**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:



***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:

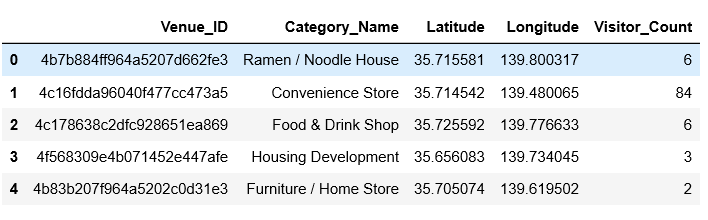
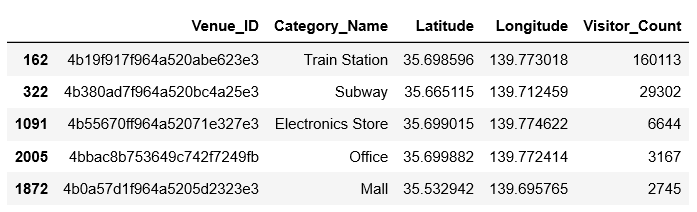


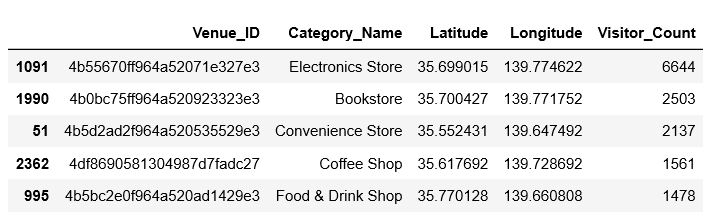
Figure -Cleaned\_dataset

*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:

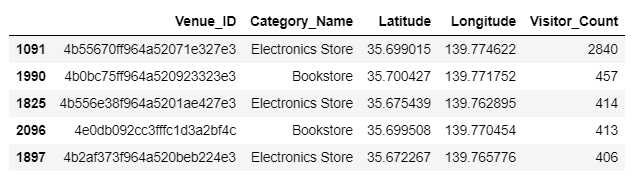


Since the dataset has some non-commercial categories (for example train station i.e. a businessman cannot open a train station: p) that might be out of our interest pool, so we need to filter out those categories and get the data for only the commercial categories. Here’s the final dataset we get:



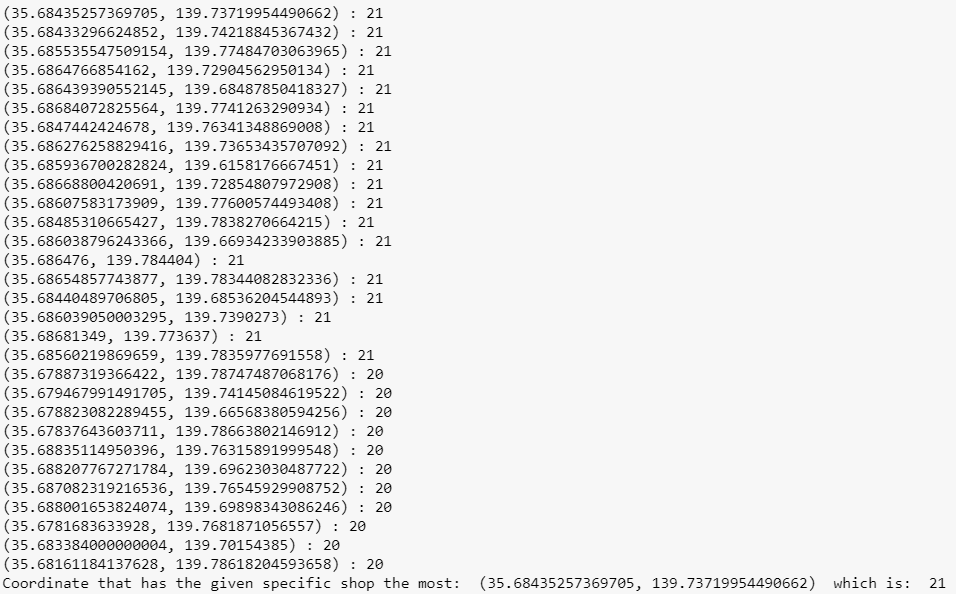
We get to know that “Electronics Store” is the top demanding commercial place at the moment based on customer visits.

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:



***Analysis and Methodology:***

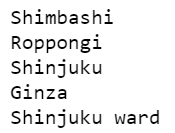
We will find the vicinity that lacks this commercial activity. To do so, for each coordinate we find the number of electronics stores present within its 3kms radius. Then we came across the coordinate that has the most number of electronics store within its 3km vicinity. Here’s what we got:



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.

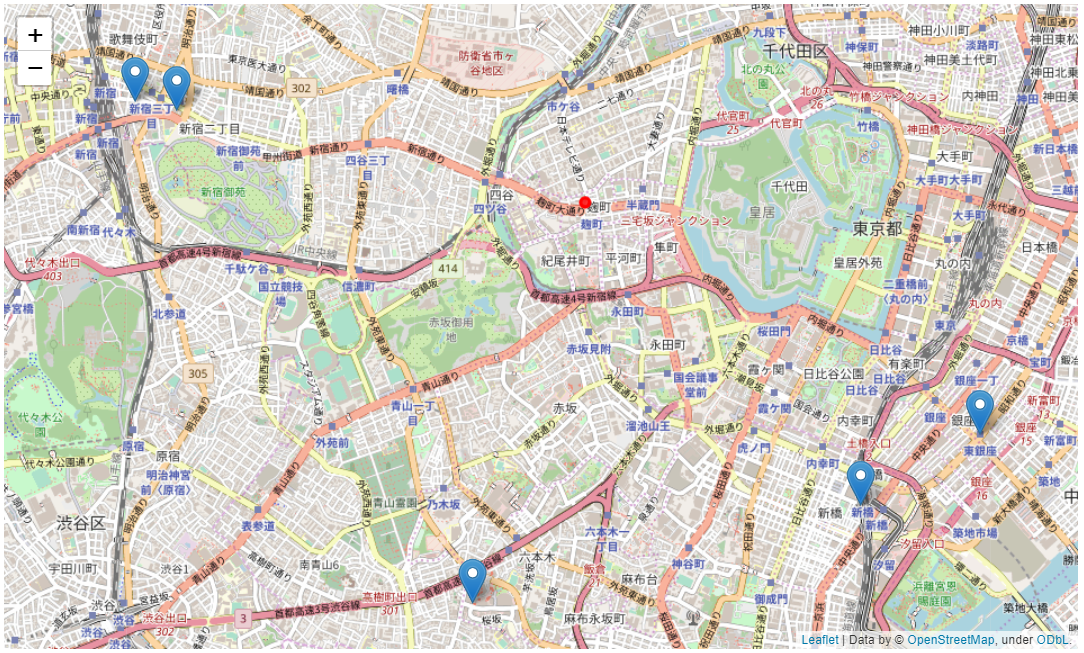
**Results**

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



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.