CMPT353 Report: OSM, Photos, and Tours

Xubin Wang, 301368109

Yangxin Ma, 301307944

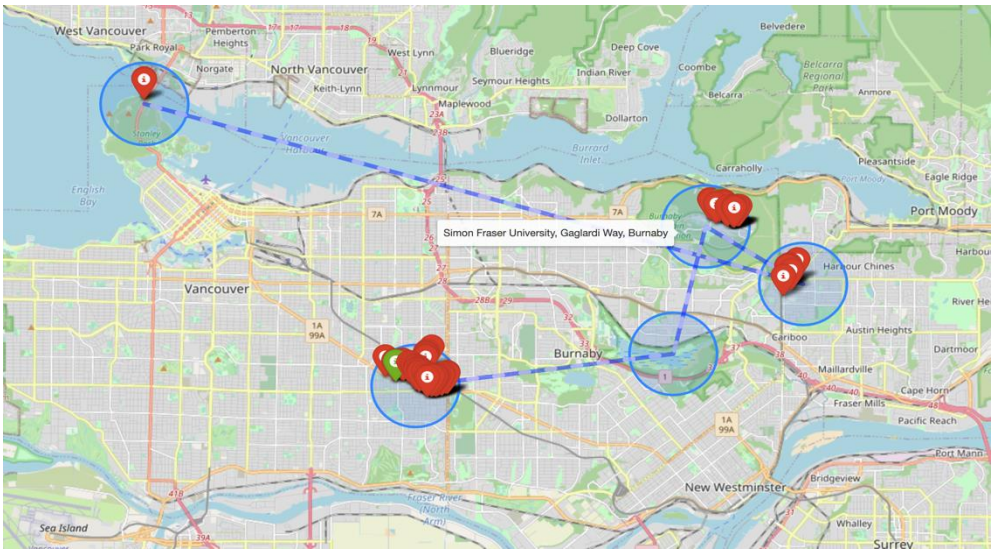
Part 1:

1. Read Pictures and Get Location Information: **read_pic.py**

    The topic we chose first is to read pictures and tell which place it is. To achieve this, we use PIL package to read EXIF information from our chosen JPEG images, extract the coordinates of each image and change it into degree, so that it can be used to calculate distance from other places as what we had done before in exercise 3. In our example, we choose 5 pictures of SFU, Deer Lake, Stanley Park, a restaurant on Kingsway and the home of one of our members in Coquitlam. And put the latitude and longitude of images into locations.csv in a file named data.

2. Find Where, Draw Route and Find Nearby Facilities: **trip_trail.py**

    In this part, we first manually separate amenities that may be food and entertainment. In this part, we find an extraordinary tool named Folium to draw amazing maps. The first step is to create a new map and locate is in Vancouver. Then, using Nominatim package from geopy.geocoders to ask for the location information from OpenStreetMap Api, it will return the whole information of this location, and we get the unit number, street name and city name from those information. This let us know where this picture was taken.



As you can see in this figure, this is where we took these five photos. If you put the cursor on the blue CircleMarker, it will shows you the name of

that place. Then we want to give a instruction of the order of these pictures being taken. Sp, we find Plugins package to draw path between tow places. The arrow between two places means the direction of our trip.

Finally, we want to check how many facilities are there around whrere we take pictures. So, we plan to calculate distances between each picture to every facilities in the given dataset. We use the same method as we have done in exercise 3 to calculate distances and filter these facilities within 1000 meters of these pictures. We use red to represent “Food”and green to demonstrate “Entertainment”.

It is obviously that we can get food easily within 1000 meters at where we take photos except in the deer lake.

3. Analyze Cities: **analize_city.py**

In this part, we want to analyze is there any relationship between amenities and cities, which is that whether a city with more amenities means this city is better developed.

We first separate the whole data file in to six cities: Vancouver, North Vancouver, Burnaby, Richmond, Coquitlam and Surrey. Unexpectedly, there around 5000 amenities that are not in any of these 6 cities, which means these amenities are located in beyond resident, so we consider these points as outliers and froppe these data. Then we found there are duplicate recorded points, which means the same amenity in the same place but recorded twice or more in different timestamp, so, we dropped these duplicates but keep the latest one. And we manually partition amenities into 7 categories:

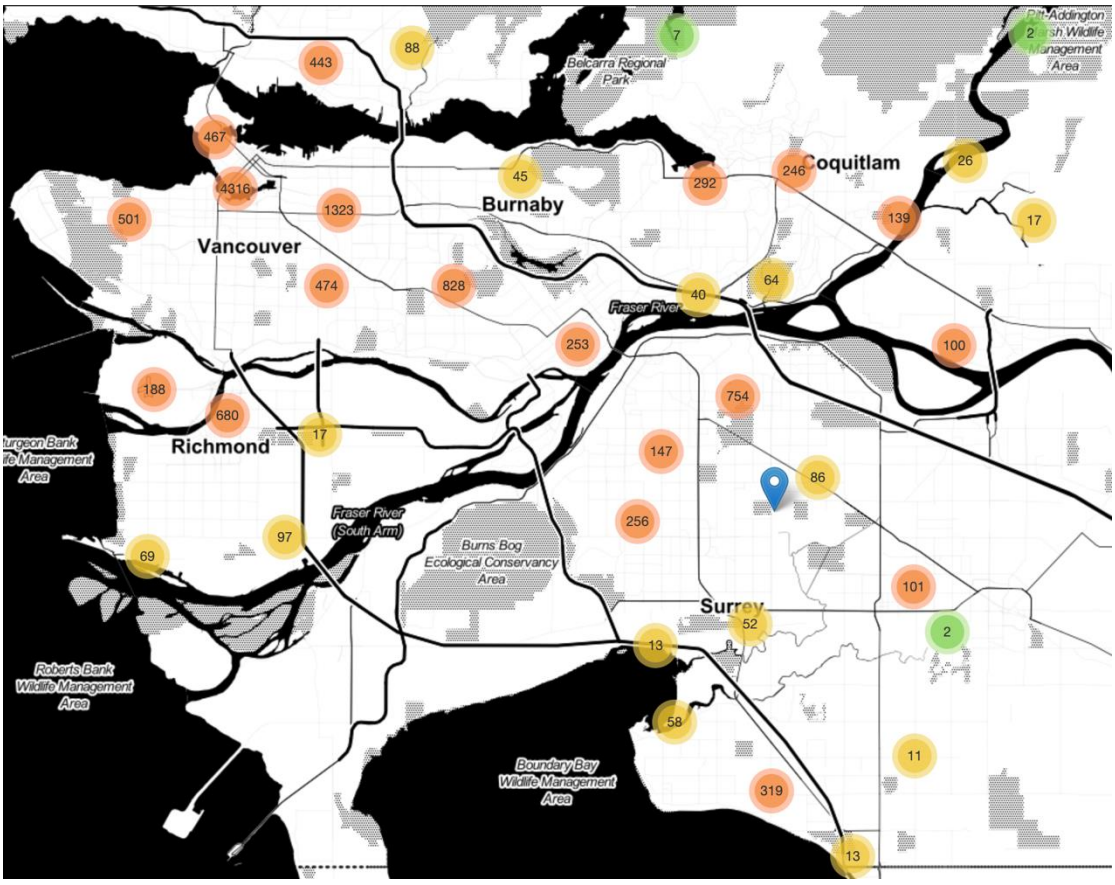
a. Food
b. Facilities: Foundation devices used actively by the public
c. Service: Places that provide life services
d. Education: Places that teach people some knowledge or skill

e. Medical

f. Entertainment: Places let people have rest and make fun.

g. Transportation: Bus Stop, Parking Space, Car Renting, etc.

In the preparation phase, we want to have a look of how these amenities are distributed in these 6 cities. So again, we use Folium to give us a whole sight of the entire dataset as bellow.



red : high density, yellow : middle density, green : rare density

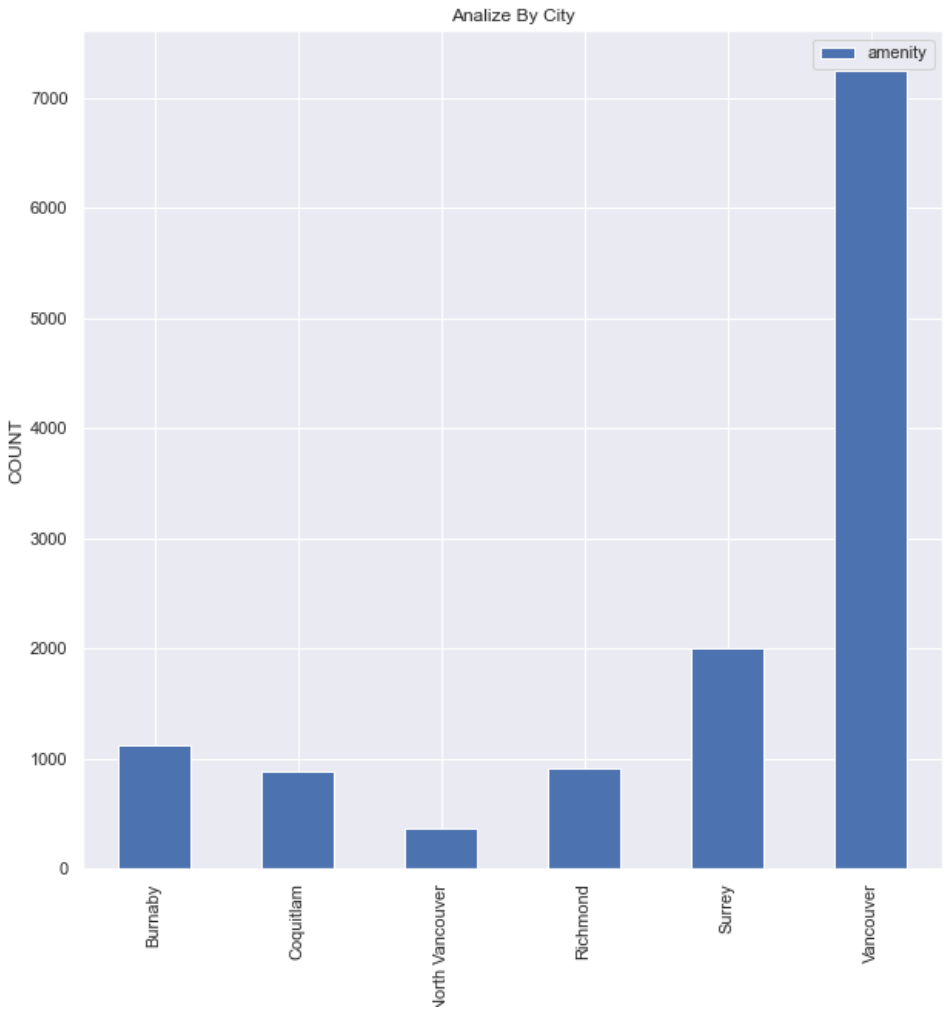
This figure gives us some initial information that Vancouver should be the most developed region. Then we start to vertify our guess.

We devide this into two parts: 1. Analyze among cities 2. Analyze within cities.

The first part is implemented by group the whole data with ‘city’ and count the number of amenties of this ciry. And we use pandas to draw a bar chart of the result. The chart has two dimentions: city and

count, the x-axis is each cities and the y-axis is the count of amenities of this city.

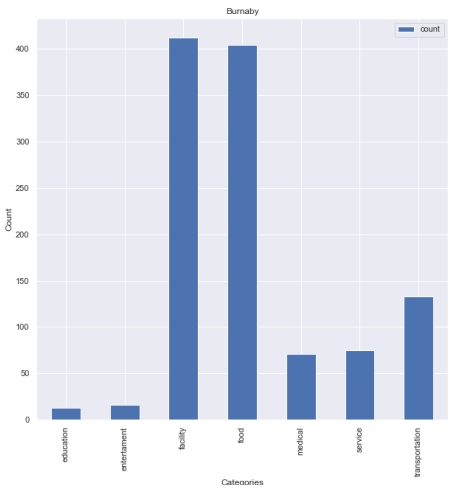
The figure is shown as below:

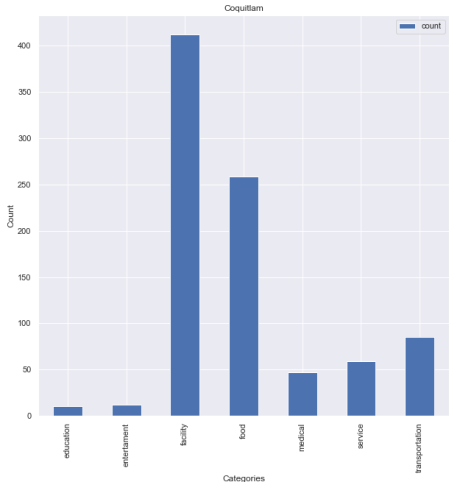


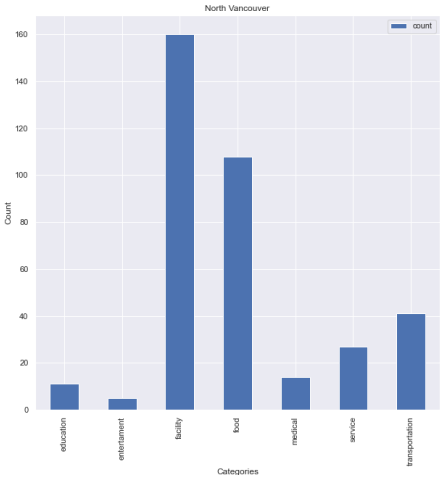
We can see that, Vancouver has its unsurpassed advantage, amenities in Vancouver is almost 7 times as other cities and 3.5 times of Surrey.

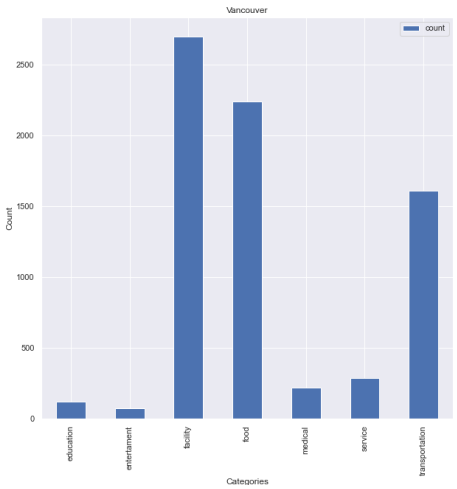
Secondly, we want to check the distribution of each categoty in each cities. This is implemented by extract data into 6 cities first and

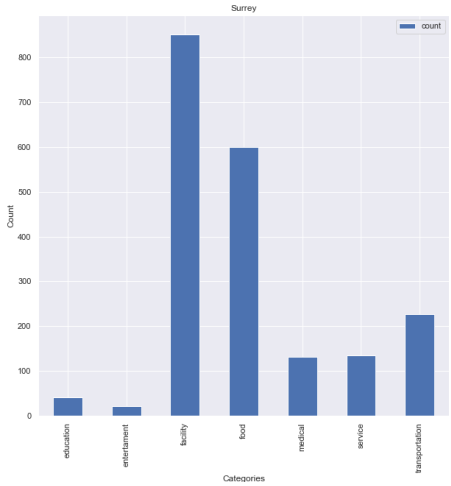
groupby each city with ‘category’ and this gives the following figures:

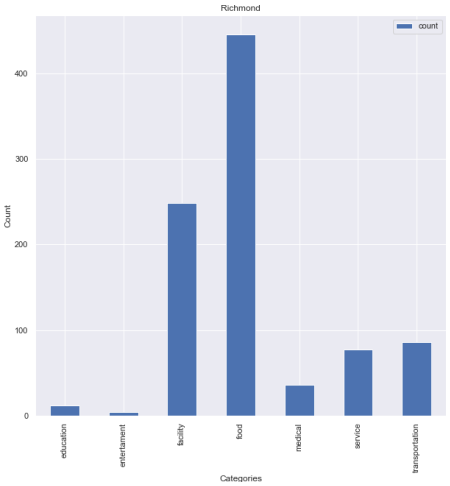












We can see that each city would contain lots of facilities and foods, so these two categories is not the determinats of the development of a

city.

So, which category can represent the determinat of a ciy, so far with these limited information, we think it should be the transportation. We can see that better cities like Vancouver, Surrey and Burnaby all have large magnitude of transportation while other cities are not. So, we can easily conclude that cities with more transportation amenities are more likely to be developed.

4. Limitations:

So far, what we have found should not be defined as rigorous, because we do not have huge amount of data and at the mean while, the data we have is not high-dimentional enough to enable us to analyze. So, our conclusion should be:

"With the data gained so far, we can say that cities with more transportation amenities are more likely to be developed.".

If we have more time, we want to gain more data in two sapects.

One is to gain more data from differnet countries and different mainland(five continents), especially data from China, US and Europe.

The other is to gain more attributes like the altitude, population, area and precipitaion.

Why we think about this? Because we saw that Surrey has more amenities and more transportations than Byrnaby, but we do not think Surrey is better than Buranby. This reason cause this situation is that Surrey has large area than Buranby. So, we think there should be other attributes to have impact on our first topic. Like altitude, why North Vancouver seems to be not developed as it has less magnitude of amenities and transportations than other cities? We consider its altitude should be one element to cause this. So, the further work is to gain more data and analyze this topic in high-dimention attributes.

5. Accomplish Statement of Xubin Wang:

   a. Learned new amazing package tools by searching online when

thinking how to demonstrate visulization and find variety of great tool to drawing map and route.

b. Knowing the logic of data analyze by thinking city analization and find more interesting aspects to explore.

Part 2:

In this part, we are going to help clients to plan a tour of the city by driving.

1. Data we use: wiki_fliter.py

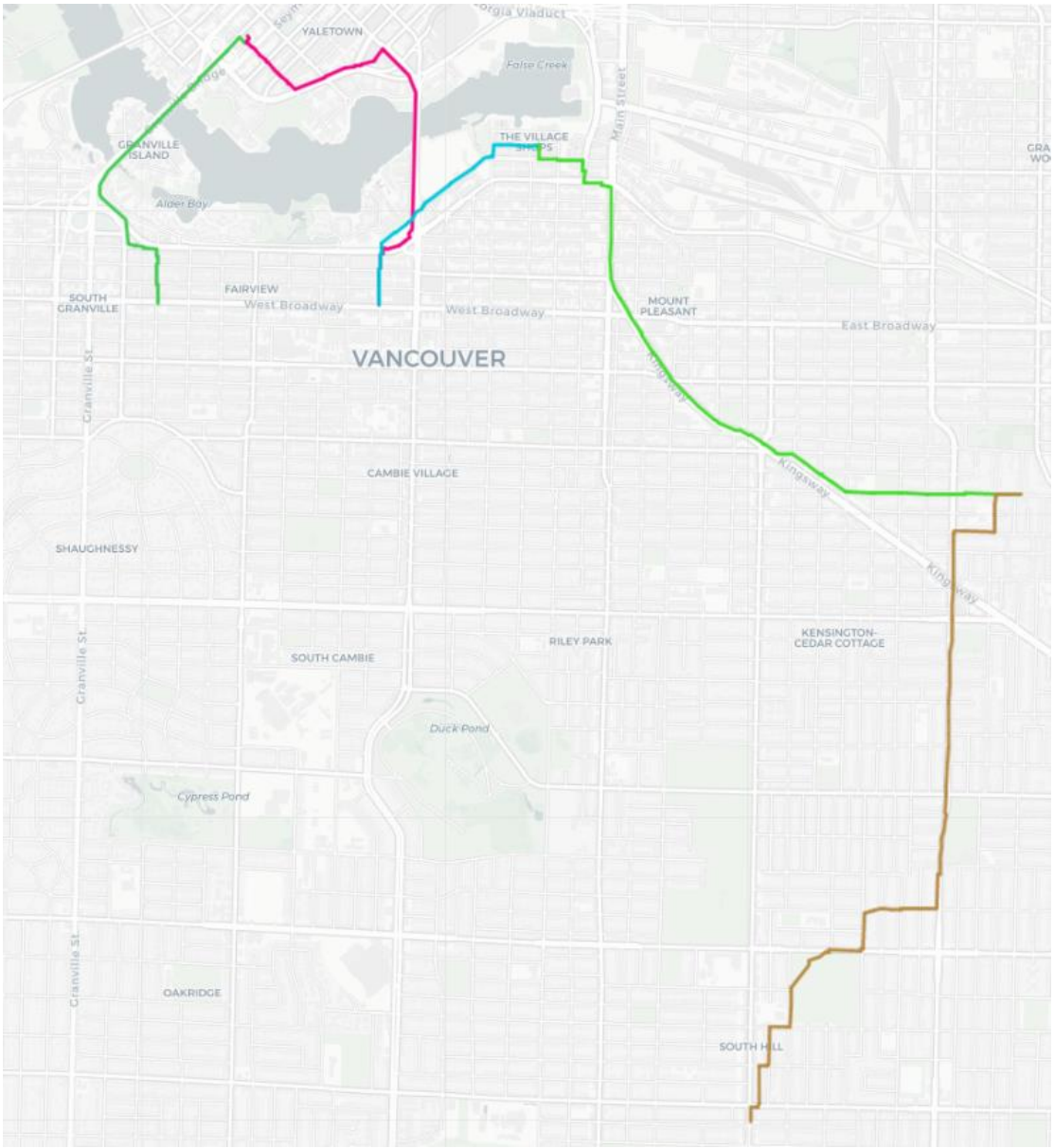
First, we categorized the whole data set into food, facility, service, education, medical, entertainment and transportation. Also, we categorized the set into different cities, like Vancouver, Burnaby and so one. For our example, we assume that the client wants a tour in Vancouver. So, we only use the records which city is equal to Vancouver. Because we need to find some famous places, we need to use Wikipedia data set. Some records contain the Wikipedia entity, some records not. We only use records which contains Wikipedia entity ID.

2. Techniques I used to analyse the data: tour_plan.py

To check a place is whether famous or not, we are going to check the Wikipedia entity completeness by using the entity ID. So, we need to use wikidata package. We got this idea from the project instruction. We found that the Wikipedia entity more complete, the length of the entity larger. We use member functions of wikidata package to get the length of all entity first. Then, if we have, we group the data set by categories. We found the max entity length for each category. We thought we were done. But we found that the famous food is Starbucks. We categorized the cafe into food. We noticed that the chain store like McDonald's or Starbucks got more completed Wikipedia entity. We think our client don't come to Vancouver for eating fast food. We think they come to Vancouver for local cuisines. So, we excluded record which shows in the data set more than ten times. Then, we

got various famous places which are not chain stores. Then, we use the distance function comes from assignment3 to get a shortest path between these famous places. We used osmnx and networkx packages to draw the map.

3. Path Picture:



The intersection of different color lines is the famous place.

4. Limitation:

Our map is too simple, we will include more information in the map, like the address of each point. The map just showing the path. Our clients may be unable to find

the correct place.

Part 3:

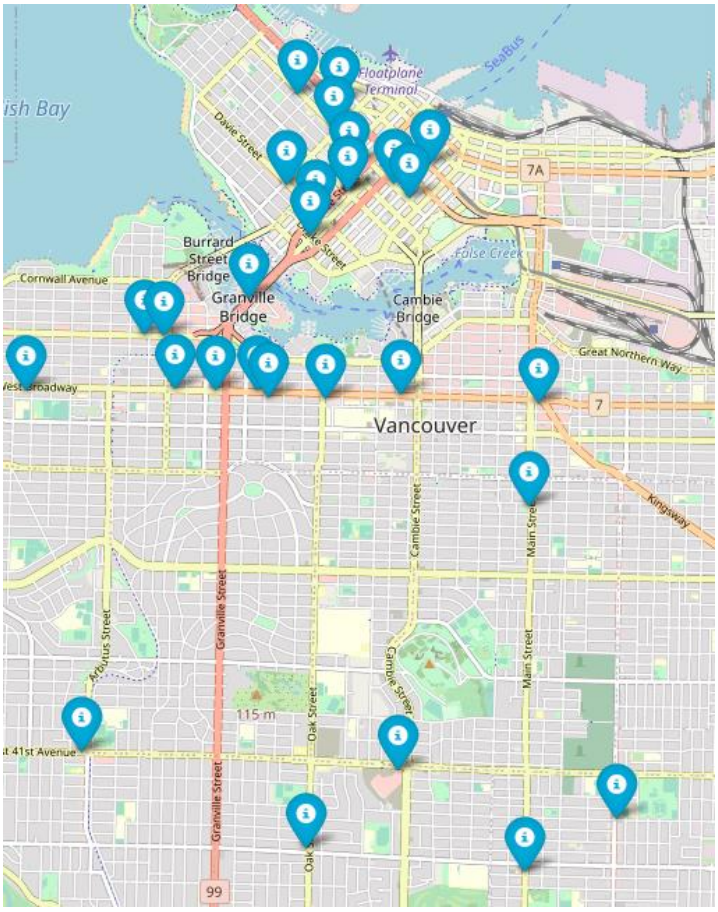
In this part, we are going to help clients to choose a hotel or Airbnb.

1. Data we use:

We assume that the client wants to choose a hotel or Airbnb in Vancouver. We categorized the set into different cities. We only use data points which locate in Vancouver. We assume that a good hotel or Airbnb is surrounded by many restaurants. So, we excluded all data records which are not restaurant.

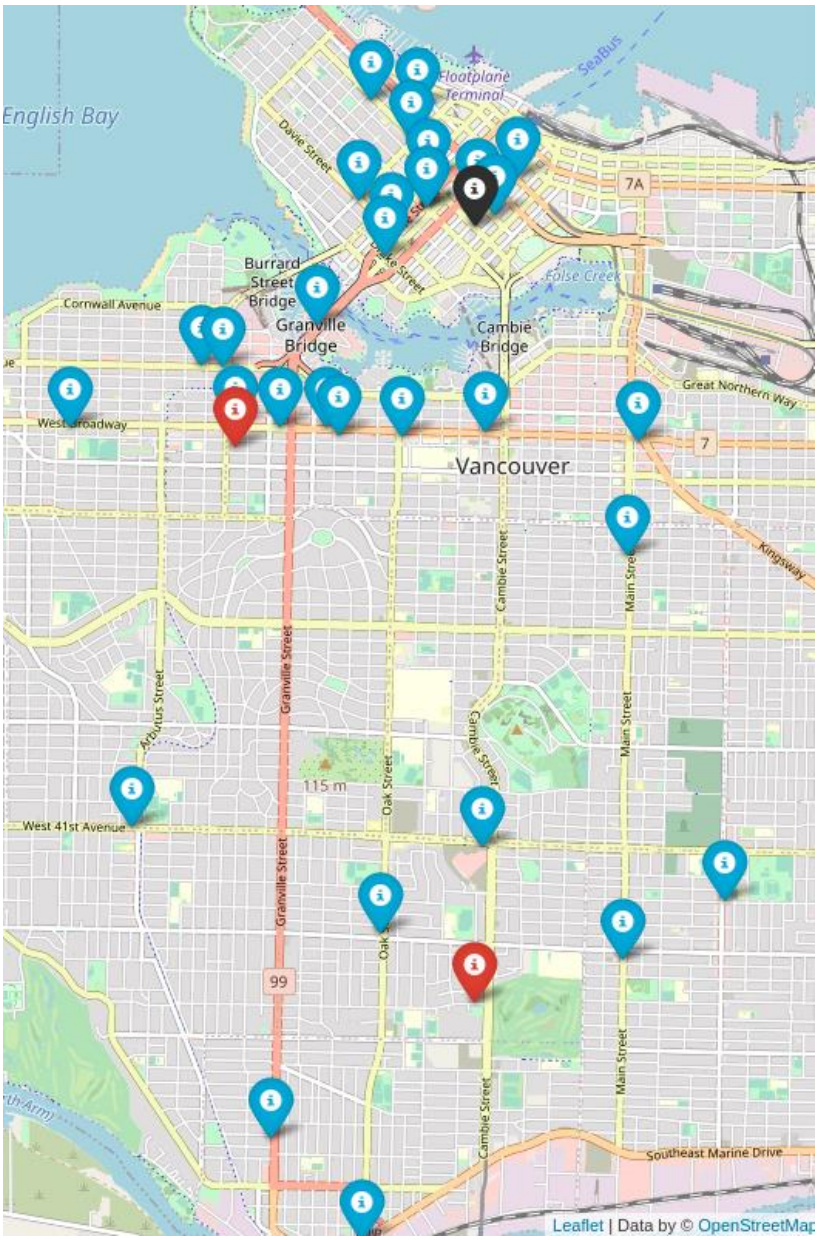
2. Techniques I used to analyze the data:

When we were going to analyze the data set, we found that there are no hotel or Airbnb data point in the data set. We thought the data set contains those point. So, we change our goal of this section. Instead finding a actual hotel of Airbnb, we planned to find a location which is surrounded by many restaurants. We created a map and show all restaurants on the map first.



After seeing the map, an idea came to our mind is clustering all the data point into several groups. The center point of each group is surrounded by many restaurants. We used the K-Means algorithm to cluster those points. In K-Means algorithm, choose a appropriate k value is very important. By doing some research on Google, we found the elbow method. Then, we used the elbow method to get the k value. We used the data set and the k value to train the K-Means model.

3. Result we got:



We got three new points. One of them is black. Two of them are red. These three points are centers points of three different cluster group. We recommend the black point to our clients because the black points are the center point of the largest cluster group.

4. Limitation:

We can't help our client to find a actual hotel or Airbnb. We can only help them to find a good location. A hotel or Airbnb nearby this location is a good choice.

Accomplish Statement of Yangxin Ma:

I created a tour path of a city, and I found some good locations of hotels or Airbnb. For creating a path on the map, I learned some knowledge of using Folium, Osmnx, and Networkx packages to create a map and a path by searching on Google. For clustering data points, I learned how to use the sklearn package which is very useful. The sklearn package contains many models. My tasks are distributed by the project leader. I learned how to work in a group. A group project is not like an individual project. Our task is relative. I used some functions from his code. This project is a great experience for me to learn how to cooperate with other people.