Toronto Neighborhood Demographics and the Potential for Restaurant Growth

Robert Morris

Introduction and Project Goal

This project seeks to answer the question: can the demographics of a neighborhood be useful in deciding whether a new restaurant of a given kind will be successful there? To answer this question quantitatively, we combined two databased: one based on Foursquare venue information, and the other one that provides demographic information about a city and its neighborhoods, in this case Toronto.

Data Sources and Preprocessing

For the demographics data we extracted a table on the Wikipedia page called “Demographics of Toronto neighbourhoods”. We reduced the columns so that the dataframe looked like this:

Out[89]:

	Name	Population	Average Income	Transit Commuting %	% Renters	2nd language
1	Agincourt	44577	25750	11.1	5.9	Cantonese (19.3%)
2	Alderwood	11656	35239	8.8	8.5	Polish (6.2%)
3	Alexandra Park	4355	19687	13.8	28.0	Cantonese (17.9%)
4	Allenby	2513	245592	5.2	3.4	Russian (1.4%)
5	Amesbury	17318	27546	16.4	19.7	Spanish (6.1%)

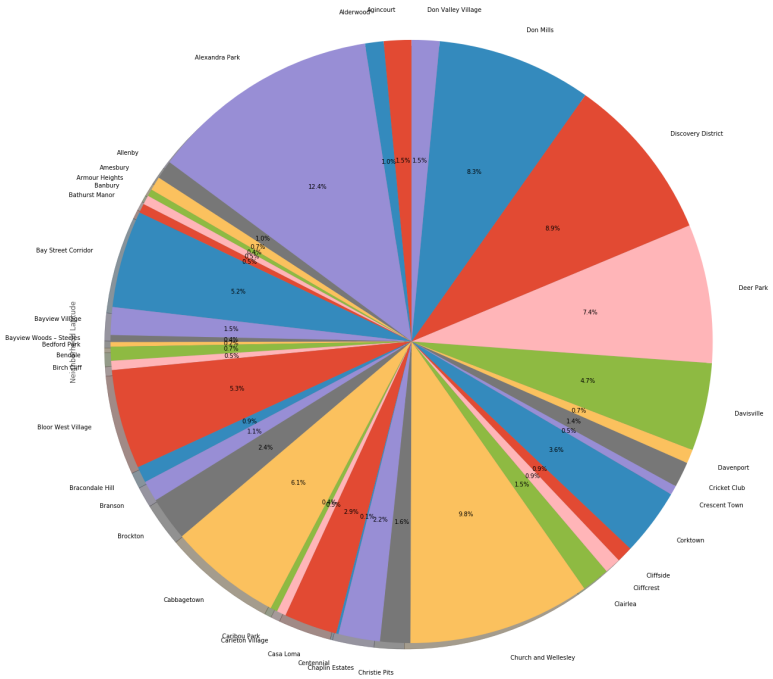
For each Toronto neighborhood, a row of this table shows its population, average income, and, just for fun, the percentage of people who rent their place of residence and who take public transportation. We also kept the column showing the second language (after English) of residents of the neighborhood.

Then, using the same techniques we used in class for New York neighborhoods, we queried and processed Foursquare data. The initial processing resulted in a grouping of neighborhoods in terms of venues. Here is the result:

Out[113]:

	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Agincourt	Chinese Restaurant	Coffee Shop	Korean Restaurant	Shopping Mall	Cantonese Restaurant	Asian Restaurant	Train Station	Hong Kong Restaurant	Food Court	Vietnamese Restaurant
1	Alderwood	Pizza Place	Dance Studio	Pub	Pharmacy	Coffee Shop	Gym	Sandwich Place	Donut Shop	Filipino Restaurant	Fast Food Restaurant
2	Alexandra Park	Bar	Furniture / Home Store	Caribbean Restaurant	Arts & Crafts Store	Coffee Shop	Café	Pizza Place	Boutique	Italian Restaurant	Poutine Place
3	Allenby	African Restaurant	Bookstore	Restaurant	Big Box Store	Fish & Chips Shop	Intersection	Café	Fast Food Restaurant	Yoga Studio	Ethiopian Restaurant
4	Amesbury	Bank	Gas Station	Coffee Shop	Intersection	Park	Athletics & Sports	Yoga Studio	Flower Shop	Fish Market	Fish & Chips Shop

We also made a pie chart that shows where most of the venues are located. This will be useful as a way to distinguish between neighborhoods that are 'residential' (with fewer venues) from those that are 'industrial' (with many venues). Here's the resulting pie chart:



For example, the Alexandra Park neighborhood has the highest percentage of venues, so it is perhaps the most industrial neighborhood. Indeed if you look at the demographics data this neighborhood has less that 5000 residents, so it one of the smaller neighborhoods in terms of residences.

Analysis and Neighborhood Clusters

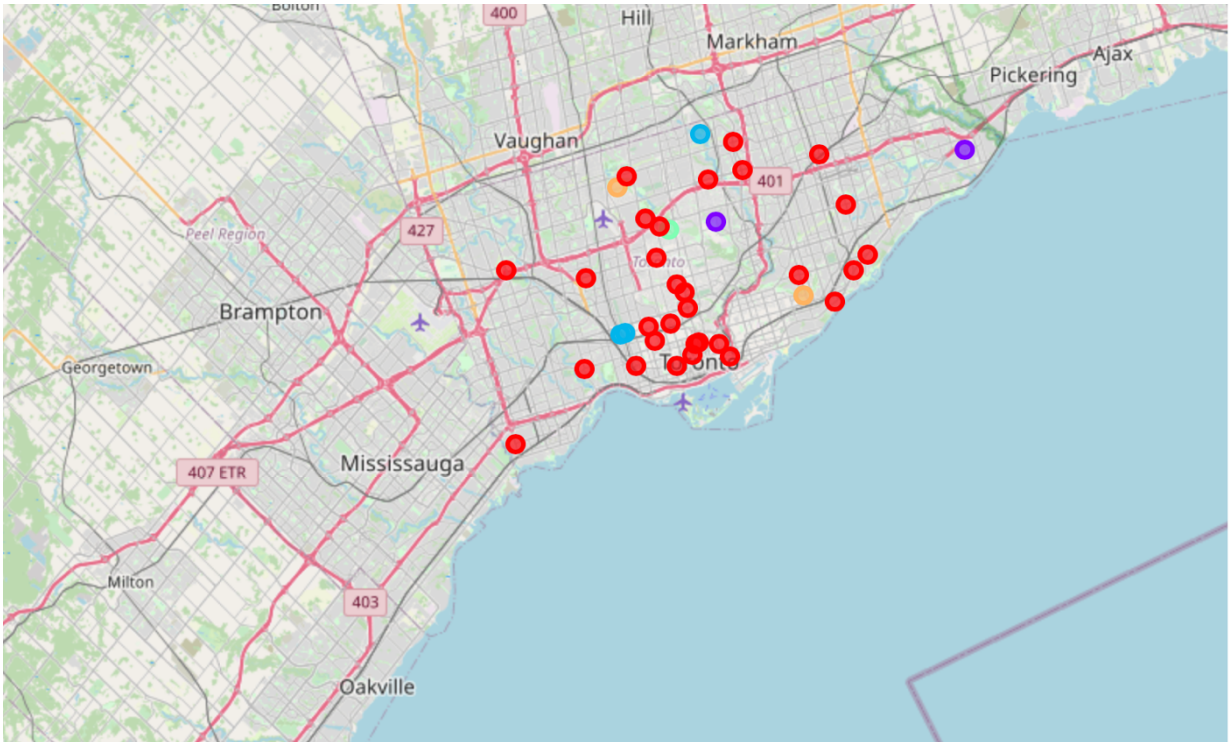
Using the same methods as the one done in the class, we analyzed and clustered the neighborhoods in terms of the venue type that was found in each neighborhood. There are 192 unique categories of venue in Toronto, and the distribution of each venue category was computed and placed in a table as shown:

```
In [119]:
toronto_grouped = t_onehot.groupby('Neighborhood').mean().reset_index()
toronto_grouped
```

Out[119]:

	Neighborhood	Afghan Restaurant	African Restaurant	American Restaurant	Animal Shelter	Arcade	Arepa Restaurant	Art Gallery	Art Museum	Arts & Crafts Store	...	Train Station	Udon Restaurant	Vegetarian / Vegan Restaurant	Vi Ga St
0	Agincourt	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.083333	0.00	0.000000	0.000
1	Alderwood	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
2	Alexandra Park	0.000000	0.000	0.010000	0.000000	0.01	0.02	0.020000	0.000000	0.030000	...	0.000000	0.01	0.020000	0.000
3	Allenby	0.000000	0.125	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
4	Amesbury	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
5	Armour Heights	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
6	Banbury	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
7	Bathurst Manor	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
8	Bay Street Corridor	0.023810	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.023810	...	0.000000	0.00	0.000000	0.000
9	Bayview Village	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
10	Bayview Woods – Steeles	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
11	Bedford Park	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
12	Bendale	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000
13	Birch Cliff	0.000000	0.000	0.000000	0.000000	0.00	0.00	0.000000	0.000000	0.000000	...	0.000000	0.00	0.000000	0.000

These vectors of values for each row provided the inputs to a k-means clustering algorithm (k=5) which tried to identify neighborhoods that were ‘close’ to others in terms of their venue profiles. Here’s a Toronto map with the 5 clusters in different colors:



Creating and Querying Merged Data

The merged demographic and Foursquare venue data, including the neighborhood's cluster number, looks like this:

Name	Population	Average Income	Transit Commuting %	% Renters	2nd language	lat	long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Mo Comm Ven
Agincourt	44577	25750	11.1	5.9	Cantonese	43.785353	-79.278549	0	Chinese Restaurant	Coffee Shop	Korean Restaurant	Shopping Mall	Cantone Restaura
Alderwood	11656	35239	8.8	8.5	Polish	43.601717	-79.545232	0	Pizza Place	Dance Studio	Pub	Pharmacy	Coff Sh
Alexandra Park	4355	19687	13.8	28.0	Cantonese	43.650758	-79.404308	0	Bar	Furniture / Home Store	Caribbean Restaurant	Arts & Crafts Store	Coff Sh
Allenby	2513	245592	5.2	3.4	Russian	43.711351	-79.553424	0	African Restaurant	Bookstore	Restaurant	Big Box Store	Fish Chi Sh
Amesbury	17318	27546	16.4	19.7	Spanish	43.706162	-79.483492	0	Bank	Gas Station	Coffee Shop	Intersection	Pa
Armour Heights	4384	116651	10.8	16.1	Russian	43.743944	-79.430851	0	Deli / Bodega	Market	Pharmacy	Yoga Studio	Electroni Sto
Banbury	6641	92319	6.1	4.8	Unspecified Chinese	43.742796	-79.369957	1	Park	Auto Garage	Tennis Court	Yoga Studio	Flow Sh
Bathurst Manor	14945	34169	13.4	18.6	Russian	43.763893	-79.456367	4	Convenience Store	Playground	Park	Baseball Field	Yo Stu
Bay Street Corridor	4787	40598	17.1	49.3	Mandarin	43.665272	-79.387531	0	Sushi Restaurant	Japanese Restaurant	Bubble Tea Shop	Mediterranean Restaurant	Yo Stu
Bayview Village	12280	46752	14.4	15.6	Cantonese	43.769197	-79.376662	0	Bank	Pizza Place	Sandwich Place	Sporting Goods Shop	Fast Fo Restaura

This table allows for complex queries to be formulated that might be useful for potential restaurateurs. For example, here's a query to find neighborhoods with an average income greater than 40000 and with second language Cantonese:

In [197]:
```
toronto_merged.loc[(toronto_merged['2nd language'] == 'Cantonese ') & (toronto_merged['Average Income'] > 40000)]
#toronto_merged['2nd language']
```

Out[197]:

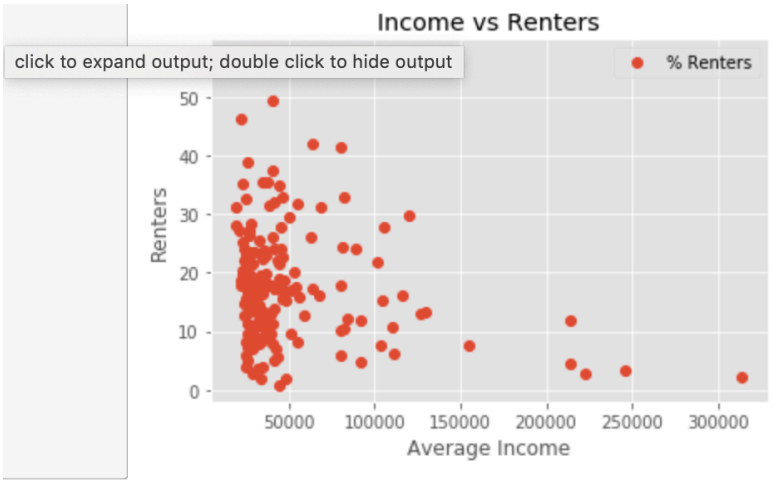
	Name	Population	Average Income	Transit Commuting %	% Renters	2nd language	lat	long	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Cor
9	Bayview Village	12280	46752	14.4	15.6	Cantonese	43.769197	-79.376662	0	Bank	Pizza Place	Sandwich Place	Sporting Goods Shop	Fast Food Restaurant	
10	Bayview Woods – Steeles	13298	41485	11.2	13.9	Cantonese	43.798127	-79.382973	2	Dog Run	Park	Trail	Eastern European Restaurant	Flower Shop	

In [154]:
```
import matplotlib.cm as cm
```

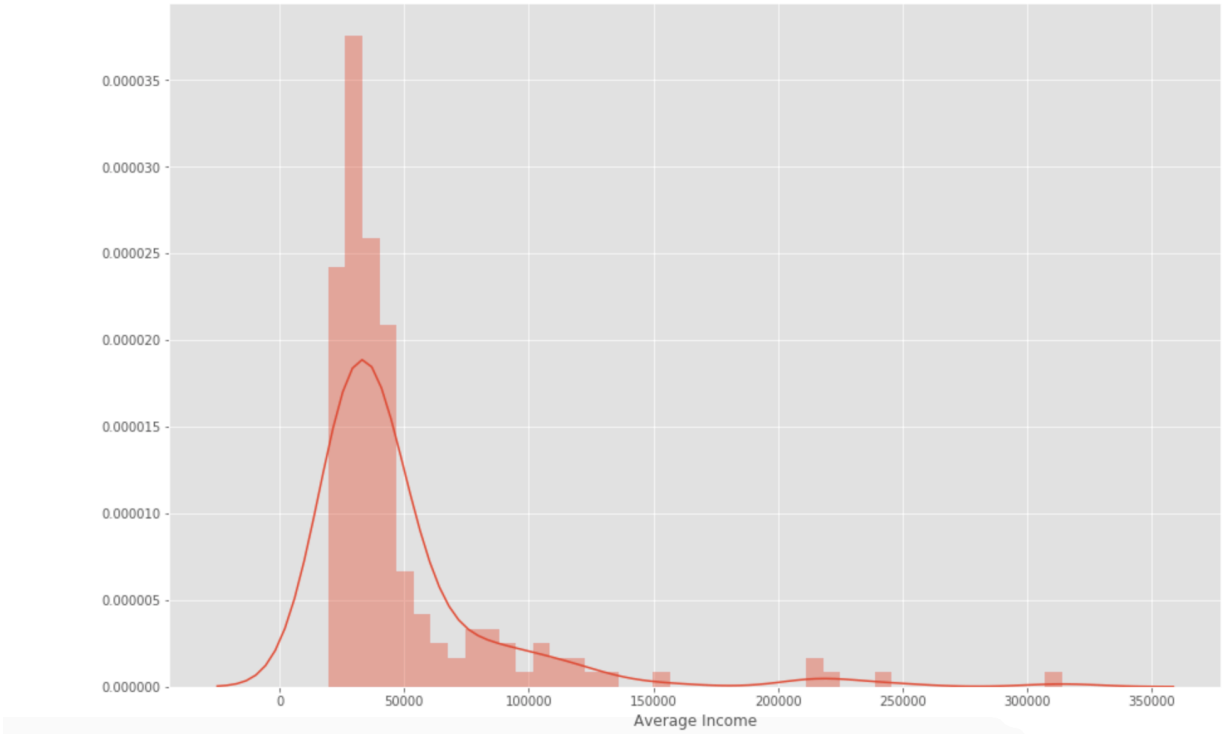
Someone who wants to open an upscale Cantonese restaurant somewhere in Toronto might find this result to suggest what neighborhood to explore.

Neighborhood Statistics

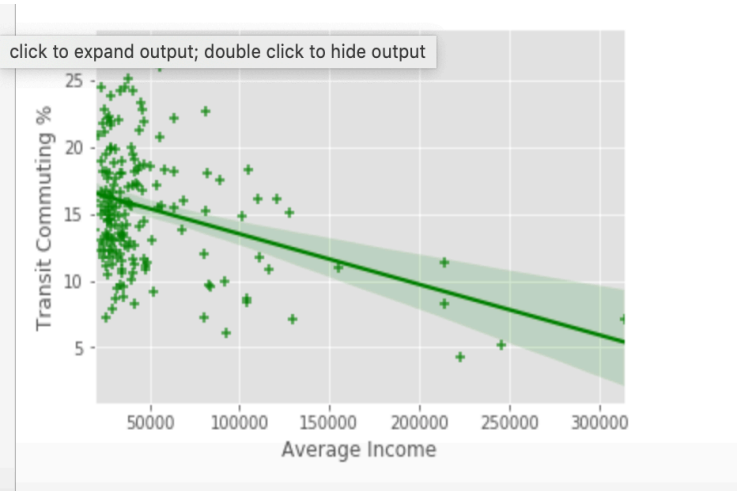
We visualized the data as histograms, as distribution plots, and as plots of a linear regression. For example, here is a histogram of income distribution:

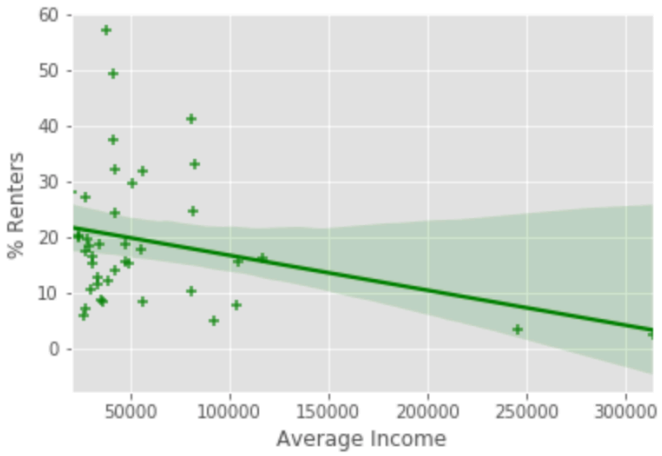


Second, using searborn.distplot, here's a distribution plot of incomes, showing the peak and tail of a Gaussian distribution:



Third, here are a couple of plots of a linear regression model, one that establishes a linear relation between income and % of people that rent their residence; another that pairs income with % of people in the neighborhood that uses public transit.





Finally, I built a linear regression predictive model that was trained on the demographic data. This model can predict income level for anyone in Toronto, based on whether the family rents and takes public transportation. Here is a slice of the code:

In [21]:
```
from sklearn import linear_model
#X = df_code['Transit Commuting\xa0%'].values.reshape(-1,1)
X = df_code[['Transit Commuting\xa0%','% Renters']]
y = df_code['Average Income'].values.reshape(-1,1)
#X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=0)
regr = linear_model.LinearRegression()
regr.fit(X, y)
print('Intercept: \n', regr.intercept_)
print('Coefficients: \n', regr.coef_)

# prediction with sklearn model
New_Commuting = 21.75
New_Renters = 21.3
print ('Predicted Income: \n', regr.predict([[New_Commuting ,New_Renters]]))
```

```
Intercept:
 [95378.89519557]
Coefficients:
 [[-2986.0825566      3.05139686]]
Predicted Income:
 [[30496.59434245]]
```

The model predicts that if a neighborhood has a 21.75 % rate of using public transportation, and 21.3% renting rate, that it will have an average income level of 30,496.

Summary

- Merging Toronto neighborhood demographic data with Foursquare venue data allows one to study how a neighborhood's income and ethnic profile is related to the restaurants in the neighborhood.
- Gaps between the ethnic profile and the list of venues might indicate whether a restaurant of a given ethnic type and price range can be supported by a neighborhood.
- Much more work should be done to isolate the most important factors and build predictive models of Toronto neighborhoods.