Sarah Hudspeth
April 30, 2020

San Francisco Bay Area Hospitals and Food Venues Study

IBM Capstone Project using Foursquare

Table of Contents

1. Introduction

1.1 BACKGROUND

The COVID-19 novel coronavirus hit regions of the United States in different ways. Shelter in Place orders have quarantined 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: 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 orders[2] [3]. In the hopes of not only keeping restaurants afloat, but also with the purpose to potentially connect healthcare workers with generous folks who would give gift certificates or other donations, this project will analyze food venues in the San Francisco Bay Area and organize hospital information. A few volunteer organizations in NY have done similar things for

[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

healthcare workers in New York[4] and allow people to send food and gift certificates to those fighting Covid-19 spread.

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 of 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

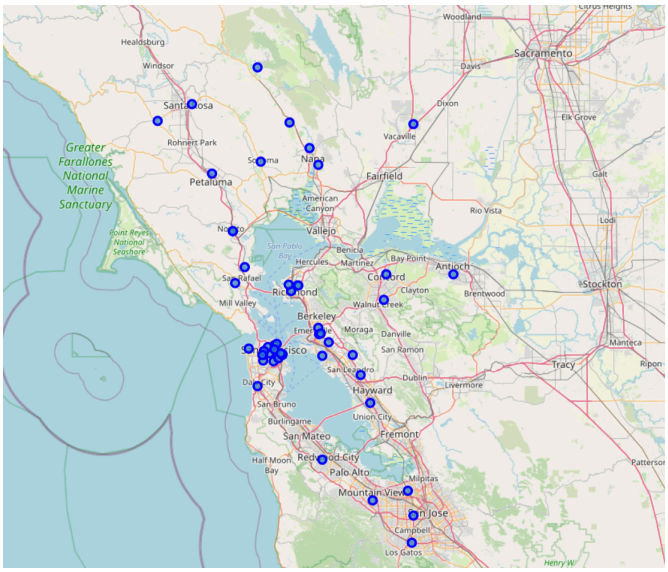
This project will use hospital geolocation data, Foursquare API food venue information as well as COVID-19 county-level data forked from the NYTIMES GitHub repo. Another source of data will be from the Yelp websites of each food venue.

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 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.

[4] https://nypost.com/2020/04/08/how-to-say-thank-you-to-essential-workers-during-covid-19/

Sarah Hudspeth
April 30, 2020



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 went back added them using a yelp website for hospitals in Marin County. The above folium map shows the geolocations of the final 51 hospitals used in the data. Below is sample code for the web-scrape of non-Marin County hospitals.

CODE TO GET HOSPITAL NAMES AND GEOLOCATIONS

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

]:
```
# Run soup and other scrape functions to get a list of hospital names in the Bay Area
soup = ' '
with open('Category:Hospitals_in_the_San_Francisco_Bay_Area', 'r') as data_file:
          for x in data_file.readlines():
            soup += x

soup_p = bs4.BeautifulSoup(soup)

list_of_hospitals = []
category = []

# gets links to Oakland and SF City hospitals
for x in soup_p.find_all('a'):
    if 'Categories:' in str(x.get('href')):
        category.append(str(x.get('href')))
    if ':' in str(x.get('href')):
        pass
    elif '.org' in str(x.get('href')):
        pass
    elif 'Main' in str(x.get('href')):
        pass
    else:
        list_of_hospitals.append(x.get('href'))

# creates soup links for oakland and SFCity hospitals
hospt_ = []
for x in category[1:3]:
    print(x)
    new_link = 'https://en.wikipedia.org/' + x
    hospt_.append(new_link)
```

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]:

	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	5127d17de4b02a1718eab694	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	4e6acf1e18a83989eca2e245	NaN
4	Burma Superstar	Burmese Restaurant	37.763652	-122.243411	Alameda	4b3fcf3cf964a52058af25e3	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', 'vietnamese 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', 'german restaurant', 'bakery',
                 'burger joint', 'sandwich place', 'noodle house', 'burrito place',
                 'asian restaurant', 'poke place', 'salad place', 'thai restaurant',
                 'chinese restaurant', 'taco place', 'cuban restaurant',
                 'wings joint', 'mediterranean restaurant', 'bagel shop',
                 'ethiopian restaurant', 'breakfast spot', 'fried chicken joint',
                 '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]:

	name	categories	lat	lng	id
41	Noah's Bagels	bagel shop	37.757246	-122.253706	4f321139198331 75d60d433e
69	House of Bagels	bagel shop	37.764104	-122.242950	4aabdfe7f964a520975a20e3
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.

Foursquare sent back 1530 food venues, 1094 being unique food venues. Research was not done to see how many venues were 'local' chains. Further analysis was done when the venues were combined with yelp data.

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. 81 venues ended up being dropped due to bad foursquare data or restaurant closings due to cover (a quick link check showed 'Closed Status' on yelp). 81 out of 1530 venues is around a 5% drop rate. Pretty good for the area and given the number of venues that consumers in the area support!

DATA SOURCE D: COVID-19 Data For the final folium cholorpleth map, a repo was forked from the NY TIMES Covid-19 repository repository on Github. A quick filter for Bay Area counties, the last scrape date (April 26th), number of cases, deaths, and sum of total cases in the area was stored in a dataframe.

3. Methodology

This project has four parts to it. First part is the hospital information scrape. Second part is the food venues via Foursquare API, third is the K-Means clustering analysis and visualizations; Fourth part includes the venue and hospital analysis using Covid-19 Data, food venue ratings, and the K-Means clustering data.

3.1 The Hospital Scrape

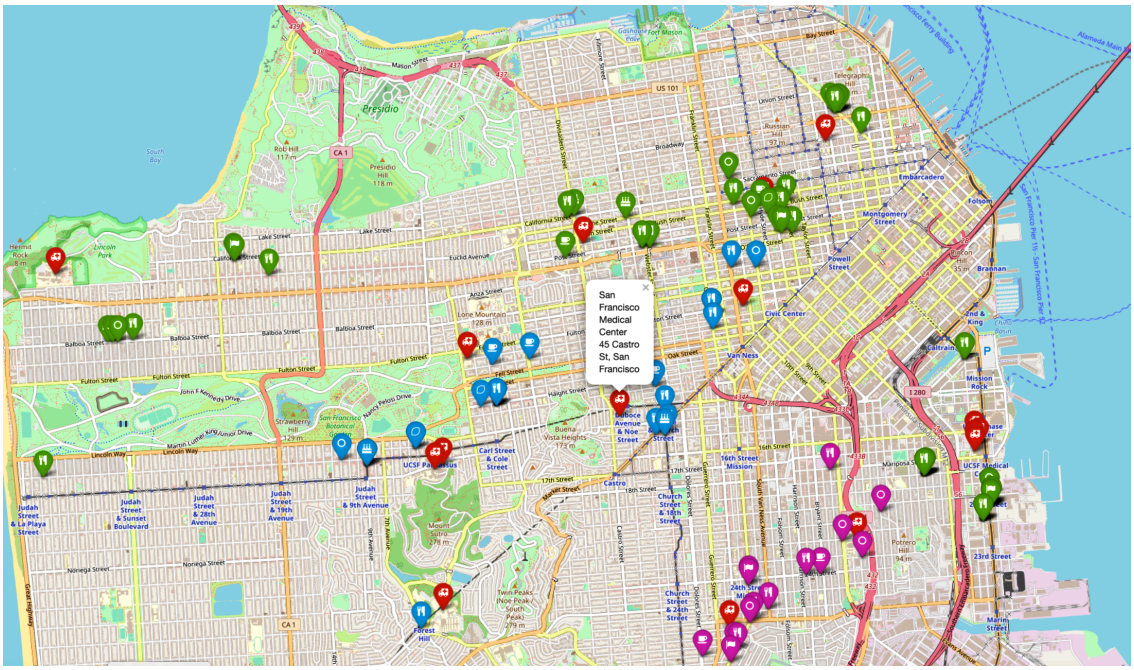
Google geolocation data was used to get the addressed and coordinates of 51 hospitals in the Bay Area.

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, and it was removed on a second go around to consolidate data. 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 1350 venues remained after dropping venues without yelp ratings and reviews. The

mean rating of these venues is 3.9 and the mean number of reviews is 674. Overall, the Bay Area has great food options with high number of customers. The median values were respectively: 4 and 427. These venues were then used to create clusters based on food venue types. Labels for each hospital were then added to the data frame (see images below).



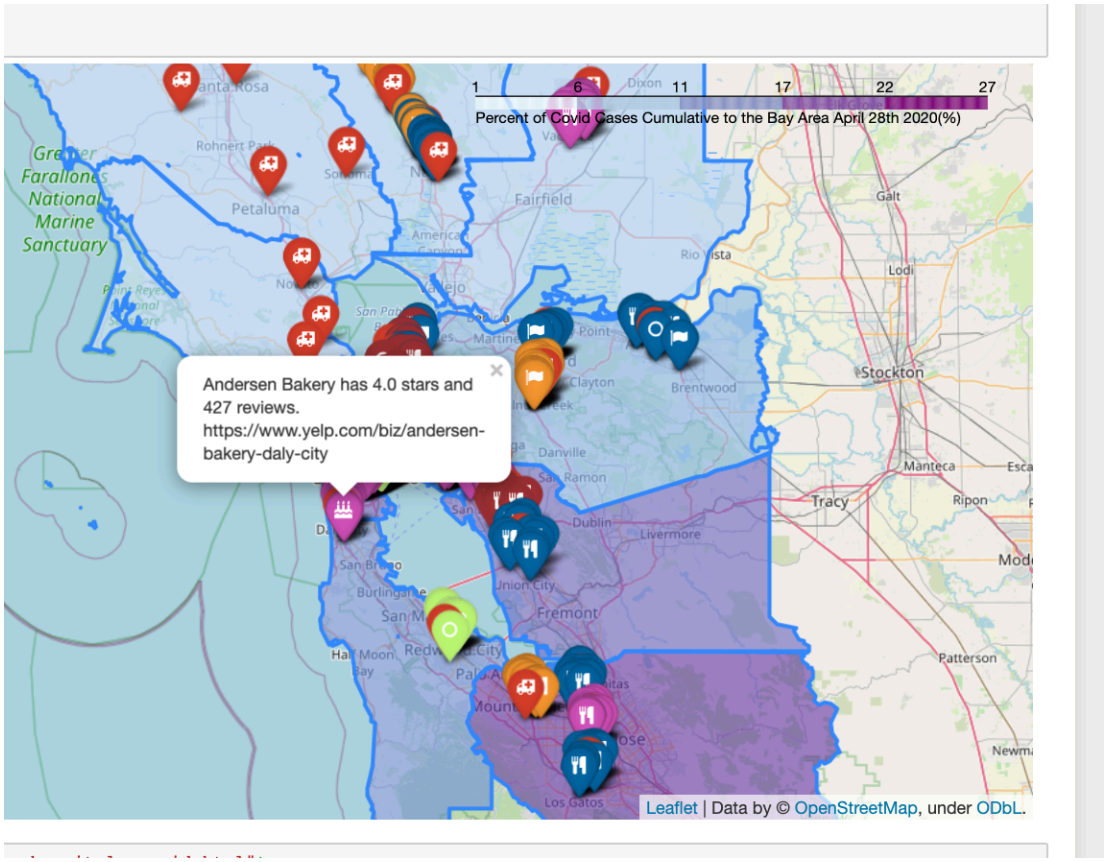
K-Means Clustering of Food Venues around Hospitals in San Francisco

3.3 Gift Card Recommendations

The third notebook contains more detailed work on determining what food venues should be included in the final choropleth map. The data was filtered for Yelp reviews with a rating of 3.5 stars and over and more than 250 posted reviews. Thus the number of venues plotted in the final folium map was 988.

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 below 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

Kmeans Clustering

To determine the number of clusters, k-means inertia scores were plotted in a graph versus number of clusters. The graph had a 'slight', and I mean, slight' elbow at 7 clusters. I believe ultimate the clusters would have arrived at an ideal clustering score equal to the number of food categories. I think more data and features to the clustering would help in making better

clusters, but there was not more time for more analysis. Therefore, there are 7 primary clusters.

The Dark Blue cluster (cluster label '0') most popular venues include Mexican, Vietnamese, Italian restaurants (possibly, dinner/evening spots) and had 13 different hospitals in its cluster. The Green cluster (label '1'), contained bakeries and breakfast spots as some of their most common venues and could may be considered more of an 'early morning' locations. There were three different hospitals in this cluster.

The Orange cluster (label '2') had 8 different hospitals and the top venues in its cluster including vegetarian options, New American restaurants, bakeries, and cafes. The Light Green cluster (label '3') could be considered a 'lunch' locale as most of the top most common venues were cafes, pizza places, and deli-bodegas. There were 9 different hospitals that formed this cluster. The second most populous (of hospitals) cluster was the purple cluster (label '4'). There were a lot of Mexican, Middle Eastern, Mediterranean. And Vietnamese restaurants in this cluster as the most common venues.

The sixth cluster (cluster label '5') had six hospitals and American, Mac&Cheese, and Korean were the most common venues in this cluster.

The seventh cluster contained one hospital and had VERY high-end, world renown New American restaurants as its top venues.

There did not seem to be much of a pattern of clustering due to location. But one could say that most Blue and Purple clusters were not in the downtown area of San Francisco City. San Francisco county had one purple cluster and no dark blue. Green and light green clusters also stayed in this more urban part of the area versus other colored clusters. However, any more correlations seems weak to make without further research and data.

Given the final choropleth, there are ample choices of quality food venues in the Bay Area that have a large number of reviews and are well rated. There did not seem to be a pattern of certain food venues where COVID-19 had increased rates of infection.

5. Results and Discussion

This project was broad in scope and is best viewed as a preliminary analysis that discovers areas for further research. 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 where Covid-19 has spread. 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).