Opening and Italian Restaurant in Toronto

- Capstone Project Week 5

Business Problem

There are a lot of neighbourhoods in Toronto

- Which neighbourhood would be ideal?
- How do we define ideal?

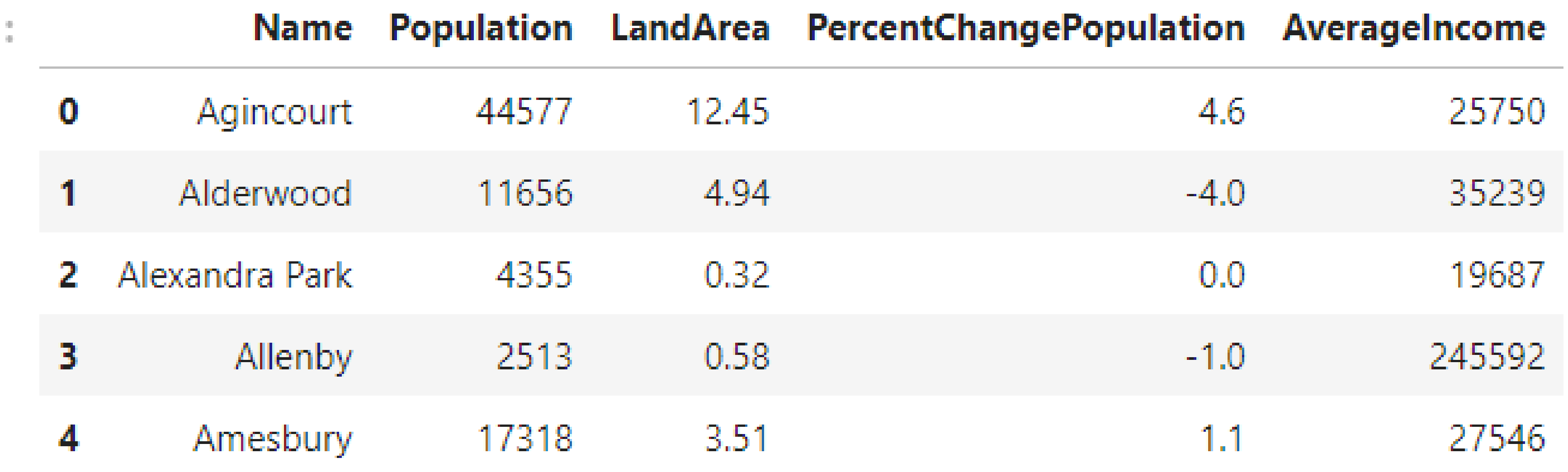
Metrics:

1) High population density
2) High average income
3) Low percentage of Italian restaurants
4) Low score of current Italian restaurants

Data Collection

We can read demographic data from Wikipedia for each neighbourhood

- We mainly care about name, population, land area, and average income



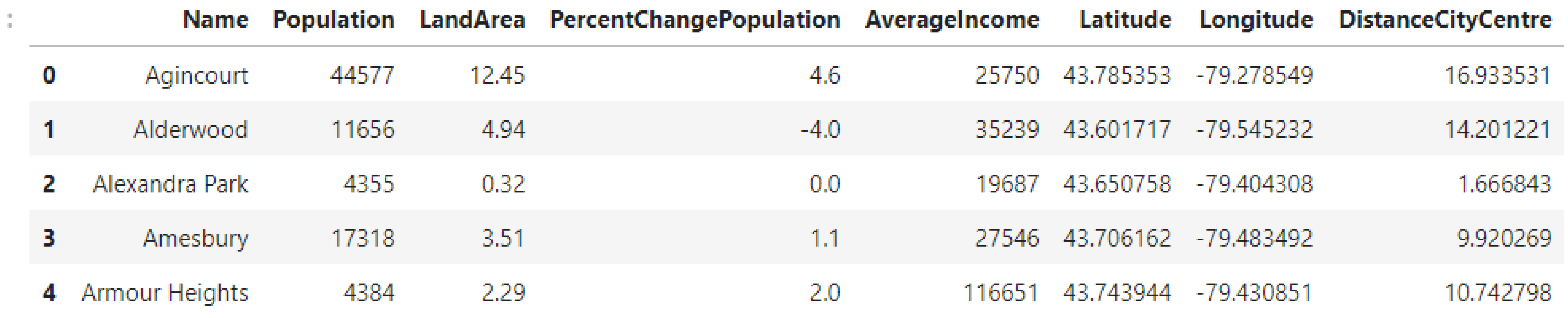
	Name	Population	LandArea	PercentChangePopulation	AverageIncome
0	Agincourt	44577	12.45	4.6	25750
1	Alderwood	11656	4.94	-4.0	35239
2	Alexandra Park	4355	0.32	0.0	19687
3	Allenby	2513	0.58	-1.0	245592
4	Amesbury	17318	3.51	1.1	27546

https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

Coordinates

We need the latitude and longitude of each neighbourhood

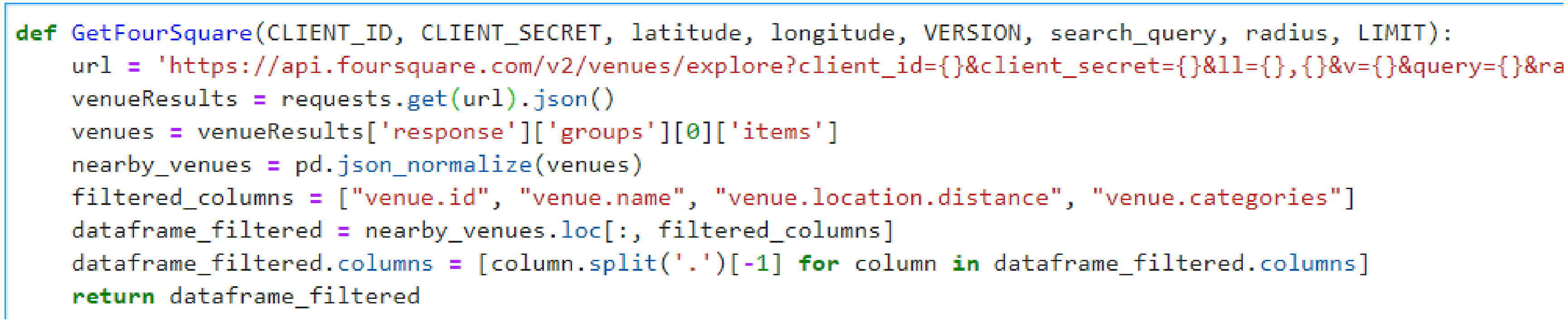
- We can use Geolocator to get the coordiantes
- We can use distance to city centre to remove far away locations



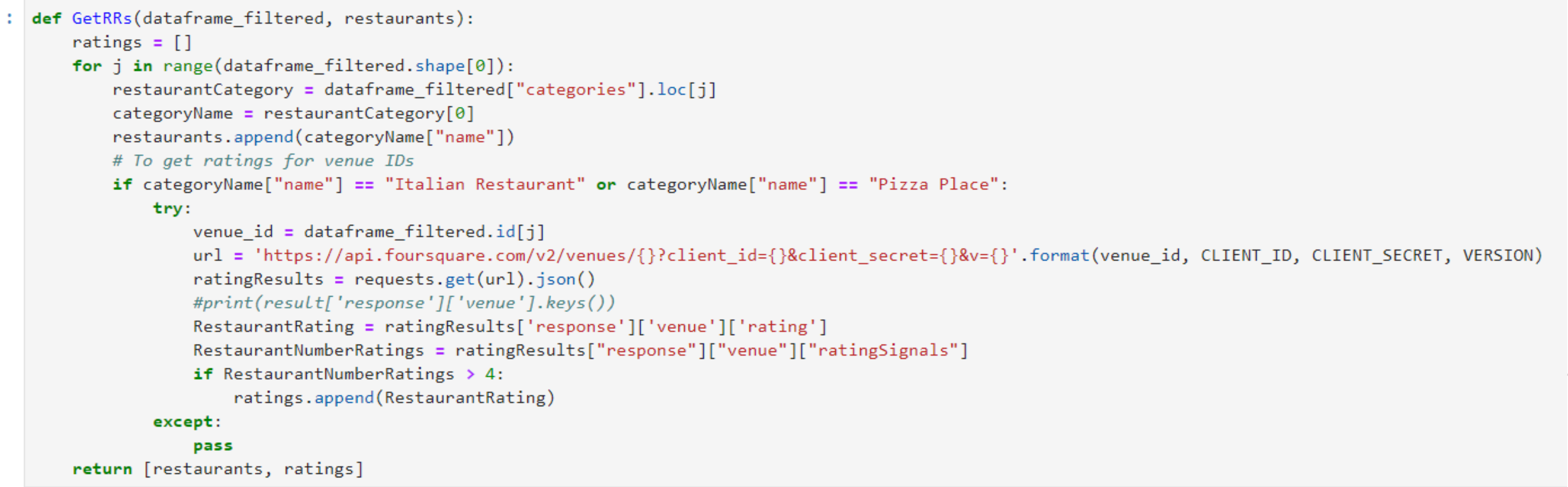
	Name	Population	LandArea	PercentChangePopulation	AverageIncome	Latitude	Longitude	DistanceCityCentre
0	Agincourt	44577	12.45	4.6	25750	43.785353	-79.278549	16.933531
1	Alderwood	11656	4.94	-4.0	35239	43.601717	-79.545232	14.201221
2	Alexandra Park	4355	0.32	0.0	19687	43.650758	-79.404308	1.666843
3	Amesbury	17318	3.51	1.1	27546	43.706162	-79.483492	9.920269
4	Armour Heights	4384	2.29	2.0	116651	43.743944	-79.430851	10.742798

FourSquare

We can define some simple functions to retrieve restaurants from FourSquare

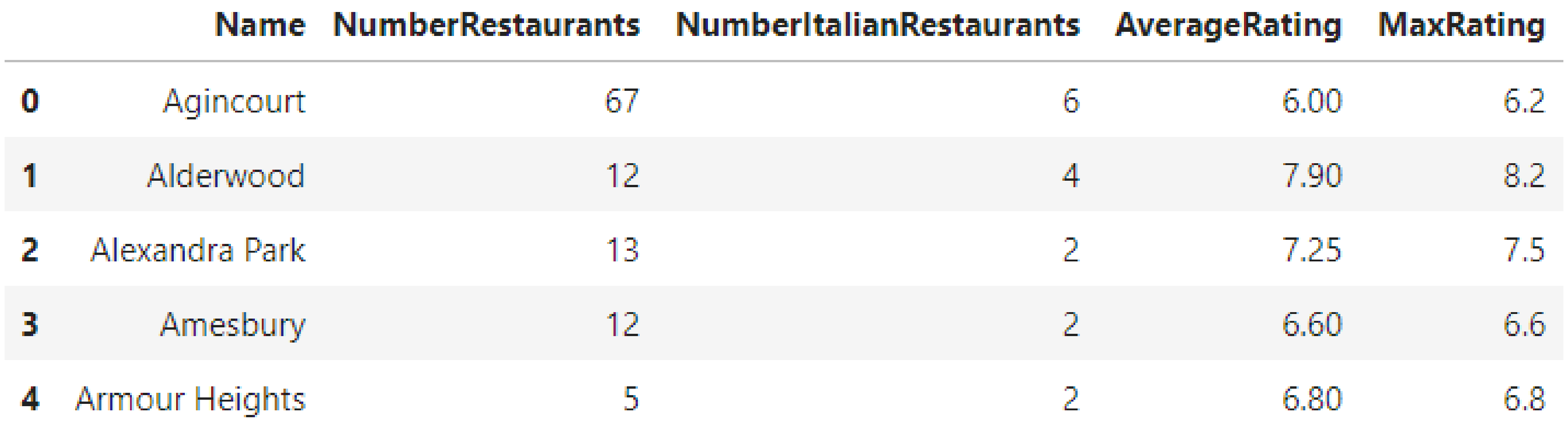


```
def GetFourSquare(CLIENT_ID, CLIENT_SECRET, latitude, longitude, VERSION, search_query, radius, LIMIT):
    url = 'https://api.foursquare.com/v2/venues/explore?client_id={}&client_secret={}&ll={},{}&v={}&query={}&ra
    venueResults = requests.get(url).json()
    venues = venueResults['response']['groups'][0]['items']
    nearby_venues = pd.json_normalize(venues)
    filtered_columns = ["venue.id", "venue.name", "venue.location.distance", "venue.categories"]
    dataframe_filtered = nearby_venues.loc[:, filtered_columns]
    dataframe_filtered.columns = [column.split('.')[-1] for column in dataframe_filtered.columns]
    return dataframe_filtered
```



```
def GetRRs(dataframe_filtered, restaurants):
    ratings = []
    for j in range(dataframe_filtered.shape[0]):
        restaurantCategory = dataframe_filtered["categories"].loc[j]
        categoryName = restaurantCategory[0]
        restaurants.append(categoryName["name"])
        # To get ratings for venue IDs
        if categoryName["name"] == "Italian Restaurant" or categoryName["name"] == "Pizza Place":
            try:
                venue_id = dataframe_filtered.id[j]
                url = 'https://api.foursquare.com/v2/venues/{}?client_id={}&client_secret={}&v={}'.format(venue_id, CLIENT_ID, CLIENT_SECRET, VERSION)
                ratingResults = requests.get(url).json()
                #print(result['response']['venue'].keys())
                RestaurantRating = ratingResults['response']['venue']['rating']
                RestaurantNumberRatings = ratingResults["response"]["venue"]["ratingSignals"]
                if RestaurantNumberRatings > 4:
                    ratings.append(RestaurantRating)
            except:
                pass
    return [restaurants, ratings]
```

Venue Information

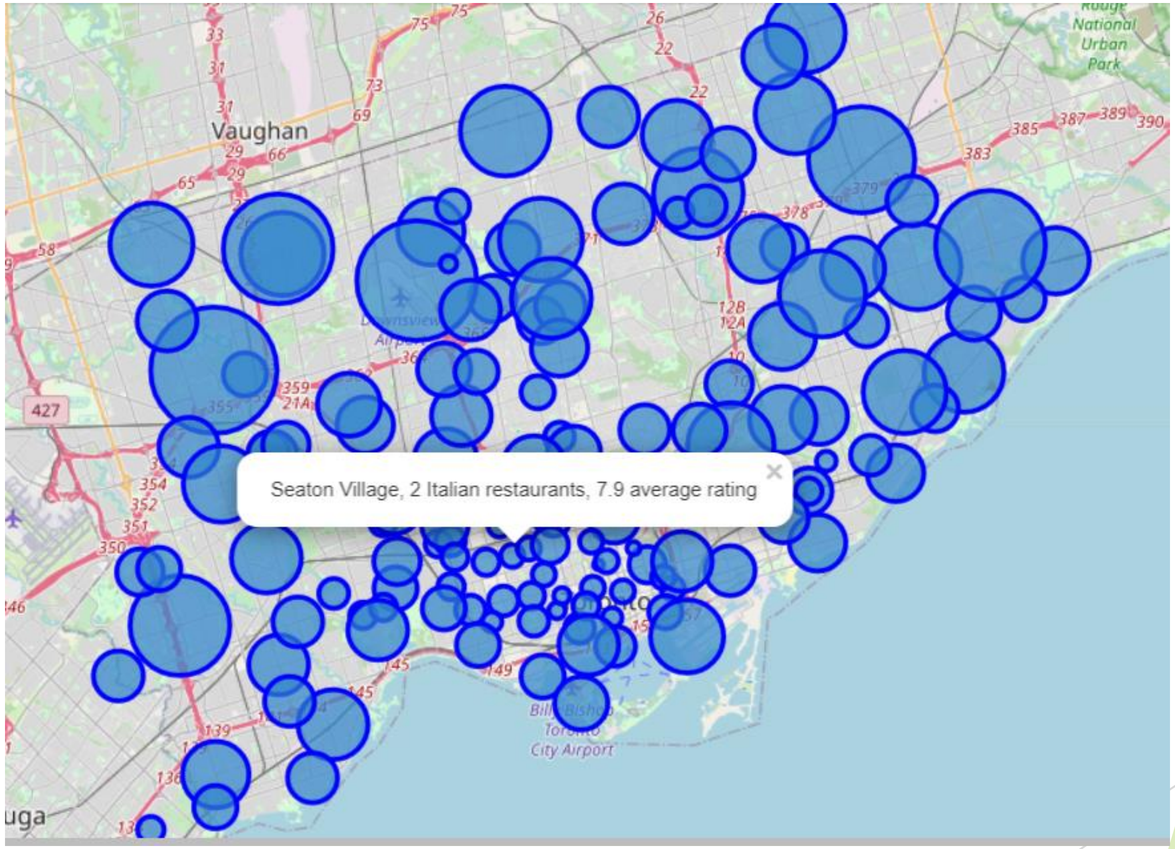


	Name	NumberRestaurants	NumberItalianRestaurants	AverageRating	MaxRating
0	Agincourt	67	6	6.00	6.2
1	Alderwood	12	4	7.90	8.2
2	Alexandra Park	13	2	7.25	7.5
3	Amesbury	12	2	6.60	6.6
4	Armour Heights	5	2	6.80	6.8

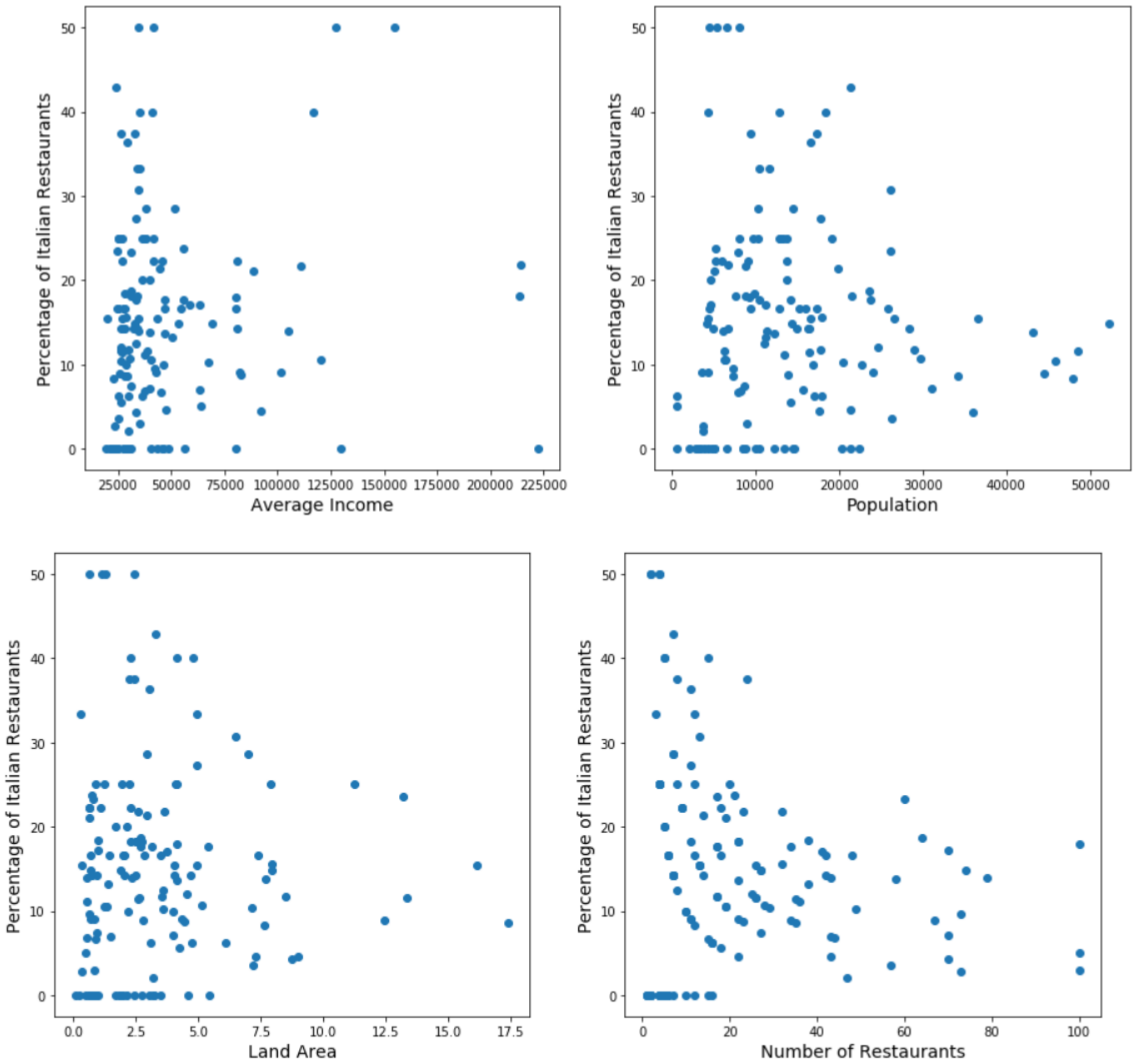
- The FourSquare data was used to generate percentage of Italian restaurants and the average rating of the Italian restaurants.

Map of Toronto neighbourhoods

- 402 Italian restaurants found
- Neighborhoods shown on map with sizes based on land area
- Can get a good idea of the layout of neighborhoods and restaurants

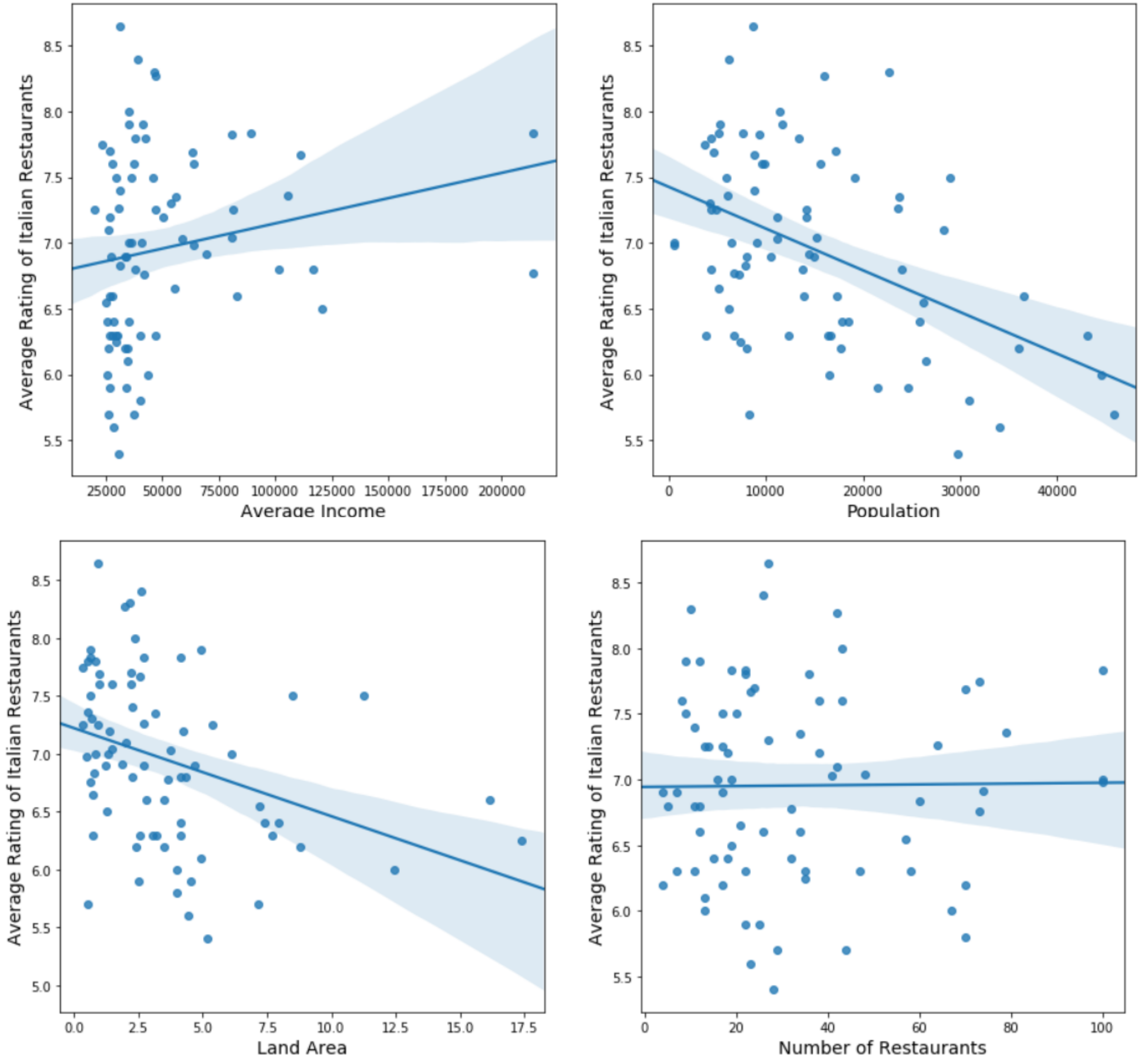

Seaton Village, 2 Italian restaurants, 7.9 average rating
Vaughan
Rouge National Urban Park
Billy Bishop Toronto City Airport
Downsview Airport
427

Correlation - % of Italian Restaurants



- Very little correlation
- This was not pursued further

Correlation – Average Rating



- Some correlation with population and land area

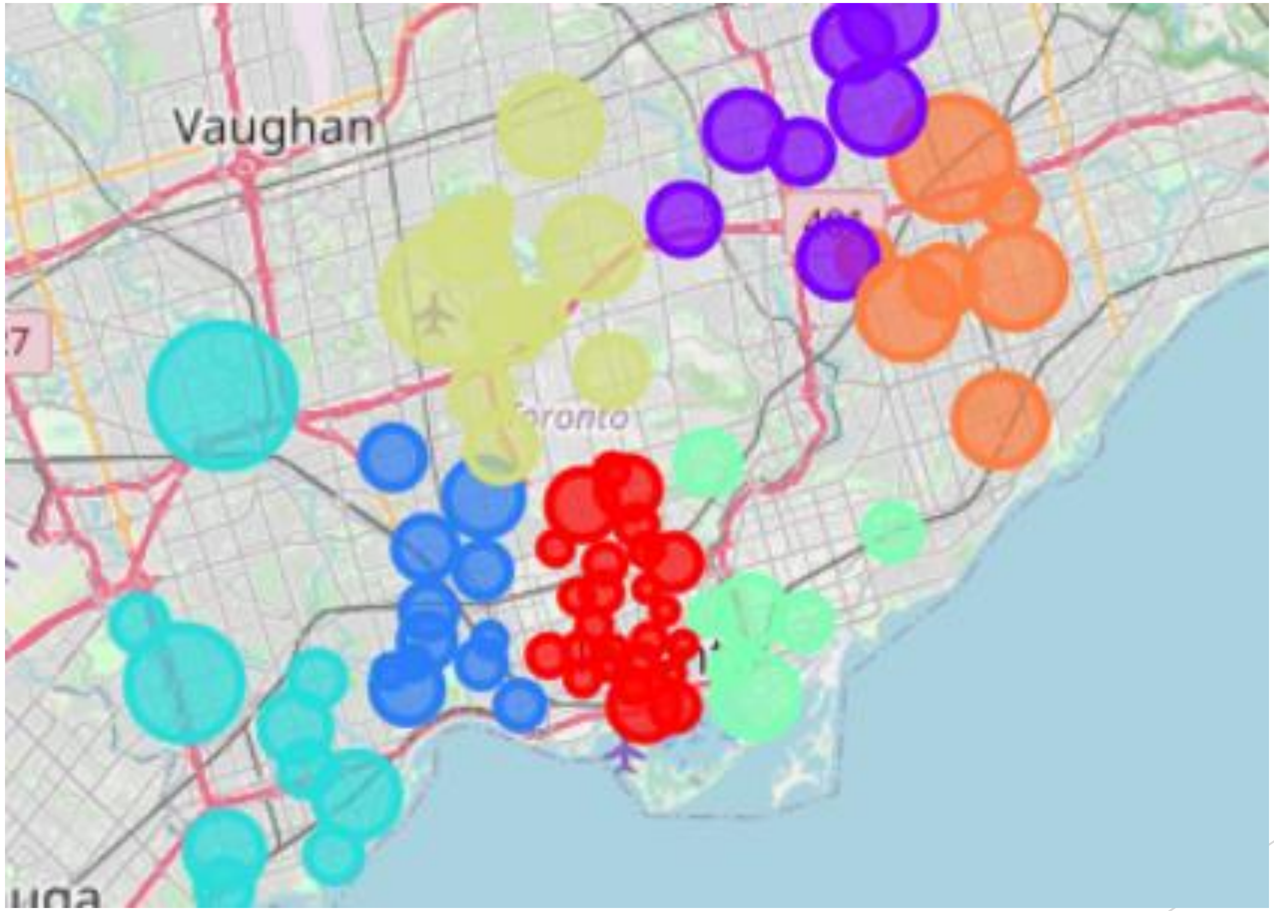
- Small sample size for some neighbourhoods, so data might be skewed

- Maybe clustering can help with this

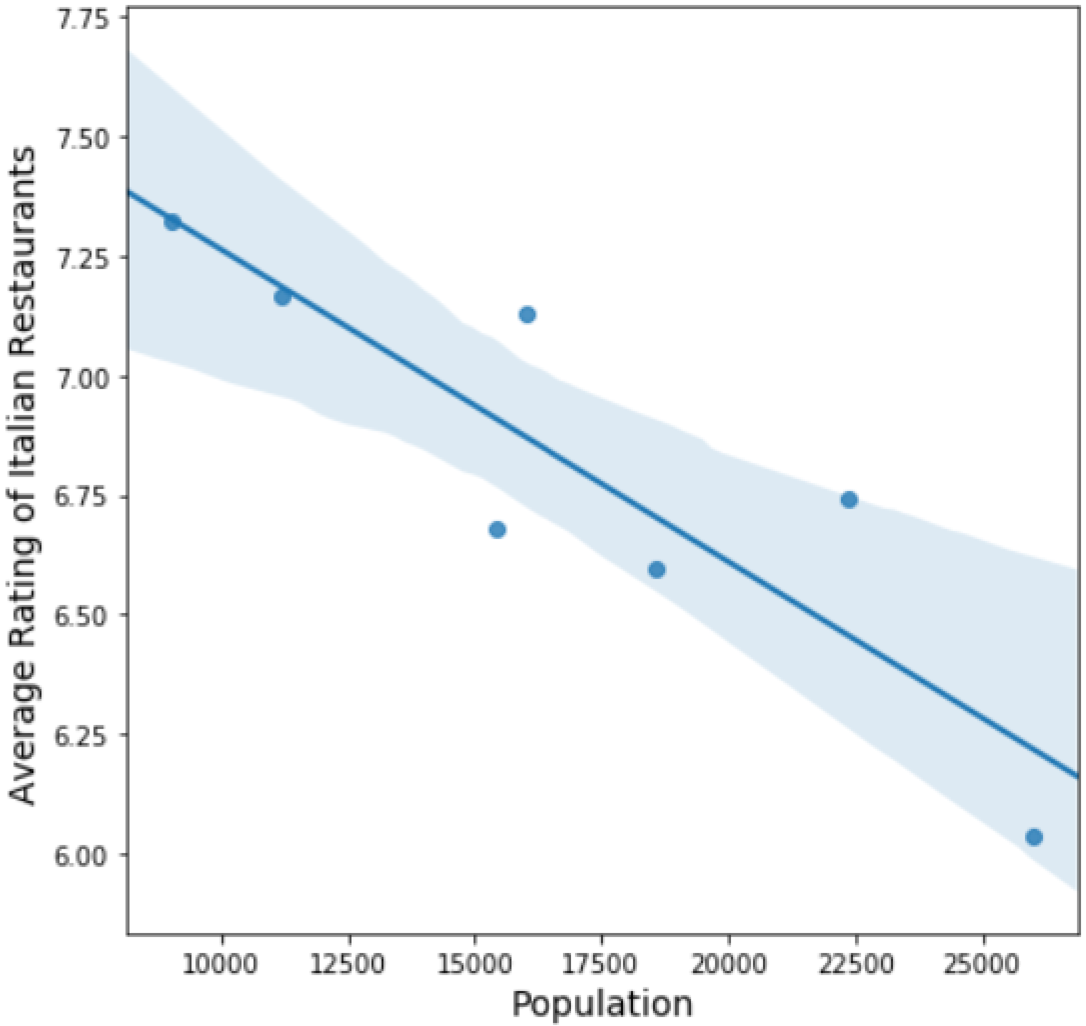
K Clusters

- Break neighbourhoods into 7 clusters

- See if this helps when correlating demographic data to average rating

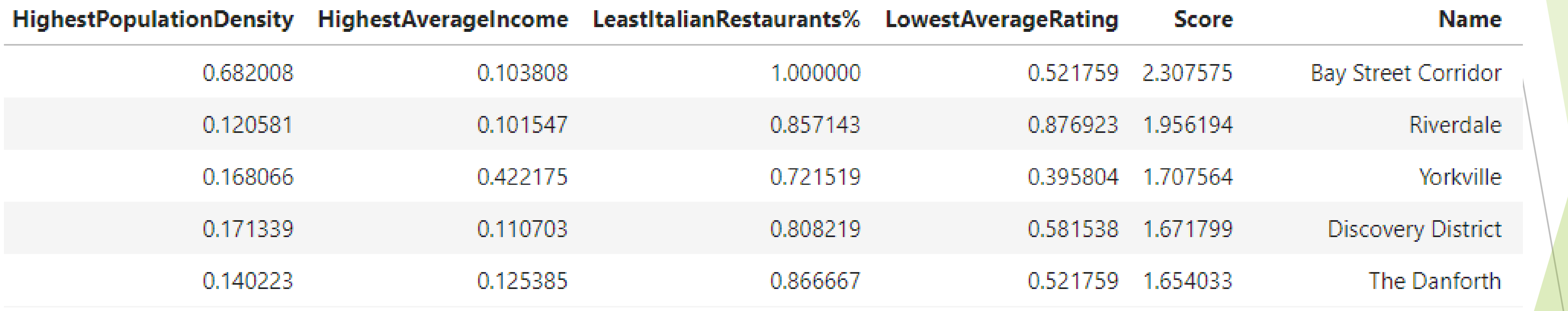


Clustered correlation



- When the neighbourhoods are clustered, population is clearly inversely proportional to the average rating

Most Ideal Neighbourhoods

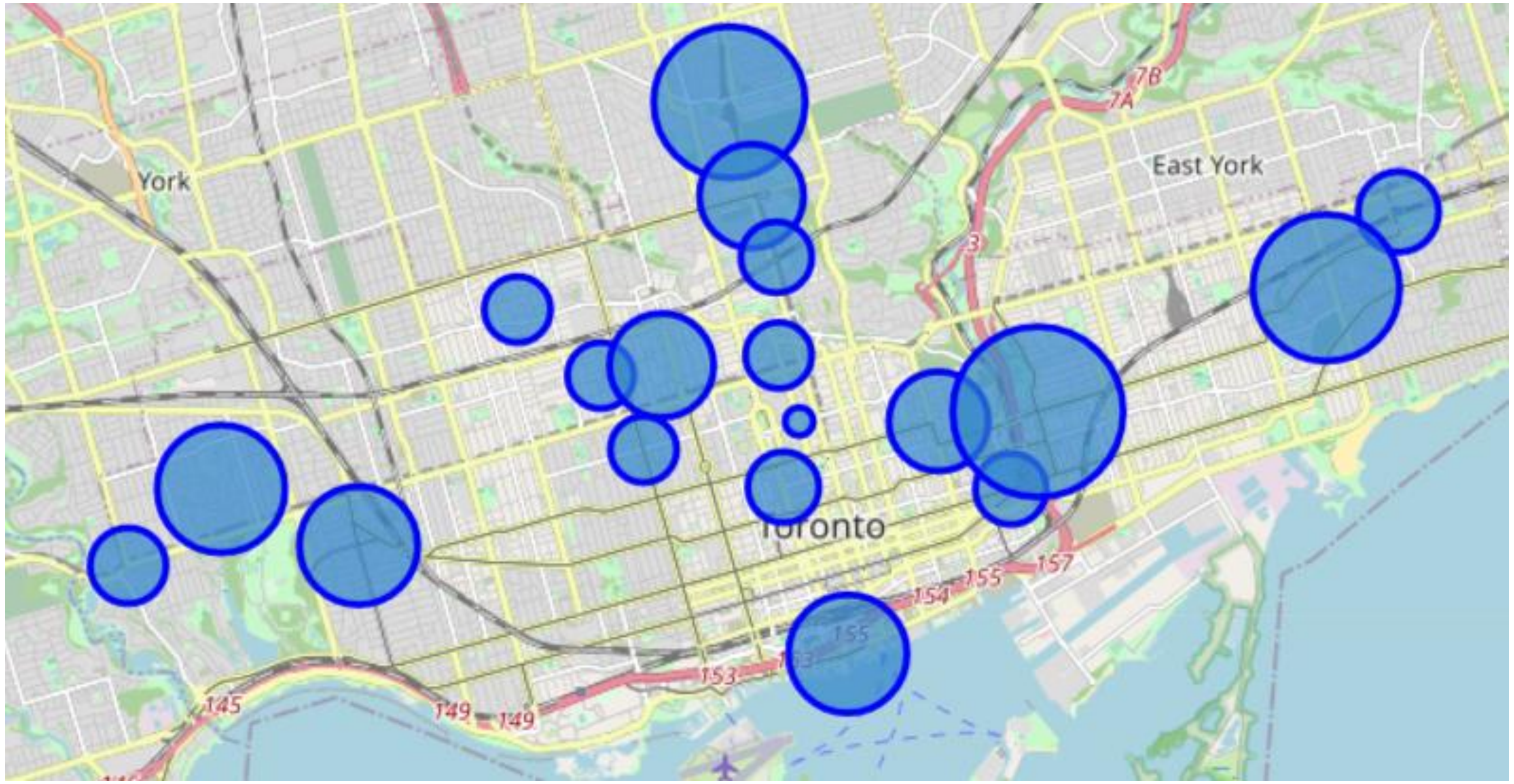


HighestPopulationDensity	HighestAverageIncome	LeastItalianRestaurants%	LowestAverageRating	Score	Name
0.682008	0.103808	1.000000	0.521759	2.307575	Bay Street Corridor
0.120581	0.101547	0.857143	0.876923	1.956194	Riverdale
0.168066	0.422175	0.721519	0.395804	1.707564	Yorkville
0.171339	0.110703	0.808219	0.581538	1.671799	Discovery District
0.140223	0.125385	0.866667	0.521759	1.654033	The Danforth

We can rank the neighbourhoods!

- Equal weighting between the four metrics

Proximity of ideal neighbourhoods



Maybe a good idea to pick an ideal location near other ideal locations!

Conclusion

- Neighbourhood data was analysed for demographics, location, number of Italian restaurants, and average rating

- Very little correlation between demographic data and % Italian restaurant and average rating
  - Some correlation between population and average rating of Italian restaurants
  - Clustering made this more obvous

- Ideal neighbourhoods were obtained by aggregated scores for each metric
  - Bay Street Corridor, Riverdale, and Yorkville were the top three choices