THE BATTLE OF NEIGHBORHOOD

PREDICTING BUSINESS OPPORTUNITIES IN NEW YORK

SHUBHAM RAI

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.

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.

INTRODUCTION

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.
- 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.
- Description of the data can be found on next slide

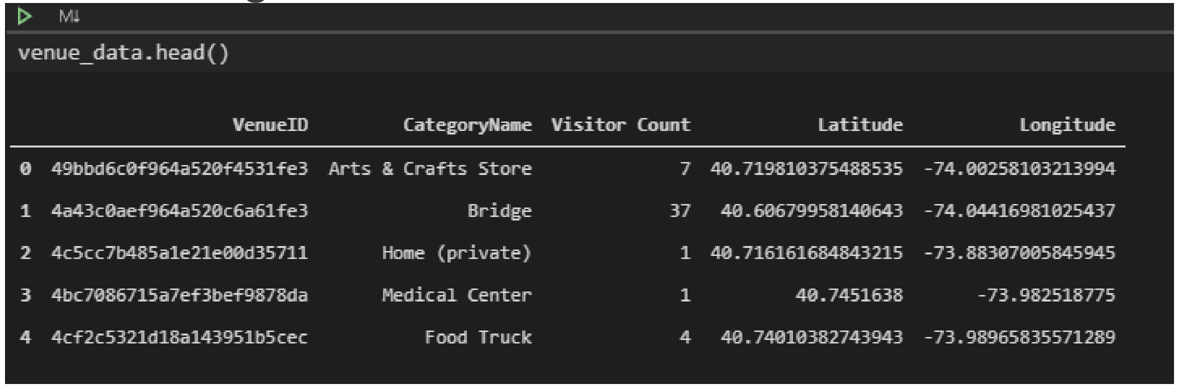
DATA DESCRIPTION

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

- User ID (anonymized)
- Venue ID (Foursquare)
- Venue category ID (Foursquare)
- Venue category name (Fousquare)
- Latitude
- Longitude
- Timezone offset in minutes (The offset in minutes between when this check-in occurred and the same time in UTC)
- UTC time

METHODOLOGY

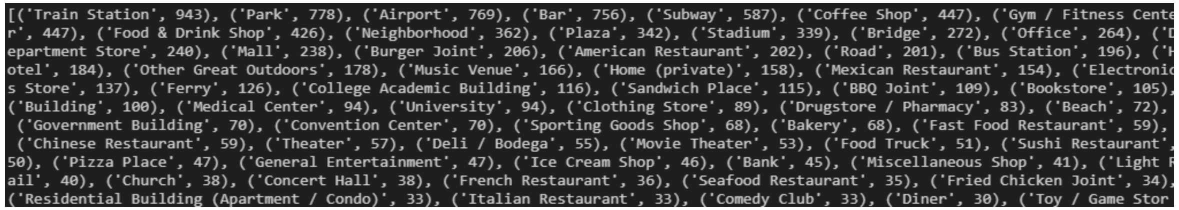
- 1. Converting it to a data-frame



```
venue_data.head()
```

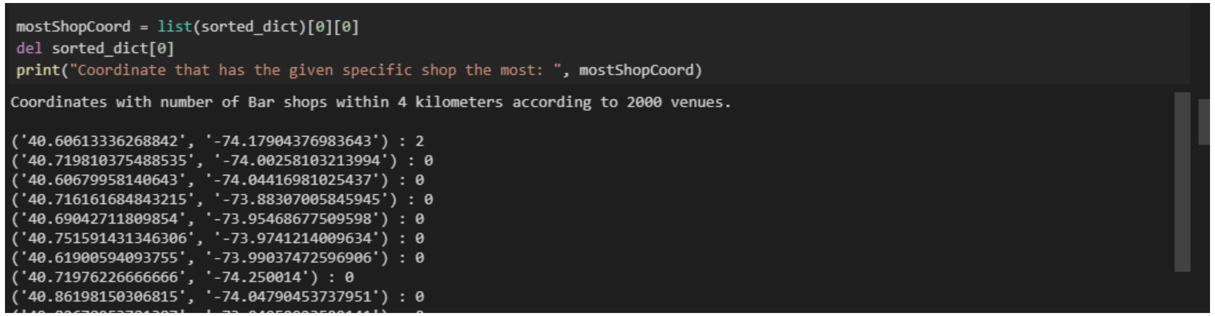
```
                    VenueID         CategoryName  Visitor Count          Latitude           Longitude
0  49bbd6c0f964a520f4531fe3  Arts & Crafts Store              7 40.71981037548853 5 -74.00258103213994
1  4a43c0aef964a520c6a61fe3               Bridge             37  40.60679958140643  -74.04416981025437
2  4c5cc7b485a1e21e00d35711       Home (private)              1 40.716161684843215  -73.88307005845945
3  4bc7086715a7ef3bef9878da       Medical Center              1         40.7451638        -73.982518775
4  4cf2c5321d18a143951b5cec           Food Truck              4  40.74010382743943  -73.98965835571289
```

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



```
[('Train Station', 943), ('Park', 778), ('Airport', 769), ('Bar', 756), ('Subway', 587), ('Coffee Shop', 447), ('Gym / Fitness Cente
r', 447), ('Food & Drink Shop', 426), ('Neighborhood', 362), ('Plaza', 342), ('Stadium', 339), ('Bridge', 272), ('Office', 264), ('D
epartment Store', 240), ('Mall', 238), ('Burger Joint', 206), ('American Restaurant', 202), ('Road', 201), ('Bus Station', 196), ('H
otel', 184), ('Other Great Outdoors', 178), ('Music Venue', 166), ('Home (private)', 158), ('Mexican Restaurant', 154), ('Electronic
s Store', 137), ('Ferry', 126), ('College Academic Building', 116), ('Sandwich Place', 115), ('BBQ Joint', 109), ('Bookstore', 105),
('Building', 100), ('Medical Center', 94), ('University', 94), ('Clothing Store', 89), ('Drugstore / Pharmacy', 83), ('Beach', 72),
 ('Government Building', 70), ('Convention Center', 70), ('Sporting Goods Shop', 68), ('Bakery', 68), ('Fast Food Restaurant', 59),
 ('Chinese Restaurant', 59), ('Theater', 57), ('Deli / Bodega', 55), ('Movie Theater', 53), ('Food Truck', 51), ('Sushi Restaurant',
50), ('Pizza Place', 47), ('General Entertainment', 47), ('Ice Cream Shop', 46), ('Bank', 45), ('Miscellaneous Shop', 41), ('Light R
ail', 40), ('Church', 38), ('Concert Hall', 38), ('French Restaurant', 36), ('Seafood Restaurant', 35), ('Fried Chicken Joint', 34),
('Residential Building (Apartment / Condo)', 33), ('Italian Restaurant', 33), ('Comedy Club', 33), ('Diner', 30), ('Toy / Game Stor
```

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

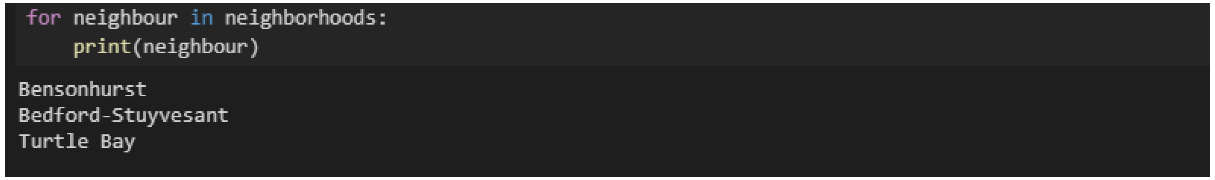


```
mostShopCoord = list(sorted_dict)[0][0]
del sorted_dict[0]
print("Coordinate that has the given specific shop the most: ", mostShopCoord)
```

```
Coordinates with number of Bar shops within 4 kilometers according to 2000 venues.

('40.60613336268842', '-74.17904376983643') : 2
('40.719810375488535', '-74.00258103213994') : 0
('40.60679958140643', '-74.04416981025437') : 0
('40.716161684843215', '-73.88307005845945') : 0
('40.69042711809854', '-73.95468677509598') : 0
('40.751591431346306', '-73.9741214009634') : 0
('40.61900594093755', '-73.99037472596906') : 0
('40.71976226666666', '-74.250014') : 0
('40.86198150306815', '-74.04790453737951') : 0
```

- 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.

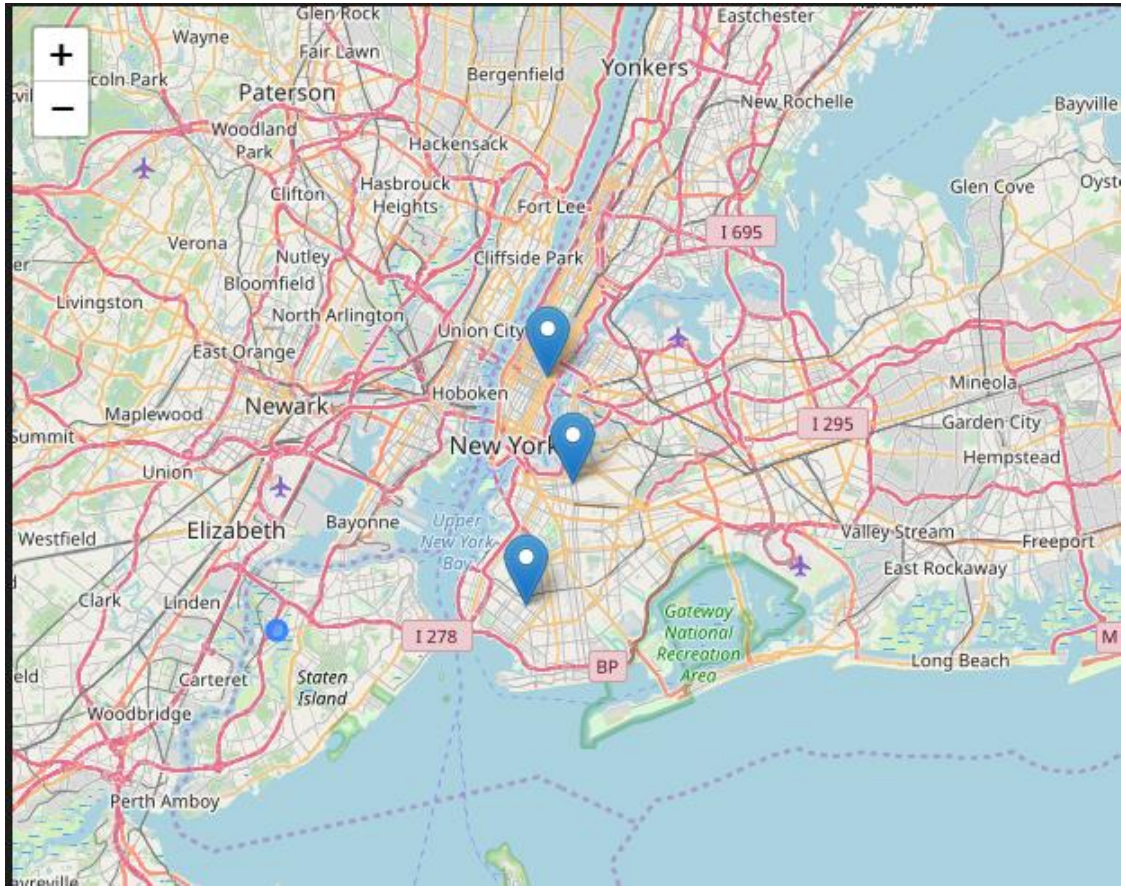


```
for neighbour in neighborhoods:
    print(neighbour)
```

```
Bensonhurst
Bedford-Stuyvesant
Turtle Bay
```

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



THANK YOU

RAISHUBHAM11@LIVE.COM