Lakshmi K

April 18, 2020

Predicting the Optimal Placement of a Hospital in Toronto

IBM Data Science Capstone Project

Introduction

For this project we will be trying to determine the best possible location to open an additional medical centre/ hospital in the city of Toronto. Many a time one might hear news about how there may not be enough space at a hospital for all the patients needing help and when there comes a time like an epidemic or a pandemic (as is the current state of things in the world), this is especially the case.

This report will be of interest to the board of directors and stakeholders of the hospital in question as well as the city of Toronto staff who would help oversee its development.

Our aim is to look for populated areas in neighbourhoods where there appears to be more young children and elderly present. Once those neighbourhoods have been found, we will then search the surrounding area for any other such health centres in the vicinity as we would like to construct the hospital in a distance far enough away from the others in an area where it would be most needed.

Using this criteria and our purpose along with relevant data to provide support, we aim to share our findings and reasonings for our choices with the city of Toronto staff, hospital board of directors, and stakeholders, on where we would advise them to construct the health centre.

Data

As mentioned above in the introductory section, the factors that will influence our decision are:

- The number of hospitals/ health centres in the area
- The number of people in the area
- Ages of the constituents of the neighbourhood

We will be using the following data sources for our analysis:

- Toronto neighbourhoods data

Attribution: Open Data License - Toronto; *Contains information licensed under the Open Government Licence – Toronto*

This csv dataset will be used to obtain the initial neighbourhood profiles and their geographical coordinates using the area_name, longitude and latitude columns.

- Toronto neighbourhood profiles data

Attribution: Open Data License - Toronto; *Contains information licensed under the Open Government Licence – Toronto*

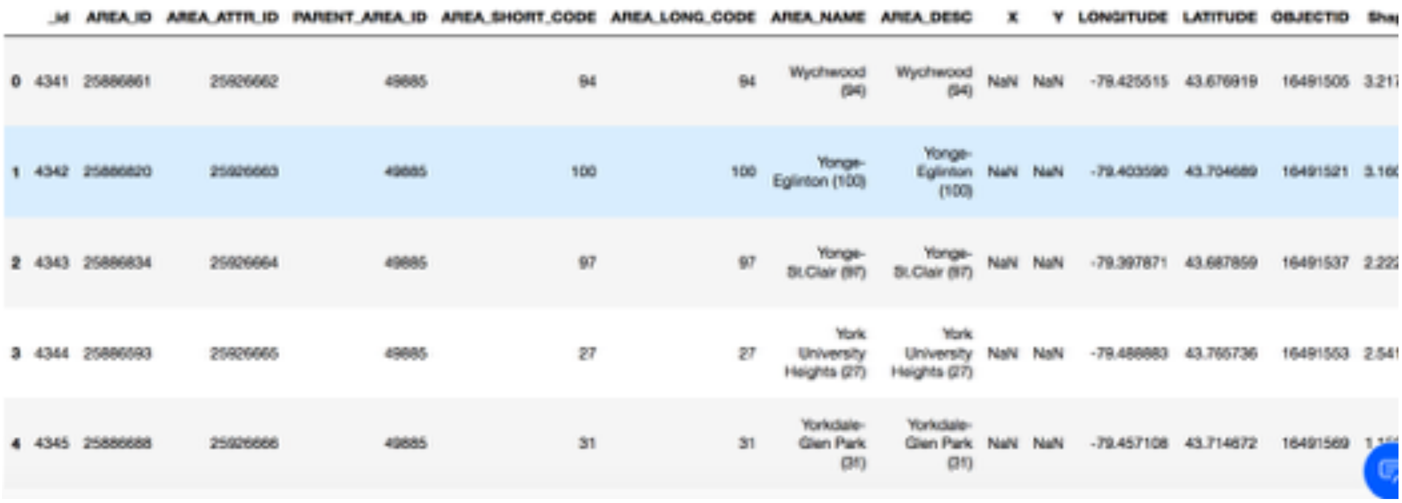
This source will be used to determine the age of the constituents in their respective neighbourhoods. In our case, we will be finding and using the appropriate rows with the population age characteristics data for seniors and children. Using the csv dataset we will also locate the number code for each neighbourhood and match it to the relevant area code column of the previous dataset in order to connect the data. Please also note that this data is from 2016.

- Foursquare API

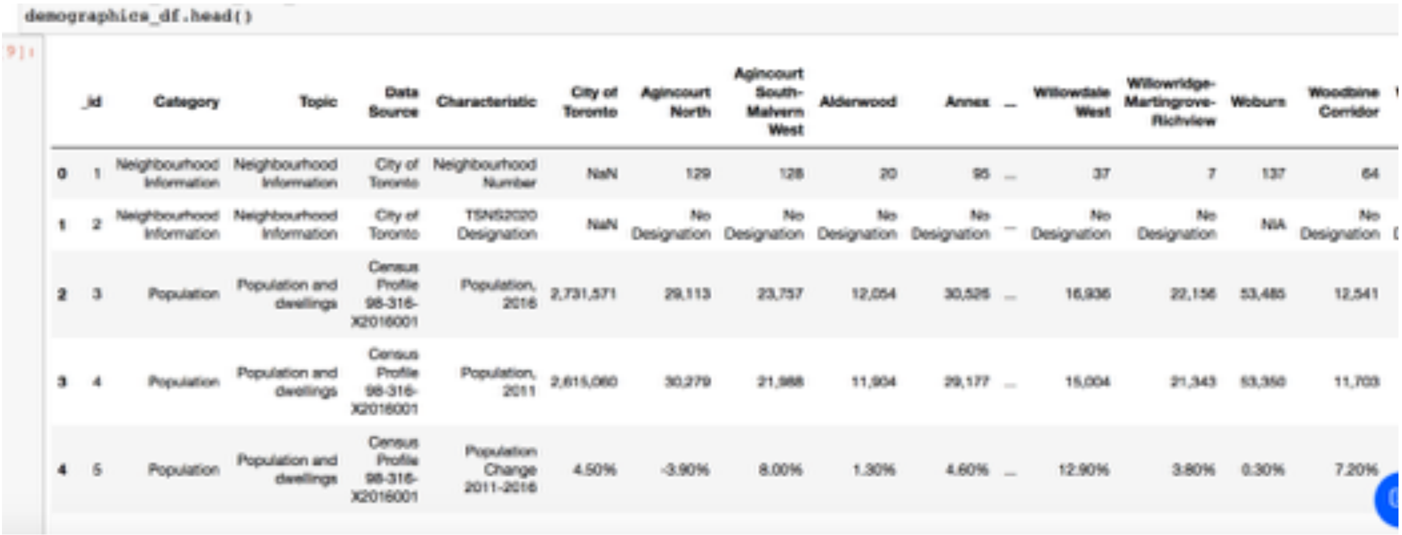
This API will be used to determine the location and number of hospitals in the Toronto area. We will then proceed to visualize this data through a map using Folium so we can see where the hospitals are situated and their distances from one another.

Methodology

This project involved me gathering neighbourhood population and neighbourhood location coordinates from two separate datasets.



	_id	AREA_ID	AREA_ATTR_ID	PARENT_AREA_ID	AREA_SHORT_CODE	AREA_LONG_CODE	AREA_NAME	AREA_DESC	X	Y	LONGITUDE	LATITUDE	OBJECTID	Sha
0	4341	25886861	25926662	49885	94	94	Wychwood (94)	Wychwood (94)	NaN	NaN	-79.425515	43.676919	16491505	3.217
1	4342	25886820	25926663	49885	100	100	Yonge-Eglinton (100)	Yonge-Eglinton (100)	NaN	NaN	-79.403590	43.704689	16491521	3.160
2	4343	25886834	25926664	49885	97	97	Yonge-St.Clair (97)	Yonge-St.Clair (97)	NaN	NaN	-79.397871	43.687859	16491537	2.222
3	4344	25886593	25926665	49885	27	27	York University Heights (27)	York University Heights (27)	NaN	NaN	-79.488883	43.765736	16491553	2.541
4	4345	25886688	25926666	49885	31	31	Yorkdale-Glen Park (31)	Yorkdale-Glen Park (31)	NaN	NaN	-79.457108	43.714672	16491569	1.1



```
demographics_df.head()
```

	_id	Category	Topic	Data Source	Characteristic	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	...	Willowdale West	Willowridge-Martingrove-Richview	Woburn	Woodbine Corridor	W
0	1	Neighbourhood Information	Neighbourhood Information	City of Toronto	Neighbourhood Number	NaN	129	128	20	95	...	37	7	137	64	
1	2	Neighbourhood Information	Neighbourhood Information	City of Toronto	TSNS2020 Designation	NaN	No Designation	No Designation	No Designation	No Designation	...	No Designation	No Designation	NIA	No Designation	D
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	53,485	12,541	
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	53,350	11,703	
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%	0.30%	7.20%	

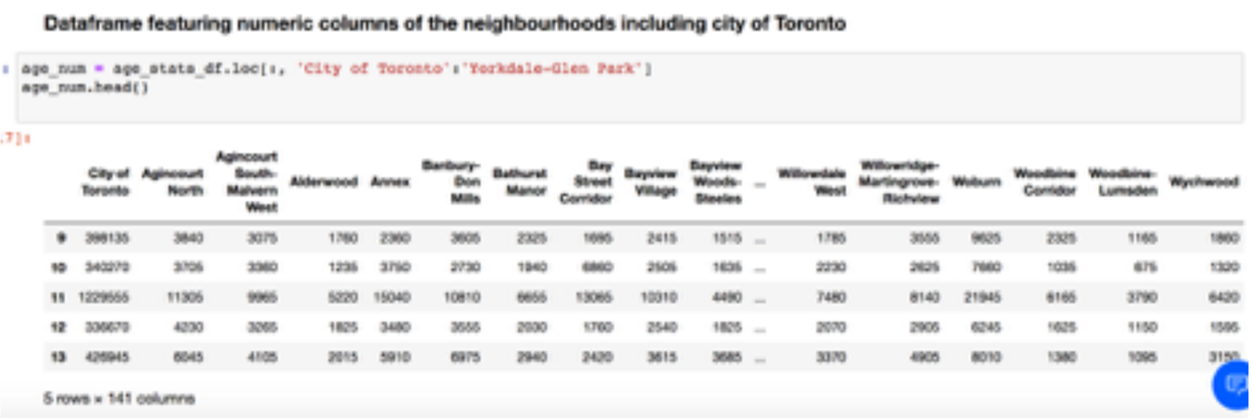
Having the idea in mind of merging these two dataframes together into a single one further on, I next selected the rows and columns in each of the dataframes that I believed would help with this process.

From neighbourhood_df, I created a new dataframe with only the 'AREA_NAME', 'LATITUDE', and 'LONGITUDE' columns and after a quick scan through demographics_df, I picked out the rows related to age characteristics by using the ‘Characteristic’ column and matching the string within.

As I wanted to differentiate between statistics from both the total population in each neighbourhood with the amount of people that could be considered to be a part of the vulnerable population, consisting of children between the ages of 0-14 and seniors 65 years and older, in each neighbourhood, I made a condensed dataframe for both.

During the creation of these new dataframes, I also noticed that what seemed to be the numerical values for the population were actually object datatypes. As such, I removed the commas, and then converted them to integer/float datatypes so that I would be able to calculate stats such as average and sum(which I added as a column to the data frames following this as well)

Example as shown below:

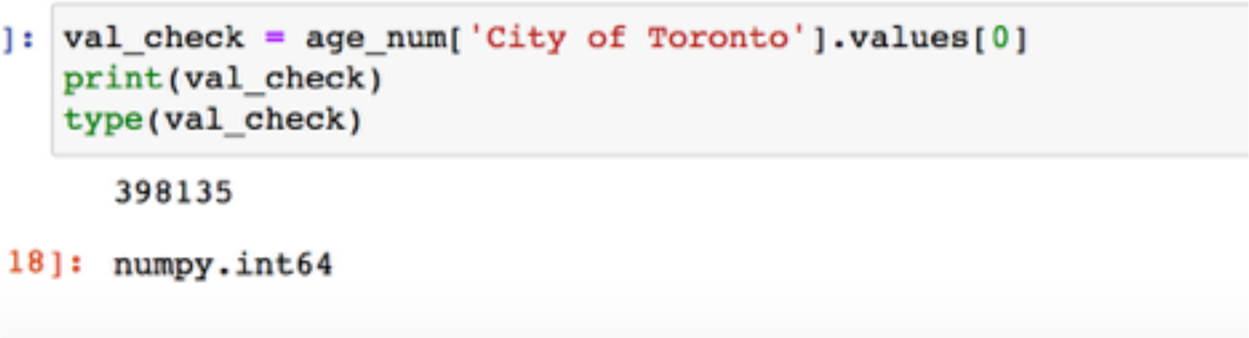


Dataframe featuring numeric columns of the neighbourhoods including city of Toronto

```
age_num = age_stats_df.loc[:, 'City of Toronto':'Yorkdale-Glen Park']
age_num.head()
```

	City of Toronto	Agincourt North	Agincourt South-Malvern West	Alderwood	Annex	Banbury-Don Mills	Bathurst Manor	Bay Street Corridor	Bayview Village	Bayview Woods-Steeles	...	Willowdale West	Willowridge-Martingrove-Richview	Woburn	Woodbine Corridor	Woodbine-Lumsden	Wychwood
9	398135	3840	3075	1760	2360	3605	2325	1695	2415	1515	...	1785	3555	9625	2325	1165	1860
10	340270	3705	3360	1235	3750	2730	1940	6860	2505	1635	...	2230	2625	7660	1035	675	1320
11	1229555	11305	9965	5220	15040	10810	6655	13065	10310	4490	...	7480	8140	21945	6165	3790	6420
12	336670	4230	3265	1825	3480	3555	2030	1760	2540	1825	...	2070	2905	6245	1625	1150	1595
13	426945	6045	4105	2015	5910	6975	2940	2420	3615	3685	...	3370	4905	8010	1380	1095	3150

5 rows × 141 columns

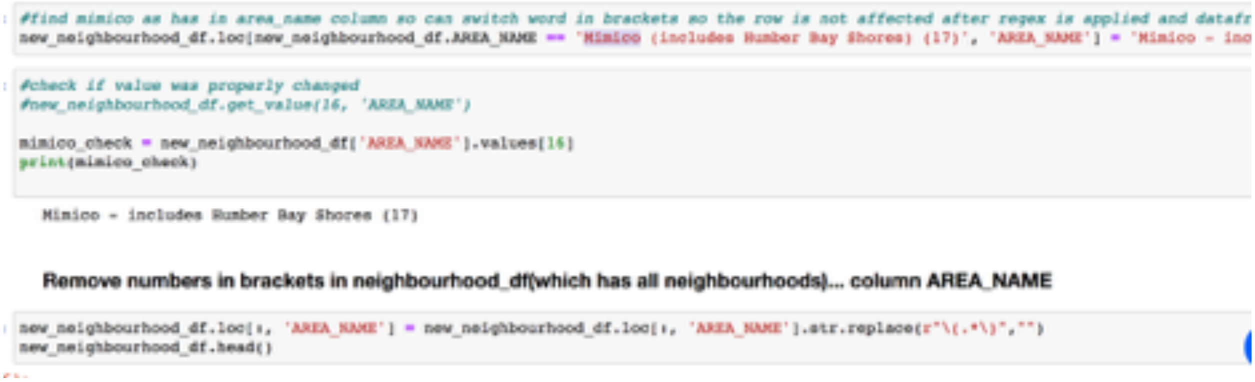


```
]: val_check = age_num['City of Toronto'].values[0]
   print(val_check)
   type(val_check)

   398135

18]: numpy.int64
```

One of the last steps I completed with the initial population dataframes(before linking them with the coordinate data) was using transpose to switch the columns and rows so that the neighbourhoods were all in the same column.

A few interesting things I took note of is pre-merging data frames using common column name 'AREA_NAME':

1 - when I was trying to remove the numbers in the brackets in the string in new_neighbourhood_df there was a value for the Mimico neighbourhood that would have caused the row to have not been included in the final dataframe had I not made a change. I used .loc to get the value and as it was the only one, used rename to change it.



```
#find mimico as has in area_name column so can switch word in brackets so the row is not affected after regex is applied and datafr
new_neighbourhood_df.loc[new_neighbourhood_df.AREA_NAME == 'Mimico (includes Humber Bay Shores) (17)', 'AREA_NAME'] = 'Mimico - inc
```

```
#check if value was properly changed
#new_neighbourhood_df.get_value(16, 'AREA_NAME')

mimico_check = new_neighbourhood_df['AREA_NAME'].values[16]
print(mimico_check)
```

```
Mimico - includes Humber Bay Shores (17)
```

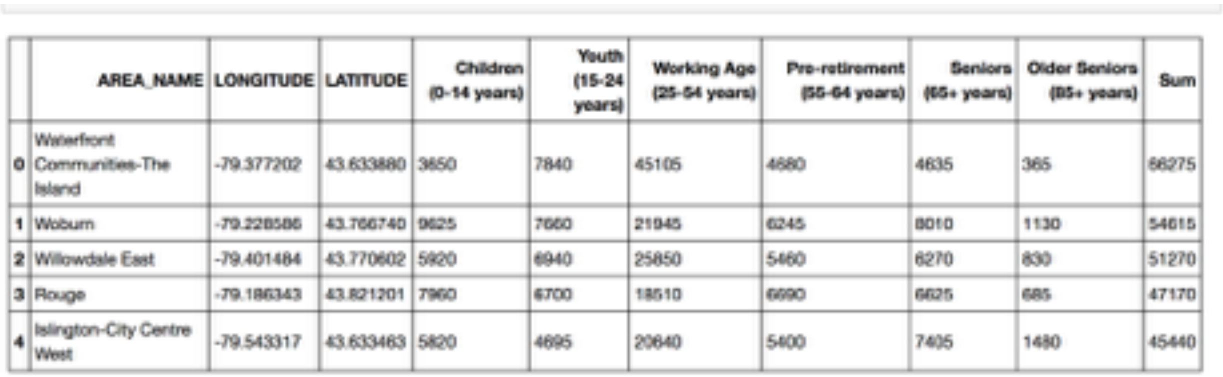
Remove numbers in brackets in neighbourhood_df(which has all neighbourhoods)... column AREA_NAME

```
new_neighbourhood_df.loc[:, 'AREA_NAME'] = new_neighbourhood_df.loc[:, 'AREA_NAME'].str.replace(r"\(.*\)","")
new_neighbourhood_df.head()
```

2 - had to further clean dataframe by removing the trailing white spaces

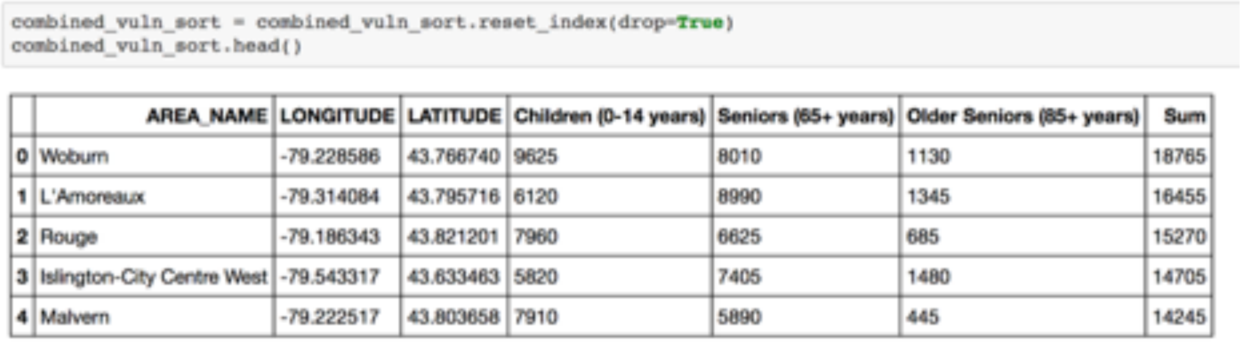
The final resulting dataframes ended up as follows:

All neighbourhoods



	AREA_NAME	LONGITUDE	LATITUDE	Children (0-14 years)	Youth (15-24 years)	Working Age (25-54 years)	Pre-retirement (55-64 years)	Seniors (65+ years)	Older Seniors (85+ years)	Sum
0	Waterfront Communities-The Island	-79.377202	43.633880	3650	7840	45105	4680	4635	365	66275
1	Woburn	-79.228586	43.766740	9625	7660	21945	6245	8010	1130	54615
2	Willowdale East	-79.401484	43.770602	5920	6940	25850	5460	6270	830	51270
3	Rouge	-79.186343	43.821201	7960	6700	18510	6690	6625	685	47170
4	Islington-City Centre West	-79.543317	43.633463	5820	4695	20640	5400	7405	1480	45440

Vulnerable Population Groupings

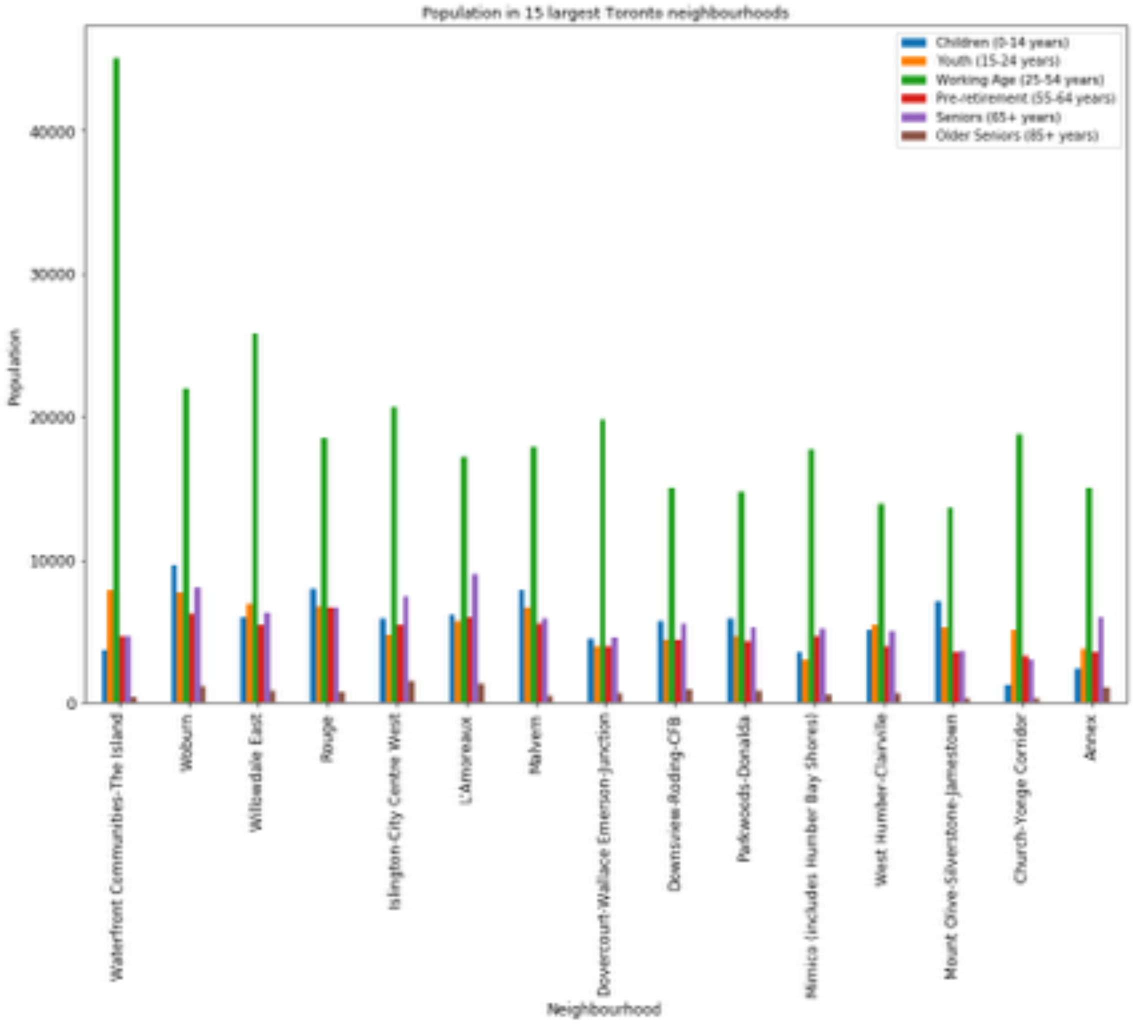


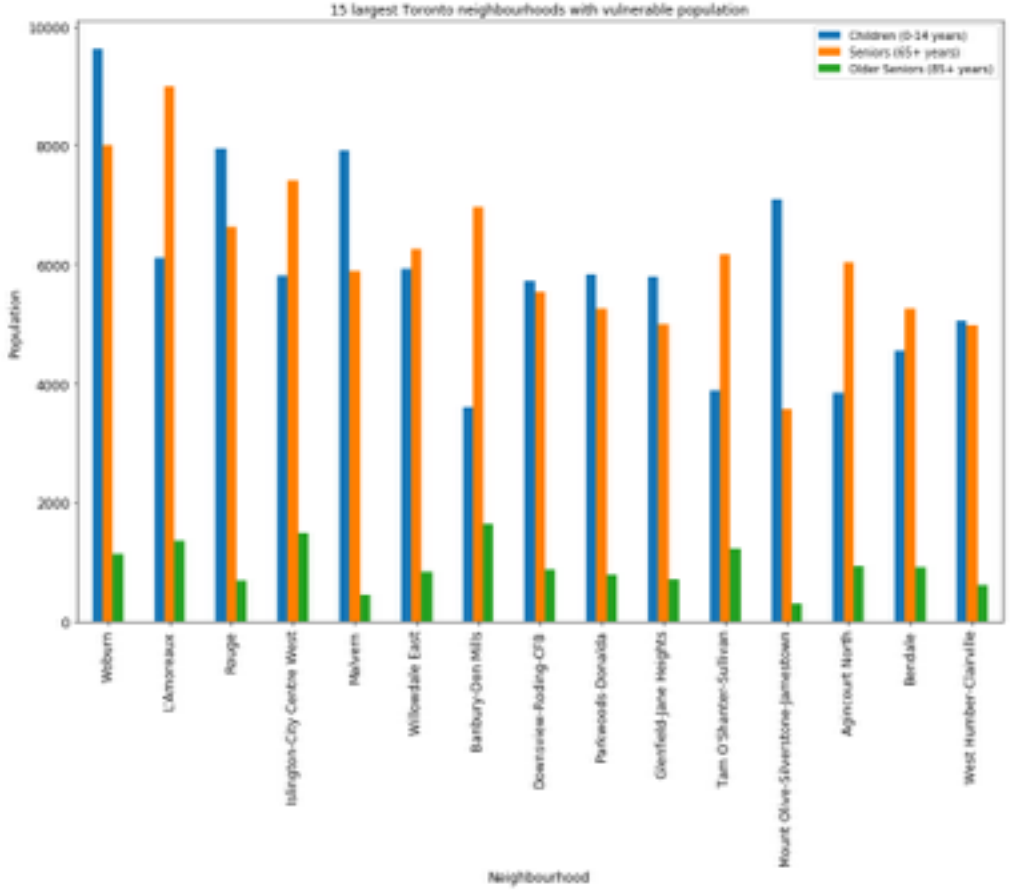
```
combined_vuln_sort = combined_vuln_sort.reset_index(drop=True)
combined_vuln_sort.head()
```

	AREA_NAME	LONGITUDE	LATITUDE	Children (0-14 years)	Seniors (65+ years)	Older Seniors (85+ years)	Sum
0	Woburn	-79.228586	43.766740	9625	8010	1130	18765
1	L'Amoreaux	-79.314084	43.795716	6120	8990	1345	16455
2	Rouge	-79.186343	43.821201	7960	6625	685	15270
3	Islington-City Centre West	-79.543317	43.633463	5820	7405	1480	14705
4	Malvern	-79.222517	43.803658	7910	5890	445	14245

As we wanted to compare between the top 15 neighbourhoods overall and the top 15 neighbourhoods with the highest vulnerable populations, I plotted two bar graphs to visualize the data.

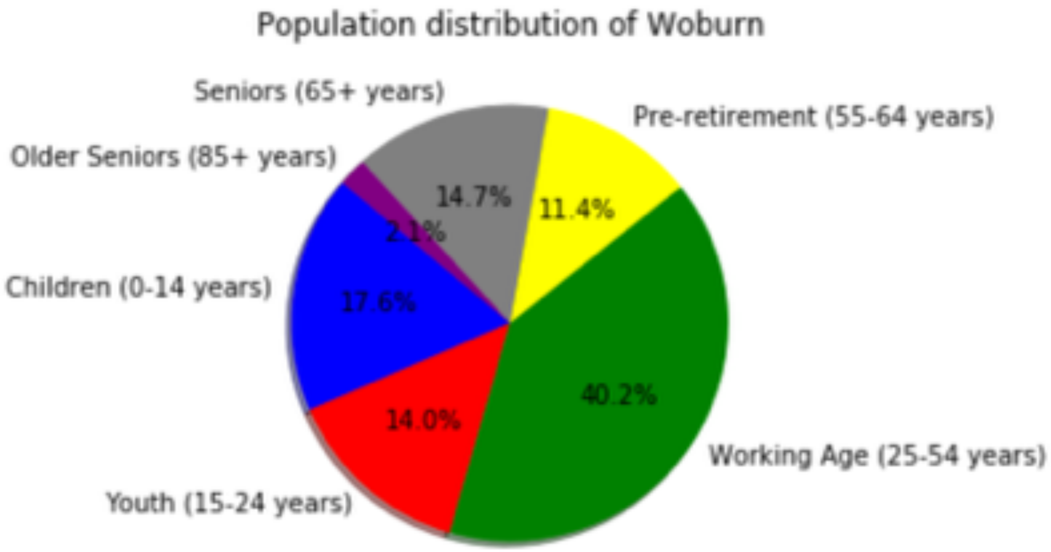
The graphs feature the neighbourhoods on the x-axis and the population on the y-axis and show the population spread between the various age groups to help more easily visualize which age group has a higher amount of people in certain neighbourhoods.

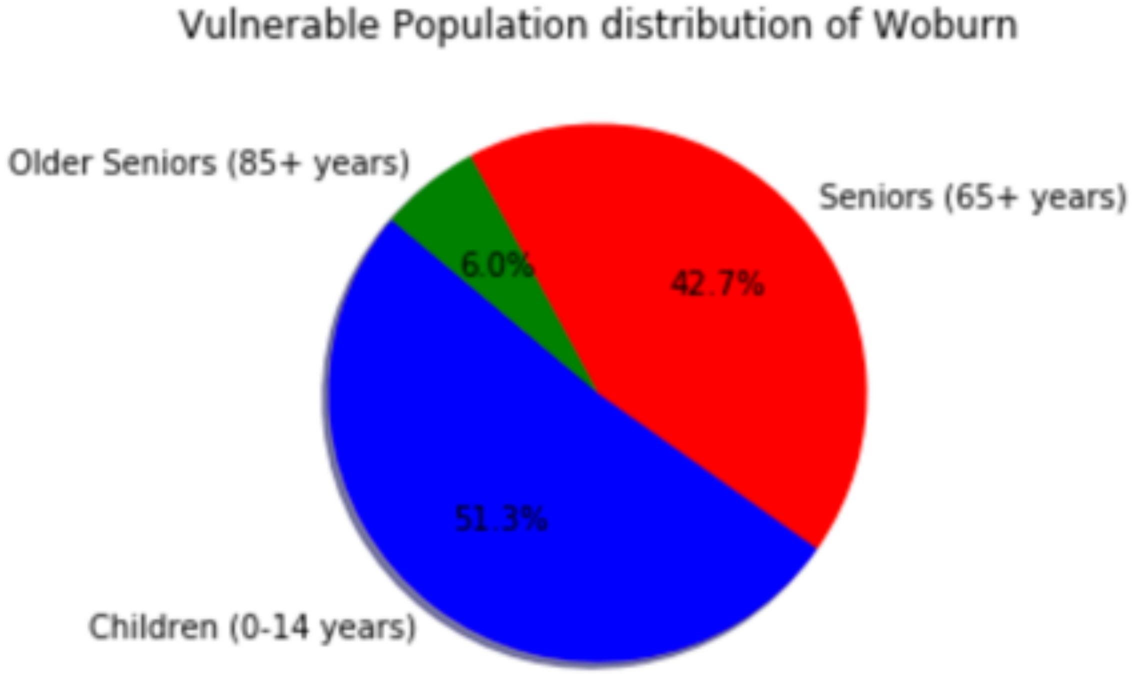

Population in 15 largest Toronto neighbourhoods
Children (0-14 years)
Youth (15-24 years)
Working Age (25-54 years)
Pre-retirement (55-64 years)
Seniors (65+ years)
Older Seniors (85+ years)
Population
0
10000
20000
30000
40000
Waterfront Communities-The Island
Woburn
Willowdale East
Rouge
Islington-City Centre West
L'Amoreaux
Malvern
Dovercourt-Wallace Emerson-Junction
Downsview-Roding-CFB
Parkwoods-Donalda
Mimico (includes Humber Bay Shores)
West Humber-Clairville
Mount Olive-Silverstone-Jamestown
Church-Yonge Corridor
Annex
Neighbourhood


15 largest Toronto neighbourhoods with vulnerable population
Children (0-14 years)
Seniors (65+ years)
Older Seniors (85+ years)
Population
0
2000
4000
6000
8000
10000
Woburn
L'Amoreaux
Rouge
Islington-City Centre West
Malvern
Willowdale East
Bathurst-Don Mills
Downsview-Roding-CFB
Parkwoods-Donalda
Glenfield-Jane Heights
Tam O'Shanter-Sullivan
Mount Olive-Silverstone-Jamestown
Agincourt North
Bendale
West Humber-Clairville
Neighbourhood

From these 2 bar graphs we can see that Woburn, Rouge, Islington City Centre West, Malvern, Downsview-Roding-CFB, Parkwoods-Donalda and Mount Olive-Silverstone-Jamestown overlap in terms of having a large population both in general as well as of the more vulnerable ages.

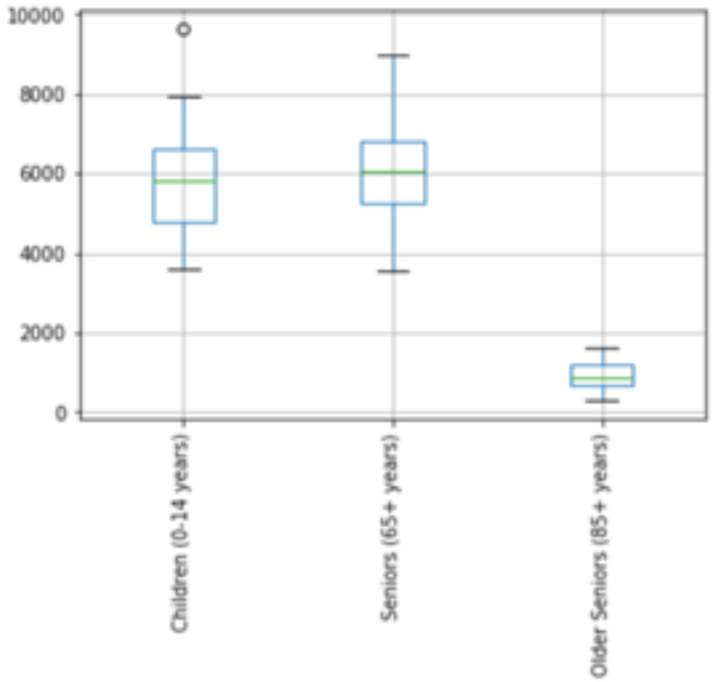
In a similar fashion to the bar graphs, I took one of the overlapping neighbourhoods - Woburn- to see the age distribution using a pie chart as well.



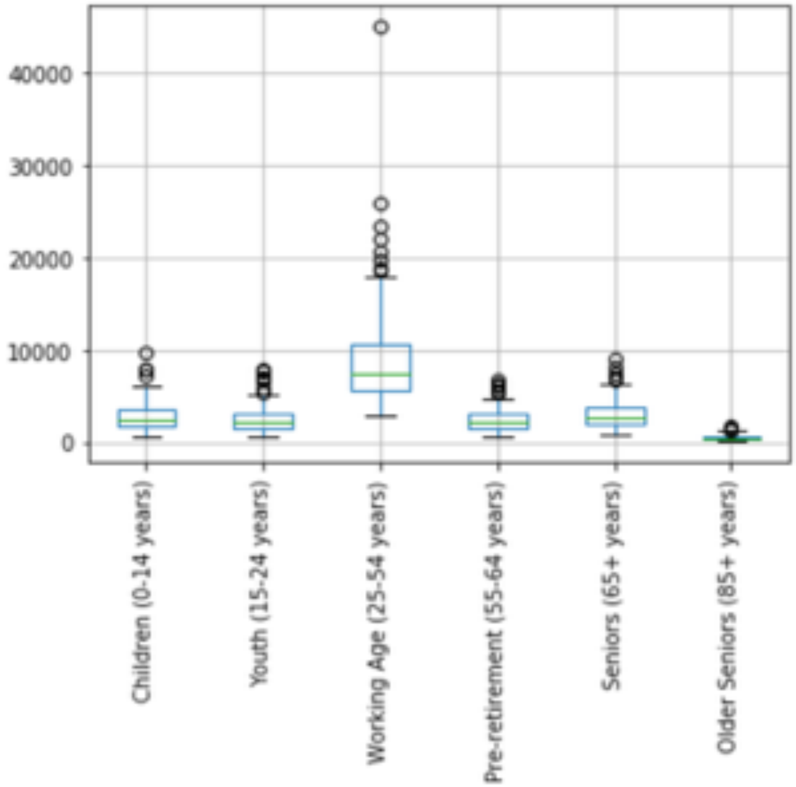


Lastly, boxplots were used to more easily visualize the statistical data (such as outliers, max, min, average, standard deviation) for the age groups amongst all the neighbourhoods.

Vulnerable groups:



All age groups:



From these box plots we can see that the majority of population in the neighbourhoods tend to be in the Working age range and then Children and Seniors appear to be the next two biggest age groups.

The Foursquare API search parameter was used along with the category ID for hospitals in order to place the data onto the Toronto map created using Folium. The JSON data from the get function result was converted into a dataframe and using the averages of the top 15 vulnerable population neighbourhoods and of all neighbourhoods as the latitude and longitude in the search request we were able to determine the distances away from the centre points using max radius of 5km.

Results

Note:

Blue circles - neighbourhoods

Red circles - neighbourhoods with highest amount of vulnerable population

Green circles - hospitals/ health centres within 5km of centre of all neighbourhoods

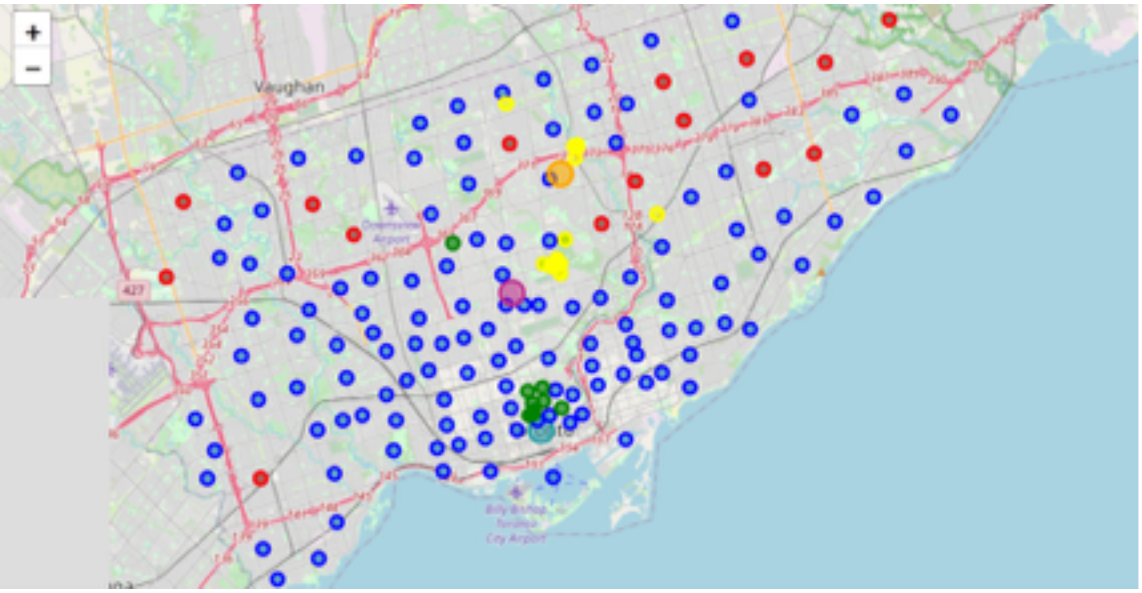
Yellow circles - hospitals/ health centres within 5km of centre of top 15 neighbourhoods with highest amount of vulnerable population

Centres:

Teal circle - Toronto, ON coordinates

Purple circle - centre of all neighbourhoods

Orange circle - centre of top 15 neighbourhoods with highest amount of vulnerable population



After having mapped out the data on the map, we can see that there is a lack of hospitals/health centres in the vicinity of the neighbourhoods on the left side of the map as we can not see any yellow or green points in that area so I would tentatively suggest placing one there after further research and approval from the hospital director and city.

Discussion

In this study, I made use of box plots, bar charts and Folium maps in order to visualize the population and neighbourhood data of Toronto juxtaposed with the locations of the hospitals in the nearby areas.

While it may not be a concrete reason, an observation that was noticed through the visualization of the bar charts is that in both graphs, for the neighbourhoods that are common, save for Islington City Centre West, the one thing these other 6 seem to all have in common is a larger number of children between the ages of 0 and 14.

As was noticed in the hospital data frames, there were many hospitals that actually overlapped. They were listed as different venues, and as such had different id's on Foursquare but were in fact different parts of the same hospital. This should be taken into account when judging the amount of hospitals in the area as viewed on the map in the future as well.

Conclusion

In conclusion, based on our observations through the data we analyzed, the differences between total population and vulnerable population and the number of medical centres in the respective areas, we would suggest that the hospital be built around the neighbourhoods more in the upper left areas of Toronto - preferably in an area in the middle(possibly achieved by averaging their coordinates and verifying the locations) of some of the top 15 vulnerable age populations.

In the future, more research in regards to the type of hospital or clinic would be ideally given as this analysis takes an approach towards a more general medical centre. We would also expand the definition of vulnerable to include other groups such as immunocompromised and the like.

References

- [1] Social Development, Finance & Administration. (last refreshed 2020). Neighbourhoods [Data file]. Retrieved from https://open.toronto.ca/dataset/neighbourhoods/

  - **Attribution:** Open Data License - Toronto; *Contains information licensed under the Open Government Licence – Toronto*

- [2] Social Development, Finance & Administration. (last refreshed 2019). Neighbourhood Profiles [Data file]. Retrieved from https://open.toronto.ca/dataset/neighbourhood-profiles/

  - **Attribution:** Open Data License - Toronto; *Contains information licensed under the Open Government Licence – Toronto*

- [3] Foursquare API - The website for the Foursquare API can be found at https://developer.foursquare.com/