Common Venues and Risk of COVID-19 in Districts of Bangkok

Sathianphong Phongsathian

January 17, 2021

1. Introduction

1.1 Background

Coronavirus disease (COVID-19) is an infectious disease caused by a newly discovered coronavirus. In Thailand, the city which has most cases of COVID-19 is Bangkok - the capital and most populous city of Thailand. The city divided into 50 districts in total.

1.2 Problem

Data needed to find the risk of COVID-19 in districts of Bangkok might include the coordinates data of each district in Bangkok, categories of venues in that district, total number of COVID-19 cases in that district. This project aims to find whether the common venues associate with the number of COVID-19 cases in each district of Bangkok.

1.3 Interest

The individual who live in Bangkok and those who in charge of health/disease control policies manager would be interested in the association of venues and risk of COVID-19 transmission for daily life planning, or invent impactful disease control policy

2. Data acquisition and cleaning

2.1 Data sources

Each COVID-19 patient data in Thailand can be requested via the official API from the Department of Disease Control of Thailand here. Coordinates data of each district of Bangkok can be found in Open Government Data of Thailand here. List of districts of Bangkok can be found in the wikipedia webpage here, with these two datasets we can create each district coordinates. Foursquare data can be requested via Foursquare API. Lastly I used GeoJSON file of districts in Bangkok which can be found here, for visualize choropleth map

2.2 Data cleaning

Data downloaded or scraped from multiple sources were combined into two tables, which will be combined with GeoJSON data to visualize the choropleth map.

There are several problems with the datasets. First, there were a lot of typos/missing values in the district column of COVID-19 patient data. I used a list of district names from the list of districts of Bangkok in the wikipedia webpage to match the COVID-19 patient data. Then find the summation of total COVID-19 cases of each district.

Second, the district name in GeoJSON data has a prefix 'เขต' (District in Thai) in keyname of the data. I had to add the string 'เขต' to every district name in total COVID-19 cases table.

Third, the coordinates data in the wikipedia webpage are not well represent the center of each district. I switched to use another dataset which is latitude and longitude data of each sub-district instead. I had to find the centroid of each sub-district to approximate the coordinate data of each district in Bangkok.

2.3 Data preprocessing

After data cleaning, there were 49 records of each district name with the number of total COVID-19 cases in the COVID-19 data, and 2,581 rows of venues data in Bangkok and 7 columns which include district name and venue category.

For the venues data, I preprocessed by one-hot coding each nearby venue into categories which result in 2,581 rows of venue data and 249 unique venue categories. Then find the mean of each district result in 50 rows of each district and 249 frequency of venue categories in each district. After that, find the 10 most common venue categories in each district result in 50 rows of each district and 11 columns included district names and 10 most common venue categories

3. Classification Modeling

3.1 Cluster districts

Using the k-means clustering method, first I find the optimal k value by using the elbow method.

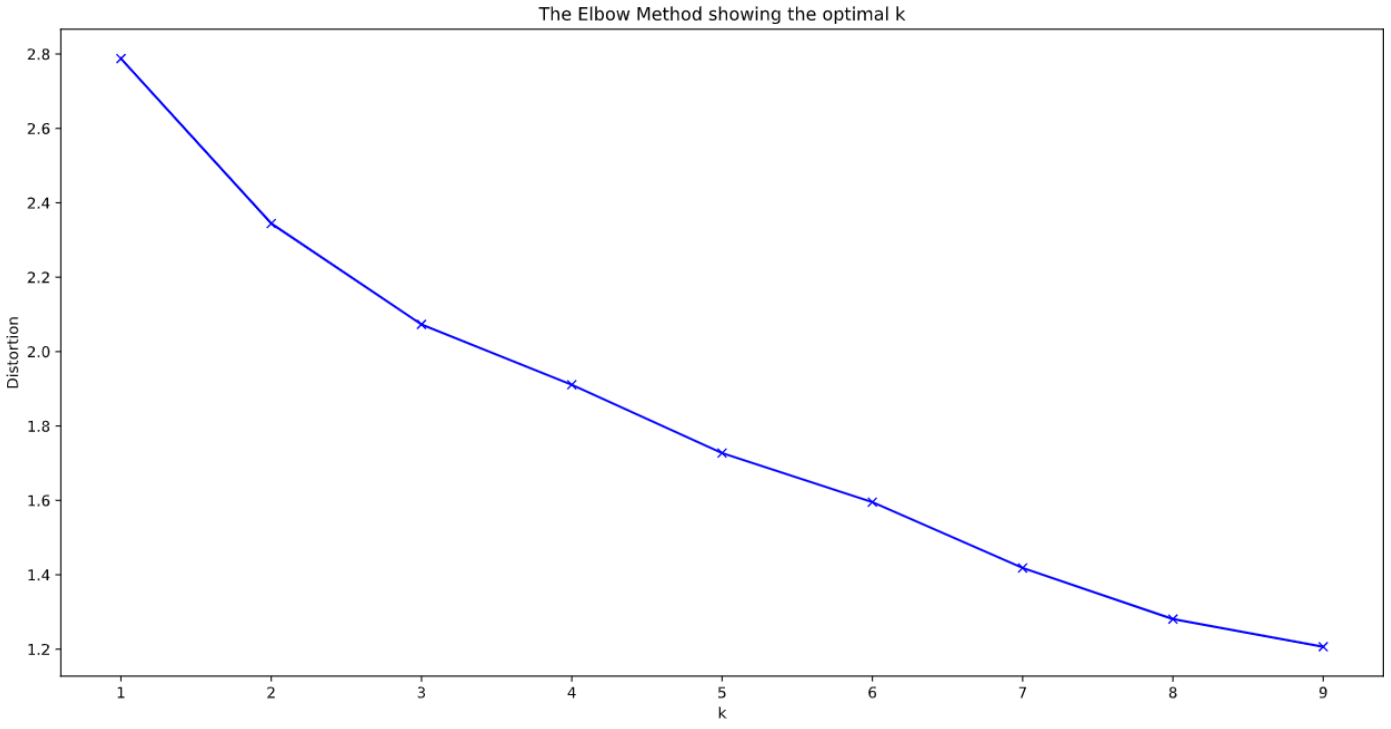

Figure 1. Line plot of distortion values of k-means model of each k values from 1 to 9

I used the optimal k = 3 to create the model. Then labeled each district with this model and merged with earlier data.

4. Data Analysis and Visualizations

Combined all the data we got to create the visualization.

4.1 Total COVID-19 cases in each district of Bangkok with colored cluster labels

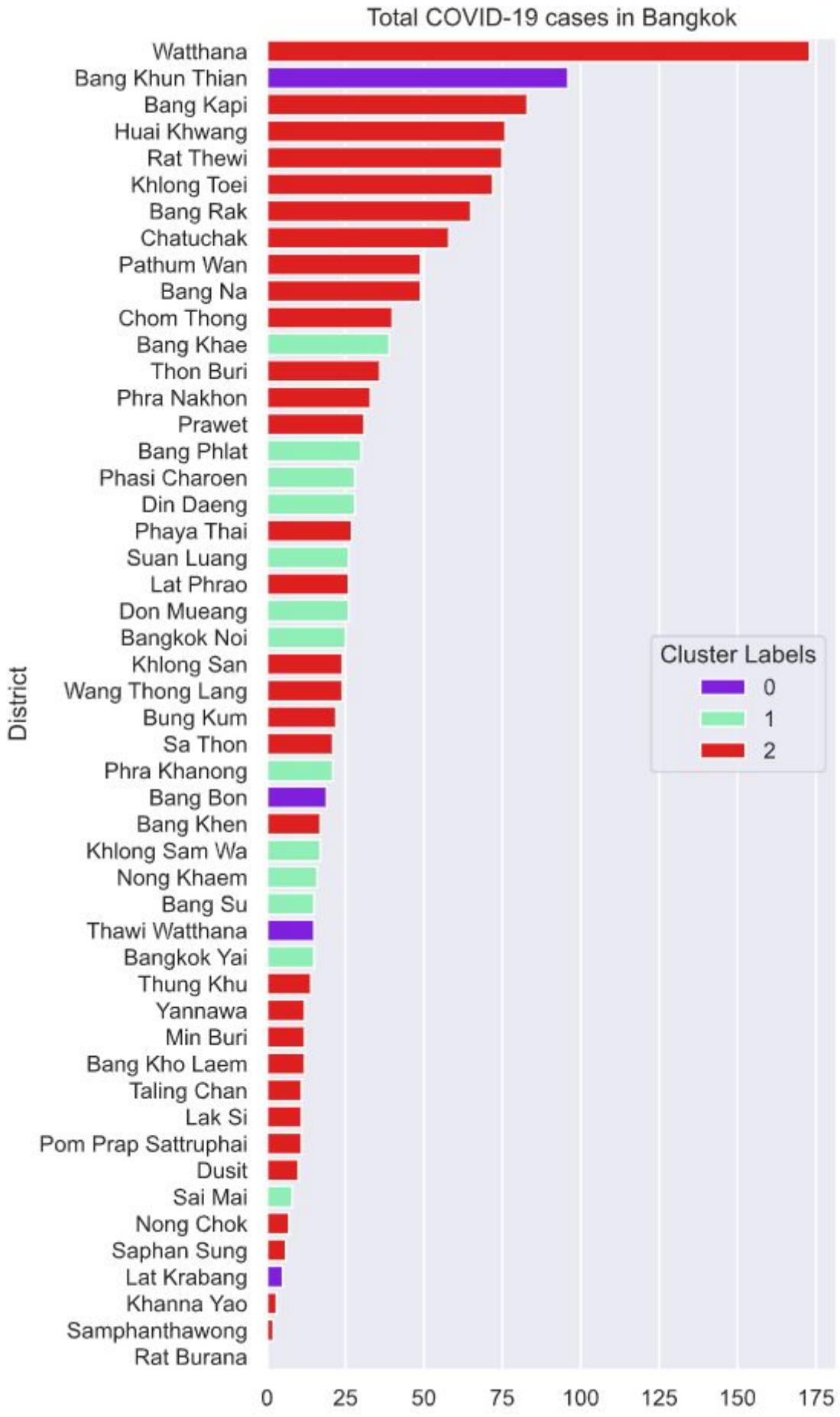

Figure 2. Horizontal bar plot of total number of COVID-19 cases in each district of Bangkok with colored cluster labels

4.2 Bangkok map shows total number of COVID-19 cases with colored cluster labels

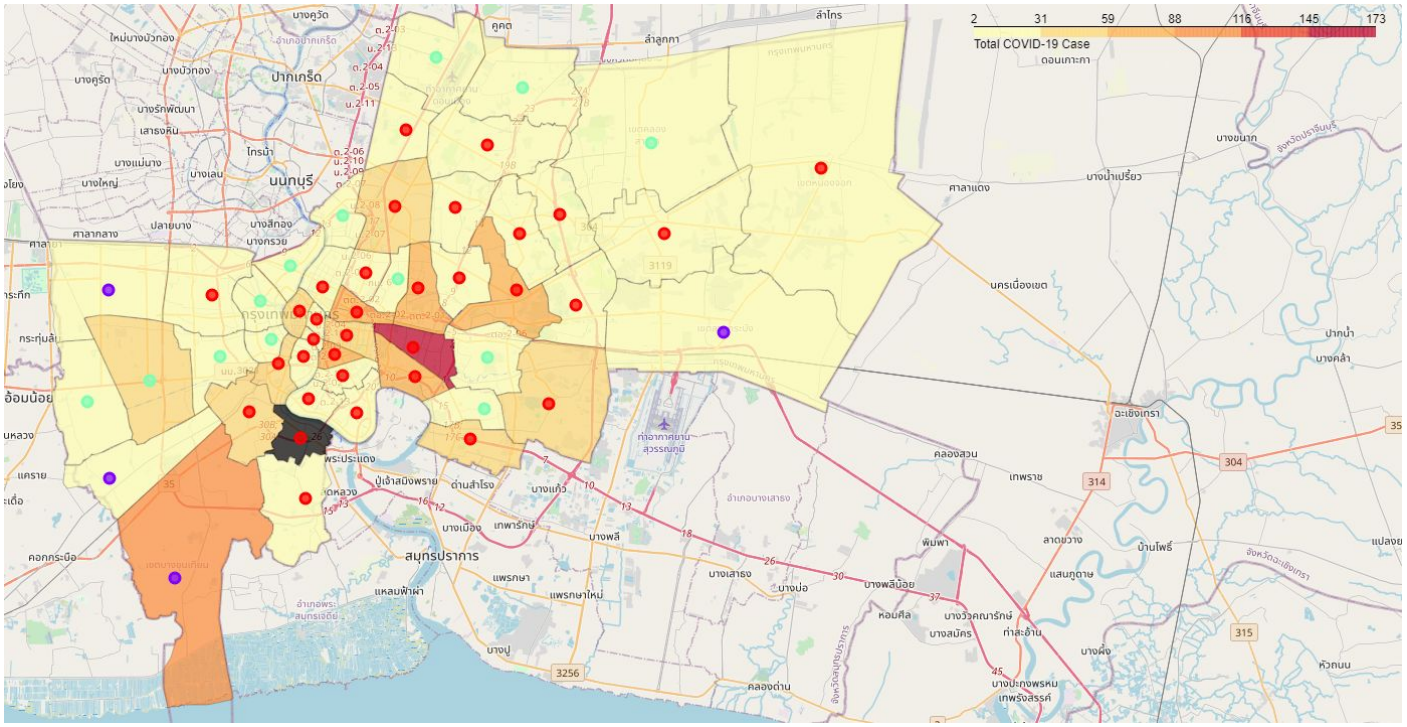

Figure 3. Bangkok choropleth map shows total number of COVID-19 cases with yellow to red color theme with colored cluster labels; Cluster 0 = Purple, 1 = Mint green, 3 = Red

4.3 Find the high risk venue categories

From the data earlier, list all 10 most common venue categories of the top 5 districts of most total COVID-19 cases and same for the least 5 districts. Then find the venue categories which include in top 5 but not in least 5. Results are as follow :

- Flea Market
- Floating Market
- Flower Shop
- Hostel
- Hotel
- Massage Studio
- Pool
- Spa
- Yoga Studio

5. Conclusions

Purpose of this project was to find the association of district venues and the number of COVID-19 cases in Bangkok. Find the summation of COVID-19 cases in each district of Bangkok. Locate each district (latitude and longitude) to find nearby venues from Foursquare data. Then create clusters of districts based on the top 10 of most common venues. Create visualizations to show the number of COVID-19 cases and clustered we created as a map and horizontal bar chart. Lastly, we identified the category of most common venues which include in the top 5 districts with the most number of COVID-19 cases but not in the districts with least cases.