CAPSTONE PROJECT – THE BATTLE OF THE NEIGHBOURHOODS

1. Introduction

According to The Economist Intelligence Unit, Calgary is among the top 5 most livable cities in the world and the most livable city in North America in the last two years [1]. Every year, it attracts tens of thousands of interprovincial and international migrants.

Information about each neighbourhood of the city will be very helpful for newcomers whether they come to study, work, or start new business. For example, a single office worker might have to choose whether to rent an apartment near downtown or buy a house elsewhere more affordable and spend time commuting by public transit.

This project aims to provide insight information about Calgary's neighbourhoods, cluster neighbourhoods based on their venues, and identify which variables most influential on house price. Hopefully, this work will help to point newcomers to the right direction and avoid costly mistakes when settling in Calgary.

2. Data collection and cleaning

- Postal codes and coordinates of neighbourhoods are web-scraped at GeoNames [2] and Mapawi [3]. The two datasets are combined as the coordinates from the former are unreliable while neighbourhood names from the latter are incomplete.
- Venue data are retrieved with Foursquare API [4].
- Neighbourhood boundaries are downloaded from Statistics Canada [5]. This boundary file is then filtered, simplified, edited and converted at mapshaper.
- Population, area, house price, crime, and school ratings data of each neighbourhood are retrieved from Wikipedia [6], The Globe and Mail House Price Data Centre [7], Calgary Police Statistical Reports, and Fraser Institute School Rankings [9], respectively. They are then aggregated to obtain several statistics such as population density, crime rate, average school ratings etc.

3. Exploratory Data Analysis

The area of each neighbourhood varies in quite a large range (Figure 1). Therefore, the radius to search for revenues within each neighbourhood should be variable, not a constant number.

The average population density of Calgary is about 2000-2500 resident/km^2 (Figure 2). The most populous neighbourhoods are located around downtown area.

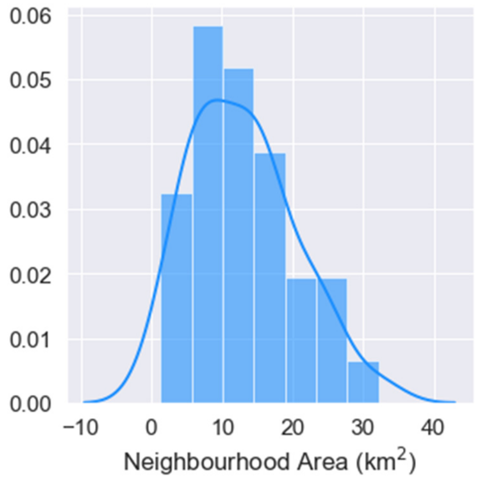

Figure 1. Neighbourhood area distribution

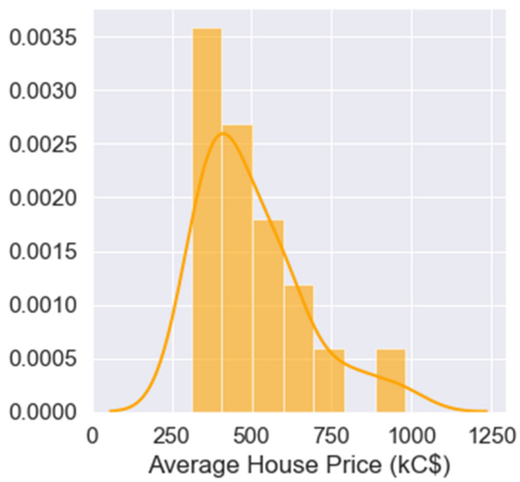

Figure 2. House price distribution

Average house price appears to be skewed right with a mode of 300 – 400 thousand Canadian dollars (Figure 2). Neighbourhoods next to downtown have the highest average house prices but the house price in downtown is not high. Northeast neighbourhoods appear to have lower price than other areas.

While downtown has the highest crime rate over the last 5 years, northeast areas appear to have high total number of crimes (Figure 3) and low school ratings (Figure 4). The relation between house price and several statistics are illustrated in Figure 5.

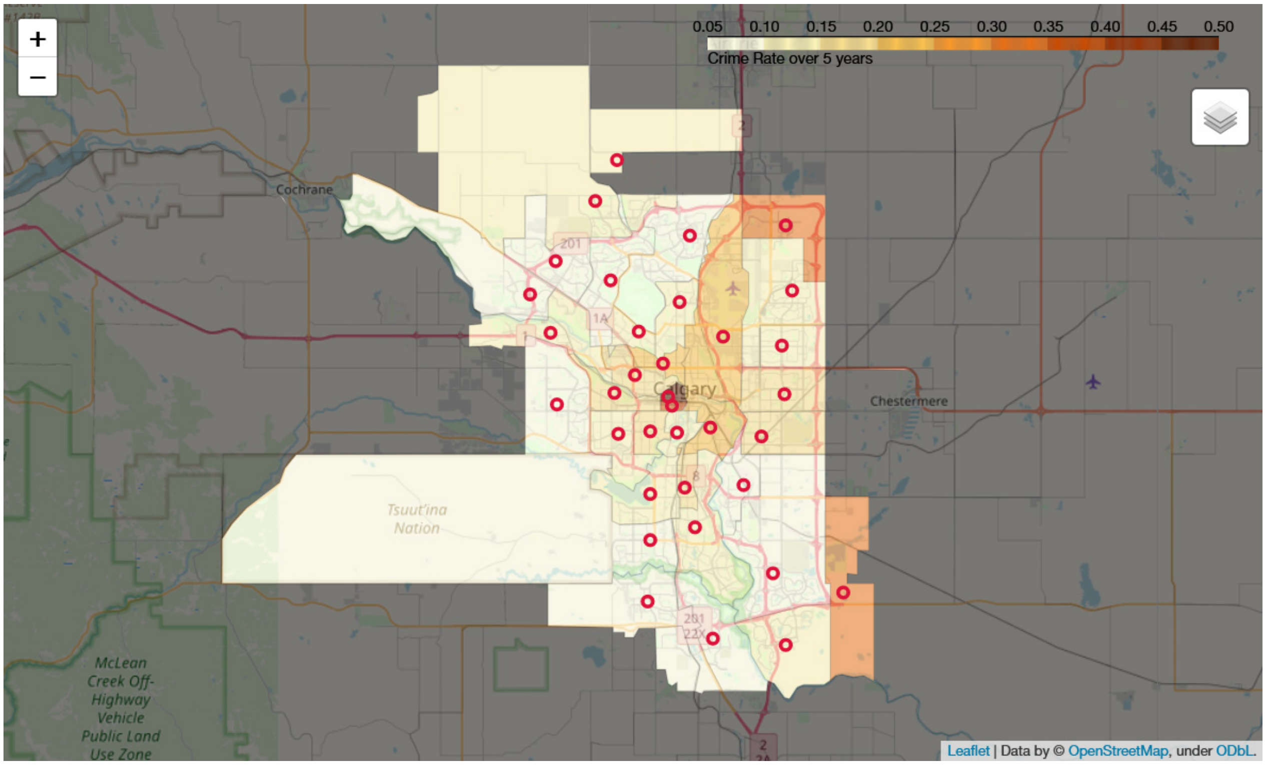

Figure 3. Crime rate over the last 5 years

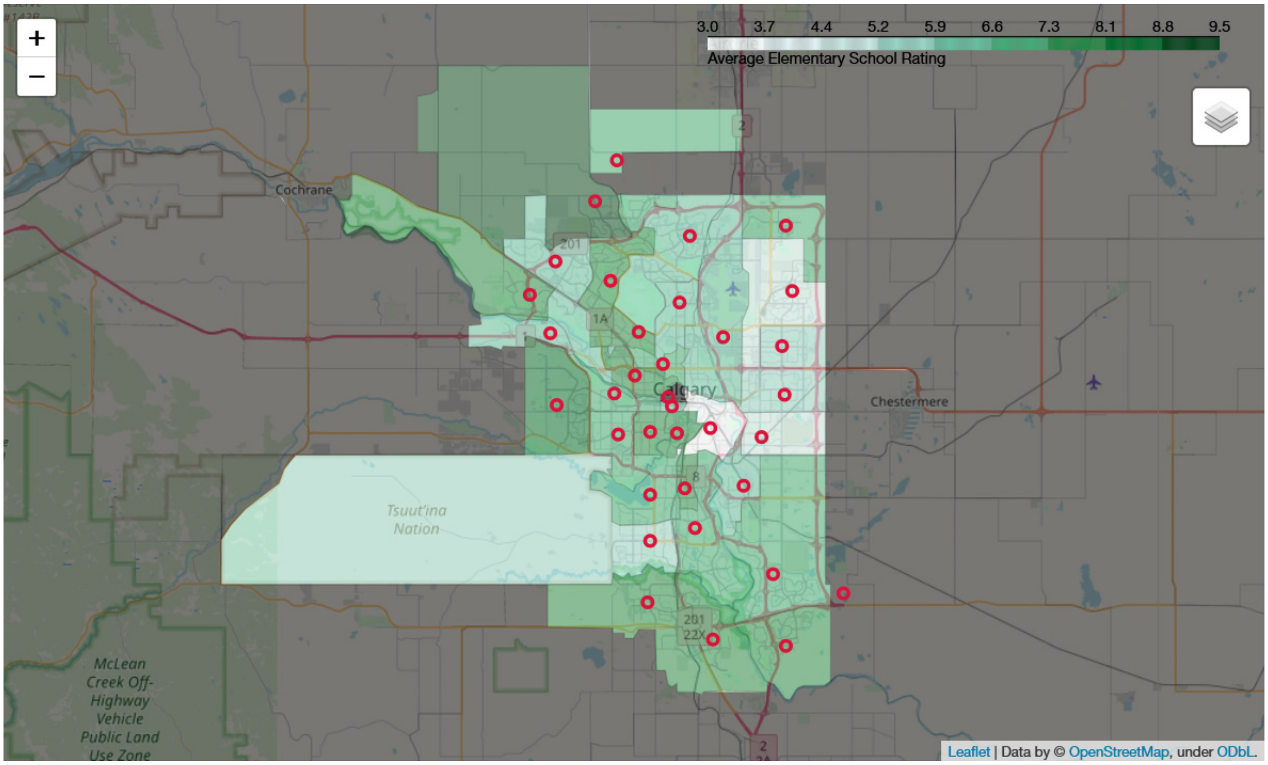
+
−
3.0
3.7
4.4
5.2
5.9
6.6
7.3
8.1
8.8
9.5
Average Elementary School Rating
Cochrane
Calgary
Chestermere
Tsuut'ina Nation
McLean Creek Off-Highway Vehicle Public Land Use Zone
201
1A
8
2
201
22X
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 4. Average elementary school rating

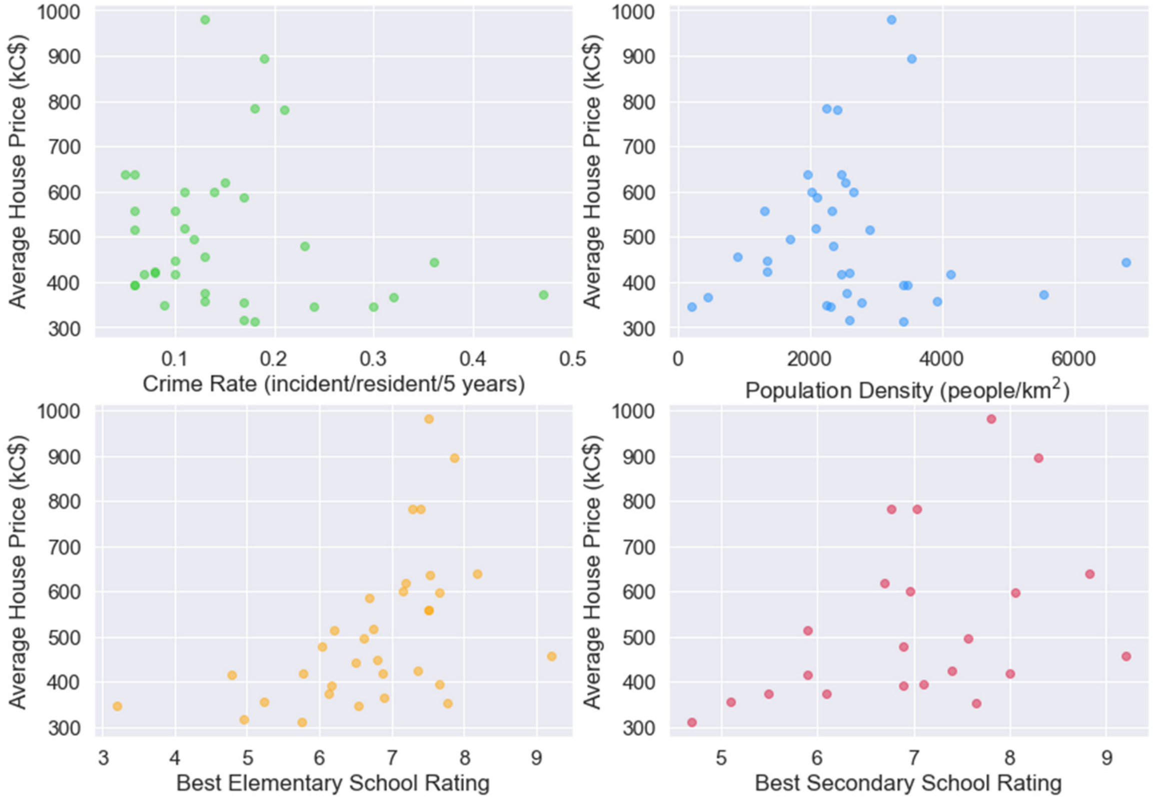

Figure 5. House price vs. crime rate, population density, school ratings

Venue data for each neighbourhood are searched within a radius of half of square root of neighbourhood area. Overall, there are 199 unique categories of venues. About 10 neighbourhoods reach the limit of 100 venues set by Foursquare. Neighbourhoods with lowest number of venues are T3S, T3P, and T3N which unsurprisingly are located at the outskirt of the city. The venue data are then one-hot encoded for clustering (Figure 6).

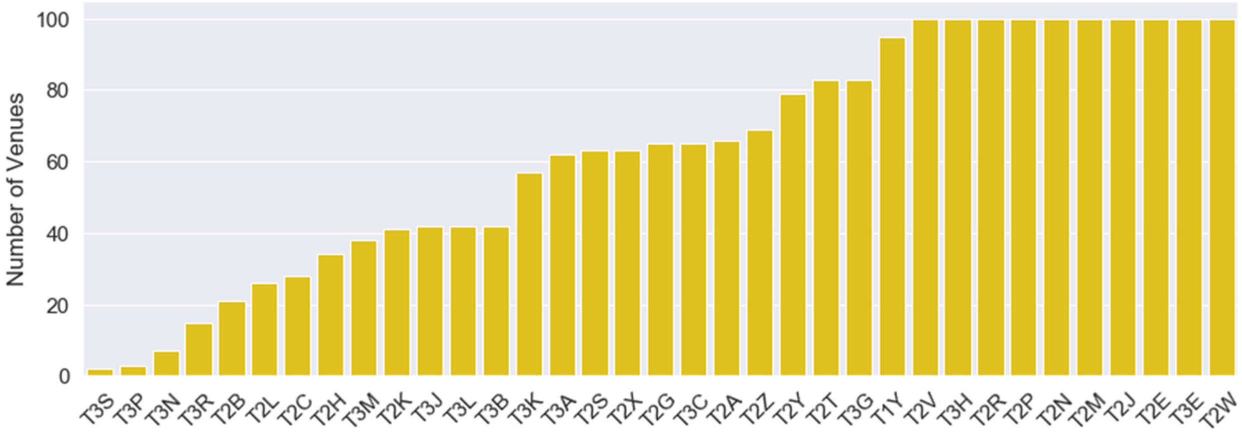

Figure 6. Number of venues of each neighbourhood

4. Modelling

4.1. Neighbourhood Clustering

Neighbourhoods are clustered using k-mean clustering. The optimal number of clusters are identified to be 5 using elbow rule of thumb (Figure 7).

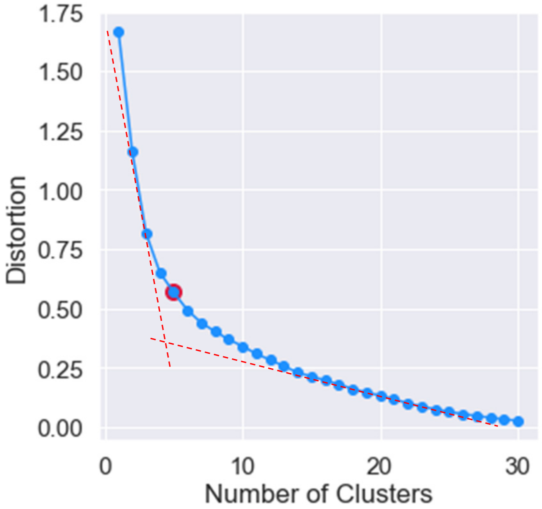

Figure 7. Optimal number of clusters

The result of k-mean clustering is shown in Figure 8. The background color represents average house price in each neighbourhood.

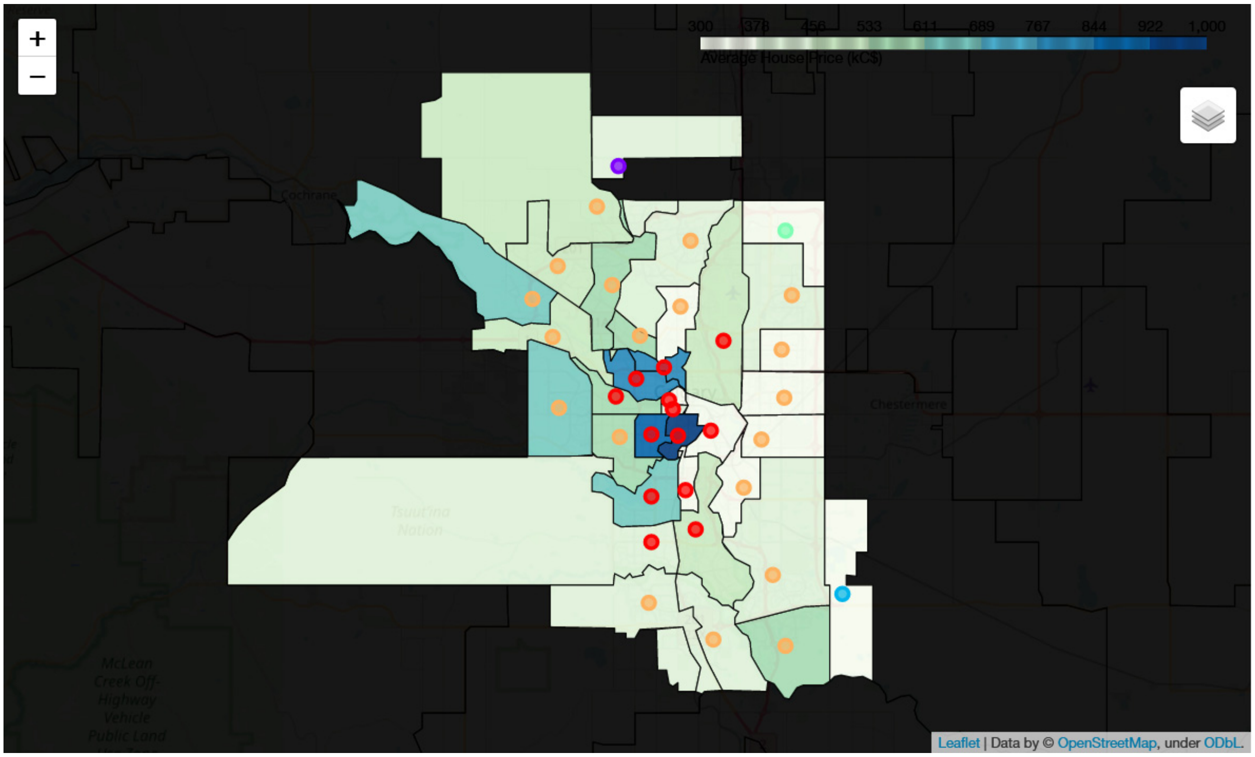
+
−
300
378
456
533
611
689
767
844
922
1,000
Average House Price (kC$)
Cochrane
Chestermere
Tsuu T'ina Nation
McLean Creek Off-Highway Vehicle Public Land
Leaflet | Data by © OpenStreetMap, under ODbL.

Figure 8. Clusters of neighbourhoods

The red cluster represents neighbourhoods in or around downtown area is characterized by venues like café, coffee shop, pub, bar, and restaurants which is suitable to single office workers who work in downtown.

The orange cluster represents neighbourhoods further from downtown area and is characterized by venues like grocery store, pharmacy, shopping mall, gym, park, restaurant which is more suitable to families or singles who prefer places that are more quiet than downtown.

The purple, green, and blue clusters represent neighbourhoods at the outskirt of the city and is characterized by low number of venues.

4.2. House Price Reasoning

As can be seen in Figure 8, houses in the same cluster based on venue data still can have very different prices. In order to identify what variables driving house price, the random forest regressor is employed to predict house price from statistics data and venue data mentioned in previous sections (Figure 9).

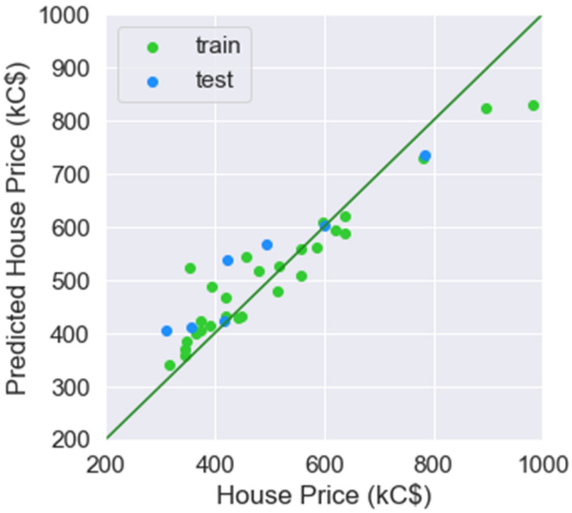

Figure 9. Performance of random forest regressor

One of the nice things about tree-based methods is that they provide importance of each input variable in the prediction output variable. In this case, elementary school ratings and venue density appear to be the most influential on house price (Figure 10).

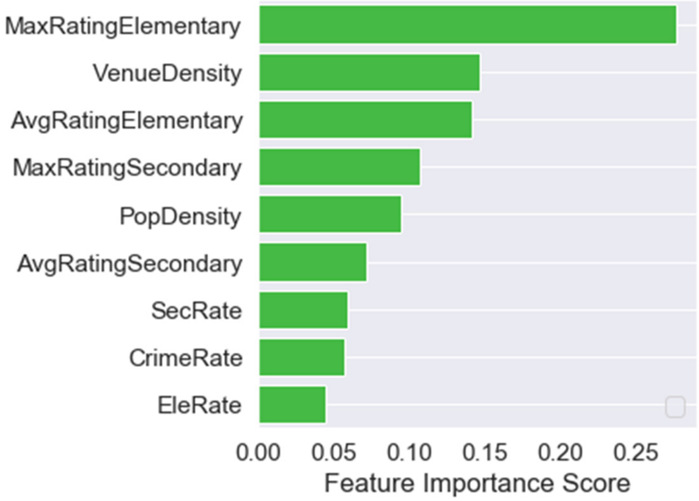

Figure 10. Importance score of variables

5. Conclusions and Discussion

In conclusion, the project provided insight information about several statistics of neighbourhoods in Calgary such as population density, crime rate, school ratings, average house prices through exploratory data analysis. More than 30 neighbourhoods (actually forward sortation areas) in Calgary were grouped into 5 clusters based on their venues through k-mean clustering. Average house price in each neighbourhood was modelled by random forest regression and driving variables were identified.

However, the results can still be improved. Firstly, coordinates of neighbourhood centers are important as they dictate how venues are searched in Foursquare. However, coordinates from GeoNames and Mapawi are not reliable. An algorithm to calculate center coordinates from boundary file will probably yield more reliable data. Secondly, Foursquare put a limit of 100 on a number of venues returned by each query. If this limit is raised and more importantly if accessing to venue ratings is granted, the result of clustering will be significantly improved as quality is more important than quantity. Last but not least, large amounts of demographic data of each neighbourhood such as age, income, occupation etc. are available on Statistics Canada. These data probably provide more useful information about each neighbourhood but collecting and cleaning them are also time-consuming.

References

[1] Calgary - Wikipedia

[2] Calgary - GeoNames

[3] Alberta - Mapawi

[4] Foursquare

[5] Community Boundary File – Statistics Canada

[6] List of Neighbourhoods in Calgary - Wikipedia

[7] House Price Data Centre - The Globe and Mail

[8] Statistical Reports - Calgary Police

[9] School Rankings - Fraser Institute