COMPUTATIONAL DATA SCIENCE

Report

Arsalan Macknojia
August 1st, 2020

Introduction

In this project, we utilized different data science tools, algorithms, and machine language standards to find concealed information from crude data to address two different problems. The analysis was performed on the data set acquired from OpenStreetMap and Airbnb listing for Vancouver. OpenStreetMap is a free editable world map that allows free access to images and all the underlying map details. Airbnb data was acquired from Inside Airbnb which provides publicly available information from the Airbnb website.

Problems

- Shortlist ideal Airbnb listings with good amenities nearby.
- Are there places in Greater Vancouver with more chain restaurants? Is there some way to find the chain places automatically and visualize their density relative to non-chains?

Shortlist ideal Airbnb listings with good amenities nearby

Data Gathering

Airbnb listing information was acquired from Inside Airbnb. The data set had over 5,500 listings for Vancouver, Canada. Some of the important attributes included name, hostname, host identity verified, latitude, longitude, neighborhood, price, number of reviews, and ratings.

The amenities data was acquired from OpenStreetMap. The data set had over 8,000 entries about different amenities types such as restaurants, banks, clinics, etc. Some of the important attributes included amenity type, latitude, and longitude.

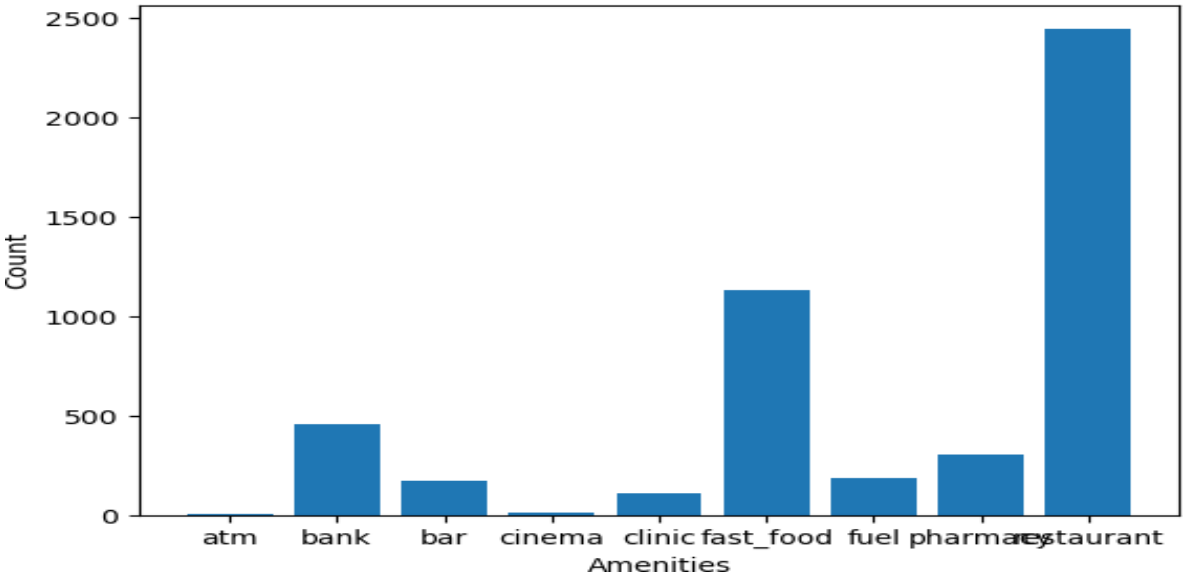
Data Cleaning

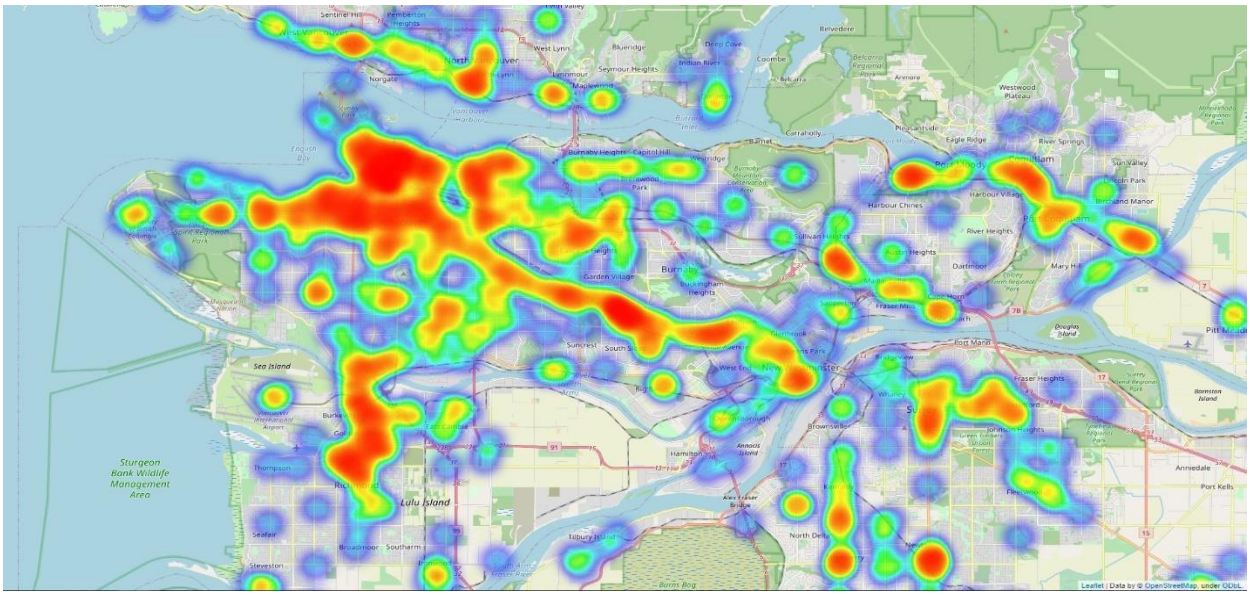
Airbnb data was filtered based on various conditions. Initially, all the entries with less than 10 reviews or an average overall rating of less than 75 were dropped. Secondly, the listings with the unverified host were also dropped. Lastly, the data was filtered on the minimum and maximum pricing per night. These values were taken from the user during runtime as an optional input with the default set to $0 and $100000. After filtering, the original data set of 5,806 was reduced to 1,233 entries.

Amenities data set was filtered based on types. The entries with restaurants, fast food, fuel, atm, bank, pharmacy, clinic cinema, and bar as amenity types were kept. The original data set of 8,169 was reduced to 4,832 entries.

Data Analysis & Results

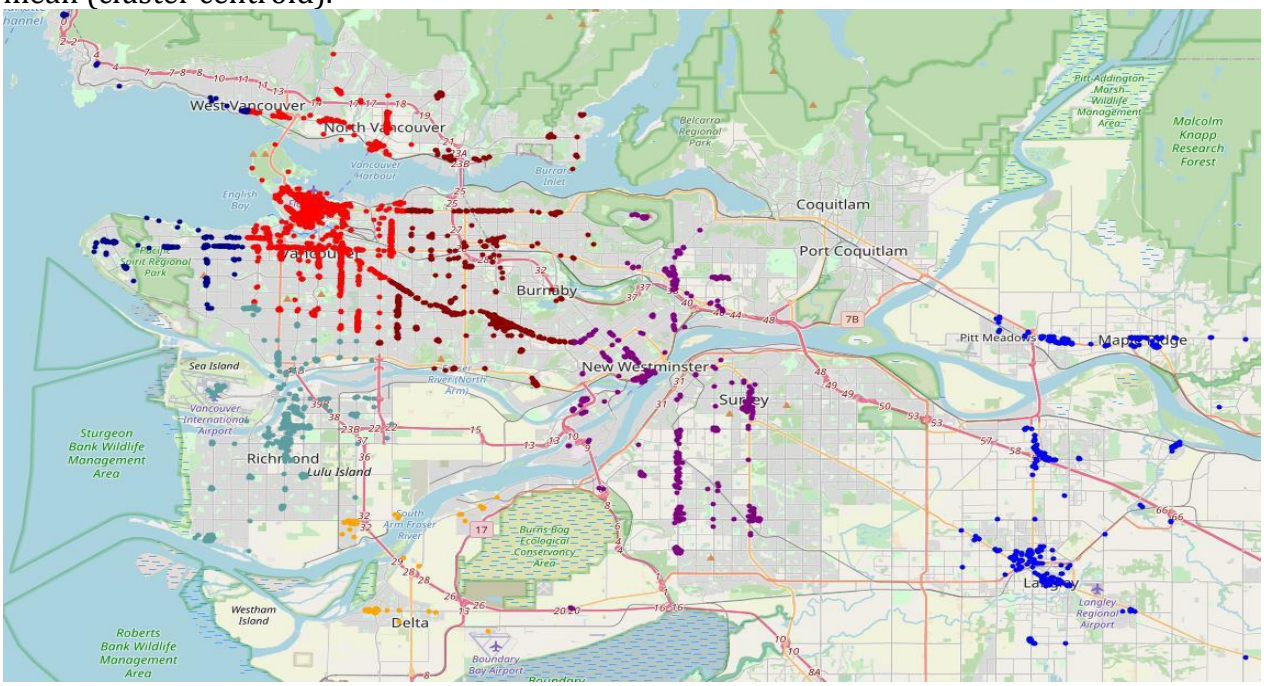
After cleaning the data, it was time to perform some data analysis. We begin by plotting the bar chart and heat map of amenities to visualize amenity types and their distribution on the map. The idea was to foresee any evident pattern before applying further statistics.



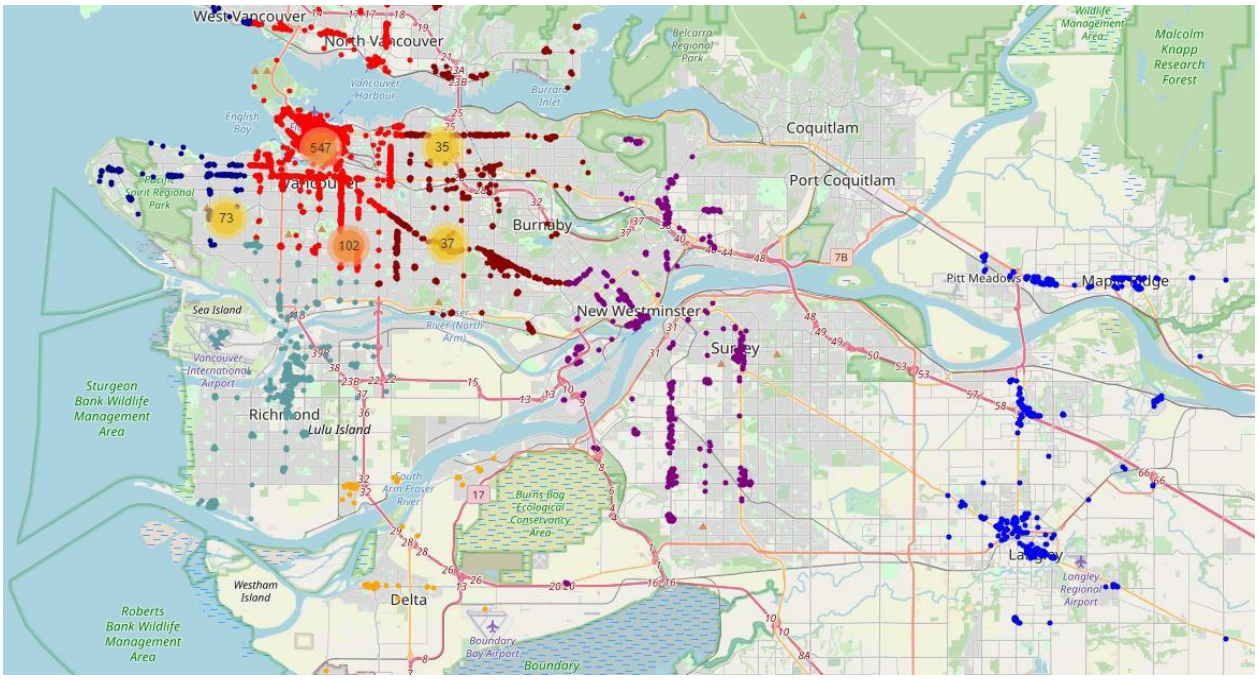


Looking at the bar chart, it is evident that the data set had more restaurant locations than any other amenity. Secondly, Downtown Vancouver has the greatest number of amenities compared to other places. Therefore, ideal listings must include majority of places in the downtown area. This was a good start but a vague visualization of ideal locations that demanded a more robust approach.

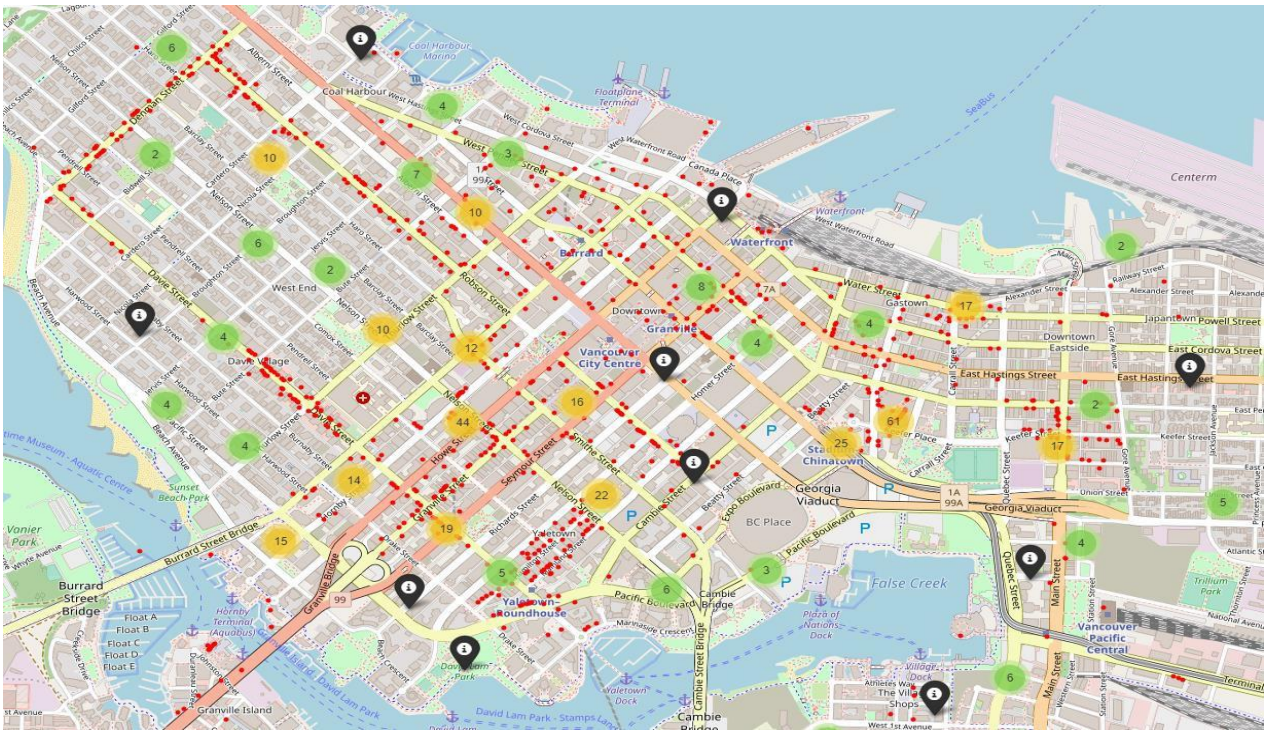
To perform an in-depth analysis, we utilized k-means clustering method to group nearby amenities into n clusters where each amenity was assigned to the cluster with the nearest mean (cluster centroid).



After splitting amenities into clusters, we used the center point these clusters as the pivot and selected Airbnb listings in the radius of 5 KMs from these clusters centroids. The listings were now significantly reduced from the original 5,500 locations to around 700, where all listings were now near amenities.



Zooming in on the map shows the exact location of the Airbnb listing along with nearby amenities.



Conclusion

To conclude, the initial prediction, after witnessing amenities heat map, of having the ideal listings near Downtown Vancouver align well with the final result. Looking at the map, it is evident that the model selected majority of listings in Downtown Vancouver which makes it an ideal location to rent places.

Limitations

The most obvious limitation was the small data set. We only had Airbnb listings for Vancouver area which prevented us from doing an in-depth analysis based on cities, amenities, and price. The overall result would have been significantly better if we had Greater Vancouver Airbnb listings as well. Secondly, the lack of user interface meant limited user interaction which prevented customizable data filtering.

Are there places in Greater Vancouver with more chain restaurants? Is there some way to find the chain places automatically and visualize their density relative to non-chains?

Data Gathering

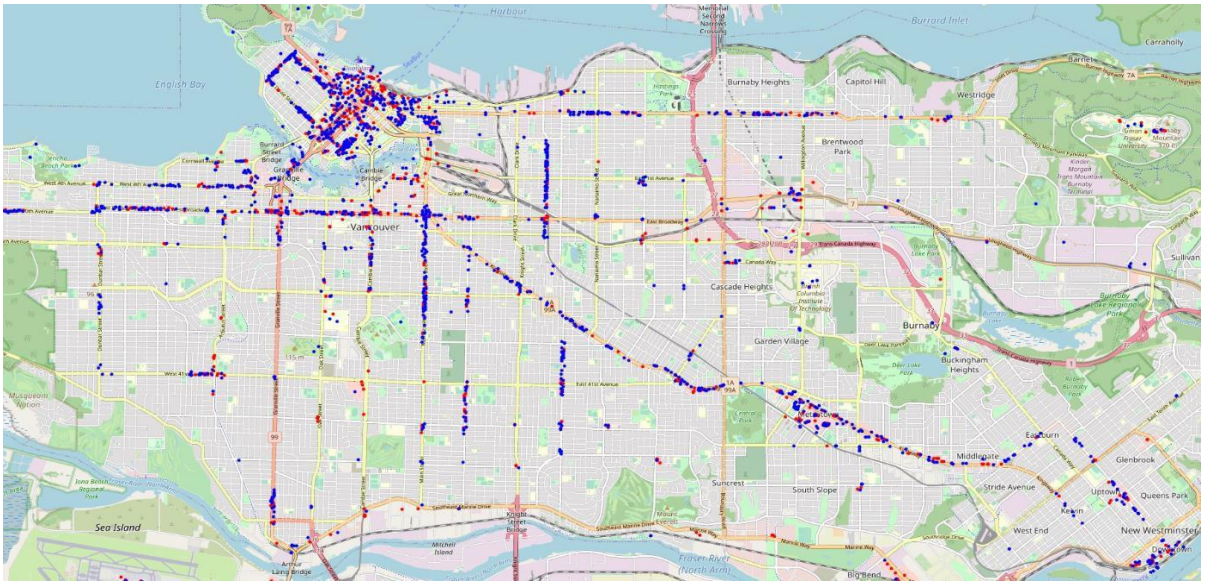
The data set to analyze the density of chain/non-chain restaurants in Greater Vancouver was acquired from OpenStreetMap. The data set had over 8,000 entries of different amenities such as restaurants, banks, clinics, etc. Some of the important attributes included amenity type, location name, latitude, and longitude.

Data Cleaning

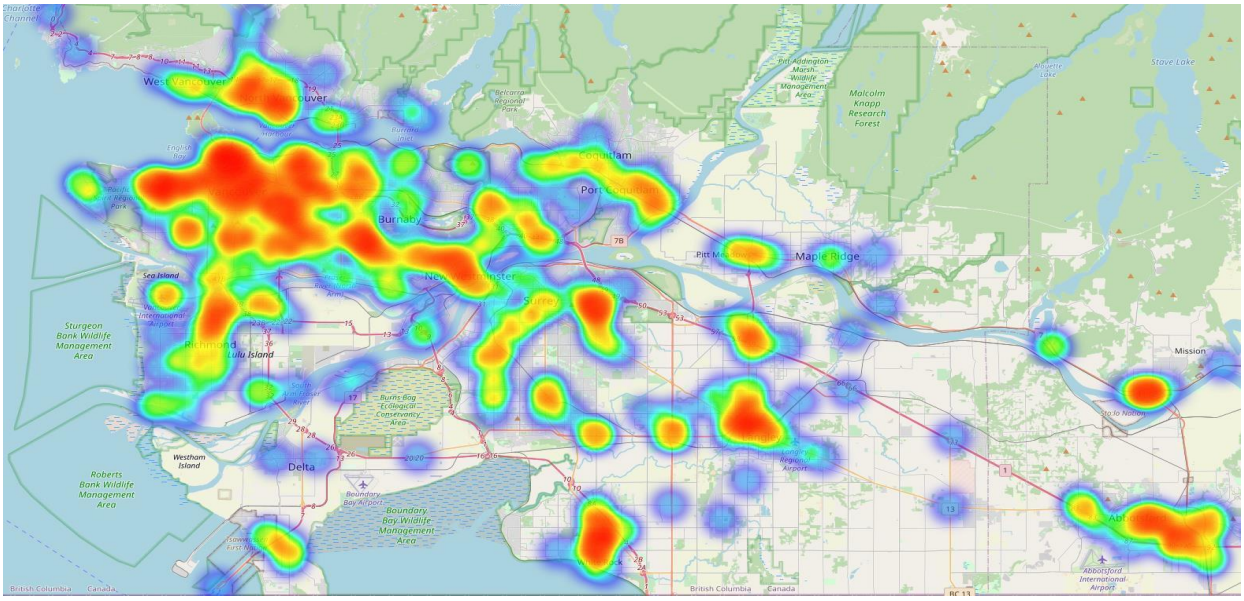
The data set was filtered based on amenity type. The entries with restaurant, café, and fast food as amenity types were selected. The original data set of 8,169 was reduced to 4,630 entries. Next, we wanted to split the data into two sets, namely chain restaurants, and non-chain restaurants. By definition, a chain restaurant has more than one location operating under the same brand name and selling similar items. The data set had limited information on the items which were being sold, therefore, we simply grouped restaurants together based on names. The groups with more than one entry were considered as chain restaurants.

Data Analysis & Results

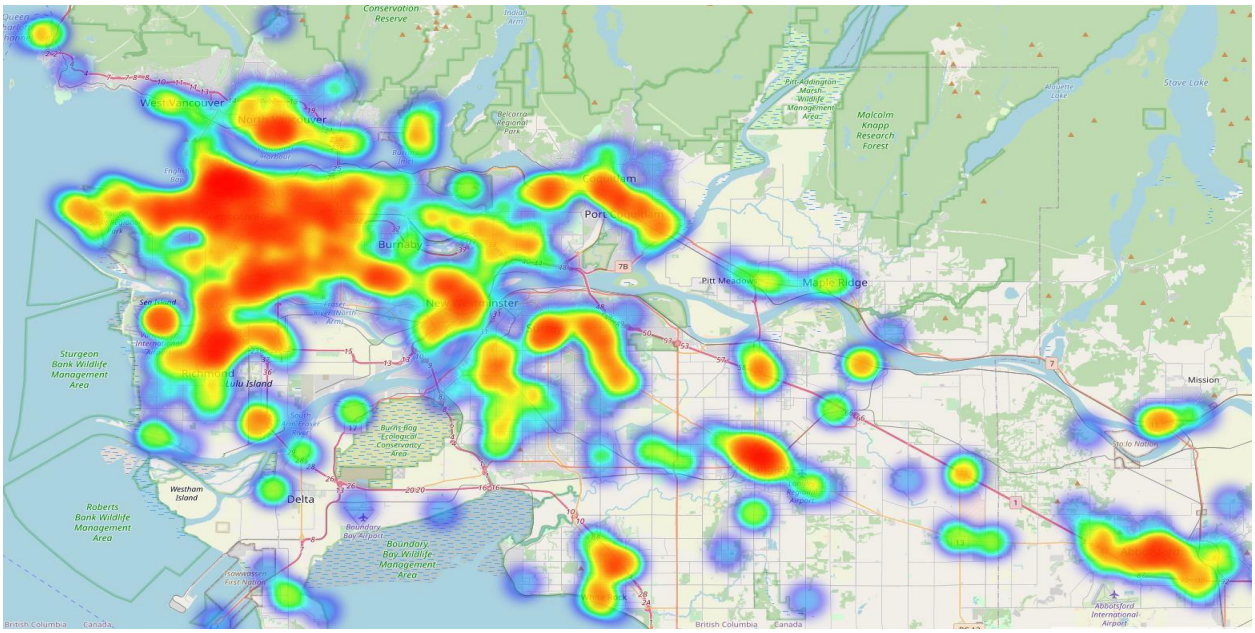
After cleaning the data, it was time to perform data analysis. We begin by marking chain restaurants as red and non-chain restaurants as blue to visualize their distribution on the map. The idea was to foresee any evident pattern before applying further statistics.



The initial visualization suggested that both chain (red) and non-chain (blue) restaurants had similar distribution with the majority of the locations concentrated in the Downtown Vancouver area. It also appeared that there are more non-chain restaurants (blue) in the Downtown area compared to chain restaurants (red). To evaluate further we generated the heat map of both the restaurant types.



Chain Restaurants

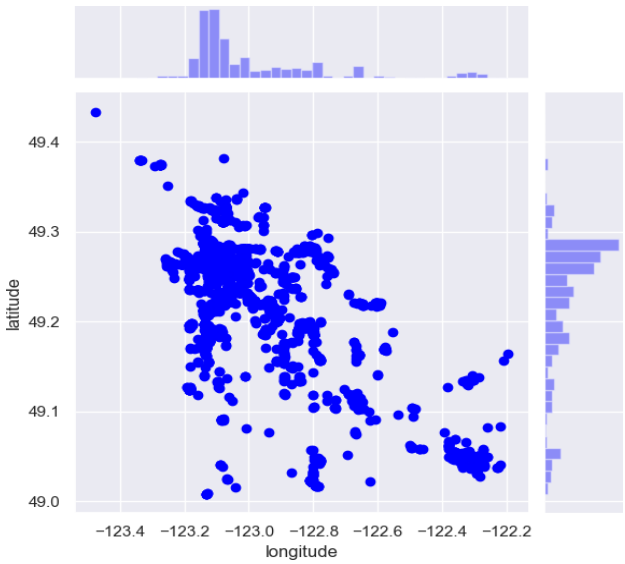


Non-Chain Restaurants

The heat map indicated that non-chain restaurants are more densely packed especially in the Downtown Vancouver area compared to non-chain restaurants which are relatively scattered. It also appeared that there are more non-chain restaurants in Vancouver, Burnaby, and Abbotsford than chain restaurants. These visualizations suggested some evident patterns, but we needed statistical proves to validate it.

We calculated the mean and standard deviation of both chain and non-chain restaurants. Chain restaurant had a mean latitude 49.21177 (SD = 0.07942) and longitude -122.94328 (SD = 0.24112). Non-chain restaurants had a mean latitude 49.22205 (SD = 0.07623) and longitude -122.98458 (SD = 0.22405). The mean latitude and longitude are very similar, but chain restaurants have a higher standard deviation which suggests that it is more scattered compared to non-chain restaurants. Following density comparison also validates the above findings.

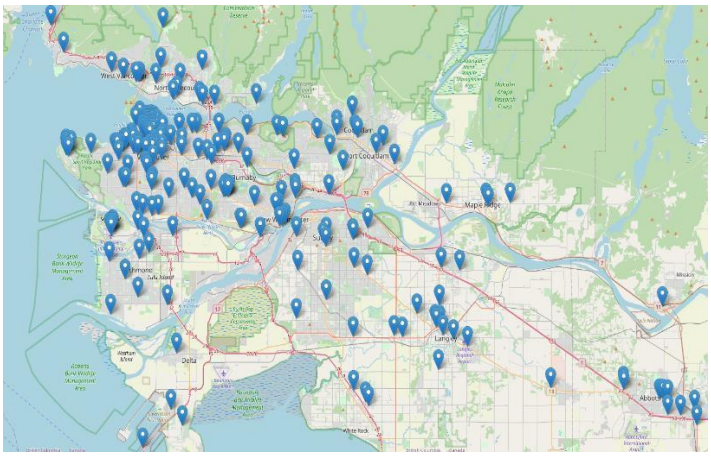
Chain Restaurants



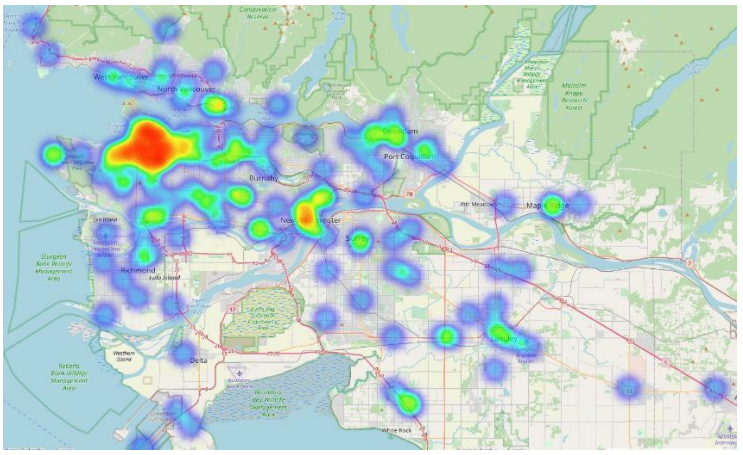
Non-Chain Restaurants

Additional Feature

We added an additional feature to allow users to search the locations of their favorite restaurants within Greater Vancouver. The user has an option to enter the restaurant's name when running the program which then pins the restaurant's locations and generates a heat map to show its distribution. The following images display all Starbucks locations in Greater Vancouver.



Starbucks locations



Starbucks heat map

Conclusion

To conclude, the visual and statistical analysis suggests that both types of restaurants have a similar distribution pattern, however, chain restaurants are relatively more scattered than non-chain restaurants. Secondly, the majority of the restaurants are concentrated in Vancouver followed by Burnaby and Abbotsford.

Limitations

The data set was quite small with only 4,630 restaurant entries. In addition, the available information about each restaurant was also limited. Restaurants were split into the chain and non-chain restaurants based on the assumption that chain restaurants will have more than one location. This might not be necessarily true as there can be chain restaurants with only one location in the Greater Vancouver area.