IBM Professional Data Science Certification

Coursera Capstone Project

On

Exploring Educational Institutions in a neighbourhood

By

Smruthy G

8th March 2020

1. Introduction

1.1 Background

Educational Institution is a place where the young minds are shaped into smart individuals. These institutions have grown manifold in most prominent cities across the globe. Getting into the right institution means better knowledge and better outgrowth in career. But there are lot of challenges in determining the right institution. One such factor is the commuting distance. Today, considering the schools, there are variety of options to choose. But those options may not be preferred by the people. A person can easily find a set of school in their area. But it is just a cluster of data with which a person might not be satisfied. Therefore, an effective data model is required to predict a list of schools within a neighbourhood that matches the user preferences.

1.2 Problem

Imagine if people have the option to explore schools that are nearby to their **Home location** as well as has a **good rating**. Based on these inputs, they could get a variety of schools to choose from. The outcome of this project is to provide a list of options for the people which will ease up the process of finding the best school in their neighbourhood.

1.3 Interests

Longer the distance, longer is the travel time and larger the transportation costs. Hence, it is a fair advantage for the people to discover a school, which is nearby to their home and also provides quality education.

2. Data Acquisition

2.1 Data Sources

Dataset on the schools in a neighbourhood can be extracted using the Foursquare API request. From this data, name of the school, rating and the neighbourhood can be used. Another set of data is user input. A table with the list of neighbourhood and rating is created manually.

2.2 Data Usage

A dataset with list of different user preferences on schools is converted to a table. The data will have the below contents.

Location	Rating
Marble Hills, Manhattan	5
York, Manhattan	4

The above dataset is prepared manually by myself. With this data, I get the required details to proceed with API request.

Initially, the latitude and longitude of each location is determined using geocode. These coordinates are used in the URL request. The list of schools from these coordinates are received from the Foursquare API request. Based on the schools listed, the corresponding user ratings are extracted. This list is passed on to another URL request which will extract the user reviews and tips data for each of the schools listed in the previous extract for a particular neighbourhood.

Further, the schools should be categorised based on the distance. Euclidean distance is used to calculate the distance between the home location and school. With this data, I will compare the distance provided by the user and the Euclidean distance to get the final set of schools.

Finally, the data retrieved in the previous step is combined with the first set of data that is extracted from Foursquare to form a final table. This table is further categorized based on the neighbourhood and the same is visualized in the Maps and is available for the User.

2.3 Data Wrangling

The data related to the schools in a city was first extracted from the Foursquare API data set. The city chosen was New York. In the request, a specific search string 'schools' was included to get the exclusive dataset only with the relative school data. Below table was generated.

	categories	hasPerk	id	location.address	location.cc	location.city	location.country	location.crossStreet	location.distance	location.formattedAddress	location.labeledl
0	[{'id': '4bf58dd8d48988d130941735', 'name': 'B...	False	4a78d55cf964a5208be61fe3	15 Barclay	US	New York	United States	btwn Broadway and Church St	194	[15 Barclay (btwn Broadway and Church St.), Ne...	[{'label': 'disp 40.712634338
1	[{'id': '4bf58dd8d48988d198941735', 'name': 'C...	False	4bbfa098f8219c748850b010	163 William St	US	New York	United States	Pace University	276	[163 William St (Pace University), New York, N...	[{'label': 'disp 40.710245400
2	[{'id': '4bf58dd8d48988d1ac941735', 'name': 'C...	False	4d5ab621e2df60fc9f09d4e5	1 Pace Plz	US	New York	United States	btwn Nassau & Gold St	135	[1 Pace Plz (btwn Nassau & Gold St), New York,...	[{'label': 'disp 40.711597788
3	[{'id': '4d4b7105d754a06372d81259', 'name': 'C...	False	4e37fc7918a8470916d00b74	1 Pace Plz	US	New York	United States	btwn Nassau & Gold St	270	[1 Pace Plz (btwn Nassau & Gold St), New York,...	[{'label': 'disp 40.711023720
4	[{'id': '4bf58dd8d48988d13d941735', 'name': 'H...	False	4b980f0ef964a520952935e3	411 Pearl St	US	New York	United States	Madison Street	426	[411 Pearl St (Madison Street), New York, NY 1...	[{'label': 'disp 40.711181831

The above table had few improper formats. Initially, I removed some columns which were not required for my analysis like venue referral id, country etc. Adding on to this, the categories column was an inbuilt array with the venue id and the category as the elements. This column was segregated and a new column was generated for Category. Below school table was generated.

	id	name	location.distance	location.address	location.lat	location.lng	Category
0	4a78d55cf964a5208be61fe3	NYU School of Professional Studies	194	15 Barclay	40.712634	-74.008312	Building
1	4bbfa098f8219c748850b010	Seidenberg School Of CSIS	276	163 William St	40.710245	-74.005968	College Academic Building
2	4d5ab621e2df60fc9f09d4e5	Actors Studio Drama School	135	1 Pace Plz	40.711598	-74.005427	College Theater
3	4e37fc7918a8470916d00b74	Lubin School Of Business	270	1 Pace Plz	40.711024	-74.003732	College & University
4	4b980f0ef964a520952935e3	Murry Bergtraum High School	426	411 Pearl St	40.711182	-74.001385	High School
5	4dbc221c6a23e294ba25a512	ABI School Of Barbering	344	Chambers	40.715366	-74.008153	Trade School
6	4d87abcc3b012d4370dff4c4	New York Law School Bookstore	516	40 Worth St	40.717360	-74.006317	College Bookstore
7	4e739d0bd22d2fd31461a9ec	OPMI Business School	582	116 John St Fl 2	40.707502	-74.005715	School
8	58220621bcf73e24b520809a	Cope school	240	225 Broadway	40.711953	-74.008670	Language School
9	536c174b498e91b15b4fb32d	The Actors Studio Drama School at PACE University	159	NaN	40.714106	-74.005519	General Entertainment
10	4fad804ee4b03d09569cb6b9	The Gym @ Spruce Street School	422	8 Spruce St	40.709118	-74.004463	School
11	50ad52a7e4b0b30cda5d7717	International School of Jewelry & Design	234	60 Reade St	40.714835	-74.006133	Trade School
12	4afc1bfbf964a520e31f22e3	Church Street School for Music and Art	437	74 Warren St	40.715103	-74.010150	School
13	4e68be2fb0fb8e94c7e9c8e8	PS.294 Spruce Street School	203	NaN	40.710967	-74.005353	School

Next, there were few invalid values in the address column. This was removed and the below table was generated.

	id	name	location.distance	location.address	location.lat	location.lng	Category
0	4a78d55cf964a5208be61fe3	NYU School of Professional Studies	194	15 Barclay	40.712634	-74.008312	Building
1	4bbfa098f8219c748850b010	Seidenberg School Of CSIS	276	163 William St	40.710245	-74.005968	College Academic Building
2	4d5ab621e2df60fc9f09d4e5	Actors Studio Drama School	135	1 Pace Plz	40.711598	-74.005427	College Theater
3	4e37fc7918a8470916d00b74	Lubin School Of Business	270	1 Pace Plz	40.711024	-74.003732	College & University
4	4b980f0ef964a520952935e3	Murry Bergtraum High School	426	411 Pearl St	40.711182	-74.001385	High School
5	4dbc221c6a23e294ba25a512	ABI School Of Barbering	344	Chambers	40.715366	-74.008153	Trade School
6	4d87abcc3b012d4370dff4c4	New York Law School Bookstore	516	40 Worth St	40.717360	-74.006317	College Bookstore
7	4e739d0bd22d2fd31461a9ec	OPMI Business School	582	116 John St Fl 2	40.707502	-74.005715	School
8	58220621bcf73e24b520809a	Cope school	240	225 Broadway	40.711953	-74.008670	Language School
9	4fad804ee4b03d09569cb6b9	The Gym @ Spruce Street School	422	8 Spruce St	40.709118	-74.004463	School
10	50ad52a7e4b0b30cda5d7717	International School of Jewelry & Design	234	60 Reade St	40.714835	-74.006133	Trade School

The required data for the analysis was cleaned and was ready for model development. The total shape of the final table was 25 rows and 7 columns.

3. Methodology

3.1 Data Analysis

Map data is basically unlabelled data, which means we cannot have a specific target variable or an independent variable. We were looking only for schools/universities that is nearby for the user. Hence, Exploratory Data analysis was not required. For unlabelled data, clustering techniques were utilised. The objective of the project was to split data in the form of different clusters based on the category obtained in the previously displayed table.

3.2 Clustering Model

Initially, the data had to be more categorized to imply the machine learning algorithms. Basically, looking at the table, I displayed earlier, we may not require latitude, longitude, name of the school for our modelling. Only the category was required.

One hot encoding technique was utilised to convert the categorical variables into numerical variables for data modelling. Upon executing the script, below table was generated.

	Address	Building	Clothing Store	College & University	College Academic Building	College Bookstore	College Classroom	College Theater	Community Center	General College & University	Gym	High School	Language School	Nursery School	Office	School	Trade School
0	15 Barclay	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	163 William St	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	1 Pace Plz	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
3	1 Pace Plz	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
4	411 Pearl St	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
5	Chambers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
6	40 Worth St	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
7	116 John St Fl 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	225 Broadway	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0

For the above data, the most suitable clustering technique was K-Means algorithm. This algorithm was chosen since we required the Euclidean distance between each of the locations. The number of clusters to be generated was set as 4. Upon applying the algorithm to the dataset, the records were split as cluster types mentioned below.

Cluster	Type
0	Others
1	Nursery School
2	High School
3	School

The below table was formed with each row belonging to one of the clusters.

	Cluster Labels	id	name	location.distance	location.address	location.lat	location.lng	Category
0	0	4a78d55cf964a5208be61fe3	NYU School of Professional Studies	194	15 Barclay	40.712634	-74.008312	Building
1	1	4bbfa098f8219c748850b010	Seidenberg School Of CSIS	276	163 William St	40.710245	-74.005968	College Academic Building
2	3	4b980f0ef964a520952935e3	Murry Bergtraum High School	426	411 Pearl St	40.711182	-74.001385	High School
3	0	4d5ab621e2df60fc9f09d4e5	Actors Studio Drama School	135	1 Pace Plz	40.711598	-74.005427	College Theater
4	0	4e37fc7918a8470916d00b74	Lubin School Of Business	270	1 Pace Plz	40.711024	-74.003732	College & University
5	0	4dbc221c6a23e294ba25a512	ABI School Of Barbering	344	Chambers	40.715366	-74.008153	Trade School
6	0	4d87abcc3b012d4370dff4c4	New York Law School Bookstore	516	40 Worth St	40.717360	-74.006317	College Bookstore
7	2	4e739d0bd22d2fd31461a9ec	OPMI Business School	582	116 John St Fl 2	40.707502	-74.005715	School
8	0	58220621bcf73e24b520809a	Cope school	240	225 Broadway	40.711953	-74.008670	Language School
9	2	4a6eff92f964a5202cd51fe3	PS 234 - Independence School	593	292 Greenwich St	40.716253	-74.011296	School
10	0	50ad52a7e4b0b30cda5d7717	International School of Jewelry & Design	234	60 Reade St	40.714835	-74.006133	Trade School
11	1	4b55fac5f964a52023fa27e3	Washington Market School	396	134 Duane St	40.716135	-74.007369	College Academic Building
12	1	4dc3f16ffa76d685cdb95dd6	Lubin School of Business	188	1 Pace Plz	40.711107	-74.005387	College Academic Building

4. Results

4.1 Predictive Modelling:

The main objective of the project is to get the location details from the user and utilise them in the constructed model to get the list of available schools for their location. As stated in the Data Usage section, I created a user list table as shown below.

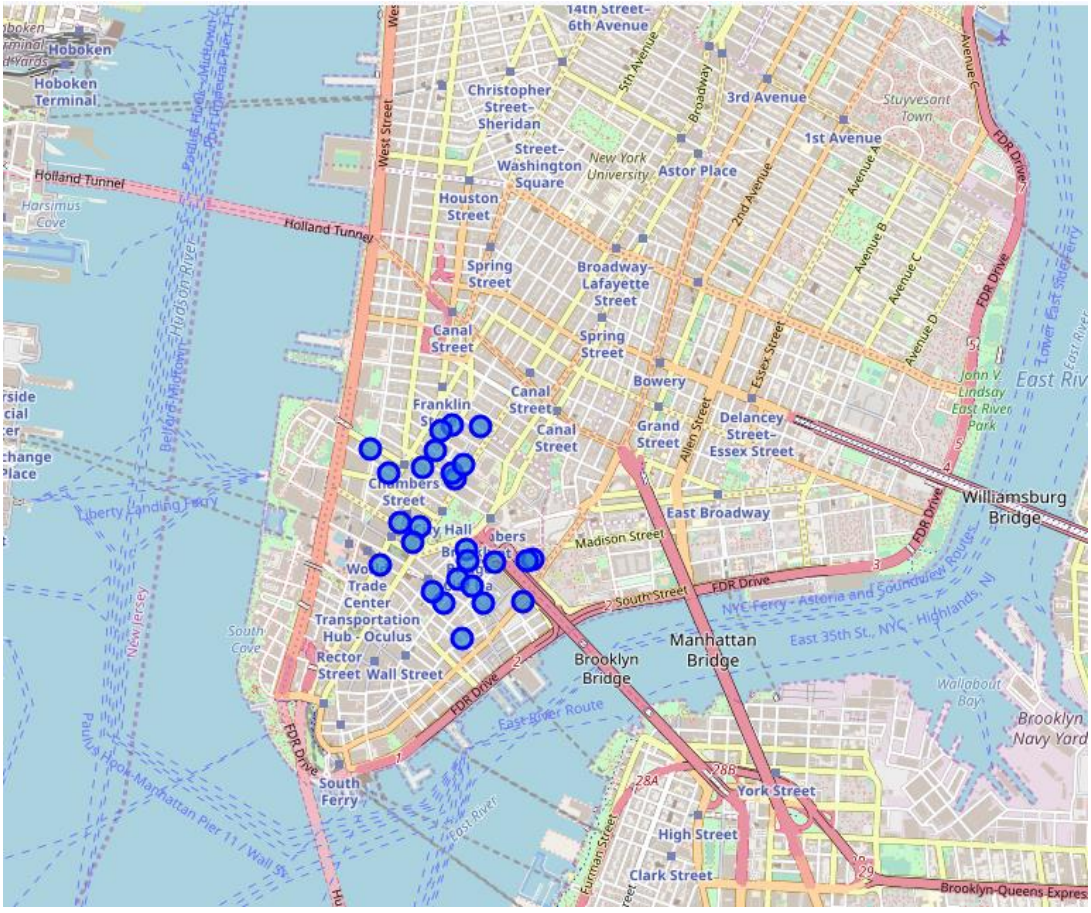
	Home Location	School Rating
0	Barclay	8
1	William St	9
2	Greenwich St	7

Using the Home location, I retrieved the related list of schools for each of the locations specified above in a separate table. The below was the final data table which had the User requested list of schools located in their neighbourhood.

	Category	Cluster Labels	id	location.address	location.distance	location.lat	location.lng	name
0	Building	0	4a78d55cf964a5208be61fe3	15 Barclay	194.0	40.712634	-74.008312	NYU School of Professional Studies
1	College Academic Building	0	4bbfa098f8219c748850b010	163 William St	276.0	40.710245	-74.005968	Seidenberg School Of CSIS
9	High School	2	4c64536211c4a593040ce911	156 William St	318.0	40.709950	-74.005116	Hawthorne Country Day School
10	School	3	4a6eff92f964a5202cd51fe3	292 Greenwich St	593.0	40.716253	-74.011296	PS 234 - Independence School
18	Nursery School	1	4fcf8f59e4b02b4665bbea22	6 Barclay St	295.0	40.712818	-74.009513	Barclay Street School
23	Office	0	53cd804e498ed3afa72314a8	123 William St	401.0	40.709171	-74.006828	Library Journal/ School Library Journal/ Junio...

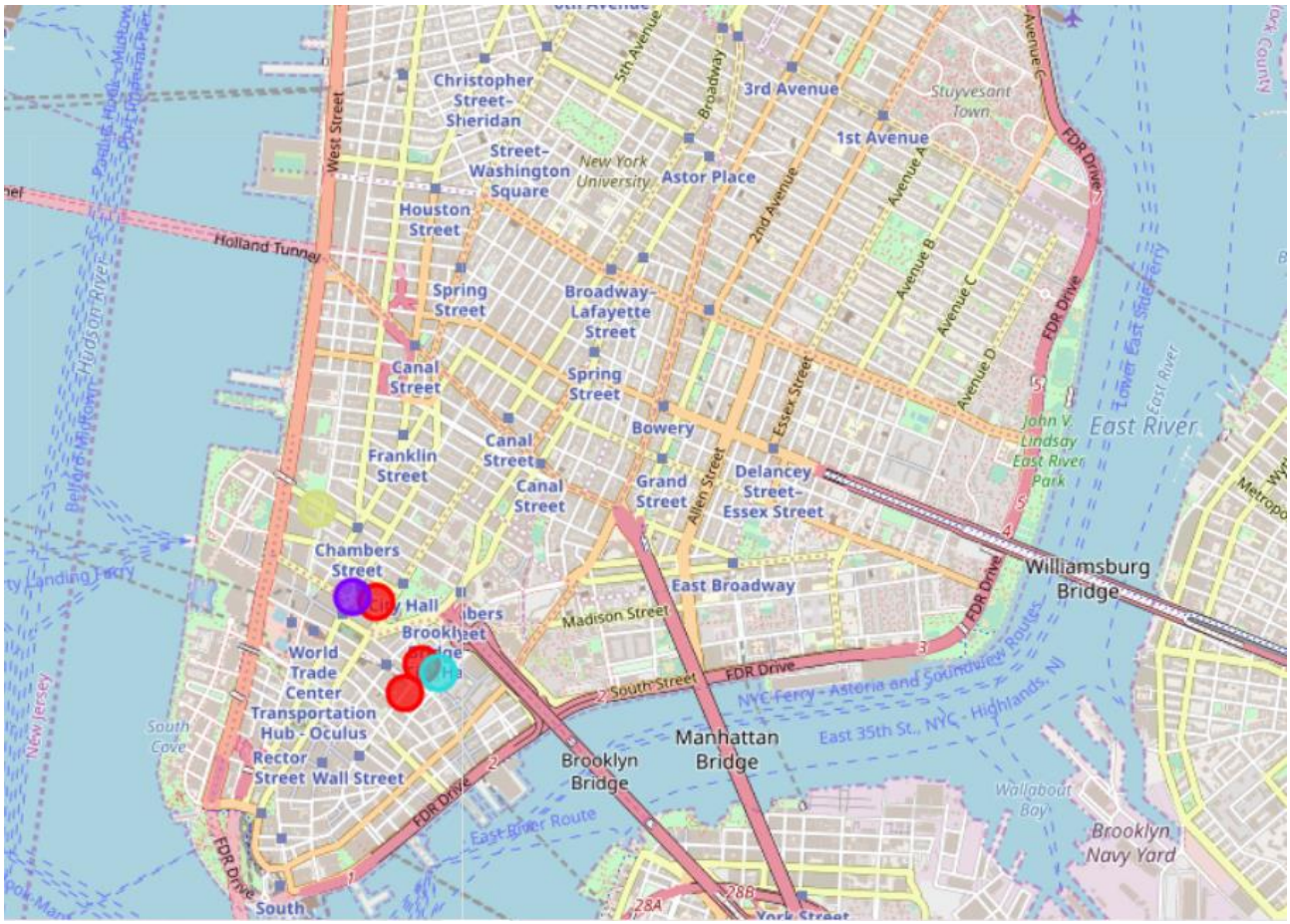
4.2 Data Visualization:

Location based data was visualised using the Folium package. The initial set of data that was retrieved via the Foursquare API request was converted as a Map to visualise the schools within the specified neighbourhood as shown below.



From the above map, it was analysed that most of the schools in the Core centre of New York city were located nearby to each other within distance variations of 0.8 Kms. This was evident to apply the clustering technique to get the places as per the user requirement.

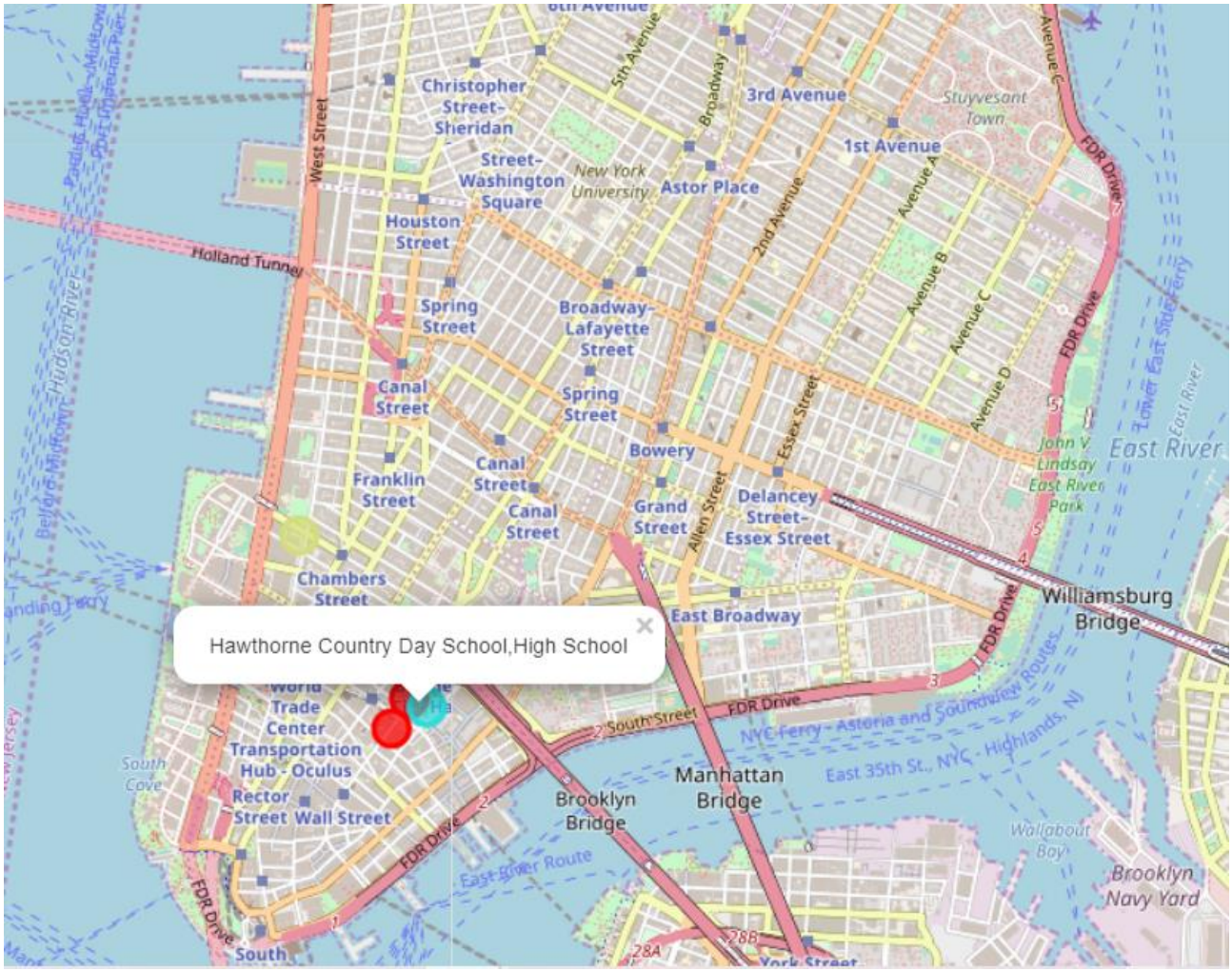
Clustering model was applied and the user data was applied to the model and the below map was generated.



In the above map, we could see the colour coding for each of the cluster. The colour coding is as follows.

Cluster	Colour Code
0	Red
1	Cyan
2	Violet
3	Yellow

Adding on to this, for user readability, popup labels were set up to get details on the name of the school and the type of school as shown below.



5. Discussion

This project had two parameters, that is, distance and rating to get the right set of educational institutions. However, I was able to predict only the distance. Rating could not be used since the Foursquare data did not have ratings for most of the schools. Yet, they had user like counts for each category. But it was difficult for me to use the likes in place of rating since likes were countable and the value was ranging from 1 to beyond 20 and rating is supposed to be from 1 to 10. Due to this limitation, I could not achieve 100% prediction.

6. Conclusion

Hence, in this project, I was able to analyse the various schools within a neighbourhood and predict the schools that would be preferred by the User according to their inputs. Based on the Foursquare API data, I was able to get the list of educational institutions prevailing in New York. Since it was unlabelled data, I used Clustering model. K-Means Algorithm provided the required distances between each cluster. The final outcome of the project to display the schools was achieved using Folium map function. The clusters are displayed at an equivalent distance from each other. This achieves one major parameter which is distance. This model would be helpful to the common people who are looking for good educational institutions in their location and provides a better insight of the options they have within their locality.

7. Future Scope

Further, this model can be enhanced to determine the schools based on likes/rating to provide more efficient results. With the ratings, the model can provide the exact school that matches with user inputs.