Clustering the City

Data Driven Neighborhood Analysis for Restaurant Location Potential - Vancouver BC Canada

Dino Rossi - November 25, 2020

1. Introduction

Vancouver is a densely populated city in the province of British Columbia on the Pacific coast of Canada. With a population of 675,000 in the city and 2,500,000 in the metropolitan area, it is the largest city in British Columbia and the third largest metropolitan area in Canada.

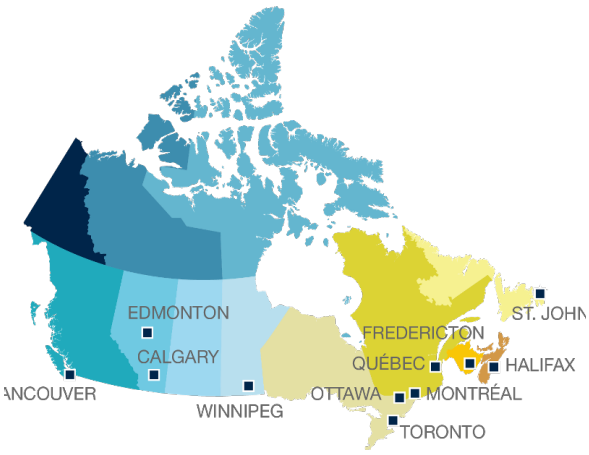


Fig. 1.1 - Map of Canadian Provences

While dense metropolitan areas bring opportunities, they also create constraints. These constraints can lead to stiff competition and high rents for prime business locations. Because of this, choosing the right location for a new business can make the difference between success and failure. The "right" location will mean different things to different people. Some might want to search out a "low" competition neighborhood where there are few restaurants, while others would prefer a "high" competition neighborhood in order to be situated within a bustling scene.

There is no substitute for local knowledge and understanding of a city and its neighborhoods, but there are often larger trends/patterns that are difficult to see. This projects sets out to take advantage of powerful data science tools and techniques in order to gain new insights into the city of Vancouver in order to understand some of those patterns and trends that are not necessarily visible or obvious. These insights will facilitate determining the best location to open a new restaurant by adding layers of information they will be complementary to local knowledge. The results will be usable by anyone looking to open a restaurant in Vancouver, and can be adapted to various use cases.

2. Data Acquisition and Handling

A *dataset* will be built through the use of *web scraping* and pulling venue data through the *Foursquare API*. The data will be *wrangled* into shape using the *Pandas* library. Analysis of the data will be carried out with the *Scikit-learn* library, in particular *K-means clustering* will be used. Finally, the results will be displayed as *maps* of the city, which will be produced using the *Folium* library. These maps can be used to narrow down potential locations for a new restaurant.

2.1 Building the Dataset

In order to build the data set it is necessary to acquire the postal codes for Vancouver. Canada uses an Alphanumeric postal code system. The country is broadly divided into 18 postal regions (see image). These regions are then further subdivided into smaller zones. British Columbia ("V" on the map) has 192 postal codes, but this project will only be looking at the postal codes in and immediately around the city of Vancouver.

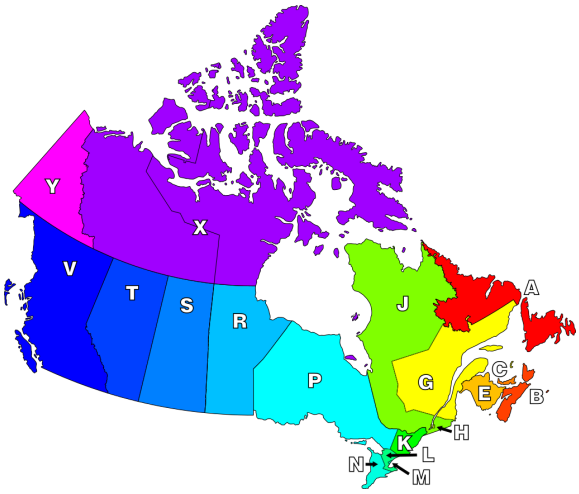


Fig. 2 - Map of Canadian postal regions

Zooming in on Vancouver enables us to select the appropriate postal codes for the areas we wish to analyze. By studying the map below, we can see that the postal codes of interest include the ones starting with V3, V5, V6, and V7.

The rough data for the postal codes can be scraped from this Wikipedia page: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V The data includes all 192 postal codes for the British Columbia region, so the data set will need to be narrowed down to include only the appropriate codes.

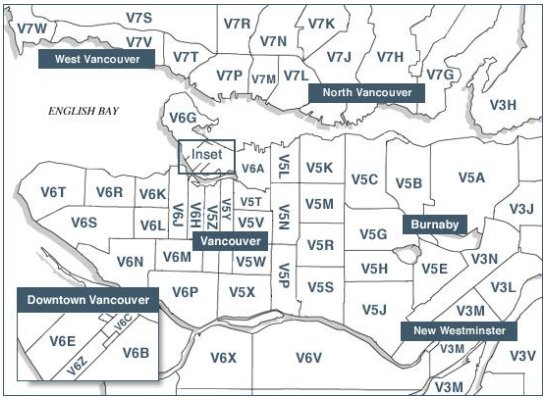
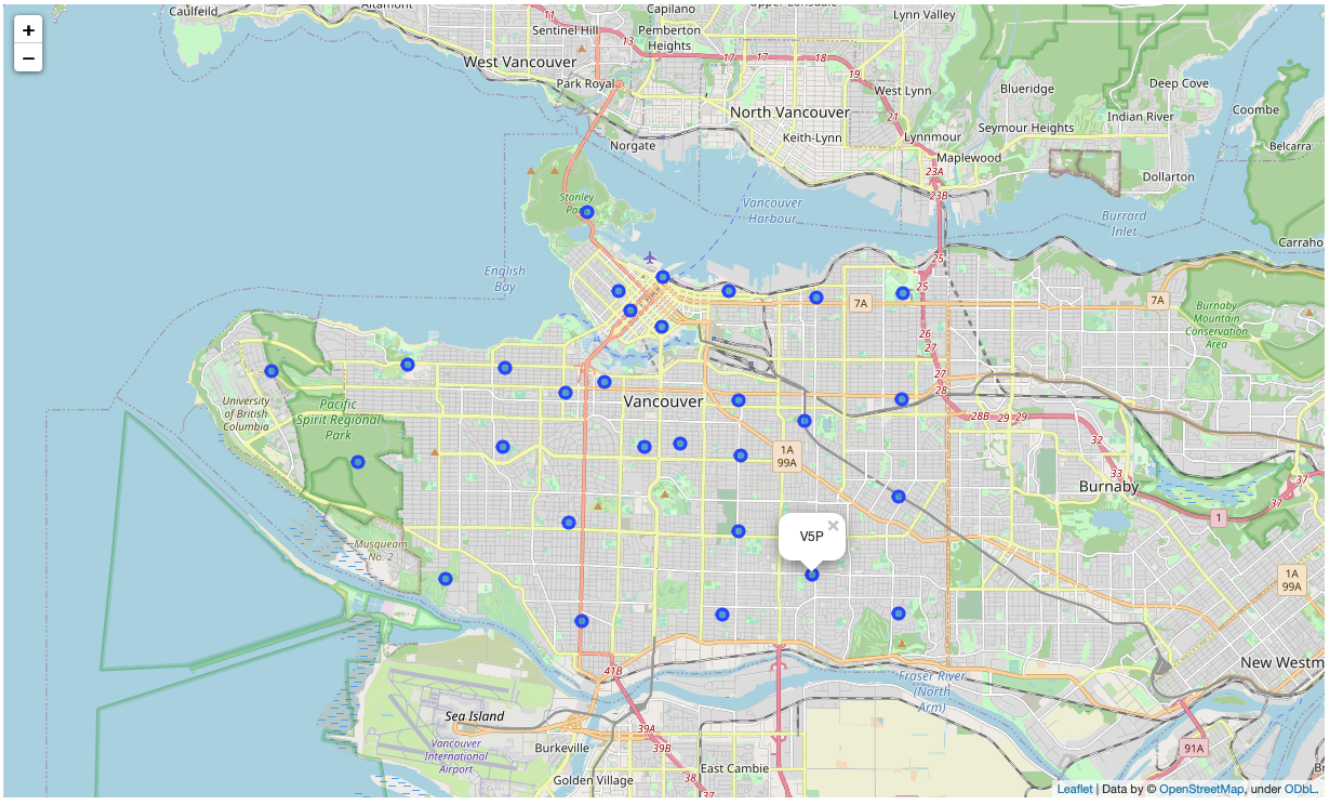


Fig. 3 - Map of Vancouver postal codes

Once a clean data-frame of postal codes and neighborhoods is created, web-scrapping can again be employed, this time to acquire the geolocations (longitude and latitude) for each postal code. The geolocations of the neighborhoods can be used to generate a map with the



Folium library. The map shows the center point of each neighborhood with a blue dot. (See Fig. 2XX). When a dot is clicked it shows a popup label with the name of the neighborhood.

After the geolocations have been added to the data-frame API calls can be made to Foursquare to acquire venue data, which in turn will be appended to the data-frame. When the data-frame contains all the necessary/relevant data analysis/exploration can begin.

3. Methodology

This section shows exploratory data analysis regarding the data gathered on Vancouver, followed by inferential statistic testing perform on the data, and finally, a description of and results from the machine learning used.

3.1 Exploratory Data Analysis

In order to better understand the data complied for this report, some basic information has been extracted and plotted. The following plots visualize various aspects of the city of Vancouver.

As we already know, the total population of Vancouver is around ~630,000. In order to understand the population distribution in more detail we looked at the population of each postal code as well its population density. Figure 3.XX shows the The top 15 postal codes by population. The most populous postal code in Vancouver is V5R.

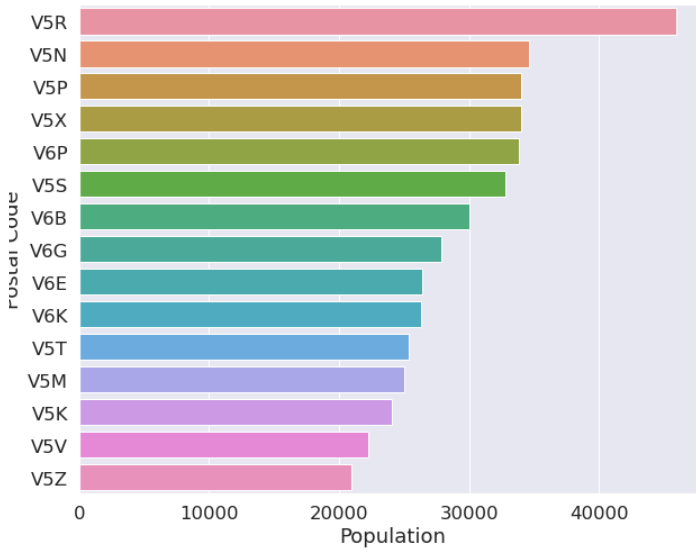


Fig. 3XX) Top 15 postal codes by population.

Figure 3.XX show the top 15 neighborhoods by population density. The most densely populated postal code in Vancouver is V6Z.

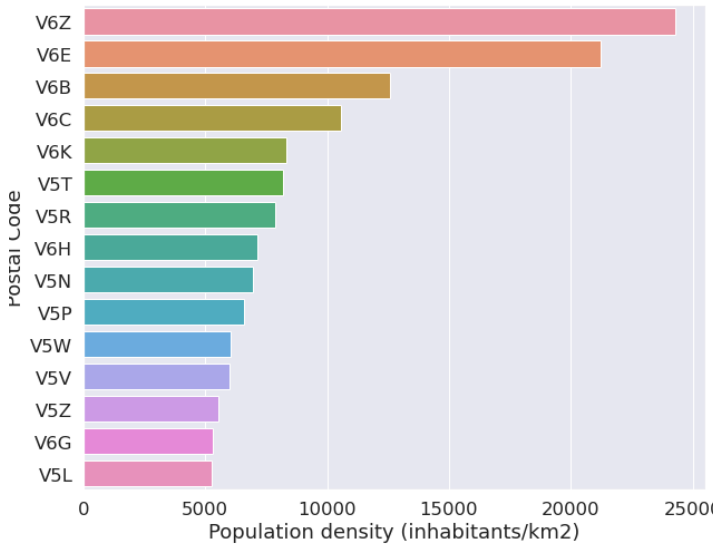


Fig. 3XX) Top 15 postal codes by density.

We can also explore a combination of total population and population density. (See Fig. 3.XX) This gives a good idea of neighborhoods with both high population and high density. The combination of the two would mean that the potential business

has a high population within a short distance. For a business dependent on foot traffic this is an ideal combination. Based on the data we can see that postal codesVB6 and V6E have both a high population and a high density.

Another interesting question is, which neighborhoods have the highest total number of venues? According to the data postal code V6E has the highest number of venues, with 100 venues (See Fig. 3.XX). Following close behind are V6Z and V6C with 99 venues each. After that there is a precipitous drop down to 51 venues in the next postal codes (V6J and V6K).

A higher venue count means high competition, but venue density can also make a neighborhood an attractive destination. A judgement would need to be made wether it is better to be in a "hotspot" location, or to try to corner the market in an area with fewer venues.

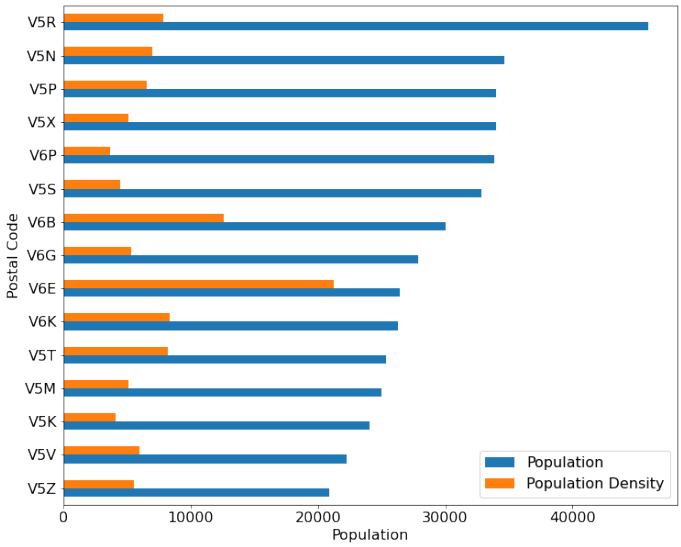


Fig. 3XX) Population and density by postal code.

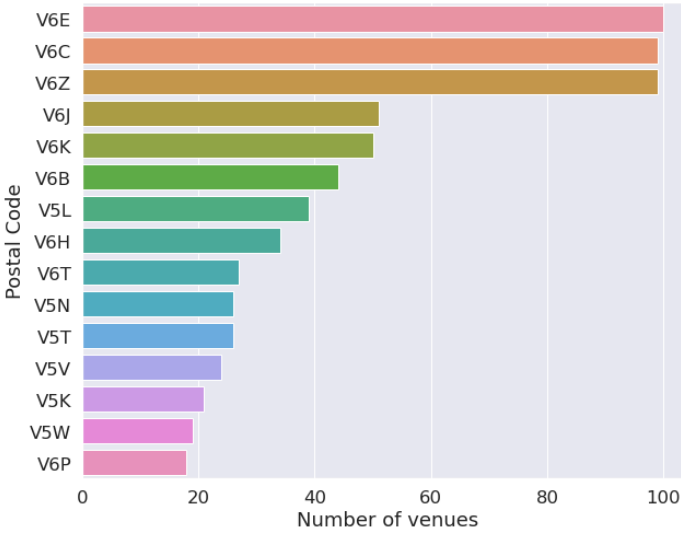


Fig. 3XX) Number of venues per postal code.

The data can also be communicated through maps which hav the advantage of visualizing the information spatially. The following four bubble maps communicate the population of each neighborhood (Fig. 3XX), the density of each neighborhood, the number of venues in each neighborhood (Fig. 3XX), and finally the venues per capita of each neighborhood (Fig. 3XX).

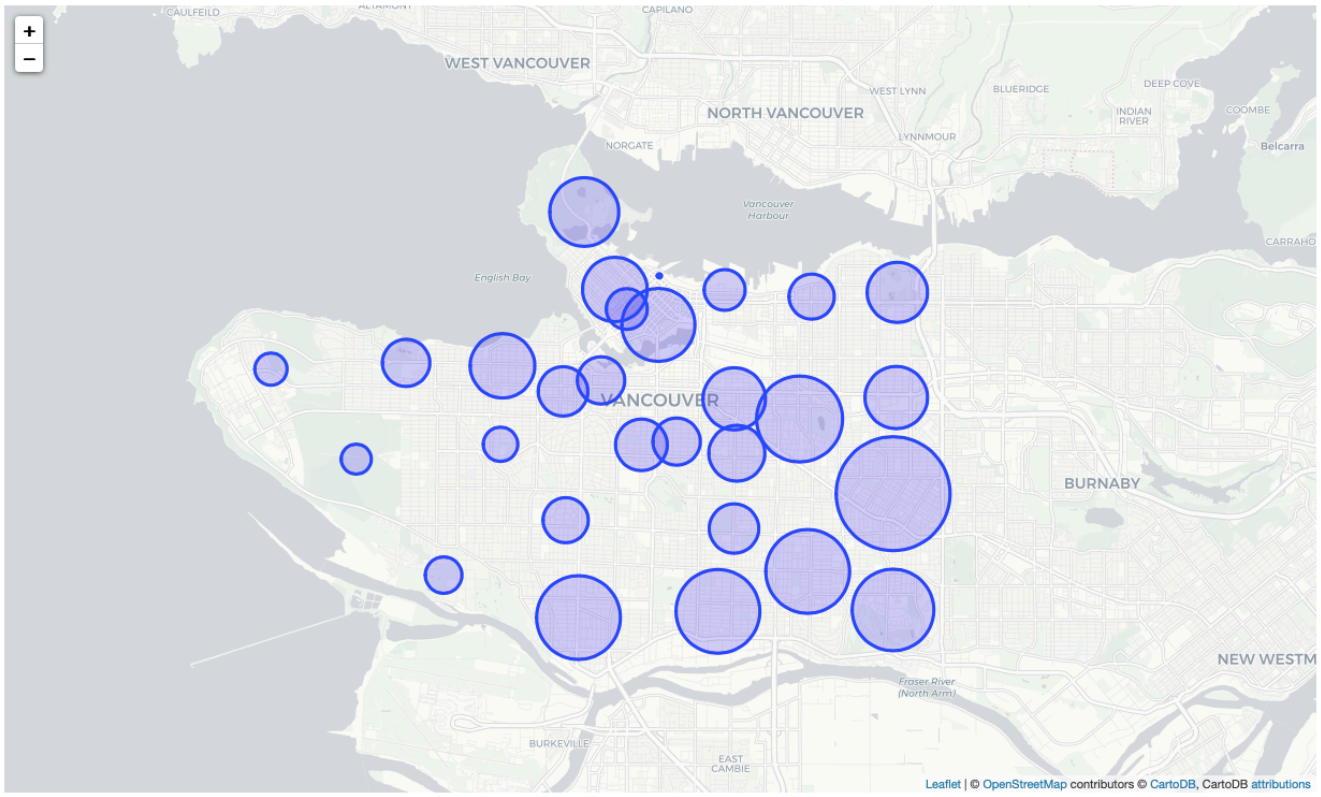


Fig. 3XX) Bubble map of population by postal code.

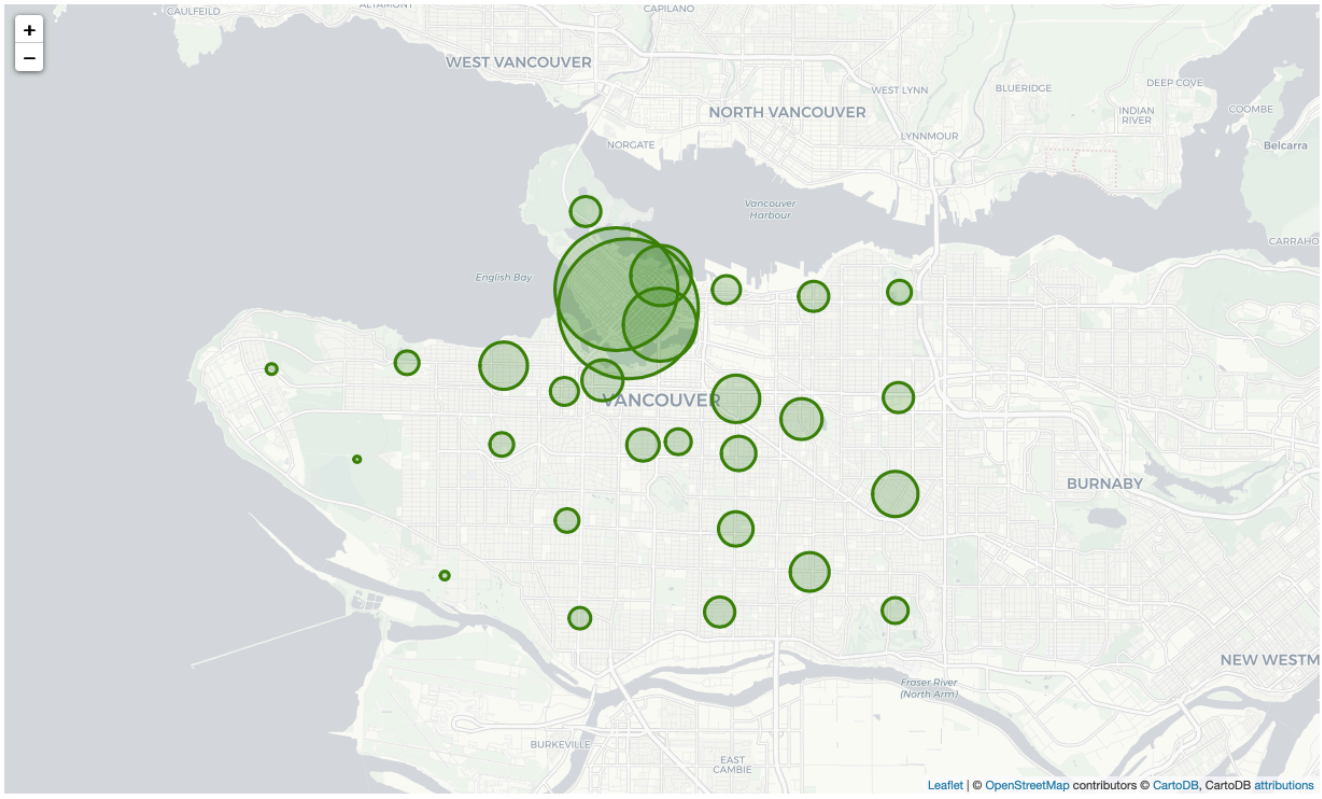


Fig. 3XX) Bubble map of population density by postal code.

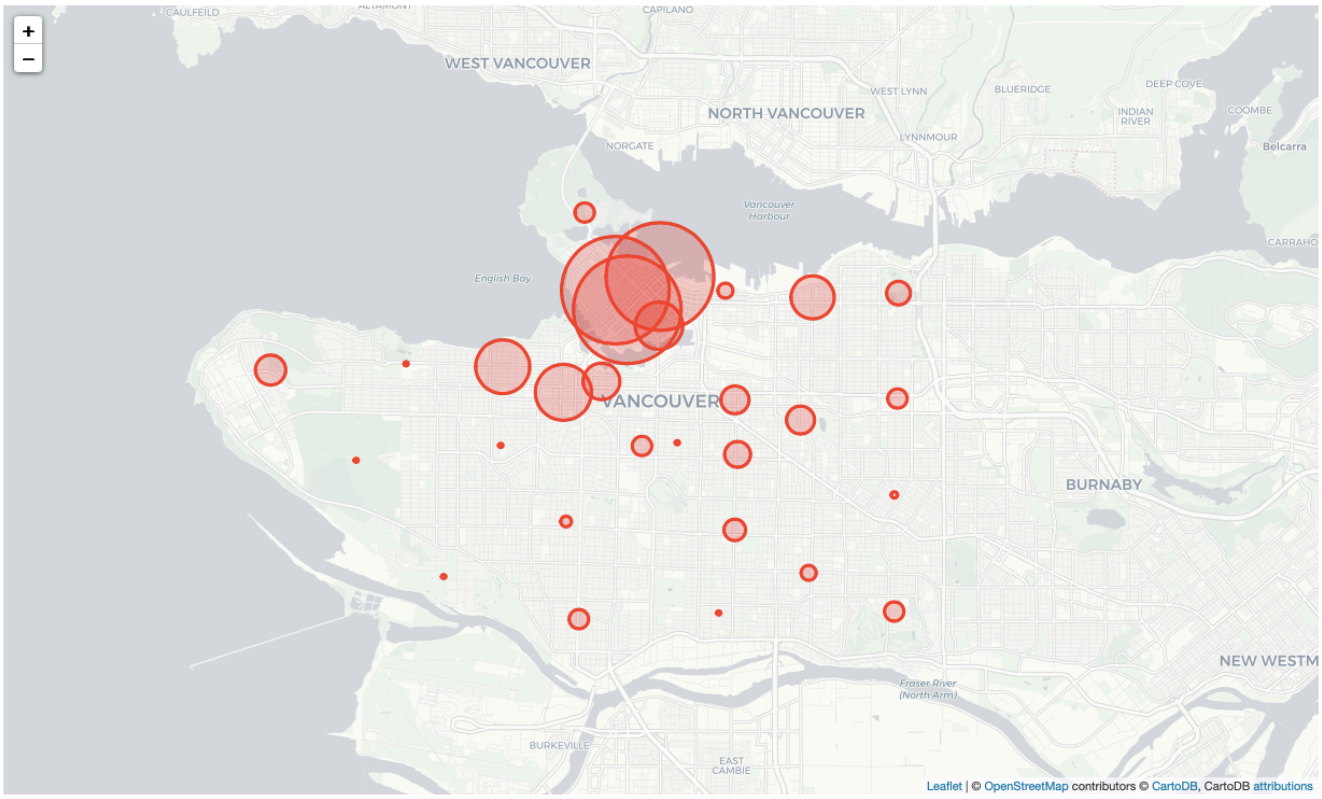


Fig. 3XX) Bubble map of total number of venues by postal code.

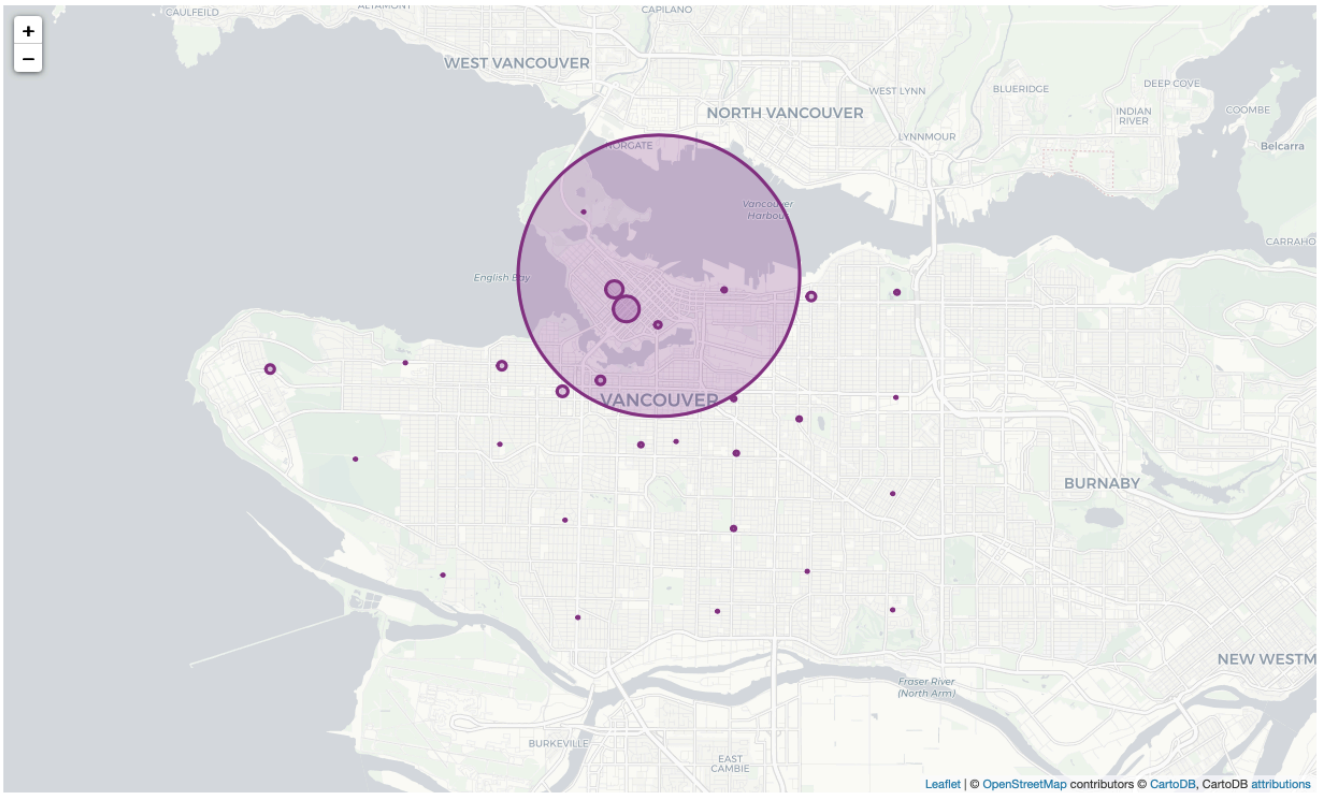
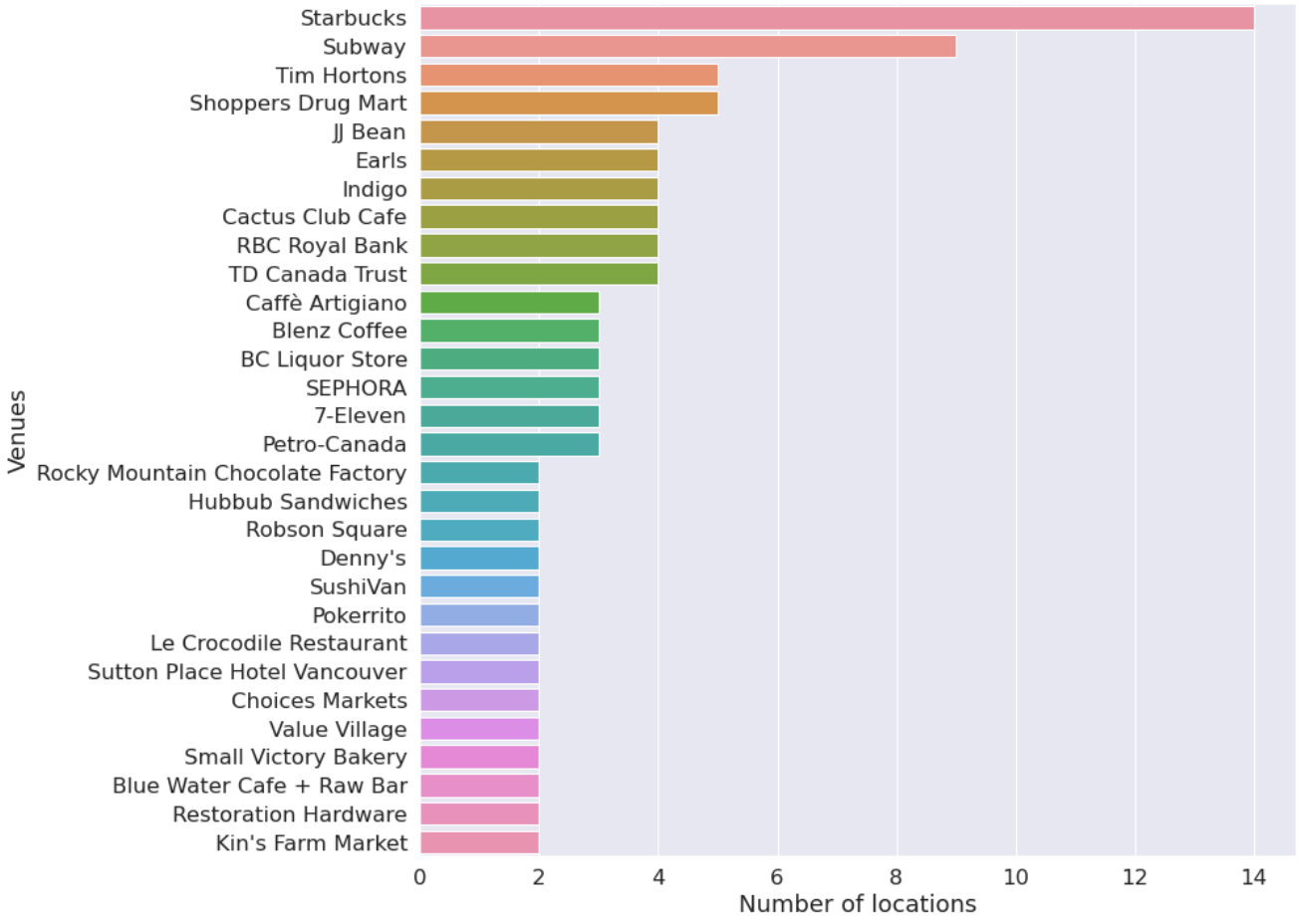


Fig. 3XX) Bubble map of venues per capita.

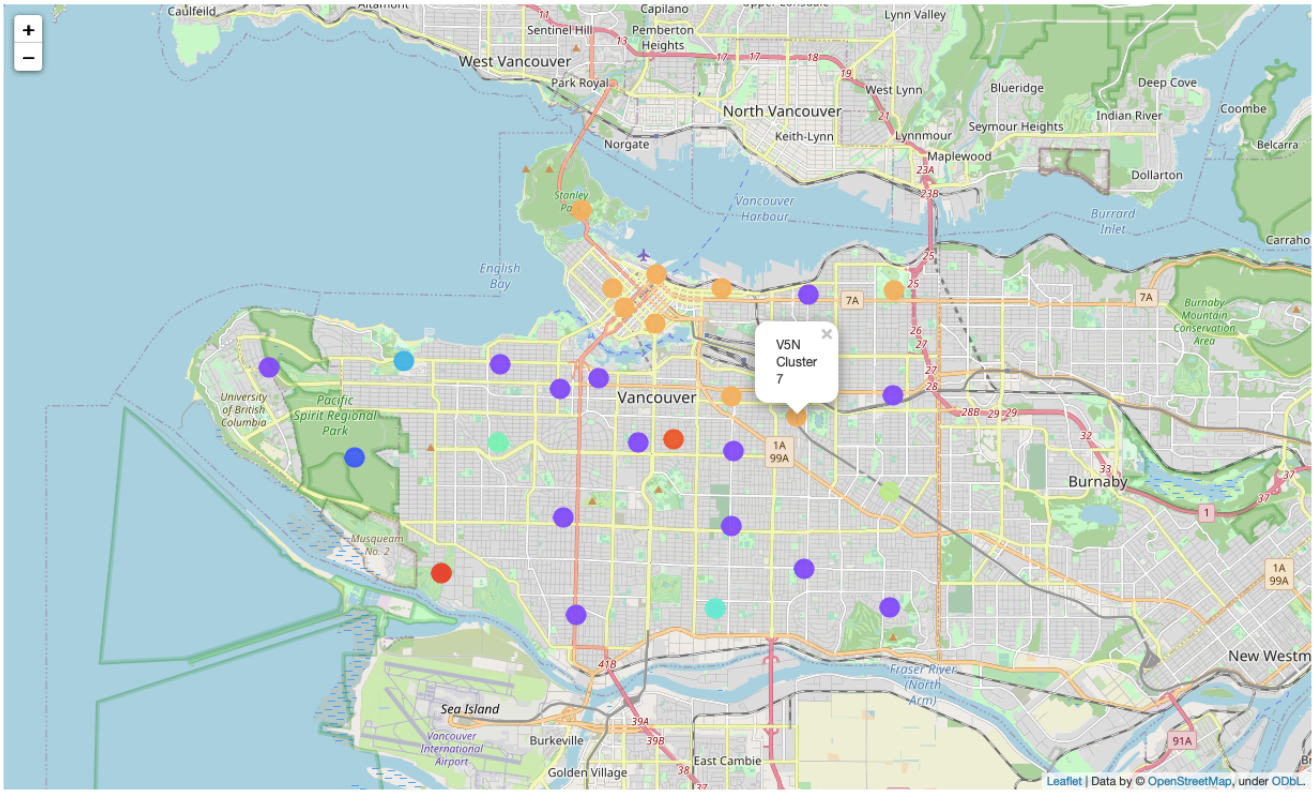
Another key aspect of Vancouver's venues to understand is the frequency with which certain types of venues occur. We can get a good sense of this by plotting the most frequent venues that occur in the data. Fig. 3XX shows the 30 most common venues in the city of Vancouver. Starbucks (a coffee shop) is by far the most frequent venue in Vancouver, and Subway (a sandwich shop) comes in second. Right away it is clear that there is stiff competition for coffee shops and sandwich shops in the city. Tim Hortons, tied for third with Shoppers Drug Mart, could also be considered a coffee shop, al be it with more of an emphasis on donuts. Looking further down the list more coffee shops jump out, Cactus Club Cafe, Caffè Artigiano, and Blenz Coffee to name a few. Based on this quick look the competition for a new coffee shops looks very tough. With out a very specific concept to differentiate a new coffee shop from the competition it looks like it would be ill advised to open one in Vancouver.



3.2 Machine Learning

In order to try to draw more information from the assembled data Machine learning was used. Specifically, the K-means algorithm was employed in order to cluster the various neighborhoods of the city. K-means works through vector quantization in order to minimize intra-cluster distance. In simple terms, the elements within a cluster are more similar to each other than they are to elements in any other cluster. What is interesting is that K-means can produce some unexpected results, in other words hard to interpret from simply looking at the data or using standard data exploration techniques.

The map in Fig. 3XX shows the results of the K-means clustering experiments. In this case the result was a division of the city into seven cluster. The clusters can be used as loose guidelines for exploring other potential locations. For example, if a certain area appears desirable for a new restaurant concept, it could be worth investigating the other postal codes that fall into the same cluster. This approach can help to vastly narrow down the search area, thereby saving valuable time during the location scouting process.



4. Results

This report lays out an analysis of the city of Vancouver in order to facilitate the location selection process for new restaurants. A number of approaches were taken in order to better understand the various areas of the city. This was accomplished by dividing the city based on postal codes. For each postal code a set of data was collected. The collected data for each postal code included: location (geo-coordinates), area, population, population density, and a listing of existing venues. Based on the collected data analysis and visualizations were produced in order to quickly and easily see key characteristics of the various areas of the city.

The goal was not to list a specific address for a new restaurant, but rather to provide a frame work for looking at the city in order to home in on area that fit the criteria of the searcher. Through studying the plots and maps produced in this report one can make some base level decisions wether a particular area would be suitable or not.

If one area becomes particularly interesting, then the clustering results can be used to find other areas with similar characteristics.

5. Discussion

section where you discuss any observations you noted and any recommendations you can make based on the results.

6. Conclusion

section where you conclude the report. section where you discuss any observations you noted and any recommendations you can make based on the results.

7. Image sources

Cover image source:
https://www.telegraph.co.uk/content/dam/Travel/Destinations/North%20America/Canada/Vancouver/vancouver-destination-guide.jpg?imwidth=1400

Fig. 1 source:
https://www.cas-satj.gc.ca/images/canada-map.png

Fig. 2 source:
https://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Canadian_postal_district_map.svg/1024px-Canadian_postal_district_map.svg.png

Fig. 3 source:
https://maps-vancouver.com/img/0/vancouver-postal-code-map.jpg

Fig. 4 source: