Prediction of potential locations for a sports shop

Pratheeksh Eapen

16th June 2020

1. Introduction

1.1 Business Problem

A person wants to start a sports shop in Manhattan, New York. For this problem we have decided to use the potential customer bases in and around Manhattan to generate locations. Potential customer bases are:

- Colleges and Universities
- Elementary schools
- Middle schools
- High Schools
- Private Schools
- Rec centers
- Stadiums

2. Data acquisition and cleaning

2.1 Data sources

- Data of **Manhattan, New York** was taken from newyork_data json file.
- Data of **potential customer bases** including coordinates were obtained using Foursquare API.
- The ID's of the various customer bases to be searched for using the Foursquare API was obtained from "https://developer.foursquare.com/docs/resources/categories"
- **Addresses of the potential locations** were obtained using LocationIQ API reverse geocoding

2.2 Data cleaning

The data of the potential customer bases once obtained using the Foursquare API was stored in dataframes. These dataframes were then individually inspected and then cleaned as each category had various categories in itself. So to decide which ones to use for the analysis, each

dataframe had to be inspected. Figure 1 and Figure 2 below shows how many categories were there in the category "College" and "Rec Center" before it was cleaned:

```
manhattancollege['Venue Category'].value_counts()
```

College Academic Building	191
General College & University	163
College Administrative Building	126
Student Center	107
College Classroom	80
University	79
Trade School	79
College Library	71
College Lab	49
College Arts Building	47
Medical School	41
College Residence Hall	40
College Gym	31
College Auditorium	29
College Cafeteria	22
Law School	21
College Theater	21
College Science Building	18
College Quad	17
Office	15
School	12
College & University	12
College Technology Building	12
College Bookstore	10
Community College	10
College Engineering Building	9
College Rec Center	8
College Communications Building	7
Doctor's Office	5
Fraternity House	5
Hotel	3
Elementary School	3
Miscellaneous Shop	3
High School	3
Library	3
Art Gallery	3

Figure 1

```
manhattanrec['Venue Category'].value_counts()
```

Gym / Fitness Center	506
Gym	476
Park	277
Yoga Studio	176
Plaza	153
Playground	114
Athletics & Sports	95
Martial Arts Dojo	82
Pilates Studio	65
Garden	57
Basketball Court	49
Scenic Lookout	35
Boxing Gym	34
Dog Run	34
Cycle Studio	31
Soccer Field	29
Tennis Court	26
Pool	25
Roof Deck	25
Weight Loss Center	20
Harbor / Marina	20
Trail	18
Gym Pool	17
Pedestrian Plaza	15
Baseball Field	14
Sports Club	13
Skate Park	10
Gymnastics Gym	8
Fountain	7
Skating Rink	7
Track	7
Golf Course	7
Recreation Center	7

Figure 2

After the dataframes of each category was cleaned they were all compiled into one dataframe. A total of 1316 venues were obtained after ignoring several categories as we wanted the cluster centers to be closer to the customer bases which could generate higher revenue.

3. Analysis

After the data was cleaned, the coordinates of these customer bases were taken and plotted on a map to get a better idea as shown in Figure 3.

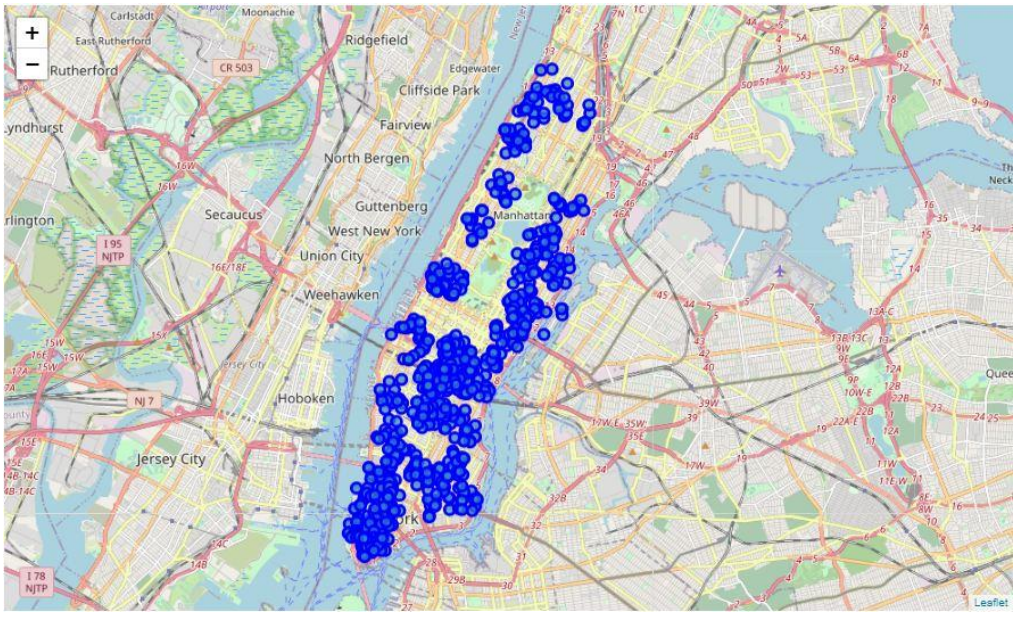

Figure 3

Now that the points have been plotted, these coordinates were used in creating clusters using k-means clustering. The number of locations required can be used as the number of centers i.e. the number of clusters formed. Here we have used 20, so 20 clusters were obtained as shown in Figure 4.

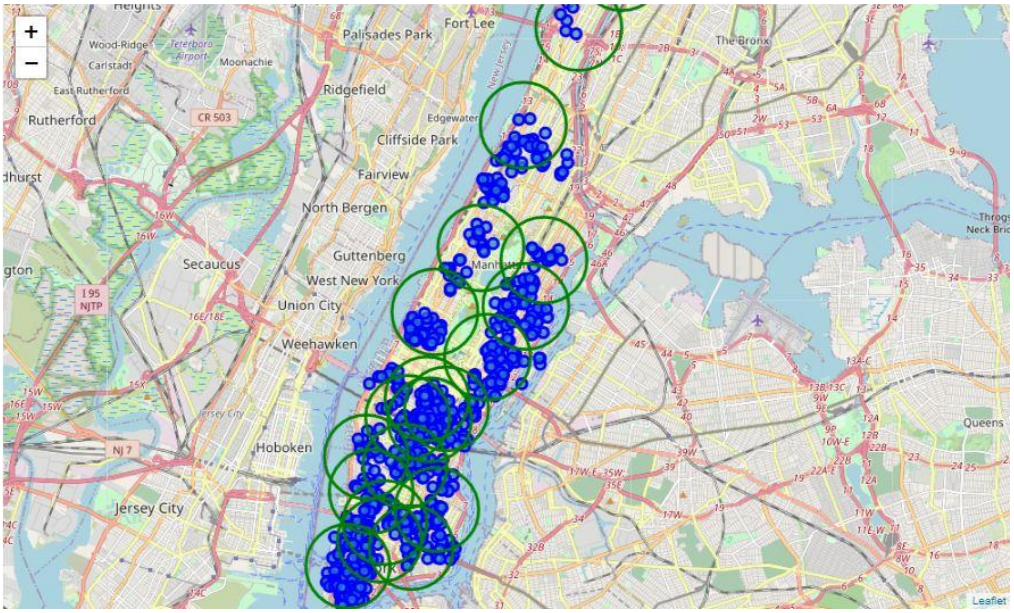

Figure 4

After obtaining these clusters, the centers of these clusters were used as the potential locations to set up the shop. These centers were then stored in a list. To obtain the addresses of these centers, the coordinates will have to be reverse geocoded. In order to do so the LocationIQ API was used. The addresses obtained after reverse geocoding the coordinates is shown in Figure 5.

```
============================================================
Addresses of locations recommended for further analysis
============================================================

386, 5th Avenue, Midtown West, New York, New York County, New York, 10018
60, Greene Street, SoHo, New York, New York County, New York, 10012
1896, 3rd Avenue, East Harlem, New York, New York County, New York, 10029
517, W 207 St, Inwood, New York, New York County, New York, 10034
600, West 182nd Street, Washington Heights, New York, New York County, New York, 10033
150, East 59th Street, Midtown East, New York, New York County, New York, 10022
606, West 143rd Street, Hamilton Heights, New York, New York County, New York, 10031
410, East 89th Street, Upper East Side, New York, New York County, New York, 10128
30, Horatio St, West Village, New York, New York County, New York, 10014
Duane Reade, Upper West Side, New York, New York County, New York, 10025
Silverstein Family Park, Tribeca, New York, New York County, New York, 10007
252, Broome St, Lower East Side, New York, New York County, New York, 10002
159, E 71 St, Upper East Side, New York, New York County, New York, 10021
One Vanderbilt, Midtown East, New York, New York County, New York, 10017
311, E 10 St, East Village, New York, New York County, New York, 10009
357, W 14 St, Chelsea, New York, New York County, New York, 10014
Duane Reade, Upper West Side, New York, New York County, New York, 10025
Citi Bike - E 43 St & Vanderbilt Ave, Midtown East, New York, New York County, New York, 10017
2, West 37th Street, Midtown West, New York, New York County, New York, 10018
```

Figure 5

4. Result

- The analysis done in this project shows that there is enough and more customer bases in and around Manhattan, New York for a sports shop.

- In this analysis we first got all the potential customers and then we clustered them and used the centers of these clusters as potential locations for the shop/store. Addresses of these locations were obtained using reverse geocoding.

- As a result of this analysis we have generated 20 potential locations. These of course are not the exact optimal locations for the stores. There may be other reasons for which there are no shops in the given location.

- This project only takes into account potential customer bases and doesn't take into account the close by shops. In this case it was done because a shop like SKECHERS being beside the shop can help boost sales. So a closer manual inspection is required. These conditions can be changed according to the objectives and requirements of the stakeholder.

5. Conclusion

- Purpose of this project was to generate potential locations to set up a sports shop depending upon the number of potential customer bases in and around the neighborhood. The potential customer bases and its coordinates were obtained using Foursquare API. The data was inspected and cleaned and then clustered using k-means clustering. The centers of these clusters were taken as potential locations and their addresses were obtained using LocationIQ reverse geocoding API.

- The final decision is up to the stakeholders and may require further manual inspection or on foot inspection as they may have other factors in mind too such as attractiveness of the location, proximity to main roads etc.