Applied Data Science Capstone

Final Project Report

By Ted Lin

May 27, 2020

INTRODUCTION

The current COVID-19 pandemic, also known as coronavirus pandemic, has been ravaging the world recently. As of this writing, over 5.6 million cases worldwide have been confirmed with approximately 355,000 death. In U.S., nearly 1.7 million cases have been confirmed with over 100,000 deaths[1].

Despite being a diverse metropolitan city with its economy supported by industries such as technologies, conventions, and tourism, , the city/county of San Francisco only has slightly over 2,400 confirmed cases with 40 deaths[2]. Early decisions by the city to close non-essential businesses and impose social distancing had played a major factor in its relatively low case counts.

This analysis attempts to see if any correlation can be found between COVID-19 case counts and Foursquare location data in San Francisco. Specifically, we will look at estimate COVID-19 case counts per 10K population by zip code, and attempt to correlate those with venue found within those zip codes from Foursquare.

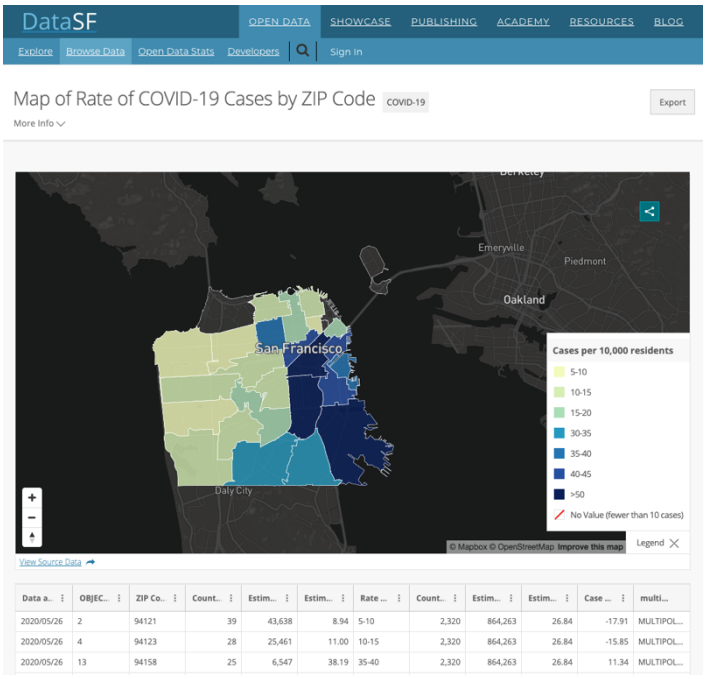
If any correlation can be found, audiences such as city/county administrators, public health officials, and general public can expand on the finding to learn how to utilize such correlation to predict and prevent future outbreak of diseases and other public health issues.

DATA

Data from this analysis will mainly come from two main sources.

First source is “Rate of COVID-19 Cases by Census ZIP Code Tabulation Area” published by DataSF (https://datasf.org/). DataSF is part of City and County of San Francisco. A description of the data set, including how the data set is created, how it is updated, definition of data field, and preview of the data set can be found on it website (see link here).

The website hosts a well-presented map visualization with data table (see image to the right; to see the actual website, see link here). Because this page is dynamically updated with embedded JavaScript, it was a hard to use Python web-scraping tool (BeautifulSoup) to gather the data table. Fortunately, the website provides links to download the data set in various format, including csv.



Data a...	OBJEC...	ZIP Co...	Count...	Estim...	Estim...	Rate ...	Count...	Estim...	Estim...	Case ...	multi...
2020/05/26	2	94121	39	43,638	8.94	5-10	2,320	864,263	26.84	-17.91	MULTIPOL...
2020/05/26	4	94123	28	25,461	11.00	10-15	2,320	864,263	26.84	-15.85	MULTIPOL...
2020/05/26	13	94158	25	6,547	38.19	35-40	2,320	864,263	26.84	11.34	MULTIPOL...

The data sets includes 27 rows and 12 columns. The 27 rows represent 27 zip codes in San Francisco, while the 12 columns contain data such as zip codes, count of confirmed cases (in a zip code), and count of San Francisco confirmed cases (for the entire city/county). We will mainly utilized two columns of data:

- Zip code is the postal code assigned by United States Postal Services (USPS) and represents geographical boundaries.
- Rate groups are segmentation of estimated cases per 10K population; for example, if the estimated cases per 10K population is 8.54, it will be categorized as “5-10” in rate groups columns. The segmentation is in increment of 5 (“0-5”, “5-10”, …etc.) with any estimated case count per 10K population greater than 50 categorized in “>50” segment. (*Estimated* case count per 10K population is used in the data set as the population by zip code data is from 2017.)

The zip code data will be augmented further by adding:

- Neighborhood name(s) – we will web-scrap San Francisco neighborhood name by zip code and append them to the data set.
- Geo-coordinates – latitude and longitude of each zip code will be extracted using Python’s GeoPy library and appended to the data set.

The second source of data is Foursquare. We will use API codes covered in earlier modules in this class. 7 columns of data are extracted from the JSON file downloaded by the API codes, however, we will mainly focus on the two columns:

- Neighborhood – neighborhood name will represent zip code.
- Venue Category – category of venue found within the zip code’s geographical area.

We will then use data wrangling techniques to clean and format these data sets and make them ready for preliminary data exploration, data visualization, and, eventually, data modeling.

REFERENECES

[1] Johns Hopkins University Corona Resource Center (https://coronavirus.jhu.edu/us-ma)

[2] San Francisco Department of Public Health (https://www.sfdph.org/dph/alerts/coronavirus.asp)