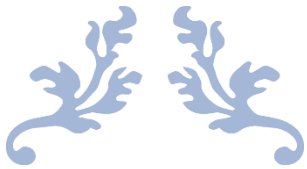


OSM, PHOTOS, AND TOURS

CMPT 353 Computational Data Science

TEAM MEMBERS

YIGAO ZHAO 301415992
TIANYU ZHOU 301426572
ZIHAO XIE 3014433

Table of Content

1. Introduction

One of the most relaxing and significant activities in life is travel. To make tours impressive, people need an informative and well-designed plan, and using data analysis will make this process more straightforward and effective. For this project, through looking for accurate location information (longitude and latitude) of several tourist attractions and accommodations, we mainly used data from Open Street Map Wiki, Airbnb Listing data, and users' photos. This project focuses on the three questions that may be helpful when people are planning a tour in Vancouver:

A. If people were going to choose a hotel (or Airbnb), where would it be? What places have good amenities nearby?
   In terms of this question, hotels built in well-equipped places are preferred by tourists. Firstly, there should be enough restaurants in these places. The traffic should be convenient. Secondly, there should be a certain number of entertainment venues around.

B. Some parts of the city have more chain restaurants: is that true?
   In this section, we investigate the distribution of chain restaurants. Specifically, we would like to find out if some parts of Metro Vancouver have more chain restaurants. If the answer is yes, we will find out these regions. This information would help people to plan their trips.

C. I plan to visit some sites based on photos taken by friends or on the Internet. Are there any suggested routes to do this?
   All the spots in the photos are included in order to find a suggested route. Furthermore, finding out a relatively close way to reach all of sites is necessary. Moreover, people need rest, so it would be better to avoid planning a route with a single thread.

2.Processing of data and techniques used

2.1 Airbnb Choice

2.1.1 Data gathering and cleaning

Extensive Airbnb data set - "Inside Airbnb adding to debate" is aggregate information and indicators of Vancouver listings, compiled on July 6, 2021. Based on this case, we preserved the name, price, minimum stay, and reviews to find the best hotel for the user. In Airbnb's data

listings, we only select properties with the shortest stays less than 3. Moreover, we selected reviews greater than 2020 and more than one review per month. (Shown below)

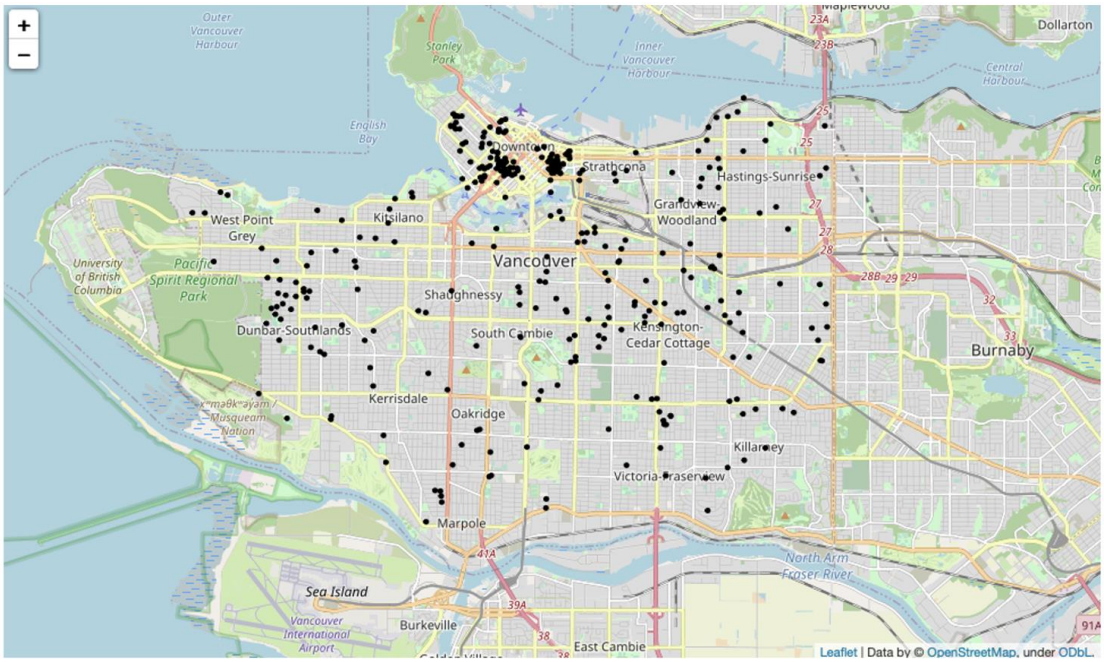


Figure1: Data cleaning

Suppose travellers are more concerned about the quality of their stay, we also did additional data cleaning, which sorted out Airbnb with prices higher than 200 separately. Travelers can choose the more expensive Airbnb. The more expensive Airbnb is better in terms of environment and service. (Shown below)

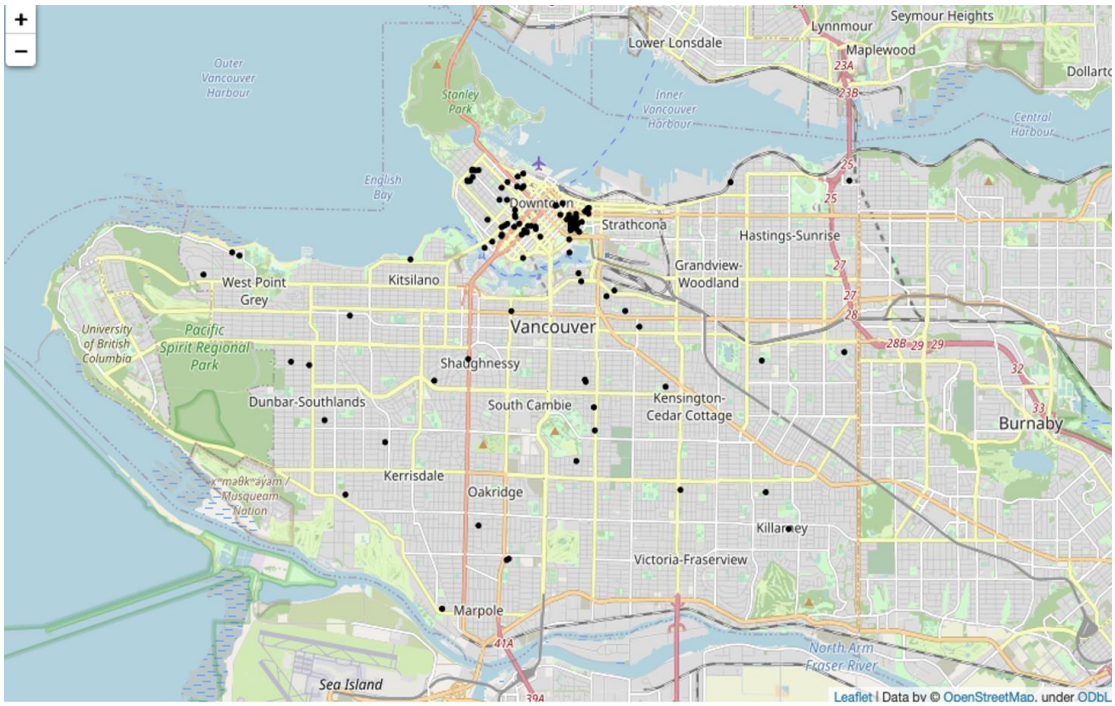


Figure 2: Price higher than 200 dollars per night

2.1.2 Technique used and visualizations result

For this report, we need to integrate the data and then analyse it. The target sites are as follows:

- Food: fast food, food court, restaurant, cafe, ice cream, bistro
- Parking: parking entrance, bicycle parking, motorcycle parking, parking space
- Recreation: pub, nightclub, bar, community center, public bookcase, library, cinema, theatre, arts Center, fountain, social center, conference center, marketplace, spa, events venue, hookah lounge, casino, dance, observation, dive center, toy library, leisure
- Traffic: bus station, bicycle rental, ferry terminal, car rental, boat rental, motorcycle rental

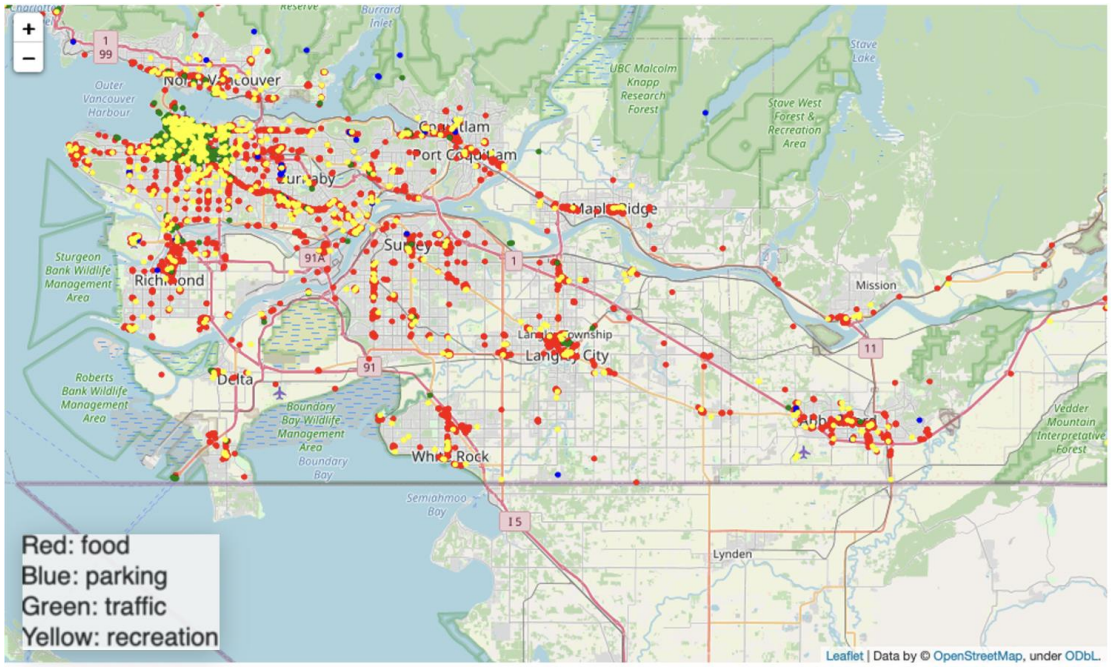


Figure 3

Based on the data visualization of Airbnb and facilities in Vancouver, it is evident that most of the classified facilities are located in downtowns such as downtown Vancouver, Richmond, Surry, and Coquitlam. Regarding the Airbnb data and facility pictures, we found that Airbnb mainly focuses on downtown and Chinatown since these two places possess many facilities. Therefore, it is a good choice for travellers to stay in these two places.

2.2 Restaurant distribution

2.2.1 Data gathering and cleaning

Before we started modeling, we selected the restaurants from the dataset. We discovered three sub-categories under the main category of "restaurant" after looking through the dataset:

"restaurant," "cafe," and "fast food." Because this is a raw dataset, the results of the categories show various informalities. For example, White Spot is a restaurant, Tim Hortons is a cafe, and McDonald's is a fast-food restaurant. This study includes all three categories. In addition, we defined chain restaurants as those with the top ten branches under each sub-category. Following these strategies, we sorted the data by names and count the number of branches. We also filtered out restaurants that possess fewer branches. Ultimately, we excluded the restaurants in Abbotsford and Chilliwack because we merely intended to concentrate on Metro Vancouver in this study.

2.2.2 Techniques used and visualizations result

To solve this problem, we formulated a clustering model and used the K-Mean Sklearn machine-learning model.

After selecting the data, we computed the optimal number of clusters in our K-Mean machine learning model. The number of cluster K is a pre-defined number that cannot be either too large or too small. After doing some research, we decided to use a famous “elbow-curve-method” to determine the number of clusters. “Elbow-curve-method” uses the “Intertia” feature of the K-Mean model, and this method chooses the K value whose corresponding Intertia value does not drop much in the next move. The Intertia computes the squared distance of each sample in a cluster to its cluster center and sums them up. The smaller the Inertia value, the more coherent the different clusters. If N is the sample size, and C is the center of a cluster, the Inertia formula is defined as follows:

$$\sum_{i=1}^{N}(x_i - C_k)^2$$

We ran our program and plotted with K between 1 to 14. The result is shown below:

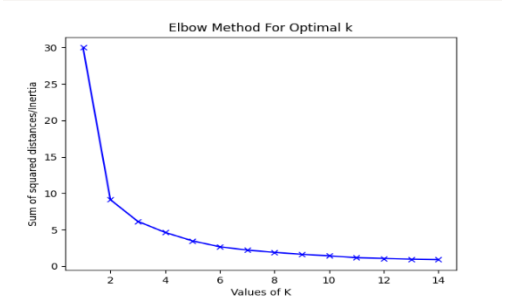


Figure 4: The change of Intertia value with increasing K value

From this plot, it should be noticed that the sum of squared distance drops a little after K reached 7.

K=7 is a good estimation for this study because this value takes the trade-off between precision and utilization of the model. If K is too big, the overfitted model will give us too small regions. If K is too big, the model will tell us a few details within each region. Using K=7, we marked these restaurants on the map with different colours, which is shown below:

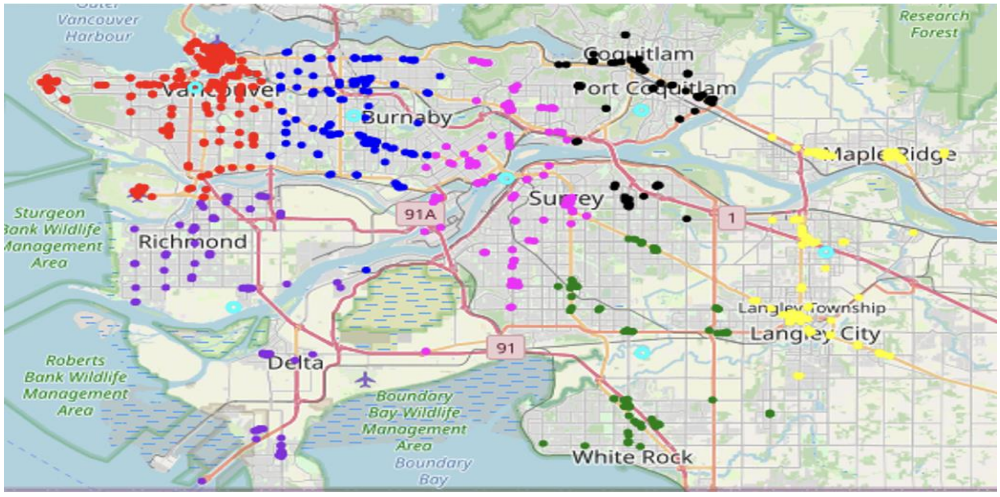


Figure 5 Distribution of chain restaurants in each colour

Cluster	A	B	C	D	E	F	G
Regions	Vancouver west, Vancouver downtown	Vancouver east, Burnaby west	Richmond, Delta	Burnaby east, New West, north Surrey	Coquitlam, Port Coquilam	South Surrey, White Rock	Maple Ridge, Langley

It should be noticed that the centre for each cluster is marked as a light blue circle. These centres are expected to represent the locations where many chain restaurants are nearby. It makes sense that the centre of area A is close to downtown Vancouver. Meanwhile, in region B, the cluster is near Burnaby's Willington Street and is in the middle of Metrotown and Brentwood, which makes sense as well. The centre of region C falls near south Richmond, where few restaurants are nearby. For regions D and E, the centres are near New Westminster downtown and Coquilam city centre, respectively, which is reasonable. The centre of region F is in south Surrey, while region G is in north Langley. For the last two centres, we think they have some deviations, but the differences are not large. We also compared this map with the heat map which can visualize the restaurant distribution better.

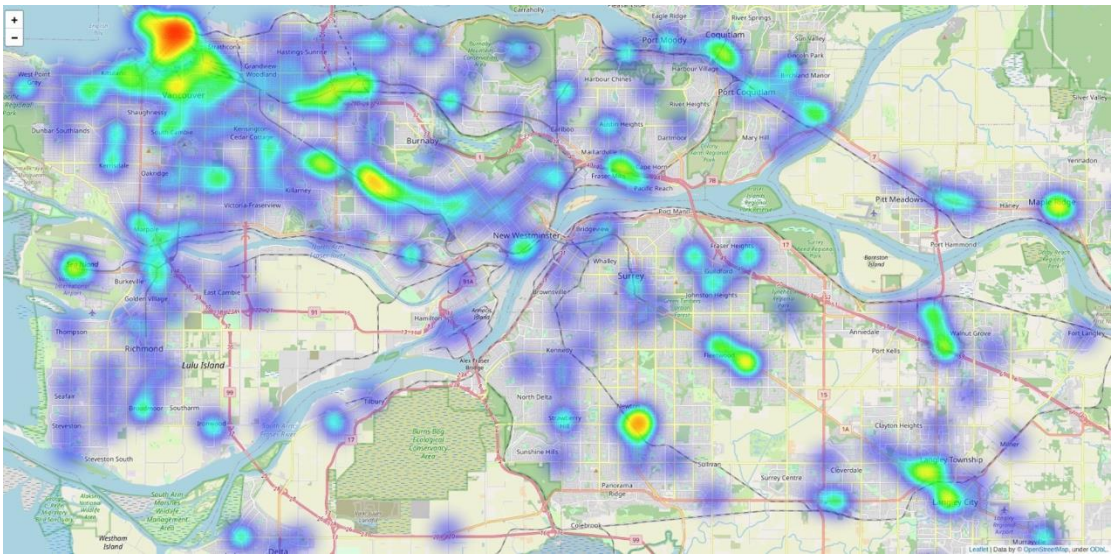


Figure 6 Heatmap for the distribution of heatmap

The results will guide people in making travel plans effectively. People can go directly to these areas if they want to find out where to eat for dinner, because some areas of the city have more chain restaurants than others. Additionally, travellers who also enjoy eating can choose hotels in these areas so that they will only need to travel a short distance to visit a well-known chain restaurant.

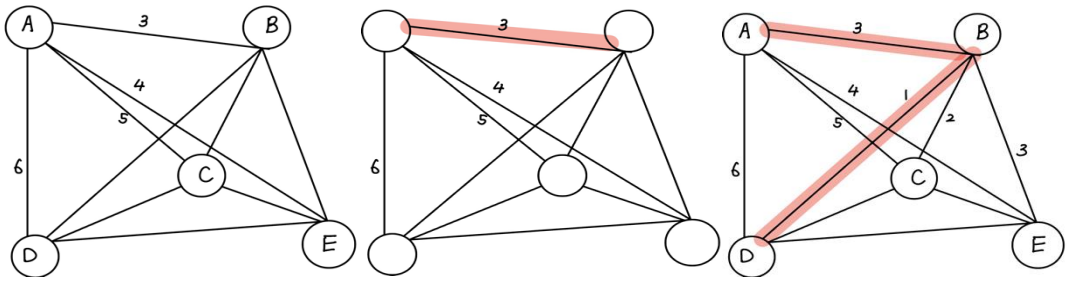
2.3 Route planning

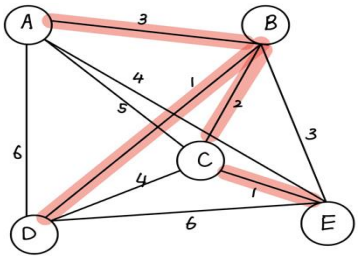
2.3.1 Data processing and cleaning

Through obtaining the EXIF data for the image and using the "Lat, Lon" column in a pandas data frame to convert it, we decide on three decimal points in order to have a comparable, useful data set. Then, we clean up the data by removing any points that are too close to others. It is reasonable that people may find some cute images and download them. Even though these photos are taken in the same place, we do not need to include duplicate ones in our research.

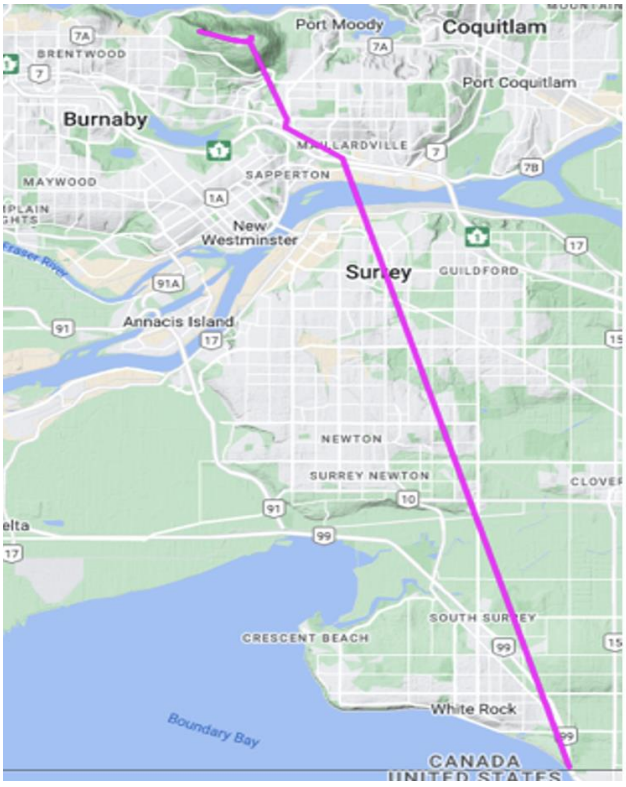
2.3.2 techniques used and visualizations result

In this study, we use the PRIM'S algorithm. The prim's algorithm is greedy, which means "always want the best" at the most basic level. Here we give some basic examples.





We employ five circles to represent the five places people would like to go. The edge between different spots is the distance between them. Starting from one point (in our code, we pick the SFU campus front gate as our starting point), pick the closest one into our “point package”, we have A and B now. Choosing the next closest one to A and B, then add it to the package. Repeat the process until all points are included. Why do we want a tree, not a thread through all the spots? In our real daily life, it is too tiring for travellers to go through many spots without rest. Therefore, we usually pick one place as our fixed rest place, such as, a hotel or B&B. The tree we get above can be excellent guidance to them.



Edge	Weight
0 - 1	0.6622764008760021
1 - 2	0.2339218609179304
6 - 3	26.883056483099683
7 - 4	0.3413908517617622
0 - 5	1.1380679469425405
4 - 6	2.4315715317367457
1 - 7	3.4791386823945567
1 - 8	0.2339218609179304

Figure 7

4. Limitations and future improvement

For Airbnb's choice, we cannot exclude the possibility of geographical factors and bias. Obviously,

most travellers will take Downtown as their first choice. Moreover, there are lots of companies that provide services for vacation rentals and tourism activities. In addition, the security of Chinatown should be highlighted. All these issues should be investigated in the future project.

For restaurant distribution, the centres of clusters C, F, and G are a bit off the town because of outliers. In region C, the centre is “pushed” miles to the south because of a few further restaurants in Delta. Similarly, the restaurants in White Rock affect the overall restaurant distribution in Surrey. From our perspective, it would be more reasonable to put Delta and White Rock in two different clusters. Another factor that we seem to ignore when formulating this problem is the unrealistic distance. All the distances we use are Euclidean distances. In the real world, however, people move along the roads. For example, north Surrey and New Westminster are very close on the map, but it takes longer time for drivers because they must take the Pattullo bridge to cross the Fraser River. In this study, we do not collect the road data, but we can adjust our model by including this factor when road data is available to us in the future. For route planning, if time permits, we can adapt the route to the actual road boundary, which will be more precise. Moreover, sometimes greedy algorithms cannot satisfy people’s travel requirements. For example, it is unfeasible to go to a museum right after we have just visited a funny amusement park. These differences in travel tastes should also be taken into account.

5.Conclusion

In conclusion, our project aims to provide visitors a greater opportunity of making an appropriated and well-designed plan when they visit Vancouver. When users search for hotels on Airbnb, we recommend hotels to them based on the data. When searching for a place to lunch, our Restaurant distribution section could help them make decision. When users decide on a route, our route planning can be a useful suggestion, too. It will save users time when organising a fantastic trip.

6. Project Experience Summary

Our group investigates the problems travelers may have and figures out some basic ideas about how to solve them. When we process and clean data, this heavy work is done by all of us. After finishing each part, we help each other with checking the output and the conclusion.

Yigao Zhao:

Extract EXIF data from the photos and build them with a useful data frame.

Make data clean with outliers and useless data.

Compare different algorithms and use prims Algorithm to find our route.

Tianyu Zhou:

Used Pandas data frame to select Vancouver's restaurants which has the greatest number of branches

Trained SKlearn machine learning model and chose the optimal cluster size by comparing Intertia

Compared the machine learning clustering with the real restaurant distribution (heatmap) and

adjusted the training data by removing outliers
Created visualized data science results by using graphs, tables and heatmaps Integrate all of the parts to finish a report.
Zihao Xie:
OSM data extraction and analysis: Extracting useful data from the given data and the Airbnb data downloaded online. Using data analysis tools such as pandas, combining with python library folium to implement map visualization.