**Capstone Project Report**

**Introduction/Business Problem section:**

When families with school age children move to a new city, they usually face different challenges at the beginning. Choosing a suitable neighbourhood to rent or buy a house or apartment is one of those challenges. In particular, new immigrants or low income families are more likely to not own a car and usually look for a school for their children at a central location where lots of amenities are nearby. Then they look for a residence in that neighbourhood.

I live in the Vancouver area and many times I have seen families with school age children looking for a suitable location in Vancouver to rent or buy a house. The family’s top priority in many cases has been proximity to the children’s school. The family also wants to be close to as many amenities as possible so that the use or need for a car is minimized. I will use the school data from the city of Vancouver to address this problem and facilitate families with data visualization and organization methods that would help them to quickly identify the schools or areas of interest.

**Data section:**

We obtain data on about 190 schools in Vancouver from the city of Vancouver’s Open Data Portal <https://opendata.vancouver.ca/explore/dataset/schools/api/>

For each school, the dataset contains the name of school, area, address, type of school, Latitude and Longitude. We will be using a combination of this dataset and data acquired from Foursquare in order to facilitate families with the information they need to make the optimum decision for their family. For example, we will be providing data/information on the number of venues around each school.

**Methodology section:**

In this project exploratory data analysis as well as the k-clustering method from Machine Learning is used to analyze and visualize the data on schools in Vancouver.

The data obtained from the city of Vancouver’s open data portal is in the format of a json file and I start by transfer the relevant data/information into a list and then subsequently converting the list into a pandas dataframe.

Using the location data on schools and folium map, I create a map with the list of all Vancouver school information on it. Then by connecting to FourSquare, I get information on the number of venues nearby each school. FourSquares gives a dataset of size 4251 on the venues nearby the schools.

I organize the venue information by counting the number of venues nearby each school and by cleaning the data and getting rid of unnecessary information/columns. A histogram is then created with 5 bins to visualize the distribution of the number of venues nearby schools.

Next we prepare the data for the clustering method (from Machine Learning). We choose 5 clusters. We then prepare the data to visualize the 5 clusters on the map with 5 different colors on the map.

We also perform some further data analysis in order to sort the schools from the ones with the highest number of venues nearby to the lowest one. We also organize the data to see the schools in each area with the accompanying number of venues nearby and also their type (Public, Independent, etc.). We also identify the schools with at least 50 venues nearby.

**Results section:**

The main results obtained from analyzing the school’s data in this project are discussed below.

The histogram shows that most schools (about 150) have fewer than 40 venues nearby and some (about 20) have 40 to 60 venues nearby and very few schools (fewer than 10) have more than 80 venues nearby.

The clustering method gives us the interactive visual map to see the 5 identified clusters. Also we get the average number of venues nearby schools in each cluster.

School Latitude School Longitude Venue\_Count

Labels

0 49.249447 -123.051519 13.408451

1 49.261256 -123.103424 45.714286

2 49.234794 -123.116393 9.250000

3 49.275732 -123.135124 89.454545

4 49.250589 -123.175956 10.266667

Sorting the schools from the ones with the highest number of venues nearby to the lowest one also gives us the list of schools with the highest number of venues nearby. King George Secondary in West End and Pattison High School in Downtown are at the top of the list.

**Discussion section:**

Based on the graphs and tables produced during the analysis, we see that most schools have fewer than 50 venues nearby and only fewer than 10 schools have more than 80 venues nearby. Families interested in this problem can visually look at the 5 created clusters and get help on identifying their child’s school as well as location of residence. For future work, I would like to also collect data on the academic ranking of the schools and redefine the clusters based on their location, the number of venues nearby as well as the ranking of school.

**Conclusion section:**

Based on our analysis of the school data for the city of Vancouver, most schools are not located in very central locations and have fewer than 50 venues nearby. The created interactive map can help families to see the location and the number of venues nearby each school in a given cluster.