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 be 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, Beautiful Soup, Request and Pandas libraries is 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.