Tauseef Kashtwari (tauseefk@sfu.ca)

CMPT 353: Computation Data Science Report

Overview / Introduction

For the final project, we used some statistics and visualization tools to analyze the amenities data that was provided to us in a reasonable JSON format. The provided JSON data set was a modified version of the original XML data obtained from **OpenStreetMaps (actually Planet.osm).** OpenStreetMaps is a free resource which provides map data for analysis and use.

Project Experience Summary (Accomplishment Statement)

- **Tauseef Kashtwari**
  - Incorporated Statistics to compute average and standard deviation on columns of a dataframe in Jupyter so as to assess variability in locations of restaurants and cafes when plotted on a map.
  - Eliminated missing values and extracted filtered data using conditional statements to create subsets of data (such as chain and non-chain restaurants) for further modeling and visualization (plotting).
  - Integrated visualization techniques and statistical tools to produce meaningful map plots such as heat maps, density plots using the Folium library so as to assess location of restaurants based on the number of branches.

Questions Asked Corresponding to the Problem:

1. Are there areas in Greater Vancouver with more chain restaurants? Is there some way to find the chain places automatically and compare their density relative to non-chains ?
2. **NEW :** Are there any differences in the density of chain restaurants amongst the amenity categories ? In essence, we are looking to map and deduce any differences between the location of chain restaurants based on their amenity types (restaurant, fast-food and cafe).

Corresponding problems being addressed:

Tauseef Kashtwari (tauseefk@sfu.ca)

A. Locate the areas in Greater Vancouver that have a greater number of **chain restaurants** (e.g A&W) and compare their density against **non-chain restaurants** (e.g 1029 Cafe), which have only one branch / outlet.
B. **Categorize** and **locate** the **chain restaurants** based on their **amenities** - restaurant, fastfood and cafe and compare their densities.

Problem A: Chain vs Non-Chain Restaurants Density Comparison

Data Gathering

The JSON data set was already provided to us as an input file, which was acquired from OpenStreetMaps. It consisted of exactly 8169 entries and 6 columns / attributes (lat,lon,timestamp,**amenity**,name and tags). We thought that the most important attribute was **amenity**. We did not collect any further data for our analysis as the provided data was adequate.

Process of Data Cleaning:

In order to clean the data, we first remove all rows with missing data (NAs) from the dataset. First, we decided to print all the distinct amenities in the data set. Then, we chose a subset from the list of amenities ('cafe', 'fast_food', 'bistro' and 'restaurant') and decided **not** to include 'juice_bar' as a restaurant. Thus, our cleaned data set consisted of four amenities (**in the same dataframe**) and had 4631 rows (roughly over 50% of the original data, which had 8169 rows).

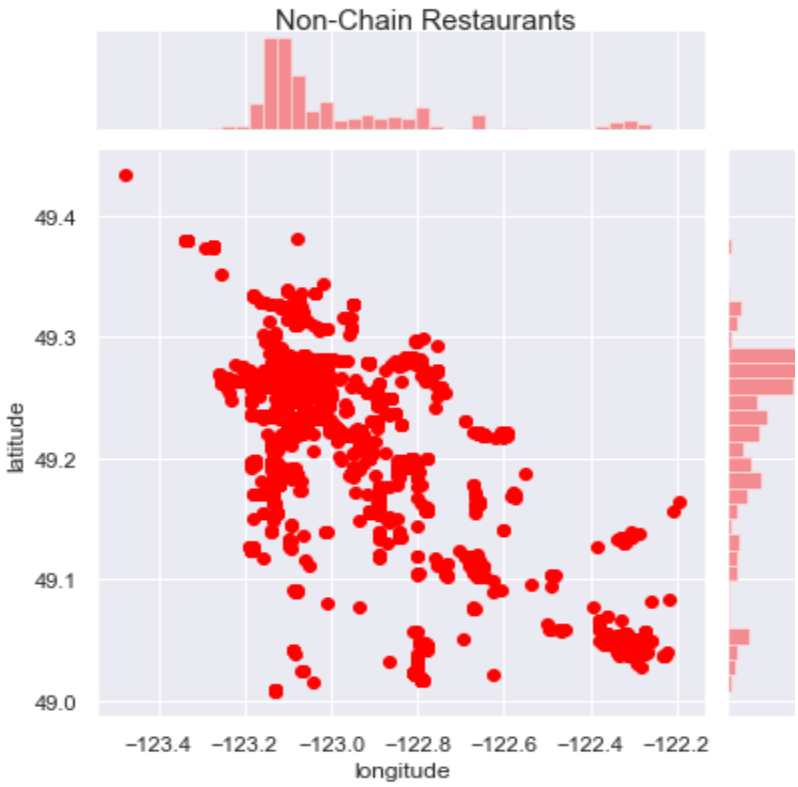
We counted the number of unique restaurant names in the dataframe and assigned it as a new column (**total_branches**). A chain restaurant by definition has more than one branch, whereas a non-chain restaurant has only 1 branch. Adding a new column allowed us to split the data frame into two: restaurant chains (restaurant names occurred more than once in the data frame i.e **total_branches > 1**) and non-chain restaurants (restaurants with a single occurrence of their name i.e **total_branches = 1**). The 333 chain restaurants (333 names) had a total of 110,313 branches. In contrast, there were only 2768 (non-chain restaurants / single branch restaurants).

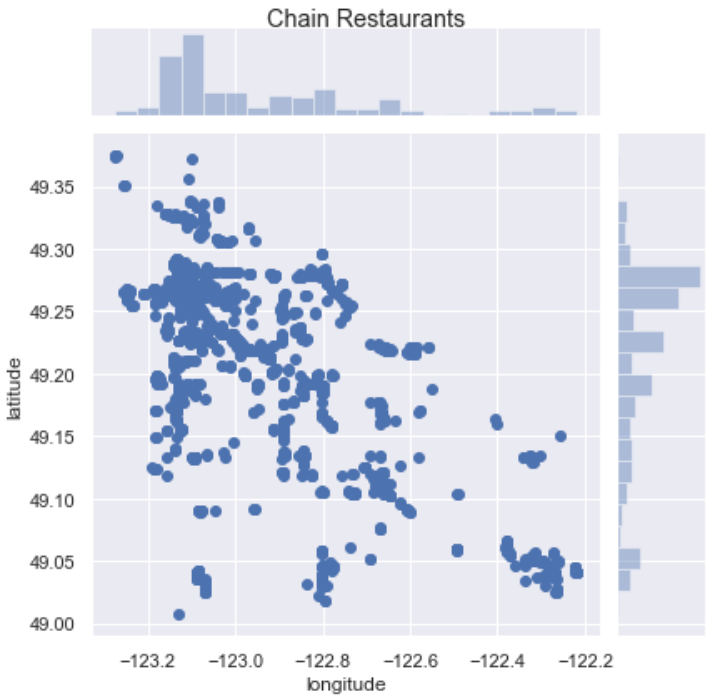
Data Analysis and Visualization

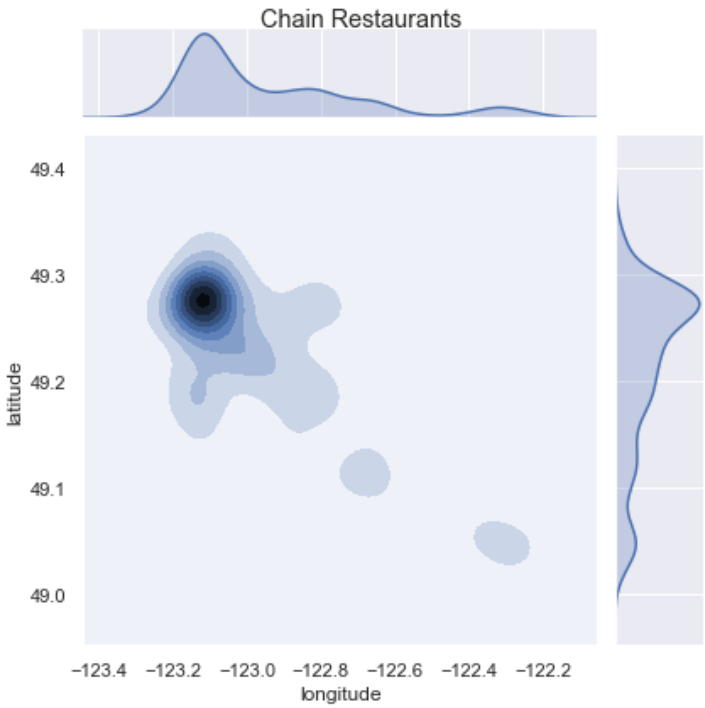
Upon completing the data cleaning process, we first analyzed the data using plots from Seaborn library. In order to check the distribution of chain restaurants, we

used a jointplot (from the SeaBorn library) and a KDE plot (to further analyze the densities). Then, we used a heatmap to visualize the number of restaurants (both chain and non-chain) on an OpenStreetMap.

Jointplot and KDE(density) plots

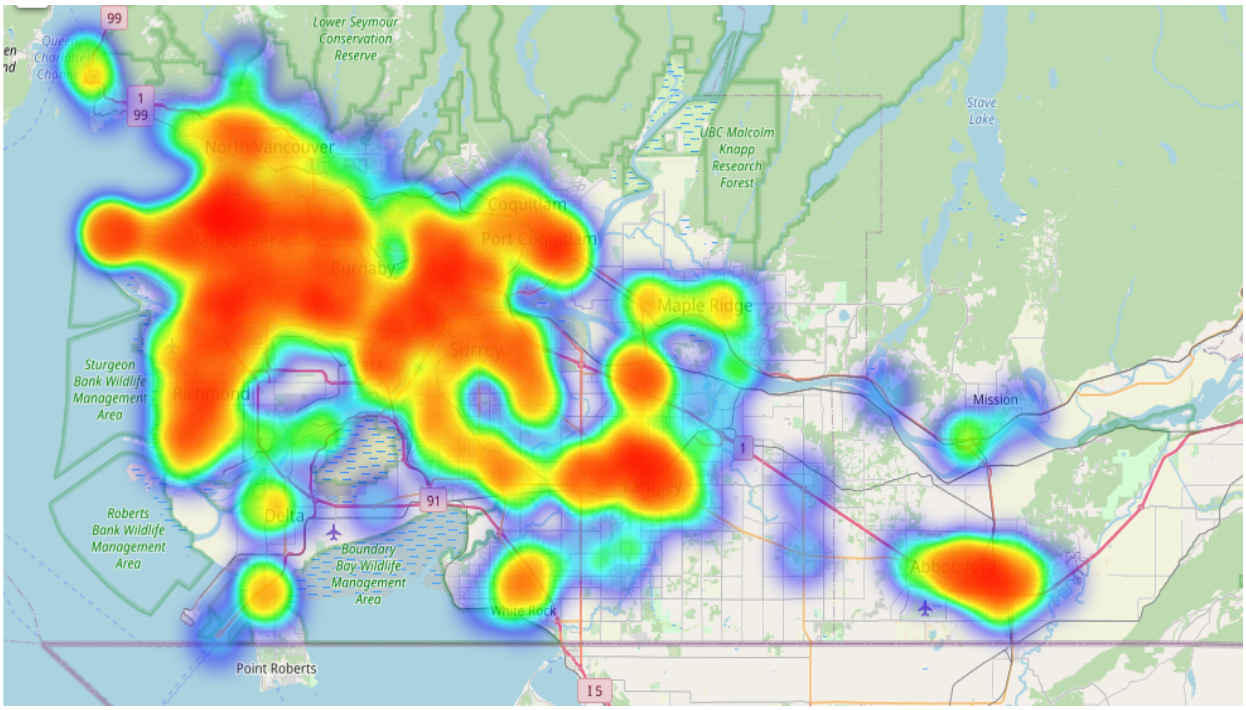


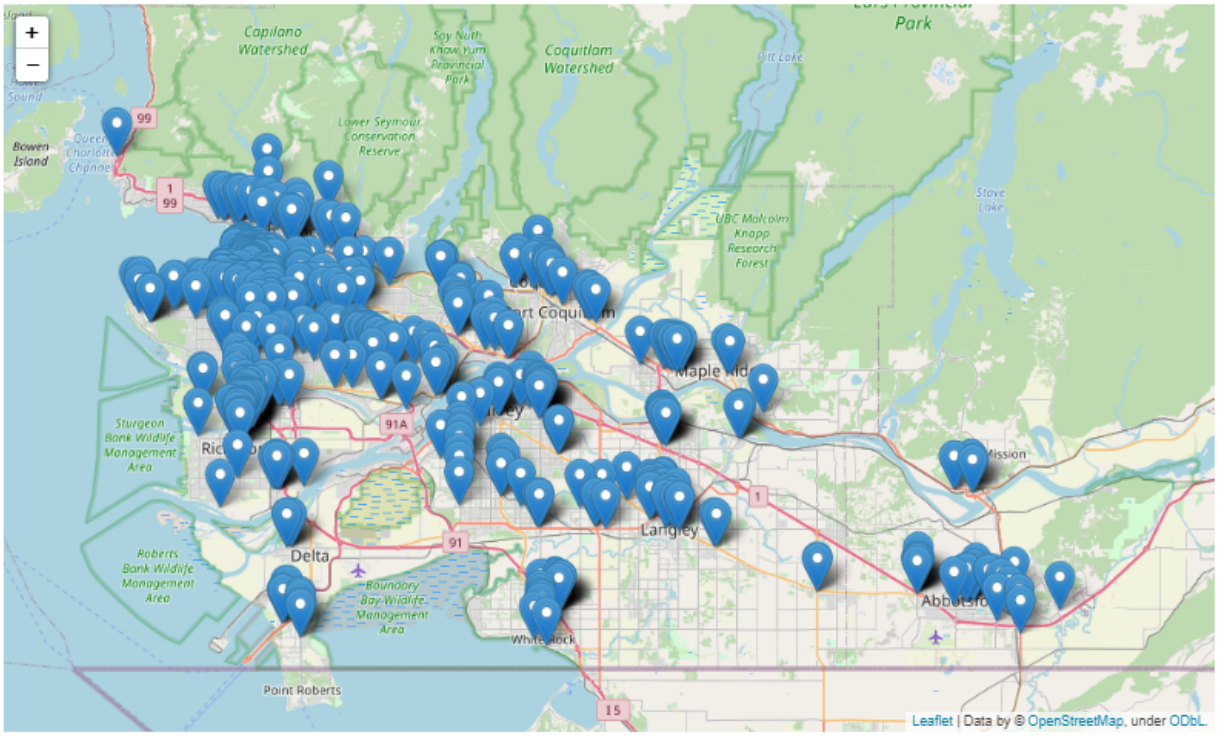




Tauseef Kashtwari (tauseefk@sfu.ca)

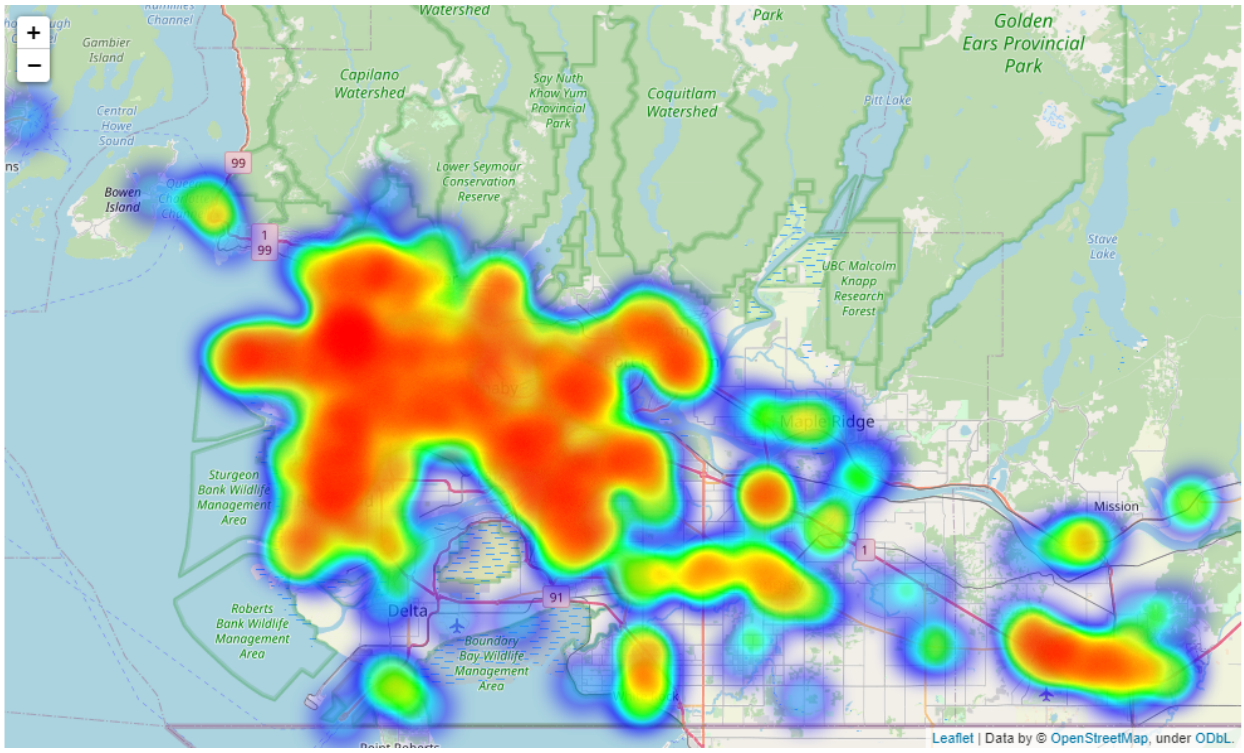
Heat Map and folium Map Chain Restaurants

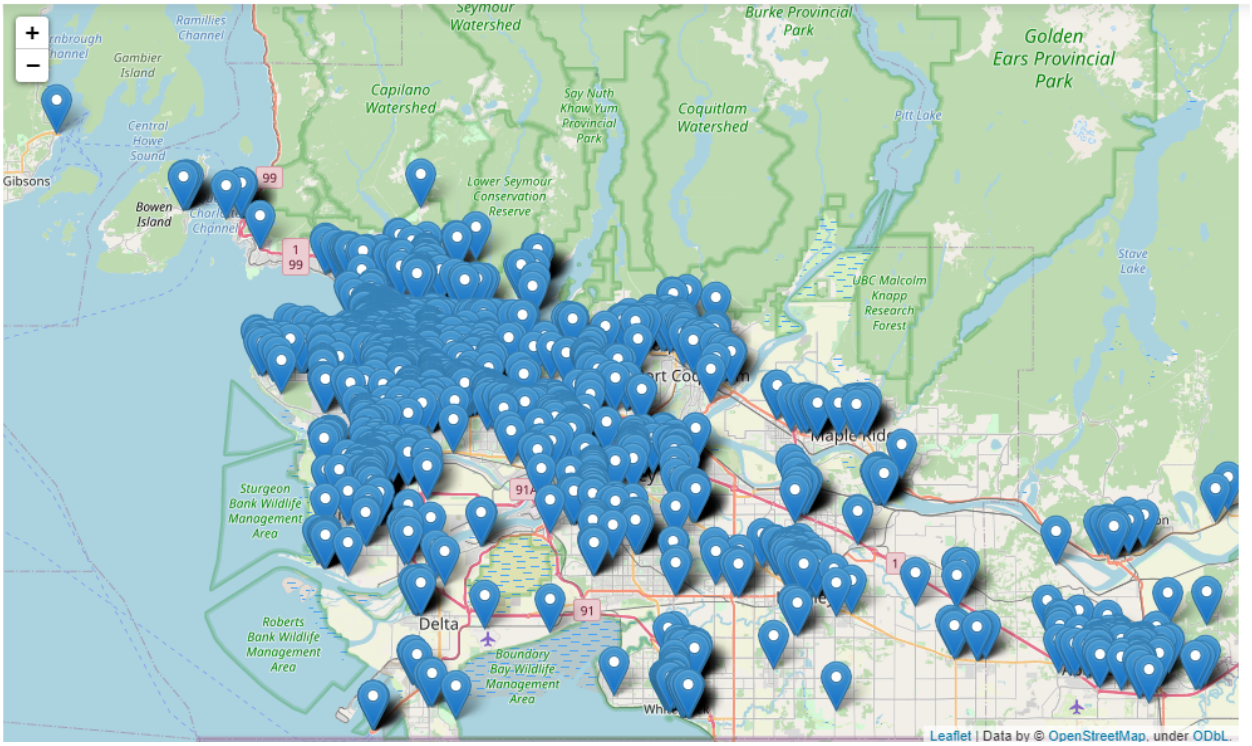




Tauseef Kashtwari (tauseefk@sfu.ca)

Heat Map and folium Map Non-Chain Restaurant





Tauseef Kashtwari (tauseefk@sfu.ca)

The jointplot (**red**) depicts a cluster of many coordinate points of non-chain restaurants around coordinates (49.25, -123.1). In addition, the jointplot for chain restaurants (**blue**) depict a cluster of coordinate points around (49.25, -123.1) as well. Furthermore, a KDE plot (of contours) suggests that there are more number of restaurants (both chain and non-chain) around those coordinates (darker shade denotes higher number) than elsewhere. This indicates that the non-chain and chain restaurants (both categories) are densely packed in Downtown Vancouver (larger area)

We computed some statistics to further validate our findings, which are summarized in the table below:

Chain Restaurants
Average (mean) Latitude: 49.21178
Average (mean) Longitude: - 122.94328
SD Latitude: 0.07942
SD Longitude: 0.24112

Non-Chain Restaurants
Average (mean)Latitude: 49.22199
Average (mean) Longitude: -122.98433
SD Latitude: 0.07629
SD Longitude: 0.22440

As we can see the mean latitude and longitude for chain and non-chain restaurants are similar. However, **chain restaurants** have a **higher standard deviation** than **non-chain restaurants**. Hence, chain restaurants are scattered around the city compared to non-chain restaurants, which are mainly located in DownTown. This is further confirmed by the **red** and **BLUE KDE plot** (and also by the bar chart located on the sides).

Conclusion

From the statistical analysis and visuals, we can confirm that both types of restaurants have a similar distribution pattern. However, from the statistics and plots above, we can confidently confirm that we've the highest number of non-chain and chain restaurants in Vancouver (most of them are in Downtown), followed by Bunraby and then Abbotsford (in decreasing number / concentration).

Extra information (Aside Note)

We also counted the number of unique entries for both chain and non-chain restaurants:

- **This tells us that there are 2768 non-chain restaurants and 333 chain restaurants (NOTE: this is not the total branches but just the restaurants).**

```
non_chain_restaurant.name.nunique()
```

```
2768
```

```
chain_restaurants.name.nunique()
```

```
333
```

Problem B: Categorization and comparing densities of chain restaurants based on amenities

Data Gathering

Using the previous problem, we already had two dataframes for chain and non-chain restaurants obtained from the restaurant dataframe (with 4631 rows). The chain restaurants dataframe consisted of 1863 rows and 7 columns (lat,lon,timestamp,**amenity**,name,tags and an additional column - total number of branches). We did not collect any further data for our analysis as the provided data was adequate.

Tauseef Kashtwari (tauseefk@sfu.ca)

Data Cleaning:

In order to filter the data, we first subsetted the **chain restaurants dataframe** using the following **unique amenities ('cafe', 'fast_food', and 'restaurant')**. In addition, we only chose This gave us three distinct dataframe (one dataframe for each unique amenity listed above).

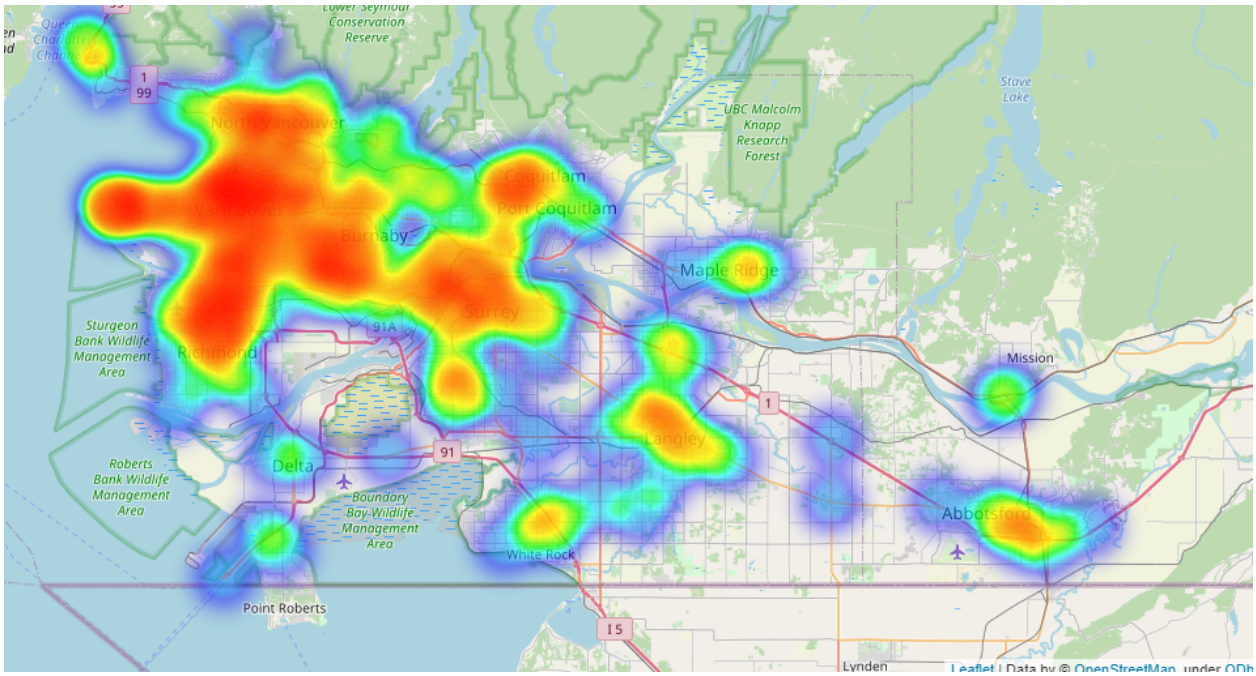
- Dimension of each dataframe
  - **Coffee (Cafe) Chains: 543 rows x 7 columns**
  - **Fast-Food Chains: 753 rows × 7 columns**
  - **Restaurant Chains: 567 rows × 7 columns**

Data Analysis and Visualization

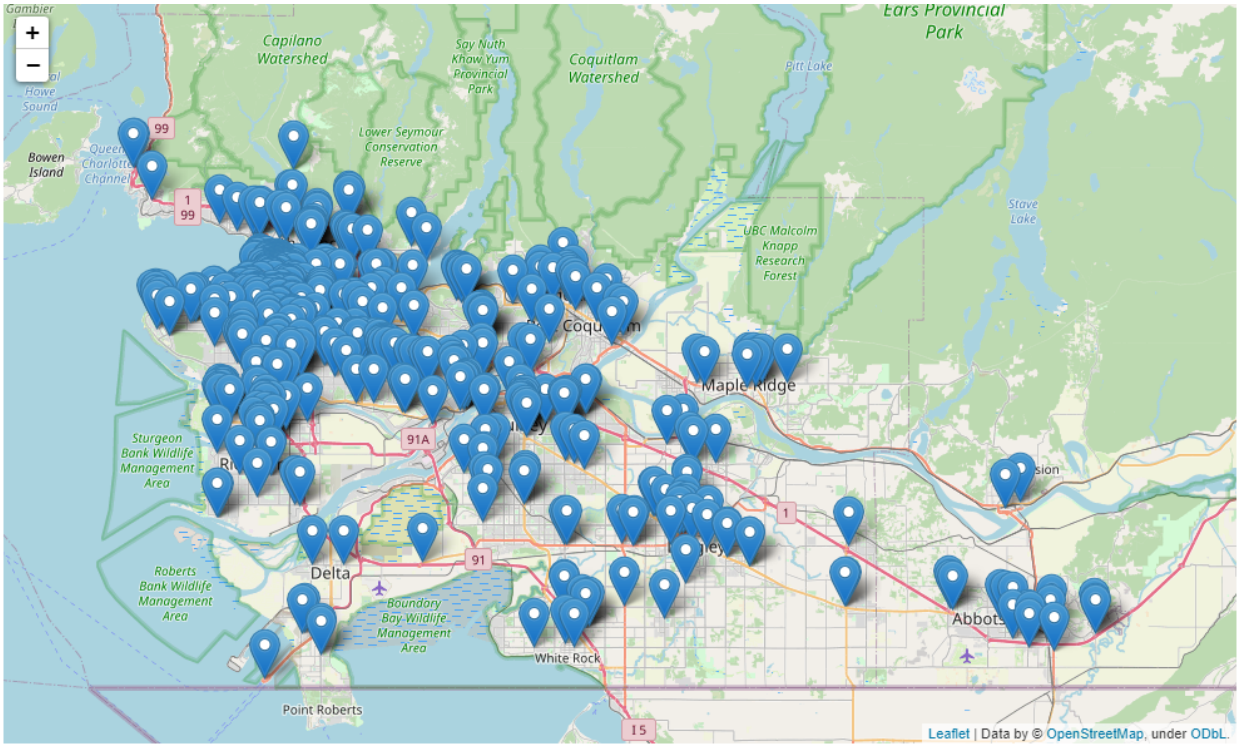
Upon completing the data cleaning process, we analyzed the data using plots from Seaborn library. In order to check the distribution of each category of chain restaurants (cafe, fast-food, and restaurant), we used a jointplot (from the SeaBorn library) and a KDE plot (to further analyze the densities).

We also used a heatmap and folium Map for visualizing the different chains, plotting it on an OpenStreetMap.

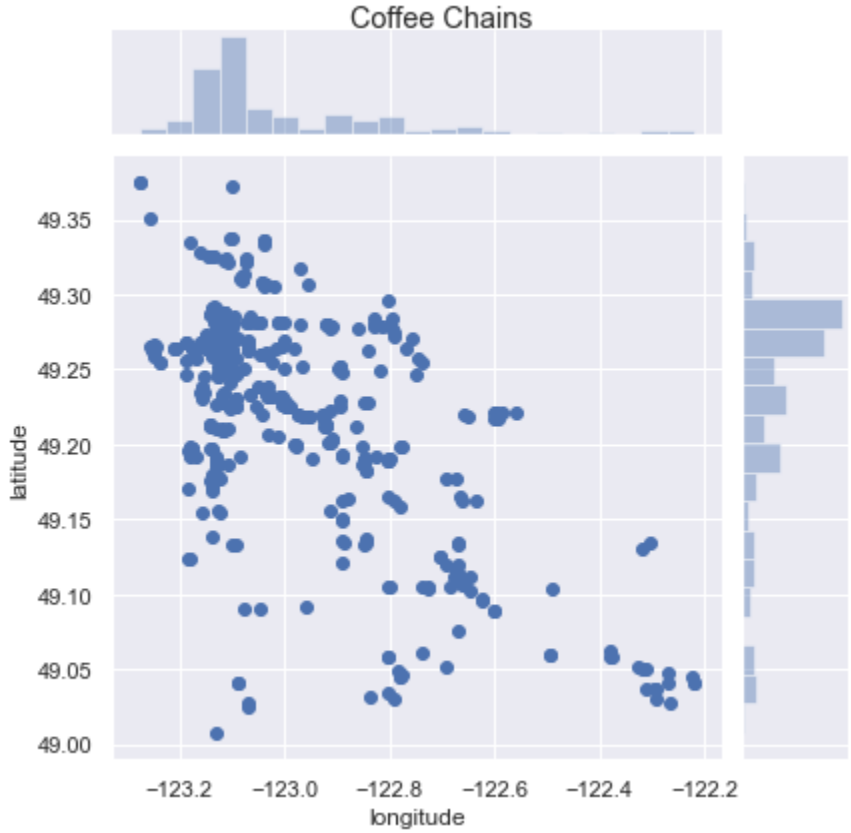
Heat Map and folium Map of Coffee Chains



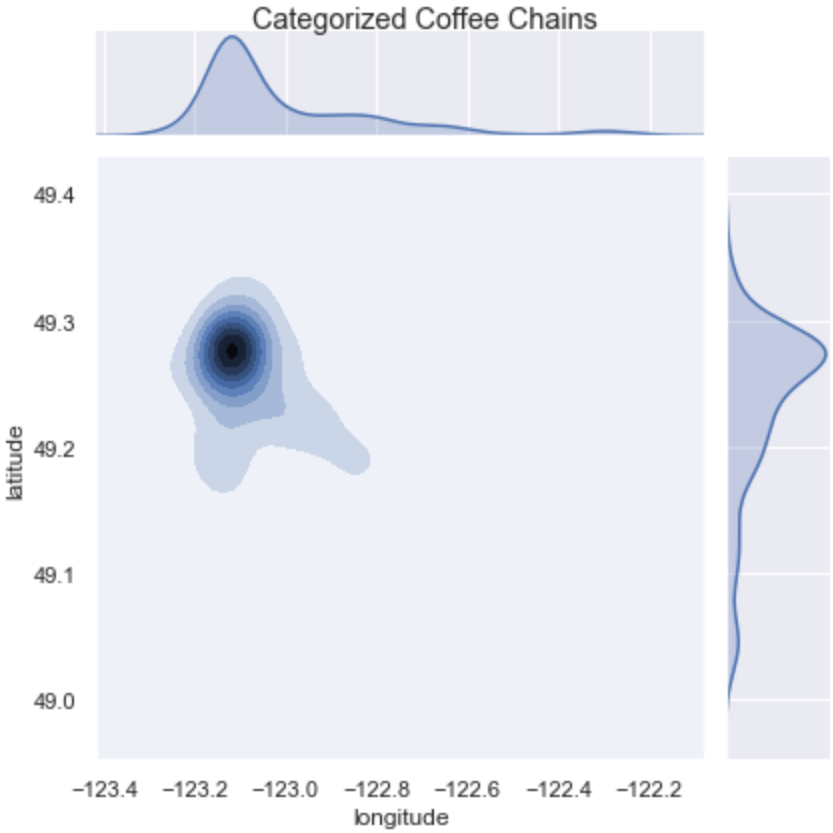
Tauseef Kashtwari (tauseefk@sfu.ca)



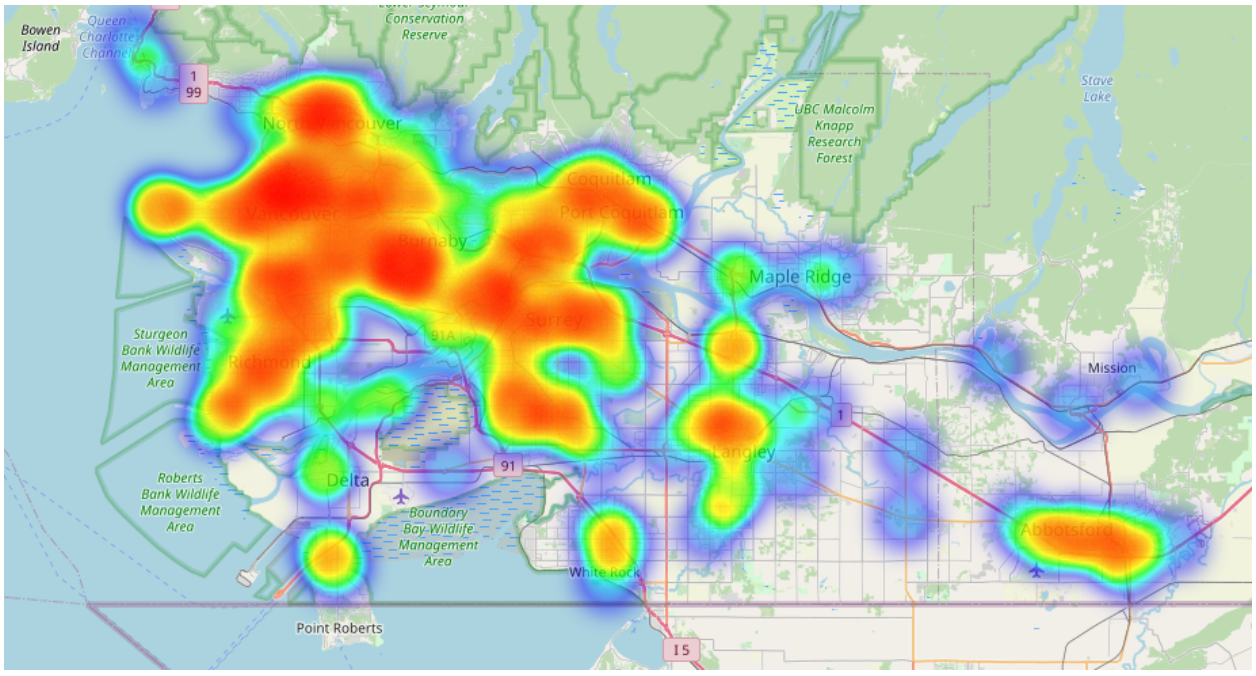
Jointplot and KDE(density) plot of Coffee Chains



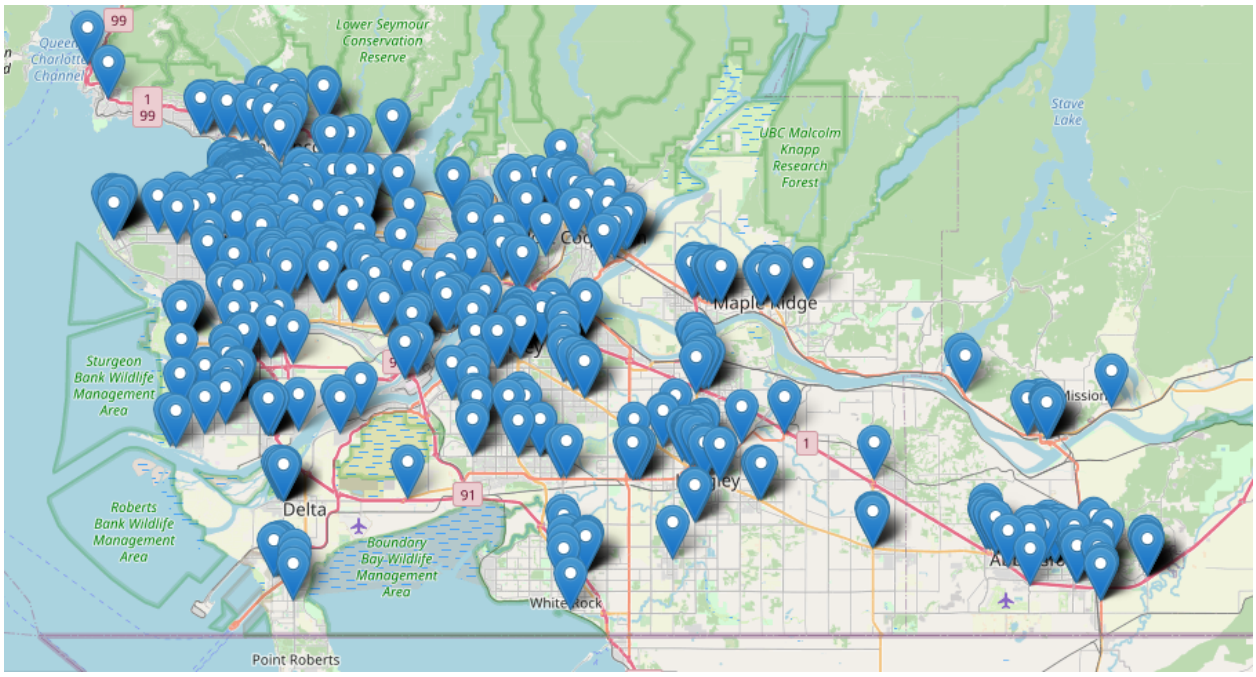
Tauseef Kashtwari (tauseefk@sfu.ca)



Heat Map and folium Map of Fast-Food Chains

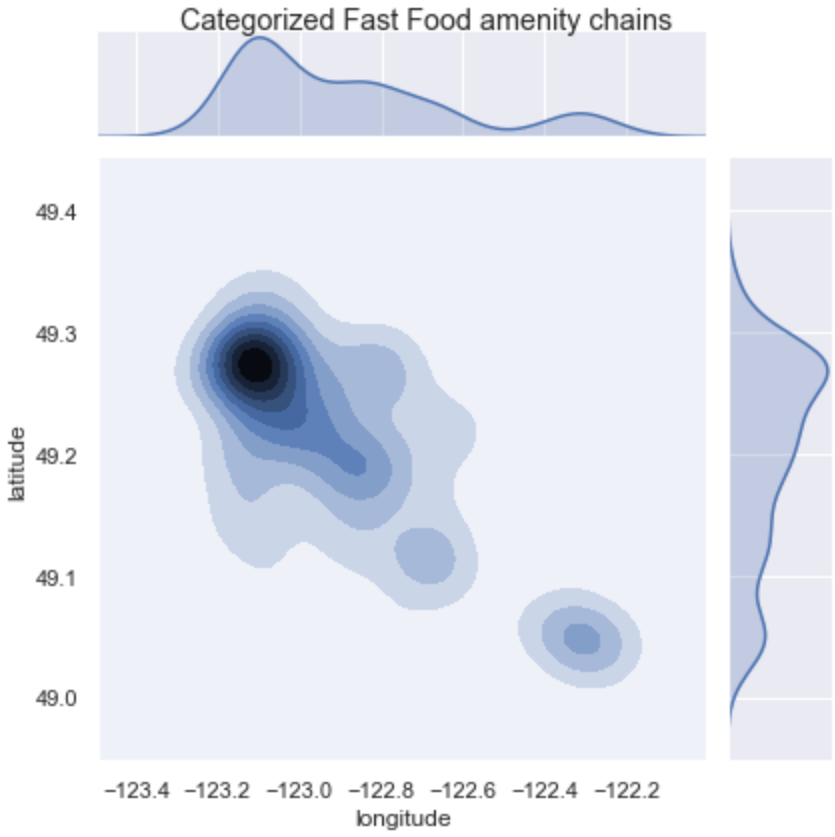


Tauseef Kashtwari (tauseefk@sfu.ca)

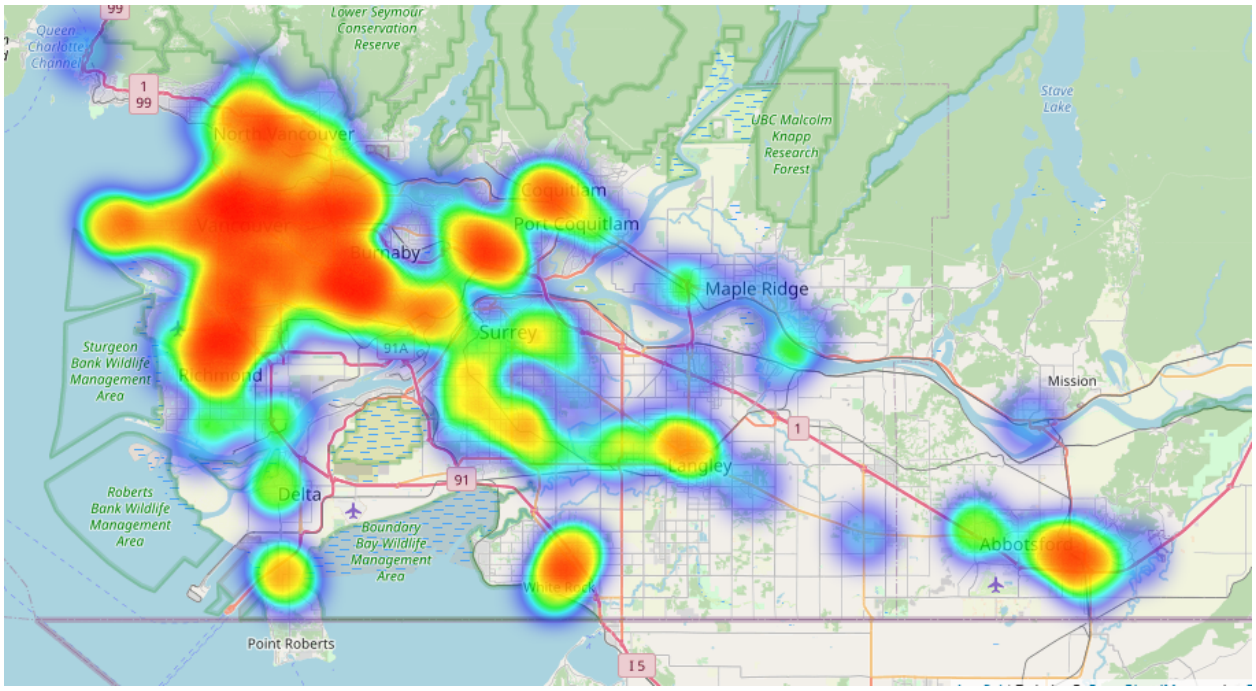


Jointplot and KDE(density) plot of Fast Food Chains

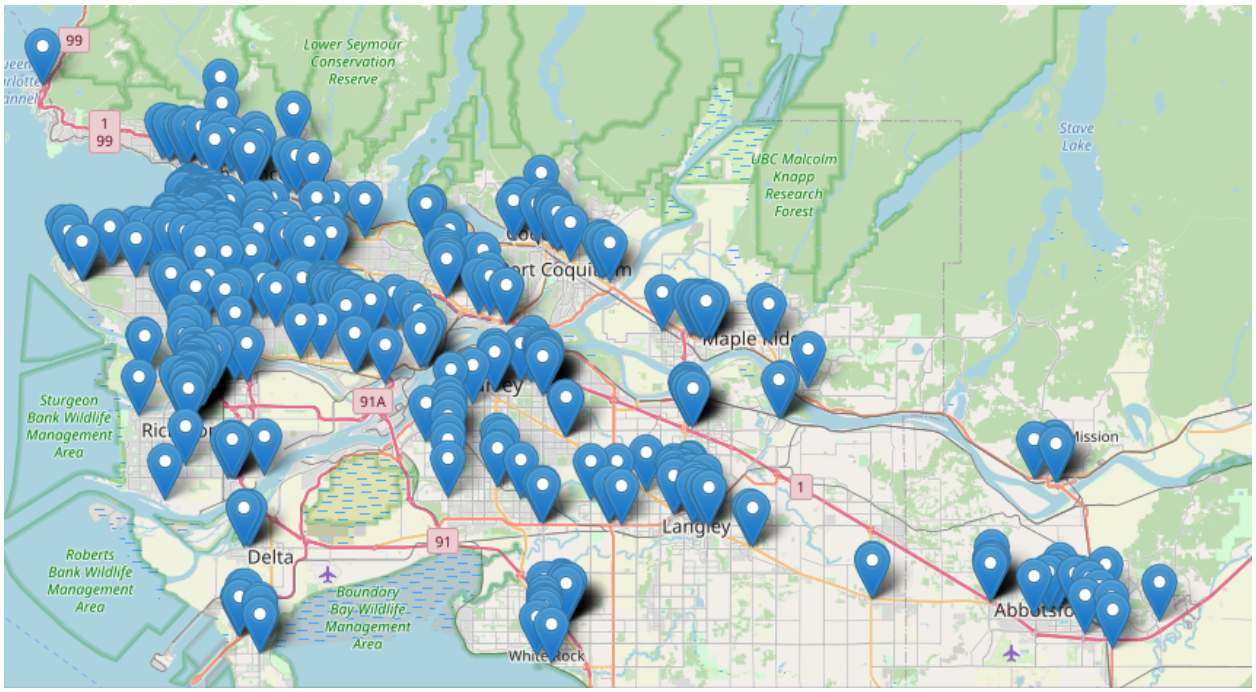
Tauseef Kashtwari (tauseefk@sfu.ca)



Heat Map and folium Map of Restaurant Chains (Amenity = Restaurant)

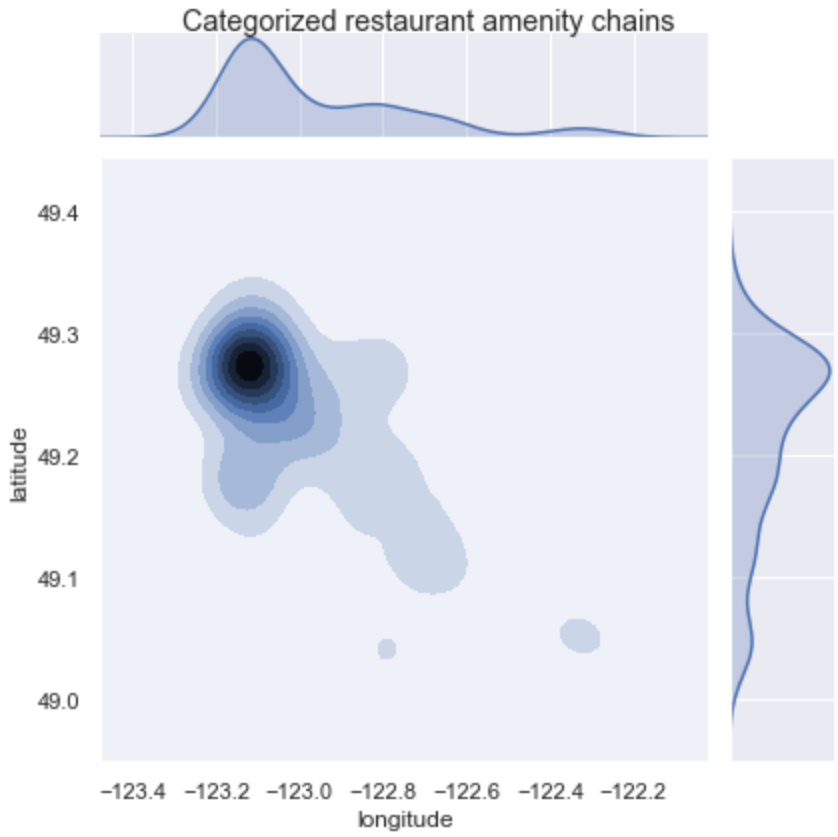


Tauseef Kashtwari (tauseefk@sfu.ca)



Jointplot and KDE(density) plot of Restaurant Chains

Tauseef Kashtwari (tauseefk@sfu.ca)



The three joint plots depict a cluster of several coordinate points (first plot - coffee chain, second plot - fast-food chains and third plot - restaurants chains) around coordinates (49.25, -123.1).

We computed some statistics to further validate our findings, which are summarized in the table below:

Coffee Chains
Average (mean) Latitude: 49.22963
Average (mean) Longitude: -123.0015
SD Latitude: 0.07185
SD Longitude: 0.20266

Tauseef Kashtwari (tauseefk@sfu.ca)

Fast Food Chains
Average (mean)Latitude: 49.19860 (rounded)
Average (mean) Longitude: -122.88868
SD Latitude: 0.082537
SD Longitude: 0.264562

RESTAURANT Chains (actual amenity = 'Restaurant')
Average (mean)Latitude: 49.21217
Average (mean) Longitude: -122.96003
SD Latitude: 0.078873
SD Longitude: 0.22680

As we can see the mean latitude for coffee (cafe) chains, restaurant chains and fast food chains are also similar (previously chain and non-chain had similar values too). However, **fast food chains restaurants** have the **highest standard deviation**. This makes sense as fast food chains are generally in multiple locations (and hence greater variability/spread). Hence, fast-food chains are scattered furthest apart around the city compared to restaurant and coffee chains. However, for all 3 chain types (restaurant, coffee and fast food chains), the means are similar as most of the branches of each chain type are located in DownTown Vancouver. This is further confirmed by the **BLUE KDE plot** (and also by the bar chart located on the sides), which shows a high concentration (proportion) of these chains (fast-food, restaurant, and coffee chains) located in DownTown Vancouver (darker shade denotes a higher proportion than lighter shades).

Conclusion

Finally, we've summarized our findings in bullet point form below:

- Restaurant, Fast-food, and coffee chains are mostly located (clustered) in DownTown Vancouver
- Therefore, the mean of the latitude and longitude for these 3 chains groups are similar.
- Also, the Standard deviation of the latitude and longitude are approximately the same
- Hence, there are not major differences in the density of chains across the amenity categories.

Limitations:

Initially, we both struggled to find partners, which took a lot of time in this virtual learning environment. Thus, time was a huge constraint, but we still managed to overcome this effectively. Upon forming a group towards the end of semester, Faisal had a family emergency a day before the due date (when we hadn't done much - did the project in a day). Thus, **I had to do the whole project myself (both implementation and report).**

If I had more disposal at time and found more effective group partners earlier, I'd answer at least one more question for further analysis of the data.