Predicting which country to choose for residence

Data

Data Sources

Chicago,USA data contains Chicago city's Borough, Neighbourhood, Latitude and Longitude. Chicago data can be found here. Toronto,CA data contains Toronto city's Postcode, Borough, Neighbourhood, Latitude and Longitude. Toronto data can be found here and here.

Data Cleaning

For chicago data, scrapped table containing Neighbourhood and Community area using pandas read_html() method. Renamed column names 'Community area' as 'Borough'. It contained only Borough and Neighbourhood names. Grouped 'Neighbourhood' values on unique 'Borough' values and listed them as one value separated by commas. To get venue details, we first need latitude and longitude coordinates of each borough. I used google search method to obtain gps coordinates of all boroughs and saved them as a .csv file. Extracted it as a new dataframe.

Both dataframe have Borough column as common, so merged both dataframe on Borough values to obtain gps coordinates. The final dataframe *chicago_data* looks like this.

	Borough	Neighbourhood	Latitude	Longitude
0	ALBANY PARK	Albany Park,Mayfair,North Mayfair,Ravenswood M...	41.9683	-87.7280
1	ARCHER HEIGHTS	Archer Heights	41.8079	-87.7236
2	ARMOUR SQUARE	Armour Square,Chinatown,Wentworth Gardens	41.8408	-87.6340
3	ASHBURN	Ashburn,Ashburn Estates,Beverly View,Crestline...	41.7479	-87.7072
4	AUBURN GRESHAM	Auburn Gresham,Gresham	41.7434	-87.6562

After cleaning there were 81 records and 4 features(Borough, Neighbourhood, Latitude, Longitude).

For Canada data, scrapped the wikipedia page with postcode, borough and neighbourhood column values using pandas read_html() method. There are few rows that has "Not assigned" value in Borough column. Delete those rows using drop() function. One entry has "Not assigned" value in Neighbourhood column. Replace its value with the value in its Borough column. Apply groupby() method, to group neighbourhoods with same postcode as one row.

Create a new dataframe to read .csv file that contains Toronto city's postcode and gps coordinates. Merging both dataframe on Postcode, to obtain gps coordinates of each postcode. The final dataframe *canada_data* looks like this.

	Postcode	Borough	Neighbourhood	Latitude	Longitude
0	M1B	Scarborough	Rouge,Malvern	43.806686	-79.194353
1	M1C	Scarborough	Highland Creek,Rouge Hill,Port Union	43.784535	-79.160497
2	M1E	Scarborough	Guildwood,Morningside,West Hill	43.763573	-79.188711
3	M1G	Scarborough	Woburn	43.770992	-79.216917
4	M1H	Scarborough	Cedarbrae	43.773136	-79.239476

After cleaning there were 99 records and 5 features(Postcode, Borough, Neighbourhood, Latitude, Longitude).