The Battle of Neighbourhoods

Toronto VS New York City

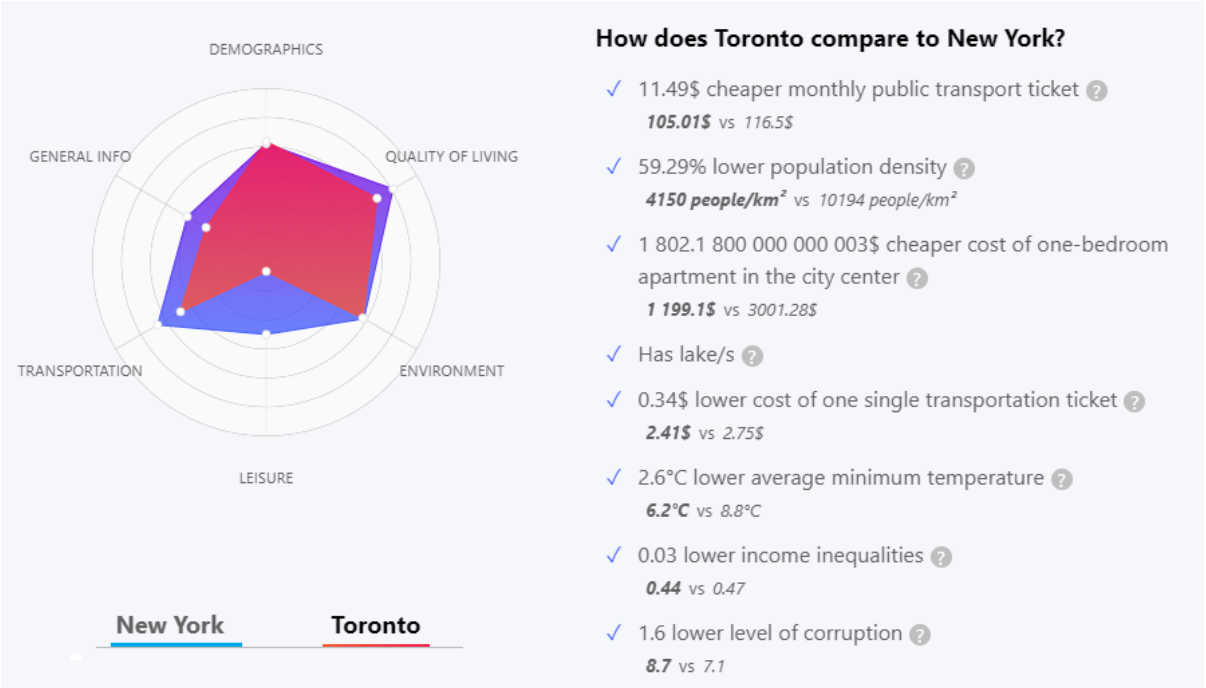
Julian Fan

1. Introduction

New York (also called New York City, short form NYC) is the biggest city in the United States, located in the state of New York. Over 8 million people live in it, and over 22 million people live in the bigger New York metropolitan area. It is in the south end of the state of New York, which is in the northeastern United States. It is the financial capital of the US since it is home to the nation's stock market, Wall Street, and the One World Trade Center. It is also the home of the United Nations Headquarters. The central and oldest part of the city is Manhattan.

Toronto is the capital city of the province of Ontario in Canada. It is also the largest city in both Ontario and Canada. Found on the north-west side of Lake Ontario, the City of Toronto has a population of over 3 million people and even more people live in the regions around it. All together, the Greater Toronto Area is home to over 6 million people making it the biggest metropolitan area in Canada. Toronto is the fourth-largest metropolitan area in North America behind Los Angeles, New York, and Chicago

For years, Toronto and New York City have been pitted against each other in a no-holds-barred comparison to determine which is the better city. New York is clearly the more famous, with its well-known history and touristy flair. But how does it compare to Toronto when it comes to cost of living? A general guess is that it's the pricer of the two, but it's not until you take a look at the numbers that you realize just how much pricier it really is. The comparison between two cities are shown in the figure below.



What we try to do in this report is to explore the neighborhoods of both cities, and show the similarity and difference of its different areas by the clustering method on the combined data of both cities. It could provide good guidance for selecting right neighborhood which fits well for their living style. It is particularly useful for people relocating from New York to Toronto or from Toronto to New York.

2. Data

For this project the Foursquare API will be used. A list of neighborhoods in New York and Toronto is downloaded and their respective location in longitude and latitude coordinates is obtained. The sources are the following:

- New York neighborhoods: https://ibm.box.com/shared/static/fbpwbovar7lf8p5sgddm06cgipa2rxpe.json
- Toronto neighborhoods: https://en.wikipedia.org/wiki/List_of_postal_codes_of_Canada:_M

The data downloaded are the neighborhoods located in New York and Toronto. Moreover, their specific coordinates are merged. Only Manhattan neighborhoods and boroughs that contain the string "Toronto" are taken into account. A Foursquare API GET request is sent in order to acquire the surrounding venues that are within a radius of 500m. The data is formatted using one hot encoding with the categories of each venue. Then, the venues are grouped by neighborhoods computing the mean of each feature.

The neighborhood data of two cities acquired from the Foursquare API will be combined and then used for clustering analysis. The similarities will be determined based on the frequency of the categories found in the neighborhoods. These similarities found are a strong indicator for a user and can help him to decide whether to move in a particular neighborhood.

3. Methodology

3.1. Feature Extraction

For feature extraction One Hot Encoding is used in terms of categories. Therefore, each feature is a category that belongs to a venue. Each feature becomes binary, this means that 1 means this category is found in the venue and 0 means the opposite. Then, all the venues are grouped by the neighborhoods, computing at the same time the mean. This will give us a venue for each row and each column will contain the frequency of occurrence of that particular category.

3.2. Unsupervised Learning

We will do a clustering analysis on the combined neighborhood data of both cities. The purpose of doing the clustering to found similarities among the neighborhoods of two cities. In this case K-Means is used due to its simplicity and efficiency.

K-Means is a clustering algorithm. This algorithm search clusters within the data and the main objective function is to minimize the data dispersion for each cluster. Thus, each group found represents a set of data with a pattern inside the multidimensional features.

Contrary to supervised learning where we have the ground truth to evaluate the model's performance, clustering analysis doesn't have a solid evaluation metric that we can use to evaluate the outcome of different clustering algorithms. Moreover, since k-Means requires k as an input and doesn't learn it from data, there is no right answer in terms of the number of clusters that we should have in any problem. Sometimes domain knowledge and intuition may help but usually that is not the case. In the cluster-predict methodology, we can evaluate how well the models are performing based on different K clusters since clusters are used in the downstream modeling.

Elbow method gives us an idea on what a good k number of clusters would be based on the sum of squared distance (SSE) between data points and their assigned clusters' centroids. We pick k at the spot where SSE starts to flatten out and forming an elbow. Then, further analysis of each cluster is done.

Hopefully the analysis results can provide a good guidance for people who want to move from Manhattan to Toronto and vice versa to get an idea what is the best suitable place for them.