The distribution of chain restaurants

Problem Description

In this section we will find out the distribution of chain restaurants. In other words, we want to find out if some parts of Metro Vancouver have more chain restaurants. If the answer is yes, we will find out these regions. These information would help people to plan for their trips.

Data Cleaning

Before we started doing our modeling, we selected the restaurants from the dataset. After going through the dataset, we found that the dataset had three sub-categories “restaurant”, “cafe” and “fast food” under the broader category of restaurant. These categories look a bit weired, but that is the raw dataset. For example, White Spot is a restaurant, Tim Hortons is a cafe while McDonald’s is a fast food restaurant. We included all these three categories for this study. Also, we defined chain restaurants as the those which have top-ten number of branches under each sub-category. Following these strategies, we grouped the data by names and counted the number of branch. We also filtered out restaurants which had less number of branches. In the end, we excluded the restaurants in Abbotsford and Chilliwack because we only wanted to focus on Metro Vancouver in this study.

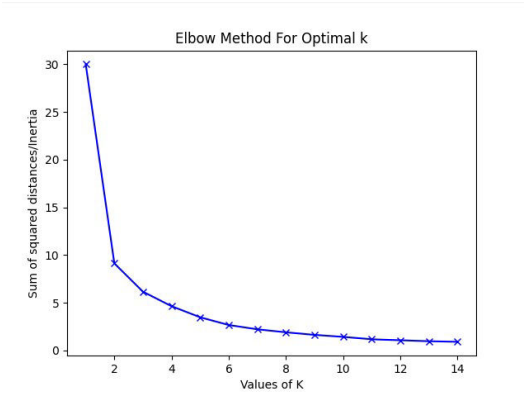
Formulation with Sklearn Machine Learning

We formulated a clustering model and used the K-Mean Sklearn machine learning model to solve this problem:

Having the data selected, we needed to compute the optimal number of clusters in our K-Mean machine learning model. The number of cluster K is a pre-defined number which cannot be too large or too small. We did some researches and decided to use a famous ‘elbow-curve-method’ to determine the number of cluster. “Elbow-curve-method” uses “Intertia” feature of the K-Mean model, and this method chooses the K value whose corresponding Intertia value does not drop much in the next move. The Intertia computes the squared distance of each sample in a cluster to its cluster center and sums them up. The smaller the Inertia value, the more coherent are the different clusters. If N is the sample size, and C is the center of a cluster, the Inertia formular is defined as follow:

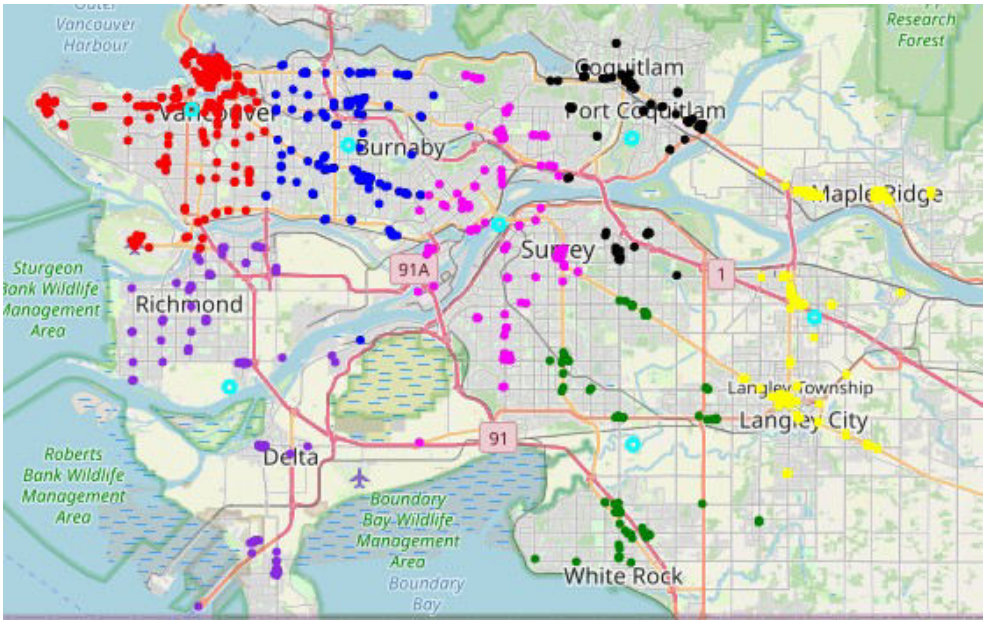
$$\sum_{i=1}^{N}(x_i - C_k)^2$$

We ran our program and plotted with K between 1 to 14. The result is shown as below:



From this plot, we saw that the sum of squared distance did not drop much after K reached 7. We believed K=7 was a good estimation for this study because this value took the trade-off between

precision and utilization of the model. If K is too big, the overfit model will give us too small regions. If K is too big, the model will not tell us much details within each region. Using K=7, we marked these restaurants on the map which is shown as below:



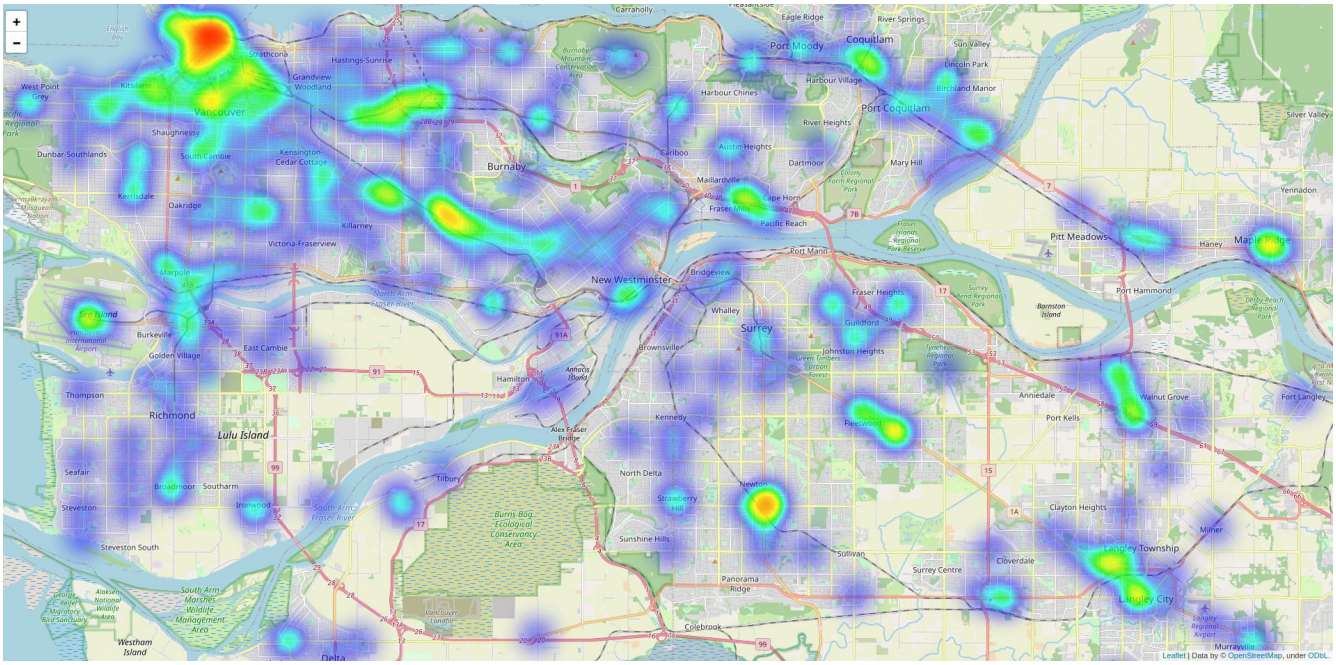
Analysis

Let's have a look at these clusters closer:

Cluster	A	B	C	D	E	F	G
Regions	Vancouver west, Vancouver downtown	Vancouver east, Burnaby west	Richmond, Delta	Burnaby east, New West, north Surrey	Coquitlam, Port Coquilam	South Surrey, White Rock	Maple Ridge, Langley

Notice that the center for each cluster is marked as a light blue circle. These centers are expected to represent the locations where there are lots of chain restaurants nearby. In region A, the center is near Vancouver downtown, which I believe makes sense. In region B, the cluster is near Burnaby Willington street and is in the middle of Metrotown and Brentwood, which I believe makes sense as well. The center of region C falls near south Richmond, where there are not many restaurants near by. For region D and E, the centers are near New Westminster downtown and Coquilam city center respectively, which are reasonable. The center of region F is in south Surrey while that of region G is in north Langley. For the last two centers, we think they are off a bit but not too bad.

We also compared this map with the heat map. The heat map can visualize the restaurant distribution better.



Outliers and Model Adjustments

The reason why the centers of cluster C, F and G are a bit off the town is because of outliers. In region C, the center is “pushed” miles to south because of a few far restaurants in Delta. Similarly, the restaurants in White Rock affects the overall restaurant distribution in Surrey. We thought it would be more reasonable to put Delta and White Rock in two different clusters.

Another factor which we ignored when formulating this problem was the unrealistic distances. All the distances we used were Euclidean distance. In real world, however, people move along the roads. For example, north Surey and New Westminster are very close on the map, but drivers drive much longer because they have to take Pattullo bridge to cross the Fraser River. In this study, we did not have road data, but we can definitely adjust our model by including this factor when road data is available to us.

The use results

The results will help people to plan for their trips. Since some parts of the city has more chain restaurants than others, people can head to these regions directly if they want to find out places for dinners. Also, for tourists who are also food lovers, they can choose hotels within these regions so that they will not need to drive too long to find a popular chain restaurant.