Young Population And Venues Data: Analysis Of Toronto City

By: Priyanka Sagwekar

For the purpose of IBM Capstone Project

Date: 1st December 2020

Github Notebook:

https://nbviewer.jupyter.org/github/PriyankaSagwekar/Coursera_Capstone/blob/main/IBM-Capstone%20project%20notebook.ipynb#3

Table of contents

1. Data acquisition and cleaning

From the given data sources, we download:
1. Neighbourhoods (geojson file) which includes information related to each neighbourhood i.e Name, Latitude, Longitude,Geometry.
2. Neighbourhood profiles (csv file) containing information related to area, population and corresponding age group in each of the Toronto neighbourhoods based on 2016 census of Canada.

All useful libraries are imported. The first datafile- neighbourhood (geojson file) is loaded and read using geopandas library. Dataset is cleaned by removing irrelevant columns, renaming columns with appropriate terminology to finally create dataset which looks as follows:

	Neighbourhood	Longitude	Latitude	Geometry
0	Yorkdale-Glen Park (31)	-79.457108	43.714672	POLYGON ((-79.43969 43.70561, -79.44011 43.705...
1	York University Heights (27)	-79.488883	43.765736	POLYGON ((-79.50529 43.75987, -79.50488 43.759...
2	Yonge-St.Clair (97)	-79.397871	43.687859	POLYGON ((-79.39119 43.68108, -79.39141 43.680...
3	Yonge-Eglinton (100)	-79.403590	43.704689	POLYGON ((-79.41096 43.70408, -79.40962 43.704...
4	Wychwood (94)	-79.425515	43.676919	POLYGON ((-79.43592 43.68015, -79.43492 43.680...

Figure 2: Neighbourhood dataset with 4 columns and 140 rows each respresenting neighbourhood in Toronto

Next neighbourhood profiles (csv file) is loaded and read using pandas library. The dataset contains a lot of socio-economic demographic information that we may not be interested in. Hence csv file is traversed, relevant rows are noted by id. This noted row details are used to clean the dataset. We drop all redundant columns, rename columns with appropriate terminology, take transpose of dataframe to make rows as columns and vice versa. Also row values datatype which were interpreted as object by python is changed to numeric. No null values were found. Further two new columns namely population and population density of people in (0-24 years) age group are added, with it's values calculated for each of the neighbourhood. For convenience it is also ensured that neighbourhood column for both datasets exactly match. So that our dataset is finally transformed from figure 3 to figure 4,

From:

]:

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	...	Willowdale West	Willowridge-Martingrove-Richview	Wobu
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	...	37	7	1
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	...	No Designation	No Designation	N
2	3	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2016	2,731,571	29,113	23,757	12,054	30,526	...	16,936	22,156	53,4
3	4	Population	Population and dwellings	Census Profile 98-316-X2016001	Population, 2011	2,615,060	30,279	21,988	11,904	29,177	...	15,004	21,343	53,3
4	5	Population	Population and dwellings	Census Profile 98-316-X2016001	Population Change 2011-2016	4.50%	-3.90%	8.00%	1.30%	4.60%	...	12.90%	3.80%	0.3

Figure 3: Neighbourhood profiles dataset with 2383 rows and 146 columns

Cleaned to:

	Neighbourhood	Population density per square kilometre	Land area in square kilometres	Children (0-14 years)	Youth (15-24 years)	Children + Youth (0-24 years)	Population density (0-24 years) in sq.Km
0	Yorkdale-Glen Park (31)	2451	6.04	1960	1870	3830	634
1	York University Heights (27)	2086	13.23	4045	4750	8795	665
2	Yonge-St.Clair (97)	10708	1.17	1210	920	2130	1821
3	Yonge-Eglinton (100)	7162	1.65	1800	1225	3025	1833
4	Wychwood (94)	8541	1.68	1860	1320	3180	1893

Figure 4: Neighbourhood population (0-24 years) profiles with 140 rows each representing neighbourhood in Toronto and 7 columns

We visualise each of Toronto neighbourhoods using Folium library. Here we create an interactive map with labels and circle markers. Another layer is added on the same map which represents a radial zone of 1.5 Km around each neighbourhood. This radial zone is explored later using Foursqaure API. After varying radius values, it was found 1.5 Km was optimum radial distance to explore neighbourhoods, as it gave optimum coverage as well as is considerate distance given our objective of assessing accessible venues nearby well within the limits of neighbourhoods.

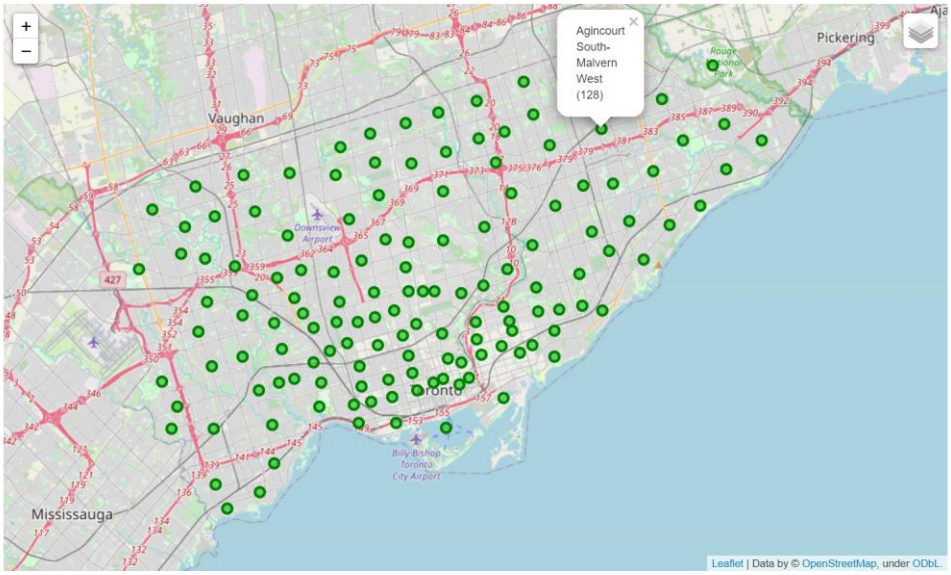

Figure 5: 140 Neighbourhoods marked with circle markers and labeled

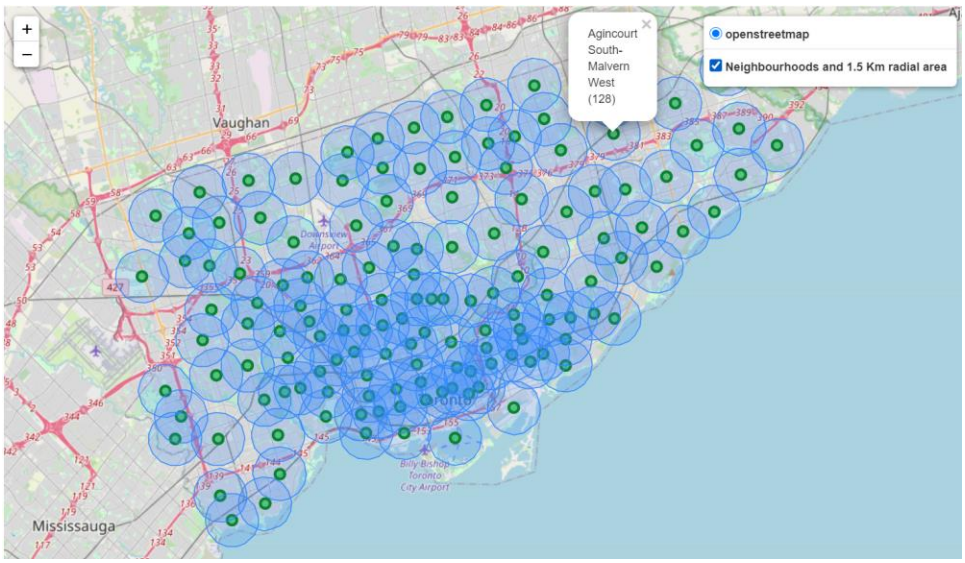

Figure 6: An interactive layer on same map depicting radial zone of 1.5 Km around each neighbourhood

We have further visualised top ten neighbourhoods with highest population density and population in age group of (0-24 years) using bar charts. Bar charts were prepared using matplotlib.pyplot library.

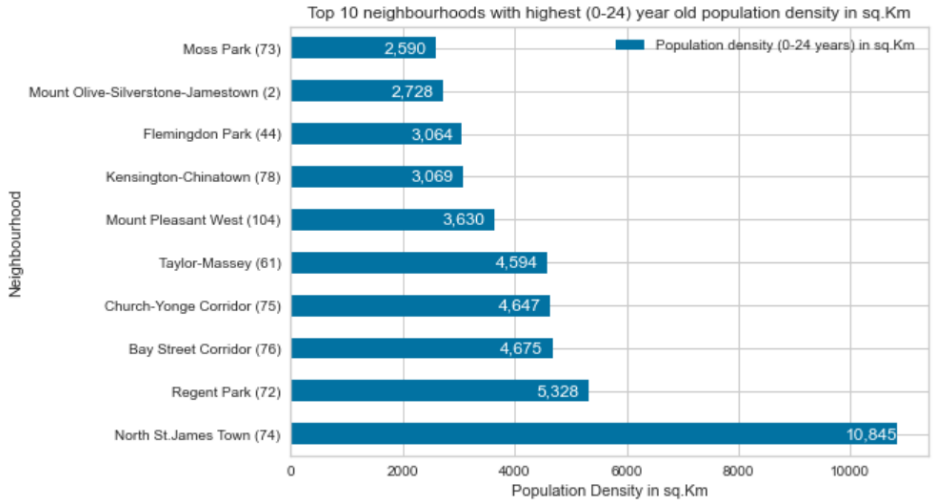

Figure 7: Top 10 neighbourhoods with highest population density in sq.Km

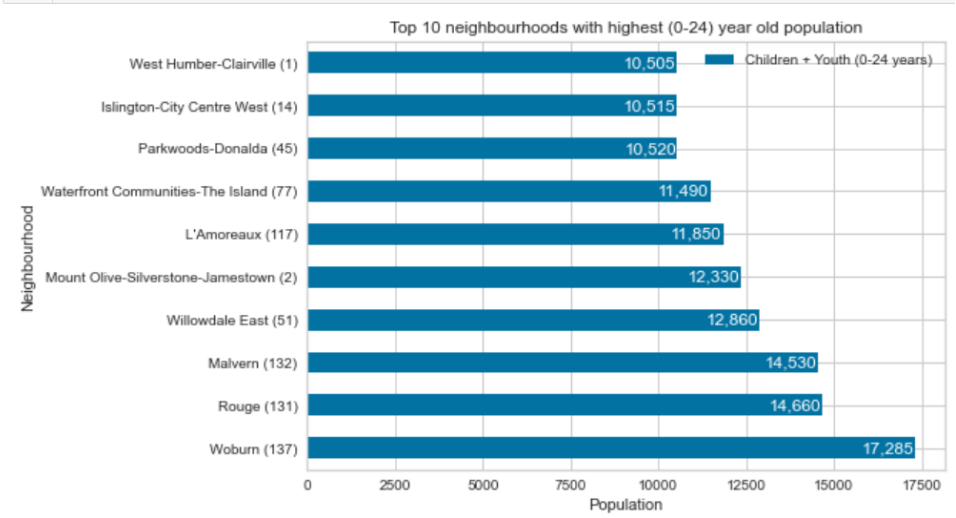

Figure 8: Top 10 neighbourhoods with highest population of (0-24) year olds

2. Exploration and analysis-Foursquare API

Now that we are interested in assessing amenities, businesses, venues that shall serve to the development, well being and health of young population as per our objective, it is a good option to use Foursquare API. Foursquare marks many venues, with all related information like name, latitude, longitude, category of venue etc in its database. Foursquare API is used to explore all nearby venues within 1.5 Km of neighbourhoods passing neighbourhood centroid geocoordinates and radius of 1500 (meters). Foursquare maximum number of venues response per neighbourhood is at it default limit of 100. Venue information is retrieved as json file from which a dataset is created including details like venue name, latitude, longitude, category for further analysis as shown in figure.

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Yorkdale-Glen Park (31)	43.714672	-79.457108	Harvey's	43.715337	-79.455396	Restaurant
1	Yorkdale-Glen Park (31)	43.714672	-79.457108	Mary Brown's Famous Chicken	43.718309	-79.455706	Fried Chicken Joint
2	Yorkdale-Glen Park (31)	43.714672	-79.457108	Krystos	43.718516	-79.455855	Greek Restaurant
3	Yorkdale-Glen Park (31)	43.714672	-79.457108	Yuki Japanese Restaurant	43.720610	-79.456119	Sushi Restaurant
4	Yorkdale-Glen Park (31)	43.714672	-79.457108	Caffè Demetre	43.720179	-79.456219	Dessert Shop

```
1 nearby_venues.shape
```

Figure 9: Dataframe with information of nearby venues obatained from Foursquare API.

It is found that there are 9082 venues belonging to 361 different categories like park, restaurant etc. There are about 41 neighbourhoods that have reached their Foursqaure venue limit.

After obtaining the data of all nearby venues located within 1.5 km radius of our neighbourhood centroids using Foursquare API, we shall proceed with data pre-processing. In the view of project objective categories of interest are related to sports, fitness and wellbeing, Entertainment,art and learning places, recreation spots, tourist places and greenspaces. For the purpose the data is screened systematically to extract all related venues.

Careful observation of the list of venue categories reveals that - the list mainly contains restaurants, stores, shops, bar, food places. These irrelevant venues that are not in conformity of project objective are removed using python code. Later the reduced list is manually traversed to create six broad categories namely: Green spaces, Sports amenities, Tourist attractions, Fitness and wellbeing, Kids stores, Entertainment, art and learning. These categories are carefully curated keeping our objective in mind, you can go through the list for better understanding of these categories.

Here's a table showing the venues corresponding to each category:

New Categories	Venues included
Sport's Amenities	Athletics & Sports, Tennis Court, Playground, Skating Rink, Hockey Arena, Hockey Field, Outdoors & Recreation, Racecourse, Badminton Court, Indoor Play Area, Skate Park, Ski Area, Racetrack, Soccer Stadium, Golf Course, Sports Club, Stadium, Baseball Field, Pool, Soccer Field, Bowling Alley, Disc Golf, Paintball Field, Golf Driving Range, Curling Ice, Ski Chalet, Pool Hall, Other Great Outdoors, Gaming Café, Moving Target, Baseball Stadium, Tennis Stadium, Volleyball Court, Basketball Stadium
'Fitness and Well-being	Gym, Gym / Fitness Center, Yoga Studio, Dance Studio, Gymnastics Gym, Pilates Studio, Massage Studio, Gym Pool, Climbing Gym, Martial Arts School, Recreation Center, Boxing Gym
Tourist attractions	Museum, Art Gallery, Zoo Exhibit, Zoo, Circus, History Museum, Science Museum, Art Museum, Theme Park Ride / Attraction, Theme Park, Monument / Landmark, Castle, Rock Climbing Spot, Historic Site, Aquarium
Green spaces	Garden, Park, Scenic Lookout, National Park, Lake, River, Farm, Field, Botanical Garden, Beach, Garden Center, Trail, Dog Run, Sculpture Garden, Hot Spring
Entertainment, Art and Learning	Music Venue, Performing Arts Venue, Movie Theater, Indie Movie Theater, Concert Hall, Event Space, Amphitheater, Street Art, Theater, Music School, General Entertainment, Casino, Jazz Club, Comedy Club, Indie Theater, Library, Community Center, School, High School, University, College Quad, College Rec Center, College Stadium, Photography Lab, Photography Studio, Street Art
Kid's Stores	Baby Store, Kids Store, Toy / Game Store, Outdoor Supply Store, Bookstore, Hobby shop, Comic Shop, Sporting Goods Shop, Video Game Store

New category is assigned to each venue, all rows that doesn't fall into any of above category are deleted. The updated dataframe has 1300 venues and 8 columns which looks as follows:

	Neighbourhood	Neighbourhood Latitude	Neighbourhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category	New Venue Category
0	Yorkdale-Glen Park (31)	43.714672	-79.457108	The Lego Store	43.725146	-79.452974	Toy / Game Store	Kid's Stores
1	Yorkdale-Glen Park (31)	43.714672	-79.457108	BATLgrounds	43.724054	-79.463398	Athletics & Sports	Sport's Amenities
2	Yorkdale-Glen Park (31)	43.714672	-79.457108	Indigo	43.727006	-79.451883	Bookstore	Kid's Stores
3	Yorkdale-Glen Park (31)	43.714672	-79.457108	Cineplex Cinemas	43.727000	-79.451290	Movie Theater	Entertainment, Art and Learning
4	Yorkdale-Glen Park (31)	43.714672	-79.457108	Playtime Bowl	43.717427	-79.458148	Bowling Alley	Sport's Amenities

Figure 10: Dataframe updated with new categories

The above dataframe is modified using groupby function. A coloumn of total venues for each neighbourhood is added. It is observed three neighbourhoods records no venue from any of the new categories. One of the limitation here is whole area of neighbourhoods is not covered, i.e just 1.5 km area from centroid is screened for venues. Also venue limit is set to 100 by default, therefore there are high chances that relevant venues might not have been retrieved giving false impression. So these 3 neighbourhoods namely Humbermede (22), Humber Summit (21), Brookhaven-Amesbury (30) must be further explored by increasing the coverage or using other sources if available.

	Neighbourhood	Entertainment, Art and Learning	Fitness and Well-being	Green spaces	Kid's Stores	Sport's Amenities	Tourist attractions	Total venues
0	Agincourt North (129)	1	0	1	0	1	0	3
1	Agincourt South-Malvern West (128)	0	3	0	1	2	0	6
2	Alderwood (20)	0	1	0	3	0	0	4
3	Annex (95)	4	3	5	1	1	4	18
4	Banbury-Don Mills (42)	1	2	3	0	1	0	7
...	...	...	...	...	...	...	...	...
132	Wychwood (94)	3	2	5	0	1	6	17
133	Yonge-Eglinton (100)	1	7	6	2	4	0	20
134	Yonge-St.Clair (97)	0	3	6	0	2	5	16
135	York University Heights (27)	0	1	1	0	3	0	5
136	Yorkdale-Glen Park (31)	2	1	0	4	3	0	10

137 rows × 8 columns

Figure 11: Number of venues for each category for 137 neighbourhoods

Using one-hot encoding a dataframe containing top 6 most common venues for each of neighbourhood is created which used for clustering neighbourhoods

	Neighbourhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Agincourt North (129)	Sport's Amenities	Green spaces	Entertainment, Art and Learning	None	None	None
1	Agincourt South-Malvern West (128)	Fitness and Well-being	Sport's Amenities	Kid's Stores	None	None	None
2	Alderwood (20)	Kid's Stores	Fitness and Well-being	None	None	None	None
3	Annex (95)	Green spaces	Tourist attractions	Entertainment, Art and Learning	Fitness and Well-being	Sport's Amenities	Kid's Stores
4	Banbury-Don Mills (42)	Green spaces	Fitness and Well-being	Sport's Amenities	Entertainment, Art and Learning	None	None

Figure 12: Top 6 most common venues for neighbourhoods

3. Data clustering- k-Means clustering

It can be observed that each neighbourhood have their own strengths like Green spaces as 1st most common venue or say Sport's amenities as 1st most common venue etc. Perhaps we might be interested to group together neighbourhoods with similar strengths and weaknesses.
Such grouping helps in decision making. For instance say while allocating resources for upgradation of sports amenities, we may be interested to know all neighbourhoods that have sport's ameneties at first place or say we want to create new sports facilities then we may be interested to know all neighbourhoods that lack such facilities. Later on we can assess them by considering other important factors.

There are many machine learning algorithms available to form clusters of similar neighbourhoods. In this study K-means clustering is used. K-means is a popular clustering algorithm that works good for sparse datasets with none values. There are methods to measure it's accuracy and optimise it's result. Here we used K-means elbow and squared error (cost) which is calculated using sklearn.metrics library and visualised using kelbowvisualiser from yellowbrick lib.

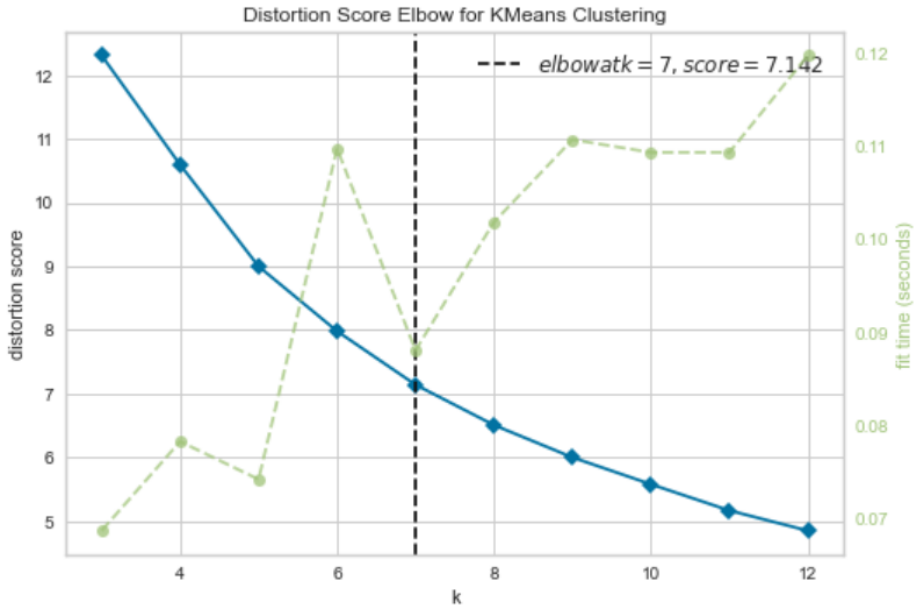

Figure 13: K-Means elbow method visualised using kelbowvisualizer

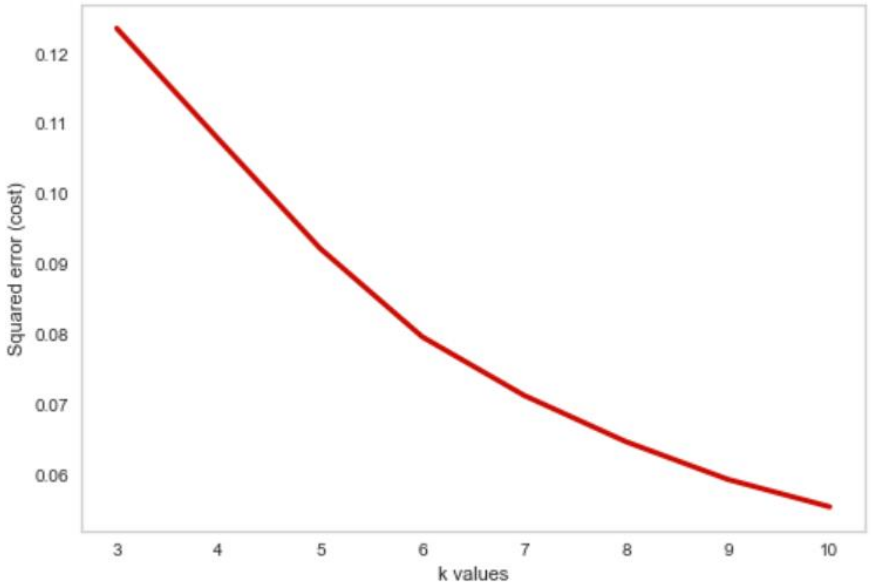

Figure 14: Squared error for k values

After evaluation k was selected as 7. K-means clustering was performed using sklearn library K-Means method. Cluster labels were added to dataset to create a dataframe shown below:

	Neighbourhood	Longitude	Latitude	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue
0	Yorkdale-Glen Park (31)	-79.457108	43.714672	3.0	Kid's Stores	Sport's Amenities	Entertainment, Art and Learning	Fitness and Well-being	None	None
1	York University Heights (27)	-79.488883	43.765736	6.0	Sport's Amenities	Green spaces	Fitness and Well-being	None	None	None
2	Yonge-St.Clair (97)	-79.397871	43.687859	5.0	Green spaces	Tourist attractions	Fitness and Well-being	Sport's Amenities	None	None
3	Yonge-Eglinton (100)	-79.403590	43.704689	4.0	Fitness and Well-being	Green spaces	Sport's Amenities	Kid's Stores	Entertainment, Art and Learning	None
4	Wychwood (94)	-79.425515	43.676919	5.0	Tourist attractions	Green spaces	Entertainment, Art and Learning	Fitness and Well-being	Sport's Amenities	None

Figure 15: Neighbourhoods clustered and labeled

To understand clusters better we visualised them as below:

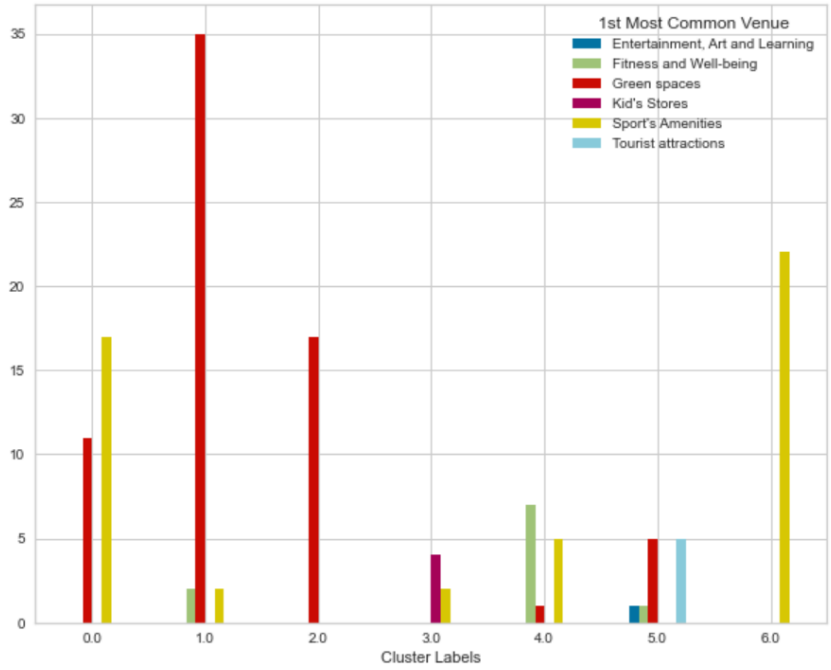

Figure 16: Cluster Label v/s 1st most common venue

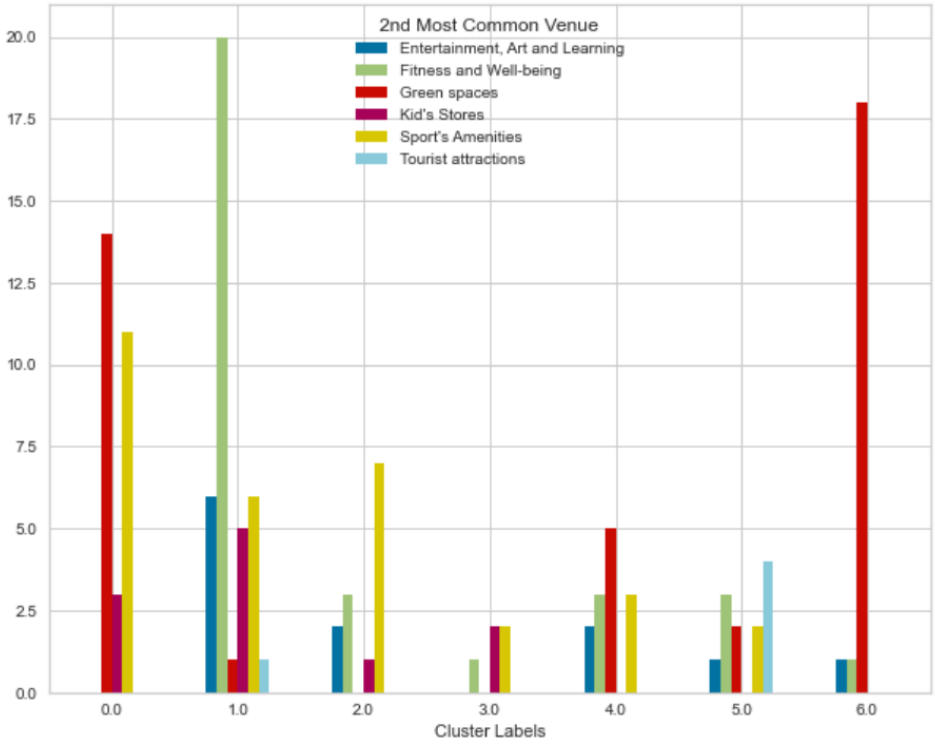

Figure 16: Cluster Label v/s 2nd most common venue

After analysis we draw following inferences:

Clusters	Strengths	Lacks
0	Green spaces and Sports amenities	Entertainment,art and learning,Tourist attractions,Fitness & Wellbeing
1	Green spaces, Fitness & Wellbeing and Sports amenities	--
2	Green spaces	Tourist attractions
3	Kid's stores and Sport's amenities	Green spaces,Tourist attractions, Entertainment,art and learning,
4	Fitness & Well being, Sport's amenities and Green spaces	Tourist attractions, Kid's Stores
5	Tourist attractions, Green spaces, Entertainment, art and learning, Fitness and well-being	Kid's Stores, Sports amenities
6	Sport's amenities and Green spaces	Tourist attractions, Kid's Stores

We now visualise clusters on map, each neighbourhood is labeled with cluster number and neighbourhood name. Moreover clusters are color coded.

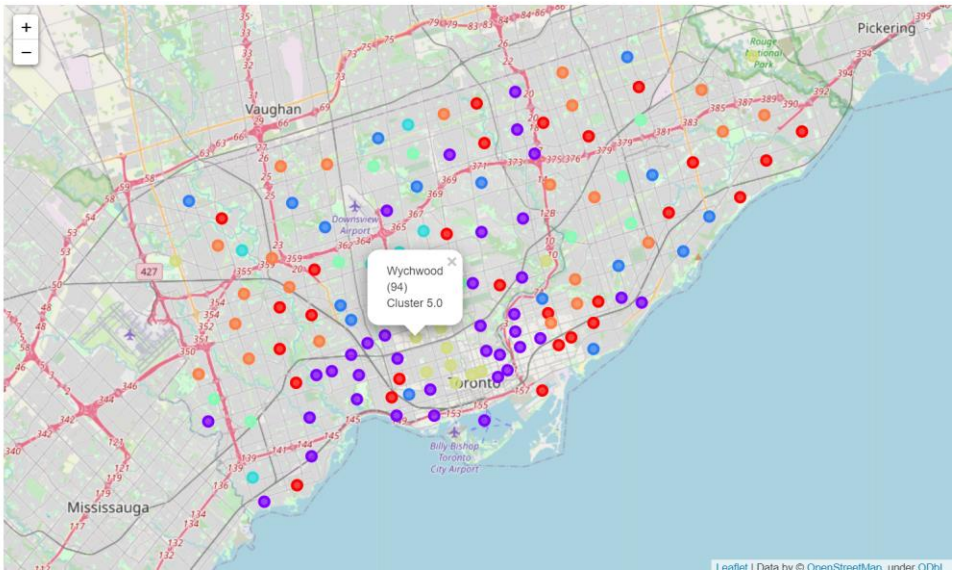

Figure 17: Neighbourhoods circle marked with different colors according to the cluster they belong

4. Data visualisation

Results are visualised for better insights using maps and charts. The visuals will be useful in drawing inferences, making reccommendations, answer the questions this study sorts to resolve. They make most of the data available with us to allign us better with this project objective.

Here's an interative choropleth map of Toronto neighbourhoods drawn using neighbourhood geojson data, colour graded according to population density and population using Folium. This choropleth is overrided with different coloured circle markers representing neighbourhoods and cluster they belong. This map makes it easier to traverse neighbourhoods and understand them. This is layered map, with options to select different layers added.

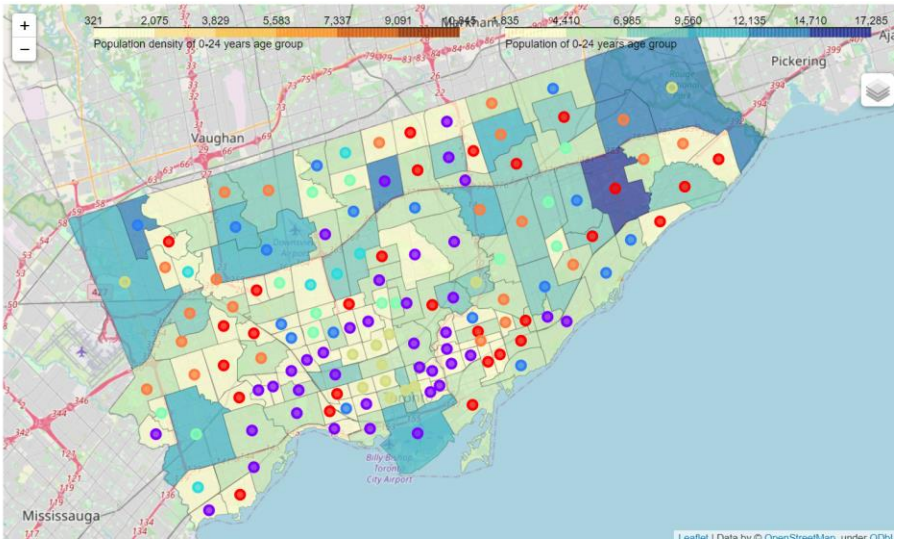

Figure 18: Choropleth map based on population overrided with neighbourhood cluster markers

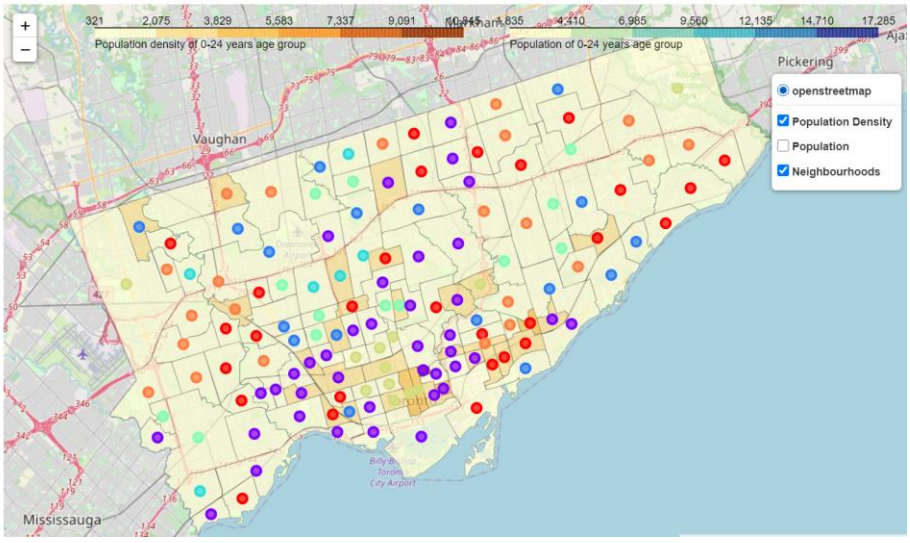

Figure 19: Choropleth map based on population overrided with neighbourhood cluster markers with layer controls currently checked at 'Population Density' and 'Neighbourhoods'

Top 10 populous neighbourhoods with cluster labels are charted with cluster labels

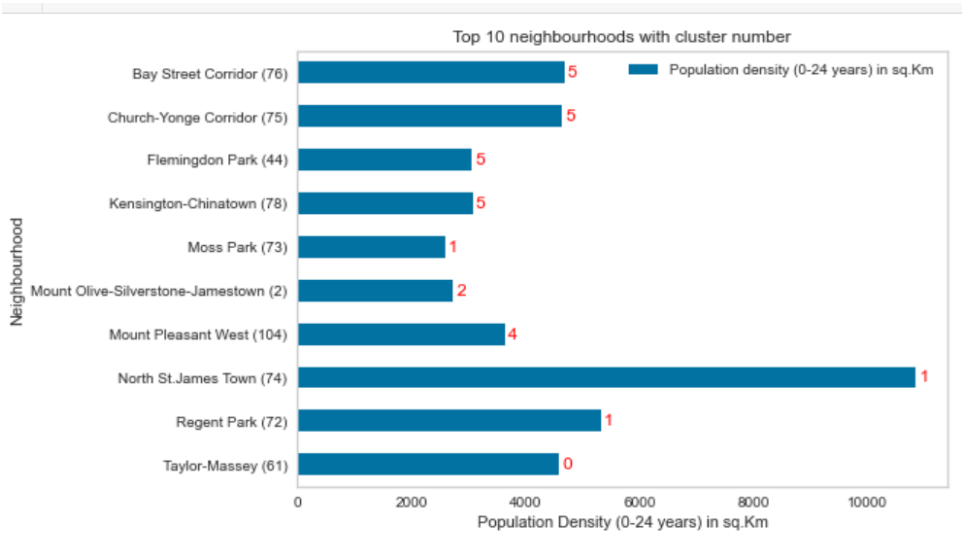

Figure 20: Top 10 neighbourhoods based on population density marked with corresponding cluster label in red

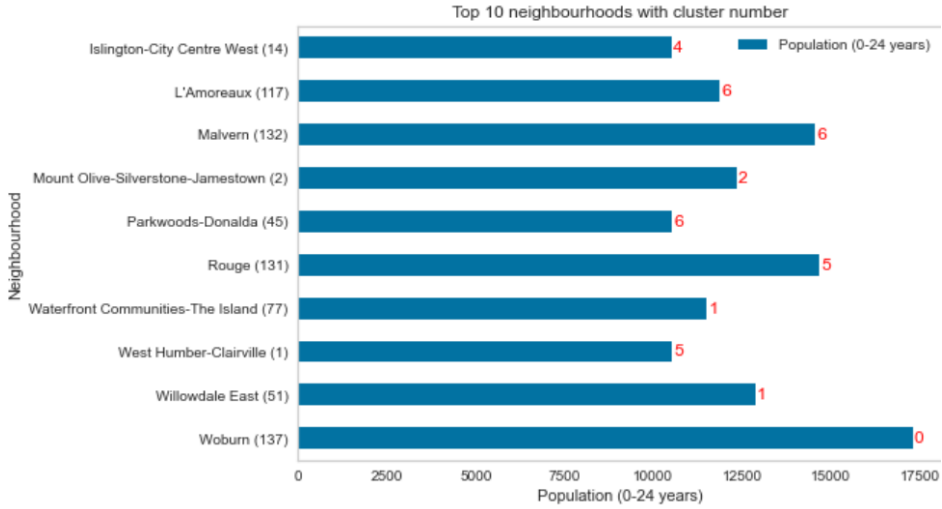

Figure 21: Top 10 neighbourhoods based on population marked with corresponding cluster label in red

5. Results and Conclusion

To summarise, we started with loading and cleaning to create a dataset with toronto neighbourhoods information like neighbourhood names, latitude, longitude, geometry, population and population density of target age group i.e (0-24 year olds). Thereafter we proceeded with data neighbourhood exploration where information related to nearby venues in a radius of 1.5 Km was collected. Once this data was available, we categorised the venues as per our objectives. Once all the required information was updated, we considered clustering neighbourhoods using K-Means technique, visualised the cluster against population and population density on an interactive map. Generated charts for top 10 neighbourhoods in terms of population and population density of target group.

Clusters	Strengths	Lacks
0	Green spaces and Sports amenities	Entertainment,art and learning,Tourist attractions,Fitness & Wellbeing
1	Green spaces, Fitness & Wellbeing and Sports amenities	--
2	Green spaces	Tourist attractions
3	Kid's stores and Sport's amenities	Green spaces,Tourist attractions, Entertainment,art and learning,
4	Fitness & Well being, Sport's amenities and Green spaces	Tourist attractions, Kid's Stores
5	Tourist attractions, Green spaces, Entertainment, art and learning, Fitness and well-being	Kid's Stores, Sports amenities
6	Sport's amenities and Green spaces	Tourist attractions, Kid's Stores

Now let's revisit our problem and try to answer the questions we intended to solve:

1. Which areas are the most suitable neighbourhoods for allocation of resources by public entities like municipalities/state/federal governments for the development of sports amenities, tourist locations and green spaces like parks?

There are three neighbourhoods namely, Humbermede (22), Humber Summit (21), Brookhaven-Amesbury (30) with considerable young population where no venues from amongst our categories were found. We need to explore these locations further. They must be on our priority for development of public amenities. Secondly public administration may be primarily interested to identify best suited neighbourhoods for development of Green spaces, Tourist attractions and sports amenities from amongst our venue categories.

Tourist attractions play a role of recreation spots, update youths with socio-cultural information, overall assist in mental development and social well being of young population. That is to say, for tourist attractions we will prefer cluster 0,2,3,4,6 with particular emphasis on following neighbourhoods as they not only lack such venues but also harbor good number of youth population: Woburn (137), Mount Olive-Silverstone-Jamestown (2), Islington-City Centre West (14), Malvern (132), L'Amoreaux (117),Parkwoods-Donalda (45),Taylor-Massey (61), Mount Olive-Silverstone-Jamestown (2), Mount Pleasant West (104)

Next category of interest is Green spaces like garden, parks, natural ecosystems etc. Such spaces are needed given rising environmental concerns and their proven impact on human life. We found Cluster 3 needs more of green spaces. It's quite commendable that none of our populated neighbourhoods fall in cluster 3. Kudos to city planners for giving due importance to this category and bringing it into implementation.

Last category of interest i.e Sport's amenities are quite crucial for youth's physical well being. Also these facilities are needed for them to excel in sports arena on international platforms. It seems city planners have succeeded in providing sufficient sport's amenities in all categories yet cluster 5 needs more sports amenities with special focus on following neighbourhoods as they have considerable young population: Kensington-Chinatown (78), Flemingdon Park (44), Church-Yonge Corridor (75), Bay Street Corridor (76), West Humber-Clairville (1), Rouge(131).

2. Which areas keep most potential to operate with success businesses like Yoga Studio, sports training/coaching services, Toy store, Theme park?

As we are quite aware that it is not only public entities but also some private entities that provide services for physical and mental well being of people in general and youths in particular. Out of our categories private entities might be interested in Entertainment,art and learning i.e music,dance studio,theater etc, Fitness and well being i.e gym, yoga studio etc, Kid's stores like sports goods shop, bookstores etc, Tourist attractions like Theme park rides/attractions.

Cluster 0 lacks in Fitness and wellbeing but tops in sport's amenities and green amenities, so it may not be suitable for say gym but is best suited for massage studio. Also it keeps potential for Entertainment,art and learning category. Interested entities can particularly explore Woburn (137), Taylor-Massey (61) neighbourhoods.

Cluster 4,5,6 is best suited for kid's stores, interestingly these clusters top in Fitness and wellbeing, Entertainment,art and learning, sports amenities so they certainly will have demand for related goods. Moreover the demand drivers i.e young population are found more in these neighbourhoods of cluster 4,5,6: Islington-City Centre West (14), Rouge (131), Parkwoods-Donalda (45), West Humber-Clairville (1), Malvern (132), L'Amoreaux (117), Kensington-Chinatown (78), Flemingdon Park (44), Church-Yonge Corridor (75), Bay Street Corridor (76), Mount Pleasant West (104).

cluster 0,2,3,4,6 are good options for developing tourist attractions like theme park, the following neighbourhoods can be considered with priority: Woburn (137), Mount Olive-Silverstone-Jamestown (2), Islington-City Centre West (14), Malvern (132), L'Amoreaux (117),Parkwoods-Donalda (45),Taylor-Massey (61), Mount Olive-Silverstone-Jamestown (2), Mount Pleasant West (104).

3. Which neighbourhoods in Toronto offer most of the children and youth oriented facilities?

As we see neighbourhoods in cluster 1 and 5 are children and youth friendly neighbourhoods with almost all required amenities for growth and development of children. Also as per our priority we can rate neighbourhoods.

With this we conclude our study here. We have effectively used freely available databases, cleaned data, used geopandas and folium libraries for visualisation, machine learning algorithm for clustering. Even though there's still a scope to further polish this study. Yet we have successfully answered our questions using data science tools and scientific analysis.

Thank you!