BeGone International

Recommendation Service Evaluation

1. Introduction

Background

The local travel company *BeGone International* is exploring rolling out destination recommendation services to its travelers.

Problem

Their research personnel are evaluating if travel recommendations can be derived from venue ratings, to influence travelers based on their venue interests. For New York City travelers, the initiative is to identify similarities of New York City and Toronto, and show that Toronto as potential destination location with similar venue categories

Interest

The company is using data science methodology and analytics to validate their hypothesis for this initiative

2. Data

For this initiative, the research team is using publically accessible information from the FourSquare platform. The primary data sources were derived from Python libraries and the data set from the FourSquare platform. The coordinate location for the selected cities was derived using Python geopy libraries. For this initiative, they are using location data and rating information for venues from these cities. Venues within the vicinity of these locations were acquired from FourSquare platform using the Places API service. The venue type is used to classify and categorize venues. The top 100 venues within 500 meters radius of the coordinates, were selected from each location

During this activity, it was required to cleanse or manage occurrence of anomalies in data. Assumptions on the data were made and rules were applied to address discrepancies, i.e. venues must valid postal code or zip code values. Toronto postal codes have a unique format; in Canada they use a 7 character string where first 3 characters represent the geographic location, the last 3 characters are used for postal delivery and separated by single space. The last 4 characters were extracted from each Toronto postal code, to remove complexity and lack of value from the postal delivery code.

As supplemental information, visual representations were created. Graphs were generated using Python graphical libraries to capture the count of venues by venue category. Location data for neighborhoods across NY and Toronto was gathered on the venues within their approximate vicinity

3. Methodology

Using statistical analysis, I used the count of venue categories by venue category for each 2 cities New York and Toronto.

Using those venues with valid postal codes recorded, a bar chart was used to depict the counts of venue categories. It was used to demonstrate the counts of venues by category, as

a basis and another data point for comparison. For demonstration purposes, the top ten venue categories were depicted

The venues were initially group by postal code. The analysis proceeded to group or cluster those venues by similarity in category type. The clustering method was applied to capture the venues and the ratings associated to those venues, to identify the top venues

Venue clusters by category were compared between Toronto and New York City

Results

Using location, all neighborhoods were identified across the New York and Toronto zip codes. Toronto zip codes were identified using zip codes

Using the FourSquare API acquires the venue information as of most current data set captured by or within the FourSquare platform. Venue and venue ratings could be subject to change.

The general location for Toronto and New York City were mapped and visualized. Once clustered, the venues were depicted based on coordinate [latitude, longitude] location. Those same venues were clustered into groups based on their venue type and mapped using those cluster associations.

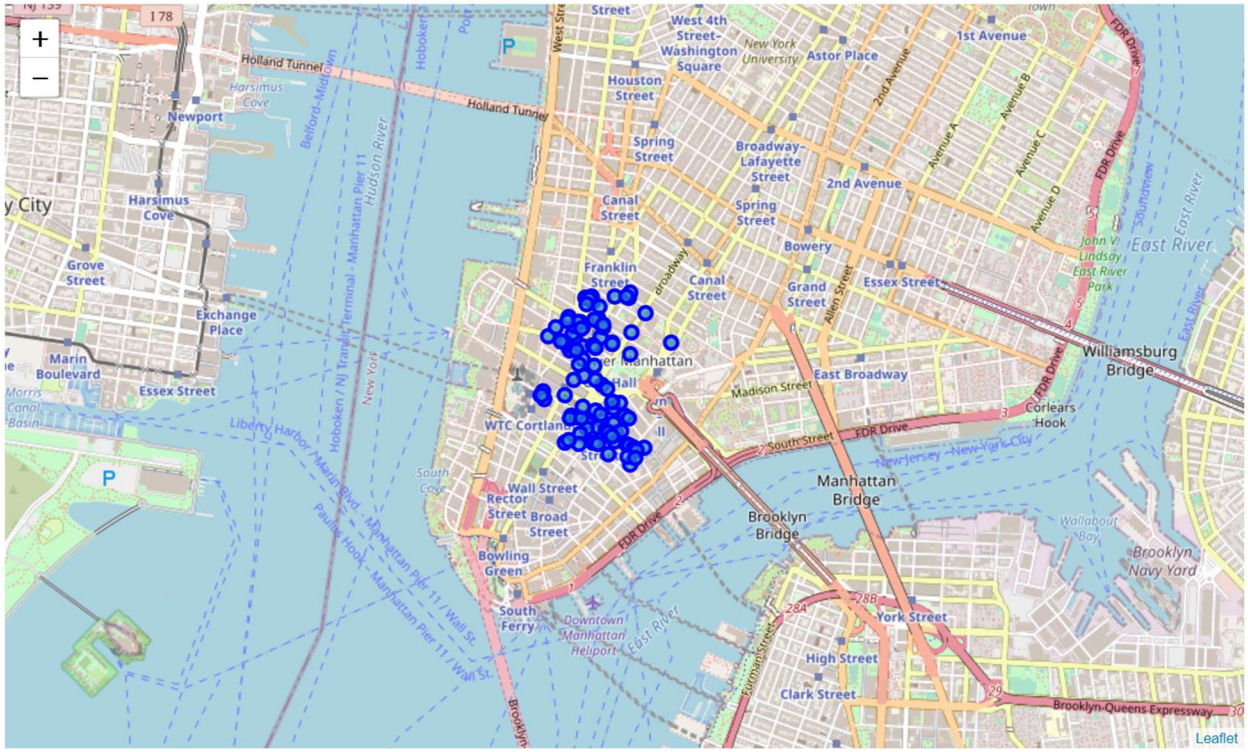
From the data sets retrieved from FourSquare, the criteria was for the top 100 venues, within the 500 meter radius. This was pulled for the specific locations. For New York City, 100 venues were retrieved and had valid postal codes. For Toronto, 100 venues were retrieved but 78 venues had valid postal values. Approximately one fifth of venues were removed due to null postal codes in the data

64 venue categories were identified in New York City compared to 56 venue categories in Toronto.

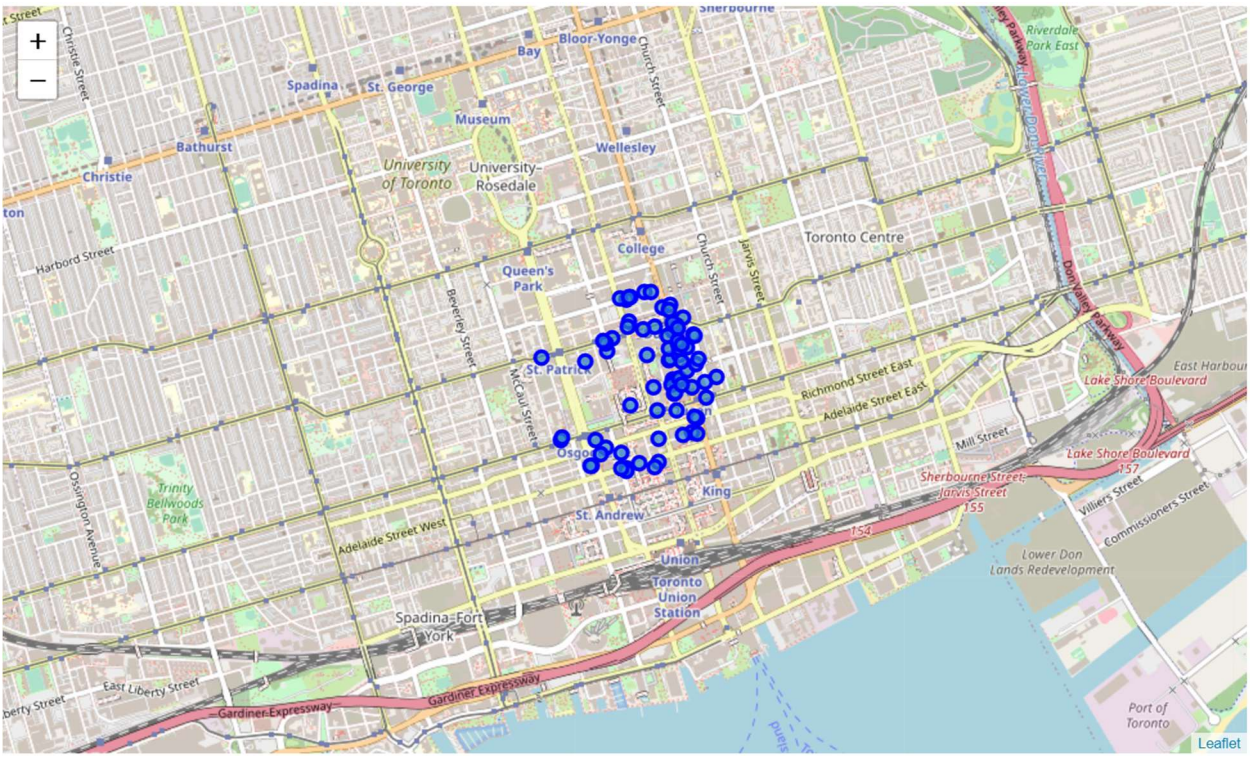
Each zip code was assessed to determine its most common venue type. 4 zip codes were assessed in New York City and 7 zip codes were assessed in Toronto.

Clustering of zip codes was generated based on the 1st most common venue type (highest frequency). Those clusters are represented visually on maps below.

Picture 1: Map of venues in New York City, New York



Picture 2: Map of venues in Toronto, Canada



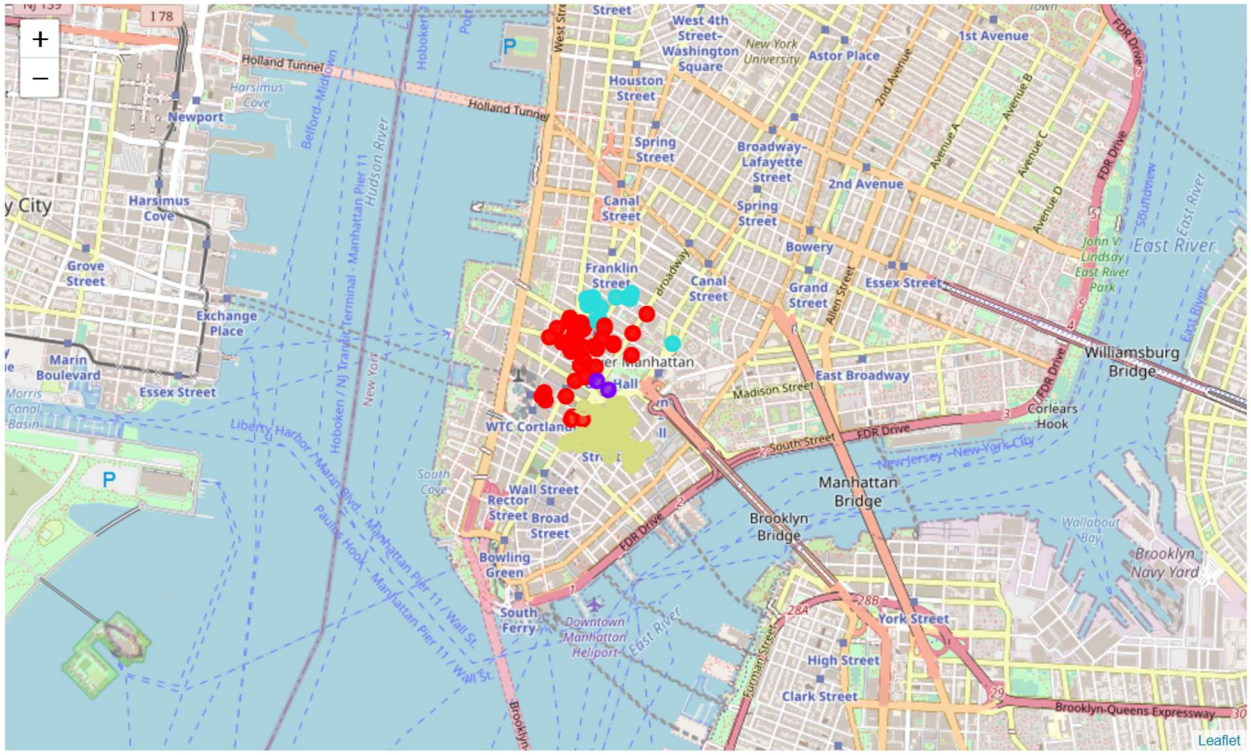
Observations

Using K-means clustering, the venues were grouped into clusters by venue types and mapped using Python GeoPy libraries. The use of N had be a minimum of the number of postal codes

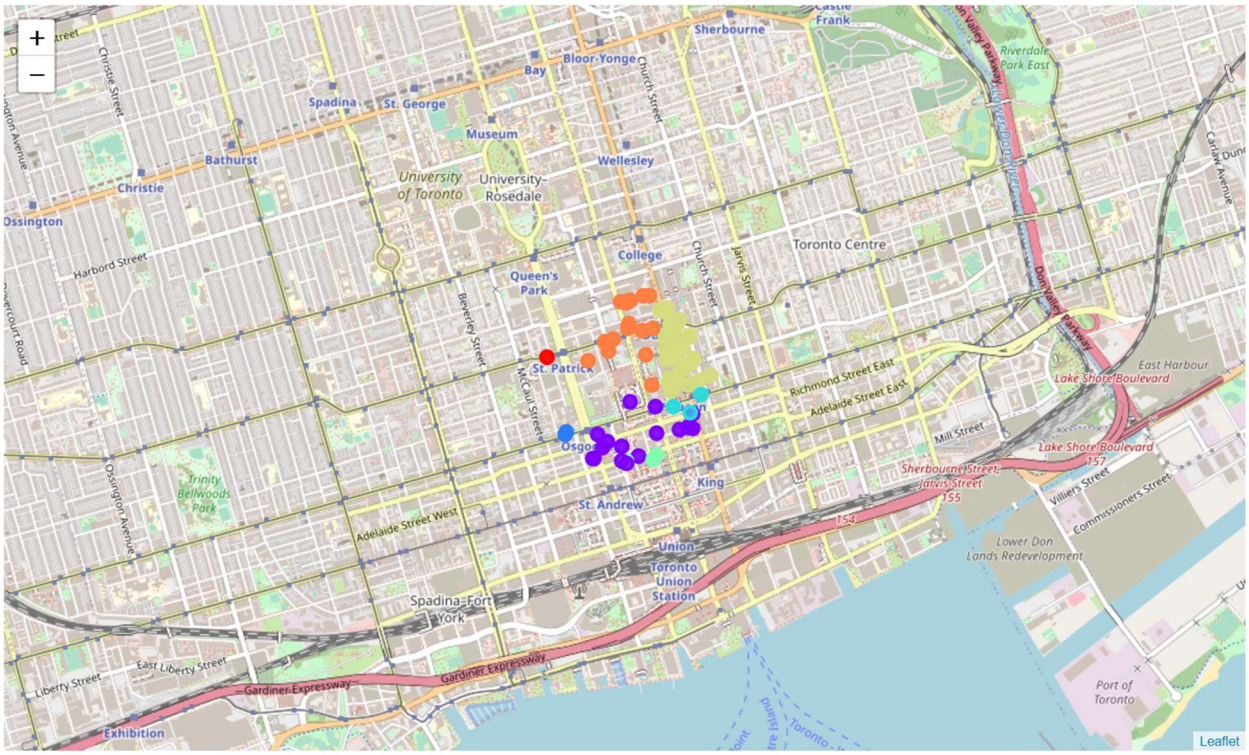
Initially, all of the valid postal codes were mapped. The clustering algorithm was applied based on the venue category .Each group was labeled and mapped to its most common venue type for New York City and Toronto.

The use of the 1st Most Common Venue category is an input into evaluation. It should not be interpreted as only point of comparison or for labeling the clusters. It was an easy or one of the most obvious attributes used to classify the cluster. The remaining most common venue categories [2nd – 10th] may also overlap in comparison between New York City and Toronto and provide input into the recommendation for NYC travelers to Toronto. In addition to the 1st Most Common Venue, that was selected.

Picture 3: Map of venue clusters in New York City, New York



Picture 4: Map of venue clusters in Toronto, Canada



For each cluster, the sum of those associated venue categories does add up the sum of total venues established in earlier in the analysis. Particularly in Data exploration section. This ensures that we reconcile back to the total number of valid venues in query from the FourSquare API

Lastly, the analysis for comparing venue results was demonstrated by using a data frame. Each data frame X (source) and Y (destination) was to show the Top 10 most common venue types by location for each cluster [label]

From the top 10 most Common Venue categories for each cluster, here are recommendations that can estimated from the analysis extracted for this activity. Toronto is similar to New York City in respect to high ratio of these venue types between these locations:

- Coffee Shops or Cafes
- Gym/Fitness Centers
- Hotels

Conclusion

The most common venue in New York City [across the venue clusters] was Coffee venues while Toronto's most common venue category are Clothing Stores.

Based on the bar graph and clusters as supplemental data sources, coffee venues and restaurants are in the top 10 of most common venues types between the 2 locations.

Further analysis on the overlap or similarities across all the top 10 venue categories between New York City and Toronto should be expanded into the scope of the recommendation solution. They would inform the recommendation solution