IBM Coursera Applied Data Science Capstone Project

The Battle of Neighborhoods

Sridharan Sadagopan | IBM Coursera Applied Datascience | June 16, 2020

Contents

IBM Coursera Applied Data science Capstone Project

Introduction

Background

Our imaginary client is a books and stationary seller based in Singapore. They would want expand their business to cities in nearby countries such as Malaysia, Indonesia, Thailand, Philippines and etc. Opening a new bookstore in a new city requires selection of the right business location with reachability for the right target customers, competition landscape and etc.

Business Problem

Our client wants us to analyse and recommend top locations for opening bookstores in those cities.

Selection of the right neighbourhood such as the ones with large number of schools, universities, shopping malls and etc is important for the bookstore business to be successful. It would also be important to analyse information on those neighbourhood about the existing bookstores to understand the competition.

As a start the results of the recommendation would need to be presented for the citiy Kuala lumpur, the capital of Malaysia the top 5 recommended locations for the bookstores.

Data Understanding

To help up us in this process we would be using the different data about venues in the locations of interest. Most important data from Foursquare API that we would depend on is the Venue Categories.

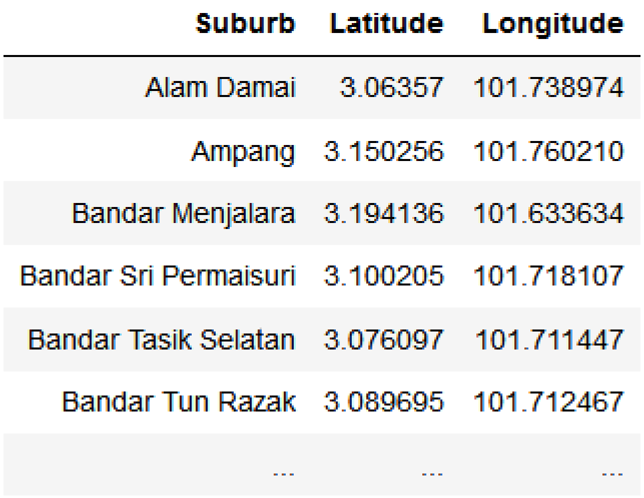
Prior to using Foursquare API, we would be getting the information about the list Suburbs/Neighbourhood and their geolocation coordinates.

For example we would getting the list of suburbs from Wikipedia for Kuala Lumpur (one of the city of interest) from the page https://en.wikipedia.org/wiki/Category:Suburbs_in_Kuala_Lumpur

Using BeautifulSoup python library we extract the suburbs of Kuala lumpur as follows:

- 'Alam Damai',
- 'Ampang, Kuala Lumpur',
- 'Bandar Menjalara',
- 'Bandar Sri Permaisuri',
- 'Bandar Tasik Selatan',
- 'Bandar Tun Razak',
- 'Bangsar'
- ...

Then using geopy.geocoders 'Nominatim' we would get the geolocation of these suburbs:



Suburb	Latitude	Longitude
Alam Damai	3.06357	101.738974
Ampang	3.150256	101.760210
Bandar Menjalara	3.194136	101.633634
Bandar Sri Permaisuri	3.100205	101.718107
Bandar Tasik Selatan	3.076097	101.711447
Bandar Tun Razak	3.089695	101.712467
...	...	...

With the Name, Latitude, Longitude we will proceed with the FourSquare API find the venues of interest (Educational Institutions) as follows:

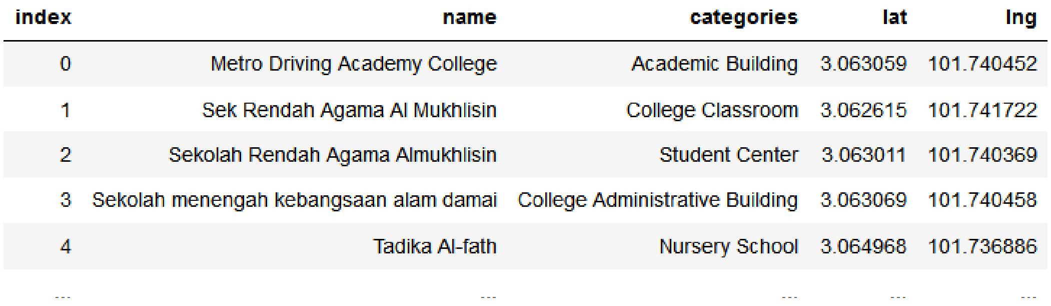
For example Fouresquare API returns the Venue categories such as schools, colleges and universities:

- School
    - Adult Education Center
    - Circus School
    - Cooking School
    - Driving School
    - Elementary School
    - Flight School
    - High School
    - Language School
    - Middle School
    - Music School
    - Nursery School
    - Preschool
    - ….
- College & University
    - Community College
    - Fraternity House
    - General College & University
    - Law School
    - Medical School
    - Sorority House
    - Student Center
    - Trade School
    - University

Using the category ids as defined by FourSquare API [https://developer.foursquare.com/docs/build-with-foursquare/categories] we can filter the venues of interest by specifying catergory of interest in the search/explore queries of FourSquare API. e.g Category Ids from FourSquare:

- College_iniversity='4d4b7105d754a06372d81259'
- Library='4bf58dd8d48988d12f941735'
- School='4bf58dd8d48988d13b941735'
- Shopping_mall='4bf58dd8d48988d1fd941735'
- Shopping_plaza='5744ccdfe4b0c0459246b4dc'

We would limit the venues that fall under the specified category as follows



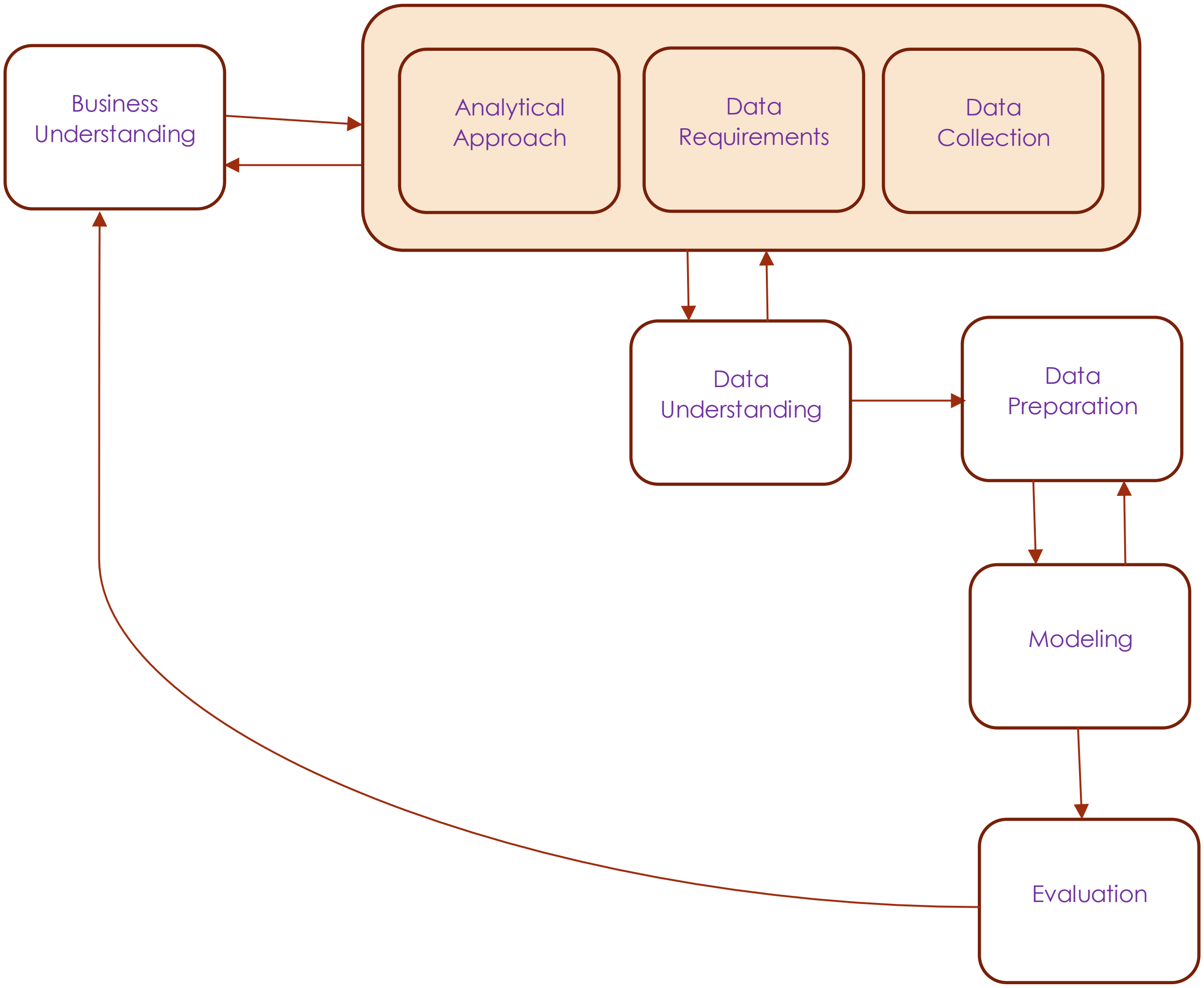
index	name	categories	lat	lng
0	Metro Driving Academy College	Academic Building	3.063059	101.740452
1	Sek Rendah Agama Al Mukhlisin	College Classroom	3.062615	101.741722
2	Sekolah Rendah Agama Almukhlisin	Student Center	3.063011	101.740369
3	Sekolah menengah kebangsaan alam damai	College Administrative Building	3.063069	101.740458
4	Tadika Al-fath	Nursery School	3.064968	101.736886
...	...	...	...	...

With the venues data returned by FourSquare API, we would be able to query the necessary nearby venues data for the neighbourhoods and proceed with data exploration and analysis. With problem to approach clearly defined and with these data that can be retrieved using Foursquare API, data requirements and correct sources of data for this project are understood. The next steps of data science methodology Data Understanding, Data Preparation, Modeling, Evaluation and Potential Deployment.

Methodology

As we learned from IBM Data Science Methodology course, we would be following the same methodology.

We started with a clear business understanding of our objectives. Based on the business understanding our objective to find neighborhoods that exhibit the desirable characteristics for our business location. So we would need to identify groups of neighborhoods and those neighborhoods are not yet labeled, so the problem that we are solving is clearly a clustering problem, so analytical approach that we would be talking would use algorithms such as KMeans clustering algorithm.



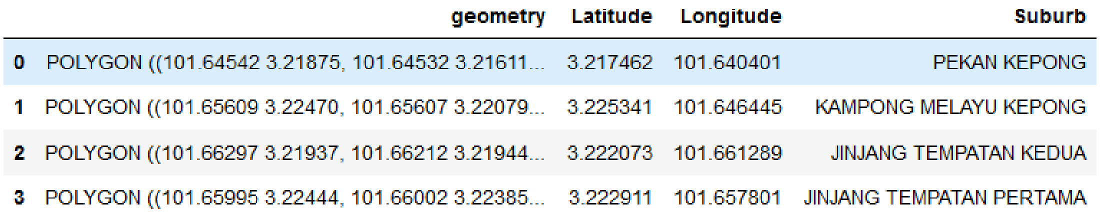
Then we proceeded to look at the kind of data we need, the source of data and how we collect them. In the subsequent sections we proceed to explore the data, prepare and start to build model and evaluate the results.

Data Analysis

With the ability to get access to the data about the venues of interest (Educational Institutions and shopping malls) we proceed to analyse the how these venues are distributed over the neighbourhoods.

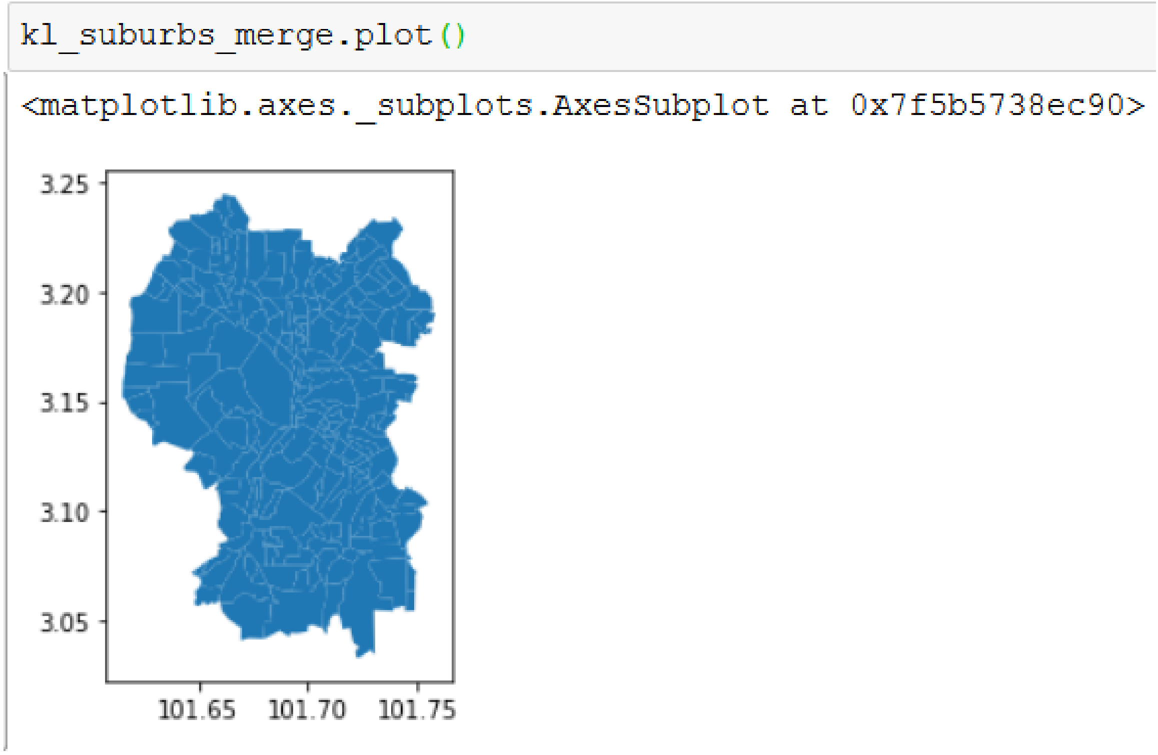
Let us pull the geojson data for Kuala lumpur from https://raw.githubusercontent.com/TindakMalaysia/Federal-Territories-Maps/master/KL/2016/MAP/MIGRATED/result/09-WPKL-New-DM-4326.geojson and read it using Geopanda package.

After dropping and renaming the columns we data fame with name of suburb/neighbourhood , latitude, longitude and its geometry polygon as follows.



	geometry	Latitude	Longitude	Suburb
0	POLYGON ((101.64542 3.21875, 101.64532 3.21611...	3.217462	101.640401	PEKAN KEPONG
1	POLYGON ((101.65609 3.22470, 101.65607 3.22079...	3.225341	101.646445	KAMPONG MELAYU KEPONG
2	POLYGON ((101.66297 3.21937, 101.66212 3.21944...	3.222073	101.661289	JINJANG TEMPATAN KEDUA
3	POLYGON ((101.65995 3.22444, 101.66002 3.22385...	3.222911	101.657801	JINJANG TEMPATAN PERTAMA

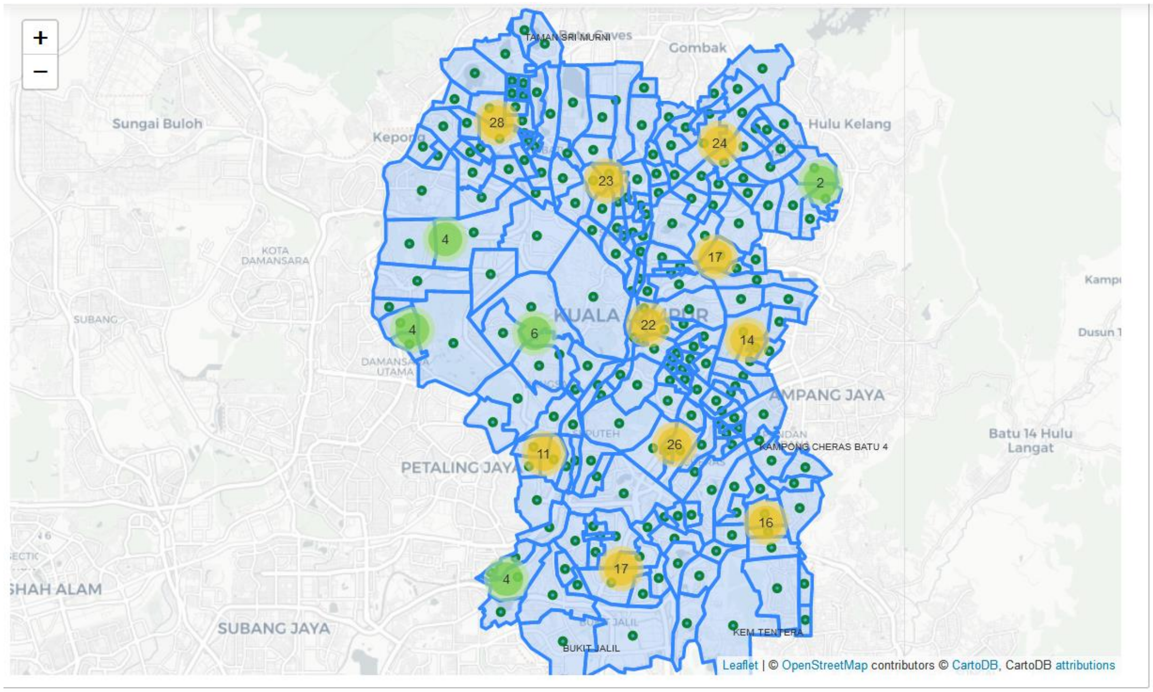
We could do a quick visualization using Geopandas



```
kl_suburbs_merge.plot()
```

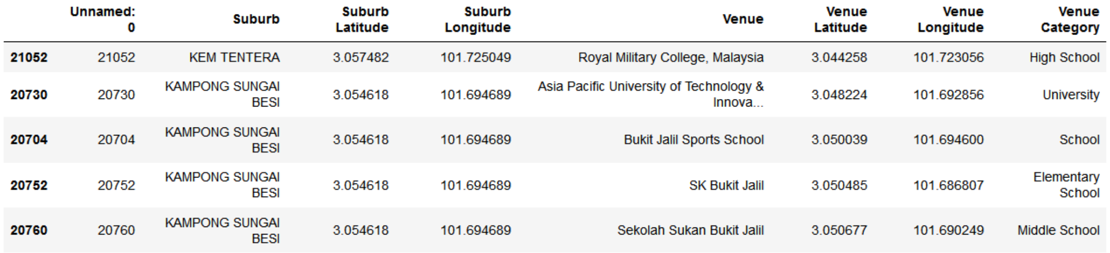
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f5b5738ec90>
```

Using Folium a map of Kuala lumper neigborhoods as shown in the map below:



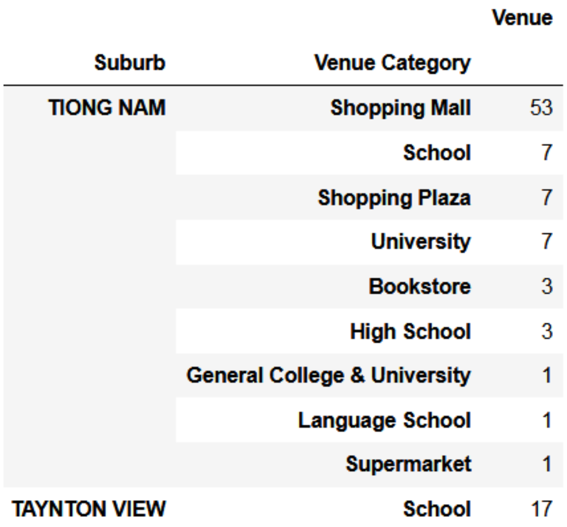
Now we have all the information that we need to proceed explore the neighbourhoods of Kuala lumpur by using geo location of the suburbs, we get the nearby venues of the neighbourhoods of Kuala lumpur using the FourSquare API.

We retrieve the nearby venues of the neighbourhoods of Kuala lumpur and explore the type of venues nearby.



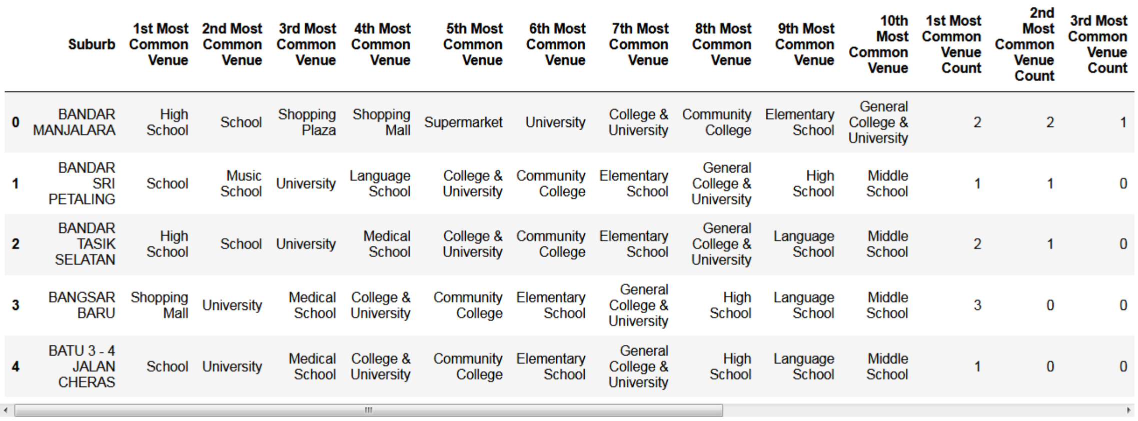
	Unnamed: 0	Suburb	Suburb Latitude	Suburb Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
21052	21052	KEM TENTERA	3.057482	101.725049	Royal Military College, Malaysia	3.044258	101.723056	High School
20730	20730	KAMPONG SUNGAI BESI	3.054618	101.694689	Asia Pacific University of Technology & Innova...	3.048224	101.692856	University
20704	20704	KAMPONG SUNGAI BESI	3.054618	101.694689	Bukit Jalil Sports School	3.050039	101.694600	School
20752	20752	KAMPONG SUNGAI BESI	3.054618	101.694689	SK Bukit Jalil	3.050485	101.686807	Elementary School
20760	20760	KAMPONG SUNGAI BESI	3.054618	101.694689	Sekolah Sukan Bukit Jalil	3.050677	101.690249	Middle School

Let us summarize the count of the categories of venues near neighbourhoods



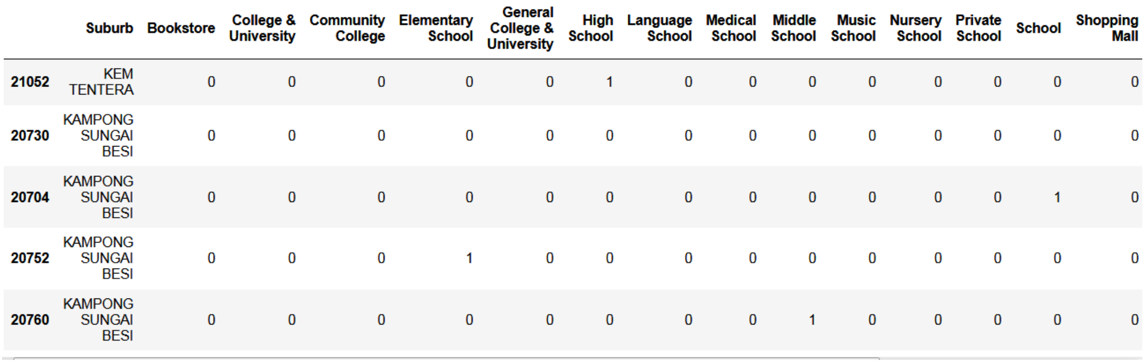
		Venue
Suburb	**Venue Category**	
TIONG NAM	**Shopping Mall**	53
	School	7
	Shopping Plaza	7
	University	7
	Bookstore	3
	High School	3
	General College & University	1
	Language School	1
	Supermarket	1
TAYNTON VIEW	**School**	17

Next we will proceed to summarize the top common venue category in the suburbs



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue	1st Most Common Venue Count	2nd Most Common Venue Count	3rd Most Common Venue Count
0	BANDAR MANJALARA	High School	School	Shopping Plaza	Shopping Mall	Supermarket	University	College & University	Community College	Elementary School	General College & University	2	2	1
1	BANDAR SRI PETALING	School	Music School	University	Language School	College & University	Community College	Elementary School	General College & University	High School	Middle School	1	1	0
2	BANDAR TASIK SELATAN	High School	School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	2	1	0
3	BANGSAR BARU	Shopping Mall	University	Medical School	College & University	Community College	Elementary School	General College & University	High School	Language School	Middle School	3	0	0
4	BATU 3 - 4 JALAN CHERAS	School	University	Medical School	College & University	Community College	Elementary School	General College & University	High School	Language School	Middle School	1	0	0

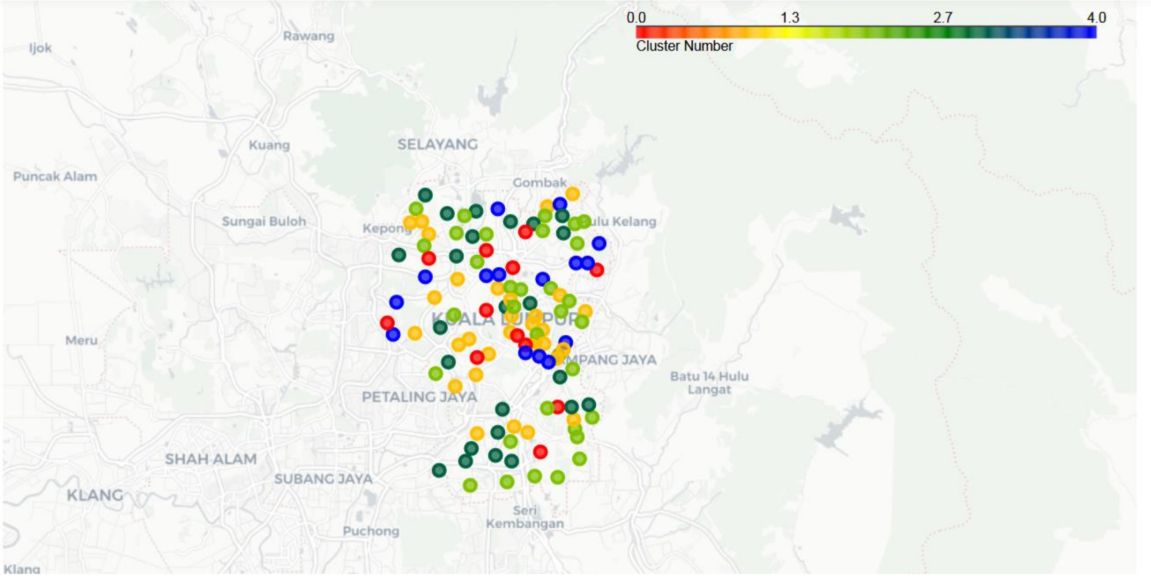
We create one hot encoding of venues for each neighbourhood



	Suburb	Bookstore	College & University	Community College	Elementary School	General College & University	High School	Language School	Medical School	Middle School	Music School	Nursery School	Private School	School	Shopping Mall
21052	KEM TENTERA	0	0	0	0	0	1	0	0	0	0	0	0	0	0
20730	KAMPONG SUNGAI BESI	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20704	KAMPONG SUNGAI BESI	0	0	0	0	0	0	0	0	0	0	0	0	1	0
20752	KAMPONG SUNGAI BESI	0	0	0	1	0	0	0	0	0	0	0	0	0	0
20760	KAMPONG SUNGAI BESI	0	0	0	0	0	0	0	0	1	0	0	0	0	0

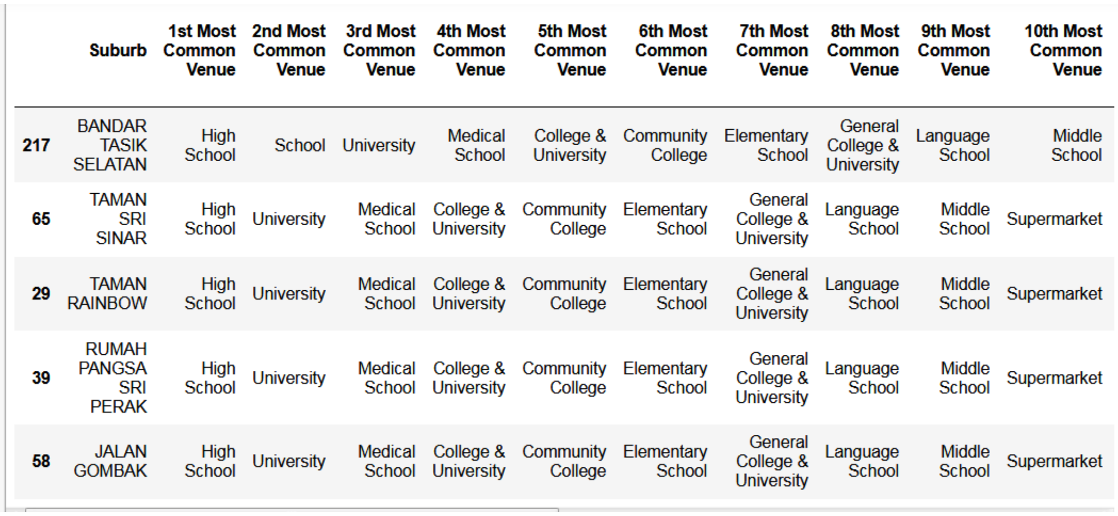
With top common venue category per suburb and one hot encoding of venue categories available we will proceed with clustering (KMeans) to cluster the neighbourhood understand the similarity pattern among the suburbs. With the number of clusters set to 5, the following map is plotted to visualize clusters.

.



Let us study further on the clustered suburbs further to understand the pattern.

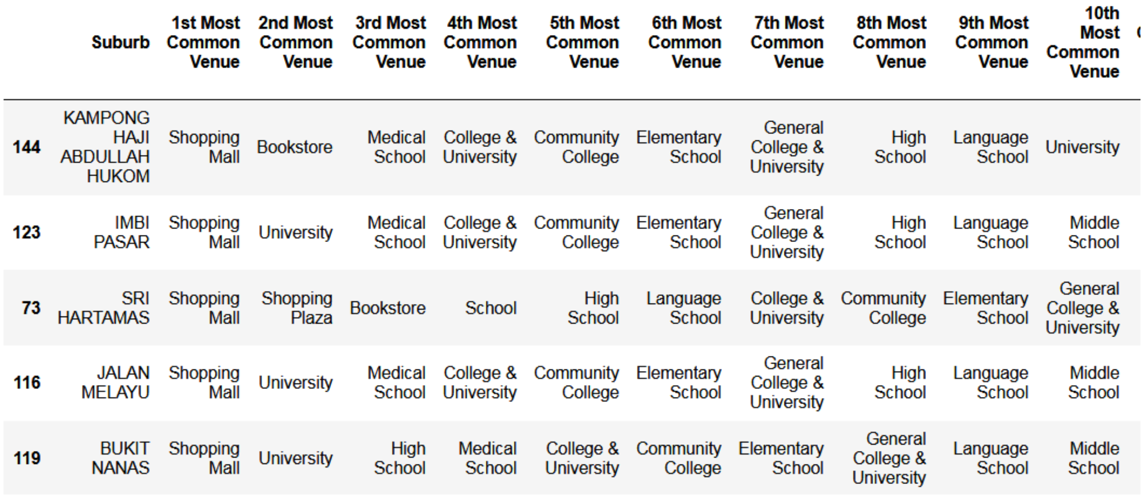
Cluster 0:



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
217	BANDAR TASIK SELATAN	High School	School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School
65	TAMAN SRI SINAR	High School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	Supermarket
29	TAMAN RAINBOW	High School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	Supermarket
39	RUMAH PANGSA SRI PERAK	High School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	Supermarket
58	JALAN GOMBAK	High School	University	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	Supermarket

Most common venues of interest in Cluster 0 seems to be “High School”.

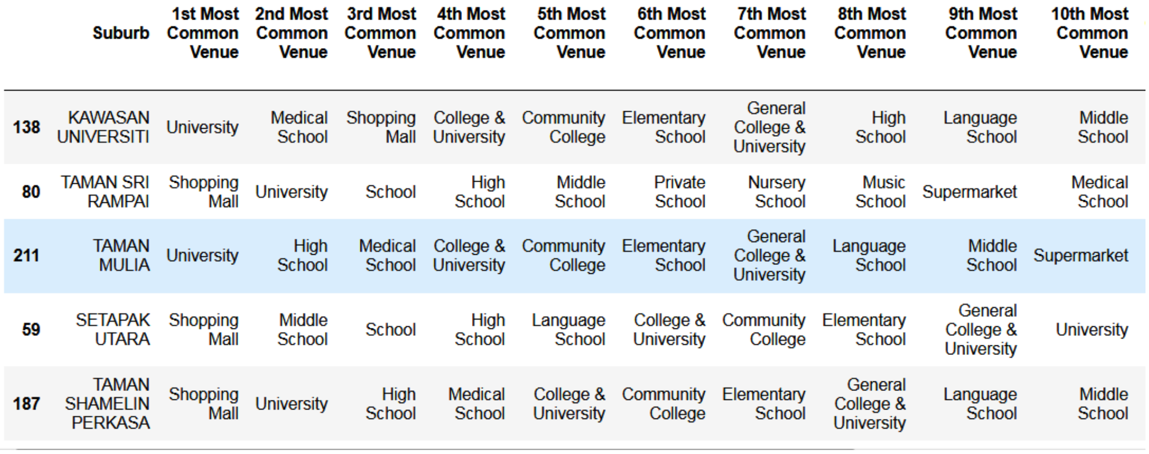
Cluster 1:



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
144	KAMPONG HAJI ABDULLAH HUKOM	Shopping Mall	Bookstore	Medical School	College & University	Community College	Elementary School	General College & University	High School	Language School	University
123	IMBI PASAR	Shopping Mall	University	Medical School	College & University	Community College	Elementary School	General College & University	High School	Language School	Middle School
73	SRI HARTAMAS	Shopping Mall	Shopping Plaza	Bookstore	School	High School	Language School	College & University	Community College	Elementary School	General College & University
116	JALAN MELAYU	Shopping Mall	University	Medical School	College & University	Community College	Elementary School	General College & University	High School	Language School	Middle School
119	BUKIT NANAS	Shopping Mall	University	High School	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School

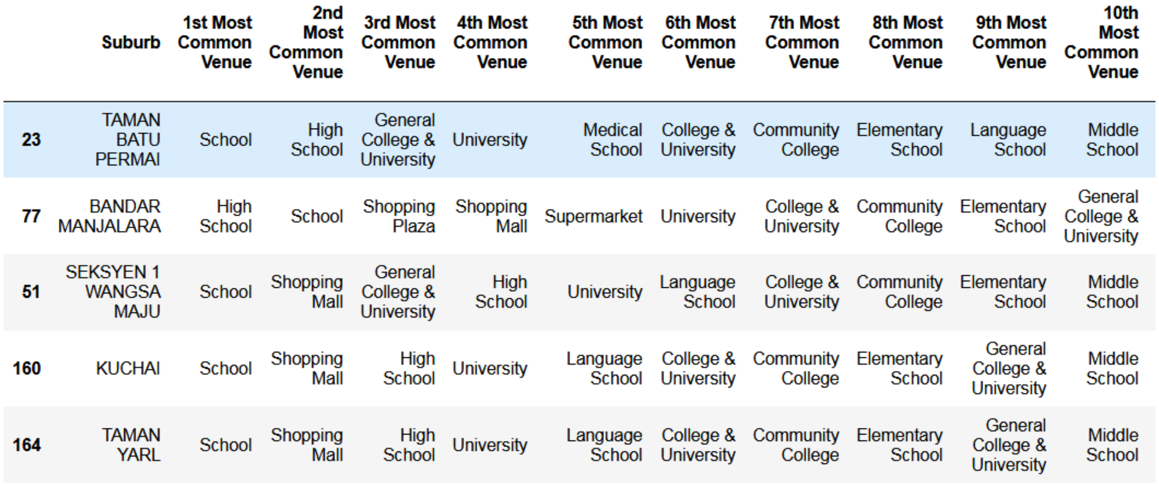
Most common venues of interest in Cluster 1 seems to be “Shopping Mall”.

Cluster 2:



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
138	KAWASAN UNIVERSITI	University	Medical School	Shopping Mall	College & University	Community College	Elementary School	General College & University	High School	Language School	Middle School
80	TAMAN SRI RAMPAI	Shopping Mall	University	School	High School	Middle School	Private School	Nursery School	Music School	Supermarket	Medical School
211	TAMAN MULIA	University	High School	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School	Supermarket
59	SETAPAK UTARA	Shopping Mall	Middle School	School	High School	Language School	College & University	Community College	Elementary School	General College & University	University
187	TAMAN SHAMELIN PERKASA	Shopping Mall	University	High School	Medical School	College & University	Community College	Elementary School	General College & University	Language School	Middle School

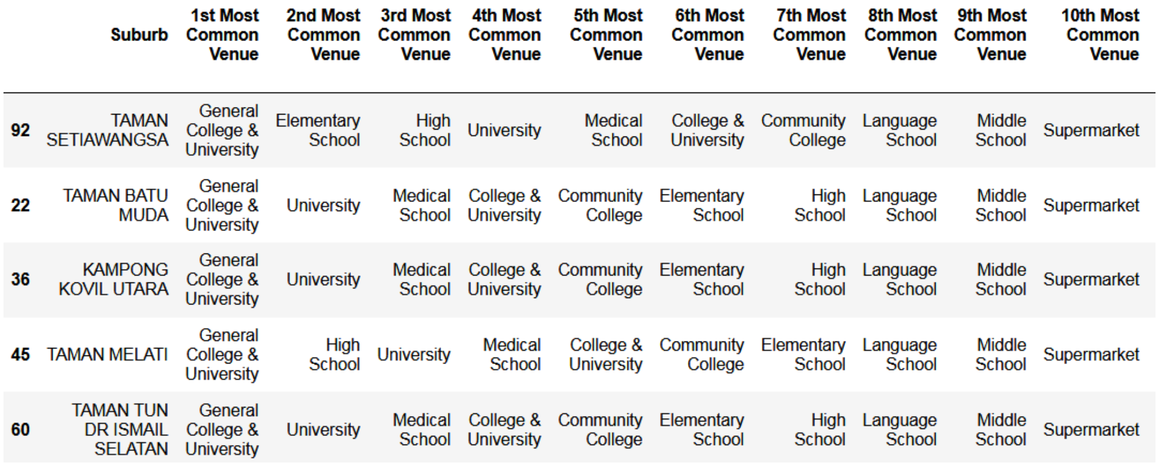
Cluster 3:



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	TAMAN BATU PERMAI	School	High School	General College & University	University	Medical School	College & University	Community College	Elementary School	Language School	Middle School
77	BANDAR MANJALARA	High School	School	Shopping Plaza	Shopping Mall	Supermarket	University	College & University	Community College	Elementary School	General College & University
51	SEKSYEN 1 WANGSA MAJU	School	Shopping Mall	General College & University	High School	University	Language School	College & University	Community College	Elementary School	Middle School
160	KUCHAI	School	Shopping Mall	High School	University	Language School	College & University	Community College	Elementary School	General College & University	Middle School
164	TAMAN YARL	School	Shopping Mall	High School	University	Language School	College & University	Community College	Elementary School	General College & University	Middle School

For cluster 3, School seems to be the most common venue of interest followed by some High Schools.

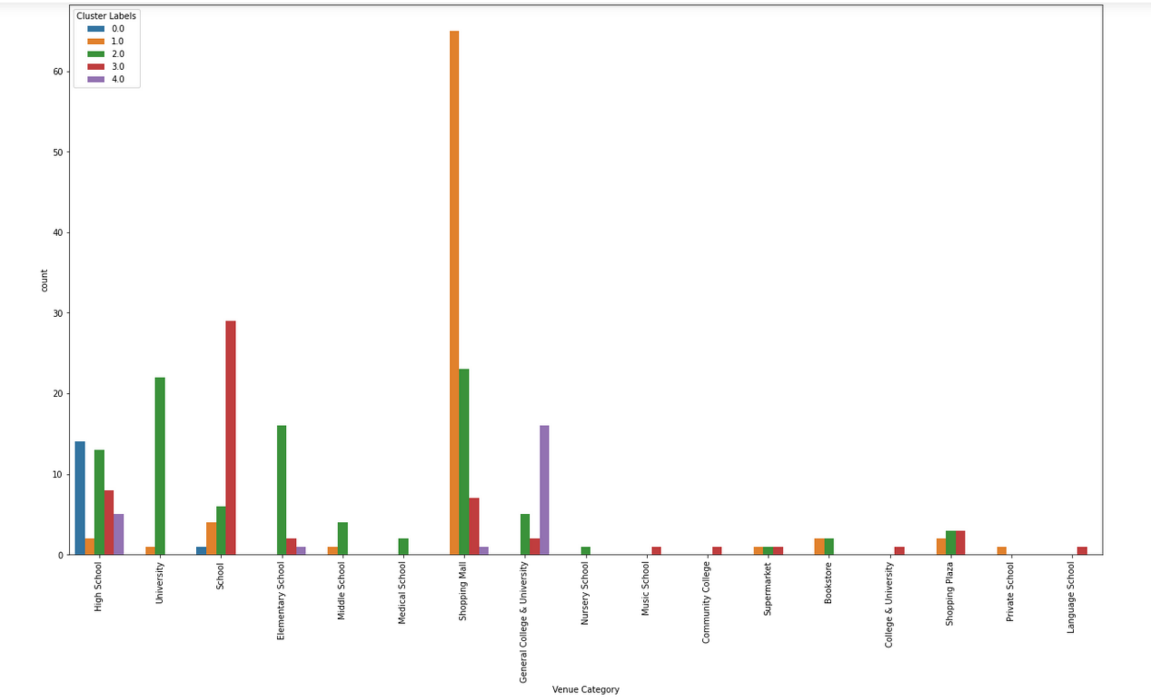
Cluster 4:



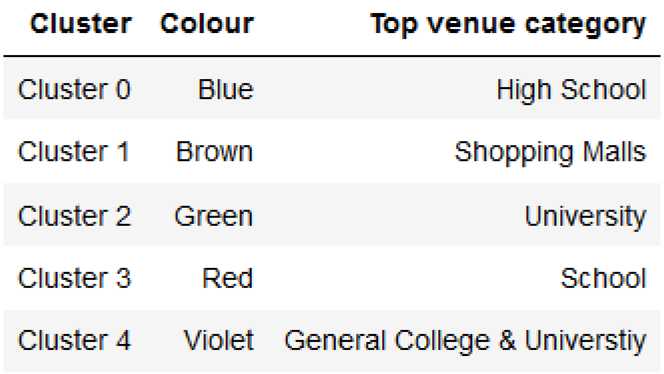
	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
92	TAMAN SETIAWANGSA	General College & University	Elementary School	High School	University	Medical School	College & University	Community College	Language School	Middle School	Supermarket
22	TAMAN BATU MUDA	General College & University	University	Medical School	College & University	Community College	Elementary School	High School	Language School	Middle School	Supermarket
36	KAMPONG KOVIL UTARA	General College & University	University	Medical School	College & University	Community College	Elementary School	High School	Language School	Middle School	Supermarket
45	TAMAN MELATI	General College & University	High School	University	Medical School	College & University	Community College	Elementary School	Language School	Middle School	Supermarket
60	TAMAN TUN DR ISMAIL SELATAN	General College & University	University	Medical School	College & University	Community College	Elementary School	High School	Language School	Middle School	Supermarket

For cluster 4, General College & University seems to be the most common venue of interest.

Now let us plot a comparative view of all the clusters with the venue categories.



This histogram show the count of top venues categories for different clusters. Based on the above histogram, we could see a pattern of Clusters as follows



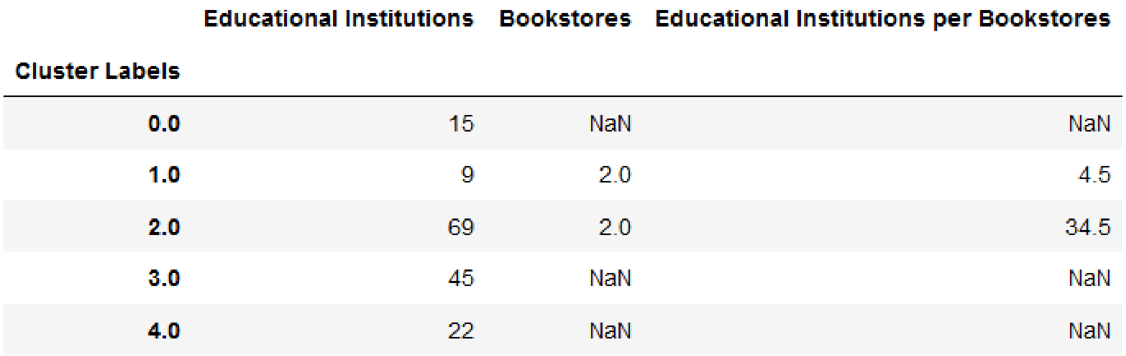
Cluster	Colour	Top venue category
Cluster 0	Blue	High School
Cluster 1	Brown	Shopping Malls
Cluster 2	Green	University
Cluster 3	Red	School
Cluster 4	Violet	General College & Universtiy

Results

With the understanding of venues of interest and their distribution among the clusters of neighbourhood done in the Data Analysis sections let us continue understand the results.

Let us start to explore the competition landscape in different clusters identified by KMeans clustering algorithm.

The following is data is obtained by filtering the "Book Stores" and "College Bookstores" in the venue categories and summarized.

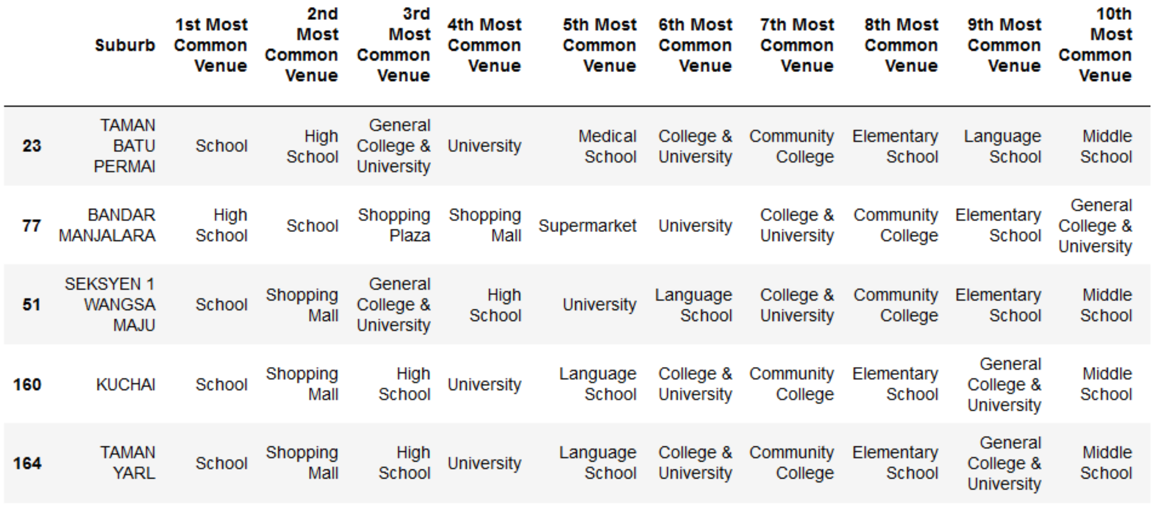


	Educational Institutions	Bookstores	Educational Institutions per Bookstores
Cluster Labels			
0.0	15	NaN	NaN
1.0	9	2.0	4.5
2.0	69	2.0	34.5
3.0	45	NaN	NaN
4.0	22	NaN	NaN

For clusters 1 and 2 there are bookstores in their vicinity, whereas for other clusters no bookstore nearby.

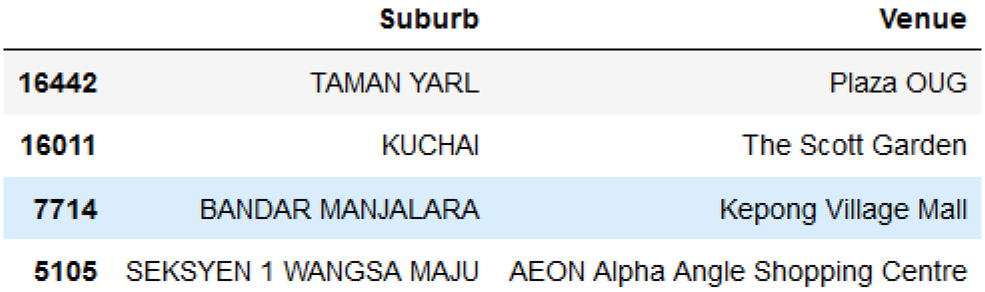
Considering the number educational institutions in cluster 3, this cluster has next higher number but no bookstore in the vicinity.

Let us examine the top 5 suburbs in the cluster 3



	Suburb	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
23	TAMAN BATU PERMAI	School	High School	General College & University	University	Medical School	College & University	Community College	Elementary School	Language School	Middle School
77	BANDAR MANJALARA	High School	School	Shopping Plaza	Shopping Mall	Supermarket	University	College & University	Community College	Elementary School	General College & University
51	SEKSYEN 1 WANGSA MAJU	School	Shopping Mall	General College & University	High School	University	Language School	College & University	Community College	Elementary School	Middle School
160	KUCHAI	School	Shopping Mall	High School	University	Language School	College & University	Community College	Elementary School	General College & University	Middle School
164	TAMAN YARL	School	Shopping Mall	High School	University	Language School	College & University	Community College	Elementary School	General College & University	Middle School

Let us look for shopping malls near these top 5 neighbourhoods



	Suburb	Venue
16442	TAMAN YARL	Plaza OUG
16011	KUCHAI	The Scott Garden
7714	BANDAR MANJALARA	Kepong Village Mall
5105	SEKSYEN 1 WANGSA MAJU	AEON Alpha Angle Shopping Centre

Potentially one of these shopping malls could be the location for consideration of choice for new bookstore.

Discussion

Location based search and exploration using services such FourSquare API, enabled me in analysing and understand a wealth of information about the neighbourhoods of Kuala lumper. One of the challenges I observed is that location based search parameters such as radius of search need to be adapted for how densely the neighbour is populated and how closes the venues are located. In a sparsely populated neighbourhoods the venues tend to be far from each other, so radius of search may need to larger. Other approach I tried to experiment is to apply filtering of the data returned by FourSquare API using use administrative boundary geometric data from such as GeoJSON and etc.

In this capstone project I only used the basic venue details Foursquare API provided, we could potentially augment it with data FourSquare Places Location Data Can Offer. For example we could use place attributes like venue ratings and reviews.

Conclusion

This capstone project provided me a very valuable experience to understand, explore, apply many different concepts of Data Science methodology and tools and services such as FourSquare Local based services.

Though it was possible to use venues of interest and their location to recommend possible locations for Bookstore as required by business understanding and objectives of the requirements, there may be more considerations such as population density, cost of running business and etc would need to be considered.

Probably in the future much more wealth data would be available, making it possible to provide a more accurate recommendations that would take all considerations into account to recommend the top suburbs for the opening bookstores.

Optionally an application can be built to help the business to continue to use the data from Foursquare API such as user ratings, reviews to continue improve the kinds of books and other items that can be stocked.