For this week, you will required to submit the following:

1. A description of the problem and a discussion of the background. (**15 marks**)

A new student is moving to Vancouver, Canada to study at UBC. To help her decide where to live, s/he would like to know which neighborhood has which characteristics and how each neighborhood differs from each other.

1. A description of the data and how it will be used to solve the problem. (**15 marks)**

In courtesy to GeoNames (https://www.geonames.org), a downloadable csv file is available and contains all Canadian postal codes and corresponding geospatial data. Only the first letters of the full postal codes are provided for copyright reasons. The data format is tab-delimited text in utf8 encoding, with the following fields:

* country code: iso country code, 2 characters
* postal code: varchar(20)
* place name: varchar(180)
* admin name1: 1st order subdivision (state) varchar(100)
* admin code1: 1st order subdivision (state) varchar(20)
* admin name2: 2nd order subdivision (county/province) varchar(100)
* admin code2: 2nd order subdivision (county/province) varchar(20)
* admin name3: 3rd order subdivision (community) varchar(100)
* admin code3: 3rd order subdivision (community) varchar(20)
* latitude: estimated latitude (wgs84)
* longitude: estimated longitude (wgs84)
* accuracy: accuracy of lat/lng from 1=estimated, 4=geonameid, 6=centroid of addresses or shape

For the second week, the final deliverables of the project will be:

1. A link to your Notebook on your Github repository, showing your code. (**15 marks**)

https://github.com/tedtkwang/Coursera\_IBM\_Capstone/blob/master/9\_Applied%20Data%20Science%20Capstone%20Project%20Notebook%20V.ipynb

1. A full report consisting of all of the following components (**15 marks**):

* **Introduction where you discuss the business problem and who would be interested in this project.**

A new student is moving to Vancouver, Canada to study at UBC. To help her decide where to live, s/he would like to know which neighborhood has which characteristics and how each neighborhood differs from each other. Other potential interested parties may include real estate agents, housing developers, etc.

* **Data where you describe the data that will be used to solve the problem and the source of the data.**

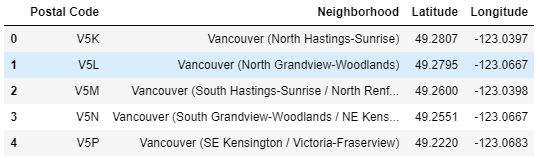
In courtesy to GeoNames (https://www.geonames.org), a downloadable csv file is available and contains all Canadian postal codes and corresponding geospatial data. Only the first letters of the full postal codes are provided for copyright reasons. The data format is tab-delimited text in utf8 encoding, with the following fields:

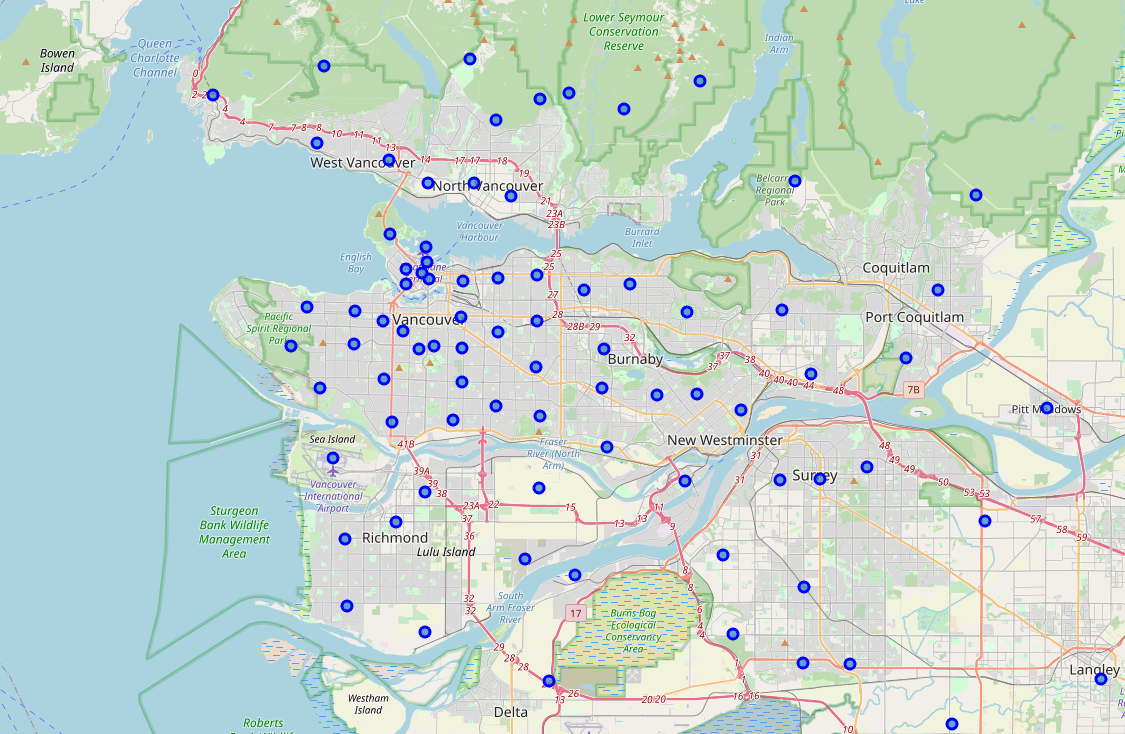
* country code: iso country code, 2 characters
* postal code: varchar(20)
* place name: varchar(180)
* admin name1: 1st order subdivision (state) varchar(100)
* admin code1: 1st order subdivision (state) varchar(20)
* admin name2: 2nd order subdivision (county/province) varchar(100)
* admin code2: 2nd order subdivision (county/province) varchar(20)
* admin name3: 3rd order subdivision (community) varchar(100)
* admin code3: 3rd order subdivision (community) varchar(20)
* latitude: estimated latitude (wgs84)
* longitude: estimated longitude (wgs84)
* accuracy: accuracy of lat/lng from 1=estimated, 4=geonameid, 6=centroid of addresses or shape

We also utilize the Foursquare, which is the location data provider we practiced using in this course, API to explore the neighborhoods and segment them.

* **Methodology and Results section which represents the main component of the report where you discuss and describe any exploratory data analysis that you did, any inferential statistical testing that you performed, if any, and what machine learnings were used and why.**

We retrieved our target locations’ geospatial data:



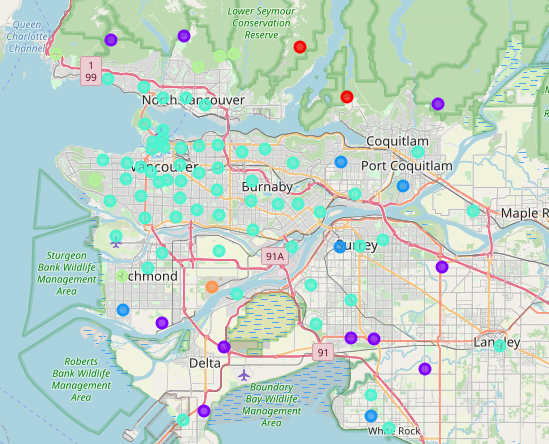


We retrieved top 20 venues within 1000-meter radius of each neighborhood in Greater Vancouver. We then checked how many venues were returned for each neighborhood. From this we got to now there are 1,223 venues in 219 unique categories across our target 88 neighborhoods: ranging from ATM, various types of restaurants, public facilities, to yoga studios, etc.

Next, we grouped results by each neighborhood by taking the mean of the frequency of occurrence of each venue category. Thus, we can show top 10 venue categories in each neighborhood.



To understand how each neighborhood differs from each other, we ran k-means clustering to cluster the neighborhoods into 6 clusters. Finally, we show visualization of the resulting clusters.



* **Conclusion and Discussion section where you discuss any observations you noted and any recommendations you can make based on the results.**

Neighborhood clustering based on Foursquare venues data effectively display each neighborhood’s venues or lifestyle characteristics.

For potential client interest, we could enhance neighborhood segmentation by including other neighborhood-specific socio-economic data. E.g. income level, education level, crime rate, housing prices, etc.