CMPT 353: Computational Data Science

FALL 2019

FINAL PROJECT: OSM,PHOTOS,AND TOURS

Jinze Wu

301414928

Yizhou Chen

301414924

Introduction

Open Street Map (OSM) is an online map collaboration project for the construction of free content. In this project, we are provided with the processed OSM data, that is, some important attributes extracted from the geotagged photos of Vancouver. In this project, we discussed 4 problems we were interested in, including the urban functional zone problem, where to choose a hotel, the distribution of chain restaurants and non-chain restaurants and tourist attractions recommendation. Throughout the document we explain how we collect and process our data, how we analyze them and what results we got.

1. Urban Functional Zone Problem

In the data there are various amenities with different functions. We are curious that will the facilities with different functions be distributed in different areas? Like CBD, industrial area, culture and education district and so on. Will we be able to predict the amenity according to its geolocation? We did visualization and tried several machine learning models to reach a conclusion.

1.1 Data Process

We check all the amenities and pick some of them for our experiment.

1. Education area: 'college', 'kindergarten', 'language_school', 'library', 'music_school', 'prep_school', 'school', 'science', 'university'
2. Entertainment area: 'arts_centre',' bar', 'casino', 'cinema', 'clock', 'fountain', 'gambling', 'leisure', 'marketplace', 'nightclub', 'pub', 'spa', 'stripclub', 'theatre'
3. Industrial area (mainly about waste): 'sanitary_dump_stations', 'trash', 'vacuum_cleaner', 'waste_disposal', 'waste_transfer_station'

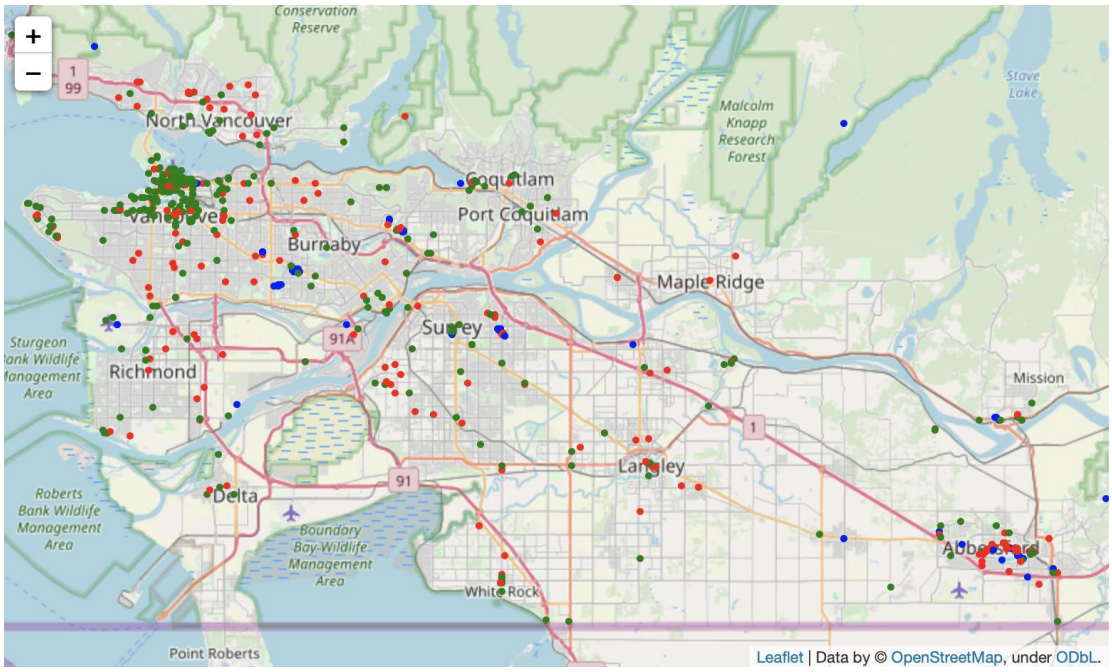
We extracted these amenities and respectively labeled them as ‘edu’, ‘ent’ and ‘ind’.

1.2 Data Analysis

- Visualization

  We plotted them on a map to see their distribution.

  Green = entertainment amenity Red = education amenity Blue = industrial amenity



  We can see from the above plot that, most of the green points gather in downtown. Blue points seem always be close to water. Red points are relatively scattered, but gather in some small group somewhere. There is some pattern to be discovered.

- Training and testing

  We split the data into training part and test part. And then tried three Machine Learning models: Naïve Bayes classifier, k-nearest-neighbors Classifier and Random Forest Classifier. Here are the results:

Classifier	Naïve Bayes	KNN	Random Forest
Score	0.48	0.64	0.69

1.3 Conclusion

- Most entertainment amenities gather in downtown, only a few scattered somewhere else. Waste treatment related facilities scatter in some place close to water. Red points scatter everywhere, but gather in some small groups somewhere.
- Random Forest Classifier with n_estimators=200, max_depth=35, min_samples_leaf=1 obtained the best score around 0.69. We thought it's a pretty good accuracy.

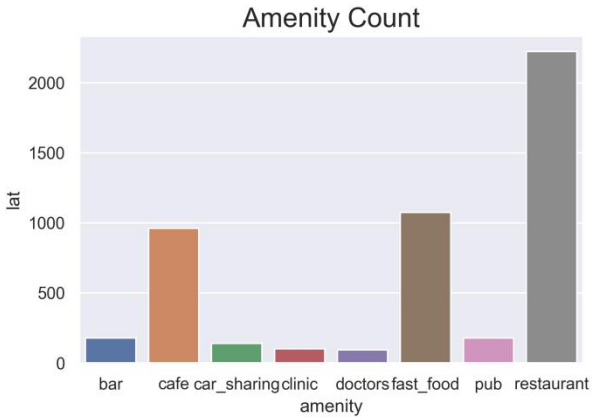
2. Where to choose the hotel

In this part, we tried to use the given data to find the habitable area. We selected some data that we thought was influential, such as restaurants, clinics. In order to get the characteristics of the data, we used latitude and longitude for clustering analysis. Finally, we visualized the results.

2.1 Find a habitable area by amenity

- Data process

  We thought there should be entertainment facilities, transportation facilities and restaurant facilities near the hotel, so we have selected the following data.

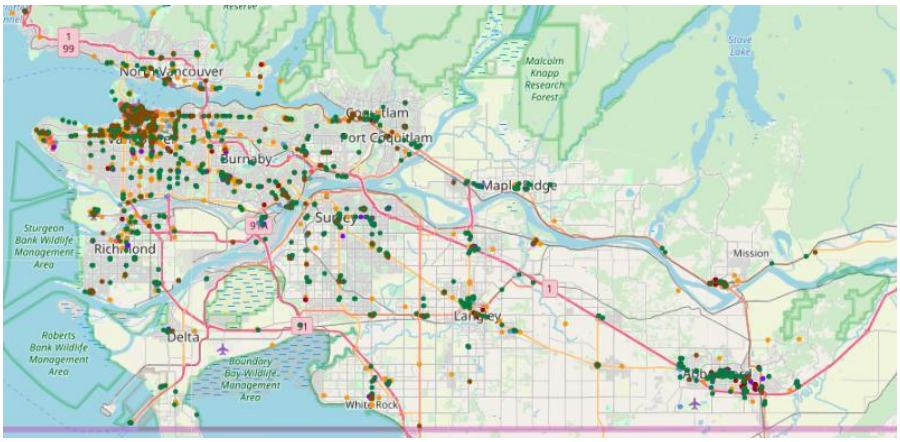


- Data analysis with KMeans

  For each amenity, such as restaurant, we wanted to know what its distribution is and divided it into several regions. So, we used KMeans Model to do cluster analysis.

- Use Map Visualization to locate areas

All amenities



This was the visualization of all amenities, and the color represents the type of amenities. We could find that there are lots of amenities around Downtown, Richmond, Abbotsford.

Clinic Distribution

After observing the overall distribution of the amenities, we could check the distribution of the amenities individually. We thought that the restaurant and entertainment facilities are everywhere, but the number of clinic facilities was relatively small. We thought there should be at least one clinic near the hotel just in case. Above was the visualization of the clinic facilities on the map. We could see that there are many clinics in the three areas we just selected.

2.2 Further precise location based on Airbnb data

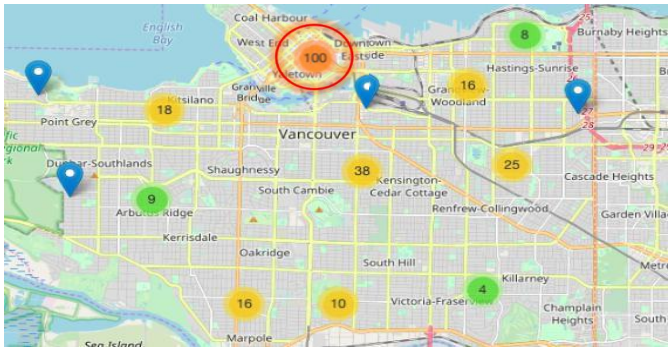
- Data collection and processing

  We found the public dataset of airbnb from 2009 to September 2019, including reviews and listings. We had integrated these data as follows:

  For Listings: Find the listings with minimum_nights <= 2

  For Reviews: Find reviews in 2019 year and select houses with more than 50 reviews in 2019

- Visulisation



- Conclusion

Finally, We plotted these listings on the map. As our Airbnb data was just in ‘Vancouver’, not Great Vancouver, we could only tell something about this small area. From the map, we could see that downtown is the best choice for living. You could find lots of different amenities and lots of Airbnb housing. In addition, from the data of Airbnb, We found that many people choose to book Airbnb in the Downtown, and the income of homeowners in the Downtown is much higher than that in other regions. Besides, I found that Surrey and Abbotford also had many good housing through my mobile phone’s Airbnb application.

3 Chain Restaurant vs. Non-chain Restaurant

Is it true that there are some parts of the city with more chain restaurants than non-chain restaurant? To solve this problem, we used visualization to directly observe the distribution of these two types of restaurant, used cluster to observe the gathering pattern of these two types of restaurant, and used statistical analysis to see if their means and standard deviation of longitudes and latitudes are different. These methods are combined to reach a conclusion.

3.1 Data Collection

In our data, there are lots of items labelled as *'café'*, *'fast food'*, and *'restaurant'*. To specify which of them are chains and which of them are non-chains, we searched in Wikidata for extra data.

This SPARQL query helps to find which restaurants are chain restaurants :

```
SELECT ?restaurant ?restaurantLabel
WHERE
{
      ?restaurant      wdt:P31      wd:Q18534542
      SERVICE wikibase:label {bd:serviceParam wikibase:language "en". }
}
```

Q18534543 is the wikidata entry for *"restaurant chain"*, and P31 represents *"instance of "*. Besides, the *"restaurant chain (Q18534543)"* has three subclasses, *"fast food restaurant chain (Q18509232)"*, *"pizzeria chain (18654742)"*, and *"café chain(Q76212517)"*. Same SPARQL query were made to collect data of fast food chains, pizzeria chains and café chains. In this way, we finally collected 4 lists of names of the chains restaurant stored in .json files: *restaurant.json, fast_food.json, pizzeria.json, cafe.json*.

3.2 Data Process

We did several steps to distinguish the chain restaurants and non-chain restaurants.

1. We first split out the items whose 'amenity' attribute is *'restaurant*' or *'fast_food'* or *'café'*.
2. If the item's name attribute is NaN, we don't take them into consideration, because we couldn't tell whether it is a chain restaurant or not. (63 unknown)
3. If the item's name is in our 4 chain restaurant lists, label it as chain restaurant. (960 in lists)

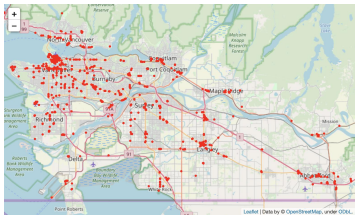
4. We found that among the rest items, there are still some restaurants with the same name. So we guessed that they may also be chain restaurants, but are not so famous to be recorded in wikidata. We labelled them as chain restaurants. (711 number > 1)
5. The rest of the items are labelled as non-chain restaurants. (2526 non-chains)
6. Split the original restaurant data set into to two set, chains and non-chains.

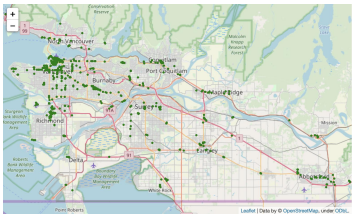
After these steps, we found 1671 chain restaurants and 2526 non-chain restaurants.

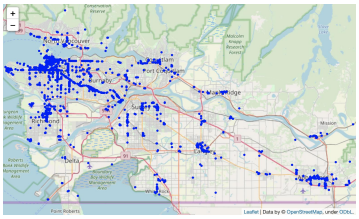
3.3 Data Analysis

- Visualization

The best method to check the distribution is to plot them on the map.


Chains (in wiki data)


Chains(not in wiki data)


Non-chains

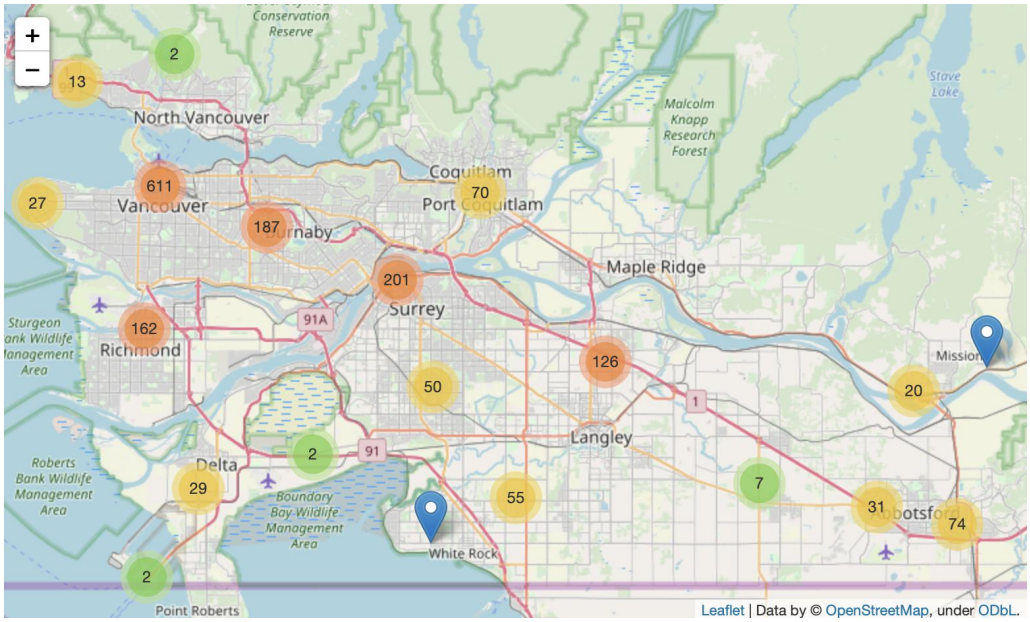
(We plot them on an interactive map. More details may be check in our code.)

From the three plots above, we can see that the distribution of these three kinds of restaurant are similar in general.

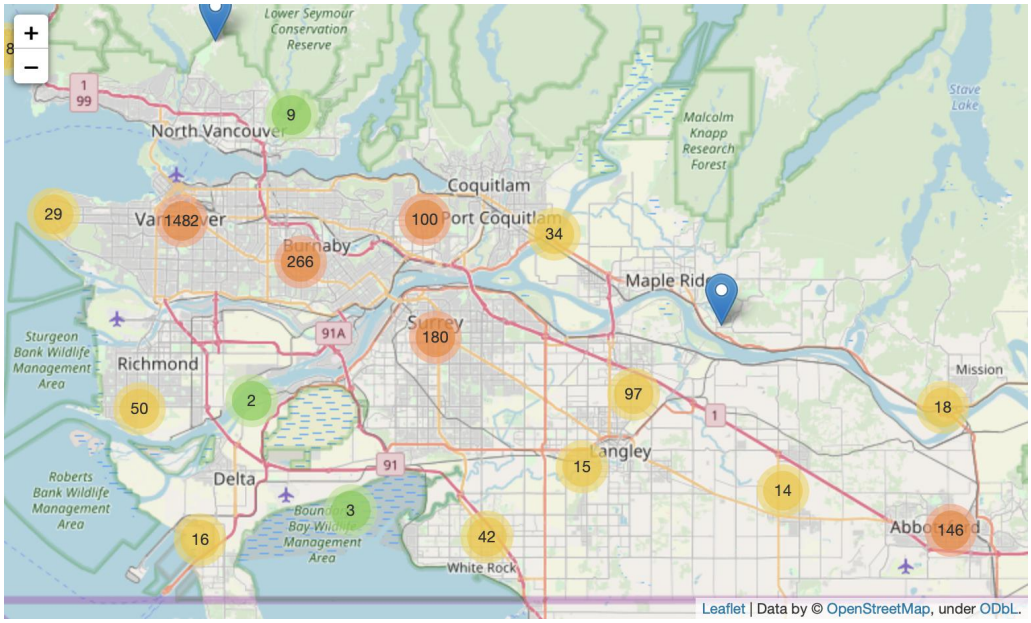
- Cluster

To further check the gathering situation, we did cluster to the restaurants according to their geolocation.

Cluster on chain restaurants

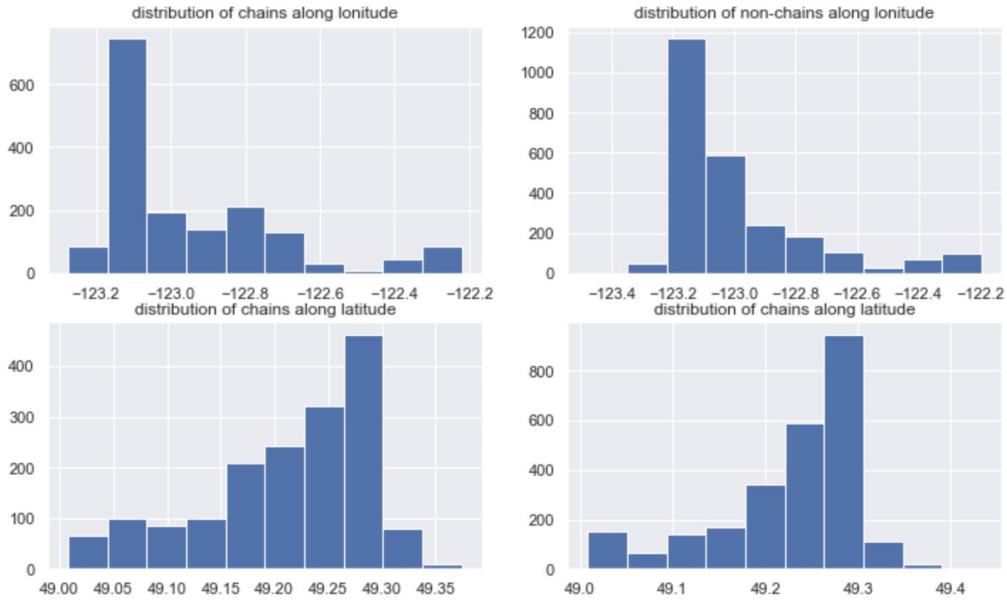


Cluster on non-chain restaurants



We can see from the cluster results that, it seems that non-chain restaurants are more inclined to gather together than chain restaurant. More than half of the non-chain restaurant (1482) gather in downtown area. Chain restaurants gather in downtown area too, but not that serious and is relatively scattered. According to the statistical data from the cluster results above, it seems that non-chain restaurants are more than chain restaurant around downtown, Burnaby, Coquitlam and Abbotsford, but are less than chain restaurant around Surrey, Langley, Shite Rock and some remote areas.

- Statistical data



We can see that their distribution patter along both sides are similar.

We calculate the samples' means and standard deviations:

	Longitude mean	Longitude std	Latitude mean	Latitude std
Chains	-122.9453	0.2429	49.2129	0.07837
Non-chains	-122.9870	0.2298	49.2249	0.07387

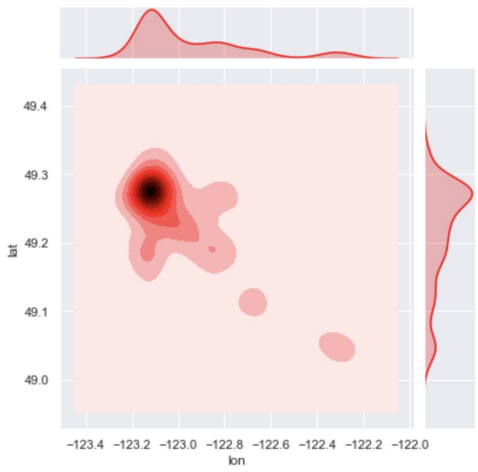
All the statistical data are very close. But we can see that the chain restaurants' standard deviation of longitude and latitude are both larger and non-chain restaurants'. So, we may infer that distribution of chain restaurants is a little bit more scatterd.

We also took 500 items each set as sample and did equal variance tests, and Mann-Whitney U-test. However, the results changes greatly each time, so let's forget about it.

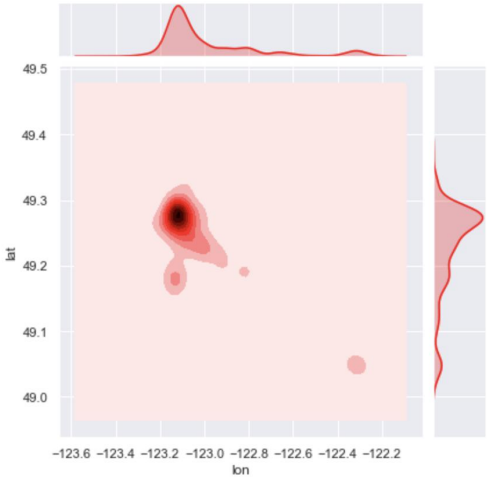
- Density comparison

  We tried heat map to compare their density:

Chains



Non-chains



3.4 Conclusions

- Non-chain restaurants are more than chain restaurants in Vancouver.
- Two kinds of restaurants have similar distribution patterns in general. They tends to gather around some same points.
- The non-chain restaurants are more inclined to gather, while chain restaurants are a little bit more scattered.
- Non-chain restaurants are more than chain restaurant around downtown, Burnaby, Coquitlam, Abbotsford, that is , somewhere like CBD and where there's more people. But they are less than chain restaurants around Surrey, Langley, Shite Rock and some remote areas.

4 Recommended Travel Route

We sift out interesting places, and pick a route according to transportation and distance.

4.1 Filter “Interesting Place”

We used two different methods, and then combined the filtering results of the two methods.

- Data collect:

I found Vancouver's open data about parks on the website “city of Vancouver”. After data cleaning, it was merged with given data (arts_centre, cinema, college and theatre).

- Data Analysis:

Method 1: Using DBSCAN Algorithm with Total Data set

I think if there are many data points in a small area, it can be regarded as an interesting place. (of course, there is also a small possibility that many benches are together) Therefore, I use DBSCAN algorithm, which can find samples of high density and expands clusters from them. Here is my filter result (DBSCAN(eps = 0.0005, min_samples = 20)):

Downtown *Metrotown* *Surrey* *Abbotford*

I use KMeans and DBSCAN to find the more precise area. I found that DBSCAN did a better job, Kmeans seems to contain most of the points, here is my result:

DBSCAN Method KMeans Method

- Conclusion

From the above data analysis, I found interesting places to visit:

Downtown	Metrotown	Granville Island	Parks near Point Grey Road
Mill Lake Park	Impact Plaza	Guildford Town Center	Parks near Grandview Woodland

a. *“Downtown, Metrotown, Granville Island, Parks near Point Grey Road, Parks near Grandviews Woodland”* are in Vancouver.

b. *“Impact Plaza, Guildford Town Center”* are in Surrey.

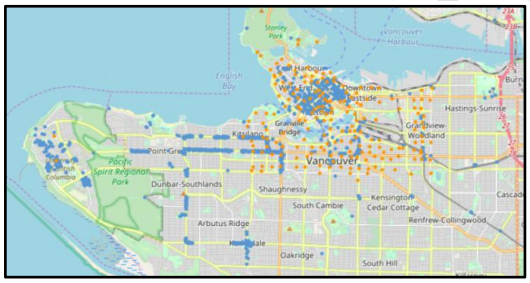
c. *“Mill Lake Park”* is in Abbotford.

4.2 Travel Route Recommended

After sifting through interesting places, we wanted to know how to travel to make travel happier. It is easy to know that at each attraction, we choose to walk the best (because the area is not very large, and the trip cannot always be spent driving). So the question now is how to travel between attractions, because we can't spend a day in a place like Metrotown. I think there are three options: biking, driving, taking a bus.

- Data Cleaning and Visualization: Biking

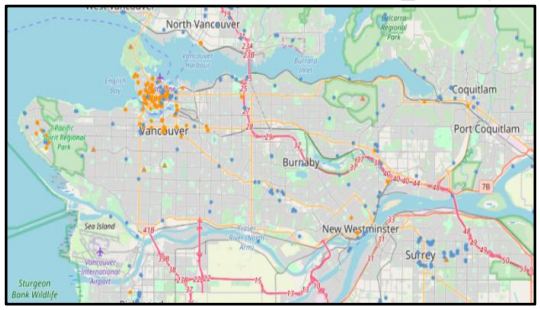
  I filter the data with amenity = 'bicycle_parking' or 'bicycle_rental' and find:



Blue: Parking
Yellow:Rental

- Data Cleanin and Visualization: Driving

  I filter the data with amenity = 'car_sharing' or 'parking' and find:



Blue: Parking
Yellow:Sharing

If you have a car, as you could see, you could park your car easily when you arrive the attraction. And if you want to rent a car, I could give you some recommended travel route.

- Data Cleaning and Visualization: Taking a bus

  I filter the data with amenity = 'bus_station' and find:

- Conclusion:

Bicycle: *Downtown —> Granville Island —> Parks near Point Grey Road*

Driving: *Metrotown —> Parks near Grandview Woodland —> Granville Island —> Downtown*

Driving: *Impact Plaza —> Guildford Town Center —> Mill Lake Park*

Bus: *Downtown —> Granville Island —> Metrotown*

References

[1] Yingjie Hu, *Extracting and understanding urban areas of interest using geotagged photos.* Computers, Environment and Urban Systems Volume 54, November 2015, Pages 240-254

[2] YanTao Zhao, *Mining Travel Patterns from Geotagged Photos*, ACM Transactions on Intelligent Systems and Technology (TIST), 01 May 2012, Vol.3(3), pp.1-18

Accomplishment statement

Jinze Wu:

- Collected park data and Airbnb data from the Internet and used python custom function to extract date and name from the raw data
- Using python third-party library folium to implement map visualization
- Read papers and learned DBSCAN Cluster analysis algorithm, then use it to extract 16 special parts from 15430 data

Yizhou Chen:

- Collected chains restaurants data on from Wikidata using SPARQL query and labelled the restaurants
- Did visualization with folium, clustering and statistical analysis on restaurants data
- Applied concepts of machine learning, using RandomForestClassifier, KNeighborsClassifier, and GaussianNB models to predict the function of an item based on its location
- Came up with the idea of using clustering to solve hotel finding