Predicting Business Opportunities in New York

# Shubham Rai

April 4, 2020

1. **Introduction**

Today entrepreneurship is becoming a culture and everyone now try to stand with there own business, people do a lot of things one may open a shop and one may open a chain. In this exercise we will explore how data science could help us explore the opportunities.

* 1. **Problem Statement**

To determine potential business opportunity in neighbourhood of New York by identifying lack of selected type of shops based on the number of check ins.

This will help people who are not clear with what to open and where to open as a shop.

1. **Data Acquisition and Cleaning**

This dataset contains check-ins in NYC and Tokyo collected for about 10 months (from 12 April 2012 to 16 February 2013). It contains 227,428 check-ins in New York city and 573,703 check-ins in Tokyo. Each check-in is associated with its time stamp, its GPS coordinates and its semantic meaning (represented by fine-grained venue-categories). This dataset is originally used for studying the spatial-temporal regularity of user activity in LBSNs.

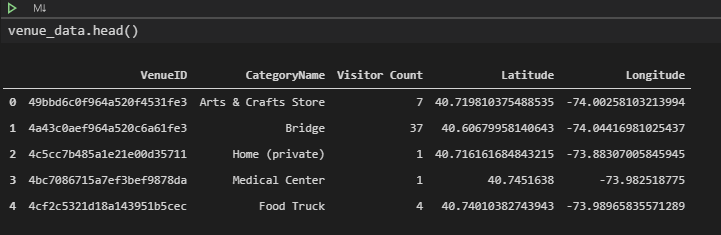
It contains two files in tsv format. Each file contains 8 columns, which are:

1. User ID (anonymized)
2. Venue ID (Foursquare)
3. Venue category ID (Foursquare)
4. Venue category name (Fousquare)
5. Latitude
6. Longitude
7. Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
8. UTC time

The data is taken from the dataset collected by **Dingqi Yang** for his and his teams’ paper:

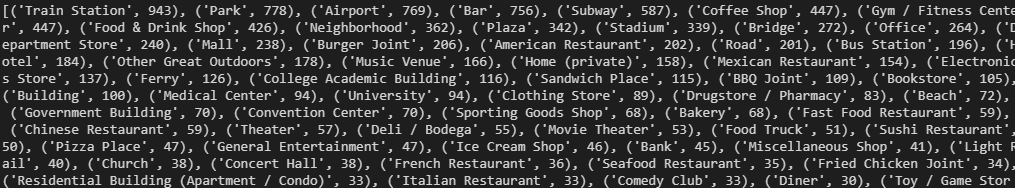
Dingqi Yang, Daqing Zhang, Vincent W. Zheng, Zhiyong Yu. Modeling User Activity Preference by Leveraging User Spatial Temporal Characteristics in LBSNs. IEEE Trans. on Systems, Man, and Cybernetics: Systems, (TSMC), 45(1), 129-142, 2015.

After data extraction we will convert it to a data frame and will use this data frame for our further investigation.

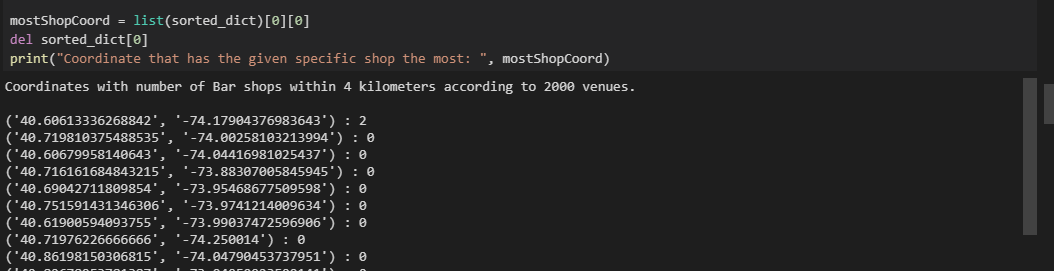


1. **Methodology**

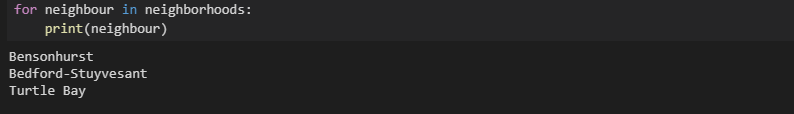
After converting data set into a pandas data-frame we will create a dictionary with most popular places based on the numbers of visitors.



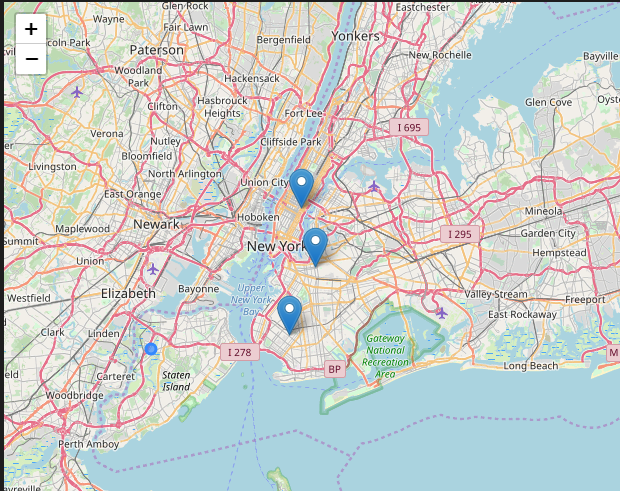
Once we have the most visited place’s we will check for the coordinates for top 2000 places within 4 KM of range.



The we will find the 3 neighbourhood’s that are closest to the coordinate which has the greatest number of the specific shop type but lacking that within our range.



Finally, with the help of folium we will plot the centre and the found newighborhoods to open the shop.



1. **Conclusion**

In our sample of 2000 venues, we did multiple coordinates that has no Bar (the most visited shop type according to sample) within 4 km range. And we did manage to get the neighbourhood’s names from foursquare database and pin down the 3 closest neighbourhood’s, ‘Bensonhurst’, ‘Bedford-Stuyvesant’, and ‘Turtle Bay’, into the map. Of course, it should not be forgotten that the data used is from 2013 hence further analysis will be required to actually use it.

Anyways, the results according to the data in hand can be checked from the map and analysis above can be of use for future entrepreneurs.