Coursera IBM Professional Certificate
Applied Data Science Capstone

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

Alina Prendes Roque

August 2021

Agenda

1. Introduction
2. Data
3. Methodology
4. Results and discussion
5. Conclusions

Agenda

1. **Introduction**
2. Data
3. Methodology
4. Results and discussion
5. Conclusions

This project aims to apply a data science methodology to determine the best location to live in Vancouver, Canada

Business problem

Deciding the best location to live in new country or neighborhood

Target audience

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

Project aim

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 based on user-specified criteria

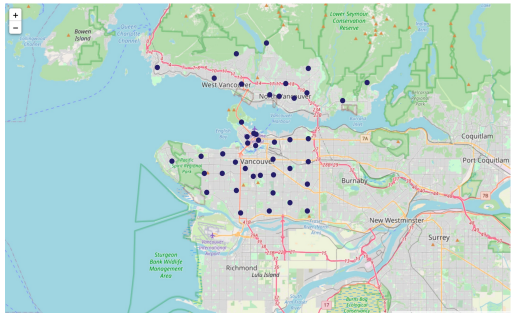
Agenda

1. Introduction
2. **Data**
3. Methodology
4. Results and discussion
5. Conclusions

The required data was obtained from two main different sources

Data set # 1

A comprehensive list of Vancouver's neighborhoods with the respective postal code, latitude and longitude



Source: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_V.

Data set # 2

Venues information for each neighborhood, which serves as criteria for further comparative analysis

FOURSQUARE

Source: https://foursquare.com

Agenda

1. Introduction
2. Data
3. **Methodology**
4. Results and discussion
5. Conclusions

The implemented methodology consists of five steps

1. Analytic approach

Define the question to be answered: if a person is planning to move to Vancouver, which would be the best location to rent or buy accommodation?

2. Data requirements

Determining which information is required, in which format and where it can be obtained

3. Data collection

Data set # 1: 44 rows and 5 columns

Data set # 2: 4379 venues were returned by Foursquare API for the 44 neighborhoods, with 237 unique categories

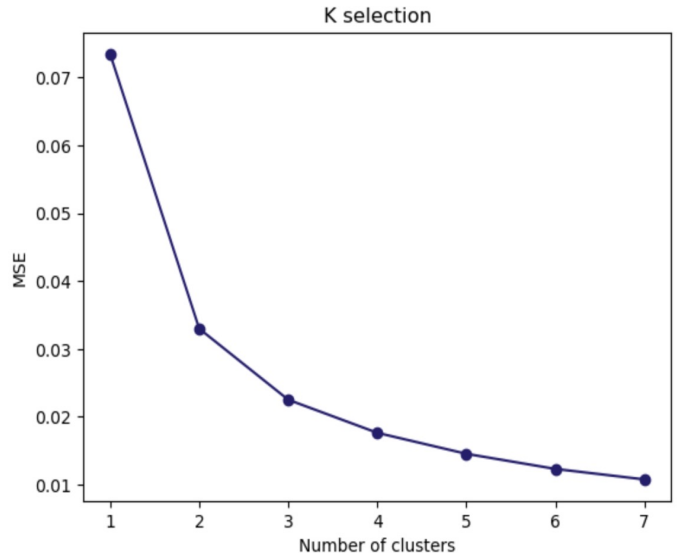
4. Data understanding and preparation

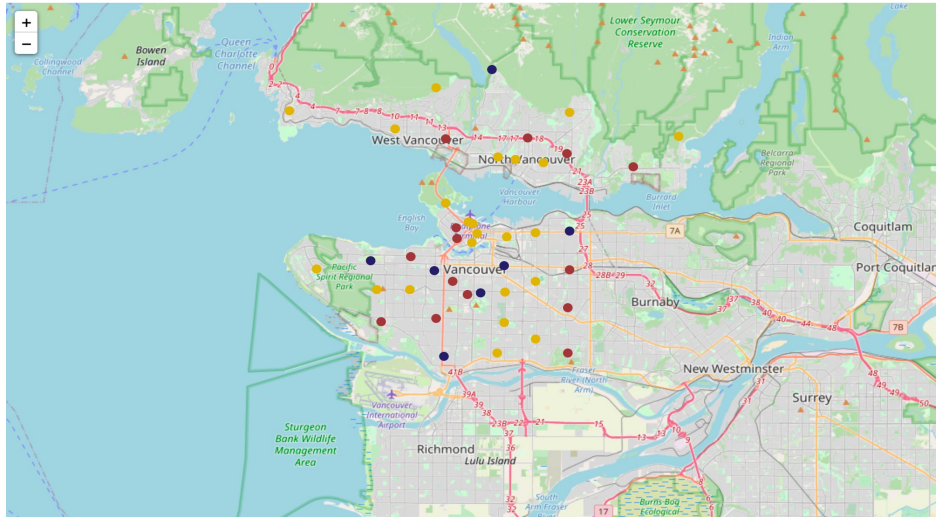
The most important criteria when comparing neighborhoods are the existence of hotels, vegetarian / vegan restaurants, coffee shops and gym / fitness centers. These criteria are considered equally weighted and the aim is to find a location which offers all these venues.

The implemented methodology consists of five steps

5. Modeling and evaluation

An unsupervised machine learning algorithm that allows to characterized the neighborhoods and cluster them according to the previously specified criteria, in this case k-means clustering is implemented, with a total of 3 clusters determined by the elbow method



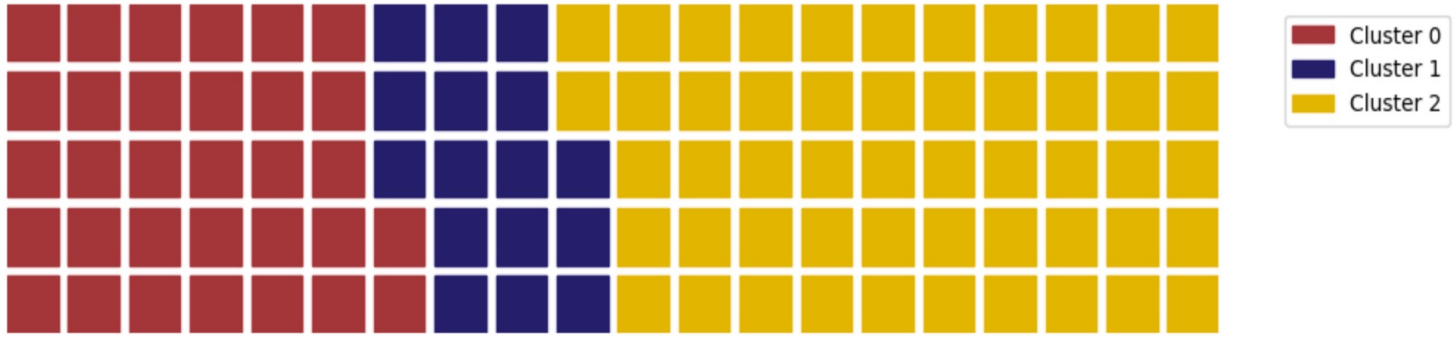


Agenda

1. Introduction
2. Data
3. Methodology
4. **Results and discussion**
5. Conclusions

Three clusters were created, each one with a specific relative density

Waffle chart



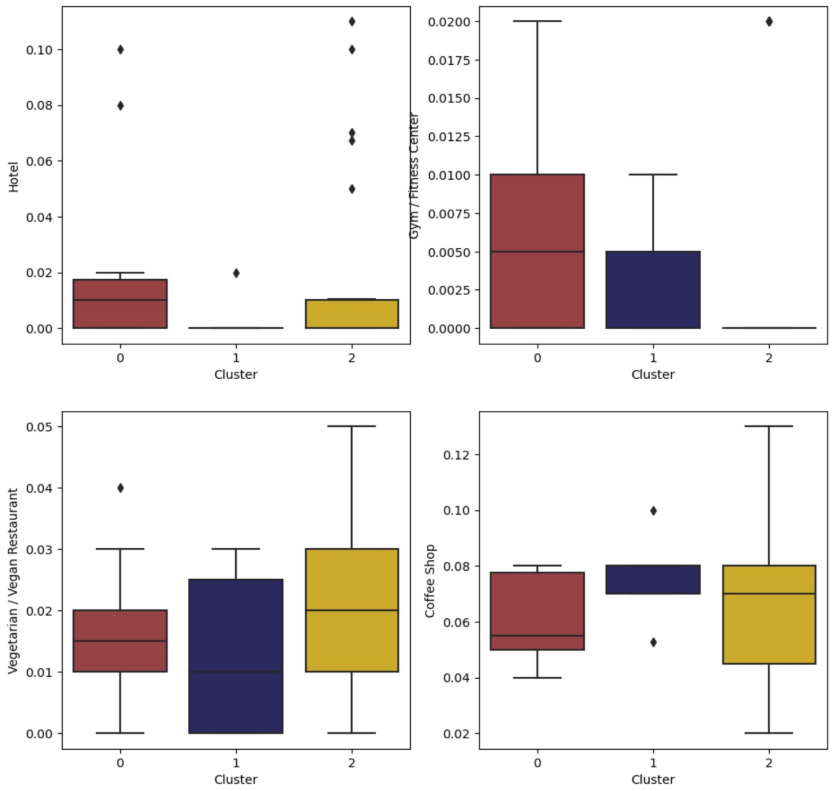
Analysis

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.

Note that the waffle chart shows the relative density (percentage), which is not the absolute number of elements of each cluster

The first cluster includes neighborhoods where all venues are present

Boxplots



Analysis

The first cluster 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.

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.

The third cluster includes neighborhoods that present relative high presence of coffee shops and vegetarian / vegan restaurants and some hotels, but almost no gym/ fitness center

Agenda

1. Introduction
2. Data
3. Methodology
4. Results and discussion
5. **Conclusions**

Conclusions

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 implemented methodology consisted on five steps, which included an analytic approach, establishing the data requirements, colleting the data, understanding and preparing the data and lastly building and evaluating the model.

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.

„All models are wrong, but some are useful“

George E. P. Box, 1976