Capstone project - Battle of Neighbourhoods

Identification of less Tennis facilitated Neighborhoods in North York, Toronto

1. Introduction/Business Problem

1.1 Background: In preparation for the Summer olympics 2024, Sports Authority of Canada in collaboration with each local government and individual sports Federations assessing the current facilities and planning to start improve the sports facilities all over Canada within two years. Their aim is not just to get medals in Olympics or international competitions, their first priority is to make awarenes among students about the importance of sports activities and how it will help them to keep a healthy body and mind. As the most populated city in Canada, they want to start their piolet project around Toronto. This study is for Tennis Federation of canada who wants to study the current facilities in Toronto and start new Tennis facilities at most needy areas.

1.2 Business Problem: This assignment is to identify the locations for new Tennis facilities in Toronto neighbourhoods. Since this facility is mainly for students in schools, colleges and Universities, need to identify the educational institutions around Toronto and the current Tennis facility in the regions. Then find out where the Tennis facility with respect to number of schools is very low. Their aim is to increase the proportion of Tennis facility to number of schools to 20%. It means, for 5 schools together at least one Facility for playing Tennis. Initially considering the neighborhoods in Noth York Borough for assessment.

1.3 Stakeholders: This project for identifying the locations for new Tennis facility in oronto, is very useful for identifying low facility and high requirement area. Sports Authority, Tennis Federation and local governments who implement new facilities in their area are stakeholders for this project. Certainly this will help the students to prove the necessity of a new Tennis facility in areas of low facility.

2. Data Acquisition and Preparation

2.1 Data Acquisition: Since the area of study is Toronto and neighbourhoods, to get the neighbourhood details, use Wikipedia as the source. Downloaded the data in html format using the link Toronto postal Codes. But this data doesn't include geographical coordinates. To get the geographical coordinates corresponding to each neighborhood, used the facility in website (Toronto neighborhood Geospatial_data). This is a csv format file. This study requires the details of schools and Tennis facilities in each neighborhood to identify the most required locations for new Tennis facility. I am using Foursquare data (Foursquare) for venues to get current details of schools and Tennis facilities in neighborhoods in North York borough (Toronto).

2.2 Data cleaning: Since the source of data for this project is from different sources and in different format, to build a clean data frame for further analysis, different data cleaning activities have to be done. The data scraped from Wikipedia for data related to Toronto neighborhood, contains neighborhood with not assigned boroughs.

Since our study is about the neighborhoods in North York only, neighborhoods without borough assigned were dropped to confirm the selected neighborhoods belong to North York itself. Since the geographical coordinates are towards postal codes, to avoid duplications in Geographical coordinates and postal codes in data, combined neighborhoods with same postal codes and made postal codes unique.

The venues list obtained from foursquare for schools include some other offices like school board and those ones which are not educational institutions were dropped. In Tennis venues list obtained from Foursquare, include shops which purchase Tennis goods also were removed from data. Also from venues data, dropped the unnecessary columns like address, city, state, cross street etc. Also dropped the redundant data like neighborhood, postal code etc.

2.3 Data preparation/ Feature Selection: Since the problem is to find out the low Tennis facilitated schools in neighboring areas, after validating various parameters, considered schools and Tennis facilities in 1KM from each neighborhood center were considered for study. The preliminary venue data types which comprise the education institutions and Tennis facilities are given below:

```
categories       object
distance          int64
id               object
lat             float64
lng             float64
name             object
Neighborhood     object
Type             object
dtype: object
```

Fig 2.1. Venue Dataframe Types

There were 38 categories of venues and for efficient analysis reduced the number of categories to 3. Various school types categorized into two groups - 'High school and below' which include elementary school, middle school and high school and 'Above High School' which include all colleges, universities and other institutions. After completing data cleaning and feature selection, conducted various exploratory and descriptive statistics and visualization to learn about the overall quality of data and to get a preliminary idea about the relationship between selected features. Conducted one hot encoding to venues and normalization to distance to make the data ready for analysis.

	categories	distance	id	lat	lng	name	Neighborhood	Type	Above Highschool	Highschool and Below	Tennis Facility
0	High School	227	4bf93961b182c9b6198e785a	43.804115	-79.366237	A.Y. Jackson Secondary School	Hillcrest Village	Highschool and Below	0	1	0
1	General College & University	456	4cb4c68e770fef3b060bb113	43.806337	-79.359022	Cliffwood Public School	Hillcrest Village	Above Highschool	1	0	0
2	Tennis Court	579	4fb2f0d6e4b0fb410b80cd27	43.798561	-79.363506	Hillcrest Tennis Club	Hillcrest Village	Tennis Facility	0	0	1
3	Athletics & Sports	297	518815ab498e99d0bb98ac96	43.805068	-79.366677	A.Y. Jackson Secondary School Track	Hillcrest Village	Tennis Facility	0	0	1
4	School	716	4c5239d099ecc9b6c8658e5a	43.797325	-79.363339	Crestview Public School	Hillcrest Village	Highschool and Below	0	1	0

Fig 2.2 Data after one hot encoding

3. Methodology

3.1 Exploratory Data Analysis:

3.1.1 Preliminary Statistical Analysis: To get an overall picture of the venues, using the describe method, analyzed the data. It gives the number of venues, mean, standard deviation, minimum, maximum and quartiles.

	distance	lat	lng	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist	sch_no
count	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000	170.0
mean	743.423529	43.753780	-79.413840	0.164706	0.723529	0.111765	111.323529	551.870588	80.229412	1.0
std	297.004189	0.025578	0.061772	0.372011	0.448574	0.316008	281.633582	421.678410	248.446726	0.0
min	144.000000	43.702148	-79.573605	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0
25%	533.500000	43.735219	-79.453806	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.0
50%	732.500000	43.753631	-79.410257	0.000000	1.000000	0.000000	0.000000	607.000000	0.000000	1.0
75%	964.750000	43.772888	-79.356112	0.000000	1.000000	0.000000	0.000000	881.750000	0.000000	1.0
max	1390.000000	43.811299	-79.312567	1.000000	1.000000	1.000000	1375.000000	1390.000000	1297.000000	1.0

Fig 3.1 High level statistics using describe method

Following is a high level summary of neighborhood statistics.

Neighborhood	Above Highschool sum	Highschool and Below sum	Tennis Facility sum	AHS_dist mean	HSB_dist mean	TF_dist mean
Bathurst Manor, Downsview North, Wilson Heights	0	4	0	0.000000	636.500000	0.000000
Bayview Village	0	4	0	0.000000	780.000000	0.000000
Bedford Park, Lawrence Manor East	0	8	0	0.000000	691.250000	0.000000
CFB Toronto, Downsview East	1	2	0	237.000000	309.000000	0.000000
Don Mills North	2	9	4	33.866667	365.066667	141.733333
Downsview Central	1	6	0	126.714286	661.142857	0.000000
Downsview Northwest	1	6	0	107.857143	620.857143	0.000000
Downsview West	1	5	0	114.333333	728.500000	0.000000
Downsview, North Park, Upwood Park	0	5	0	0.000000	843.600000	0.000000
Emery, Humberlea	0	1	0	0.000000	680.000000	0.000000
Fairview, Henry Farm, Oriole	3	9	1	184.769231	594.000000	33.538462
Flemingdon Park, Don Mills South	0	2	0	0.000000	821.500000	0.000000
Glencairn	2	7	2	151.545455	646.181818	133.000000
Hillcrest Village	1	6	2	50.666667	465.111111	97.333333
Humber Summit	0	3	0	0.000000	727.666667	0.000000
Lawrence Heights, Lawrence Manor	1	5	1	158.142857	469.714286	170.285714
Newtonbrook, Willowdale	3	4	0	345.285714	476.285714	0.000000
Northwood Park, York University	0	1	2	0.000000	258.666667	624.000000
Parkwoods	1	8	1	109.700000	592.600000	92.400000
Silver Hills, York Mills	0	4	1	0.000000	300.200000	150.400000
Victoria Village	0	4	0	0.000000	779.250000	0.000000
Willowdale South	8	9	3	215.400000	344.300000	122.400000
Willowdale West	1	5	0	36.833333	785.166667	0.000000
York Mills West	2	6	2	170.300000	559.500000	155.000000

Fig 3.2 North York Neighborhoods

Different high level analysis were done for learning the data. A high level venue analysis based on categories done to get the distribution status of the education institutions and Tennis facilities.

	Type
Highschool and Below	123
Above Highschool	28
Tennis Facility	19

Fig 3.3 Count of venues in North York

3.1.2 Preliminary Visual Analysis of venues using map: To see the geographical locations of 24 neighborhoods in North York, plotted a map. The following map shows the locations of neighborhoods.

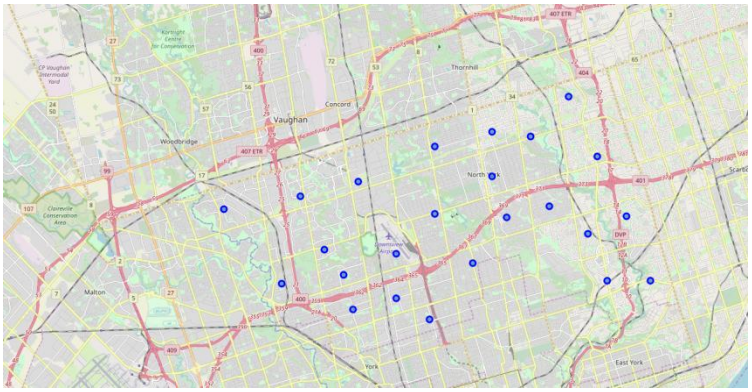

Fig 3.4 Neighborhoods in North York

3.1.2 Detailed Analysis: More analysis done on data to get the information about the most required neighborhoods for new Tennis facilities. The below pie chart shows the over all distribution of schools and Tennis facilities in North York. The figure shows the current ratio of Tennis facility is only 11% when compared with the Education Institutions which is very low with respect to the target.

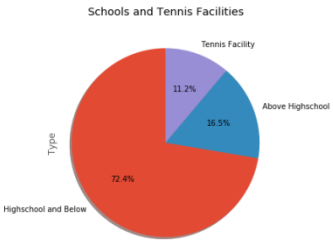

Fig 3.5 Education Institutions and Tennis Facility

The scatter plot between the venue distance from neighborhood centers shows the distributions of venues with respect to distance. Also this figure gives an idea about the less availability of Tennis Facilities with respect to Education institutions.

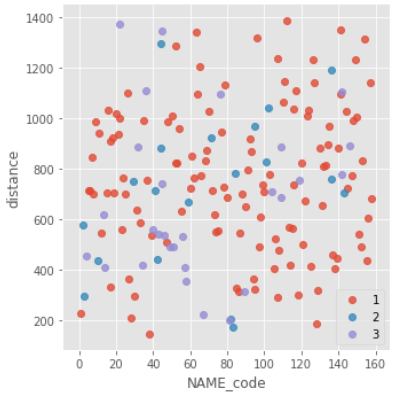

Fig. 3.6 Distance from neighborhood center to school/Tennis

Below box plot shows the quartile distribution of various venues with respect to distance.

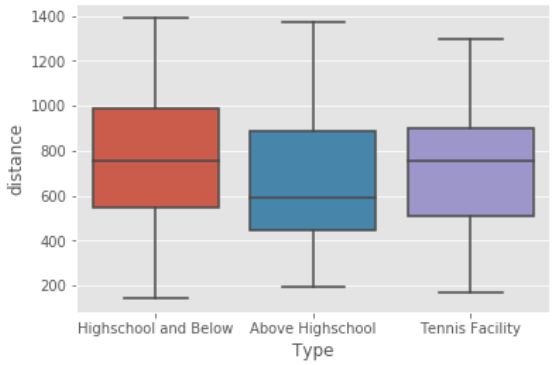

Fig. 3.7 Venues with distance

Below Bar chart shows the number of Highschool and below educational institutions are very high with respect to above high school and Tennis facilities.

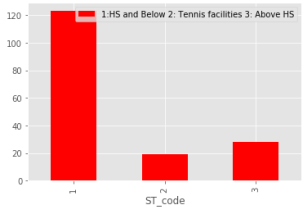

Fig 3.8 Number of Schools and Tennis Facilities in North York Borough

3.2 Machine Learning :

In further analysis used the machine learning technique. K-means cluster analysis done on data for segmenting the venues to find out the groups with less Tennis Facilities. Since the most affected feature is the number of various schools in neighborhoods used the number of venues and tennis facilities in neighborhood are taken for clustering. The data are segmented into 5 clusters based. Analysis shows some neighborhood sets have good facilities and some have no facilities. Below shows the details of clusters.

:

	Cluster Labels		Neighborhood	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist
				sum	sum	sum	mean	mean	mean
10	0	0	Fairview, Henry Farm, Oriole	3	9	1	184.769231	594.000000	33.538462
12	0	0	Glencairn	2	7	2	151.545455	646.181818	133.000000
13	0	0	Hillcrest Village	1	6	2	50.666667	465.111111	97.333333
14	0	0	Humber Summit	0	3	0	0.000000	727.666667	0.000000
15	0	0	Lawrence Heights, Lawrence Manor	1	5	1	158.142857	469.714286	170.285714
16	0	0	Newtonbrook, Willowdale	3	4	0	345.285714	476.285714	0.000000
18	0	0	Parkwoods	1	8	1	109.700000	592.600000	92.400000

Fig 3.9 Cluster1(L0)

	Cluster Labels		Neighborhood	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist
				sum	sum	sum	mean	mean	mean
0	1	1	Bathurst Manor, Downsview North, Wilson Heights	0	4	0	0.000000	636.500000	0.000000
1	1	1	Bayview Village	0	4	0	0.000000	780.000000	0.000000
2	1	1	Bedford Park, Lawrence Manor East	0	8	0	0.000000	691.250000	0.000000
3	1	1	CFB Toronto, Downsview East	1	2	0	237.000000	309.000000	0.000000
4	1	1	Don Mills North	2	9	4	33.866667	365.066667	141.733333
5	1	1	Downsview Central	1	6	0	126.714286	661.142857	0.000000
6	1	1	Downsview Northwest	1	6	0	107.857143	620.857143	0.000000

Fig 3.10 Cluster2(L1)

	Cluster Labels		Neighborhood	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist
				sum	sum	sum	mean	mean	mean
21	2	2	Willowdale South	8	9	3	215.4	344.3	122.4

Fig 3.11 Cluster3(L2)

	Cluster Labels		Neighborhood	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist
				sum	sum	sum	mean	mean	mean
10	3	3	Fairview, Henry Farm, Oriole	3	9	1	184.769231	594.000000	33.538462
12	3	3	Glencairn	2	7	2	151.545455	646.181818	133.000000
13	3	3	Hillcrest Village	1	6	2	50.666667	465.111111	97.333333
15	3	3	Lawrence Heights, Lawrence Manor	1	5	1	158.142857	469.714286	170.285714
16	3	3	Newtonbrook, Willowdale	3	4	0	345.285714	476.285714	0.000000
18	3	3	Parkwoods	1	8	1	109.700000	592.600000	92.400000

Fig 3.12 Cluster4(L3)

	Cluster Labels		Neighborhood	Above Highschool	Highschool and Below	Tennis Facility	AHS_dist	HSB_dist	TF_dist
				sum	sum	sum	mean	mean	mean
7	4	4	Downsview West	1	5	0	114.333333	728.500000	0.0
8	4	4	Downsview, North Park, Upwood Park	0	5	0	0.000000	843.600000	0.0
9	4	4	Emery, Humberlea	0	1	0	0.000000	680.000000	0.0
11	4	4	Flemingdon Park, Don Mills South	0	2	0	0.000000	821.500000	0.0
14	4	4	Humber Summit	0	3	0	0.000000	727.666667	0.0

Fig 3.13 Cluster5(L4)

From the clusters, it is understood that cluster5 which include 5 neighborhoods have no Tennis Facilities. Next, cluster2 with 7 neighborhoods, only one neighborhood has a Tennis facility and no facilities to other 6 schools. So this cluster too need to taken into consideration when taking decisions about the neighborhood for new Tennis facilities. Cluster 3 with 1 neighborhood is the best in facility. Cluster1 and cluster 4 have medium facilities.

4. Results

Based on various exploratory analysis and machine learning analysis on data for 24 neighborhoods within 1Km radius in the selected borough North York, it is understood that Tennis facilities are available to only limited number of neighborhood education institutions. Among 24 neighborhoods, there are total 128 high school or below institutions and 28 above high school education institutions. Number of Tennis facility is only 19 for the total 151 education institutions. The scatter diagram and box plot, shows more schools and facilities are distributed between 300m and 700m from neighborhood centers.

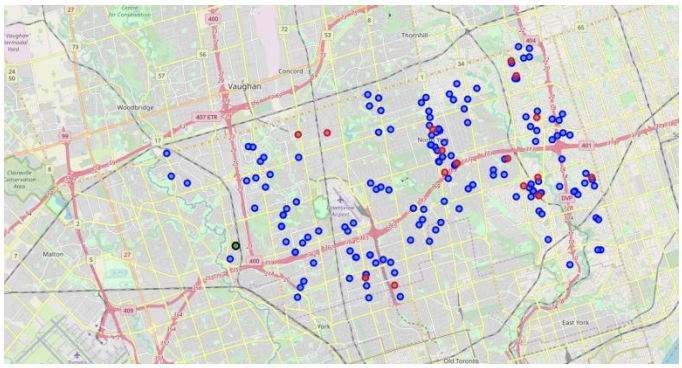

Fig 4.1 Map showing education institutions(blue) and tennis facilities(red)

Cluster analysis shows cluster5 with 5 neighborhoods and cluster2 with 7 neighborhoods are the two sets of neighbors without Tennis facility. From cluster 2 we can exclude the neighborhood Don Mills North, since it has some Tennis facility. Now we are considering the 11 neighborhoods for further discussion.

5. Discussion with observations and recommendations

Based on cluster analysis, we selected 11 neighborhoods for considering to new Tennis facilities. Since the authorities preference is to get maximum benefits from the new facility, the selection of new facility is based on number of education institutions can be benefitted from it. So the preferences will be given to the neighborhood with maximum number of schools with no facility. Table below shows the preference order for new facility. If we are considering for three new facilities, the recommendation is Bedford Park, Lawrence Manor East (8 institutions benefitted), Downsview Central and Downsview Northwest (7 institutions benefitted).

Neighborhood	Total School	Preference
Bedford Park, Lawrence Manor East	8	1
Downsview Central	7	2
Downsview Northwest	7	3
Downsview West	6	4
Downsview, North Park, Upwood Park	5	5
Bathurst Manor, Downsview North, Wilson Heights	4	6
Bayview Village	4	7
CFB Toronto, Downsview East	3	8
Humber Summit	3	9
Flemingdon Park, Don Mills South	2	10
Emery, Humberlea	1	11

6. Conclusion

In north York borough among 24 neighborhoods, 11 neighborhoods have no facility for Tennis. By conducting various analysis, recommended three neighborhoods with more than 6 educational institutions benefitted from the new facility for first face implementation of project. Those are Bedford Park/Lawrence Manor East, Downs view Central and Downs view.

References

1. https://www.coursera.org
2. **https://en.wikipedia.org**
3. https://www.google.com
4. https://stackoverflow.com
5. https://www.markdownguide.org
6. https://matplotlib.org/index.html
7. https://pandas.pydata.org/pandas-docs/stable/reference/index.html