Applying Data Science in Python to find the Neighborhoods in Tokyo with Potential Business Opportunity

Muhammad Sheheryar Naveed

Introduction

The aim of this project is to identify the areas in Tokyo city which has the potential of a business opportunity. The study is based on the dataset consisting of all the commercial places in Tokyo City with their number of checks-in by the customers. This dataset can be boiled down to find the type of commercial with the most demand in the industry and market. Thereafter finding the neighborhoods in Tokyo which do not have such category of commercial places could be something really useful for stakeholders who are in search of finding the potential business opportunity.

Target Audience

The target audience here are going to be any businessman interested in investing in a commercial activity. Provided a dataset of check-ins of commercial places within a city, this model can help visualize 5 places for a businessman to choose from that have a potential for the commercial activity based on the number of customer visits i.e. market demand.

Dataset

Source:

The data comes from the following source:
https://sites.google.com/site/yangdingqi/home/foursquare-dataset
It consists of data of Tokyo city listing the venues and their respective customer visits accompanied by date time stamps. The data is for ten months from 12 April 2012 to 16 February 2013. It has around 573,703 records for Tokyo. The file is a text file stored as tab separated values. As per the source, this dataset was originally used for studying the spatial-temporal regularity of user activity in LBSNs.

The column names and the first five rows of the dataset are as follows:

	User_ID	Venue_ID	Category_ID	Category_Name	Latitude	Longitude	Time Zone(off. mins)	UTC Time
0	868	4b7b884ff964a5207d662fe3	4bf58dd8d48988d1d1941735	Ramen / Noodle House	35.715581	139.800317	540	Tue Apr 03 18:22:04 +0000 2012
1	114	4c16fdda96040f477cc473a5	4d954b0ea243a5684a65b473	Convenience Store	35.714542	139.480065	540	Tue Apr 03 19:12:07 +0000 2012
2	868	4c178638c2dfc928651ea869	4bf58dd8d48988d118951735	Food & Drink Shop	35.725592	139.776633	540	Tue Apr 03 19:12:13 +0000 2012
3	1458	4f568309e4b071452e447afe	4f2a210c4b9023bd5841ed28	Housing Development	35.656083	139.734045	540	Tue Apr 03 19:18:23 +0000 2012
4	1541	4b83b207f964a5202c0d31e3	4bf58dd8d48988d1f8941735	Furniture / Home Store	35.705074	139.619502	540	Tue Apr 03 19:20:09 +0000 2012

Refinement:

We will then try to refine the data and extract the useful data elements from the imported dataset and load it into a fresh data frame. The new data frame will list the venues along with their respective number of check-ins *i.e. dataset grouped by venues along with each venue_id's visitor count.* Here's a snapshot of the refined data:

	Venue_ID	Category_Name	Latitude	Longitude	Visitor_Count
0	4b7b884ff964a5207d662fe3	Ramen / Noodle House	35.715581	139.800317	6
1	4c16fdda96040f477cc473a5	Convenience Store	35.714542	139.480065	84
2	4c178638c2dfc928651ea869	Food & Drink Shop	35.725592	139.776633	6
3	4f568309e4b071452e447afe	Housing Development	35.656083	139.734045	3
4	4b83b207f964a5202c0d31e3	Furniture / Home Store	35.705074	139.619502	2

Figure 1-Cleaned_dataset

Note: For the sake of simplicity, we will just use first 2000 venues in the dataset.

We now need to group the data by category names, and we get the following dataset after doing that:

	Venue_ID	Category_Name	Latitude	Longitude	Visitor_Count
162	4b19f917f964a520abe623e3	Train Station	35.698596	139.773018	160113
322	4b380ad7f964a520bc4a25e3	Subway	35.665115	139.712459	29302
1091	4b55670ff964a52071e327e3	Electronics Store	35.699015	139.774622	6644
2005	4bbac8b753649c742f7249fb	Office	35.699882	139.772414	3167
1872	4b0a57d1f964a5205d2323e3	Mall	35.532942	139.695765	2745

Since the dataset has some non-commercial categories (for example train station i.e. a businessman cannot open a train station: p) that might be out of our interest pool, so we need to filter out those categories and get the data for only the commercial categories. Here's the final dataset we get:

	Venue_ID	Category_Name	Latitude	Longitude	Visitor_Count
1091	4b55670ff964a52071e327e3	Electronics Store	35.699015	139.774622	6644
1990	4b0bc75ff964a520923323e3	Bookstore	35.700427	139.771752	2503
51	4b5d2ad2f964a520535529e3	Convenience Store	35.552431	139.647492	2137
2362	4df8690581304987d7fadc27	Coffee Shop	35.617692	139.728692	1561
995	4b5bc2e0f964a520ad1429e3	Food & Drink Shop	35.770128	139.660808	1478

We get to know that "Electronics Store" is the top demanding commercial place at the moment based on customer visits.

Please note that the above refinement was just to find out the most visited and active commercial place. For further analysis we will use the *Figure 1-Cleaned_dataset1* above and remove all the unwanted entries from it i.e. train stations etc. and we get the following dataset as our input for analysis:

	Venue_ID	Category_Name	Latitude	Longitude	Visitor_Count
1091	4b55670ff964a52071e327e3	Electronics Store	35.699015	139.774622	2840
1990	4b0bc75ff964a520923323e3	Bookstore	35.700427	139.771752	457
1825	4b556e38f964a5201ae427e3	Electronics Store	35.675439	139.762895	414
2096	4e0db092cc3fffc1d3a2bf4c	Bookstore	35.699508	139.770454	413
1897	4b2af373f964a520beb224e3	Electronics Store	35.672267	139.765776	406

Analysis and Methodology:

We will find the vicinity that lacks this commercial activity. To do so, for each coordinate we find the number of electronics stores present within its 3kms radius. Then we came across the coordinate that has the most number of electronics store within its 3km vicinity. Here's what we got:

```
(35.68435257369705, 139.73719954490662) : 21
(35.68433296624852, 139.74218845367432) : 21
(35.685535547509154, 139.77484703063965) : 21
(35.6864766854162, 139.72904562950134) : 21
(35.686439390552145, 139.68487850418327) : 21
(35.68684072825564, 139.7741263290934) : 21
(35.6847442424678, 139.76341348869008) : 21
(35.686276258829416, 139.73653435707092) : 21
(35.685936700282824, 139.6158176667451) : 21
(35.68668800420691, 139.72854807972908) : 21
(35.68607583173909, 139.77600574493408) : 21
(35.68485310665427, 139.7838270664215) : 21
(35.686038796243366, 139.66934233903885) : 21
(35.686476, 139.784404) : 21
(35.68654857743877, 139.78344082832336) : 21
(35.68440489706805, 139.68536204544893) : 21
(35.686039050003295, 139.7390273) : 21
(35.68681349, 139.773637) : 21
(35.68560219869659, 139.7835977691558) : 21
(35.67887319366422, 139.78747487068176) : 20
(35.679467991491705, 139.74145084619522) : 20
(35.678823082289455, 139.66568380594256) : 20
(35.67837643603711, 139.78663802146912) : 20
(35.68835114950396, 139.76315891999548) : 20
(35.688207767271784, 139.69623030487722) : 20
(35.687082319216536, 139.76545929908752) : 20
(35.688001653824074, 139.69898343086246) : 20
(35.6781683633928, 139.7681871056557) : 20
(35.683384000000004, 139.70154385) : 20
(35.68161184137628, 139.78618204593658) : 20
Coordinate that has the given specific shop the most:  (35.68435257369705, 139.73719954490662)  which is:  21
```

So now we find the neighbors (a total of 5 neighbors) which are closest to this coordinate but has one less than (at optimum) the most number of shops i.e. less than 21 shops. FourSquare api comes into play here for providing us the neighbor when we feed in the coordinates (longitude, latitude) into the api request.

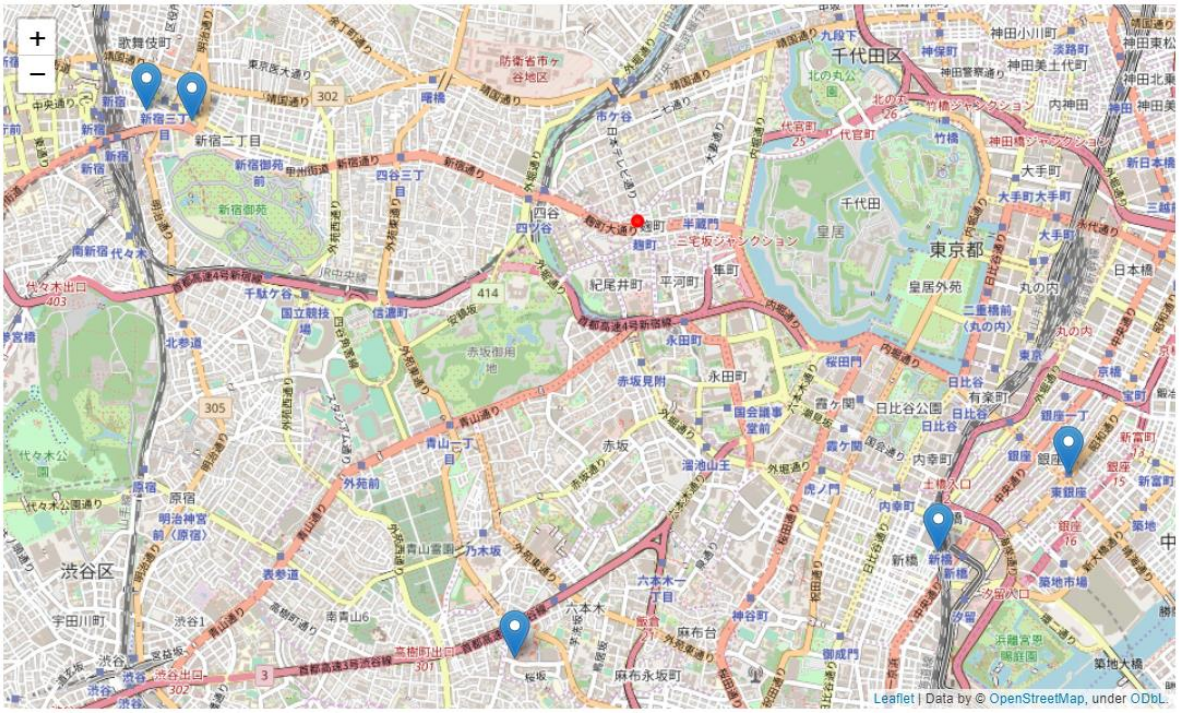
Results

Based on the methodology above, we came to know that these are the 5 target places to open the Electronic Stores:

```
Shimbashi
Roppongi
Shinjuku
Ginza
Shinjuku ward
```

Foursquare returned the results in a Chinese language so we used a translator API to convert it back to English and thus get the above names in English.

The map is as follows, where the red dot denotes the place having the most number of Electronics Store where as the blue markers are the target sites for opening electronic stores to get the customer attention:



Limitations

The data used is a bit old and the market trends changes fairly quicker sometimes. Likewise, we are just analyzing 2,000 venues at the moment for the sake of simplicity. To get a detailed analysis we should not only use the most updated dataset but also increase the magnitude of our input data.

Conclusion

Data Science can be used to enhance business initiation. While there could be limitations in the age, trends, location and collection of data, data science can help cope up with the loopholes and analyze large scale data to provide some valuable insights.