|  |
| --- |
|  |
| Best neighborhood hunting for constructing a new indoor young child playing center in City of Calgary |
| May 15 by Jeff Jia |

1. Introduction

1.1 Background

City of Calgary, I have lived for years, has a population of 1.2 million as per most recent census. It is the second largest municipality in Western Canada with reasonably good economic and very safe environment. The city offers so many things to enjoy, such as parks, rivers, trails and abundant snow activities as the nearby magnificent Rock mountains.

On the other hand, since Calgary is in northern 51-degree latitude, in general, its winter lasts longer than many other places. During the winter season, outdoor activities is not as enjoyable as summer for many young families. So, if you have visited or lived in the city, you know there are likely crowded shopping malls, café, restaurants and indoor facilities. Therefore, from economic point of view, for family with young children or business investor or stakeholders, you will see there is a reasonable need for indoor playground for playing, doing birth party (once coronavirus pandemic ends in the future).

1.2 Business problem & interest

The purpose of this project is to select optimal location or neighborhood to invest & construct indoor young children playground, inside the city of Calgary. The following lists the specific requirements for such a business investment:

* + This indoor playground is mainly for young kids playing, holding birthday party and so forth. Main age ranges are from new-born to 14 years old.
  + The best area or neighborhood should have relatively more young kids.
  + There is no or should be far away from existing indoor playground
  + The target neighborhood should be less crimes & good household income

This study is targeted to stakeholders or parties who are interested in investing in indoor recreation or early children education in the City of Calgary, Canada.

## **2. Data Source, Acquisition & Cleaning**

2.1 Data sources

Data will be collected from various source, including public website and City of Calgary Open Data portal and others. There are total Five classes of raw data we have collected.

* Calgary official neighborhood or communities from [HERE](https://en.wikipedia.org/wiki/List_of_neighbourhoods_in_Calgary) . Only residential neighborhoods were selected since we are dealing with indoor playing center for young children.
* City of Calgary social demographic & economical data such as neighborhood population, area in square kilometer, children group counts. The data can be found from [THIS SITE](https://data.calgary.ca/Demographics/Census-by-Community-2019/rkfr-buzb). We have downloaded the data and saved on GitHub for easy access. Check the the accompanying Jupiter Notebook code part for more details.
* Community crime & disorder statistic data from Calgary Police Service Open Data portal, the data link and file header explanation can be found in this [LINK](https://data.calgary.ca/Health-and-Safety/Community-Crime-and-Disorder-Statistics-to-be-arch/848s-4m4z). This data covers all incidents from 2012 to 2019. During this project, we extracted data for 2019 only for sake of quick data handling purpose.
* Community household average income history data was extracted from this [WEBSITE](https://great-news.ca/demographics). This data may not most recent statistics.
* Indoor recreational venue data were collected by using Foursquare API, from our neighborhood locations, with two Venue Categories as of “Arts & Entertainment” and “Outdoors & Recreation” only as per [Foursquare website](https://developer.foursquare.com/docs/build-with-foursquare/categories/). Then those outdoor recreation facilities are excluded since we are only care about indoor recreation for children business only.

In order to view the neighborhood data nicely, we have acquired City official community/Sector boundary Geojason files.

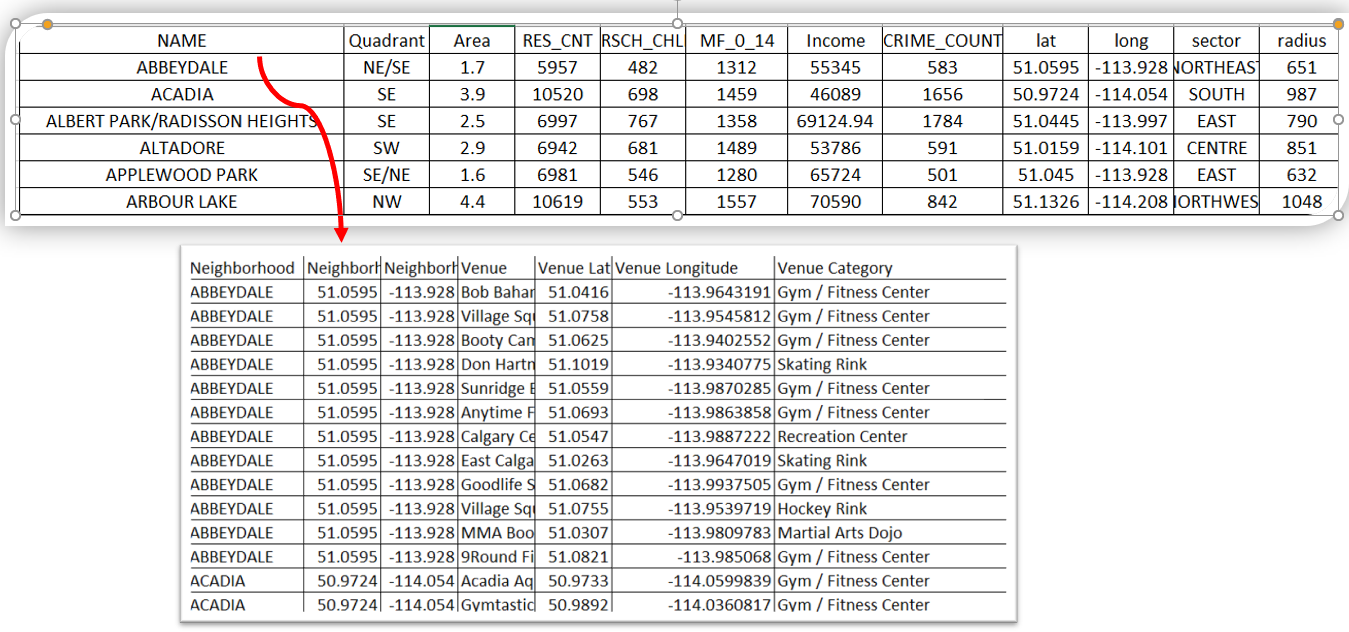
All necessary data & Python codes were loaded up to GitHub repository in my study for easy online access.

2.2 Data acquisition & cleaning

Two ways in data acquisition were adopted. The first was to use Python to read table directly from the public website. The other was to download the data files first, then uploaded to Github so Python can access.

As said in previous paragraph, data were acquired from various sources, bringing many issues ranging from different format, different naming and so forth. It is very tedious to describe what had done in the data cleaning. Instead, data cleaning part is detailed in the accompanying Jupiter Notebook code part with detailed comments & annotations. Our goal was to clean the data and combined them into one table. Here is the list of the major steps involved:

* Work out official residential neighborhood table, which should cover the entire City, from the City center to every suburbs.
* Adding social and economical data to the neighborhood table.
* Appending crime/disorder data to the neighborhood table,
* Appending central location in geographic coordinate format (latitude and longitude) to the neighborhood candidate tables
* Querying indoor recreation facilities from Foursquare API only. Check the code part for how this was implemented in detail. However, it is worth of mentioning about the key point in this process: to get only the indoor recreation venues. Our business is to construct a new indoor young child playing center or facility, so the rest outdoor or a lot of Foursquare location venues are not related to our business goal at all. Researches and trials have been performed during this project over the Foursquare so that there is no deviation from our business requirements.

The below shows data features and searched venue in one neighborhood after data were concatenated into one final table:

The main table has 19 features, mainly including indoor recreational venues but also social & economic data as well. This table has about 200 neighborhoods.

As mentioned before, we also constructed a second dataframe table. It is covering the entire City but only have 8 official Sectors, namely as of East, West, Central, North, South, Northeast, Southeast, Northwest. We have summed or averaged all included communities in each Sector to work out those features in this table. This table served us for some viewing and qualitative analysis.

## **3. Data analysis & methodology**

3.1 Neighborhood candidates and existing indoor recreation facilities:

Neighborhood candidates are displayed in blue circles in Figure 1 below and those queried indoor recreation facilities are displayed in red dots. We can see that our neighborhood candidates are covering the entire City, including the center and all suburbs, which is exactly our business outlined.

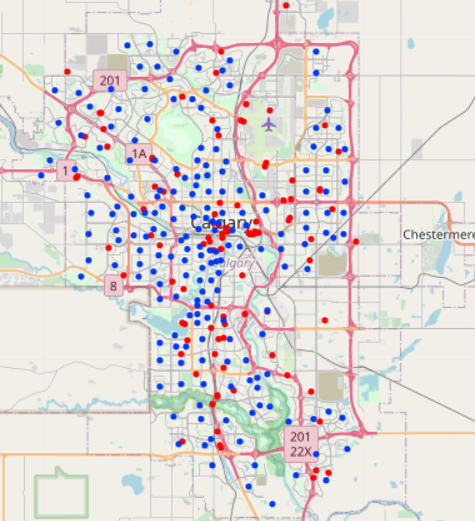
On the other hand, those existing indoor recreation centers plotted in red circles as in Figure2, are not evenly located in the City. It seems some zones or neighborhoods are better for constructing a new indoor child playing center, such as south, southeast. 

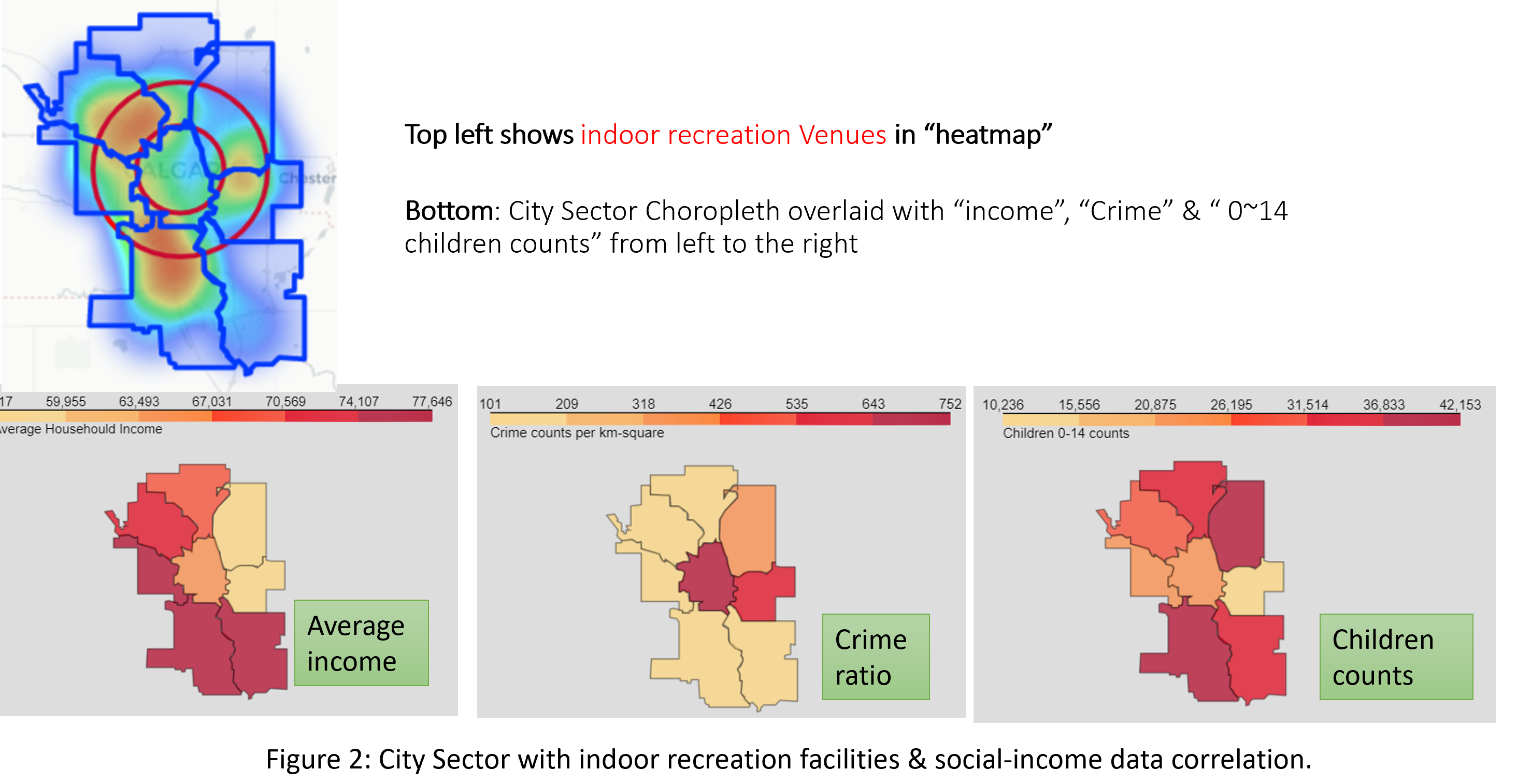
Figure 1: Neighborhood candidates (Blue dots) and nearby indoor recreation facilities in Red dots

3.2 Exploratory data analysis

Recommending a business decision is a complex issue because it involves social and economical consequences. Our methodology is to use data driven method, both qualitatively and quantitively.

Let us look at data we have in a qualitative way.

We can present the indoor recreation facilities by using “Heatmap” as Figure 2. Heatmap is a different data visualization technique that shows the magnitude of our recreation venue distribution in different dense of color. This is better than just plotting them as a dot on the map as we have seen on Figure 1. This information can help to point out the lower density of indoor playing centers. In addition, once we loaded more economical data, it indicated the potential neighborhood would be limited to the south part of the City as per business criteria outlined in previous chapter. With money as a constraint factor as always, our preference would be focused on south part of the City by taking all feature qualitatively.



3.3 Quantitative methodology applied to all data features & their results

There are basically 4 steps applied in this part:

* Normalize all 16 data features with Z-score method
* Secondly assign different “weight factor” to different features, according to our outlined business criteria
* Calculate the total score for each neighborhood
* K-means clustering the neighborhoods

The combined data table has total 16 features, 13 of them are related to indoor recreation venues while the other 3 are young children population, income and crime rate. The numerical data value range of all these features are ranging from 0 up to 100,000 as each has different unit. For example, the income is 5 digital numbers with dollar unit, but swimming pool venue may just be around 1.2 pure number. The goal of data normalization is to change the value of numerical columns to a common scale. In our dataset, Z-score calculation is applied to all data feature so we leveled all features without any “discrimination”.

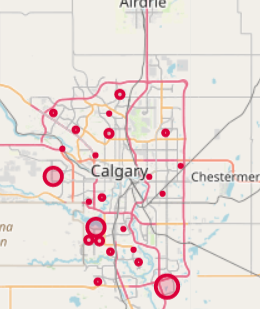
After Z-sore normalization applied, our next unique step was to apply customized scaling, as per the requirements of business set out in the introduction part. To meet our business needs, higher important factors (i.e, higher weight) had be assigned to the neighborhoods with high density of children. Similarly, if there is existing indoor play center nearby, we would assign a very lower weight accordingly since we have no desire to build playing facility nearby at all. Weight is in range of -1 to positive 1, indicating total negative to absolute positive correlation. Once we had done this application, we calculated the total scale so that

Figure 3, recommended future neighborhood in big circles

Neighborhoods could be compared to each other so preferable communities can be inferred t construct a playing center. Figure 3 shows these results. It can be inferred from Figure 3 that the preferable neighborhoods to start a new indoor playing center would be Southwest and Southeast.

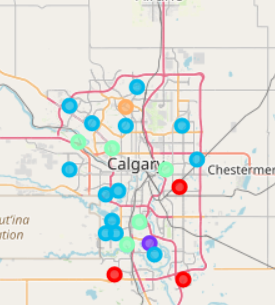
 Finally, unsupervised K-means algorithm was selected and applied for neighborhood clustering. The main reason is that there are many common indoor recreation venues for our City Sector or neighborhoods but there are no target variables. Figure 4 shows the clustering results.

Figure 4, Neighborhood K-Means clustering results.

This red cluster is the cluster with no or lower density of recreation facility zone but there are lot of young family with young kids. Their data properties could be seen below in Figure 5.

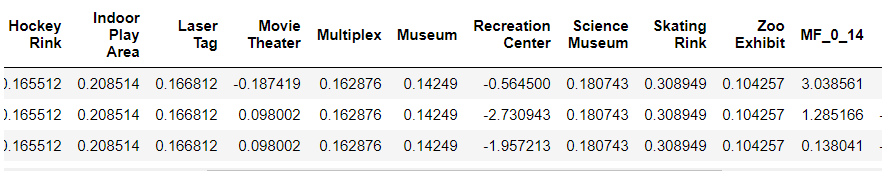


Figure 5, Feature values for the Red cluster neighborhoods of Figure 4.

4: Conclusion and discussion.

In this capstone project, I have selected all the neighborhoods from the City of Calgary and extensively searched all recreation venues in very large radius. Secondly, I also use all social & economic data, closely related to our business needs, from the City Census database. Therefore, we have a good grasp about not only the recreation facility distributions but also more business-oriented features (specially, number of young children, income and crime).

Our goal is that we are searching low or very low-density zones with indoor recreation but good social-economical status in the City.

Scientifically, 19 data **features** were selected we normalized the data and then we applied business important factor (or called “weight”) to each feature before doing quantitively analysis.

Figure 2 summarized our qualitative analysis results; the less density zones of recreation facilities are in West & South part of the City. Figure 3 ~4 depict those analysis results in quantitative way. These results are pointing to **deep South West & South East** of the City for the indoor playground investment.

Therefore, we recommended to build indoor playground in **deep South East, South West** of the city. Specific neighborhoods are called Abbeydale, Bridlewood, Erin Woods, where there are a lot of young families.

In fact, in the last couple of years, the City have built a large billion-dollar hospital and recently a light rail train line is being constructed to the south east sectors from City center. This makes sense because that area is the under rapid developments in the last 10 years.

5: Future improvement.

During this project, I was able to collect various data to do the analysis such as location data, geographic, social data. However, these data qualities are not the same. For example, the household income was hard to find and the data we got may not reflecting the true nature of current status.

Secondly, different selection of various features can lead to slightly different results. There are different options to choose from as always.

Finally, as time goes by, searched Venues from Foursquare API can be changed a lot, in additional to the changes in term of venue category definition. My free API query results may not have all the information or have them correctly categorized. So, a fee-paid-account to Foursquare API will improve the results. Anyway, the education purpose is well served with the free account totally.

Importantly, it should be noted that recommended zones or locations shall be treated only as a starting point since more detailed analysis & new constructed facilities (may not present in Foursquare API search list) will change results in a possible big way.

Thank you.