SFU SIMON FRASER UNIVERSITY
ENGAGING THE WORLD

Simon Fraser University

CMPT 353 D100: Computational Data Science

Project Report

Topic: OSM, Photos, and Tours

Instructor: Dr. Greg Baker

Group Doge

Tian Xiao, 301355843, txa11@sfu.ca

Yuqing Wu, 301368555, ywa292@sfu.ca

Zeyong Jin, 301353174, zeyongj@sfu.ca

August 13, 2021

Table of Contents

I. Introduction 3
II. Problems 3
III. Analysis 3
Problem 1 3
1.1 Data gathering and cleaning 3
1.2 Refine the idea 4
1.3 Techniques 4
1.4 Visualization 4
1.5 Findings 5
1.6 Limitations 5
Problem 2 5
2.1 Data gathering and cleaning 5
2.2 Refine the idea 5
2.3 Techniques 6
2.4 Visualizations 6
2.5 Findings 8
2.6 Limitations 8
Problem 3 9
3.1 Data gathering and cleaning 9
3.2 Refine the idea 9
3.3 Techniques 9
3.4 Visualizations 10
3.5 Findings 12
3.6 Limitations 12
IV. Conclusions 12
V. Limitations 13
VI. Contributions 14
VII. Project Experience Summaries 14
Tian Xiao 14
Yuqing Wu 14
Zeyong Jin 14
VIII. References 15

I. Introduction

This project is using data from OpenStreetMap that is provided by the instructor on the course page [1]. To decrease the running time, especially the reading time, of the program, we uncompressed the file of amenities-vancouver.json.gz in the beginning. The project also uses Wikipedia [2] [3], Inside Airbnb [4] and 2016 Census Data from Statistics Canada [5].

OSM data contains buildings' latitudes, longitudes and kinds of amenities, Wikipedia dataset contains names of chain restaurants, Inside Airbnb dataset contains hotel's latitude, longitude and price, Statistic Canada's 2016 Census dataset contains immigrations' original nationality and the number of them. In this project, we are going to analyze choosing hotels and the relationship between restaurants distributed and immigrations.

II. Problems

1. Taking the current location, how to choose a "good" hotel?

2. Whether some parts of greater Vancouver have more restaurants? And where has the higher possibility to find a restaurant?

3. What is the distribution of restaurant types in greater Vancouver? Does this distribution have anything to do with the ethnicity of immigrants in greater Vancouver?

III. Analysis

Problem 1

Taking the current location, how to choose a "good" hotel?

1.1 Data gathering and cleaning

For this problem, we use data from OSM and Inside Airbnb. Also, we collect some photos in Vancouver with GPS information. Considering that we only need the location and price of the hotel, we just selected the latitude and longitude in the data as well as the price information. To check which hotel is "good", we selected several amenity types of medium size in the OSM data. We did not include restaurants, as the restaurant data is large and will be analyzed more specifically later in the problem.

1.2 Refine the idea

It is hard to check whether the hotel is “good” through visual data, so we select food_court, atm, bus_station, clinic and fast_food from the OSM dataset as said above. By using a heat map to show the number of amenities in each area, the nearest hotel is selected as the "good" hotel.

1.3 Techniques

First of all, we use the ‘exifead’ package to help us to extract the photo’s GPS information and the ‘folium’ package to draw the map.

Then, we draw a heat map to show the number and distribution of these amenities, and we will show all the hotels within a radius of 300 meters with the current location as the center of the circle on the map. After choosing six amenities, the heat map could help us find the location with more amenities (the area with the colour ‘red’).

Given that the factor to judge whether a hotel is good or not is the density of amenities next to the hotel, we also support filtering hotels by price. The user can input the highest price to further filter hotels. We also set max_budget as the input variable for selecting hotels and set “all” for viewing all prices of hotels. But the max_budget only accept integer.

1.4 Visualization

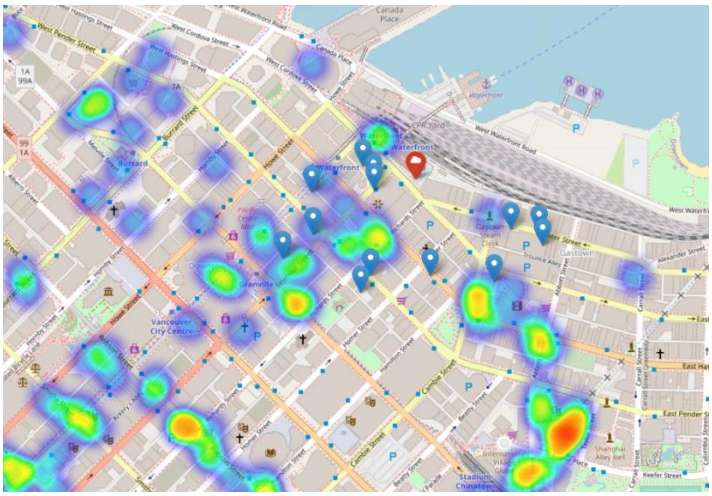

Figure 1: Heat map of restaurants near Waterfront.

Explanation of Figure 1: The ‘red cloud’ icon represents your current location, and the other icons are the Airbnb near you within 300 meters. If you click the icon, the popup will show

the name of Airbnb which makes it easy to get other information in “hotel.csv” (collected hotel information after filtering).

1.5 Findings

According to Figure 1, we know that coloured spots mean the density of amenities in this area. And the hotel which is closer to the coloured spots would be better because there are more amenities near it.

1.6 Limitations

1. The dataset of listing.csv only contains information on the hotels located in the city of Vancouver instead of greater Vancouver, and therefore our development and findings are based on the downtown. This phenomenon may lead to the low precision of the results.

2. If we had more time, we would like to collect all information from the website, so that our findings would be more meaningful.

3. The max_budget only accept interger as input, if we have more time, we would like to make it accept float.

Problem 2

Whether some parts of Vancouver have more restaurants? And where has the higher possibility to find a restaurant?

2.1 Data gathering and cleaning

For data collected from Wikipedia, we select from line1 until the third line from the bottom. Moreover, to make sure their names correspond to the names in the OSM packet, we remove the brackets from the name. Considering that there are too many kinds of amenities in the OSM dataset, we only gather seven kinds of amenities (bbq, restaurant, fast_food, café, bar, juice_bar and food_court) and their location information.

2.2 Refine the idea

By visualizing the overall distribution of the restaurant, then dividing the restaurant into chains and non-chains and then visualizing the details to further determine the distribution of the restaurant. By calculating the average number of restaurants within 0.5km to determine where it is easier to find restaurants.

2.3 Techniques

We use scatter plot, hist2D and histogram to make the visualization clear.

Precisely, to observe whether some parts of Vancouver have more restaurants, we use scatter and hist2D plots to see the density of the restaurants. And for finding where has a higher possibility to find a restaurant, we calculate the average number of restaurants owned within 0.5 km of other types of restaurants other than the above seven types and select the top 20, and drawing a histogram to make it easier to view.

2.4 Visualizations

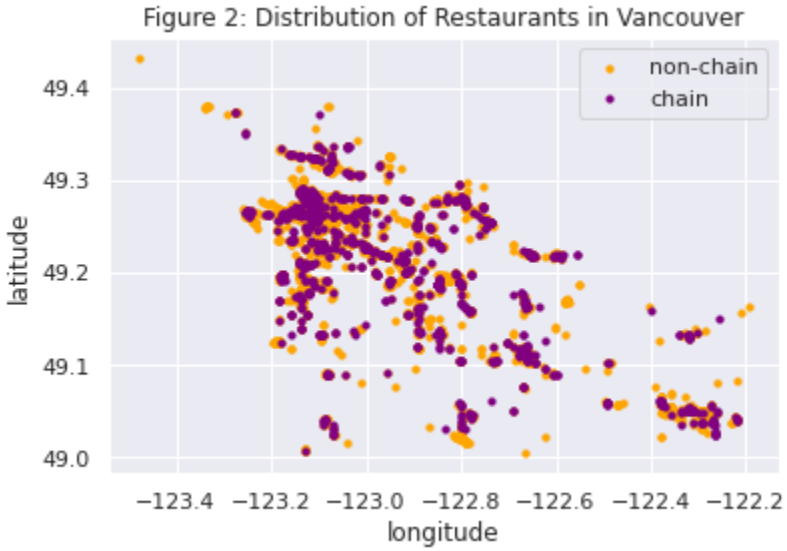

Figure 2: Distribution of restaurants in Vancouver.

Explanation of Figure 2: The purple dots represent chain restaurants, while the orange dots represent non-chain restaurants.

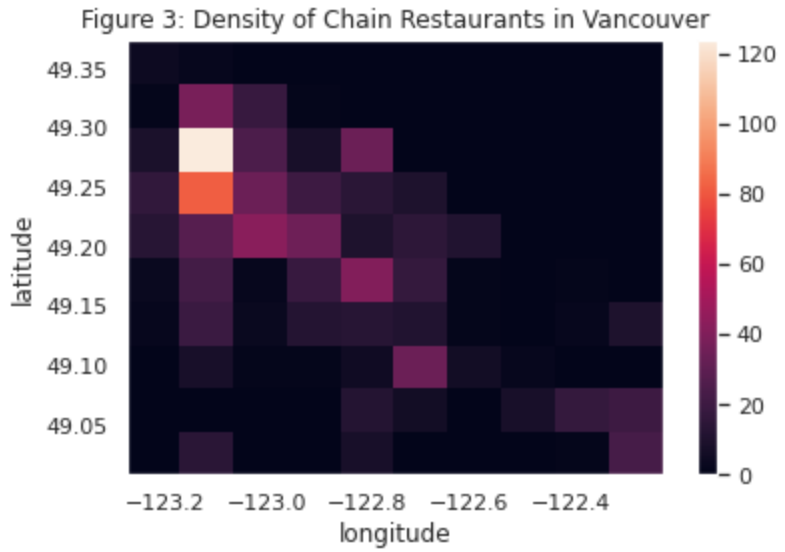

Figure 3: Density of chain restaurants in Vancouver.

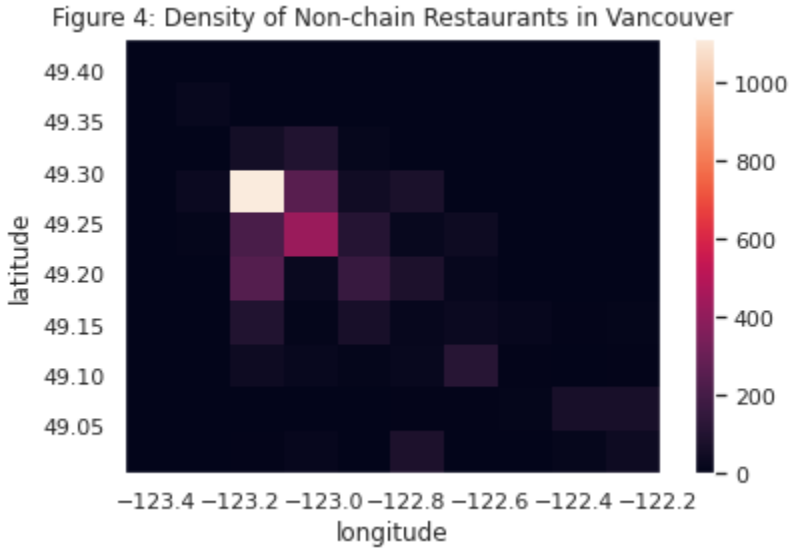

Figure 4: Density of non-chain restaurants in Vancouver.

Explanation of Figures 3 and 4: The colour bars on the right side of the figure shows that the lighter the colour is, the more chain/non-chain restaurants the place has.

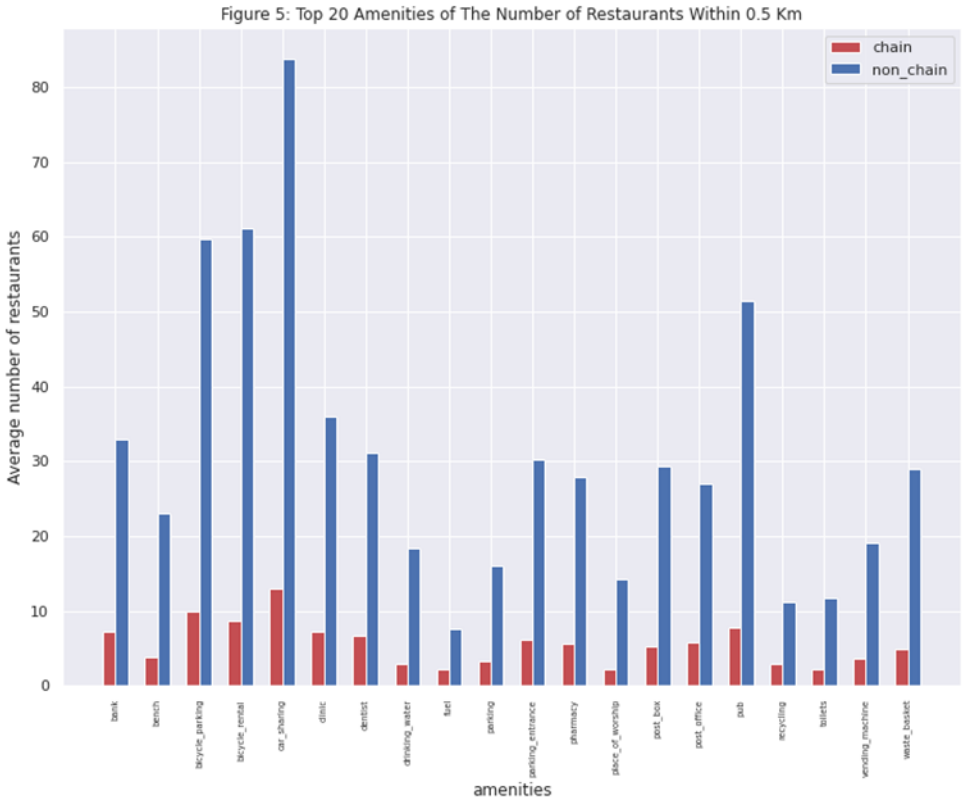

Figure 5: Top 20 amenities of the number of restaurants within 0.5 km.

2.5 Findings

1. From Figure 2, we find out that chain restaurants and non-chain restaurants are most densely distributed at latitudes of 49.2 to 49.3 and longitudes of -123.2 to -123.0. To make the density clearer, we use hist2D to plot chain and non-chain restaurants respectively.

2. From Figures 3 and 4, we can see that at (49.30, -123.2) chain restaurants are most which are around 120, at (49.30, -123.2) non-chain restaurants are also most which are around 1000.

3. From Figure 5, we can easily find that near car_sharing around 0.5 kilometres, there has the most average number of chain and non-chain restaurants. So, near car_sharing will have the highest possibility to find a restaurant.

2.6 Limitations

1. The dataset of amenities-vancouver.json is not completed, and therefore our development and findings may not be able to reflect the authentic situation of greater Vancouver.

2. The dataset of amenities-vancouver.json may be out of date, since amenities may be replaced in a short time. Given our development and findings are based on past data, the results may not be able to reflect the authentic situation of greater Vancouver.

3. If we had more time, we would like to collect up-to-date and detailed information from online resources, so that our findings would be more meaningful.

Problem 3

What is the distribution of restaurant types in the Vancouver area? Does this distribution have anything to do with the ethnicity of immigrants in Vancouver?

3.1 Data gathering and cleaning

To get the cuisine type, we firstly find out all restaurants which have “cuisine” in their tags. Then extract their corresponding cuisine types into a new DataFrame. After this, considering that some of the cuisine types are mutually inclusive, we combine those types that can be combined into one item to reduce the variety of pie charts and make them more viewable.

3.2 Refine the idea

Through visualizing immigrants to Vancouver by cuisine and by country and comparing these figures, we can see whether cuisine distributions are affected by immigrants various.

3.3 Techniques

We use a histogram and pie chart to make the visualization clear.

Precisely, through histogram and pie chart, we show the percentage of immigrants to Vancouver by cuisine and by country. By comparing these figures, we can see whether cuisine distribution is affected by immigrants various.

3.4 Visualizations

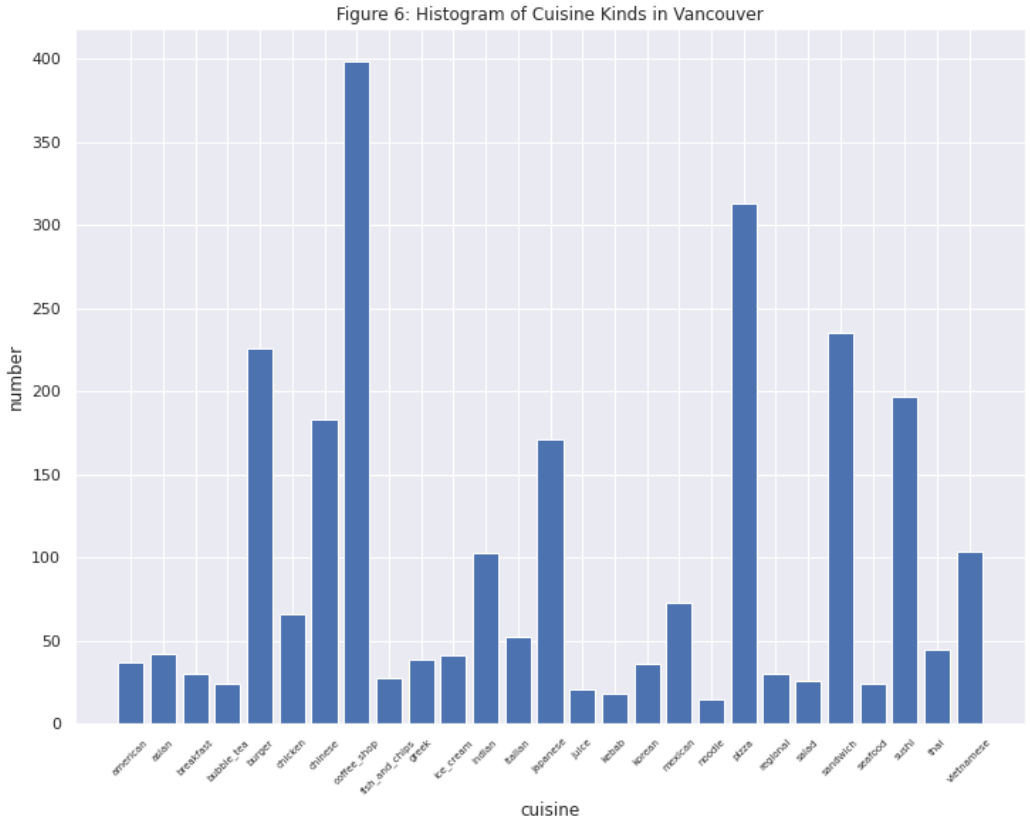

Figure 6: Histogram of cuisine kinds in Vancouver.

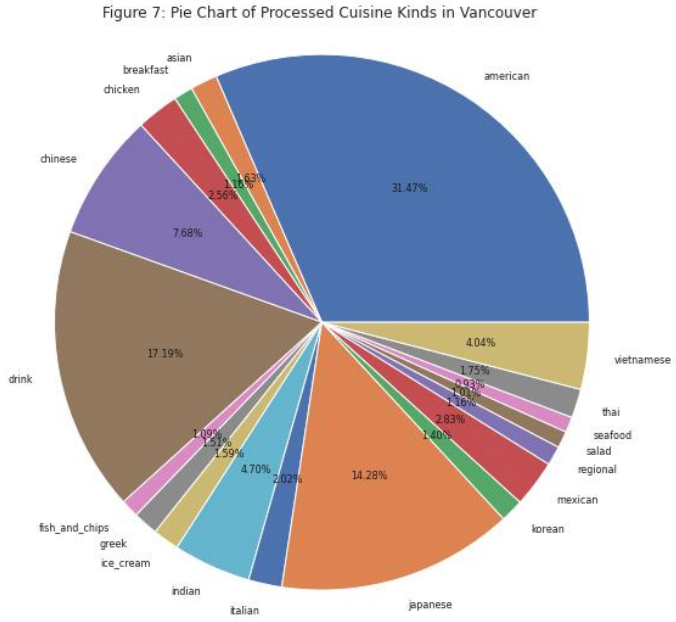

Figure 7: Pie chart of processed cuisine kinds in Vancouver.

Explanation of Figure 7: To make a pie chart having not so many categories, we combined bubble_tea, coffee_shop and juice as a drink, noodle as Chinese, sushi as Japanese, kebab as Indian, pizza, sandwich and burger as American.

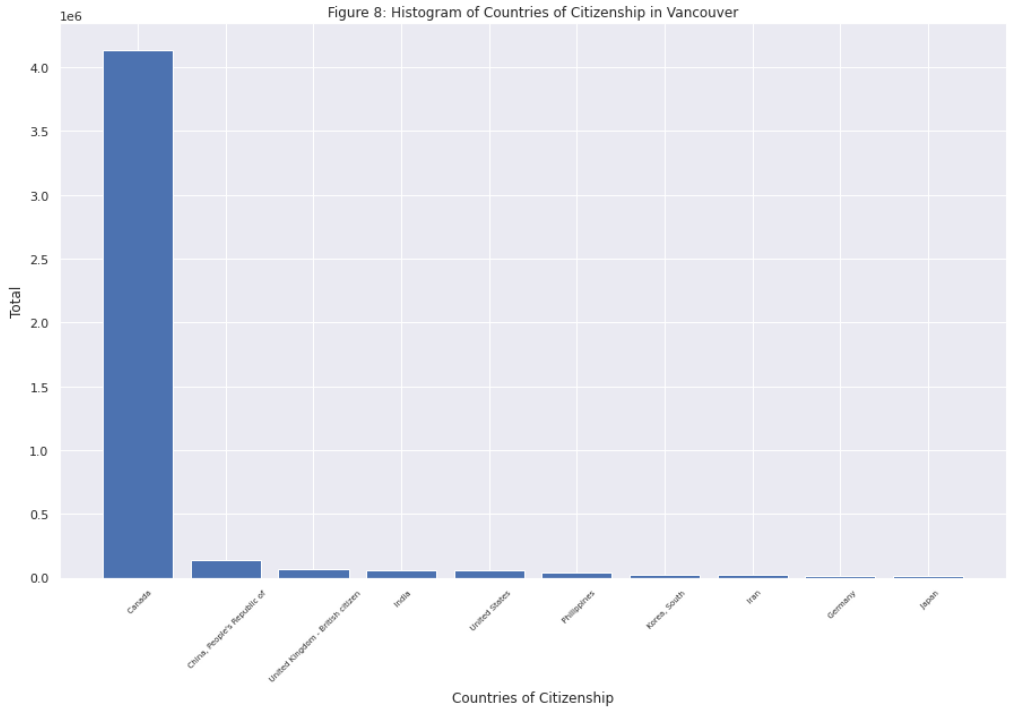

Figure 8: Histogram of countries of citizenship in Vancouver.

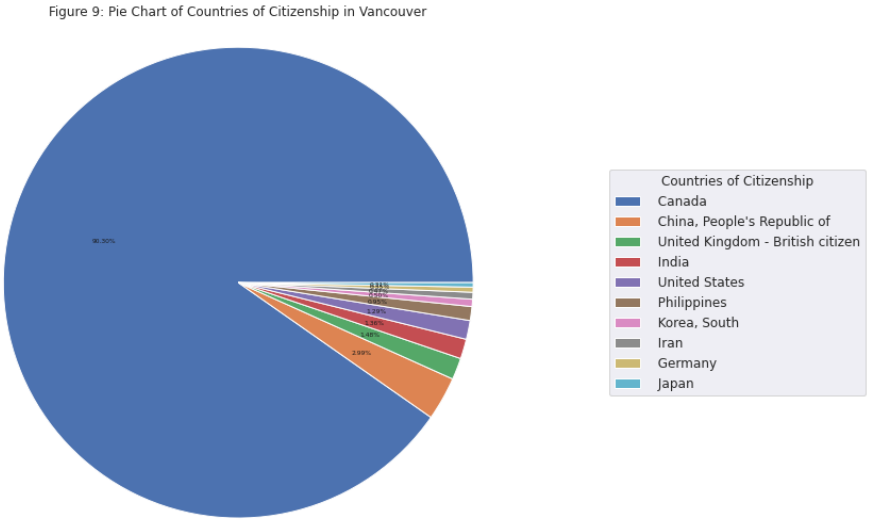

Figure 9: Pie chart of countries of citizenship in Vancouver.

3.5 Findings

From Figures 6 and 7, we can see that American food has the largest share. Drinks are the second one, Japanese is the third.

From Figures 8 and 9, we can find out that the majority of nationality in greater Vancouver is Canadian (over 90%). The percentage of Japanese is less than 0.50%, ranked 10th. However, Japanese cuisine is the 3rd welcome cuisine in greater Vancouver. This phenomenon indicates that the distribution of cuisines has nothing to do with the ethnicity of immigrants in greater Vancouver. And we guess that it is because of the diversity and tolerance of the community of Metro Vancouver. Although there are not too many Japanese here, Japanese culture and food are accepted by the community.

3.6 Limitations

1. We only prove that the distribution of cuisines has nothing to do with the ethnicity of immigrants in greater Vancouver. But for the hypothesis we put forward, we lack evidence. And there might be some other confounding factors that we do not take into consideration.

2. We used the data of the 2016 Census to analyze. But currently, it is 5 years after the data, in other words, the data is out of date. Also, some detailed data, like population distribution in a certain area, are classified and not accessible. So, even if we have a good hypothesis, we cannot find some evidence to support it. And there might be some significant changes in the population now. And therefore, the findings may not be able to reflect the authentic situation now.

3. We should have understood what data is provided by Statistics Canada in the beginning. And based on the available information, we put forward some reasonable hypotheses and try to validate them.

IV. Conclusions

1. The hotel which is closer to the coloured spots would be better because there are more amenities near it.

2. Chain restaurants and non-chain restaurants are most densely distributed at latitudes of 49.2 to 49.3 and longitudes of -123.2 to -123.0.

3. At (49.30, -123.2) chain restaurants are most which are around 120, at (49.30, -123.2) non-chain restaurants are also most which are around 1000.

4. Near car_sharing will have the highest possibility to find a restaurant.

5. American food is the most popular cuisine. Drinks are the second, Japanese is the third.

6. The majority of nationality in greater Vancouver is Canadian (over 90%). The percentage of Japanese is less than 0.50%, ranked 10th.

7. The distribution of cuisines has nothing to do with the ethnicity of immigrants in greater Vancouver, further research is needed.

V. Limitations

1. Data are out of date, or not accessible. Hence the results may not reflect the authentic situations. Or we cannot find enough data to support our hypotheses.

2. Data are not completed. Hence the results may not reflect the authentic situations.

3. Due to the time limit, we do not have enough time to finish all 4 tasks, and we could not put forward more creative and meaningful questions. Our research is limited.

4. We did not fully implement the feature of inputting max_budget for filtering the hotels, so the user experience of running the program may be affected.

5. If we had more time, we would first examine the data, try to find more detailed and up-to-date data. Also, we need barnstorms to solve all the tasks, and put forward some other questions. For each question, we should also provide a basic guideline about how to find and use the data, what analysis should be done, what outcomes do we expect and so on.

6. If we had more time, we would improve and fulfill the features of our program.

7. We should work in cooperation, and maintain no silos of knowledge. Each member should be familiar with other's components so that even if someone met issues, others could help this member. In this way, we could track the progress of the project, and maintain all the works are done by the deadline.

VI. Contributions

1. Tian Xiao finished the part of Problem 1.

2. Yuqing Wu and Zeyong Jin were responsible for Problems 2 and 3.

3. Tian Xiao, Yuqing Wu and Zeyong Jin wrote the report by cooperation.

VII. Project Experience Summaries

Tian Xiao

- Gathered Airbnb hotel data from the internet.
- Made three limitations to select the hotel we want.
- Applied exifread and folium to draw the map.
- Visualized which places have more amenities on the map.
- Used longitude and latitude to locate hotels and the traveller.
- Finished the first draft of a report for Question1.

Yuqing Wu

- Cleaned and gathered data for better analysis.
- Made the data visualized for better analysis.
- Proposed and realized restaurant cuisine and population distribution.
- Organized all outputs into specified folders.
- Searched Downtown Live Photos with Geographic Coordinates.
- Organized the first draft of the report.

Zeyong Jin

- Grabbed data from the web.
- Improved the image storage path to ensure that images can be overwritten.
- Completed readme, license, contribution and changelog.
- Made suggestions for data visualization and helped do the implementation.
- Searched Downtown Live Photos with Geographic Coordinates.
- Organized all the reports and optimize the layout to complete the final draft.

VIII. References

[1] G. Baker, "ProjectTourData," 7 April 2021. [Online]. Available: https://coursys.sfu.ca/2021su-cmpt-353-d1/pages/ProjectTourData. [Accessed 13 August 2021].

[2] Wikipedia, "List of Canadian restaurant chains," Wikipedia, 14 July 2021. [Online]. Available: https://en.wikipedia.org/wiki/List_of_Canadian_restaurant_chains. [Accessed 13 August 2021].

[3] Wikipedia, "List of restaurant chains," Wikipedia, 2 August 2021. [Online]. Available: https://en.wikipedia.org/wiki/List_of_restaurant_chains. [Accessed 13 August 2021].

[4] Inside Airbnb, "Get the Data," Inside Airbnb, 6 August 2021. [Online]. Available: http://insideairbnb.com/get-the-data.html. [Accessed 13 August 2021].

[5] Statistics Canada, "Data tables, 2016 Census," Statistics Canada, 17 June 2019. [Online]. Available: https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/dt-td/Rp-eng.cfm?TABID=2&LANG=E&A=R&APATH=3&DETAIL=0&DIM=0&FL=A&FREE=0&GC=59&GL=-1&GID=1341689&GK=1&GRP=1&O=D&PID=112048&PRID=10&PTYPE=109445&S=0&SHOWALL=0&SUB=0&Temporal=2017&THEME=120&VID=0&VNAME. [Accessed 13 August 2021].