

Coursera IBM Professional Certificate
Applied Data Science Capstone

Using Data Science to determine the best location to live in Vancouver, Canada

Author: Alina Prendes Roque
August 2021

1. Introduction

It is a well-known fact that deciding the best location to live in can be exciting and at the same time extremely stressful, mostly for people who are moving to a new country or neighborhood. The purpose of this project is to use a data science methodology and tools to allow people to make data-driven decisions on selecting the best location.

Given the importance of selecting the best location, a thorough research is required in order to maximize the chances of being successful. For this, this project focuses on Vancouver, Canada, by comparing the different neighborhoods in elements such as supermarkets availability, different types of restaurants or stores and in general any criteria or category that could be relevant for the user.

1.1. Business problem

This project provides information that supports the decision-making process of people who are deciding the best neighborhood for living in Vancouver, Canada. For this, an analysis is carried out through the use of Exploratory Data Analysis and Machine Learning techniques, being able to cluster the different neighborhoods of Vancouver according to different criteria. The following question is to be answered: if a person is planning to move to Vancouver, Canada, which would be the best location to rent or buy accommodation?

It is important to note that the criteria for deciding whether a neighborhood is convenient or not are highly individually determined, as for one person it could be important to have a supermarket nearby, whether for other it could be something secondary and the most important thing is to be near a gym. Therefore, this project provides a sample solution based on personal interests, which can be then adapted to personal preferences.

1.2. Target audience for this project

The main audience are people who are considering to live in Vancouver, Canada and would like to have information to better decide where to buy or rent an accommodation. There are three main target groups or segments:

- People who live in other countries and are moving to Canada, specifically to Vancouver (either temporarily or permanently)
- People who live in some other city in Canada, but are thinking about moving to Vancouver
- People who live in Vancouver, but are planning to move and would like to get as information as possible to make the best decision

Furthermore, the information provided by this project can also be used by people who want to get information best location in Canada, such as tourists who are going for a few days and want to make the best out of it by staying close to the places that are most relevant for them. Last but not least, this project could also be useful for people who are going to carry out a similar analysis in another city and would like to adapt the created code for similar analysis in other locations.

2. Data

In order to solve the previously mentioned business problem, the following data is required:

- A comprehensive list of Vancouver's neighborhoods with the respective postal code, latitude and longitude
- Venues information for each neighborhood, which serves as criteria for further comparative analysis

First of all, a comprehensive list of Vancouver's neighborhoods with boroughs and postal codes is obtained through the following link: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V. In order to extract the data from the previously mentioned website, several libraries are used. For retrieving data from the website, BeautifulSoup, Requests and Pandas libraries are used, as they allow to obtain information of each neighborhood and to store it in a structured way in a Python data frame. After that, the Geocoder library package is used to obtain the latitude and longitude of each neighborhood (this requires the postal code to be provided as an input).

After that, the venues information is obtained by using Foursquare API. Foursquare is a location data provider, which offers the following information about venues within a specified distance of

the latitude and longitude of the neighborhoods: neighborhood, latitude, longitude, venue, venue name, venue latitude, venue longitude and venue category.


This project encompasses diverse data science tools, from web scraping to data cleaning, data wrangling, machine learning (K-means clustering) and data visualization. The methodology section below explains more into detail the steps followed to carry out the project, as well as the results obtained for each one of the analyses.

3. Methodology

The applied methodology consists of the following five steps:

- 1) Analytic approach: this requires first of all to clarify the question. After the question is fully understood, it will lead to determine the most suitable approach. In this case it would be possible to use an unsupervised machine learning algorithm that allows to characterized the neighborhoods and cluster them according to some specified criteria.
- 2) Data requirements: it is necessary to determine which information is required, in which format and where it can be obtained. As mentioned before, in this case it would be necessary to have a comprehensive list of Vancouver's neighborhoods with the respective postal code, latitude and longitude, as well as venues information.
- 3) Data collection: this steps has be to be carefully carried out, as the information is to be found online in different sources. Different tools are used, ranging from web scraping to using different libraries and making Foursquare API calls. This step requires constant review and data cleaning. For instance, in case a postal code is not assigned to any borough, it needs to be deleted from the created data frame. Furthermore, it is necessary to constantly check the format of the retrieved information to make sure that it meets the requirements, being understandable and easy to read.

The first data frame is obtained from the Wikipedia page that contains the postal codes. An overview of the website is provided in Figure 1, where is it possible to see that the postal codes with their corresponding neighborhoods and boroughs are provided via a table. Furthermore, some boroughs contain more than one neighborhood and for some postal codes are not active (see V4H and V8H in Figure 1).



WIKIPEDIA

The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)
[Contact us](#)
[Donate](#)

[Contribute](#)
[Help](#)
[Learn to edit](#)
[Community portal](#)
[Recent changes](#)
[Upload file](#)

[Tools](#)
[What links here](#)
[Related changes](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Cite this page](#)
[Wikidata item](#)

[Print/export](#)
[Download as PDF](#)
[Printable version](#)

[Languages](#)
[Français](#)
[Edit links](#)

[Article](#)
[Talk](#)

[Read](#)
[Edit](#)
[View history](#)

Figure 1. Overview of the website

After having created a clean data frame with all relevant postal codes and corresponding boroughs and neighborhoods, it is necessary to determine the geographical coordinates, i.e., latitude and longitude for each one of them. In order to achieve this, a function is created and applied to each row of the data frame (each postal code). Then, these two independent datasets are joined (by using the postal code as key), and a complete dataset of 44 rows and 5 columns is obtained and exported to excel, so that it can be stored independently and shared if necessary.

Afterwards, it is necessary to determine which one of these neighborhoods is the best, for which relevant criteria are defined. In this case, Foursquare API is used to explore the boroughs and neighborhoods. A limit of 100 venues is set, as well as a radius of 500 meters.

- 4) Data understanding and preparation: requires the use of descriptive statistics, as well as data cleaning and data transformation. The data quality is highly determined by this exploratory analysis. Then comes the data preparation, where missing or invalid values will be analyzed and duplicates removed. In this step it will also be determined the frequency with which each venue happens, as well as the different venues' categories.

During this step, visualization techniques support the data understanding. The first visualization is a map of Vancouver's neighborhoods. In order to do this, the latitude and

longitude previously obtained are used to create a map with folium library. The resulting map is shown in Figure 2.

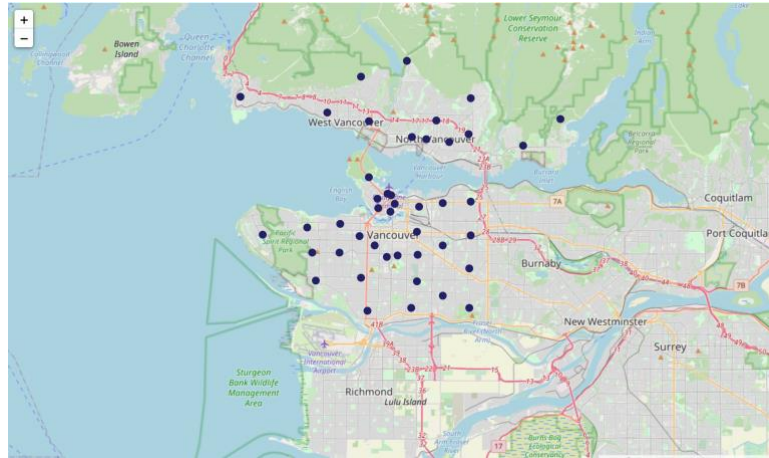


Figure 2. Map of Vancouver's neighborhoods

After the first collected dataset was collected, understood and prepared for further analysis, the venues information has to be analyzed and prepared. In this case, 4379 venues were returned by Foursquare API for the 44 neighborhoods, with 237 unique categories. After the venues are obtained, a list of all the categories is shown, so that it is possible to select which ones are relevant. Although it would also be possible to carry out further analysis with all the 237 dimensions, it is also possible to reduce the dimension of the dataset by implementing techniques such as Principal Components Analysis (PCA) or by selecting the criteria that are relevant. In this case, it is determined that the most important criteria when comparing neighborhoods are the existence of hotels, vegetarian / vegan restaurants, coffee shops and gym / fitness centers. These criteria are considered equally weighted and the aim is to find a location which offers all these venues.

- 5) Modeling and Evaluation: requires that the data is visualized in order to answer the initial question. The model would be based on the collected data and would, through a sequence of criteria, determine the answer of the question. Then the model needs to be evaluated. In this case, K means clustering is the selected algorithm used to be able to differentiate neighborhoods according to defined criteria / categories.

Before applying the Kmeans clustering algorithm, it is necessary to determine the number of clusters that should be created, for which the elbow method was used. The results of this

analysis are depicted in Figure 3. According to the elbow method, it was decided to create three clusters.

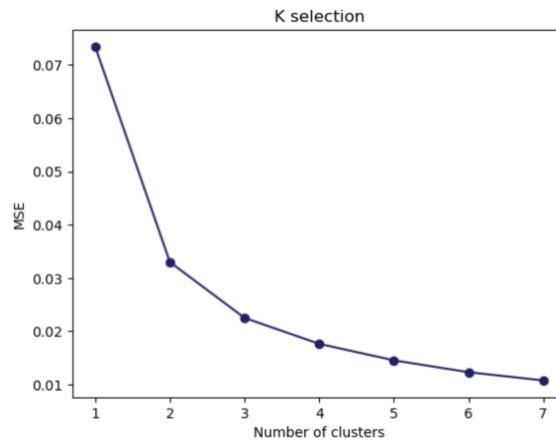


Figure 3. Elbow method for determining the number of clusters to be created

It is essential to notice that the previously mentioned steps are iterative, i.e., it is not a linear process that it is not uncommon that after reaching one step, it is necessary to go back to the previous one(s) and make necessary arrangements. For instance, in this project, the first data set was collected, cleaned, understood and visualized before collecting the second one and merging it with the first. Furthermore, it was also necessary to iterate between the fourth and fifth steps, as a first model was built, but during evaluation it was clear that the model was not sufficiently accurate, so it was necessary to understand and prepare the data again in a format that was suitable for further analysis and model building.

4. Results

After applying the previous methodology, it was possible to get a dataset with all the neighborhoods, their geographical coordinates and venues for each one of the specified categories. Furthermore, three different groups / clusters of neighborhoods were created, each one of them having different characteristics allow to better decide which one is the most suitable for living. The two most valuable outputs of the project are the data frame with all the information, as well as the map shown in Figure 4, which provides a visualization of the different clusters.

Furthermore, it is also valuable to establish the methodology, which can be implemented for any similar analysis in another city.

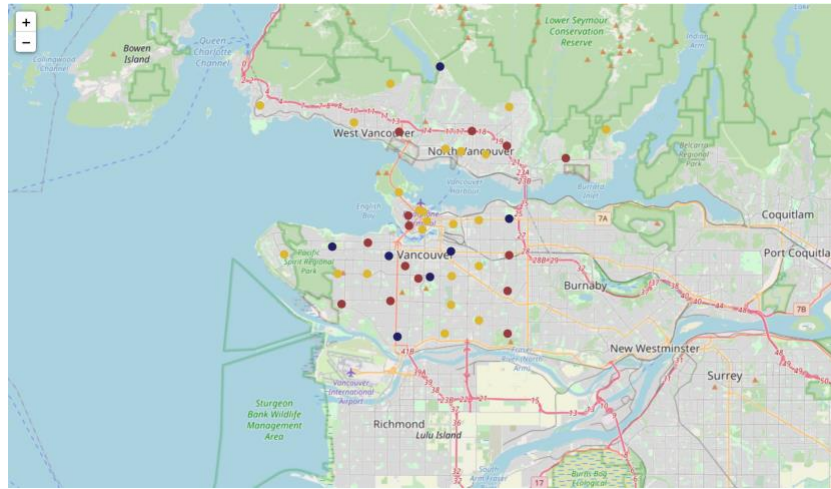


Figure 4. Map of Vancouver's clustered neighborhoods

5. Discussion

According to the carried-out analysis, there are three groups of neighborhoods, each one with different relative density. The third cluster is the one that includes the higher number of neighborhoods (23), followed by the first one (14) and lastly the second one with 7 elements. The waffle chart shown in Figure 5 shows the relative density of the clusters.



Figure 5. Waffle chart of the cluster's relative density

5.1. Observations

Even if the waffle chart allows to see which cluster includes more neighborhoods, it is still necessary to further determine the differences or main characteristics of each cluster, which is why Boxplots are built to differentiate each cluster and to have a better idea of which cluster contains the neighborhoods with the highest average presence of each venue's category. In this

case, the first group includes the neighborhoods that on average present venues on the four provided categories, i.e., hotels, gym/fitness centers, vegetarian/vegan restaurants and coffee shops, which correspond to the following postal codes: V6K, V7J, V5Z, V6N, V7H, V5S, V6H, V7T, V6M, V5R, V5M, V6Z, V6E, V7N. This allows to narrow the search from 44 possibilities to 14, i.e., eliminates 70% of the postal codes, making the search less complex and more focused on the target. While the second and third clusters have some higher average level of venues for some categories, they did not include all of them. The second cluster includes neighborhoods that have a relative high amount of coffee shops and some vegetarian /vegan restaurants and gym/fitness centers, but no hotels, which could be highly inconvenient when looking for a temporary accommodation. Regarding the third and last cluster, the neighborhoods included here present relative high presence of coffee shops and vegetarian /vegan restaurants and some hotels, but almost no gym/fitness centers.

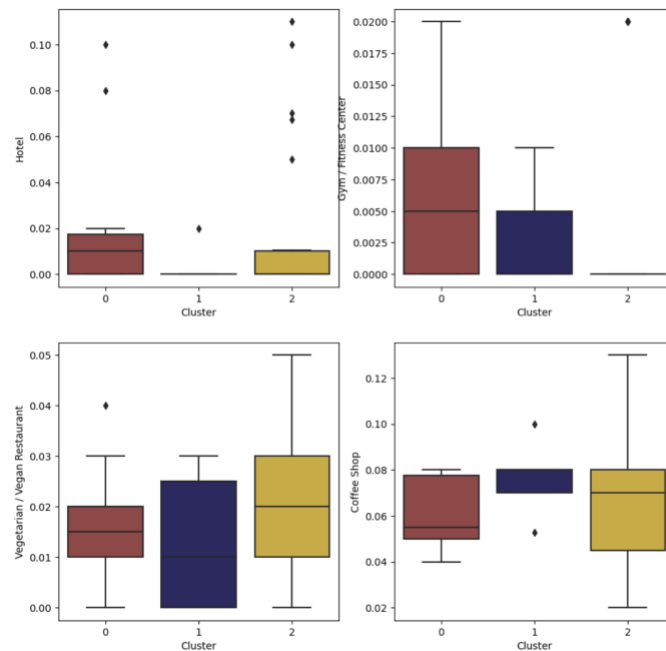


Figure 6. Boxplots for cluster's comparison

5.2. Recommendations for future research

It is recommended to carry out the research with a higher number of venues, as the higher number of venues provides more accurate solutions / models. Furthermore, it could also be convenient to take into account other criteria and compare results, as well as using reduction techniques such

as PCA, which could allow to cluster neighborhoods in a relatively easy way but still taking into account most of the information available.

6. Conclusions

As mentioned in the introduction, the aim of the project was to use a data science methodology and tools to allow people to make data-driven decisions on selecting the best location to live in Vancouver, Canada. The selected relevant criteria for comparing neighborhoods were the existence of hotels, vegetarian / vegan restaurants, coffee shops and gym / fitness centers. These criteria were equally weighted and the most convenient location would be one offers all these venues. The carried-out analysis allowed to determine that neighborhoods that were included in the first cluster were the ones where, on average, all the venues' categories considered relevant for this project could be found, which is why neighborhoods included in this cluster would be the recommended locations.

7. References

- Aljarah, I., Faris, H., & Mirjalili, S. (2021). Evolutionary Data Clustering: Algorithms and Applications (Algorithms for Intelligent Systems) (1st ed. 2021 ed.). Springer.
- Alpaydin, E. (2020). Introduction to Machine Learning, fourth edition (Adaptive Computation and Machine Learning series) (fourth edition). The MIT Press.
- Cielen, D., Meysman, A., & Ali, M. (2016). Introducing Data Science: Big Data, Machine Learning, and more, using Python tools (1st ed.). Manning Publications.
- Decaria, A., Petty, G. W., & Weidemann, L. (2021). Python Programming and Visualization for Scientists (2nd ed.). Sundog Publishing, LLC.
- Lee, J., & Sainsbury, B. (2020). Lonely Planet Vancouver & Victoria 8 (City Guide) (8th ed.). Lonely Planet.
- Schmidt, S. (2012). Geolocation mit PHP - Foursquare-API, Google Places & Qype (German Edition). entwickler.press.
- Thogarcheti, H. S., Pulabaigari, V., & K, M. (2020). Improvements over k-means clustering methods for large datasets: Prototype based hybrid techniques. LAP LAMBERT Academic Publishing.
- Wikipedia contributors. (2021, May 6). List of postal codes of Canada: V. Wikipedia. https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V