# The Battle of Sydney Suburbs

# Lie Zhou

## Introduction

Australia has been affected by the corona-virus since February 2020. Sydney's retail industry has suffered a huge blow because of the social distance restriction. Among all the retail industry, the catering industry has been greatly affected. Many restaurants have not been operating for three months. These restaurants are facing bankruptcy crisis. Fortunately, Australian government's response to the corona-virus is very good. Since the mid-May, the Australian government is trying to gradually loosen its control over the retail industry. Hence, restaurant have begun to re-open. However, because the global epidemic has not subsided, there still exist a risk of spreading the virus when people are crowded. The company I work for is selling anti-epidemic materials such as hand sanitize and masks to various retail companies in Sydney. However, this company does their business in B2B, which means it is difficult for many small companies to purchase hand sanitize directly and efficiency. This project aims to provide more potential customer to the company I'm working with, so that both the company and Sydney's catering industry can benefit.

## Data acquisition and cleaning

Sydney is one of the largest cities in the Southern hemisphere, but the scope of the Sydney city council is very small. This project will study the entire Great Sydney Area. Therefore, the first step is to collect the suburb information in the Great Sydney Area. Next, similar to the homework the example in the third week, this project will collect relevant catering industry information through Foursquare. Because the retail industry is greatly affected by population, this project also needs to cluster the data based on population in different suburbs.

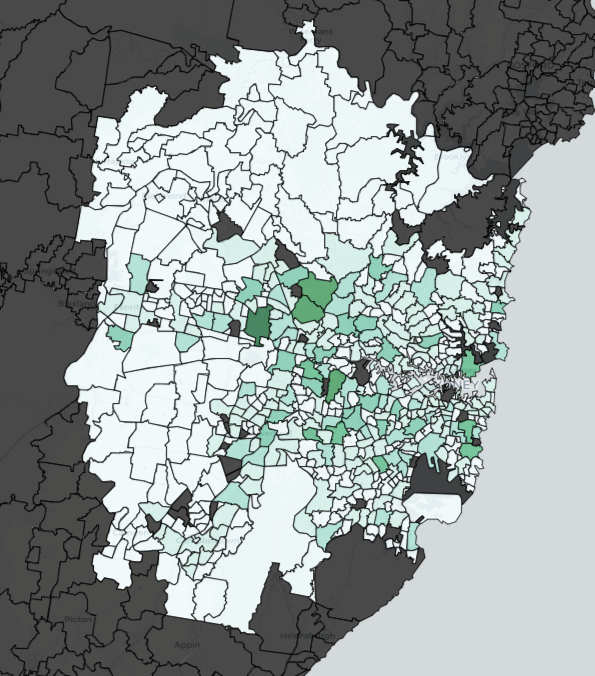
Through a simple research, I found that there is no data on the Internet that includes the names and coordinates of all suburbs in Sydney. However, the Wikipedia website and the NSW state government website provide the names of Sydney’s suburbs and the coordinates of NSW’s suburbs, respectively. In addition, the Australian postal system also provides coordinates corresponding to each postal code. I first analyzed the postal code and found that a postal code was used by multiple suburbs at the same time. So I decided to use web scrape skill to get the name and coordinates of the suburb from Wikipedia and NSW state government.

As for the data cleaning steps, I first compared the postcode and coordinates of the collected NSW with the information of Sydney’s suburbs to get a table containing the postcodes and coordinates of all suburbs in Sydney. Then, using crawlers, collect demographic data of Sydney suburb from domain.com. I did a manual check on the suburb which has no demographic data, and delete some duplicate named suburbs (these suburbs has same suburb name but different location).

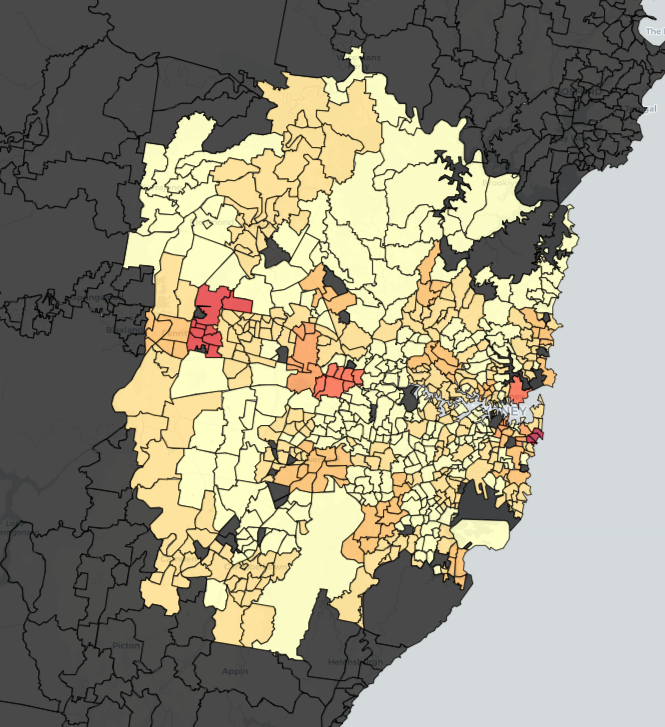
After completing this step, I downloaded all the infection data of Sydney since the outbreak of COVID-19 from the NSW government website, and added these data to the demographic data frame I made.

## Exploratory data analysis

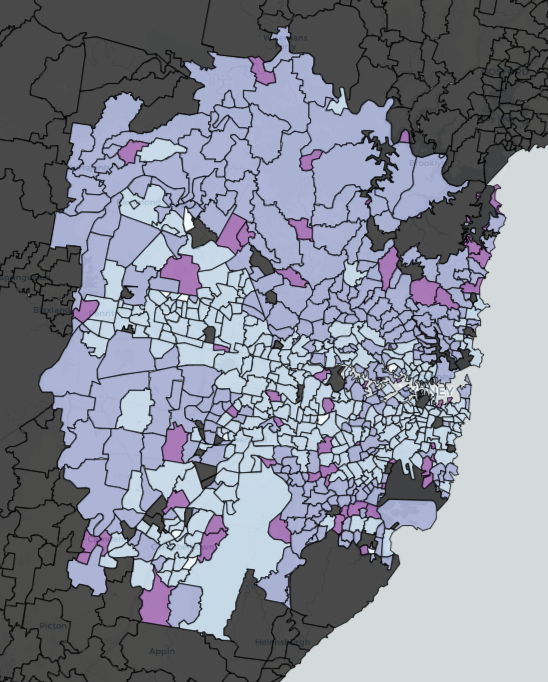
After completing the data collection steps, I conducted a correlation analysis of the number of COVID-19 patients and the population of suburb. The result is disappointing, COVID-19 has no correlation with the population of Suburb. The previous report on COVID-19 mentioned that the elderly may be more susceptible to virus infection, so I analyzed the age group of Sydney COVID-19 patients. The results showed that there were 2944 COVID-19 patients in Sydney. Among them, 118 people aged 0-19 years, accounting for 4%; 1119 people aged 20-39 years, accounting for 38%; 794 people aged 40-59 years, accounting for 27%; age 60 913 people over the age, accounting for 31%. This result shows that in Sydney, COVID-19 does not show a stronger ability to infect the elderly. However, relatively speaking, children are not easily infected. The reason for the relatively large number of adult infections may be because adults have higher social needs and more frequent contact with people. However, children and young people have no chance of being exposed to the virus due to the timely closure of schools by the NSW government.



Population by suburbs



Patient by suburb



Age group by suburb

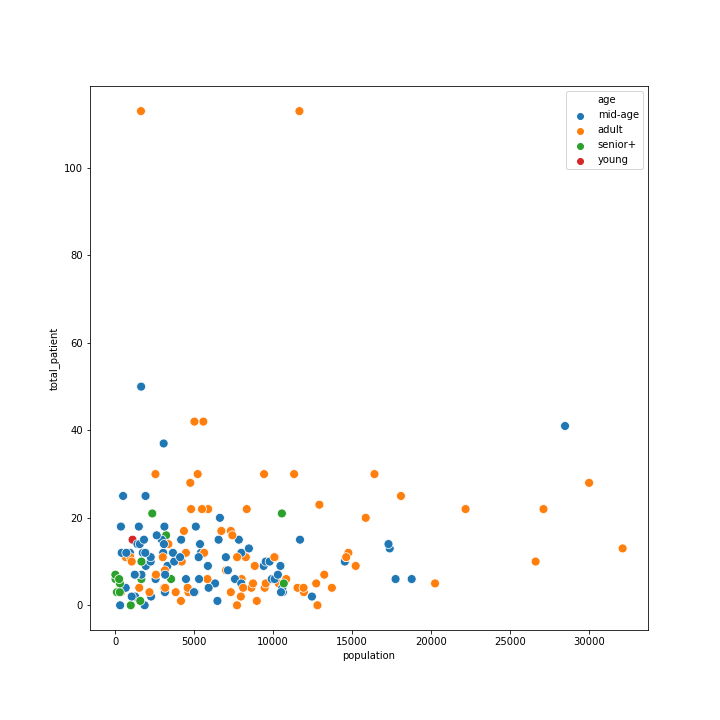
By using the code learned in the previous week3, the venue information of each suburb was collected from the Foursquare website, and the top 10 venues of each suburb were selected.

## top 10 venues

Top 10 venues for Sydney suburbs

Using the method of K-mean clustering, the Sydney suburbs are divided into 4 clusters according to the types of top 10 venues (a total of 5 clusters without the venues information are counted).

After analyzing, I found that cluster 3 contains more restaurant compares to other 4 clusters. although I already check the correlation between population and COVID-19 total patient number in previous steps, and found no linear relation, I'm still going to make a scatter plot to show the potential connection between population, age group, and patient for cluster 3.



By visually checking the scatter plot, I found out that the distribution of COVID-19 patients in cluster 3 is characterized by various suburbs and low frequencies. Moreover, the graph suggest that the number of patients in suburbs seems to be divided by 30 or 40. Therefore, the area where more than 30 patients appear in the suburb is regarded as vital area.

After screening, I left 7 important suburbs. I analyzed these 7 important suburb restaurants and found that Suburb Girraween, Llandilo, and Pemulwuy have no restaurant. Suburb Mosman and Queens park only have a few. By manually checking the data, I found that the most common venue for these five suburbs are cafes.

## Conclusion

In this project, I obtained the coordinates, demographic data, and the number of patients of COVID-19 from all NSW state government, domain.com, and other network sources. Through the Foursquare API to obtain the restaurant information of each suburb. Through K-means cluster analysis, I screened the collected information. The screening results show that there are a large number of restaurants in Bondi and Tamarama, and there are many COVID-19 patients. The number of restaurants and the population of the region mean that there are more potential promotion goods customers. In these areas, if there are COVID-19 patients higher than those in other areas, it will stimulate merchants to buy more promotion goods related to epidemic prevention after Sydney releases lockdown.