OpenStreetMap Data Analysis

Students: TRUNGLAM NGUYEN (301326848)
QUAN NGUYEN (301326163)

Introduction

OpenStreetMap (OSM) is an open source project that is built by the community of contributors who voluntarily provide and maintain data about places, facilities, transportations, etc all over the world. OSM has an enormous data collection, fortunately, the professor provided us a subset of data containing information of places and amenities in Vancouver. The data contains information about latitude, longitude, timestamp, amenity type (restaurant, cafe, bench, etc), name (McDonalds', Tim Hortons, etc) and tags of each place. In this project, we will apply statistics, machine learning and other data analysis skills to solve interesting problems.

Problems And Ideas

We have come up with some questions and hypotheses about the data:

1. Given a list of locations (latitudes and longitudes) that a person went by, can we guess what places that the person must have seen?
2. Is it true that there are some specific parts of the city where most fast food restaurants are located? Can we visualize these areas?
3. Can we conclude that independent restaurants are more highly rated than fast food restaurants? (We need to somehow get rating scores for each place in this dataset)
4. If someone is about to book a hotel or AirBnb, where should it be? Which places have good amenities nearby?
5. With data of all schools in Vancouver, we determine the area where students live in which commute conveniently to schools. Are we able to see these areas?
6. If drivers stop by at a gas station, there are some amenities drivers expect to have close to the station, can we determine accessibility score of each gas station brands?
7. Can we conclude when is a good time to go biking to work,to school or even to travel around the city as tourists ?
8. Is it true that the bike parkings are checked-in more often during weekdays than weekends ?

Data Collection And Extraction

The dataset contains many data (bike stations, bus stations, benches, etc) that are not quite useful to solve our problems. We decided to only choose data about entertaining places such as restaurants, schools, night clubs, pubs, bars, casinos, etc. We also wrote out the transportation data that contains amenities like parking, fuel, bus_station, etc.
In addition to original data, we found it quite interesting if we can have rating scores for entertaining places. Therefore, we utilized Google Map API to collect rating scores for every

place in the dataset. Google Map API allows us to collect ratings of places based on its location (latitude and longitude) and its name.

Data Analysis

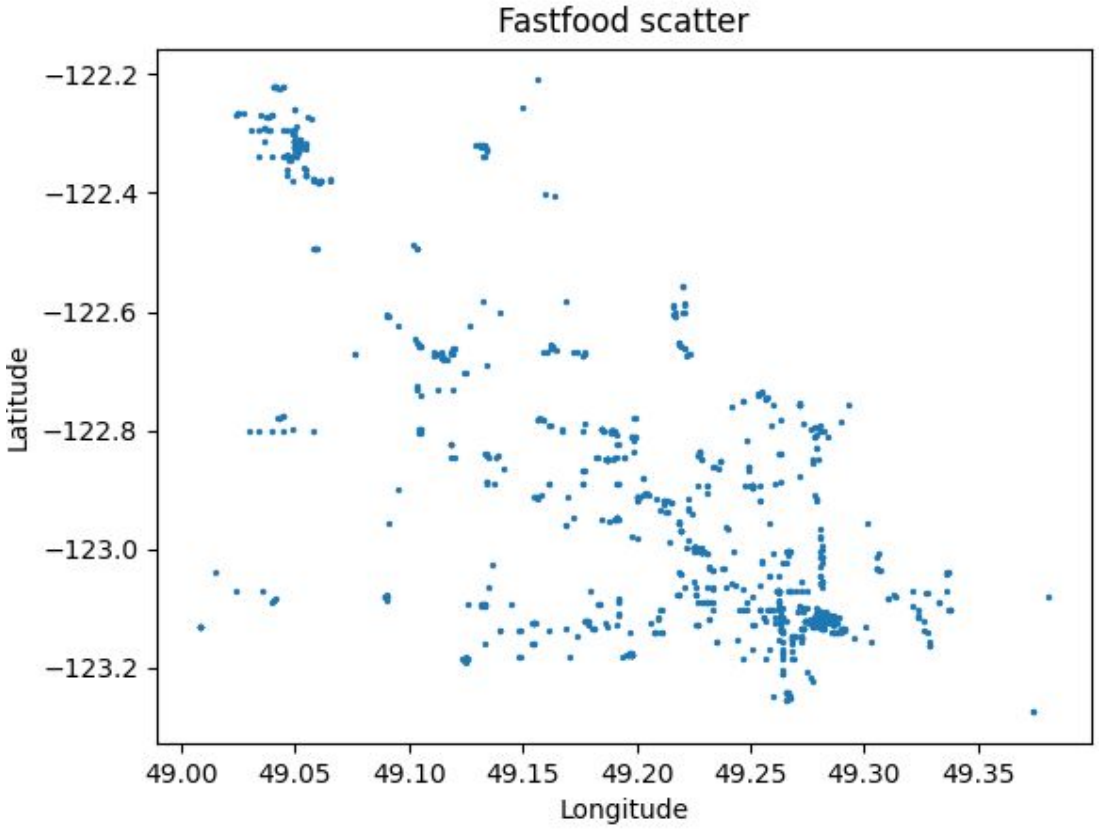
1. Given a list of locations (latitudes and longitudes) that a person went by, can we guess what places that the person must have seen?

Our approach for this problem is that for each location we will find places in our data that are near to this location. We calculate the distance of each data with each location, and only select the ones that are within 50 meters from the location. There can be duplicated places so we only choose the unique ones for the final result. Below is the sample input and the result we got:

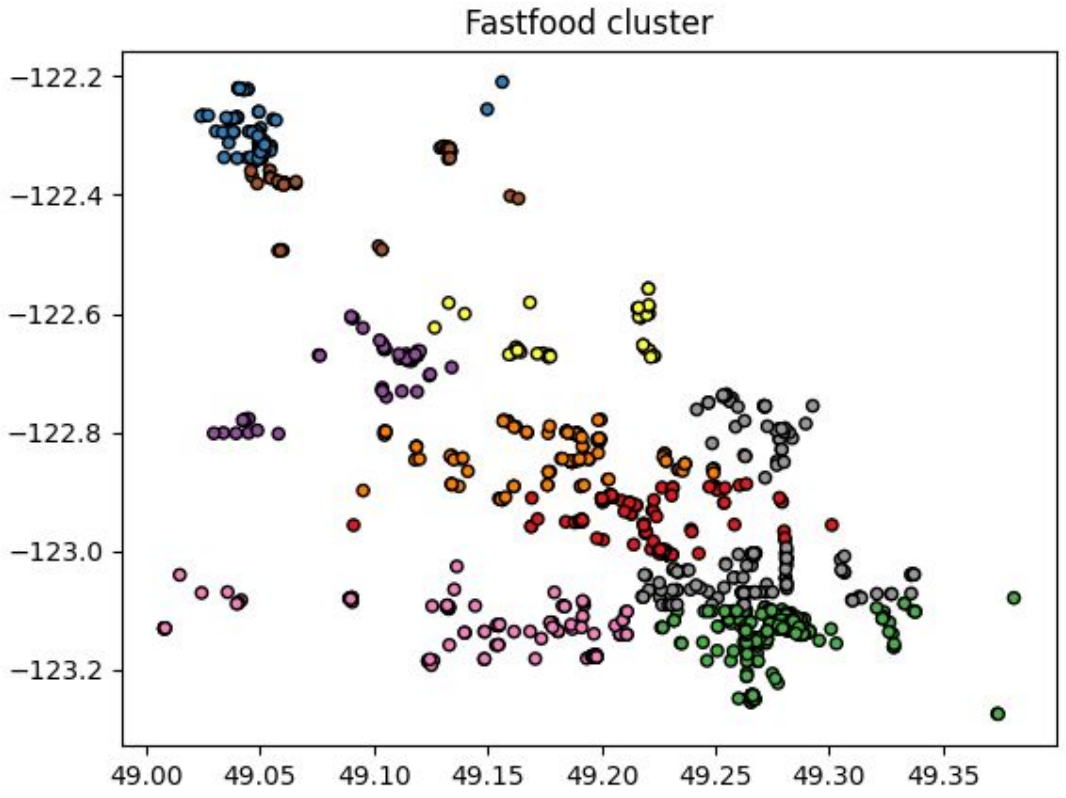
Input latitude	Input longitude	Output nearby places
49.268882	-123.046485	Fuku ramen
49.204056	-123.134982	Wick’s cafe and Cafe de l'Orangerie
49.245791	-122.891637	Pho 99 and Insadong
49.235220	-123.115715	Starbucks
49.211249	-123.140028	Red Star Seafood Restaurant and Sushi King House
49.263003	-123.138347	Sushi Van

2. Is it true that there are some specific parts of the city where most fast food restaurants are located? Can we visualize these areas?

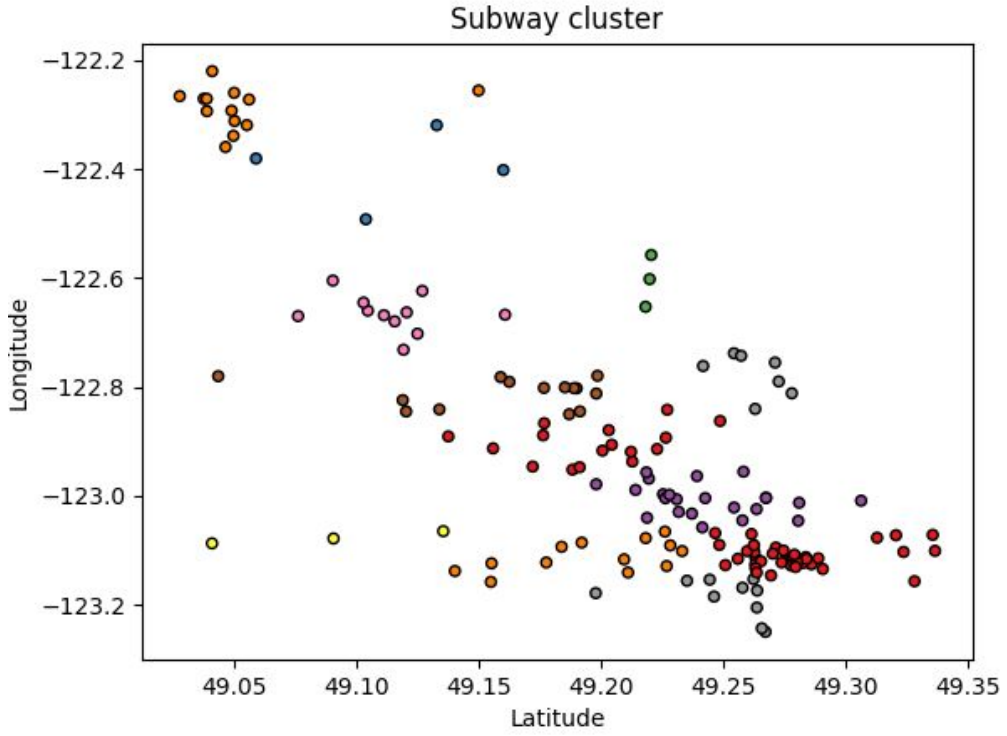
We first tried to get an idea of how fast food restaurants scatter:

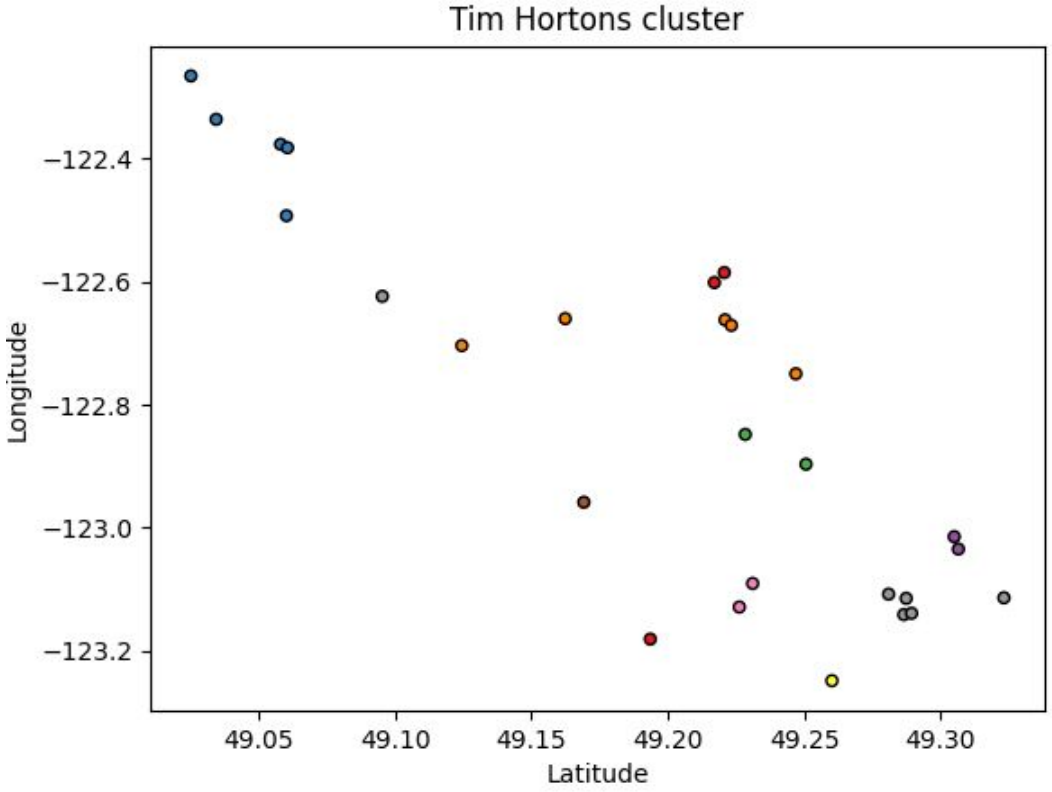


We realized that there might be some clusters in the above chart so we decided to use KMeans clustering algorithm for this dataset:

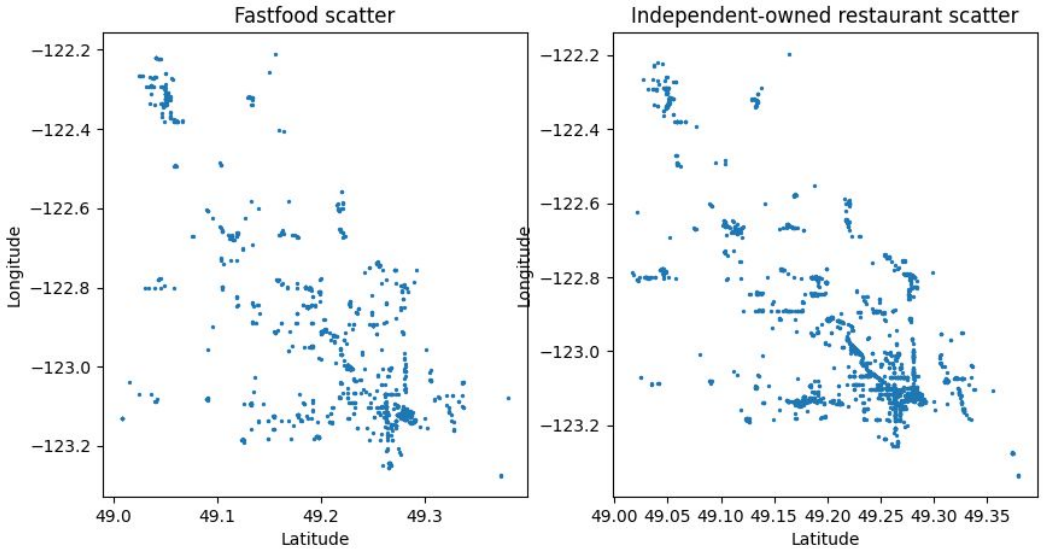


There are roughly 10 clusters in the above graph, each corresponding to a busy area of Vancouver. In addition, we were also curious about how specific fast food branches are scattered. Below is the result for Subway and Tim Hortons:





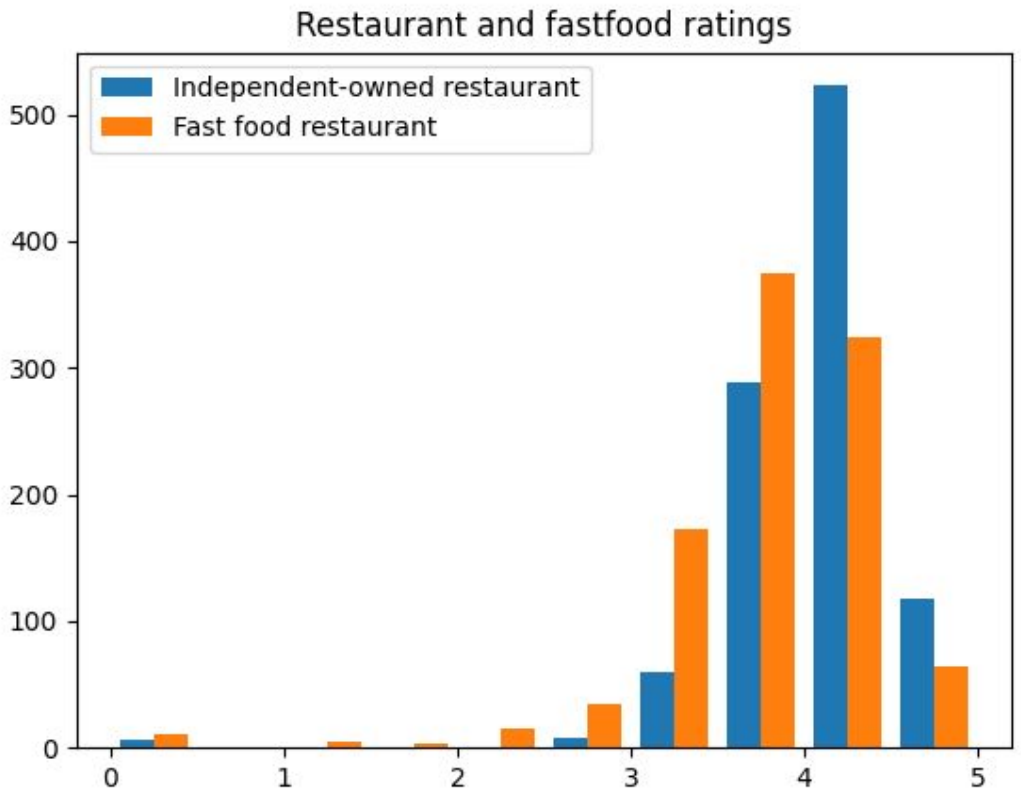
Finally, we want to compare and visualize the density of fast food restaurants and non-chains (independent-owned) restaurants.



It is quite clear that there are more non-chains restaurants than fast food branches but they seem to have similar scatter patterns.

3. Can we conclude that independent-owned restaurants are more highly rated than fast food restaurants?

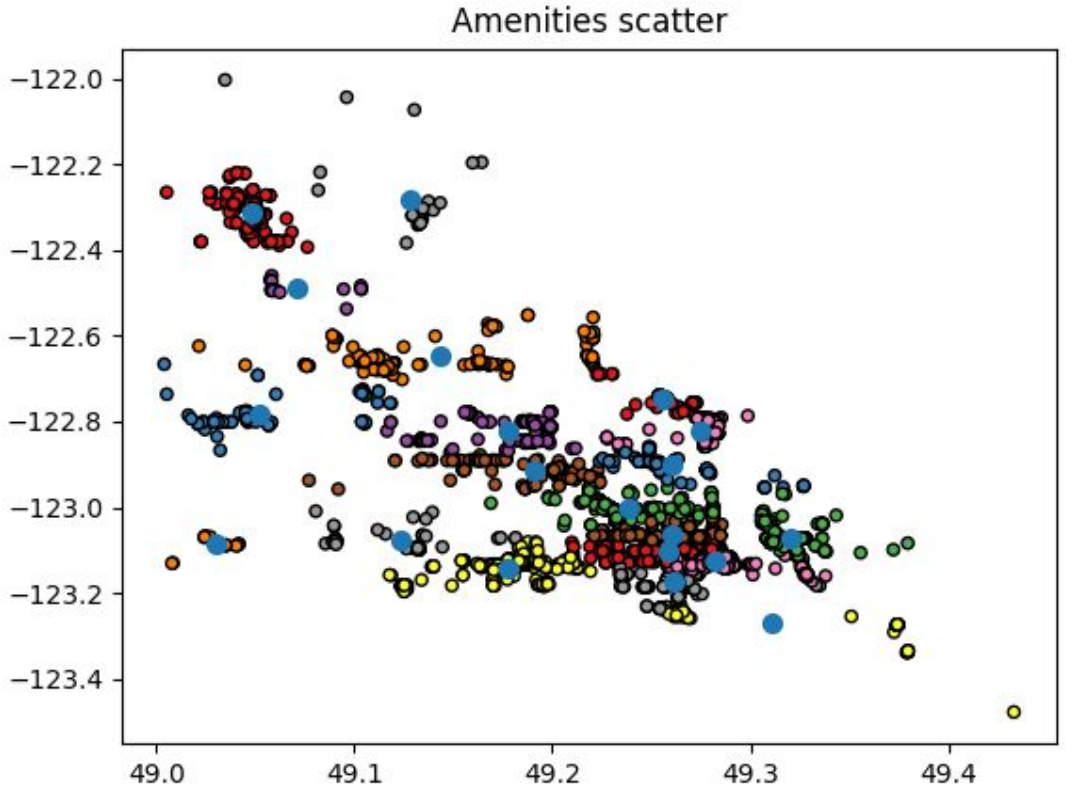
In order to test our hypothesis about the ratings of independent-owned restaurants being higher than fast food restaurants, we extracted ratings of both kinds of restaurants, then applied T-Test. Our null hypothesis is: The average ratings of both independent-owned restaurants and fast food restaurants are the same. The pvalue of the T-Test we got is 5.836947116204433e-54, which indicates that the average ratings must be different. Below is the histogram of ratings:



The average rating for non-chain restaurants is 4.04 and for fast food is 3.71. Based on the result of the T-Test, we can conclude that the average ratings of non-chain restaurants is higher than the fast food restaurants.

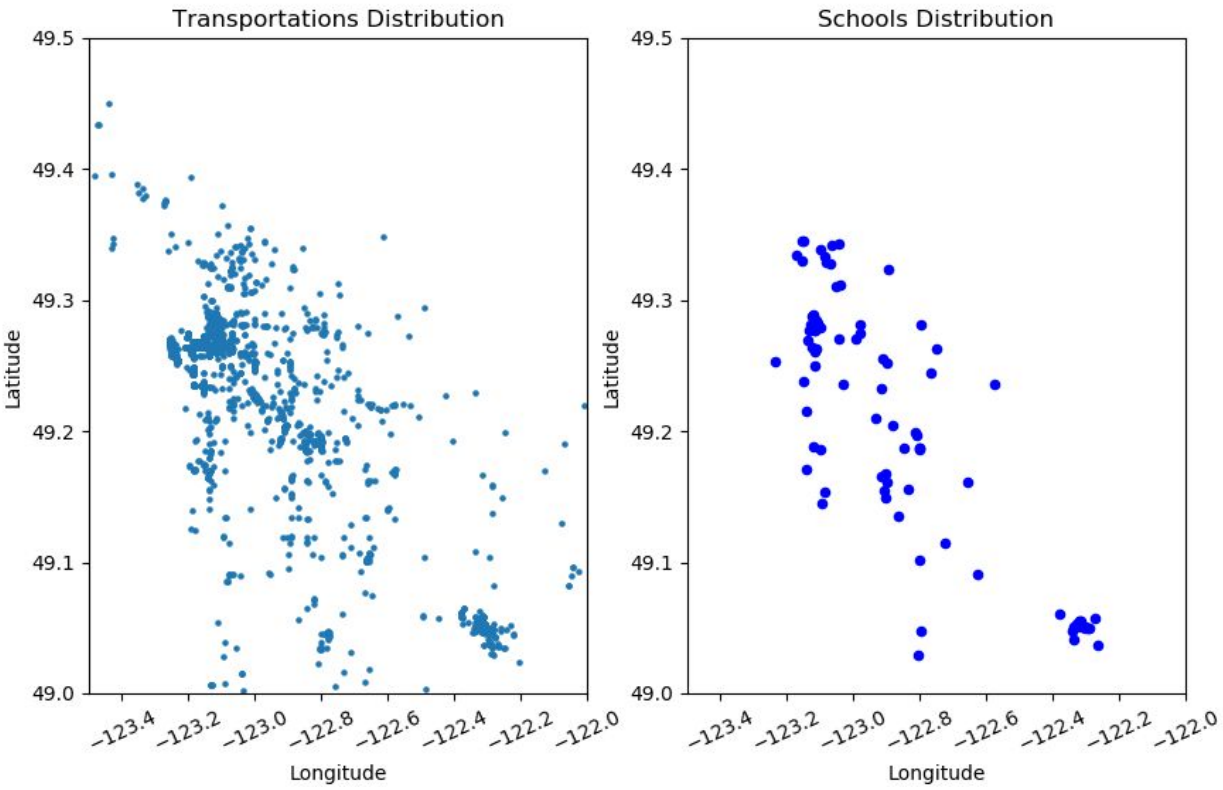
4. If someone is about to book a hotel or AirBnb, where should it be? Which places have good amenities nearby?

When a person books a hotel in the Vancouver area, he/she will mostly care about whether the hotel is near entertaining places such as restaurants, pubs, famous visitor attractions, etc. Therefore, we filtered the data with all of the interesting amenities, excluded other boring places and found the centre of each cluster of these amenities. These centres would be ideal locations for the visitor to find a hotel nearby. The result we got is presented below:

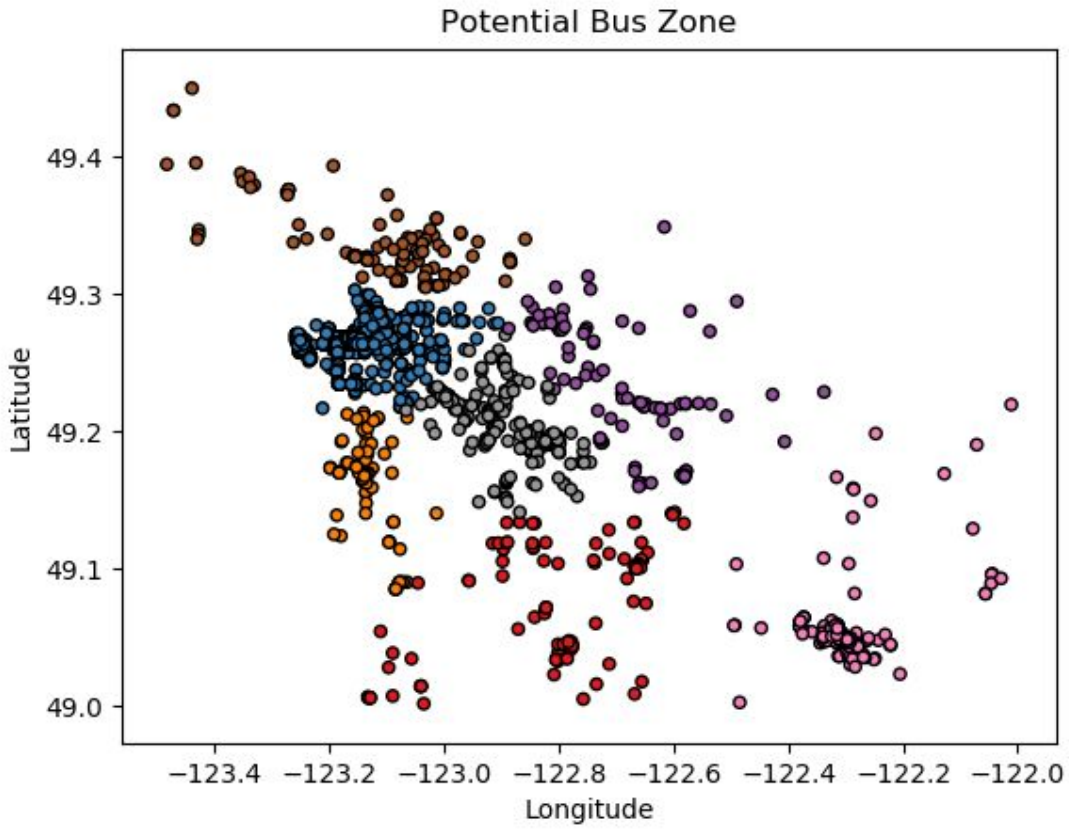


We applied KMeans clustering algorithm to find out all the clusters of entertaining places. The big blue dots are the centers of each cluster and so they are the ideal areas for a visitor to look for a hotel.

5. We are trying to find the relationships between schools and transportations which determine the areas where students are able to commute to school with ease. We use scatter to see transportation locations and school locations:

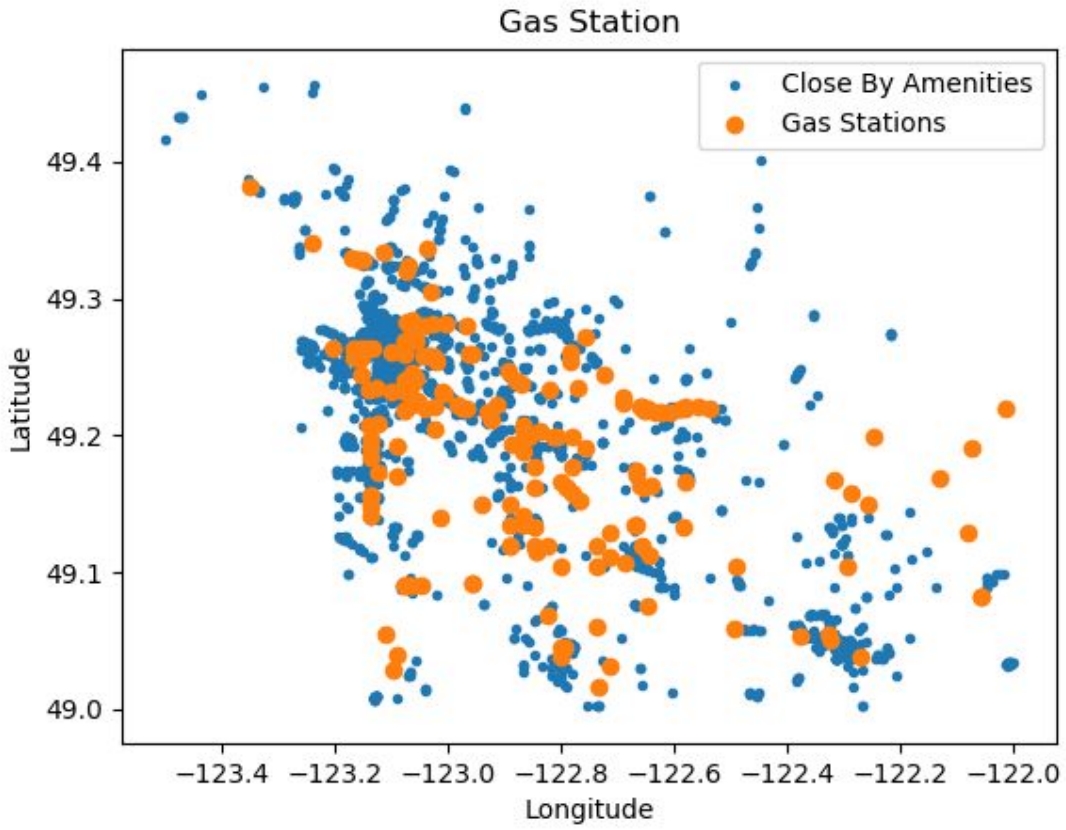


We found that these public utilities were built upon the locations of schools so that students have several ways to commute to school. We also determined which areas are the most beneficial for students to live in order to access schools. We discovered similar clusters in both transportation distribution and school distribution. Therefore, we used MinMaxScaler to transform the features to a range and perform KMean to identify clusters.



As we adjusted the numbers of clusters, 7 clusters are the most reasonable clusters which depicts school regions in Vancouver. The scatter helps the government figure out which regions to improve on as well as build more transportation services.

6. If drivers stop by at a gas station, there are some amenities drivers expect to have close to the station, can we determine accessibility score of each gas station brands?
We extracted gas station data and accessible amenities should be around a gas station. These are cafes, toilets, vending machines, car wash, fast food, banks, etc. We came up with this graph:



Then we used the haversine formula to calculate the accessible score for each gas station all close by amenities. We only mark down those amenities within a radius of 200m. We chose a radius of 200m because we think drivers are easy to see in the area of 200m. Here is the sample result we got from the data:

Longitude	Latitude	Name	Score
49.2302419	-123.0060917	Esso	8
49.3285566	-123.1571712	Esso	7
49.0511229	-122.3239513	Petro Canada	7
49.3279474	-123.1563947	Petro Canada	6
49.2638545	-123.1686333	Petro Canada	6
49.3242961	-123.0719424	Esso	6
49.1626075	-122.6600762	Esso	4
49.0383021	-122.269923	Husky	4

49.2713968	-122.7567067	Petro Canada	4
49.3371539	-123.0385578	Petro Canada	4

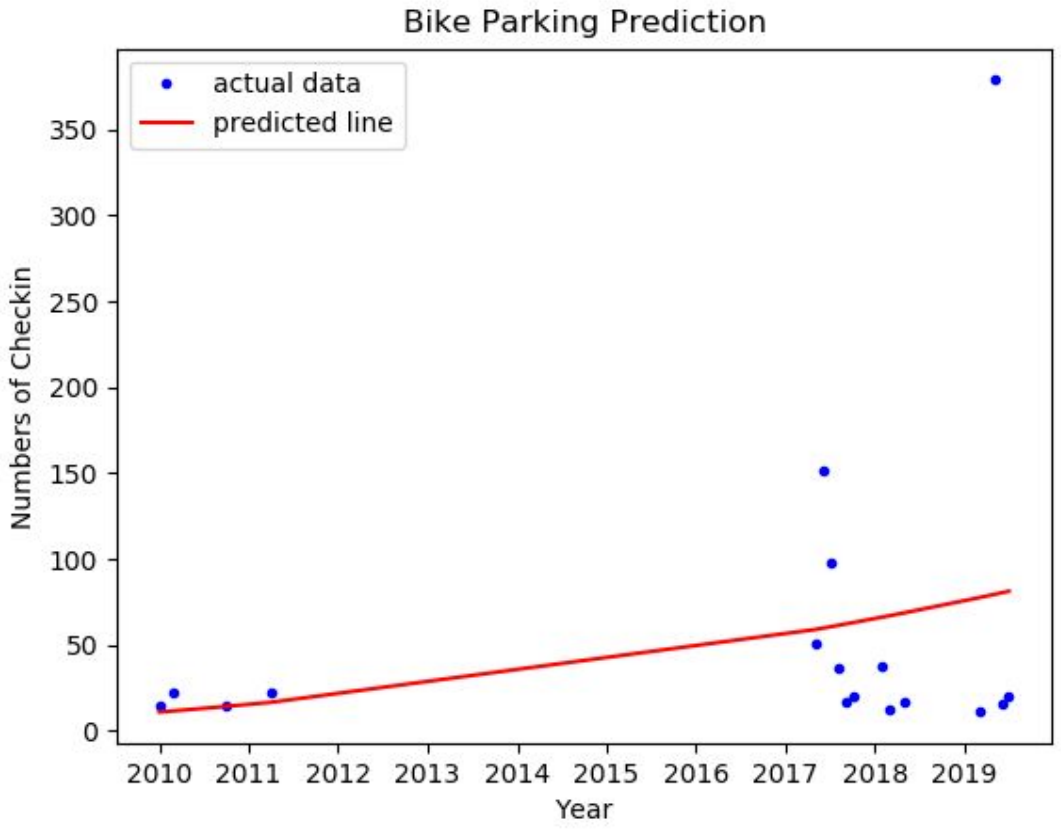
From the table above, we can 2 hypotheses:

1. If the amenities are built before gas stations, we can conclude that Esso and Petro Canada have made good decisions to reside their stations. Therefore, they are mutually benefiting each other.
2. If the gas stations are built before the amenities, the gas company has great power on the market which drives other amenities to be built around them.

7. Can we conclude when is a good time to go biking to work,to school or even to travel around the city as tourists?

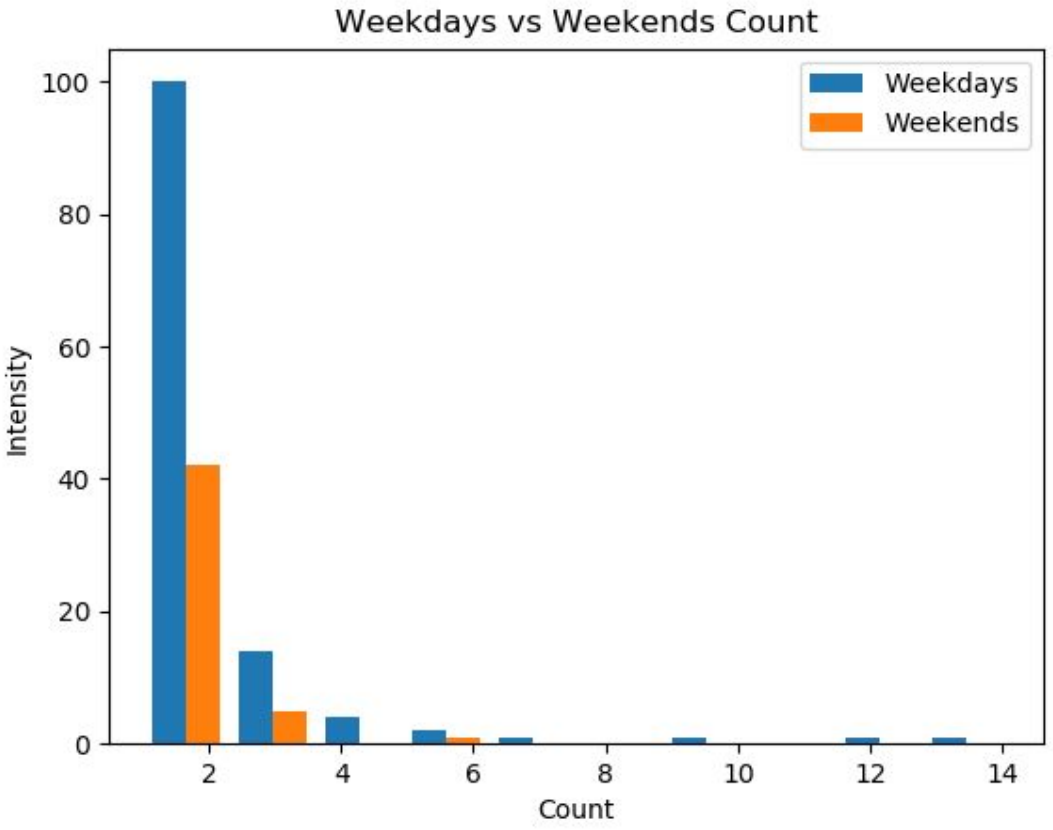
Assume the timestamp is the time when people are checking in the place. Vancouver is a city with an extremely active biker and they are willing to ride to work when the weather outside is good enough. Therefore, we can use the data of biking to conclude the weather of Vancouver based on the intensity of bike parking check in. Here is the graph where we counted the numbers of check-ins per month.

We found that in 2017 and 2019, the number of bike parking check-ins is increasing dramatically. From May, 2017 to September, 2017 and from March, 2019 to June, 2019, bike parkings are recorded with high intensity in use. This indicates that the weather during that time is good for biking. There is not a clear reason why summer 2018 the bike parkings are doing well. However, we are going to predict the demands of bike parkings in the following years so that the city would consider whether to develop more bike parkings as well as build more roads for bikes. We used machine learning tools such as Linear Regression to predict its pattern and we came up with the following result:



8. Is it true that the bike parkings are checked-in more often during weekends than weekdays ?

We dived deeper into the bike parking data above by dividing into 2 categories: weekdays and weekends count. Using datetime library and weekday function, we divided them into 2 dataframes and printed out the histogram with check-in count. Here is the result from it:



We used T-Test on these two data frames which returned a value of 0.22787. We can state that both data frames are having the same mean value but they both failed the normal tests. Therefore, our hypotheses are not satisfied due to failure of normal tests.

Conclusion And Retrospect

OSM data is a big collection of data that we can analyze and make lots of good insights from data. We come up with several questions derived from the data and perform data science analysis in order to answer our targeting questions. There is not always enough content from data to draw conclusions to some problems but we can hope that our predictions are going in the right direction. We attempted to predict the upcoming quantity of bike parkings in the following years but the given data did not provide enough information to confidently write out predictions. Bike parking is just an unfortunate problem that we confronted, we actually collected ratings from restaurants which we were able to make comparisons between independent-owned restaurants and fast-food chains. We applied statistical tests to determine its independence test as well as normality test.

Individual Work And Experience

TRUNGLAM NGUYEN:

- Extracted useful data from the dataset
- Collected rating data by using Google Map API
- Solved problem 1-4 by using statistic, data analysis and data visualization with python libraries: numpy, pandas, scikit-learn, spark, matplotlib
- Refactored and documented source code
- Wrote reports with detailed explanation
- Source code contribution: entertainments_nospark.py, entertainments_spark.py, fastfood.py, fastfood_scatter.py, fastfood_vs_restaurant.py, get_ratings.py, get_ratings_spark.py, nearby_amenities.py

QUAN NGUYEN

- Extracted transportation-related data from the data set
- Solved problems 5-8 by using statistical tests, machine learning techniques, clusters with pandas and numpy libraries as well as sklearn, spark and matplotlib.
- Wrote reports with comprehensive explanations.
- Source code contributions: bike_parking.py, commuter_schools.py, fuel_stations.py, transportations_schools.py, weekend_weekday.py