OSM, Photos, and Tours

CMPT 353 Project

SFU - Fall 2022

Professor: Greg Baker

David Ligocki - 301391784
Dmitrii Beliaev - 301435234
Minjae Shin - 301383936

Introduction

In this project we look at data from OpenStreetMap in the Vancouver region. We wanted to visualize the relative densities of chain and non-chain restaurants to analyze how their density changes throughout different parts of the city.

Problem

We are trying to solve the problem of figuring out which parts of Vancouver contain relatively higher amounts of chain restaurants compared to non-chain restaurants. This data can give value to residents or visitors of Vancouver who are looking to find restaurants of either type. We felt this was a good issue to cover because we can make good use of the knowledge gained in this course to clean the data, analyze with machine learning tools, and present some features about the restaurant locations with plots.

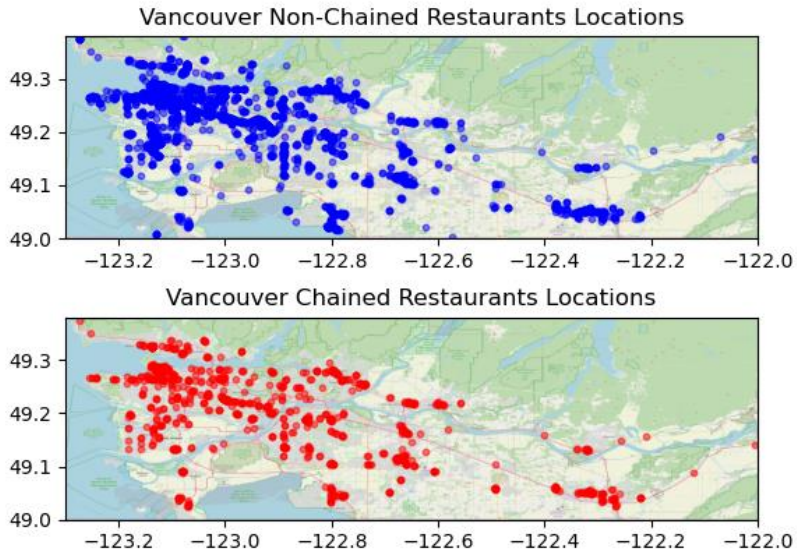
Data Gathering

We started with data provided by the professor on all of the amenities in Vancouver that were contained in OpenStreetMap. We filtered through the data to keep only restaurants and fast food, which means we also removed cafes as we felt like they generally don't offer the sort of food we are searching for. For simplicity, we may just say "restaurants" in this paper when we are referring to both restaurants and fast food.

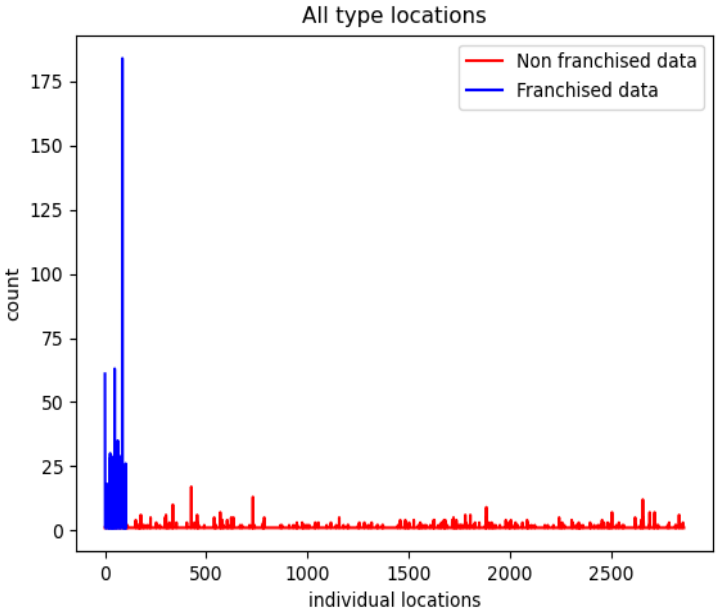
We identified many chain restaurants by scraping the internet for lists of restaurant franchises, but some chains were missed due to not being included in the scraped data or having their name formatted differently. To resolve this issue we searched the list of restaurants again and decided that all restaurants with a "brand" in their tags should also be included as a chain.

Analysis Methods

We started by superimposing the locations of restaurants over the map of Vancouver with a scatter plot. Scatterplot focused on both (Non-) Chained Restaurants & Fast-food locations all around the Vancouver (BC) area.

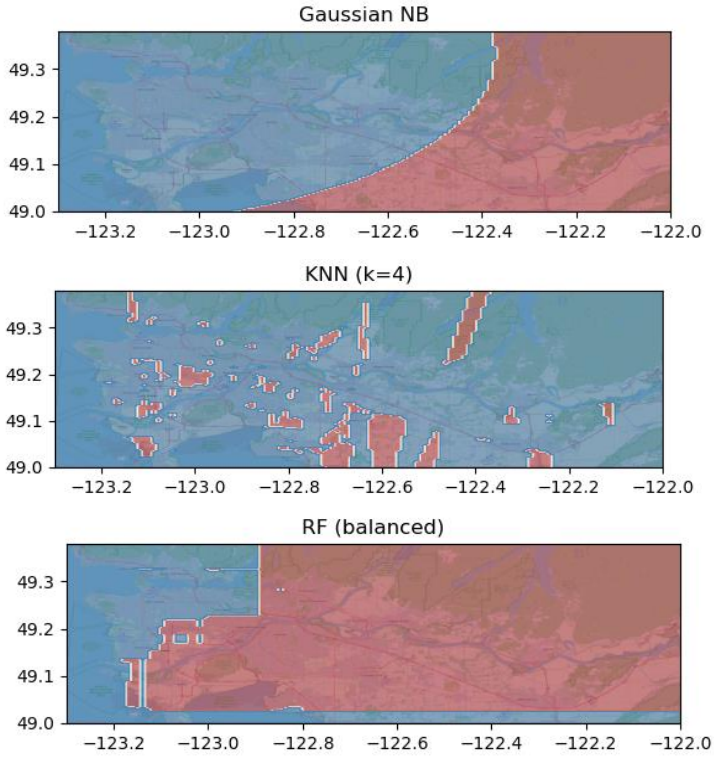


It allowed us to actually look at the data distribution in accordance with the map of the general locations. Moreover, with the pattern distribution given by these graphs, we can already make clear predictions and conclusions. For example, all the restaurants and fast-food locations are generally located in similar areas and it is due to commercial zones all around BC. Furthermore, the area with the thickest density is based around Vancouver Downtown which is not exactly a surprising result.



This graph shows the quantity of both official franchises and non-franchised locations representation. Franchise and non-franchised chains are paired against each other. They have a completely different visual interpretation as official franchises are not that large in individual or original brands but are extremely developed with the amount of branches per franchise. Non-franchised locations have it all in reversed notion. As there are not many branches or open restaurant locations under one name, they certainly beat it in the sheer number of unique locations. Certain conclusions come up right away.

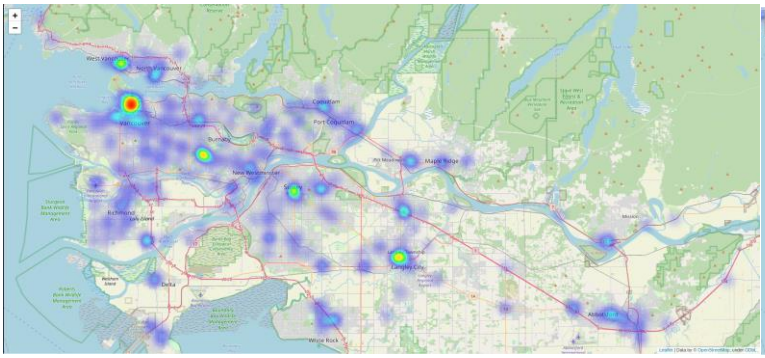
Results



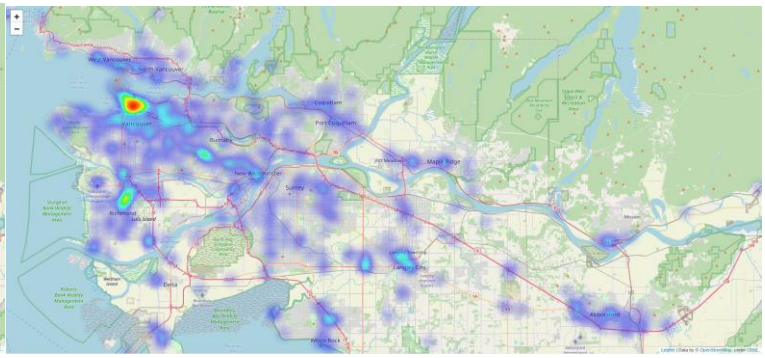
Here we tried to use Gaussian NB, K-Nearest Neighbours, and Random Forest classifiers to gain insight into the data. Gaussian NB gives us some indication that franchises (red) tend to be located in the South East portion of Vancouver, relative to non-franchises (blue). From the KNN model, we see that in most places of Vancouver, you can expect to find mostly non-franchises, but within red areas of the figure it is more likely to be near a franchise restaurant. Finally, we used a weighted Random Forest model to get a more precise idea of the difference in locations of franchises and non-franchises. We can see that the North West of Vancouver is particularly more dense in non-franchises than the rest of Vancouver.

This part gives a more clear perspective of the location densities and its distribution. We used several different visualization methods to get density as clear as possible. In this case it was a heat map that allowed us to see the clear zones of the highest density.

Franchised data

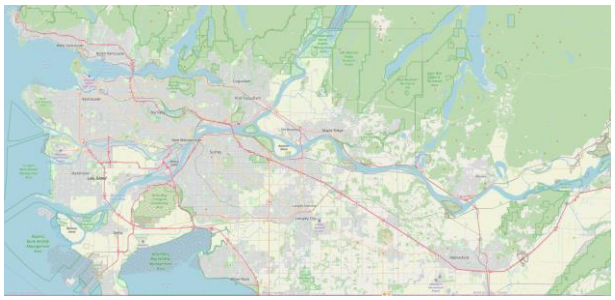


Non-Franchised data

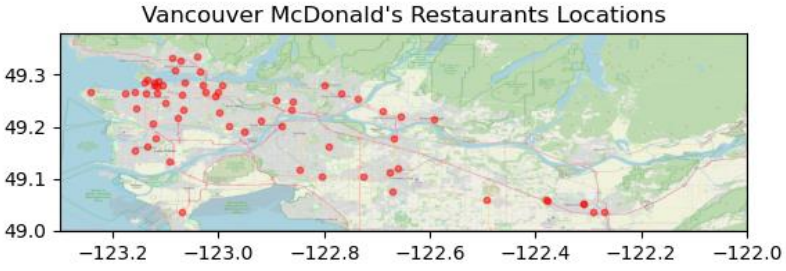


The downtown area can be considered the “hottest” by the result of these heat maps for both Franchised and Non-Franchised locations. The biggest difference between them is identified in that Franchised chains are located more densely to each other and are mostly placed in certain city areas (either city centers or dense commercial areas). At the same time, Non-Franchised locations are less condensed to one point (except downtown) and are scattered almost evenly all around the Vancouver area in lesser qualities everywhere else.

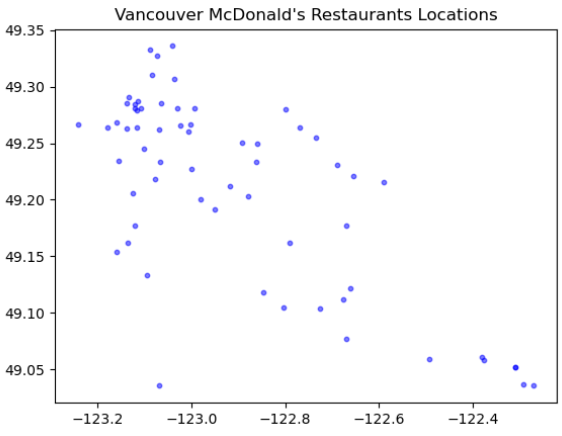
Additional Visualization



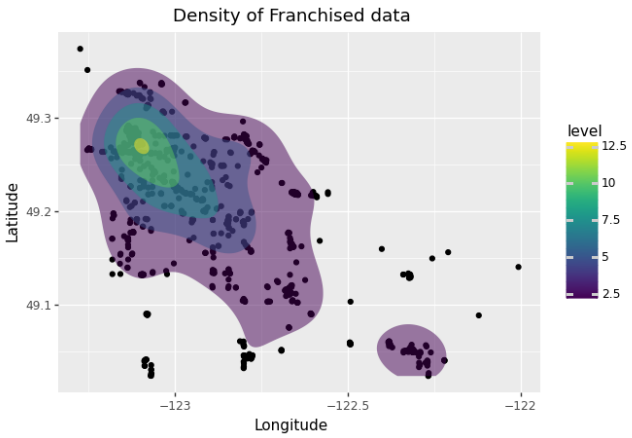
This map image was used for several map-oriented plots in this project. It was essential for the data interpretations and distribution visualization. In addition, it provides a sense of reality to the data visualization that gives it a feeling of credibility on top of being a great piece of world cartography.

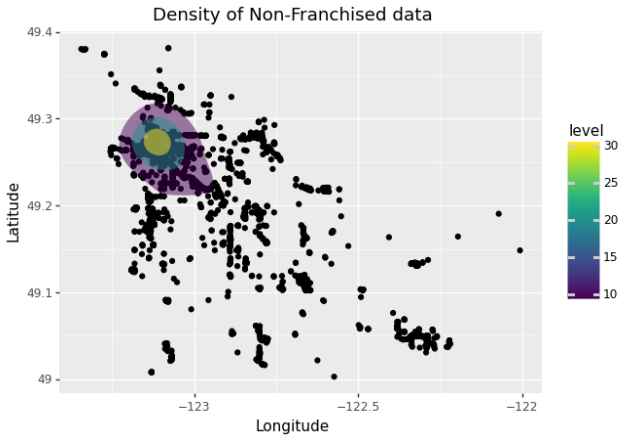


As an example of looking at one chain of restaurants/ fast-food places we took McDonald’s because it is arguably and statistically one of the largest of the eating options in terms of branch numbers. Especially with the addition of a plot where all locations were counted and its results, this plot gives a great perspective about how popular and common some of the famous dining options are.



This is another example of McDonald’s franchise representation. Here on this scatterplot it is easy to see the major density spots for this particular chain. In particular, we can clearly distinguish longitude and latitude corresponding to the area of Downtown Vancouver and around it (that is 49.2820° N, 123.1171° W according to Google).





This density representation method showed results of plotnine's ggplot tool. It is a great perspective of density for both data sets. (Even though the levels set differently for both data sets)

Limitations

Limitations, Issues, Problems that were experienced throughout project development.

Data collection:

- Data provided by an external source which could have mistakes. For example, a list of all the franchise and non-franchise names that were used to filter out initial data was not easy to find and check for truthfulness. After some testing we decided on the source that is used in the project and even though it was our final choice it still contains certain errors. In case of non-franchise names, there are several that are not categorized properly (as an example of such, we found "cactus club" that certainly can be considered as a chain with more than 10 locations in BC but is not marked as an official franchise in our identified data).
- Data will not match reality in the future as restaurant locations open and close. We might expect plenty of single-restaurant places closing and opening at high rates in the current economy. Meanwhile, popular franchises like McDonald's, Burger King and other quickly growing fast-food chains are known to prevail in an economical crisis. It would not be a surprise if in the next decade many of the non-franchise locations that we categorized will either close or change to some other marketing strategy while franchised chains will remain unaffected.

Data cleaning:

- We assumed that all chains contain a “brand” in their JSON tags, and all non-chains do not
    - We know there may be room for improvement because some definitions suggest that 3 or more restaurants is a chain, and the “All type locations” figure shows that some data is labeled as non-chains but have multiple locations nonetheless. We could improve by forming a clear definition of what constitutes a chain and labeling more thoroughly and selecting several official sources of franchise names in BC to include every potential chained franchise.

Data representation:

- Data was in a particular form that made choosing and making plots hard (for example box-plots and histogram were hard to implement with longitude and latitude and did not have any good shapes). In particular, histograms through matplotlib did not give any good result with all possible bar values.
- Pandas and matplotlib plotting options provided results with a little difference overall. After trying both methods while trying to get the best data representations, the view that matplotlib library provided was more customizable and easy to approach, so it was decided to mostly use it for plots.
- Density representation is not an easy characteristic of data to illustrate without a mistake. Several methods and representation methods were tried and used to find the best visual format.
- Plotnine’s ggplot was hard to work around and assign specific level values for both data sets. With chain and non-chain data being plotted exactly the same way, graphs show a significant value difference that makes visualization look biased.

Project Experience Summary

Minjae Shin

Worked on data refining process that involves refined Open Street Data and webscraping for chained restaurants' names, and combined them by matching and joining with new feature called "is_franchise" that stores Boolean value to represent whether the place is chained or not. Also, involved in visualizing chained and non-chained restaurants on the map by bounding the map size with minimum lat, lon and maximum lat, lon value to exactly fit and scale the map for the scatter plot of restaurants' locations.

David Ligocki

Helped with data cleaning by searching through the JSON data for specific tags that would indicate a franchise. Created Random Forest, K-Nearest Neighbor, and Gaussian NB classifiers to find trends in the locations of each type of restaurant, and tweaked values to improve performance. Plotted classifier predictions onto the map of Vancouver to make it easier to visualize the data. Analyzed classifier results to understand what trends exist in the data and how they can be interpreted.

Dmitrii Beliaev

Focused on data representation methods. Made an example of visualization of one of the largests chain restaurants and its branches using both data representation with a real map and clean data with statistical plotting tools. Prepared data and plotted visualization of data points' densities and density estimation using techniques and libraries like heat maps via folium and ggplot's stat_density2d, which allowed for a clearer further data analysis. Made certain predictions and conclusions based on these visual representations. Worked and helped on the report's overall structure, content and format.