Analysis of chain restaurants in Greater Vancouver

Yufei Wang 301338190
Xinyue Ma 301297142

1 Introduction

This project is based on the data from OpenStreetMap and Wikipedia[1][2]. OSM data contains the latitude and longitude of buildings and their tags. In this project, we will do some analysis of chain restaurants and other amenities in Greater Vancouver.

1.1 Problems

The questions we are going to address:

1. Plot chain restaurants and non-chain restaurants on the map to visualize the density of restaurants.
2. Whether some parts of Greater Vancouver have more chain restaurants?
3. Do amenities related to tourism attract chain restaurants?
4. Are there some amenities that affect the number of chain restaurants?

1.2 Provided data

The original data is from OpenStreetMap project. We use the data provided by the instructor, Greg Baker. The original data has been turned into a .json.gz file with amenities in Greater Vancouver.

1.3 Data about chain restaurants

We use web scraper to get data about the names of chain restaurants in Canada[1] (e.g. McDonald's is not in this list) and international chain restaurants[2] from Wikipedia and save the list of chain restaurants as .csv file, named "chain_restaurants.csv".

1.4 Data Cleaning

The only problem of the provided data is that there are some amenities that do not have names, but this will not affect the analysis we will do in this project, so we do not do any data cleaning for the provided data. For the data that we scrape from websites, there are duplicates in the two lists of chain restaurants, so we delete duplicates. And some names in the original list is not the same as the name in OSM data. For example, "A&W", which is the name in OSM data, is recorded as "A&W(Canada)" in the list of Canadian chain restaurants. We remove the brackets and the content in it. As a result, we get a list of 164 different chain restaurants that Greater Vancouver could have.

2 Methodology, Results and Analysis

We use both the provided data and data about chain restaurants for extracting restaurants. After importing the provided data as pandas dataframe, we select amenities that are fast food, restaurants and cafe to get all restaurants in Greater Vancouver since we find that some chain restaurants in the list have been categorized into fast food and cafe instead of restaurants, such as McDonald's and Tim Hortons. We separate chain restaurants and non-chain restaurants by deciding whether the restaurant is in the list of chain restaurants scraped from websites. We use this basic extracted data for all problems.

2.1 Plot chain restaurants

Problem 1: Plot chain restaurants and non-chain restaurants on the map to visualize the density of restaurants.

2.1.1 Methodology

We screenshot the Great Vancouver from OSM website which has latitude between 49 and 49.5 and longitude between -123.5 and -122. Then we scale every restaurant according to its longitude and latitude. And plot them on the screenshot with different colors (red dots for chain restaurants and blue dots for non-chain restaurants).

2.1.2 Results and Analysis

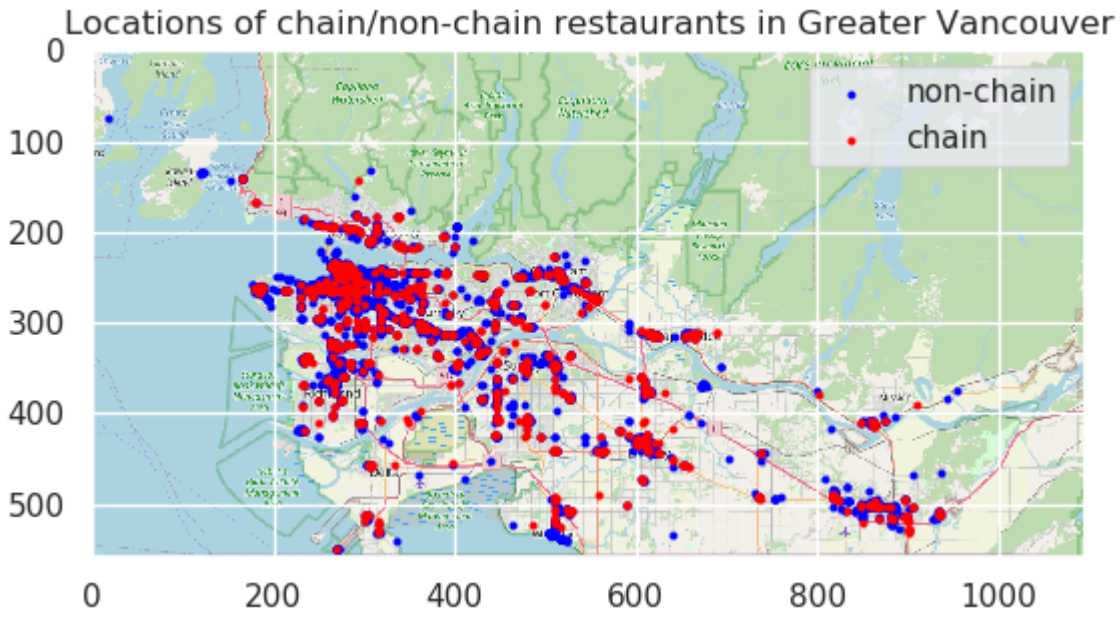


Figure 1: The plot of chain and non-chain restaurants on the map of Greater Vancouver (red for chain, blue for non-chain)

From the result, we can see that there are many more chain restaurants and non-chain restaurants in Vancouver than in other cities. And there are also many restaurants located closely in Abbotsford.

2.2 More chain restaurants

Problem 2: Whether some parts of Greater Vancouver have more chain restaurants?

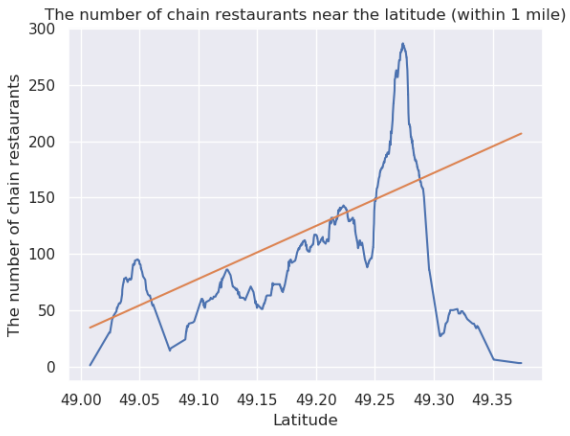
2.2.1 Refine the Problem

The question concerns the relationship between the number of chain restaurants and locations. Since we are going to figure out whether this relationship exists, we decide to use linear regression to find the slope of the best fit line. If the slope of the best fit line is zero then there is no relation between the number of chain restaurants and locations, otherwise, the number of chain restaurants is related to locations. However, the representation of locations is 2-dimension, latitude and longitude, so it is impossible to fit this data into linear regression because linear regression only contains 2 dimensions. Then we decide to figure the relationship between the number of chain restaurants and latitude/longitude separately. Then we will have two fit lines. The original question is transformed into whether the number of chain restaurants is related to latitude or longitude. Thus, if one of the slopes of the fit lines is non-zero, then we can conclude that there exists a relationship between the number of chain restaurants and locations.

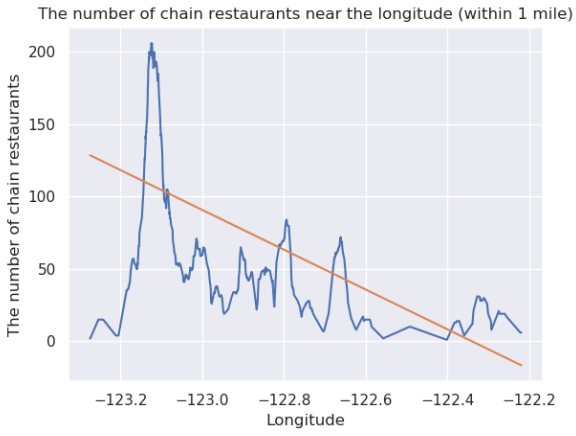
2.2.2 Methodology (Linear regression)

Since we are going to figure out whether there are more chain restaurants in some parts, the density is important. In order to calculate the number of chain restaurants in some parts, we calculate the number of chain restaurants within one mile for every chain restaurant. As we refine the question above, we calculate the number of chain restaurants within one mile for the latitude/longitude of every chain restaurant. Then we use linear regression to find the fit line between the number of chain restaurants and longitude/latitude. Then check if any p-value (the probability of the slope of the fit line is zero) is less than 0.05. If so, we can conclude that there is a relationship between the number of chain restaurants and locations.

2.2.3 Results and Analysis



(a) fig: The number of chain restaurants near the latitude (within 1 mile)



(b) fig: The number of chain restaurants near the longitude (within 1 mile)

Figure 2: The number of chain restaurants near latitudes and longitudes

The two figures in Figure 2 indicate that the number of chain restaurants near the latitude/longitude within one mile. From the figures, we can see that there are global maxima around latitude 49.27 and longitude -123.12.

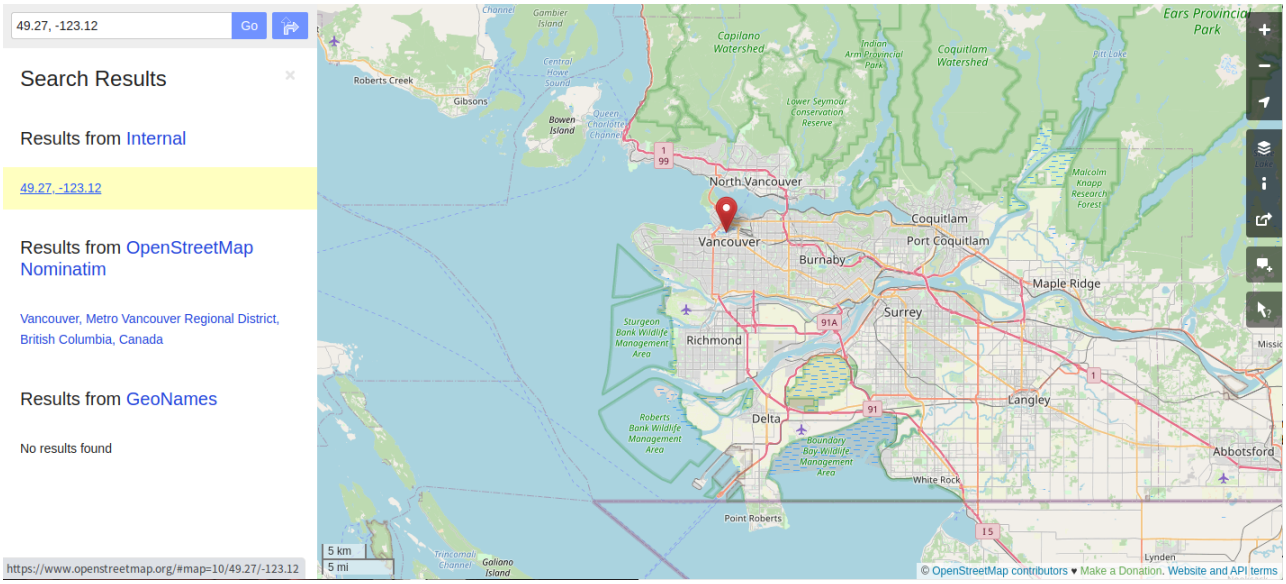


Figure 3: The location of latitude 49.27 and longitude -123.12

After checking the latitude and longitude on the OSM website, we can see that the location is in Vancouver. Compared with figure 1, we can clearly see that there are more chain restaurants in Vancouver which are around the location (49.27, -123.12) we find in OSM.

	Fit line for latitude	Fit line for longitude
p-value	5.450732524201113e-86	9.560425724731466e-84

Table 1: P-values for fit lines

The p-values of both fit lines are less than 0.05, which indicates that there exist relationships between the number of chain restaurants and both latitude and longitude. Thus, there exists a relationship between the number of chain restaurants and locations.

2.3 Tourism and Restaurants

Problem 3: Do amenities related to tourism attract chain restaurants?

2.3.1 Refine the Problem

We are going to figure out whether the amenities related to tourism affect the locations of chain restaurants. We refine the problem to whether there are more chain restaurants than non-chain restaurants near any amenities related to tourism. And more precisely, whether there is a different fraction of chain restaurants near any amenities related to tourism (we will use 'tourism' for 'the amenities related to tourism' for the rest).

2.3.2 Methodology (Chi-Square)

We are going to figure out the fraction of chain restaurants, so we decide to use Chi-Square. We still use the 1 mile as the range to count the number of chain restaurants. First, we choose to count the number of chain restaurants near every tourism within 1 mile. But there is a problem with this method, which is that it is hard to choose latitude and longitude for areas that do not contain any tourism since the chain restaurants located randomly. And the number of chain restaurants would vary dramatically due to the choice of latitude and longitude. Then we decide to check whether there is at least one tourism nearby for every chain restaurant. There is a tag named 'tourism' to indicate that the amenity is related to tourism. We extract every amenity related to tourism by finding the tag 'tourism' in tags. After getting the list of tourism, we count the number of chain restaurant as indicated before. We construct the contingency for Chi-Square indicated in the table. If the p-value of Chi-Square is less than 0.05, then we can conclude that there is a different fraction of chain restaurants with tourism nearby.

	The number of chain restaurants	The number of non-chain restaurants
With tourism nearby	209	1037
Without tourism nearby	785	2654

Table 2: Contingency

2.3.3 Results and Analysis

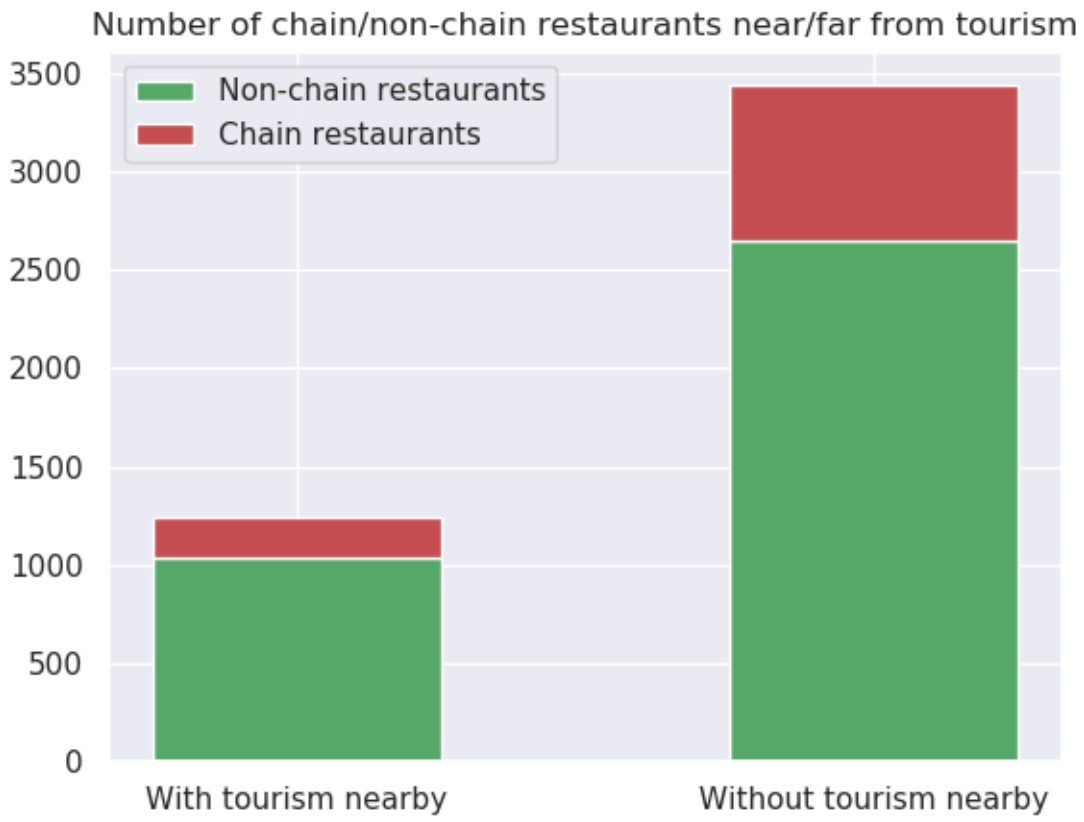


Figure 4: Tourism and chain/non-chain restaurants

P-value	Percentage of chain restaurants near tourism	Percentage of chain restaurants far from tourism
9.128168469857012e-06	0.16773675762439808	0.22826403024134923

Table 3: Result of Chi-Square and percentages of chain restaurants near/far from tourism

The figure provides a visualization of the comparison between the number of chain restaurants and non-chain restaurants. The p-value of Chi-Square is 9.128168469857012e-06, which is less than 0.05. Then we can conclude that there is a different fraction of chain restaurants near tourism. The percentage of chain restaurants near tourism is around 16.8% which is less than the percentage of non-chain restaurants far from tourism, 22.8%. After checking the fractions of chain restaurants under different conditions, we can give a more precise conclusion that tourism does not attract chain restaurants.

2.4 Determining Factors

Problem 4: Are there some amenities that affect the number of chain restaurants?

2.4.1 Refine the Problem

In order to figure out what kind of amenity can affect the number of chain restaurants, we need to know how many chain restaurants are around each type of amenities. Then we refine this problem into whether there exist some amenities that have more chain restaurants around them than other amenities.

2.4.2 Methodology (Post Hoc)

First, we extract non-restaurant amenities from all amenities. In order to do Post Hoc, we need the data to be normally distributed and equal variance. According to Central Limit Theorem (if the data has more than 40 data points, and the distribution is not too skewed, then the data can be seen as the normal distribution), we only get those types of amenities that have more than 40 locations in OSM data. Then for each type of amenity, we count the number of chain restaurants around each amenity in that type. The number of chain restaurants for each amenity is the data point we are going to use for analysis. And for each amenity, we only store the number of chain restaurants around it and its type of amenity. We use those data to find whether there exist some amenities that have larger average numbers of chain restaurants around them. If so, then we can conclude that there is a relationship between the type of amenity and the number of chain restaurants. After getting the Post Hoc analysis, we decide to count the number of rejections for the null hypothesis (null hypothesis: the means of two data sets are same) for each type, calculate the percentage of rejections for each type and sort the types of amenities by the percentage of rejections (descending). We do this step since we want to see how different the average number of chain restaurants for one type is from other types. This can help us get a more precise conclusion. We can try to conclude that one type of amenities is related to the number of chain restaurants if the average number for that type is different from and larger or less than more than half of other types.

2.4.3 Results and Analysis

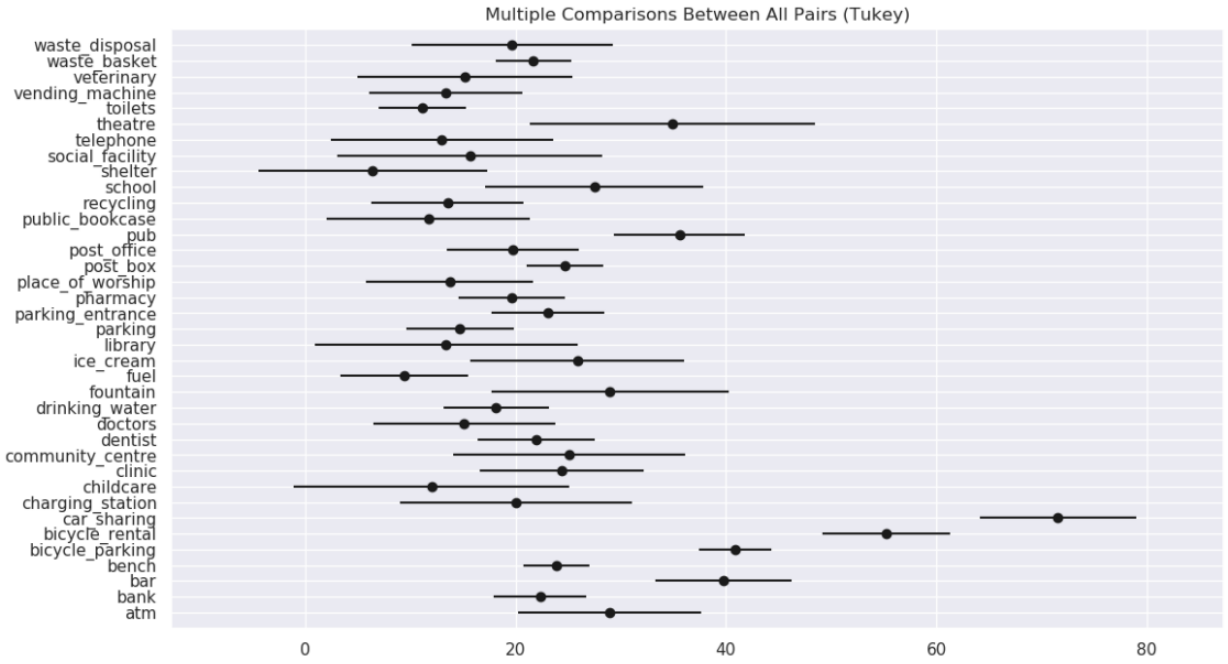


Figure 5: Analysis of Post Hoc

Amenity type	False	True	Percentage of rejections
bicycle_rental	0	36	1.00
car_sharing	0	36	1.00
bicycle_parking	6	30	0.833
bar	8	28	0.778
pub	11	25	0.694
toilets	20	16	0.444

Table 4: First 6 rows of the results of calculating the percentage of rejections

The plot indicates the analysis of Post Hoc. The table indicates that the number of rejections for each type of amenities and the percentage of successful rejections. 'True' means successful rejections.
From the table, we can see that the average numbers of chain restaurants for 'bicycle_rental' and 'car_sharing' are different from that of other types for sure. And from the plot of Post Hoc analysis, we can see that the average numbers of chain restaurants for 'bicycle_rental' and 'car_sharing' are larger than that of other types. We can conclude that the amenities 'bicycle_rental' and 'car_sharing' are related to the number of chain restaurants. According to the table and the plot, we can also conclude that the amenities 'bicycle_parking', 'bar' and 'pub' might be related to the number of chain restaurants.

3 Conclusion

In Greater Vancouver, there are more chain restaurants in Vancouver. Tourism has effects on the locations of chain restaurants and non-chain restaurants. With amenities related to tourism, there are fewer chain restaurants compared to those areas without those amenities related to tourism. And there are more chain restaurants located around bicycle rental systems, car-sharing stations, bicycle parking areas, bars and pubs.

4 Limitations and Discussion

For the first two problems related to the locations of chain restaurants, we have to fit two lines to figure out the results. With this method, the range for counting the number of chain restaurants may be too wide and the results may be affected by the other factor. For example, when we count the number of chain restaurants between two latitudes, there is no limitation on longitude. Thus we have to analyze two plots to figure out the relationship between the number of chain restaurants and locations. It is hard to visualize the result straightforward.
For the third problem related to tourism, we can only know that the fraction of chain restaurants near tourism is less than that far from tourism. From the data we have now, we can not figure out whether tourism has negative effects on the number of chain restaurants or tourism has more positive effects on the number of non-chain restaurants than that of chain restaurants. Both factors can result in a decreased fraction of chain restaurants.
For the fourth problem related to determining factors, we can only know that there are more chain restaurants around some amenities, but it is hard to figure out whether those amenities choose to construct near chain restaurants or chain restaurants choose those locations due to those amenities. It is hard to know whether chain restaurants or amenities are the cause for the other.

5 References

[1] https://en.wikipedia.org/wiki/List_of_Canadian_restaurant_chains
[2] https://en.wikipedia.org/wiki/List_of_restaurant_chains

6 Project Experience Summary

Yufei Wang

- Cleaned data for lists of chain restaurants for extracting chain restaurants from OSM data
- Added a new list to the lists of chain restaurants to extract more chain restaurants
- Reran each program to get new results according to the new lists of chain restaurants
- Implemented the program for Post Hoc to analyze whether there are more chain restaurants near some amenities
- Implemented the program for Chi-Square to analyze the fraction of chain restaurants near amenities of tourism

- Wrote the parts of report related to data cleaning, the problem of tourism and chain restaurants and the problem of types of amenities and chain restaurants

Xinyue Ma

- Scraped lists of chain restaurants from websites for extracting chain restaurants from OSM data
- Implemented the program for plotting restaurants on the map of Greater Vancouver to visualize their density
- Implemented the program for linear regression to figure out the relationship between the number of chain restaurants and locations
- Analyzed the results of the problem of locations and chain restaurants to make a conclusion
- Wrote the parts of report related to collecting data, plot restaurants and the problem of locations and chain restaurants