Computational Data Science Project Report

Analyzing Restaurants in Greater Vancouver

Abay Kulamkadyr

Aditya Kulkarni

Kenneth So

The Problem.

From the suggested topics, we chose to derive a different problem from the OSM idea. Using OSM data, we wanted to analyze the prominence of restaurant types, cuisines, and chains in cities within the Greater Vancouver area. We approached this with the idea of providing analysis for prospective business owners by answering these kinds of questions:

- *Is this restaurant idea common in the area?*
- *Does this area have competitors?*
- *Does this restaurant idea satisfy an unfulfilled opportunity space?*
- *How does the cuisine compare to the demographics of the city?*

To provide “answers” to those questions, we wanted to provide plots, charts, and OSM map visualizations. We used the OSM data as a starting point, and made use of other datasets to make up for where that data is insufficient. We chose to split up and explore several different possible avenues for expanding the breadth and depth of our analysis.

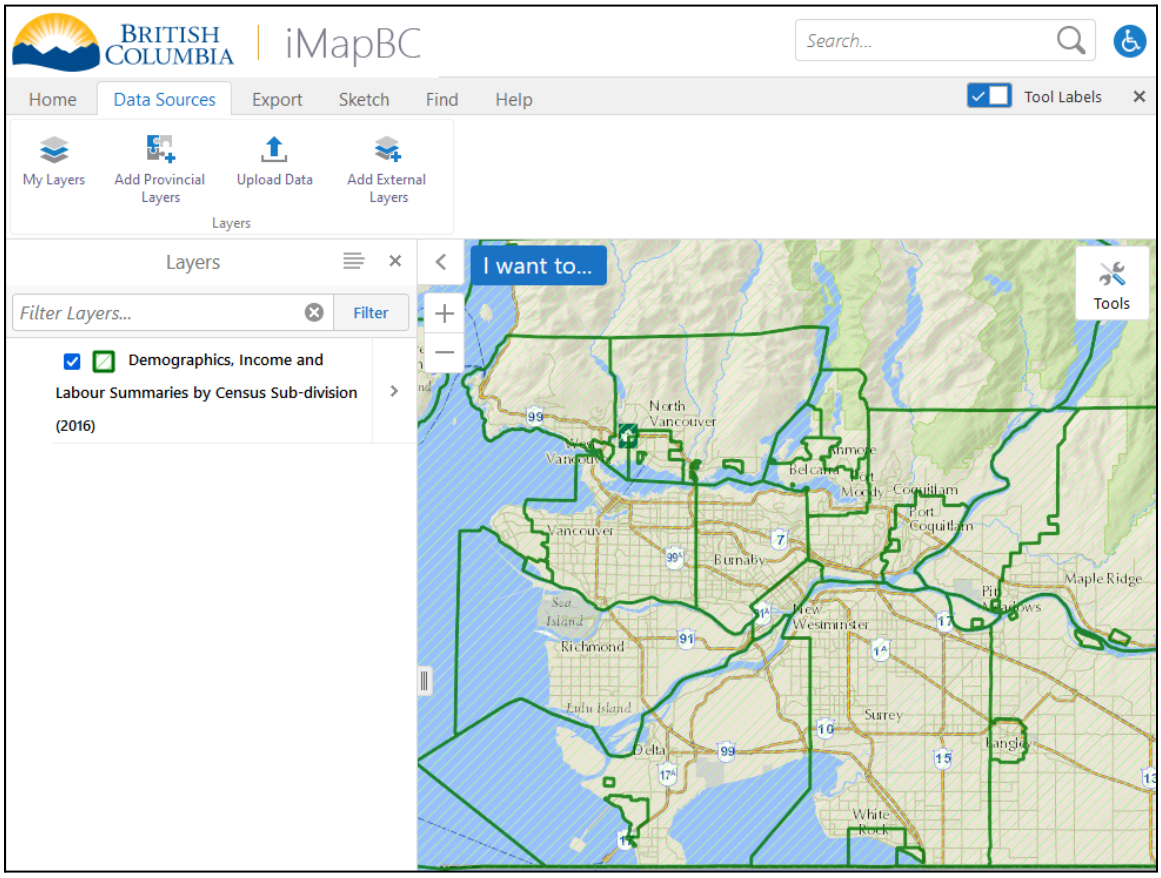
These approaches were as follows: Yelp restaurant data, Vancouver crime data, and Vancouver demographics data. We found a comprehensive Yelp dataset on Kaggle. We wanted to use this data to complement the OSM data to determine if there is a dominance of certain restaurant types in Greater Vancouver. We also found a dataset of crimes committed in Vancouver, which could help us see the amount of crime that happens in the area of a given restaurant. Lastly, we wanted to explore possible connections between the demographics of each city and the distribution and characteristics of its restaurants.

The Data.

The data we used was an updated version of the OSM Vancouver amenities, census data acquired from the British Columbia government, and Kaggle data related to restaurants on Yelp, and the prominence of crime in Vancouver.

For the OSM data, we noticed that the provided amenities JSON file only had data up to June 29, 2020. Given that the pandemic has resulted in significant turnover in the service industries, we felt it would be appropriate to use more recent data. We downloaded the latest OSM data from Geofabrik and used the provided python / pyspark scripts to generate an updated JSON of Vancouver amenities.

To acquire demographics data, we used the iMapBC API to download the data as shapefiles. iMapBC has an extensive catalog of the province's data, and each dataset has an overwhelming amount of information. We ultimately settled on using data for municipality boundaries, general demographic census data, and languages spoken at home census data. To use the data, we simply loaded the data's shapefiles as GeoDataFrames using Geopandas, extracted relevant columns, and joined together as needed.



iMapBC API Interface

For the Yelp data, we used a data set provided by Kaggle (version 3) that is available for academic purposes. The dataset contains approximately 11 GB of JSON files which provide information on businesses, their locations, user reviews, tips, star ratings, and other features. For our analysis, we extracted data related to food and restaurants businesses. The dataset contains information about businesses across eight metropolitan areas in the USA and Canada, but our project focuses on the areas within the Greater Vancouver area.

For the crime data, we also used a Kaggle data set. The dataset contains a CSV file with details about the crimes which occurred in Vancouver. It provides the type of crime, location of crime, year of crime and the neighborhood where the crime happened.

Techniques for Analyzing Data.

The results of these techniques can be found in the next section, Results and Findings.

OSM Data.

We were fortunate to find that the amenities data was well-suited for our purposes. The majority of registered amenities are restaurants, many of which have their cuisine listed in their tags. This made it convenient to filter and plot restaurants based on their cuisine.

cuisine 4692 addr:street 2703 addr:housenumber 2674 brand 1811 brand:wikidata 1782 brand:wikipedia 1755 takeaway 1743 addr:city 1719 opening_hours 1290 phone 1231 website 1173 addr:postcode 915 ...	coffee_shop 624 pizza 572 sandwich 378 burger 367 chinese 253 sushi 249 japanese 189 indian 145 vietnamese 139 mexican 130 chicken 116 bubble_tea 92 ...
The occurrences of each unique tag	*The occurrences of each unique cuisine*

The effectiveness of this dataset also made it convenient to use in conjunction with our other datasets. Our primary goal with this data was to analyze the spatial relations between restaurants of the same cuisine.

One of our first uses for the OSM data was to track the nearest competitors. First, we acquired the data of all food amenities. The user inputs the location of where they plan to open their amenity (e.g., cafe). We use the formula of a circle and remove all amenities that are beyond that circle (with the amenity as the center). All other amenities within the circle are potential competitors. We use plotly express to plot the locations of these amenities.

British Columbia Demographics.

From the iMapBC API, we acquired multiple pieces of shape data. First, we acquired the municipality boundaries of BC so that we could plot them on a map and visually separate the amenity data by these borders. From the demographic census data, we used the household income medians and population totals per city to derive a possible correlation between these statistics and the prominence of certain restaurants. In addition, population density was also added to the plot, and these statistics were used in conjunction with the OSM amenity data to calculate the density of specific cuisines in each municipality.

We also constructed bar plots for the overall distribution of the household incomes per city and the population distribution by age. For another dataset, we plot the distribution of languages spoken at home to see the ethnic demographics of a given city. An issue we faced with this data however was the scale of it - the census tracked hundreds of languages spoken. This meant we had to select languages, so we decided on languages that are commonly spoken in British Columbia. We visualized the distribution of these languages using a bar plot.

Crime in Vancouver.

From kaggle, we acquired data of crimes (version 3) that have occurred throughout Vancouver from 2006 to 2021. The user inputs the location of where he/she plans to open his/her amenity (for example cafe). We use the formula of a circle and remove all crimes that happened beyond that circle(with the cafe as the center). We also filtered out all the crimes that would not affect the cafe like bicycle theft or vehicle collision.

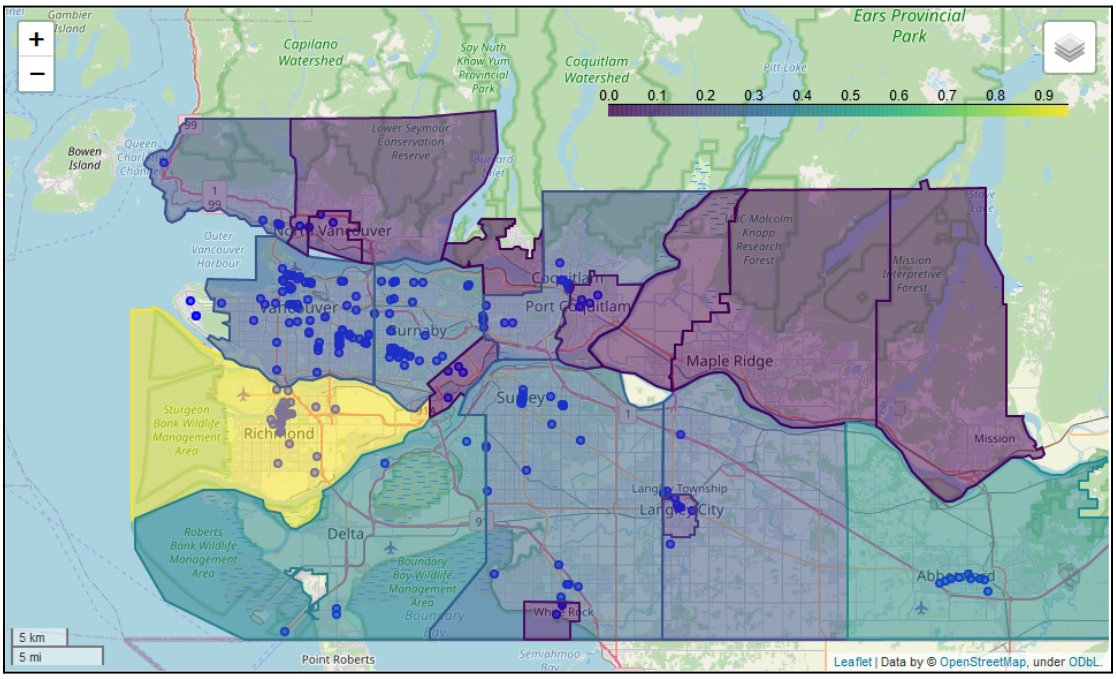
Yelp Data.

From Kaggle.com, we acquired 11 Gb of JSON files. To extract data on businesses related to the food industry which are within the geographical bounds of BC, we used the municipality dataset's shapefiles and latitude/longitude columns from the Yelp JSON files. We created a Spark program that determines if the latitude/longitude coordinates of a businesses' location are within the bounds of the municipality shapefiles. Then the program matches Yelp's 'categories' column with a list of categories related to food and cuisines that are provided on Yelp's website.

Results and Findings.

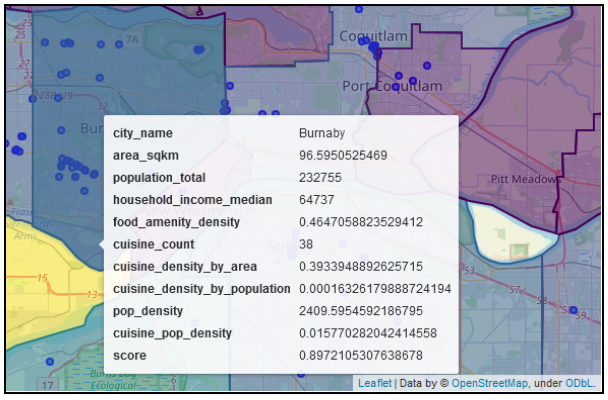
OSM Amenities and Demographics Data.

Using the OSM amenities data combined with the municipality and demographics data from iMapBC, we constructed an interactive map using Geopandas and Folium that visualizes the distribution of restaurants based on cuisine, and their relative densities per city. In the figure below, the distribution of Chinese restaurants is plotted on the map, and each municipality is coloured according to the density of these restaurants relative to its population density.



Geopandas Folium plot - distribution of restaurants, where cities are coloured based on density

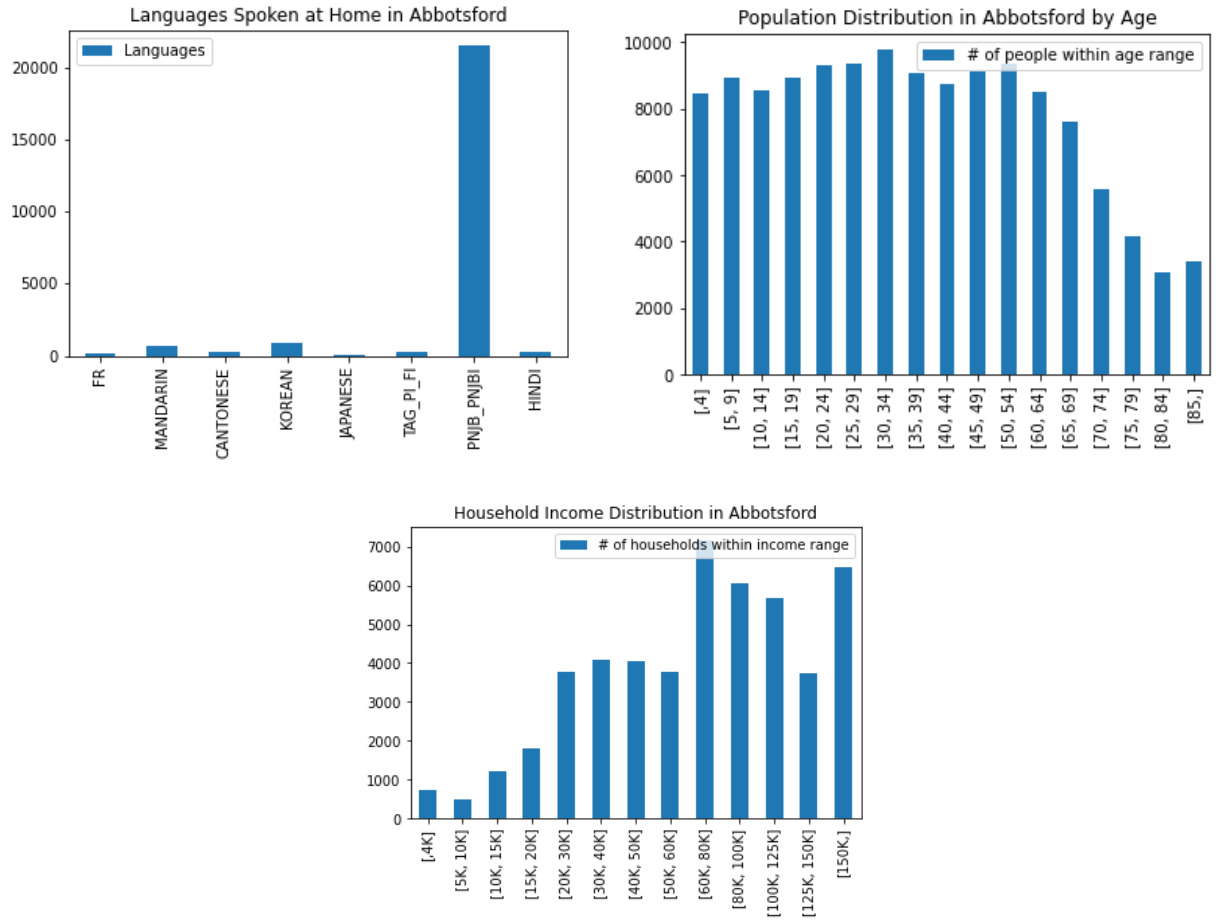
You can hover over each municipality to view specific information like household income, population, cuisine counts, and multiple density related statistics.



Pop-up after hovering mouse over Burnaby

A core aspect of this map is that two cuisines can be plotted at the same time, allowing the visualization of the distribution of two cuisines together. This, along with the aforementioned functionalities, can be seen using [this example instance of our map], since Github / Gitlab cannot natively render the maps in the notebook.

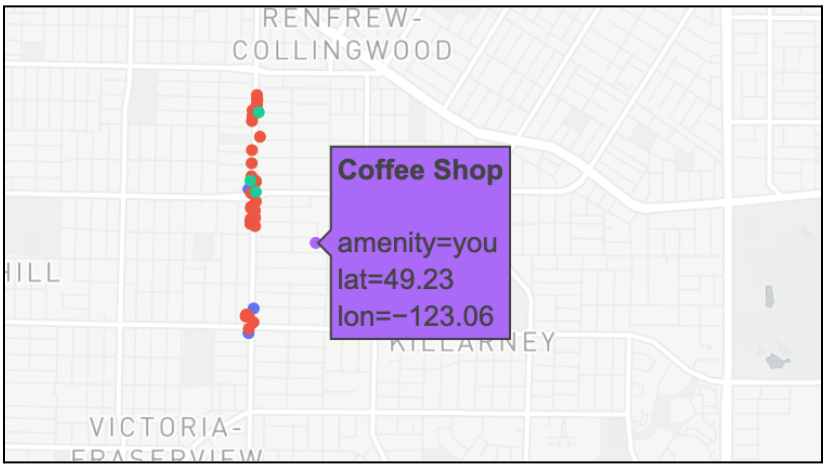
While the map makes use of the basic demographics to inform the connection between a city and cuisine type, it doesn't relate the overall distribution of these demographics. As an addition, we also generated bar plots to showcase the distribution of some of these statistics, such as household income, population, and languages.



Bar plots of relevant demographic distributions.

Finding Competitors

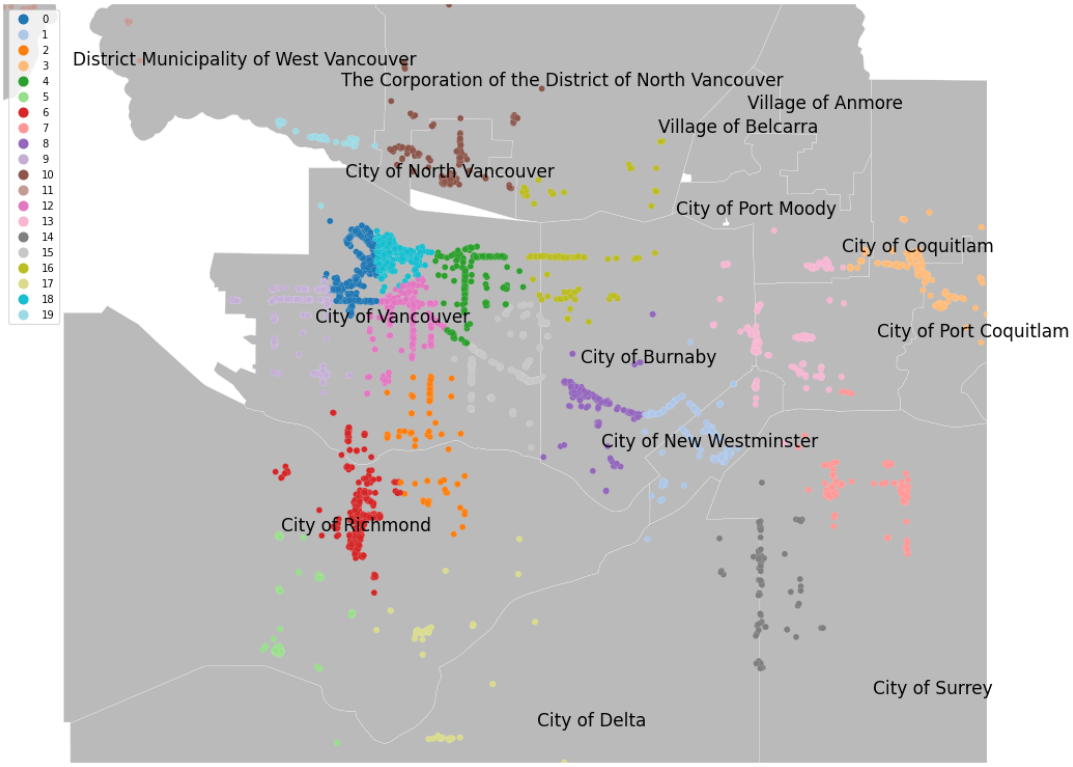
Using OSM data and the users input, we find all food amenities that could be a potential competitor. We then plot all points using plotly express. Given below is a sample result.



Yelp Data and K-means Clustering.

From the filtered Yelp dataset, we plot the coordinates of each restaurant on a map using Geopandas. When viewing the distribution of restaurants, it seems that their locations are sometimes clustered around areas that cross over city boundaries. There may be some interesting analysis that can be derived by going outside of these pre-defined boundaries. Therefore, this is a perfect candidate for K-Means Clustering.

After applying the Elbow method, we discovered that twenty clusters (K=20) yields the best result. The figure below shows twenty distinct clusters which we will individually analyze to determine the most prominent cuisines for each of the clusters. Each color corresponds to a particular group identified by the scikit-learn library.



Cluster number	1'Most Common Cuisine	2'Most Common Cuisine	3'Most Common Cuisine	4'Most Common Cuisine	5'Most Common Cuisine	6'Most Common Cuisine	7'Most Common Cuisine	8'Most Common Cuisine	9'Most Common Cuisine	10'Most Common Cuisine
0	Coffee & Tea	Japanese	Breakfast & Brunch	Pizza	Sushi Bars	Chinese	Fast Food	Bakeries	Sandwiches	Specialty Food
1	Breakfast & Brunch	Coffee & Tea	Chinese	Pizza	Indian	Sushi Bars	Specialty Food	Japanese	Vietnamese	Cafes
2	Chinese	Indian	Coffee & Tea	Sandwiches	Bakeries	Fast Food	Seafood	Burgers	Pizza	American (Traditional)
3	Japanese	Pizza	Sushi Bars	Chinese	Coffee & Tea	Sandwiches	Bakeries	Fast Food	Vietnamese	Burgers
4	Coffee & Tea	Specialty Food	Pizza	Breakfast & Brunch	Bakeries	Sushi Bars	Japanese	Cafes	Chinese	Vietnamese
5	Japanese	Sushi Bars	Pizza	Coffee & Tea	Ice Cream & Frozen Yogurt	Chinese	Sandwiches	Bakeries	Cafes	Fish & Chips
6	Chinese	Coffee & Tea	Japanese	Desserts	Bubble Tea	Sushi Bars	Seafood	Fast Food	Cafes	Bakeries
7	Fast Food	Pizza	Sushi Bars	Coffee & Tea	Mexican	Korean	Vietnamese	Sandwiches	Chinese	Japanese
8	Chinese	Coffee & Tea	Japanese	Pizza	Taiwanese	Ice Cream & Frozen Yogurt	Bakeries	Bubble Tea	Sushi Bars	Cafes
9	Coffee & Tea	Japanese	Sushi Bars	Bakeries	Desserts	Chinese	Specialty Food	Cafes	Breakfast & Brunch	Italian
10	Coffee & Tea	Pizza	Sushi Bars	Japanese	Burgers	Specialty Food	Chinese	Bakeries	Cafes	Fast Food
11	Seafood	Coffee & Tea	Specialty Food	Breakfast & Brunch	Ice Cream & Frozen Yogurt	Bakeries	Barbeque	Ethiopian	Tapas/Small Plates	Beer
12	Coffee & Tea	Chinese	Japanese	Sushi Bars	Pizza	Breakfast & Brunch	Cafes	Bakeries	Specialty Food	Vietnamese
13	Korean	Chinese	Pizza	Japanese	Breweries	Vietnamese	Bakeries	Sushi Bars	Fast Food	Coffee & Tea
14	Indian	Chinese	Sushi Bars	Fast Food	Chicken Wings	Greek	Breakfast & Brunch	Specialty Food	Japanese	Middle Eastern
15	Chinese	Vietnamese	Japanese	Fast Food	Taiwanese	Coffee & Tea	Beer	Bakeries	Breakfast & Brunch	Filipino
16	Japanese	Coffee & Tea	Pizza	Chinese	Sushi Bars	Fast Food	Specialty Food	Bakeries	Cafes	Barbeque
17	Breweries	Vietnamese	Coffee & Tea	Sandwiches	Chinese	Ice Cream & Frozen Yogurt	Wineries	Fast Food	Convenience Stores	French
18	Coffee & Tea	Pizza	Sandwiches	Japanese	Specialty Food	Fast Food	Cafes	Breakfast & Brunch	Seafood	Sushi Bars
19	Coffee & Tea	Breakfast & Brunch	Burgers	Sushi Bars	Vietnamese	Specialty Food	Portuguese	Fast Food	Japanese	Thai

Result of Cluster Analysis

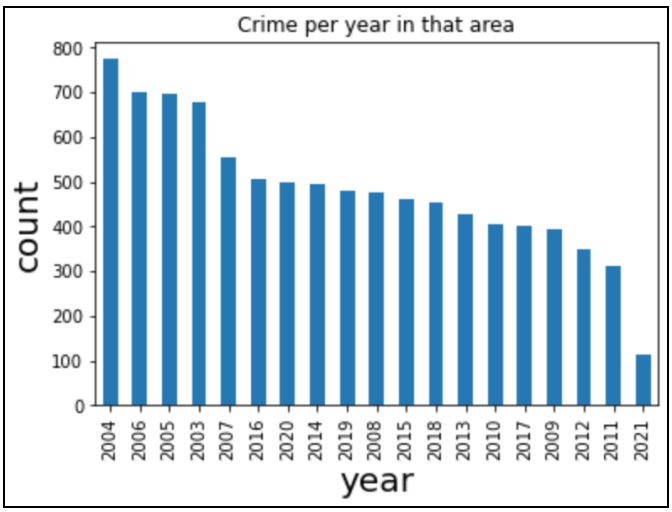
The first step for analyzing the clusters was to count the number of each cuisine type for each cluster. With these counts, we can now calculate the frequencies of each cuisine in that cluster so that we can assign ranks and sort cuisines by popularity. The table below shows the obtained top 10 restaurant cuisines for each identified area.

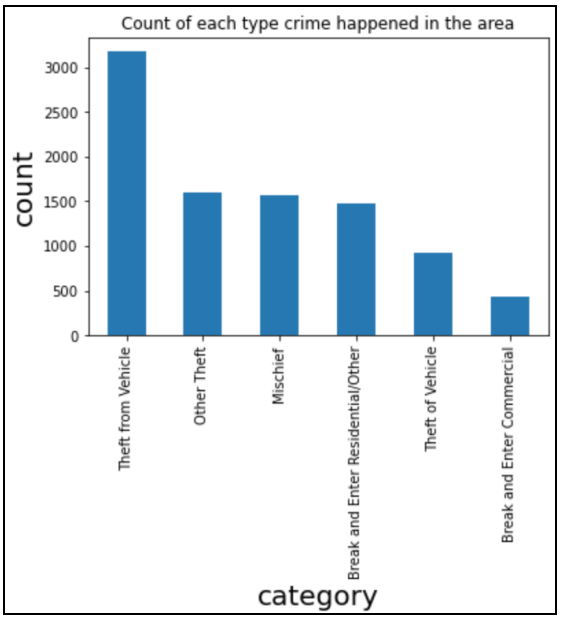
Crime in Vancouver.

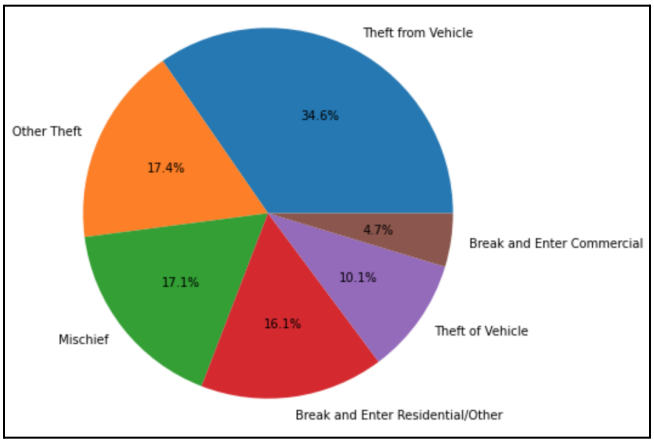
After all the analysis, we used folium to create a heatmap of crime that has happened in the area of that amenity. To make the heatmap visually better, we made 3 heatmaps of the same area which represent the crimes that have occured in the last 3 years. Below is a sample of what our map looks like:

Heatmap of crime around selected amenity in Vancouver (2020)

We constructed a pie chart to find out the distribution of the type of crime. We did an analysis to see if crime has increased or decreased in the neighborhood where our amenity is. We also did an analysis of the number of crimes that happen in that neighborhood per day. Sample outputs are given below.


Crime per year in that area
count
year
800
700
600
500
400
300
200
100
0
2004
2006
2005
2003
2007
2016
2020
2014
2019
2008
2015
2018
2013
2010
2017
2009
2012
2011
2021


Count of each type crime happened in the area
count
category
3000
2500
2000
1500
1000
500
0
Theft from Vehicle
Other Theft
Mischief
Break and Enter Residential/Other
Theft of Vehicle
Break and Enter Commercial


Theft from Vehicle
34.6%
Other Theft
17.4%
4.7%
Break and Enter Commercial
17.1%
10.1%
16.1%
Mischief
Theft of Vehicle
Break and Enter Residential/Other

Limitations.

Our initial premise for the project was to consider amenity density on a neighborhood scale. We were not able to find data that provides these boundaries for the entirety of Greater Vancouver. We had to work with municipalities instead.

While the OSM data was fruitful for restaurant data, the OSM data was not useful for deriving any consistent information on non-restaurant amenities, which could have helped us to develop a better profile of the city in terms of its amenities (e.g., tourism presence). We could have modified the provided scripts for generating the amenities data to instead generate "non-amenity" data, but our initial efforts did not work due to the sheer scale of the raw OSM data.

The iMapBC data was very extensive, and there was great opportunity to use it to form a more in-depth demographic profile of each city. Our use of the demographics data in this final submission was ultimately basic. We wanted to return to it and do more, but we ended up spending more on designing around plotting the maps instead.

The analysis on clusters takes into account only the number of restaurant cuisines as the only measure of popularity within each cluster. The analysis could have been more complete by incorporating Yelp's reviews and performing sentiment analysis to determine if a restaurant of certain cuisine is actually popular in the area. We have extracted Yelp reviews but given the short amount of time to complete the project we were not able to perform an analysis on the reviews.

Given more time to work on this project, we would first focus on combining efforts more effectively. While we drew from each other's results and efforts, we ultimately did not have enough time to combine our findings into a single, coherent result. Conflicting schedules and other commitments made it difficult to consistently work together on this project throughout the semester. Our hope was that if we could have put together our results better, we may have found more opportunities to expand the depth of our analysis.

Project Experience Summary.

Abay's Accomplishments

- Cleaned and prepared Yelp's dataset for the analysis using Spark
- Reduced 6Gb of Yelp reviews data to 180 MB by joining on restaurant business ids using Spark
- Applied K-Means clustering based on latitude and longitude coordinates of the businesses using the scikit-learn library to identify relationships between restaurants' locations and cuisines
- Visualized clusters of restaurants and restaurants' locations on a map using Geopandas and Shape Files
- Computed frequencies of restaurant cuisines for each cluster to construct a table of the top 10 most popular restaurant cuisines within a certain area

Aditya's Accomplishments

- Cleaned and prepared the amenities data for analysis.
- Plotted all amenities on map using plotly_express. All amenities are color coded.
- Found all competitors near the users amenity.
- Cleaned and prepared the Vancouver crime data for analysis.
- Analyzed the crime in the area where the amenity is.
- Generated heatmaps of the crimes near the users amenity using folium.

Kenneth's Accomplishments

- Improve the amount of data on each city by extracting demographics data from iMapBC
- Extract meaningful columns from each dataset and plot the distributions using a bar plot
- Clean and connect the separate data on each city into one DataFrame using Pandas and GeoPandas to reorganize and join
- Use city demographic data with the OSM amenities to count the prominence of cuisines within each city
- Visualize the distribution of cuisines and their relative density within each city using interactive Folium map

- Expand Folium map to be able to plot two cuisines at once for comparison of their distribution as well as their respective densities within each city