**Relocation to another city**

Alexey Sytnikov

April 24, 2020

**1. Introduction**

In this project we will try to solve how to find the best place to move from your hometown. Particularly, this work will be targeted to people willing to move from Toronto (Canada) to New York (the USA).

There are many reasons why people decide to relocate, e.g. job offers, weather conditions, relationships and etc. Moving to another city, people still want to feel at home. That is why it is important to find the same conditions for living and choose a house with similar places next to it that people used to visit before they moved. So, the main beneficiaries of this problem solution are considered to be people who are willing to relocate for some reasons and to find comparable urban environment at the same time.

We will use several data science approaches and tools to find out the best places for relocation to New York for people from Toronto. To reduce the amount of computation power used we will explore one borough in each city. As our example task we can consider the following situation: there is a very successful trader who lives in Downtown of Toronto and who got a job offer from one of the best firms of Wall Street, and now he has to decide which Manhattan neighborhood to choose to move to.

**2. Data**

To implement the following analysis, we need information about both of the cities (New York and Toronto) representing as tables with the data about boroughs and neighborhoods of the cities and their geographical coordinates. Specifically, we are going to focus on one borough in each city (Manhattan and Downtown Toronto) Also, the venues that are located next to every neighborhood in selected boroughs should be included in the tables.

To create these tables, we will use some information sources:

* The dataset about New York City is extracted from the spatial data repository of New York University (<https://geo.nyu.edu/catalog/nyu_2451_34572>). This dataset contains all the necessary information: boroughs, neighborhoods and geographical coordinates of neighborhoods.
* The data about boroughs and neighborhoods of Toronto with the corresponding postal codes is scraped from the Wikipedia page (<https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M>).
* There is a file that contains all the postal codes of Toronto with the corresponding geographical coordinates.
* Information about centers of each city, borough and neighborhood is obtained using Nominatim, open-source geocoding tool.
* Data about all venues for chosen neighborhoods in each city is retrieved using Foursquare API.

As mentioned above, the data retrieved and scraped from different sources should be represented in the form of tables that can be used for analytical purposes. Two types of tables are formed. The first type of table represents each neighborhood being analyzed (Manhattan and Downtown Toronto) in the following shape (Fig. 1).

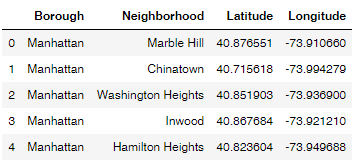


Fig. 1. Data about Manhattan neighborhoods and their coordinates

The other type of tables shows all the venues with their corresponding geographical coordinates and categories that are found for each neighborhood in a borough.

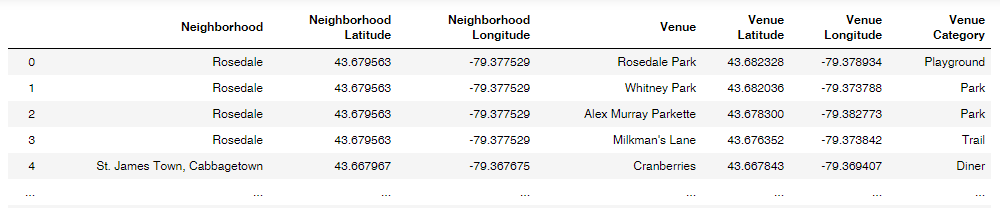


Fig. 2. Data about all venues pertaining to Downtown Toronto neighborhoods

It should be considered that there were some restrictions taken. First, all the venues were found within the radius of 500 meters from the center of a neighborhood. Second, there was a limit to the number of venues selected that was set on 50 by default.

**3. Methodology**

In this relocation project, we are directing our efforts on finding the Manhattan neighborhoods that are comparable to the Downtown Toronto ones as much as possible to offer these options to people interested in relocation.

First, we have collected all the required data: name, location and category of every venue within 500 meters from centers of all considered neighborhoods using Foursquare API requests.

As we are going to compare neighborhoods of two different cities, these neighborhoods should be compatible to each other. Since our analysis is based on the infrastructure similarities it is important to consider only those venue categories that are represented in each borough (Manhattan and Downtown Toronto). So, in second step, we have to analyze which categories we can leave and which ones we have to drop from our initial datasets.

Then, we are going to describe each neighborhood in terms of top 10 venues characterized them. One hot encoding technique will be used to identify the weights of each type of venues in the neighborhoods.

During the last step, we will focus on one of the Downtown Toronto neighborhoods and compare it to all the Manhattan neighborhoods by clustering them using k-means clustering technique. After that, the map with the initial Downtown Toronto neighborhood and the Manhattan neighborhoods that are suited for relocation will be presented by means of Folium, map rendering library.

**4. Analysis**

Choose common categories

According to the exploratory analysis, there are 279 and 181 different venue categories in Manhattan and Downtown Toronto, correspondingly. Intersecting them it was found that these boroughs have 142 venue categories in common. The following analysis is based on the assumption that this number of venue categories is enough to choose the best place to relocate as this set contains a lot of essential venue categories such as grocery stores, different kinds of restaurants, financial organizations, sport venues, chemist's, cultural places and etc.

All datasets with venue information (Fig. 2) are modified by dropping the rows that do not contain venue categories from common venues list.

Get top 10 venues

Understanding more popular venues in each neighborhood is very useful at the final stage of making a decision about which neighborhood to choose to move to because the final range of chosen neighborhoods will much more likely be consisted of more than one option.

Choosing top 10 venues for each borough includes the following steps. First, one hot encoding technique is performed. Next, all the rows are grouped by neighborhood names. And last, the venue categories are sorted within each neighborhood (Fig. 3).



Fig. 3. Manhattan venues sorted (extract)

Cluster neighborhoods

As our data is unlabeled it is justified to use clustering tools. The one that is called k-means clustering has been chosen for this purpose.

To start clustering it is necessary to preprocess some data and create a new concatenated dataset that includes all the data about Manhattan neighborhoods and one of the Downtown Toronto neighborhoods that we move from. That is one of the main reasons why having equal venue categories in each neighborhood is crucial.

As the example task, we assume that that our trader living in Downtown Toronto is the resident of the **Regent Park, Harbourfront** neighborhood. First, we have to extract the rows with the name of this Downtown Toronto neighborhood from datasets about Downtown Toronto neighborhoods and concatenate these rows with the corresponding Manhattan datasets (Fig. 4, Fig. 5, Fig. 6).

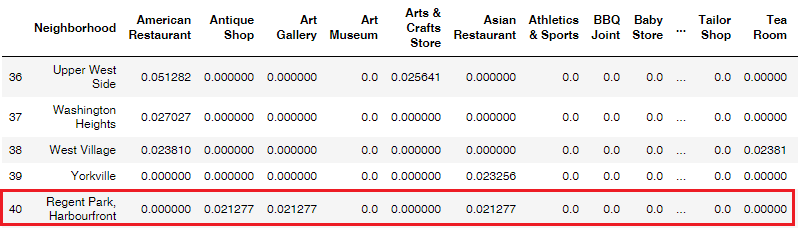


Fig. 4. Concatenated grouped dataset for clustering purposes (extract)



Fig. 5. Concatenated venues sorted dataset (extract)

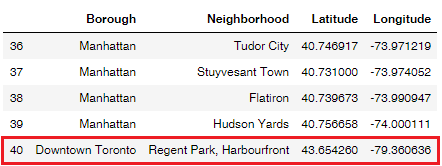


Fig. 6. Concatenated neighborhoods with coordinates dataset (extract)

Using the dataset represented in Fig. 4, all the chosen neighborhoods were divided into 5 clusters by means of k-means clustering technique. The number of clusters was defined by the elbow method (Fig. 7).

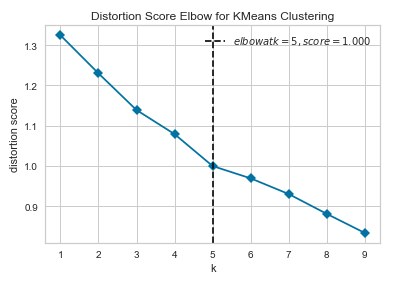


Fig. 7. Elbow method for choosing k number

Aggregate and map the findings

The concatenated datasets were merged and cluster labels were added to the merged dataset (Fig. 8).

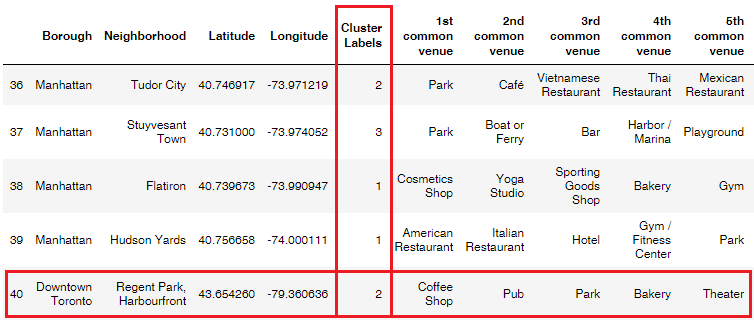


Fig. 8. Merged dataset with cluster labels (extract)

According to the figure above, the initial **Regent Park, Harbourfront** Toronto neighborhood was assigned to the cluster 2. Only those rows that were assigned to the cluster 2 were extracted from the merged dataset and represented in the filtered one (Fig. 9).

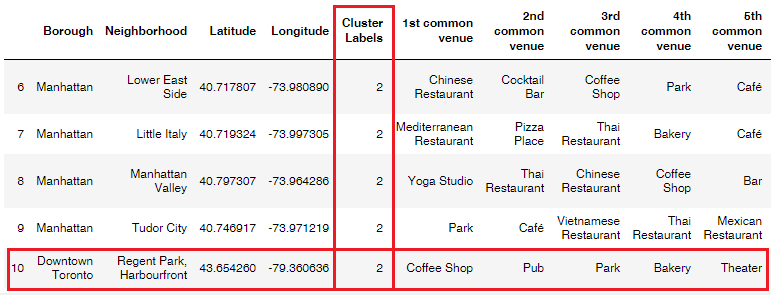


Fig. 9. Filtered by the cluster number merged dataset (extract)

All the neighborhoods from the filtered dataset were then placed on a map of Manhattan (Fig. 10).

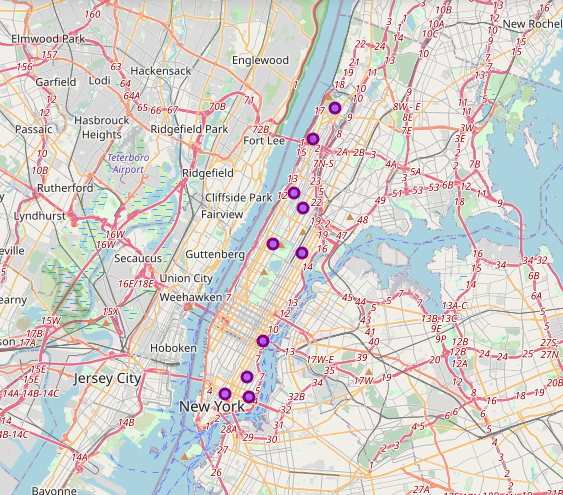


Fig. 10. All the recommended neighborhoods for relocation

On the map presented in Fig. 10, we can find 10 Manhattan neighborhoods that are recommended as similar ones to **Regent Park, Harbourfront** neighborhood from the perspectives of venues located next to and within these neighborhoods.

**5. Results**

The main result of this work is the final set of 10 Manhattan neighborhoods that can be recommended as potential places for relocation to Manhattan from **Regent Park, Harbourfront** Toronto neighborhood. These recommendations are based on the idea of similarity of venues that are typical for these neighborhoods. Initially, there were 40 Manhattan neighborhoods that we started analysis with.

Any of the Downtown Toronto neighborhoods can be chosen as the initial one. According to the initial neighborhood, a unique set of similar Manhattan neighborhoods can be received. Neighborhoods from the generated set are presented on the map and labeled with their names.

**6. Discussion**

This project can be considered as the first step for making a recommendation system for people willing to relocate to another city and finding a place that is similar to their previous residence.

Of course, there are plenty of things that can be done to make this system stronger. Here are some of them:

1. This system shouldn't be restricted by two possible cities. It could be a good idea to start with one of the region and then to expand it to a country level or higher.

2. While making a decision about a place to relocate to, the system should take into account not only the information about the venues of the neighborhoods but also the data about real estate prices, crime rate, level of noise and etc.

3. For this project, it is also a good idea to standardized venue categories and use their variety to make the results more accurate, because in our analysis we had to drop some venue categories reduced their number to 143 although there were initially 279 and 181 different venue categories in Manhattan and Downtown Toronto, correspondingly.

**7. Conclusion**

Purpose of this project was to identify the range of best places to relocate to Manhattan for a trader from Downtown Toronto who's recently got the job on Wall Street.

To describe each neighborhood in these boroughs, the data about venues next to and within these neighborhoods was extracted from Foursquare database. One of the Downtown Toronto neighborhoods then was chosen as the initial one where we suppose our trader lives. To match Manhattan and Downtown Toronto neighborhoods it was necessary to standardize their descriptions. After standardizing descriptions of neighborhoods by dropping unequal venue categories, there was a clustering stage which helped to divide all the neighborhoods including the initial one into five clusters. The cluster containing the initial neighborhood became a target one.

This target cluster and datasets with top 10 venues in neighborhoods can provide the trader from the example task with the information that can help him to make a better final decision on choosing the place to relocate.