Coursera Project: Capstone

Where to open a new stadium in Shanghai, China

By: Zhe Chen

Jan. 2020

Introduction

For many residents, stadium is a great place to have tons of fun. They can enjoy the games, do exercise and even find restaurants and do shopping. Nowadays, modern stadiums are more like a shopping mall with sport functions luring all types of people. For the shopkeeper, the attraction of stadium bringing the large crowd provides a great distribution channel to sell their products and services. The stadium developers are also taking advantage of this trend to build more stadiums to cater to the demand. As a result, there are some stadiums in Shanghai, China. Opening stadium allows developers to earn consistent rental income. Of course, as with any business decision, opening a new stadium requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the stadium is one of the most important decisions that will determine whether it will be a success or a failure.

Business Problem

The objective of this capstone project is to analyze and select the best locations in Shanghai, China to open a new stadium. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the business question: In Shanghai, China, where would be recommended, if a property developer is looking to open a new stadium?

Data

To answer the question above, we need the following data:

- List of districts in Shanghai
  We can get it from Wikipedia(https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai).
  To do that, we choose to use web scraping techniques to extract the data with Python requests and beautifulsoup packages.

- The coordinates of these districts
  Then we will get the geographical coordinates of the districts using Python Geocoder package which will give us the latitude and longitude coordinates of the districts.

- Venue data related to the current stadium
  We will use Foursquare API to get the venue data for those districts.

Methodology

First, we need to get the list of districts in Shanghai. The list is available in the Wikipedia (https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shanghai). We will scrap using Python requests and beautifulsoup packages to extract the list of districts data. However, this is just a list of names. We have to get the geographical coordinates in the form of latitude and longitude in order to be able to use the Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas DataFrame and then visualize the districts in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in Shanghai.

Next, we will use Foursquare API to get the top 200 venues that are within a radius of 5000 meters. We need to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the districts in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each district and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each district by grouping the rows by district and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the “Stadium” data, we will filter the “Stadium” as venue category for the districts.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the districts into 3 clusters based on their frequency of occurrence for “Stadium”. The results will allow us to identify which districts have higher concentration of shopping malls while which districts have fewer number of stadiums. Based on the occurrence of shopping malls in different districts, it will help us to answer the question as to which districts are most suitable to open new stadium.

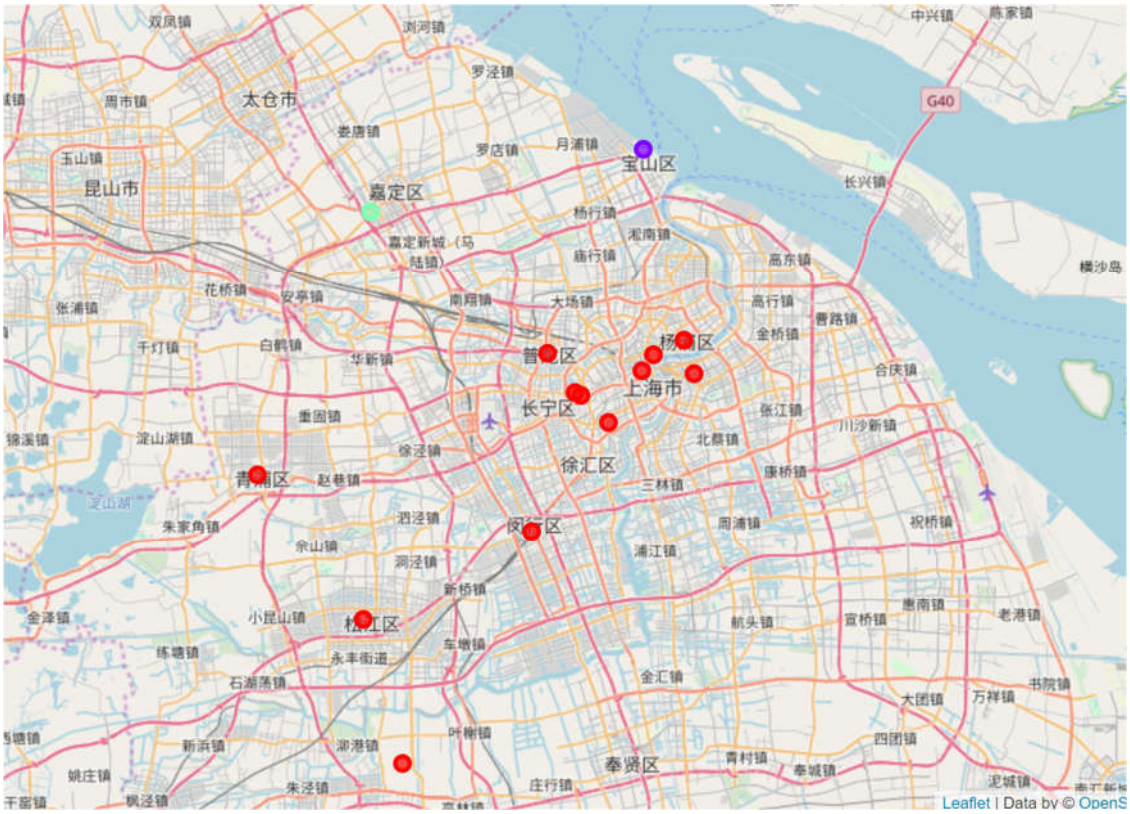
Results

The results from the k-means clustering show that we can categorize the Districts into 3 clusters based on the frequency of occurrence for “Stadium”:

- Cluster 0: Districts with low number of shopping malls
- Cluster 1: Districts with moderate number to no existence of shopping malls
- Cluster 2: Districts with high concentration of shopping malls

The results of the clustering are visualized in the map below with

- cluster 0 in red,
- cluster 1 in purple, and
- cluster 2 in mint green.



Discussion

From the map, we can be aware of that there're one stadium sitting in the Baoshan district and there're two in the Jiading district. The rest districts have zero. According to calculation, cluster 0 has no stadium, which represents a great opportunity and high potential areas to open new stadiums as there is no competition from existing stadiums. Meanwhile, stadiums in cluster 2 are likely meet the residents' demand in that district. Therefore, this project recommends stadium developers to capitalize on these findings to open new stadiums in the districts in cluster 0.

Lastly, stadium developers are advised to avoid the districts in cluster 2.

Conclusion

In this project, we have gone through the process of identifying the business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into three clusters based on their similarities, and lastly providing recommendations to the relevant stakeholders i.e. property developers and investors regarding the best locations to open a new stadium. To answer the business question that was raised in the introduction section, the answer proposed by this project is: The districts in cluster 0 are the most preferred locations. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding highly competitive areas in their decisions to open a new stadium.