Clustering the City

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

Dino Rossi - December 2, 2020

1. Introduction

Vancouver is a densely populated city in the province of British Columbia on the Pacific coast of Canada. (See Fig. 1.1) With a population of ~650,000 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.

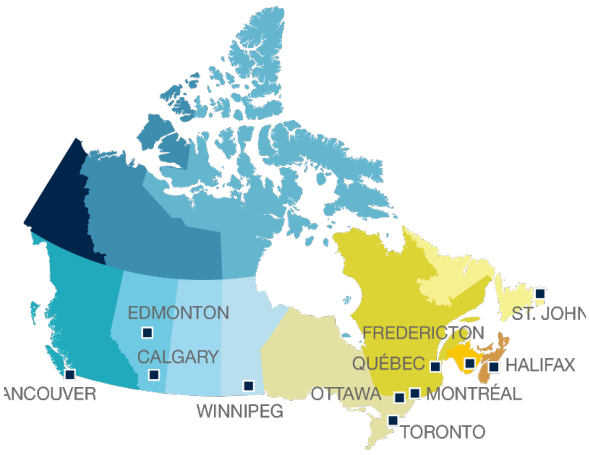

Fig. 1.1 - Map of Canadian Provences

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 at “street level. 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 areas to open a new gastronomy concept by adding layers of information that will be complementary to local knowledge. The results will be usable by anyone looking to open a gastronomy concept in Vancouver, and can be adapted to various use cases. The approach can be easily adapted to other cities and expanded to include other data categories as needed/desired. The aim is to create a “bird’s eye view” of the city in order to broadly understand the various neighborhoods, thereby making it easier to home in on desirable areas.

2. Data Acquisition and Handling

In order to better understand the city of Vancouver a dataset will be built through the use of web scraping and pulling venue data through the Foursquare API. Once the data is collected it will be wrangled into shape and explored using the Pandas library and Seaborn plotting library. Analysis of the data will be carried out with the Scikit-learn library, in particular K-means clustering will be applied. Results will be displayed as plots as well as maps of the city, which will be produced using the Folium library. These maps can be used to quickly narrow down potential locations for a new restaurant from a bird's eye view or 30,000 ft perspective.

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 Fig 2.1.1). 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 within the city limits of Vancouver.

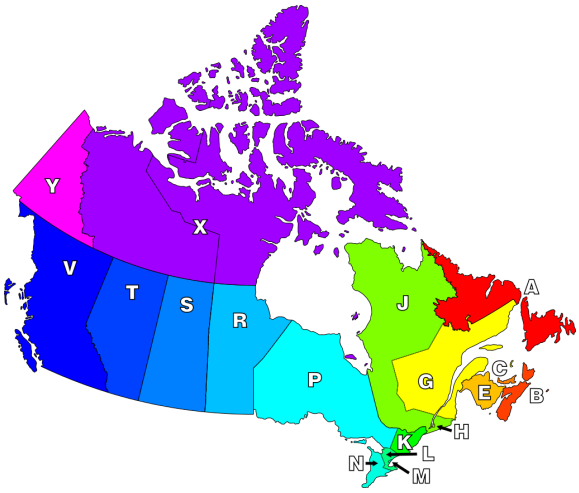

Fig. 2.1.1) Map of Canadian postal regions

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 V5 and V6.

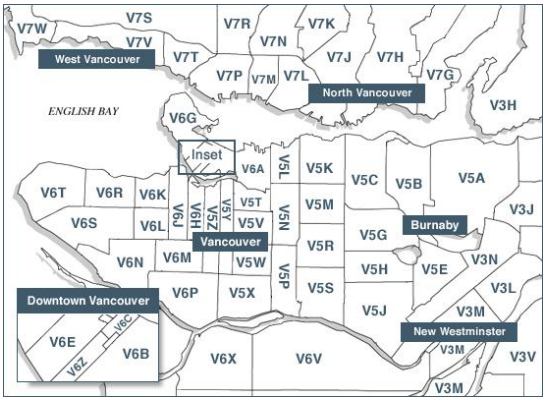

Fig. 2.1.2) Map of Vancouver postal codes

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.

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 postal codes can be used to generate a map with the Folium library. The map shows the center point of each postal code with a blue dot (See Fig. 2.1.3). When a dot is clicked it shows a popup label with the postal code for that 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 categories data analysis/exploration can begin.

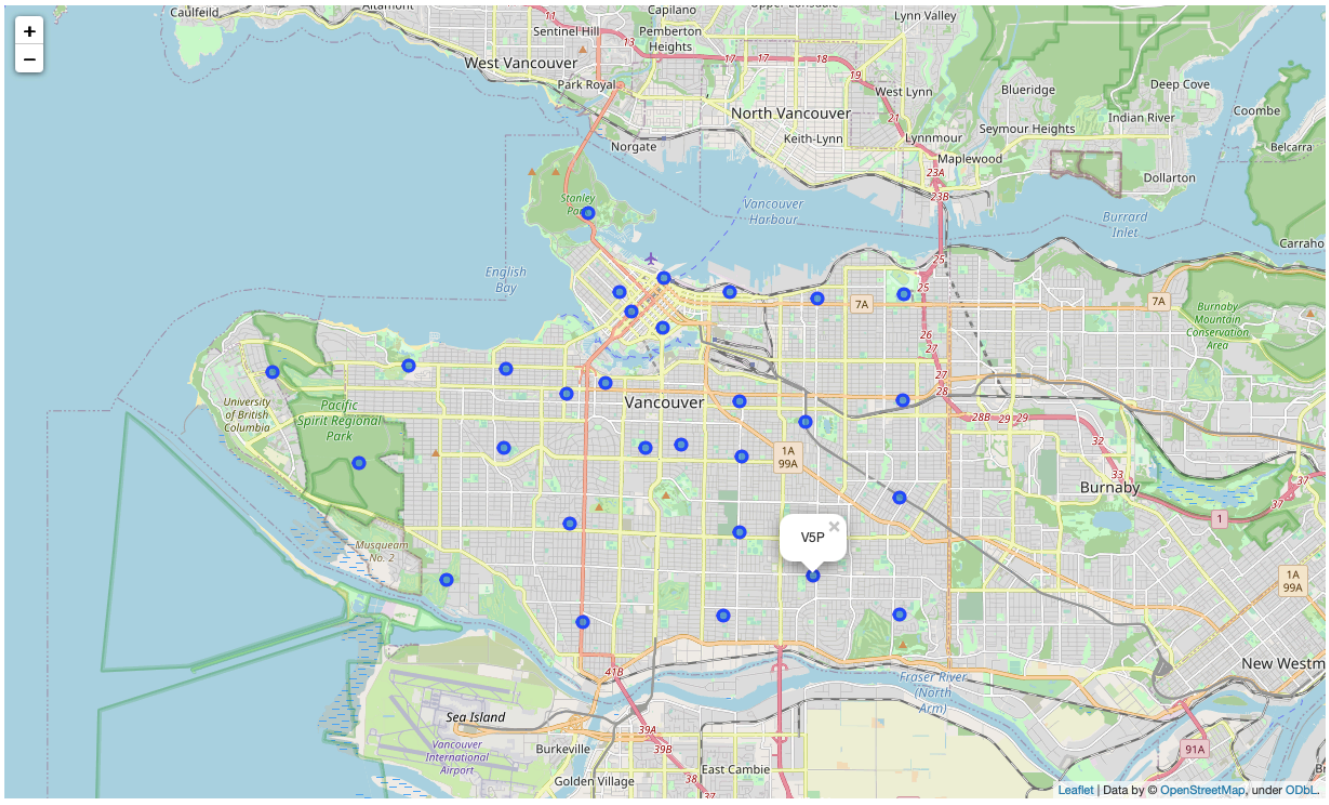

Fig. 2.1.3) Map of Vancouver, each blue marker is the center point of a postal code.

3. Methodology

This section shows exploratory data analysis regarding the data gathered on Vancouver, followed by a description of and results from the machine learning implementation.

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 ~650,000. 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 3.1.1 shows the the top 15 postal codes by population. The most populous postal code in Vancouver is V5R.

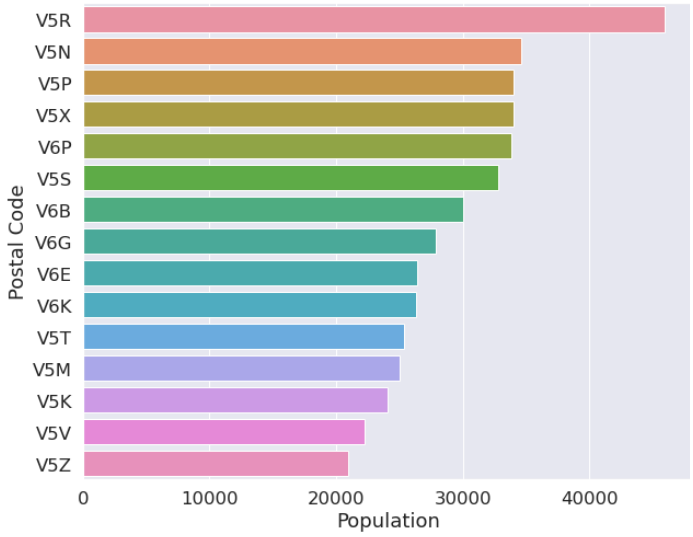

Fig. 3.1.1) Top 15 postal codes by population.

Figure 3.1.2 show the top 15 neighborhoods by population density. The most densely populated postal code in Vancouver is V6Z.

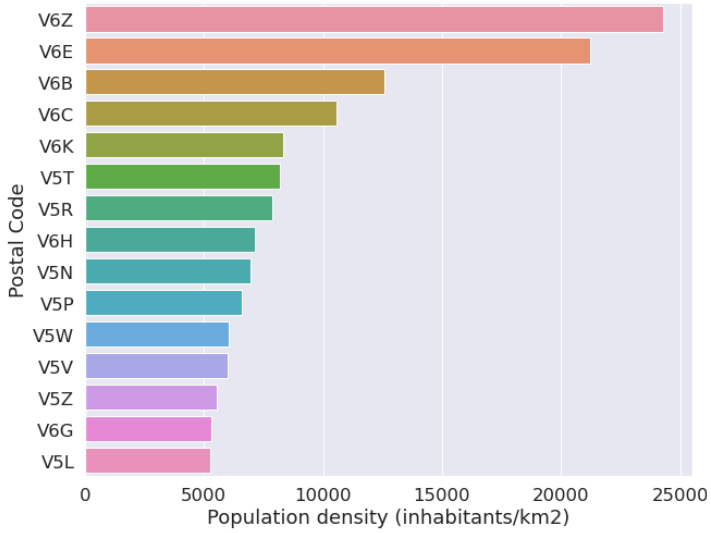

Fig. 3.1.2) Top 15 postal codes by density.

We can also explore a combination of total population and population density (See Fig. 3.1.3). 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 codes V6B 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. 3.1.4). Following close behind are V6C and V6Z 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.

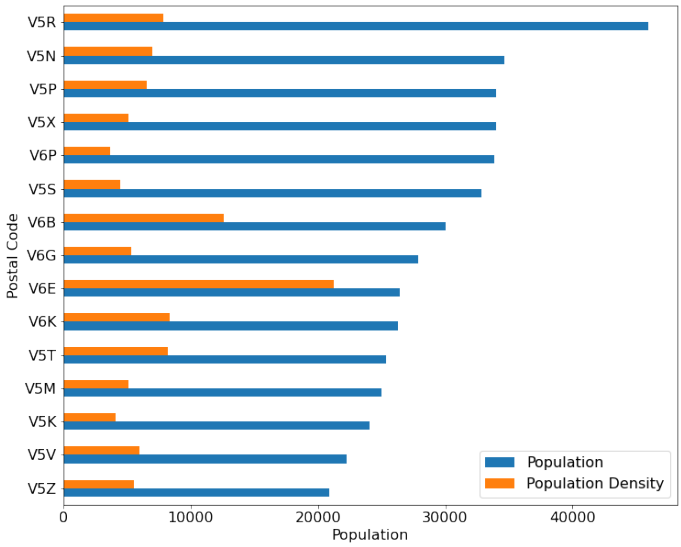

Fig. 3.1.3) Population and density by postal code.

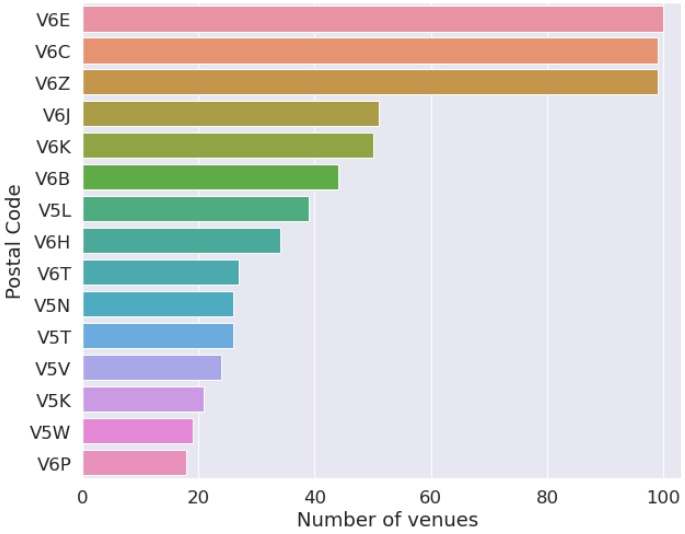

Fig. 3.1.4) Number of venues per postal code.

The data can also be communicated through maps which have the advantage of visualizing the information spatially. The following four bubble maps communicate the population of each neighborhood (Fig. 3.1.5), the density of each neighborhood (Fig. 3.1.6), the number of venues in each neighborhood (Fig. 3.1.7), and finally the venues per capita of each neighborhood (Fig. 3.1.8).

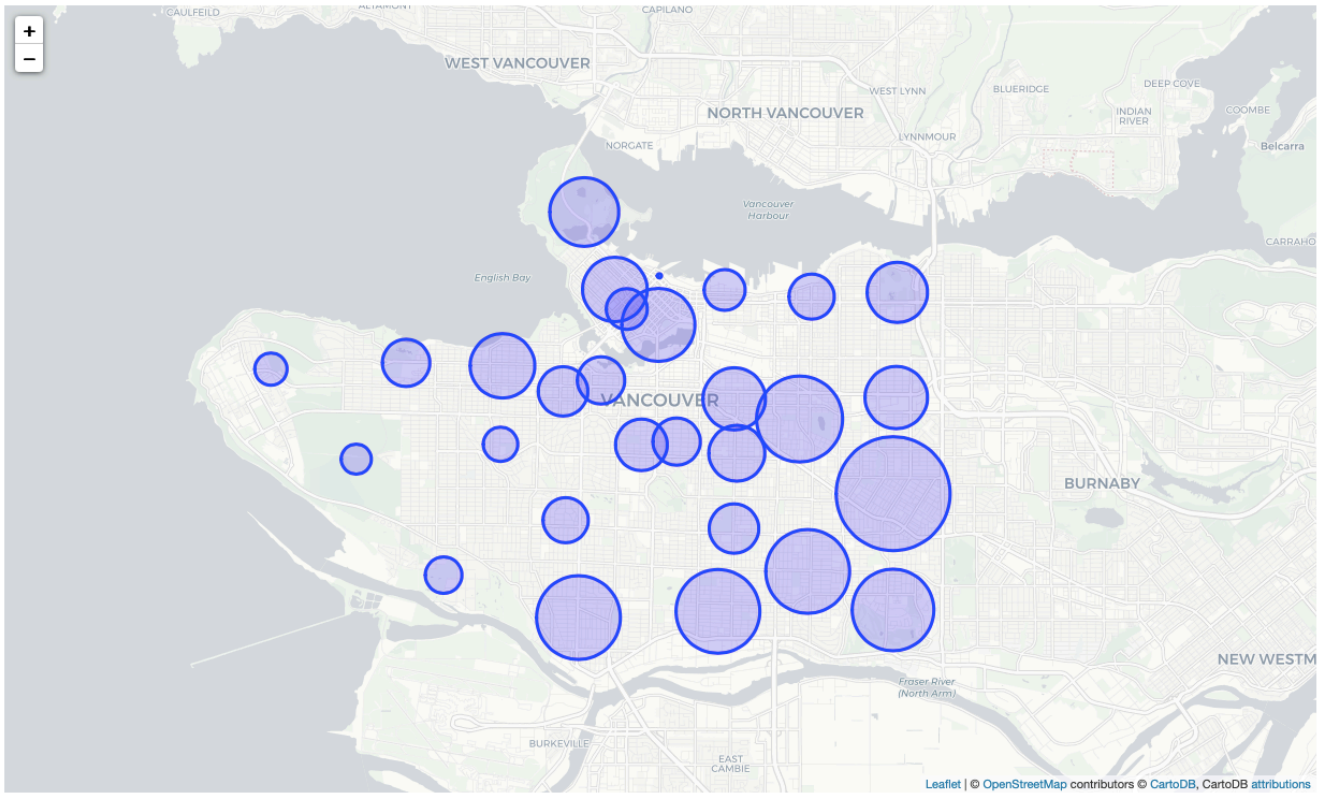

Fig. 3.1.5) Bubble map of population by postal code.

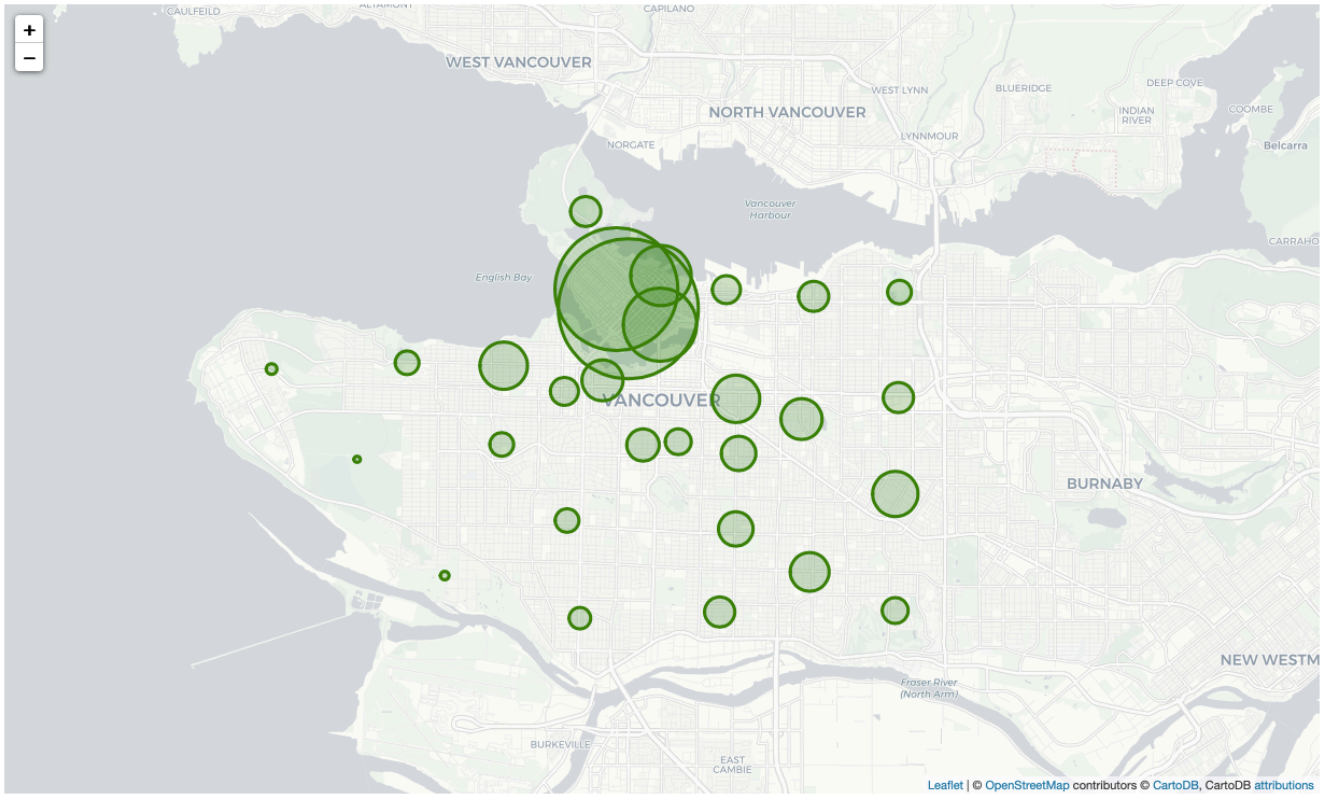

Fig. 3.1.6) Bubble map of population density by postal code.

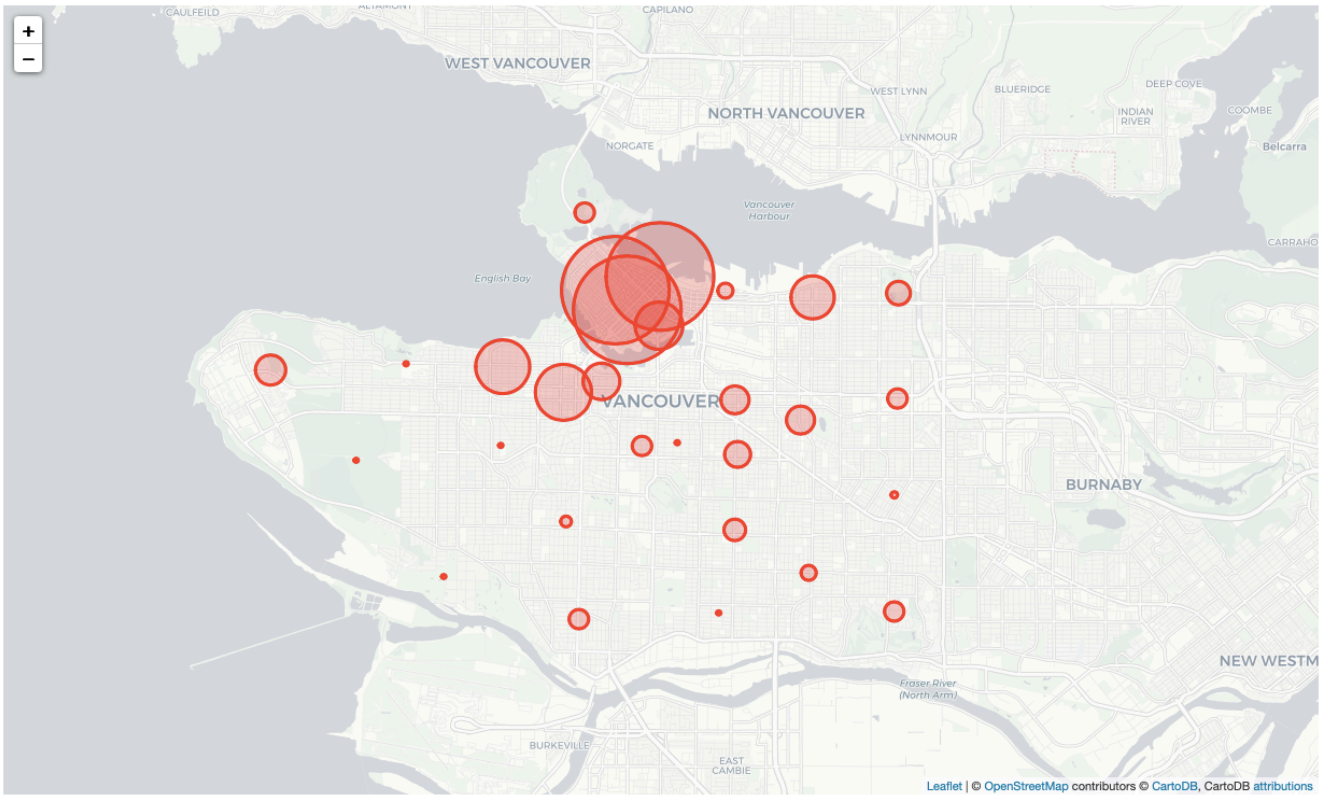

Fig. 3.1.7) Bubble map of total number of venues by postal code.

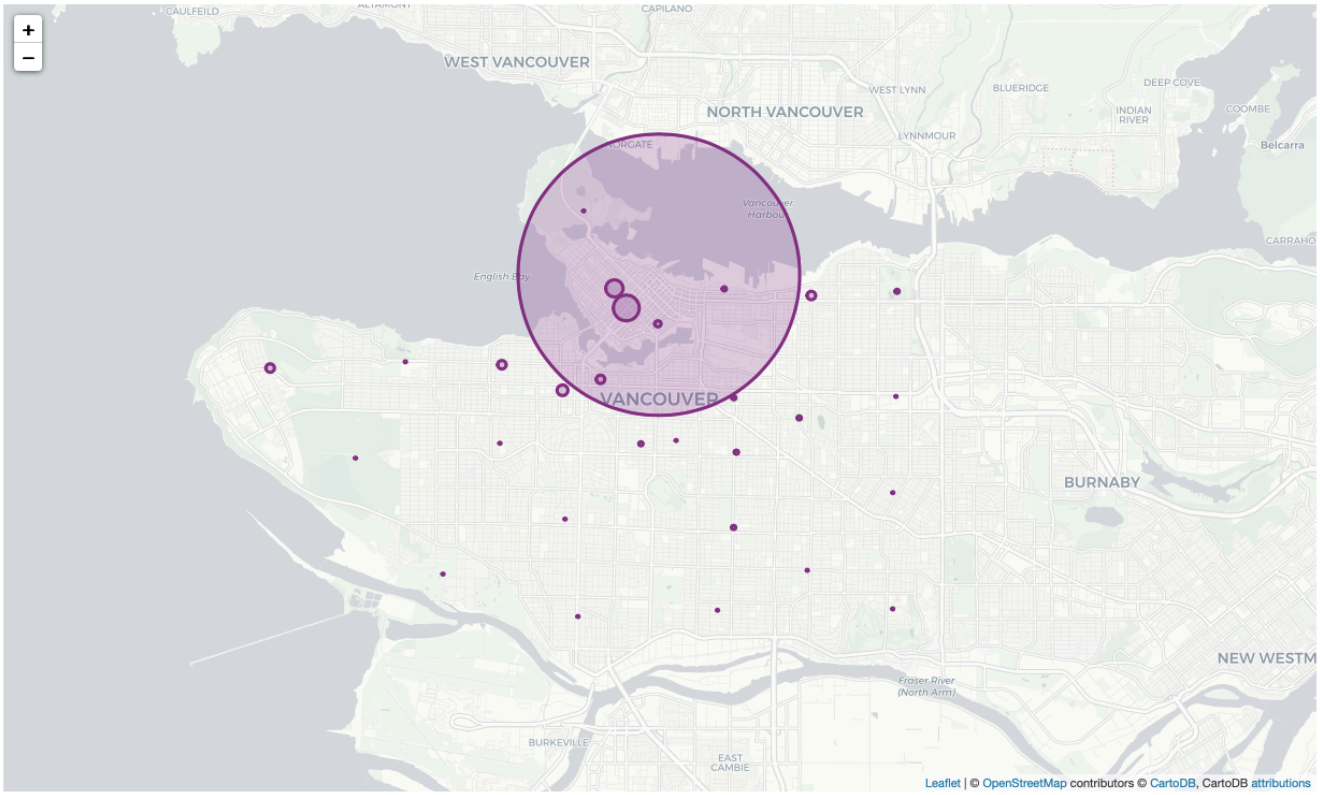

Fig. 3.1.8) Bubble map of venues per capita.

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. Figure 3.1.9 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, albeit 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.

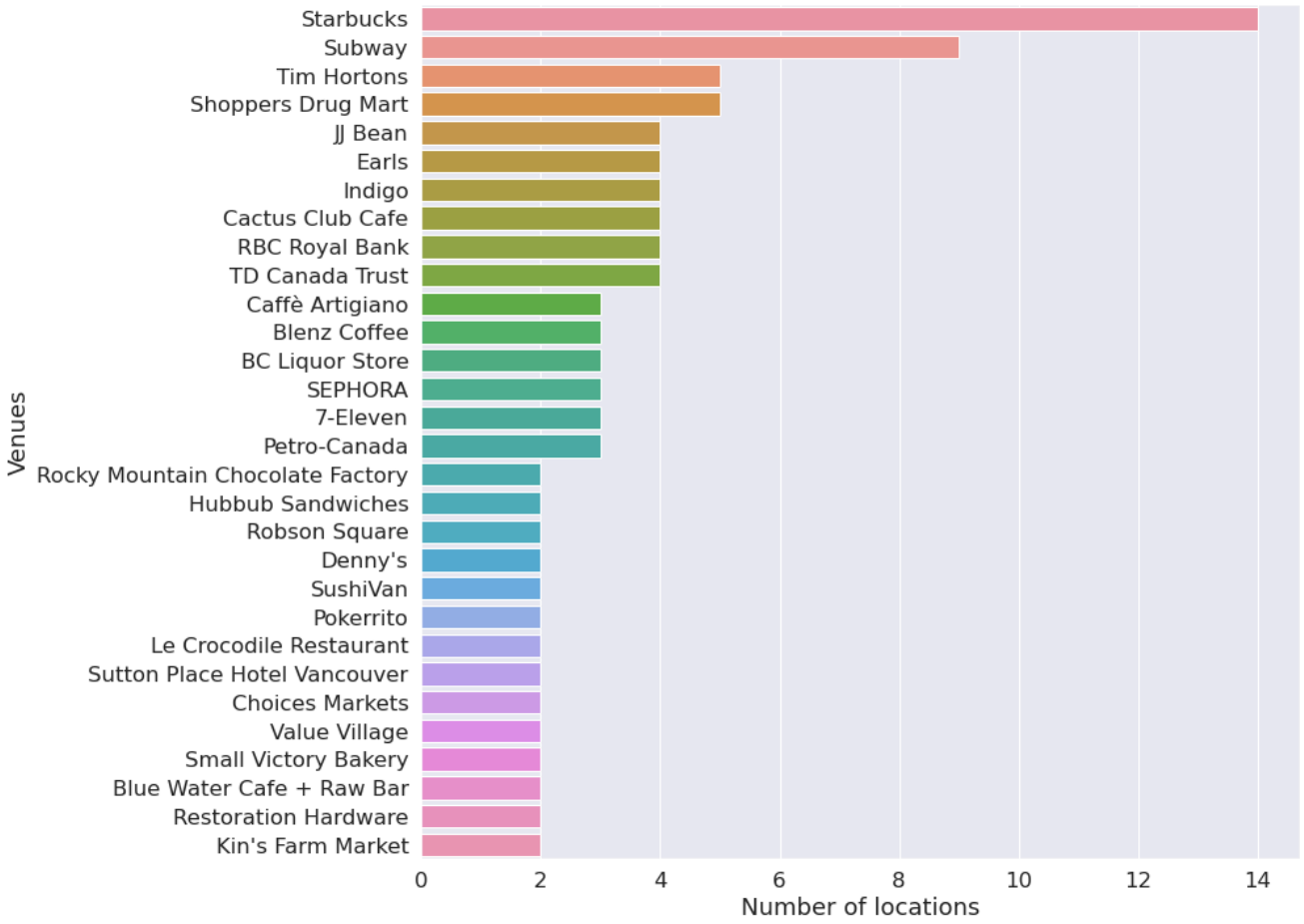

Fig. 3.1.9) The top 30 most frequent venues in Vancouver.

3.2 Machine Learning

In order to try to draw more information from the assembled data set 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. 3.2.1 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 gastronomy 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.

Postal code V5R has been highlighted on the map. It is in its own cluster. Looking back at the plots we can see that V5R has a high population and medium density, but very low venue count.

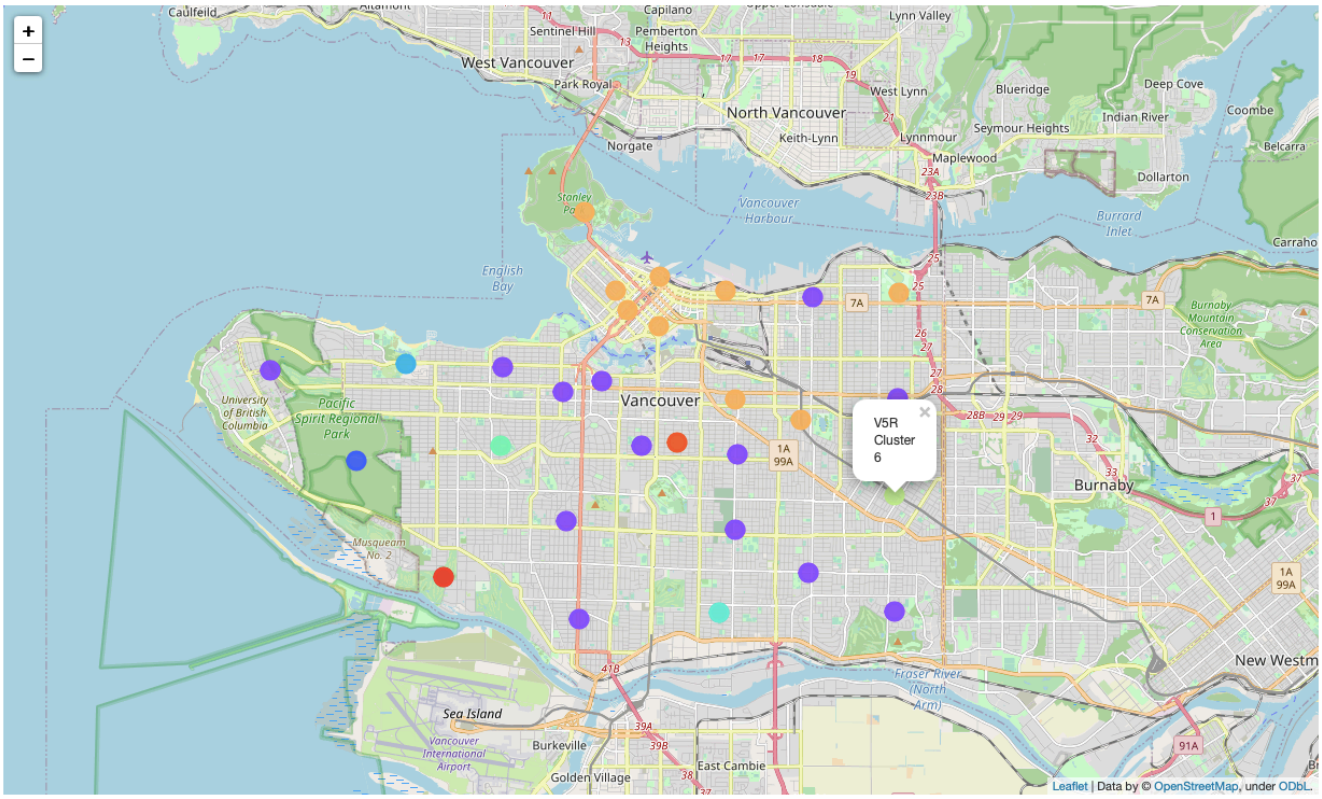

Fig. 3.2.1) K-means clustering of Vancouver based on postal codes / venues.

4. Results

This report lays out an analysis of the city of Vancouver in order to facilitate location selection process for new gastronomy concepts. 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 (km2), population, population density, and a listing of existing venues. Visualizations were produced based on the collected data in order to easily understand 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 as a whole in order to home in on areas 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 for their concept.

If one area becomes particularly interesting, then the clustering results can be used to find other areas with similar characteristics. This approach can be used to quickly narrow down the search area.

5. Discussion

As is to be expected, the highest density postal codes in Vancouver are near the downtown are, the core of these are V6B, V6E, and V6Z. The downtown area also has the highest concentration of venues. These factors make the postal codes in and around the downtown attractive in terms of high volume of foot traffic, proximity of clientele, and the fact that high-density/high-venue areas tend to draw visitors in from other areas. The down side is that these areas have more competition (a lot of venues), and typically have higher rents. To offset the high rents, and take advantage of high foot traffic, restaurant concepts with low overhead and small small space requirements would be recommended for these areas. Potential examples include: lunch oriented take-away shops catering to mid-day service/evening delivery, or cocktail bars focused on afterwork clientele and “bridge and tunnel” business. Coffee shops and sandwich shops should be avoided unless a very well well thought through concept has been produced in order to distinguish the new business from the high level of pre-existing competition, as shown in Figure 3.1.9.

Looking a bit farther out from the downtown area, there are some postal codes with high populations but very low venue counts. The clearest example is postal code V5R,

highlighted in Figure 3.2.1. It has a high population, medium density and very low venue count. This may be why the K-means algorithm separated V5R into its own cluster.

These areas, while having less foot traffic, could offer lower competition and lower rents, making them attractive for gastronomy concepts with a neighborhood focus and higher space requirements.

potential examples include: full service restaurants which may include an integrated produce market and/or butcher, or a beer garden/brewpub which needs outdoor space and/or brewery space.

Given more time, the data set could be expanded to include age and income categories to further narrow down the "character" of each area. Another possibility would be to include public transportation access/proximity. The categories included could be adapted from client to client and city to city in order to best meet the needs of the project.

6. Conclusion

This report provides a bird's eye view of the city of Vancouver. This was achieved through the collection, organization and analysis of data relating to population, area, density, and venue counts of the postal codes within the city limits. The data was communicated through plots and maps, and can be used by any person or party trying to narrow down areas of the city for a new gastronomy concept. The framework established here can be expanded with other data categories such as age/income/public transportation access/etc. to gain a more fine grained view. The framework is also easily expandable to other cities, especially within Canada.

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

Fig. 2.1.1 source:
https://upload.wikimedia.org/wikipedia/commons/thumb/e/e4/Canadian_postal_district_map.svg/1024px-Canadian_postal_district_map.svg.png

Fig. 2.1.2 source:
https://maps-vancouver.com/img/0/vancouver-postal-code-map.jpg

8. Source Code

An rendered Jupyter Notebook of the source code used in this report is available for viewing at: https://nbviewer.jupyter.org/github/dprossi/Capstone_Project_IBM_Data_Science/blob/master/Applied_Data_Science_Capstone_Project_Vancouver.ipynb