CMPT353 Final Report - The Best Live Place in Burnaby

Preface:

In this final report, we will explore and analyze the best live place in Burnaby. To achieve this, we divided the work into three parts:

Part 1: Jingxiao Zhang

- In this section, I concentrated on the considerations of a prospective home buyer.

Problem solved:

- Questions:

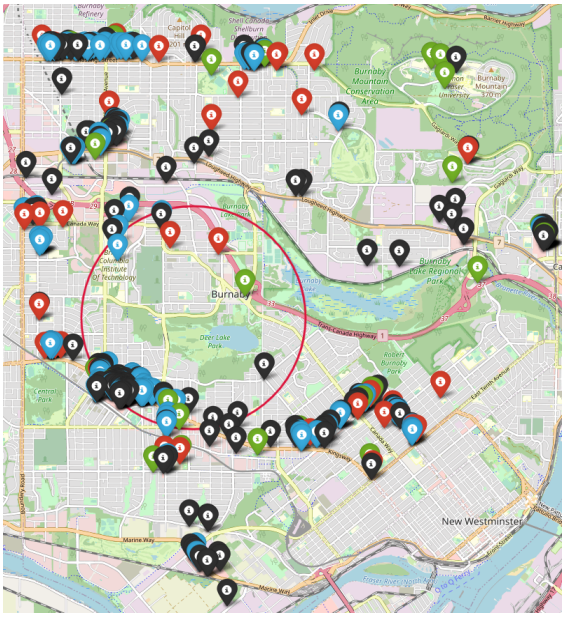
  What is the optimal location to purchase a house in Burnaby?

- The ideal living location should be in proximity to shopping malls, healthcare facilities, entertainment venues, and educational institutions. :
  - Education: Properties located in school districts are often close to high-quality schools with superior teaching standards, well-qualified teachers, and abundant educational resources. Due to the scarcity and popularity of these properties, their prices are usually higher. For some buyers, investing in a property within a school district may lead to potential appreciation in the future. It Contains university, school, library, kindergarten…etc.
  - Entertainment: Entertainment venues serve as gathering places for social activities, allowing people to meet new friends, participate in community events, and foster neighborhood interactions. It contains such as park, garden, social_center, cinema…etc
  - Healthcare: Living near medical facilities allows for quicker access to medical assistance during emergencies. This is crucial for addressing sudden health issues for oneself or family members, ensuring timely and professional medical care. It contains such as clinic, hospital, pharmacy…etc
  - ShopMall: Living near shopping malls allows for easy access to a wide range of goods and services required for daily life. The convenience of shopping saves time and effort, which is especially valuable for busy families or individuals with tight work schedules. It contains such as wholesale, supermarket, seafood…etc

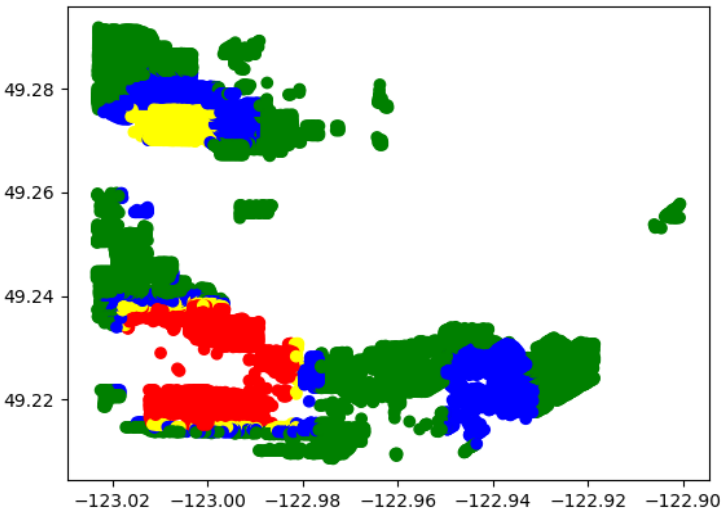
- *Data Format(transform_json.py & transform_building.py):*
  - The Orginal data is contains: ‘type’ , ‘id’ , ‘lat’ , ‘lon’ , ‘tags’.
  - I simply need to use Python to extract the data I need from the 'tags' column of the dataframe, and then create new elements with this data.

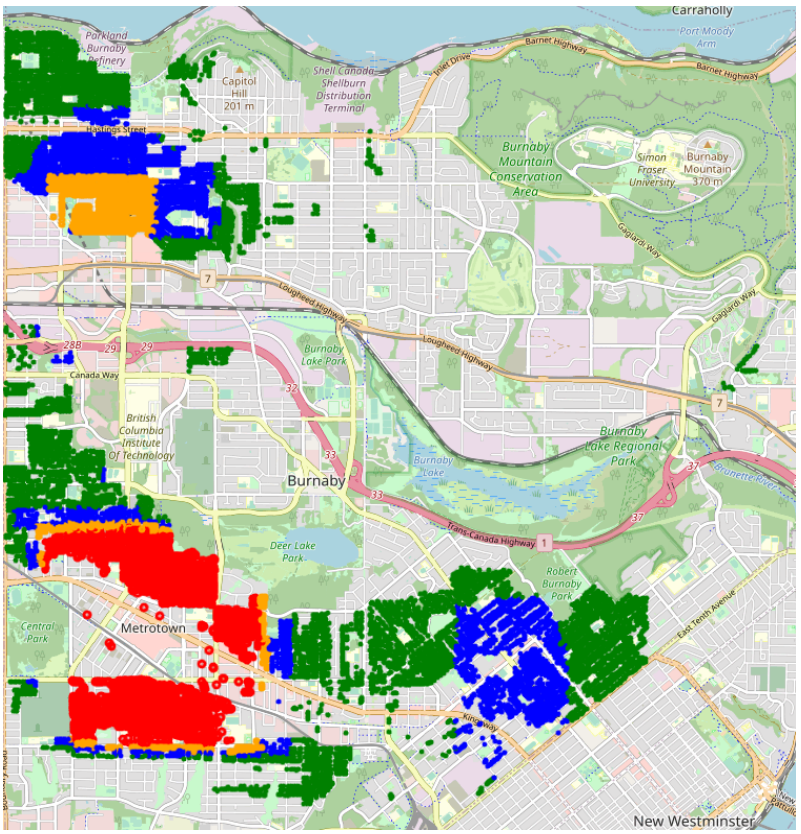
- *Data cleaning(buyer_data_cleanning.py):*
  - Remove useless data:
    - such as bar , fast_food, cafe, parking… etc that shouldn’t be our consider in problem.
  - Group the remaining amenities and classify them as above four groups:
    - Education: childcare , place_of_worship, kindergarten, language_school, library, toy_library, school, university are in amenity.
    - Healthcare: clinic, dentist, doctors, hospital, pharmacy are in amenity.
    - Entertainment: cinema, community_centre, conference_centre, events_venue, exhibition_centre, music_venue, theatre, police, social_centre, gym are in amenity. playground, garden, sports_centre, sports_hall, stadium, swimming_pool, park, track are in leisure.
    - ShopMall: department_store, wholesale, supermarket, bakery, convenience, dairy, farm, frozen_food, greengrocer, health_food, seafood, general, mall, baby_goods, clothes, shoes are in shop.
  - Remove the rows which name is Nan.

    - Remove outlier:
        - I used the fucation called remove_outlier, its consider by latitude and longitude of the points.
- *Data Written (buyer_data_cleanning.py):*
    - After data cleaning process, the cleaned data will be written into cleaned_data.csv and four separate data: education.csv, healthcare.csv, entertainment.csv, shopMall.csv
    - All processed data in the directory named ‘data’.
- *Plot the data on education.csv, healthcare.csv, entertainment.csv,shopMall.csv(show_data_points_on_map.py):*
    - Use python folium to present the map
        - The circle is how far from the center of each building
        - Black: shopMall_data.csv
        - Red: education_data.csv
        - Blue: healthcare_data.csv
        - Green: entertainment_data.csv
    - This will be used for comparison in subsequent analyses



- *Analysis and Conclusion (analysis_result_by_building.py):*
    - My primary concept is that each building is assigned a score and is surrounded by a circle with a radius of 2km. I then search for groups that fall within this circle. If a group is found within the circle, I add the corresponding score to that building. The buildings are then displayed with their respective scores, allowing me to identify the buildings with the highest scores.
    - I will assign 3 points to education and healthcare groups, as these are of significant importance. Entertainment groups will be given 2 points, and shopping malls will be assigned 1 point.
    - I will remove any buildings that have a score lower than 50 points.
    - I will use different colors to represent them on the map and plot. Red will represent scores higher than 200 points. Yellow or Orange will represent scores between 150 and 200 points. Blue will represent scores between 100 and 150 points. Green will represent scores between 50 and 100 points.
    - I will create a plot of the analyzed data.





- As we know, there are two major shopping malls in Burnaby, one in Metrotown and the other in Brentwood. In the graph, the red points represent scores higher than 200 points, which are typically near Metrotown. This is because Metrotown has a high concentration of shopping malls, educational institutions, healthcare facilities, and entertainment venues, which is why these properties score higher than others.
- Orange points, representing scores between 150 and 200, are shown in Brentwood because it hosts the second major shopping mall, which contributes to the higher scores. Additionally, blue points, representing scores between 100 and 150, appear to be concentrated around the Edmonds area.
- In conclusion, if you are considering purchasing a house, the areas marked in red on the map should be your primary focus. These areas have scored more than 200 points in our analysis, indicating a high concentration of essential amenities such as shopping malls, educational institutions, healthcare facilities, and entertainment venues. Choosing to reside in these areas could significantly simplify your daily routines and enhance your lifestyle by providing easy access to these facilities.
- In addition, the areas marked in orange and blue are also worth considering. The orange-marked areas, scoring between 150 and 200 points, are typically located around Brentwood, which hosts a major shopping mall contributing to the higher scores. The blue-marked areas, scoring between 100 and 150 points, are concentrated around the Edmonds area. While these areas might not have as high a concentration of amenities as the red-marked areas, they still offer a good mix of facilities that can cater to most of your daily needs.
- Overall, choosing a location with a high score according to our analysis can provide you with a more convenient living environment and a higher quality of life. It's not just about the house itself, but also about the community and facilities surrounding it. Therefore, when making a decision about where to buy a house, these factors should be taken into consideration to ensure that the chosen location aligns with your lifestyle and preferences.

Part 2: [Huayu Wang]

Introduction:

- The purpose of this report is to present a data-driven analysis focused on identifying optimal areas for long-term rental accommodation.

Problem Idea:

What is the optimal location for long-term renting in Burnaby?

A reasonable location for long-term rental should ideally be in close proximity to various key amenities, some of which are indispensable, while others are preferable but not mandatory:

- Food and Beverage: A suitable rental location should be near an assortment of eateries such as restaurants, cafes, pubs, and bars. Availability of fast-food chains, bakeries, delis, and ice-cream parlors also adds to the appeal of the neighborhood, providing a variety of food and beverage options.
- Market: For daily necessities and groceries, the vicinity of markets is essential. This includes supermarkets, grocery stores, butchers, greengrocers, convenience stores, and pharmacies.
- Service: Access to various services significantly enhances the convenience of a neighborhood for long-term rental. These services include fuel stations, banks, post offices, ATM locations, laundry services, car repair centers, healthcare facilities like clinics, hospitals, and dentists.
- Shops: Proximity to various shopping outlets such as bookstores, clothing and electronics stores, furniture stores, and pet stores adds to the appeal of a rental location, offering a wide range of products and services to the renters.
- Leisure: A neighborhood with recreational facilities such as parks, sports centers, fitness centers, playgrounds, and swimming pools can greatly enhance the living experience for long-term renters, offering options for relaxation, physical fitness, and entertainment.
- Entertainment: The presence of entertainment facilities such as libraries, arts centers, theaters, and cinemas, as well as shopping malls can make a location more desirable for long-term living.
- Education: For families in the academic field, proximity to educational institutions like universities, schools, kindergartens, and childcare centers is crucial.
- Public: Being near public buildings like community centers, town halls, and churches can provide a sense of community and belonging, contributing to the quality of long-term living.

Solving steps:

- A well structure pipeline is created based on the task, which is separate by following steps, in two files:

1. The first code is cleaning.ipynb. This file contains the steps of cleaning missing fields, categorizing and generating features, with input of total_data.json.gz, and output a csv file of categorized_buildings.csv.

   In the first part it reads the total_data.json.gz file, and then extracts the elements to make a Dataframe. The total_data.json.gz has three main types, nodes, way, and relation:

   a. **Nodes**, there are two types of nodes, one is pure nodes, which contains a node id and a group of latitude and longitude. Another type is building node, which is a node representing a building, which contains a tag storing all the useful information and a group of latitude and longitude.
   b. The **Way** type has a list of nodes and a tag, the list of node contains the node id, which can be read in the pure node part, and the tag contains the useful information.

c. **Relation** type is only used as polygon shape, which is not often used, there are only few data that are using relation type to represent and since it's hard to get its latitude and longitude, I clean them on purpose.

In this part, I also create two dataframe, one storing the building(building_elements_df), and another save all the pure nodes(pure_nodes_df).

In the second part, I extract the useful information stored in tag column from building_elements_df, which can be used for filtering and categorizing

The next part, I use the extracted data to generate a name, category and type column for the later filtering. I noticed most houses do not have a name, so I decided to use the house number combined with the street name to give them an identical name.

Since we now have the category and type, we can apply the filter to remove the unnecessary data to maximize the performance. In this step I only keep the useful types for long-term renting.

The 'way' type of building does not have their corresponding coordinates, instead they have a list of nodes that mark out a shape to represent this building. In this step, by reading the list of node id, and finding the corresponding coordinates in the pure_nodes_df, finally using the mean function to find a geometric center, and write it to the latitude and longitude column to the dataframe.

Now the data frame is complete, but for the next part of the train-model, it's easier to categorize the type. The categories are: food_beverage, market, service, shops, leisure, entertainment, education and public.

2. The second code is ML.ipynb, which contains generation, model training, and data visualization, below is a brief description of important part:
   - **Generate-features**: The function find_optimal_living_area takes in a residential type as input, reads in the processed data, and computes a new feature, 'Importance', based on the 'New_Category' column. It then filters the data for the specified residential type.
   - **Train-model**: The script also performs a model training operation using KMeans clustering to find centers for each category. It then uses NearestNeighbors to find the closest residential buildings to each center.
   - **generate-plots**: Lastly, it plots the results on a folium map, marking the centers for each category and the optimal living areas for the specified residential type. This visualization task aligns with the generated plots stage of the pipeline.

A significant aspect of the presented script is the user's ability to customize the residential type and the weights associated with each category of amenities.

For the residential type, users have the option to choose among different types such as 'house', 'residential', and 'apartments'. The selected residential type guides the algorithm to find the optimal living areas specific to that type. This feature increases the versatility of the script, allowing it to cater to different living preferences.

On the other hand, the weights assigned to each category reflect the user's personal preferences for different amenities. For instance, someone who places a high value on food and beverage amenities could assign a higher weight to the 'food_beverage' category. The algorithm takes these weights into account while identifying the optimal living areas, thereby providing results that are tailored to the user's preferences.

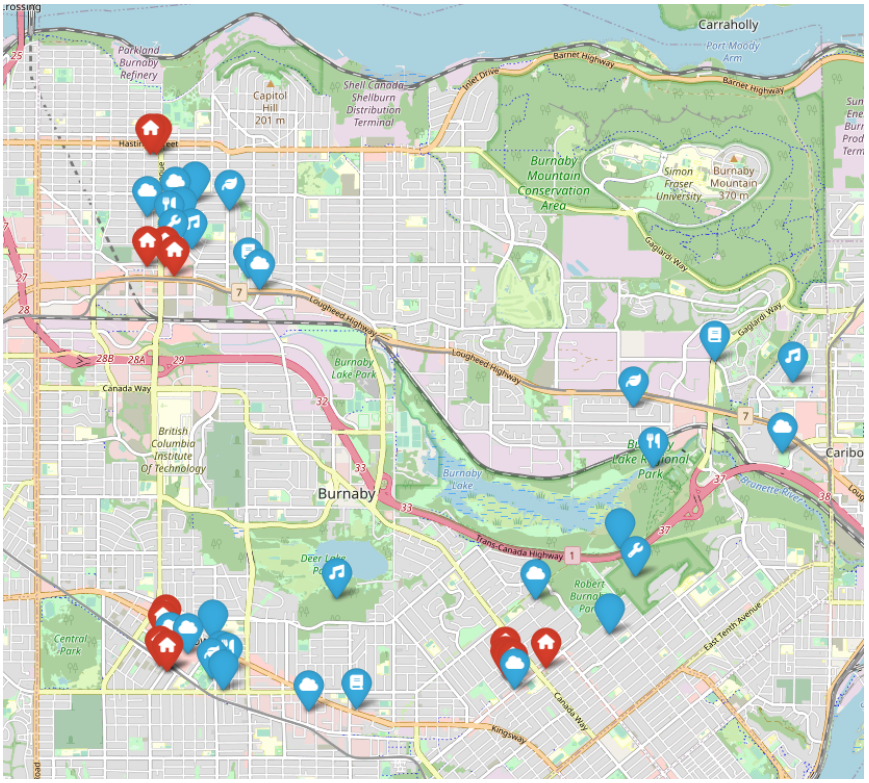
The ability to customize the residential type and importance weights ensures that the generated results are not only accurate but also relevant to the user's lifestyle and preferences. This customization brings a level of personalization to the tool, making it a practical and versatile solution for identifying optimal living areas.

Result & Graph:

In the task of identifying optimal living areas, a useful feature has been incorporated which uses unique icons to represent the centers of various categories on the map.In addition, the application offers the ability to assign importance weights to each category. A higher weight signifies greater importance of that category in the decision-making process.

By adjusting these weights, users can tailor the search results to align better with their lifestyle and priorities. In the analysis, the residential type can also be chosen using the function find_optimal_living_area(**residential_type**), where residential_type is a parameter representing the desired housing category.

The available housing types are determined based on the categories identified in the initial data and are stored in a dataframe for easy access and reference. These categories include types like 'house', 'apartment', 'residential'. Below is the graph of best long-term renting properties based on the selected weighted and house type:



As the graph shows, each facility category will generate few center points, and based on those center points and the weighted dictionary, a weighted center can be located. Then finding the nearest few houses based on the weighted center, which indicated the best long-term renting places based on the surrounding facilities.

Summary:

Throughout the project, I was able to gain hands-on experience with geospatial data analysis, web scraping, data cleaning, feature engineering, machine learning, and data visualization. This not only solidified my understanding of these areas but also allowed me to observe how these diverse fields could come together to solve a practical, real-world problem.

Part 3: [Qiting Wang]

Problem:

Questions:

What is the optimal location for short term rental in Burnaby?

The purpose of this report is to analyze the data to find out the best places for short term accommodation in the burnaby area.

Process steps:

file: data_cleaning.py
command: python3 data_cleaning.py original_data.json

Step 1. Use a unified code to obtain data in overpass turbo and choose to download it as a compressed file in .json format.

Step 2. Filter the node_data and way_data in original_data by the type of amenity (because these information contain a lot of useless information) and only keep some node_data and way_data that contain amenities that are usually more popular.

Step 3. The project further filters the information filtered in step 2 for the short-term rental apartment. Divide the information of the apartment into two parts (one part is the node_data of the apartment, and the other part is the way_data of the apartment).

Step 4. Extract the first node_id from the way_data of the apartment as the address of the apartment and then merge with the node_data of the apartment to obtain the latitude and longitude information that the apartment did not have originally.

file: short_term_rental.py
command: python3 short_term_rental.py

Step 5. Import the distance formula between two points to calculate the distance between the candidate apartment and the relevant amenity within the selected range

Step 6. Perform weighting operations on different types of amenity within the scope to select more suitable apartments for different groups (this project only weighs a certain common situation) and output information about all ideal short-term rental apartments through the rank function.

Step 7. Visualization: first display the heat map distribution of all apartments, and then display the specific locations of the most ideal n apartments (unweighted, marked in blue) according to the rank list in step6, and then display them in step6 The rank list of the most ideal n apartment (weighted, marked in red) specific locations.

Results (shown as graphs):

Display the cleaned original data and sample apartment dataframe.

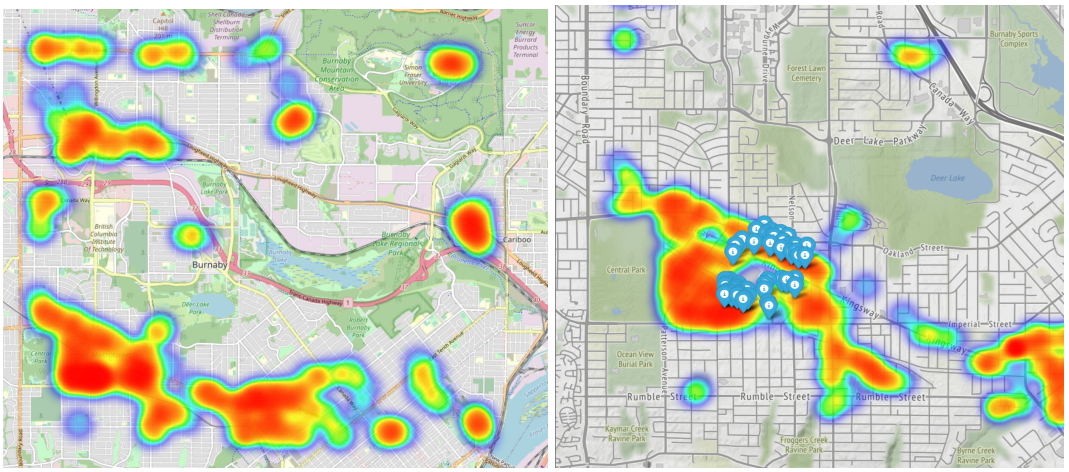
	id	lat	lon	tags.amenity	tags.leisure	tags.name
0	411900947	49.277918	-122.912361	restaurant	NaN	Pho 99
1	411900948	49.278119	-122.912285	cafe	NaN	Starbucks
2	482696846	49.279490	-122.967047	cafe	NaN	Starbucks
3	482697087	49.280070	-122.969768	restaurant	NaN	White Spot
4	482697092	49.280046	-122.966568	restaurant	NaN	Cockney Kings Fish & Chips
...	...	...	...	...	...	...
2040	9398267431	49.223620	-122.993450	parking	NaN	None
2041	1431318014	49.225209	-122.990745	parking	NaN	None
2042	9735573823	49.265836	-122.993446	parking	NaN	None
2043	9735573827	49.265773	-122.992958	parking	NaN	None
2044	9735573831	49.266199	-122.993140	parking	NaN	None

[2045 rows x 6 columns]

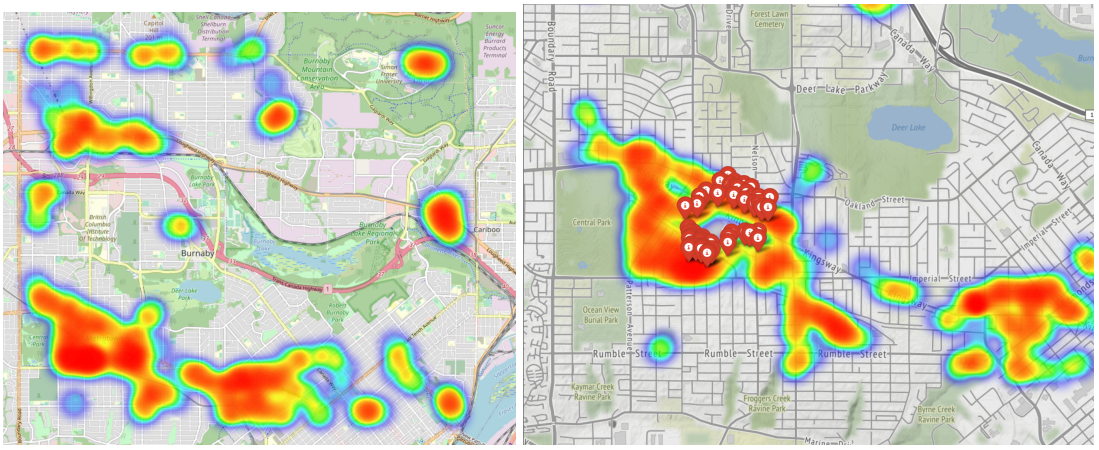
	tags.building	tags.name	lat	lon
0	apartments	Royal View Apartments	49.227774	-122.989755
1	apartments	The Hub	49.277906	-122.910665
2	apartments	Bonsor Avenue Place	49.224017	-122.998408
3	apartments	None	49.224959	-122.997862
4	apartments	The Bonsor	49.224521	-122.998262
..	...	...	...	...
769	apartments	Cirrus	49.265438	-123.005804
770	apartments	None	49.280026	-122.908527
771	apartments	None	49.272196	-123.008595
772	apartments	Holdom Place	49.282319	-122.981194
773	apartments	None	49.280836	-122.981689

[774 rows x 4 columns]

Display the specific locations of the most ideal n apartments (unweighted, marked in blue) according to the rank list in step6.



Display them in step6 The rank list of the most ideal n apartment (weighted, marked in red) specific locations.



Display the rank list for graph(unweighted and weighted)

	tags.building	tags.name	lat	lon	num_of_amenity
336	apartments	Spectrum	49.228542	-122.997507	0.757578
322	apartments	Arbour Place South	49.229177	-122.997039	0.748028
513	apartments	None	49.228096	-122.995987	0.748028
320	apartments	None	49.228723	-122.999832	0.748028
219	apartments	La Mirage Tower I	49.229528	-122.996053	0.744845
337	apartments	None	49.227764	-122.994508	0.744845
713	apartments	Hazel	49.229273	-122.997437	0.741662
714	apartments	Sussex	49.229465	-122.998050	0.741662
221	apartments	The Evergreen	49.228750	-122.995701	0.741662
323	apartments	Maple Glade	49.228302	-122.993941	0.738479
357	apartments	None	49.224502	-123.002073	0.738479
326	apartments	Horizon Towers	49.227748	-122.993862	0.735296
3	apartments	None	49.224959	-122.997862	0.735296
4	apartments	The Bonsor	49.224521	-122.998262	0.735296
321	apartments	Arbour Place North	49.229734	-122.997468	0.735296
222	apartments	None	49.228355	-122.994816	0.735296
353	apartments	None	49.223995	-123.000895	0.732113
2	apartments	Bonsor Avenue Place	49.224017	-122.998408	0.732113
220	apartments	La Mirage Tower II	49.229273	-122.995044	0.728930
358	apartments	None	49.224277	-123.002346	0.728930

	tags.building	tags.name	lat	lon	num_of_amenity
336	apartments	Spectrum	49.228542	-122.997507	1.330535
219	apartments	La Mirage Tower I	49.229528	-122.996053	1.314620
322	apartments	Arbour Place South	49.229177	-122.997039	1.311437
320	apartments	None	49.228723	-122.999832	1.311437
513	apartments	None	49.228096	-122.995987	1.305071
714	apartments	Sussex	49.229465	-122.998050	1.298704
221	apartments	The Evergreen	49.228750	-122.995701	1.298704
713	apartments	Hazel	49.229273	-122.997437	1.298704
337	apartments	None	49.227764	-122.994508	1.295521
321	apartments	Arbour Place North	49.229734	-122.997468	1.292338
357	apartments	None	49.224502	-123.002073	1.285972
353	apartments	None	49.223995	-123.000895	1.279606
323	apartments	Maple Glade	49.228302	-122.993941	1.276423
222	apartments	None	49.228355	-122.994816	1.273240
592	apartments	None	49.230387	-122.998373	1.266873
358	apartments	None	49.224277	-123.002346	1.266873
4	apartments	The Bonsor	49.224521	-122.998262	1.263690
3	apartments	None	49.224959	-122.997862	1.263690
218	apartments	The Madison	49.229898	-122.999456	1.263690
220	apartments	La Mirage Tower II	49.229273	-122.995044	1.260507