CMPT353 Report: OSM, Photos, and Tours

Xubin Wang, 301368109

Yangxin Ma, 301307944

Xiaohang Hu, 301353291

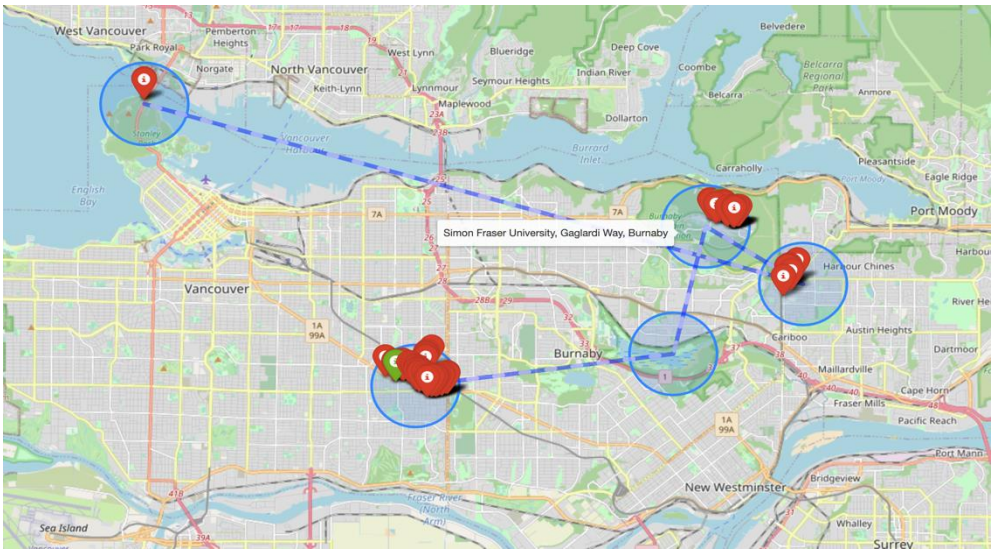
Part 1:

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:

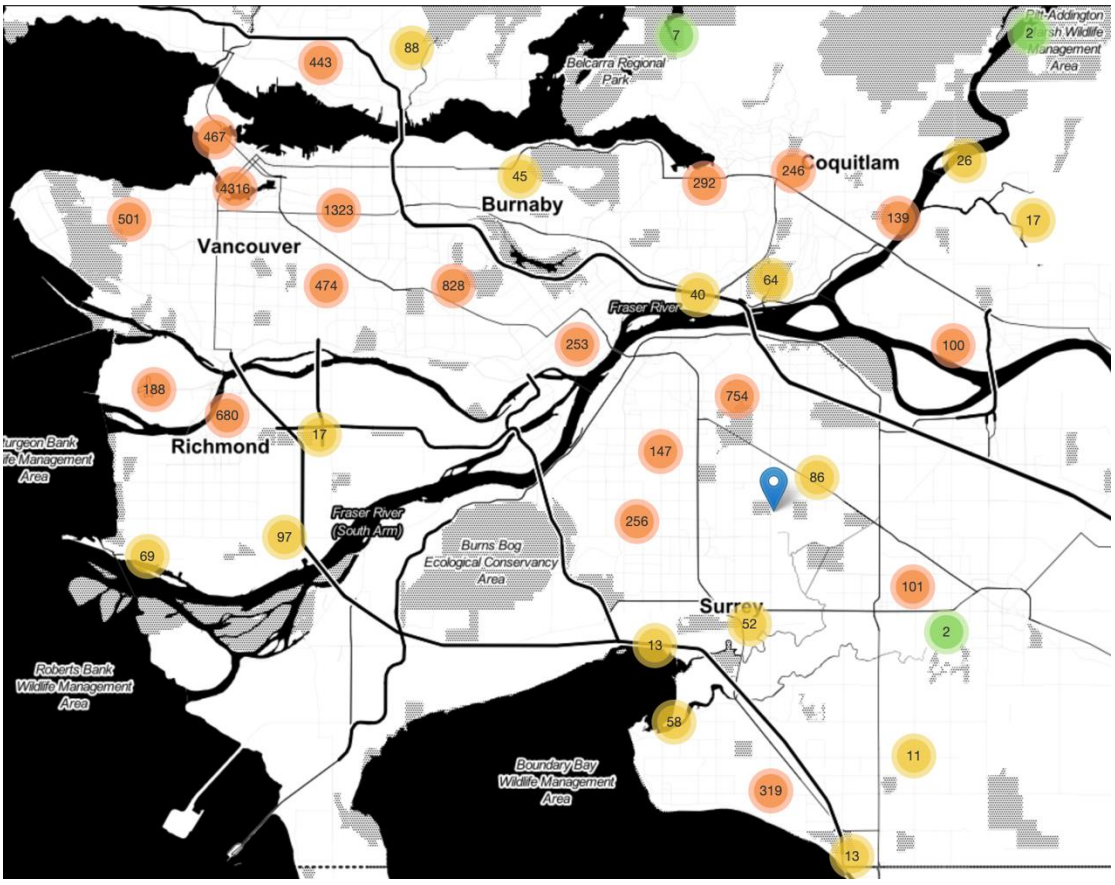
a. Food
b. Facilities: Foundation devices used actively by the public
c. Service: Places that provide life services
d. Education: Places that teach people some knowledge or skill

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

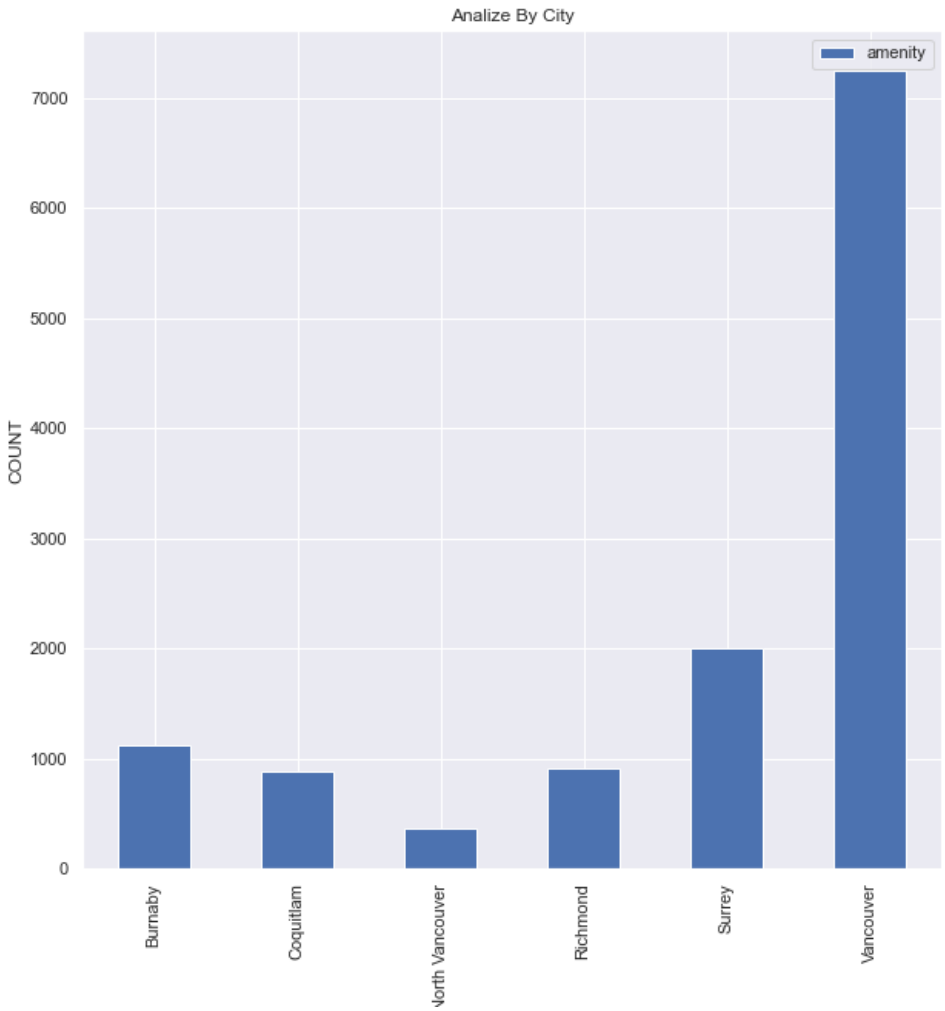
This figure gives us some initial information that Vancouver should be the most developed region. Then we start to vertify our guess.

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.

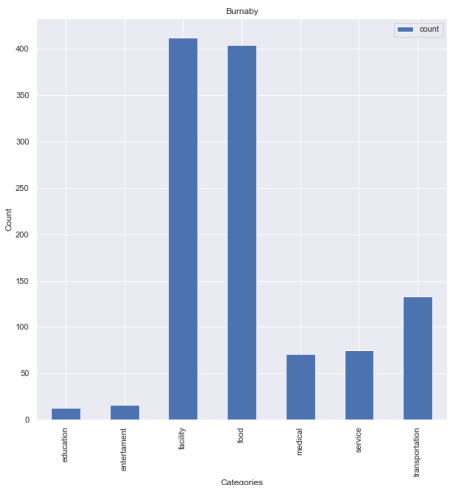
The figure is shown as below:

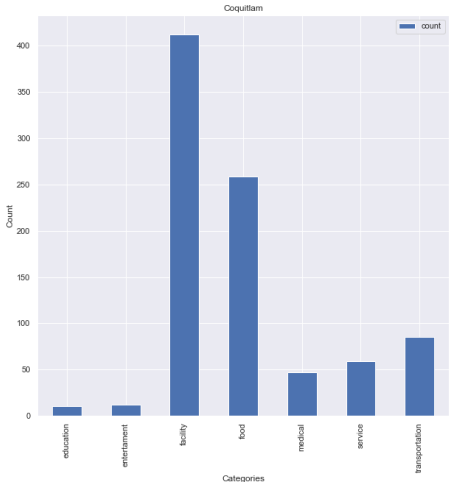


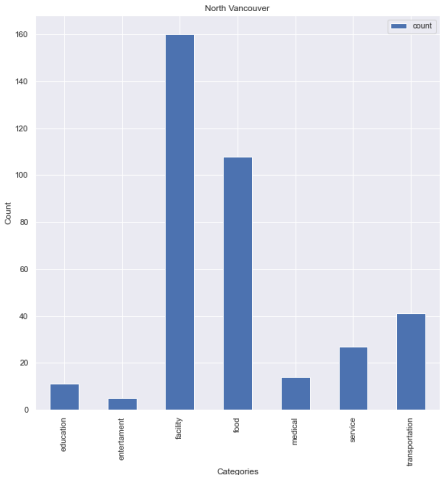
We can see that, Vancouver has its unsurpassed advantage, amenities in Vancouver is almost 7 times as other cities and 3.5 times of Surrey.

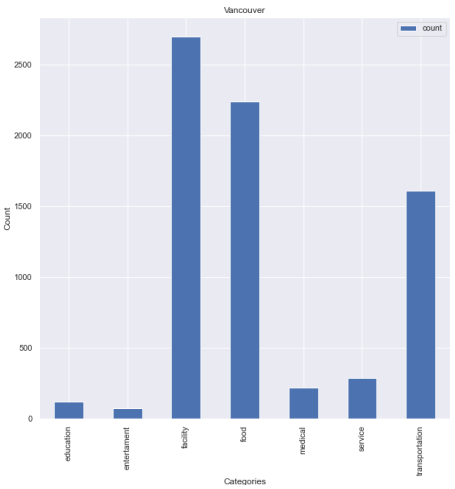
Secondly, we want to check the distribution of each categoty in each cities. This is implemented by extract data into 6 cities first and

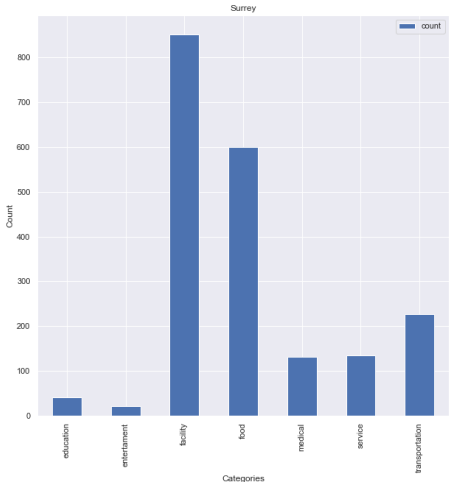
groupby each city with ‘category’ and this gives the following figures:

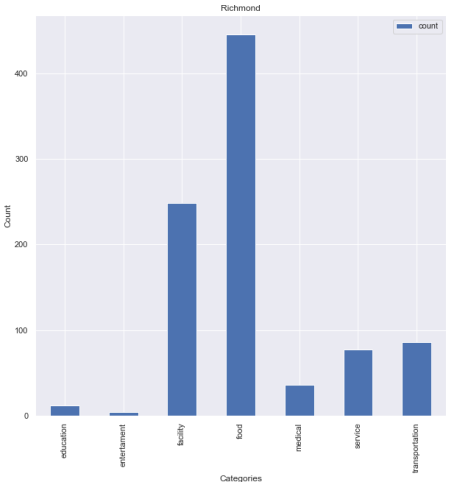












We can see that each city would contain lots of facilities and foods, so these two categories is not the determinats of the development of a

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**

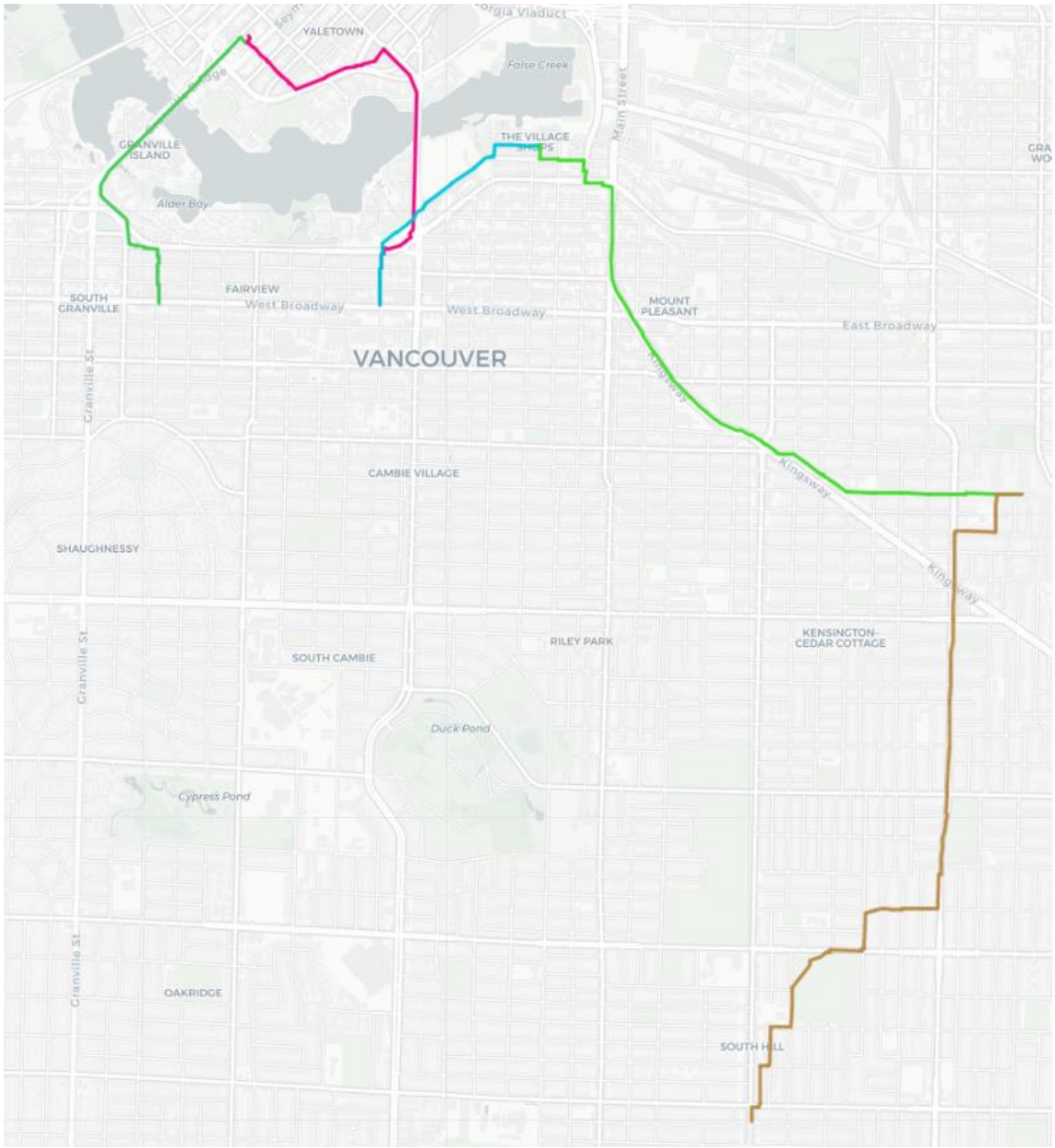
First, we categorized the whole data set into food, facility, service, education, medical, entertainment, and transportation. Also, we categorized the set into different cities, like Vancouver, Burnaby, and so on. For our example, we assume that the client wants a tour in Vancouver. So, we only use the 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 do not. We only use records that contain Wikipedia entity ID.

2. Techniques I used to analyze 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 the wiki data package. We got this idea from the project instruction. We found that the Wikipedia entity is more complete, the length of the entity is larger. We use member functions of the wiki data package to get the length of all entities 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 chain stores like McDonald's or Starbucks got more completed Wikipedia entity. We think our client doesn’t come to Vancouver for eating fast food. We think they come to Vancouver for local cuisines. So, we excluded records that show in the data set more than ten times. Then, we got various famous

places that do not have chain stores. Then, we use the distance function comes from assignment3 to get the 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:

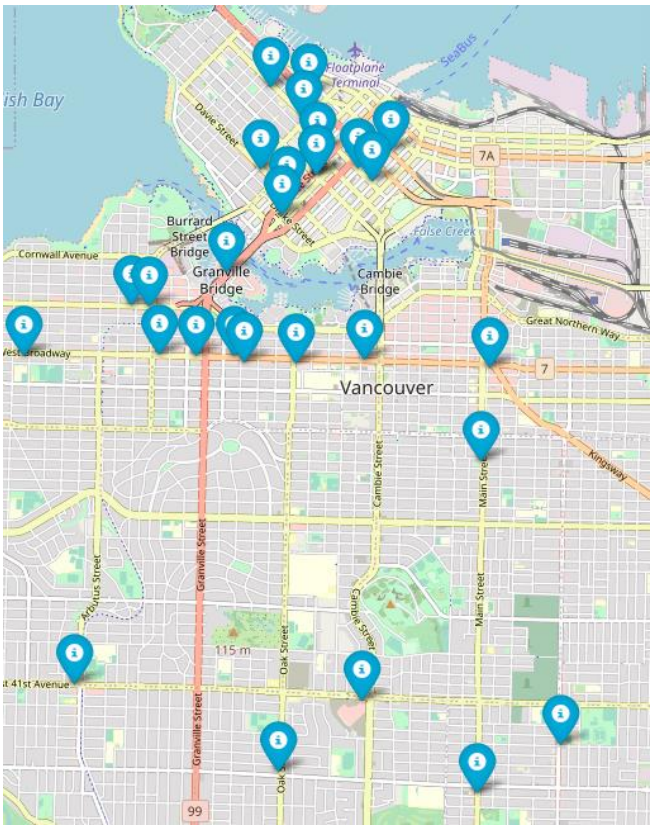
In this part, we are going to help clients to choose a hotel or Airbnb.

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 a restaurant.

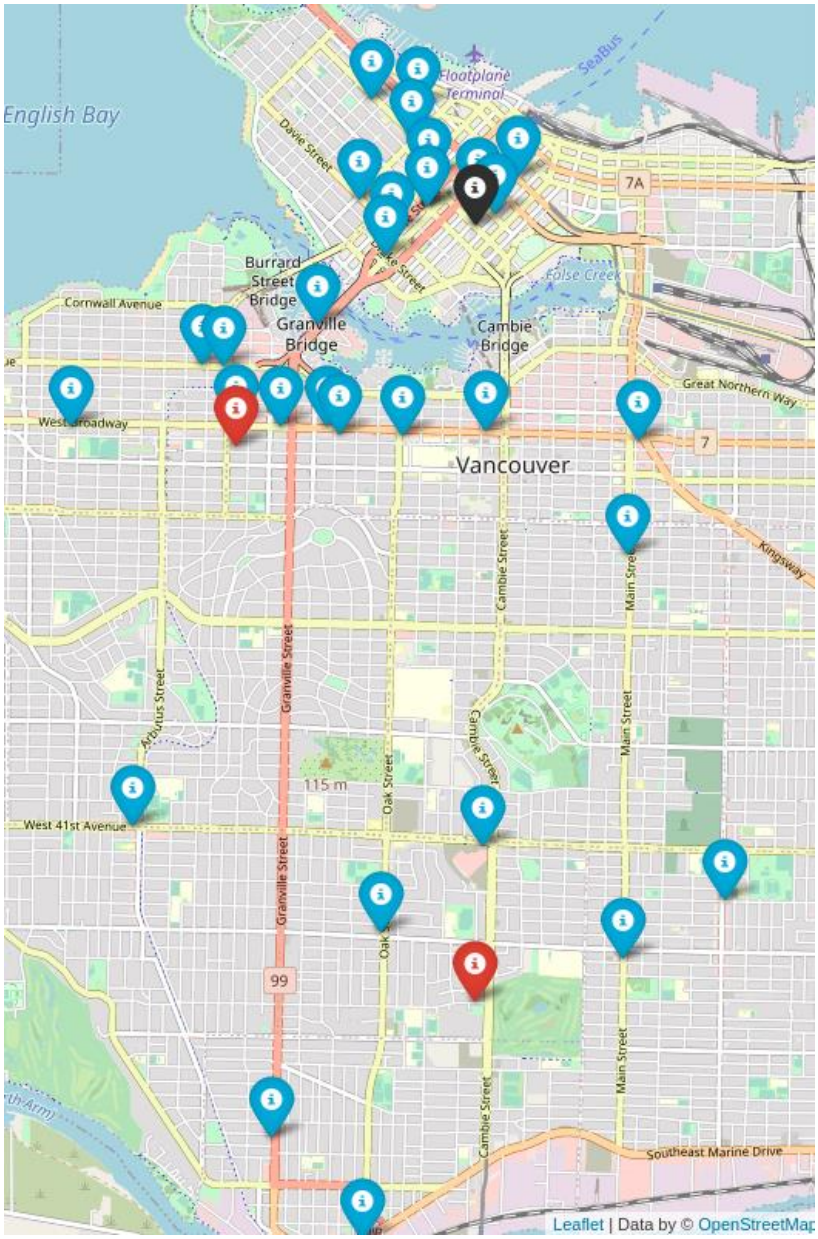
2. Techniques I used to analyze the data:

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



After seeing the map, an idea that 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 the K-Means algorithm, choose an 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 groups. We recommend the black

point to our clients because the black point is the center point of the largest cluster group.

4. Limitation:

We can’t help our client to find an 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.

5. 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 Networks 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.

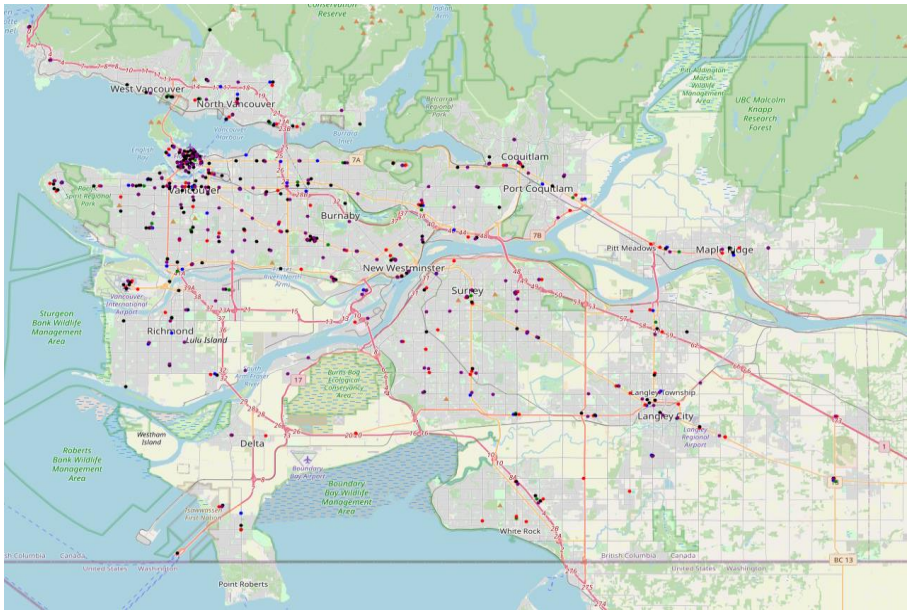
Part 4:

Question: Is some part of city with more chain restaurants?

1. Which belong to chain restaurants? **Rest_rain.py**

   If we want to know the distribution of restaurants in different cities, the first thing we need to do is pick chain restaurants. I only choose restaurants that have at least 20 restaurants. (McDonald's, White Spot, Tim Hortons, Starbucks, A&W, Subway) I combine them into one Data frame for future analysis. Of course, those restaurants do not belong to chain restaurant also will be put into a Data frame. I also remove some data from food categories, such as juice bars. In my view, they cannot be count as restaurants.
2. Show chain restaurants on map: **Rest_rain.py**

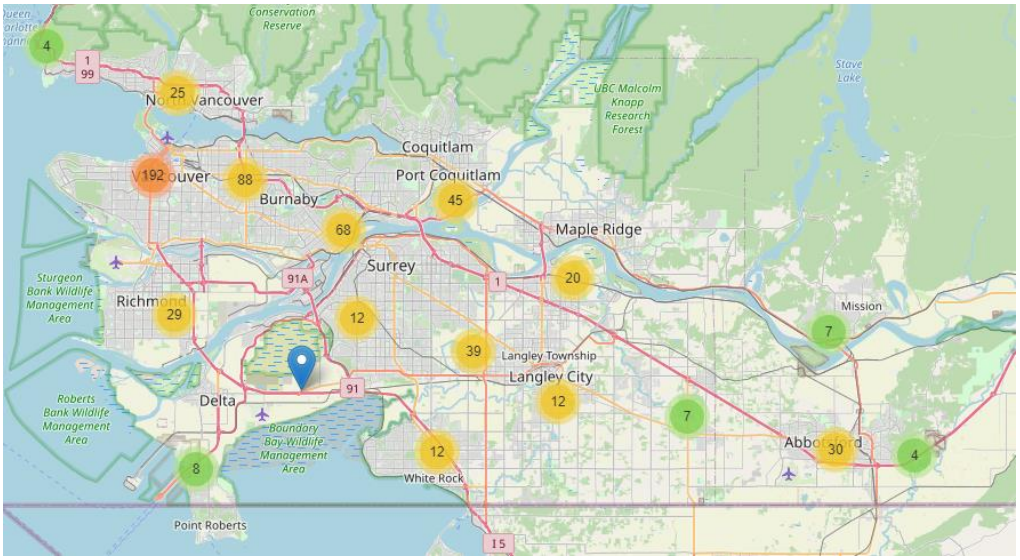
We want to know and analysis the place of chain restaurants. The histogram is of course a good choice, but it is not intuitive enough. Visualizing these restaurants on the map will help to understand the relationship between the city and the chain restaurants. First, import folium and location of Vancouver to map we need to use. Then I marked different chain restaurants in different color in the map.



Through viewing this figure, we find those chain restaurants distribute in every city. However, the number of chain restaurants in Vancouver is far greater than those in other cities.

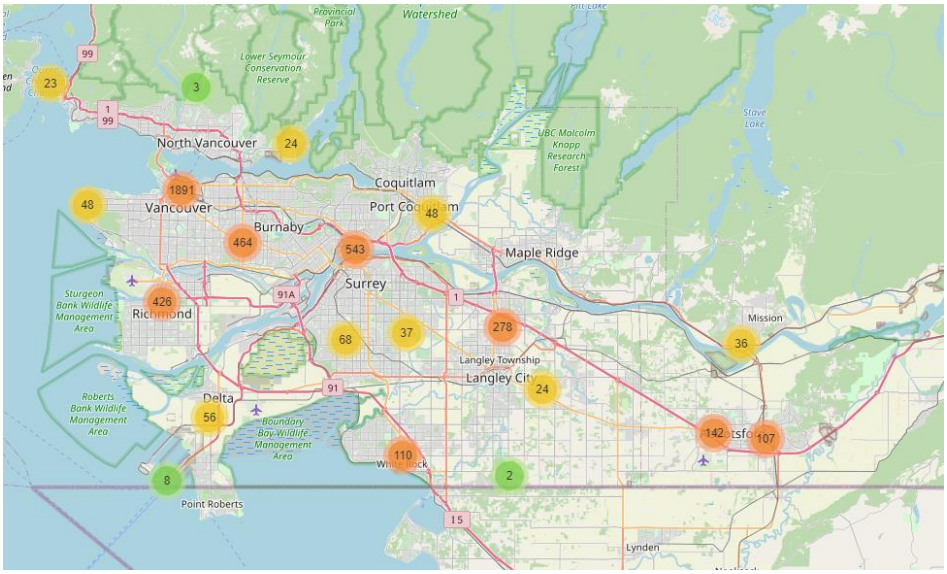
3. Using cluster to show the density of chain restaurants and not chain restaurants in different city: **Rest_rain.py**

To better show the density of chain restaurants in different cities, I used



Cluster in the map. It not only visualizes the density better but also displays the number of chain restaurants in different cities.

The orange color on Vancouver which means there has the highest density.



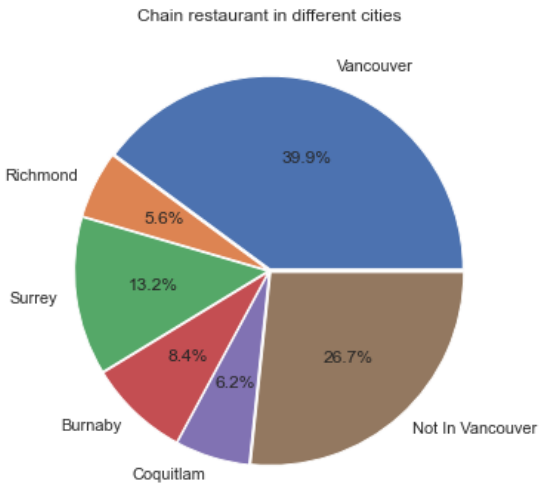
Other city has the middle density. (yellow color) The green color shows rare density which means little chain restaurants between different cities. Through this picture, we have seen that there are more chain restaurants in some places. (Vancouver) I also create the cluster figure of not chain restaurants through the same way.

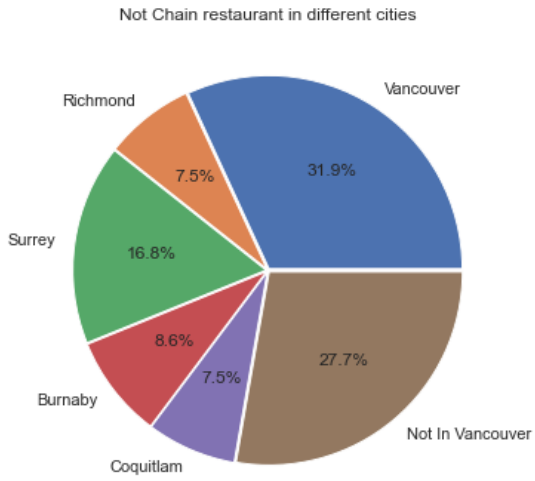
Similar, the Vancouver city also has the highest density. Moreover, the closer to

the city center, the greater density of restaurants. From the above two figures, we can see that when a city has more chain restaurants, there are also more non-chain restaurants. Do they have the similar proportion in each city?

4. Draw Pie graph to show the percentage: **ChainVsNotChain.py**

To find out the proportion of chain restaurants and non-chain restaurants in different cities, I created pie chart for them respectively. My partner has created the way to show city of each restaurant. Therefore, I create Data frame of restaurants with column[‘city’] firstly. Then split elements to Data frame of chain restaurants and Data frame of Not chain restaurants. Then created the Pie graph





based on different cities.

We can see that, Whether it is chain or non-chain, Vancouver has the largest proportion which followed by Not Vancouver area, Surrey, Burnaby, Richmond and Coquitlam. The two graphs can more intuitive to show that there are some city with more chain restaurants. As my team member Xu Bing described, Vancouver is the most prosperous city in the surrounding area. It has more amenities and traffic, so the density of chain restaurants are denser than other cities.

In addition, I also found an interesting phenomenon. Although there are more

data for non-chain restaurants than chain restaurants, their proportions in different cities seem to be very close.

5. The P-value of chi2_test for Chain restaurants and not Chain restaurants in different cities.: **ChainVsNotChain.py**

   Moreover, I also used the p-value of chi2 test to show quantity of Chain restaurants and not Chain restaurants is different.

```
#chi test of restaurants in different cities
obs=[[Vlen,Rlen,Slen,Blen,Clen,NVlen],[Vlen2,Rlen2,Slen2,Blen2,Clen2,NVlen2]]
chi, p, dof, ex = chi2_contingency(obs)
```

   Take the number of Chain restaurants and not Chain restaurants in six city into chi2 test.

```
In [39]: run ChainVsNotChain.py
"Are there some citites have more restaurant ?" p-value:  0.0377
```

   The p-value is 0.037 which is smaller than 0.5. It further means number of restaurants are different in the six cities.

6. Limitations

   The limitation of data:

   In my experiment, I discarded those restaurants with fewer than 20 , and chose some well-known chain restaurants（such as AW and Starbucks). Join those chain restaurants with few numbers, and there may be more discoveries. The relation between Chain restaurants and Not chain.

   Just as I found through viewing Pie graph. The percentage of chain restaurants and not chain in six place seems to be similar. But I did not know how to use a suitable test to get the p-value to show their relationship.

7. Accomplish Statement of **Xiaohang Hu:**

   a. I learned how to visualize data point on the map in different color and cluster them.

b. Learned how to create a Pie Graph.

c. I found the a amazing thing the percentage of chain restaurants and not chain in six place seems to be similar. This will inspire me to explore further.