# IBM Data Science Capstone Project

Part of IBM Data Science Professional Certificate

Opening a craft beer tap room in Vancouver

By: Agnieszka Jugowicz

July 2020

Table of Contents

[IBM Data Science Capstone Project 1](#_Toc46736952)

[1. Introduction 2](#_Toc46736953)

[2. Data 2](#_Toc46736954)

[3. Methodology 3](#_Toc46736955)

[4. Analysis 3](#_Toc46736956)

[5. Result and Discussion 3](#_Toc46736957)

[6. Conclusion 3](#_Toc46736958)

# 

# Introduction

Vancouver is a coastal seaport city in the mainland of British Columbia, Canada. Vancouver is consistently ranked one of the most livable cities in the world. Its location near ocean, tucked up against mountains, makes the city a go-to location for year-round exploration.

Vancouver’s metropolitan area, with population of over 600,000 inhabitants, is an ideal place for investment, especially in the hospitality industry. The tourism is flourishing, as the city combines an urban getaway with outdoor adventures. Vancouver has also a rich cultural background and is third-largest film production center in North America, holding the moniker of “Hollywood North”.

This project will attempt to explore the patterns of neighbourhoods within Vancouver by categorizing them into clusters in order to identify existing trends within each neighbourhood. Using data science and machine learning techniques, this project aims to provide an answer to the business question: If an investor is looking to open a new craft beer tap room, where should it be opened?

The recommendation is aimed at investors wanting to open or expand their business in the hospitality sector of Vancouver.

# Data

In order to analyze the trends within each neighbourhoods of Vancouver, the list of forward sortation areas (FSAs) was downloaded from the website https://www.geonames.org/postalcode-search.html?q=vancouver&country=CA as CSV file. Forward sortation area is a geographical region in which postal codes start with the same three letters. It enables a geographical grouping of city areas, creating neighbourgoods.

The database, as seen below, consists of the following columns: *CountryCode*, *FSA*, *Neighbourhoods*, *Province*, *Latitude*, and *Longitude*.



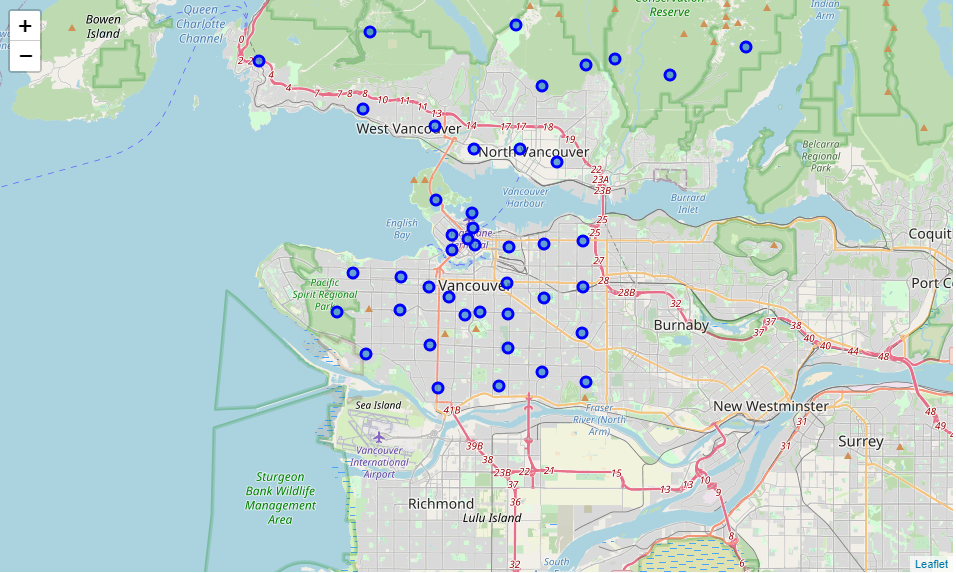
Venue queries will be performed using FourSquare APIs. The results of the venue queries will then be used to observe commonalities between neighbourhoods. The clusters will then provide insights on which cluster is most suitable for the investment. K-means clustering algorithm will be used to find patterns between the neighbourhoods.

# Methodology

Firstly, the data retrieved from [www.geonames.org](http://www.geonames.org) website will be loaded into pandas dataframe and then visualized in a map using Folium package in order to validate the correctness of data. Next, the FourSquare API will be used to obtain the venues data. API calls to FourSquare will be made, passing in the geographical coordinates of the neighbourhoods in a Python loop.

The K-means clustering algorithm will be used to categorize neighbourhoods within Vancouver. A one-hot encoding will be performed on the venue dataframe and the dataframe will be grouped by each neighbourhood. The encoding will provide information on venue categories (in columns) per neighbourhood. This data will then be grouped to provide weighting of venue type occurrence per neighbourhood. Next, the encoded dataframe will be filtered into top venues, before K-means algorithm is implemented.

# Analysis



# Result and Discussion

# Conclusion