effects. Enough groundwork was done for creation of an app that recommends food nearby hospitals. The next steps would be to have a website interface where someone could pick a hospital and send doctors and health care workers gift cards of nearby venues.

6. Conclusion

This project took on a lot and has plenty of areas for further development and focus. It is a basic snapshot of the area and food and healthcare. Some further project ideas including the ratings and number of reviews in the clustering models, weighting venues by the hours open (and aligned to shifts at hospitals), including more distributions on employment and occupations in the counties, and including transit information (do most health care workers live in the counties they work in).

My next step is to take the information and post it online somewhere in the case that good-hearted individuals could send gift cards or takeout to hospitals in the area to cheer on the coronavirus efforts (and help restaurants and food workers in the area).