Popularity Score Measurement for Cities

Applied Data Science Capstone Project

By

Hari Kanagala

1. Introduction

For most of us, it is a bigger factor to find place to live that is accessible and well connected within the city. It becomes especially important to compare the accessibility between the current city and new city before accepting a new job offer in a different city. In this business problem, we attempt to help customers find a way data analytics can help with providing comparative analysis between two different cities. For instance, the current analysis was tried out between Toronto and New York cities by generating a popularity score for each of the cities based on the popularity of the neighborhoods in each of the cities.

The popularity score can be defined by the proximity and number of venues in each of the cities. More venues in a neighborhood increases the popularity score, and in the same way, more restaurants in a neighborhood increases the score further. For instance, if the overall popularity score of New York city is higher than Toronto, then one would prefer to choose a new job offer in New York City than in Toronto. This popularity score can help as a deciding factor when combined with salary offers in each city to finalize job offers. In order to simplify the objective, the location of top 50 popular places in each of the cities are identified along with their location in each city. By knowing how closer these popular places are to the city and how accessible are these from the center of the city, this analysis would define the popularity score of New York City and Toronto.

2. Data Extraction and Cleaning

As a first step, the Geocode libraries are used to obtain the location specific latitude and longitude data for two cities, and we obtained New York and Toronto City location data for this analysis. FourSquare APIs are used to request data of the top 50 popular venues in all the neighborhoods around the center of these two cities. A radius of 10,000 meters is used to identify the top 50 popular venues to explorer around the center of each of the city. Specific client ID and codes are generated to make the API calls with FourSquare. The data obtained in the JSON format from FourSquare, which is sorted by the distance from the center of the city, is converted into panda dataframes and then cleaned to extract relevant data into other dataframes. The most relevant data that is used in this analysis for each of the 50 popular venues is the latitude, longitude, category, distance from the cities, and unique ID of the venue.

3. Exploratory Data Analysis

For each category type of the venue, a score is assigned on a scale of 1 to 10 depending on the popularity of those venues if the customer would choose that city. To simply the analysis the venues are categorized into 4 different buckets – Scenic, Parks, Entertainment, and Necessities. For instance, venues classified as 'necessities' would get score 10, while venues classified as 'scenic' would get 7. For each of the venue the score is then divided by the distance of the venue from the center of the city, because the popularity diminishes when the venues are farther to reach from the center of the city. These scores for individual venues get added to calculate the total popularity score of each city. Higher the score implies more popular venues are existing nearby in the city. Based on the comparison of the popularity score of Toronto and New York cities, a customer can decide which city to choose as one of the factors to accept a new job offer. Another approach tried is to cluster these top 50 venues into different clusters and calculate the mean distance between the centers of these clusters. Depending on how distant apart the clusters are the mean distance would be able to give us an approximate idea of how close the popular places on one city are compared to other. For better visualization folium maps are used to understand the spread of these venues across the cities. Folium maps are also used to identify the clusters into which these venues are separated based on the proximity to each other.

4. Modeling

Using the dataframes that contain cleaned information of 50 popular venues around New York City and Toronto, the unique categories of these venues are identified for each city. These categories are binned into four different buckets - Scenic, Parks, Entertainment, and Necessities. Depending on the need for each of these venues for a customer, a score has been assigned to each of these buckets – Scenic got a score of 7, Parks got 8, Entertainment got 9, and Necessities bucket got score of 10. The score is normalized by the distance from the center of city in kilometers. This normalized score is added for all the 50 popular venues to get the final popularity score. To visually understand the results, these venues are clustered using KMeans algorithm is used based on the proximity from the center of the city. After multiple iterations, it is observed that 4 clusters can represent the venues in a better way for both the cities. Some of PIP packages installed to obtain and identify the geographic locations are Geopy and Folium.

	name	categories	lat	lng	distance	score	city
0	Prospect Park	Park	40.661938	-73.969617	6434	8.0	Brooklyn
1	Central Park	Park	40.783076	-73.965497	8544	8.0	New York
2	Brooklyn Bridge	Bridge	40.705967	-73.996707	1087	7.0	New York
3	Bryant Park	Park	40.753621	-73.983265	4940	8.0	New York
4	Hudson River Park	Park	40.733747	-74.010425	2369	8.0	New York

Table 1 First 5 rows of Popularity Venue dataframe for New York City

	name	categories	lat	lng	distance	score	city
0	High Park	Park	43.646479	-79.463425	6449	8.0	Toronto
1	Woodbine Beach	Beach	43.663112	-79.306374	6337	7.0	Toronto
2	Humber Bay Park	Park	43.622396	-79.478389	8359	8.0	Toronto
3	Taylor Creek Park	Park	43.696599	-79.306693	7855	8.0	Toronto
4	The Distillery Historic District	Historic Site	43.650244	-79.359323	2014	7.0	Toronto

Table 2 First 5 rows of Popularity Venue dataframe for Toronto

5. Evaluation

Among the four different buckets of categories, both the cities have highest number of parks than any other category of venues. However, New York City has more number of venues in the 'necessities' bucket than Toronto.

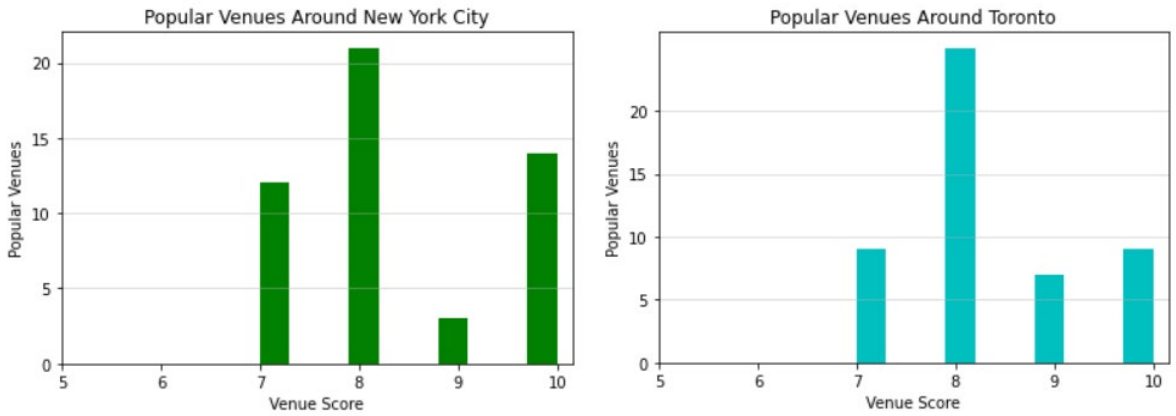


Figure 1 Geographical Location and clustering of 50 venues in New York City

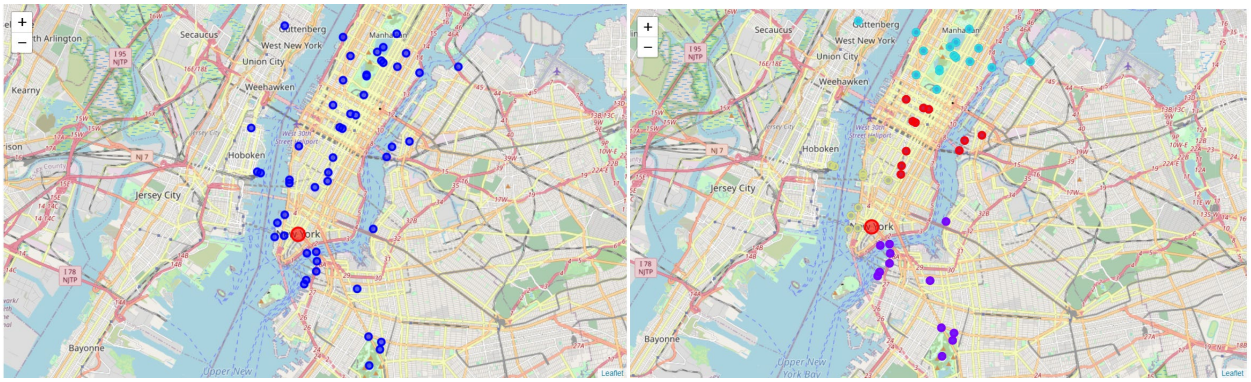
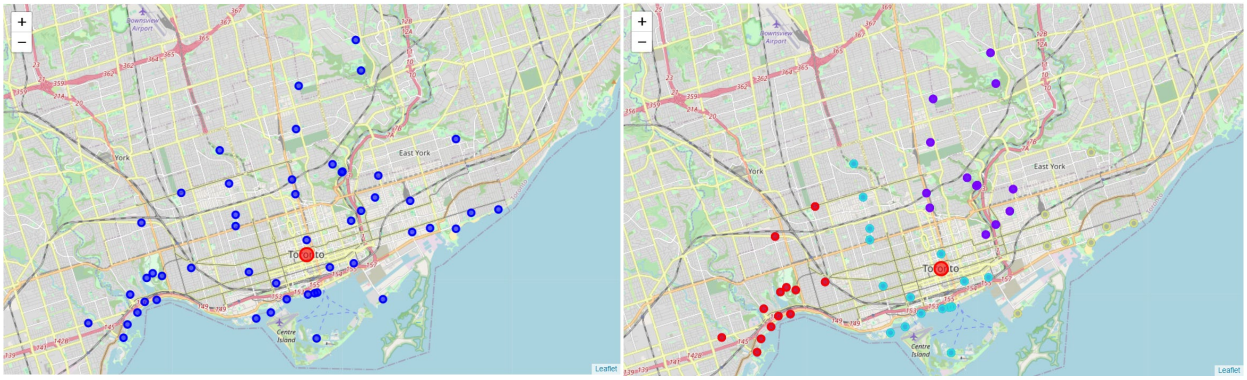


Figure 2 Geographical Location and clustering of 50 venues in Toronto



From Figures 1& 2, the venues are closer to the center of the city in New York City than in Toronto. Also, the New York City has more venue that have higher score than the Toronto. From Table-3, the popularity score calculated for both the cities by adding the normalized score for all the 50 venues.

City	Popularity Score
New York City	135.9
Toronto	125.7

6. Conclusion

The popularity score for New York City is about 10% higher than that of Toronto. This indicates that the customer would have 10% better chances to find specific venues that are popular and cater to the needs of the customer. This score combined with other factors such as cost of living and salary offer in the new

city, would allow the customer to take a better judgement whether the move from Toronto to New York City or vice versa would be an overall benefit.

7. Future Scope

In the current data analysis, the scores are calculated for two big cities, however, the same methodology can be used to calculate popularity score for any other cities. This methodology can be made robust by bringing in other factors, along with the popularity score, about the venues around the city that would help in the better decision making for the customers. The correlation for the popularity score can be refined by identifying the correct weightages for the impact of popularity and distance of the venues from the center of the city.