Finding out the similarity between Toronto and Cherryhill

Jingwei Li

October 27, 2019

1. Introduction
   1. Background

Cherryhill is a community in London, Ontario. There are over ten apartment buildings in this area, making Cherryhill the most popular renting choice for Western University’s students. Most of those students will work in GTA area after they graduate and rent an apartment near their workplace. Some of them want to find out a place that is like to their old apartment.

* 1. Problem

This project aims to calculate the similarity between Cherryhill and Toronto’s different areas and find out the most similar area.

* 1. Interest

Cherryhill is the place where I currently live. However, after I graduate from university, I plan to find a job in Toronto. So, I am interested in this problem. Other students who facing the similar situation may also be interested.

1. Data acquisition and cleaning
   1. Data sourcesThe postal code, borough and neighborhood of Great Toronto’s Area can be found from Wikipedia (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>). The latitude and the longitude of each postal code can be found from ([http://cocl.us/Geospatial\_data](https://cocl.us/Geospatial_data)). The postal code, borough, neighborhood, latitude and the longitude of Cherryhill can be find using google search. The venues around each postal code can be found by requiring from Foursquare.
   2. Data cleaning

Several steps are needed before we can get a useful dataframe.

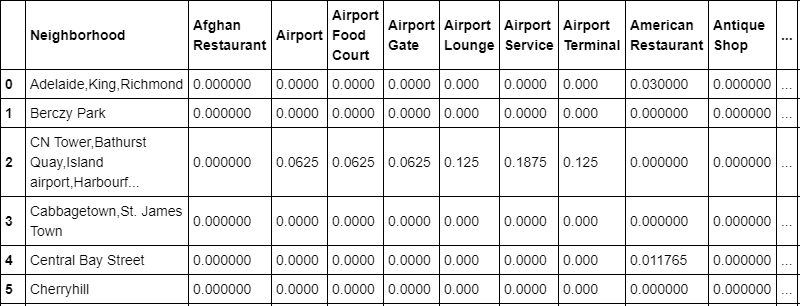
1. Extracting columns ‘PoatalCode’, ‘Borough’ and ‘Neighborhood’ from original dataframe.
2. Ignoring cells with a borough that is ‘Not assigned’.
3. Combining rows with same postal code.
4. Renaming the neighborhood’s cell as borough’s name if the cell was ‘Not assigned’.
5. Merging the dataframe with another dataframe which contains latitude and the longitude of each postal code.
6. Extracting rows only have ‘Downtown Toronto’ and merge with Cherryhill’s dataframe.

Table 1. Geomatic information of downtown Toronto and Cherryhill



1. Requiring venue information around each postal code from Foursquare.
2. Converting venue’s category into dummy variable
3. Grouping rows by neighborhood and taking the mean of the frequency of occurrence of each category

Table 2. Frequency of each category’s occurrence based on postal code



1. Exploratory Data Analysis