Determining the best neighborhood to set up a indian restaurant.

Anish S Ghiya

Data

Data acquisition and cleaning

2.1 Datasources

The main data was obtained from wikipedia ie the postal codes of Canada.But this was not sufficient to get the required attributes that were required. So for the remaining data ie the latitudes and longitudes were taken from the geospatial data that was provided in the previous modules.The remainong columns were from wikipedia ie Demographics of toronto by neighborhood.

The fousquare api was alo used for this to attain the datasets for the venues in the particular neighborhood.

The links to each of theses are :

i) https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M
ii) https://en.wikipedia.org/wiki/Demographics_of_Toronto_neighbourhoods

2.2 Data cleaning :

Data downloaded and the scraped ones from wikipedia had some data elemnts which were not defined.So for that set of data i had to remove the rows where the borough of the neighborhood was not assigned and then use the remaining data.

The next obstacle was with the demographic data which had the neighborhoods in a different pattern so i had to convert the other negihborhood data into csv and then manually change the data and get it to that particular format that was compatible.But with that also i only managed to get the data for 47 neighborhoods in the city of toronto.

The demographic data had the secong language as the combintion of the percentage of the people who comrised of that population. So to change that i had to convert it to a list and the convert then into 2 parts. The first one was to to use the split funtion based on space and then use the list created to extract the language. Similarly there was a problem with the population % , to handle it i used the same approach but the change was just that i used the deliminator for the spit function as the “% “symbol.

Now the next problem was that the percentages got saves as a object instead of a float so i had to change the variable type to numeric

There were no outliers as such in the data as this was census data and so there was not much of a trouble there.

2.3 Feature Selection :

The features set had 18 features which were not enough to understand the data so after dropping the columns of the second highest spoke language and the percentage the data was removed of the redndant data of these.

Now the other features like the map was not available for all the values and so that column had to be dropped.now this left me with just 43 rows ie 43 neighborhoods. Don't worry this is just the basic data. The main data comes in when the foursquare calls are made.

Once the foursqaure calls are made the total number of columns increase by 4 and all of them are important as they include the venue name and the venue latitiude and longitiude and the category of that shop

I also designed another feature of population percentage ie the actual population which spoke the second highest language as there might be areas with less population density to which might not be so significant.

Name	FM	Population	Land area (km2)	Density (people/km2)	% Change in Population since 2001	Average Income	Transit Commuting %	% Renters	Postal code	Borough	Latitude	Longitude	Language	Percentag
Agincourt	S	44577	12.45	3580	4.6	25750	11.1	5.9	M1S	Scarborough	43.794200	-79.262029	Cantonese	19.
Bayview Village	NY	12280	4.14	2966	41.6	46752	14.4	15.6	M2K	North York	43.786947	-79.385975	Cantonese	08.
bagetown	OCoT	11120	1.40	7943	5.3	50398	18.5	29.6	NaN	NaN	43.667967	-79.367675	Unspecified	01.
hurch and Wellesley	OCoT	13397	0.55	24358	8.8	37653	25.1	57.0	M4Y	Downtown Toronto	43.665860	-79.383160	Spanish	01.
Davisville	OCoT	23727	3.14	7556	4.5	55735	26.0	31.7	M4S	Central Toronto	43.704324	-79.388790	Persian	01.

Out[81]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Agincourt	43.7942	-79.262029	Panagio's Breakfast & Lunch	43.792370	-79.260203	Breakfast Spot
1	Agincourt	43.7942	-79.262029	El Pulgarcito	43.792648	-79.259208	Latin American Restaurant
2	Agincourt	43.7942	-79.262029	Twilight	43.791999	-79.258584	Lounge
3	Agincourt	43.7942	-79.262029	Mark's	43.791179	-79.259714	Clothing Store
4	Agincourt	43.7942	-79.262029	Commander Arena	43.794867	-79.267989	Skating Rink