OSM, Photos, and Tours

Samantha Tse
Raymond Chan

Introduction

Exploring a new city is an adventure filled with endless possibilities waiting to be discovered around every corner. However, sometimes it is easy to miss certain points of interest if you do not plan ahead. Planning a tour of a city means finding exciting places to visit, such as landmarks, entertainment, and food. It's also helpful to know where chain restaurants are compared to local ones, to see the city's food scene. Furthermore, when choosing a place to stay, it's important to pick somewhere close to fun things to do. This project will aim to enhance the city exploration experience by offering information on must-see attractions, navigating diverse culinary landscapes, and selecting convenient accommodations near key amenities and attractions. Specifically, the project aims to solve the following questions.

1. If I was planning a tour of the city (by walking/biking/driving), where should I go? Are there paths that take me past an interesting variety of things?
2. I feel like there are some parts of the city with more chain restaurants (e.g. McDonand's or White Spot franchises, not independently-owned places): is that true? Is there some way to find the chain places automatically and visualize their density relative to non-chains?
3. If I was going to choose an Airbnb, where should it be? What places have good amenities nearby?

Gathering Data

We used five datasets for this project:

1. amenities-vancouver.json.gz provided from the project
    - A compressed JSON file containing amenities within the Greater Vancouver area, extracted from the larger OpenStreetMap (OSM) data.
2. listings.csv
    - A CSV file obtained from https://insideairbnb.com/get-the-data/
    - Shows listing information from all Vancouver Airbnbs during December 13, 2023.
3. Geotagged photos
    - Photos taken around Vancouver to represent a walk that contain EXIF tags for latitude and longitude information.
4. all_restaurants.osm
    - Contains recent OSM data from the Overpass Turbo for nodes containing amenity=restaurant, amenity=fast_food, amenity=cafe https://overpass-turbo.eu/s/1JS1
5. census_tract.shp
    - Contains census tract boundaries for Canada, obtained from Statistics Canada https://open.canada.ca/data/en/dataset/b5a4adbc-5c56-4acd-b3f9-9a177c047a0e/resource/9d80eae5-2f37-4363-af25-8febe1ff26f0

Cleaning the data

Tour of the City

The provided data contains too many amenity types and we only want to know amenities that are interesting. To do this, we first read the JSON file into a Pandas DataFrame. Then, we create a dictionary called categories to map amenity types to more interesting categories. For instance, ‘cafe’ and ‘restaurant’ would belong to the food category. We then add a new column 'category' to the DataFrame and assign the corresponding category to each amenity based on its type. After categorization, data is grouped by category and saved into separate compressed JSON files. This makes it easier to analyze and access the data later on.

For cleaning the Airbnb data, we used PySpark on listings.csv. First, we read the file into a DataFrame and removed any duplicate rows by identifying them with a unique identifier called 'id'. Next, we decided to drop the column 'host_name' as it wasn't necessary for our analysis. We made sure to get rid of any rows that had missing values in key columns such as 'name', 'price', 'number_of_reviews', and 'reviews_per_month'. Additionally, we filtered out rows with less than 10 reviews or less than 0.5 reviews per month, ensuring we work with listings that are actively reviewed and maintained by guests, indicating a level of popularity. Furthermore, we removed rows where the minimum nights required exceeded 30 or where the property availability was less than 30 days in a year to ensure that we focus on listings that are suitable for short-term stays and are actively used. Finally, after all the cleaning steps, we saved the cleaned data into a new CSV file named 'cleaned_listings'.

Distribution of Chain Restaurants

To use the OSM file exported from Overpass Turbo containing all restaurants, we processed the XML into compressed JSON format, extracting all tags as separate fields so that it would be easier to work with later.

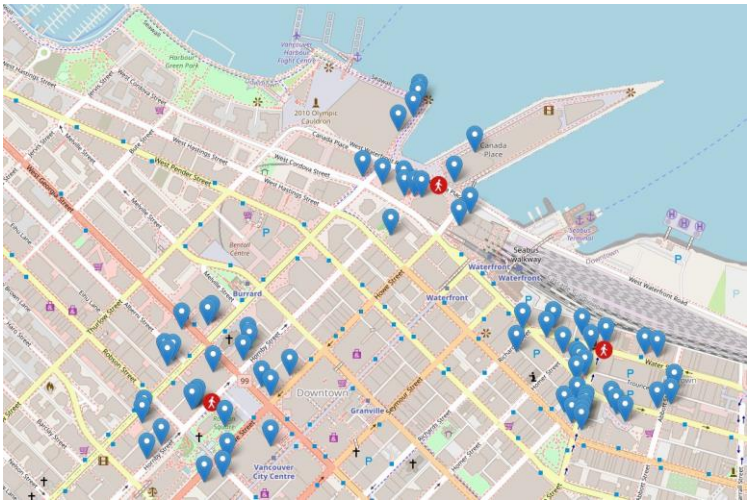
The census tract boundaries downloaded from Statistics Canada contain all census tracts within Canada. To keep only those within the Vancouver census metropolitan area (CMA), we first loaded the shapefile into a GeoPandas DataFrame, which is great for handling geometric data. Then, we retained only the rows associated with Vancouver, keeping the columns for CTUID (unique ID of the census tract) and geometry (boundaries of the census tract). Finally, we saved the resulting DataFrame to a GeoJSON file named ‘vancouver_census_tract.geojson’.

To align the restaurant data with the census tract data, we utilized GeoPandas GeoDataFrames. This enabled us to perform a spatial join between the two datasets, matching them based on whether the restaurant point (latitude, longitude) falls within the census tract polygon.

Techniques used to analyze the data

Tour of the City

Users are able to take photos of their walk and put them into a folder. The program will then take the folder of geotagged photos and identify the location of each photo. By using the Haversine formula, the program calculates distances between each specified amenity and each photo's location, filtering out establishments falling within a specified search radius based on if the user is walking, biking, or commuting. Folium is used to generate HTML maps pinpointing both the image location (red walk icon) and nearby amenities (blue marker) along with their distance. This helps us check what we might have missed during our walk, and find nearby amenities so we can plan our path better. The map below shows food amenities within 200 meters of each image location.



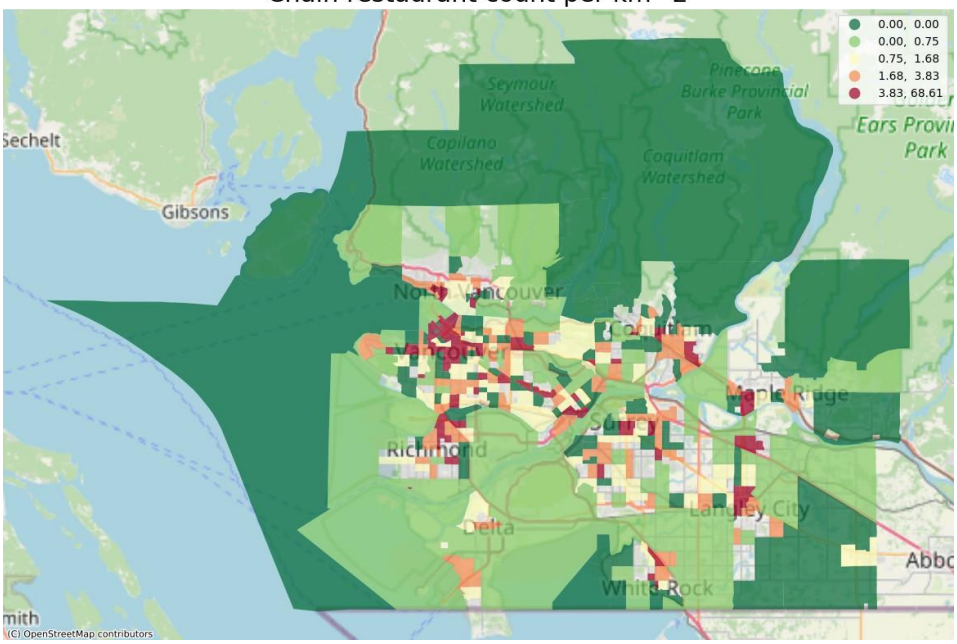
A similar approach is used to find Airbnb's with the most amenities nearby. Marker clustering is used to enhance map readability when numerous amenities are clustered closely together. Below is the Airbnb that has the most food amenities within a 50-meter radius.

Distribution of Chain Restaurants

In our analysis of chain restaurants, when we mention restaurants, we are specifically referring to locations in OSM data where the amenity value is either 'restaurant', 'fast_food', or 'cafe'. Within OSM, there is a 'brand=*' tag utilized to denote the primary brand of goods, often trademarks. For restaurants, this typically indicates chain or franchise names. Therefore, we classified restaurants based on the presence or absence of the 'brand' tag.

To begin, we used a choropleth map to explore the spatial patterns in our data. We mapped the density of chain restaurants (restaurants/km^2) within each census tract. This approach allows us to discern if certain areas within Greater Vancouver have a higher concentration of chain restaurants.

Chain restaurant count per km^2



Through visual analysis alone, it is clear that spatial clusters exist, indicating certain areas within the city have a higher density of chain restaurants. However, it is known that clusters can occur randomly; so we are interested in investigating whether the observed pattern of clustered chain restaurants is more than just random chance. Therefore, in addition to relying solely on visual analysis, we sought additional statistical evidence to support our visual observations.

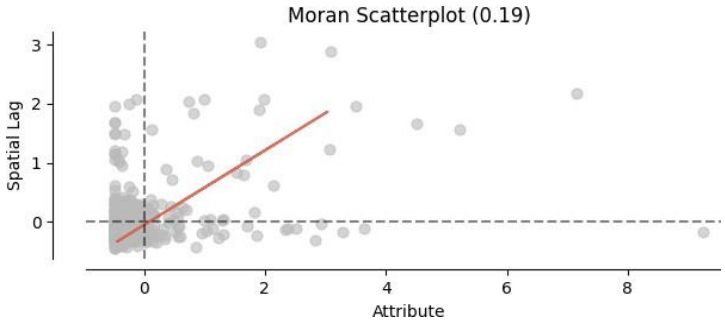
Chi-Square Test

The null hypothesis for a Chi-Square test states that the categories are independent, meaning that the proportions of chain and non-chain restaurants will be the same across various parts of the city. After conducting a Chi-Square test, we got a p-value of 2.085e-50. Since our p-value is less than the significance level of 0.05, we reject the null hypothesis. This leads us to conclude that different parts of the city have some impact on the prevalence of chain restaurants within them. This suggests that the distribution of chain restaurants is not uniform across the city, indicating that some areas may have a higher concentration of chain restaurants compared to others.

Global Spatial Autocorrelation

The Global Moran's I evaluates whether the data exhibits clustering or dispersion. The null hypothesis states that the data is randomly distributed. This test investigates the correlation between the distances separating features and the differences in their values. The Moran's I value indicates whether we have positive or negative autocorrelation. A positive value suggests clustering, while a negative value indicates a dispersed pattern.

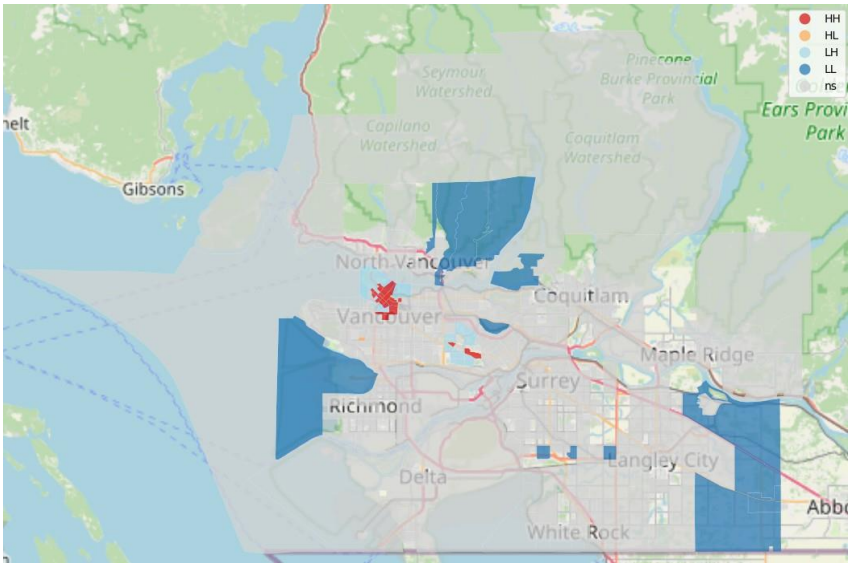
In our data, each census tract polygon will contain a feature representing the number of chain restaurants within the region. However, to enable more meaningful comparisons between different census tracts, we will normalize this count by the area of each census tract. After conducting the Global Moran's I test, we obtained a Moran's I value of approximately 0.187, with a corresponding p-value of 0.001. Since the p-value is less than 0.05 and the Moran's I value is positive, we can reject the null hypothesis. This indicates that if one area has a high density of chain restaurants relative to its size, neighbouring areas are likely to also have similarly high densities, forming clusters of high chain restaurant density. Conversely, areas with low chain restaurant density relative to their size are likely to be clustered together as well.



The scatterplot reveals a positive slope in the regression line, suggesting a correlation between the feature value and the values of their neighbours.

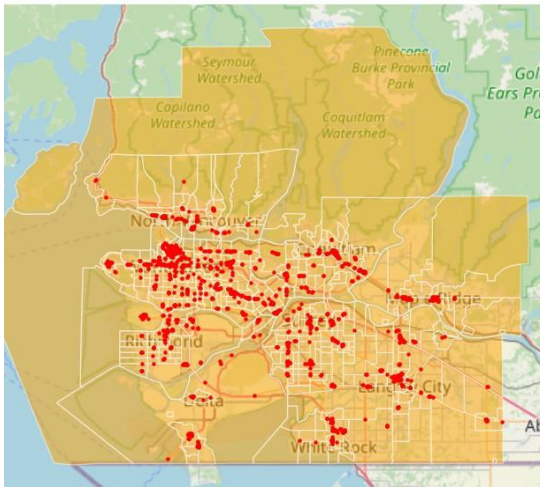
Hot spot analysis

While the Global Moran's I indicates whether the data is clustered, it does not tell us where these clusters occur. Therefore we wanted to further our analysis by performing a hotspot analysis, to identify the specific locations where these clusters occur. In the map below, the coloured sections represent census tracts that have a p-value less than 0.05, meaning that these are statistically significant and spatially autocorrelated. The red sections indicate hotspots, meaning that there are high values near other high values, and the dark blue sections indicate coldspots, where low values are near low values.

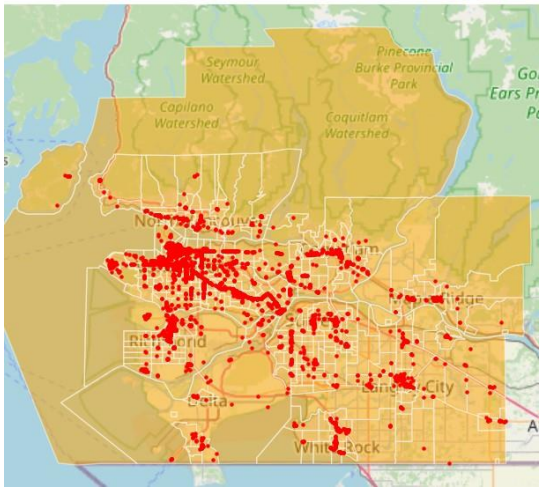


Finally, in terms of the density of chain restaurants versus non-chain restaurants, it is apparent that there are more independently owned restaurants. However, both types of restaurants appear to cluster in similar areas, such as Downtown Vancouver and popular streets like Commercial Drive.

Chain



Non-Chain



Conclusions

The program processes geotagged images to extract GPS coordinates and identifies nearby amenities of a specified type within a given radius depending on if you are walking/biking/driving. This enables the identification of points of interest along your tour route. The generated folium map provides a clear overview of where the amenities are situated in relation to your tour path so you can plan an optimized tour route that takes you past a variety of interesting things.

Based on our analysis of chain restaurants in Greater Vancouver, we have found that Downtown Vancouver has a higher concentration of chain restaurants compared to other areas within the metropolitan region.

Limitations and Future Improvements

- The JSON file does not have many exciting tourist spots. We could add more by looking for extra data on interesting places, even ones farther from downtown Vancouver.
- The program could have generated an ideal path on the map based on the most interesting amenities near each image.
- The tags in the OSM data could be missing the `brand` tag even though the restaurant is indeed part of a chain. In the future, we could look for better ways to classify whether the restaurant is part of a chain.
- For restaurant density, we normalized the data using census tract areas, however, it may be beneficial to use the population within those regions instead

Project Experience Summary

Raymond Chan

- Organized the amenities JSON file and utilized Spark to clean the Airbnb listings data.
- Extracted GPS coordinates from geotagged images by using GPSphoto library.
- Utilized the folium library to generate a map that visualizes the filtered amenities with markers, providing a clear overview of nearby points of interest.

Samantha Tse

- Converted XML data into a user-friendly JSON format, making it easy to analyze with Pandas.
- Generated graphs and maps, leveraging the Contextily library to incorporate detailed base maps, which enhanced spatial visualization and analysis.
- Conducted research into spatial analysis methodologies and performed spatial autocorrelation tests to investigate spatial patterns and relationships within the dataset