Opening a new Gym in Austin, TX

IBM Applied Data Science Capstone Project

By

Jalal Uddin

March, 2020

Introduction:

Located in Central Texas within the greater Texas Hill Country, it is home to numerous lakes, rivers, and waterways. Austin was recently voted the No. 1 place to live in America. It was also named the fastest growing large city in the U.S.A. Austin and its suburb have an estimated population of 2.20 MM. Austin is a hotbed for technology, startups and innovation. A number of Fortune 500 companies have headquarters or regional offices in Austin, including Dell, 3M, Amazon, Apple, Google, IBM, Intel, Oracle, Texas Instruments, and Whole Foods Market.

People in the city and its neighborhood are affluent, health conscious, and willing to spend high dollar to take care of their wellbeing. That's why I have selected Austin, TX for my project to open a new gymnasium.

Business Problem

The objective of this project is to analyze and identify a suitable location in Austin, TX that will have an excellent potential to open a new Gym. It is extremely important to choose an appropriate location where there is less or no Gym at all to ensure the success of the new Gym. Using Data Science Methodology and Machine Learning techniques like clustering, we should be able to recommend the best location for opening a new Gym in the neighborhood of Austin, TX.

Target Audience

Anyone who is looking to open a new Gym in the Austin, TX area is the target audience. Whether a single location for an individual entrepreneur or multiple locations for a big business, it’s a good investment to fulfill the health and wellbeing needs of a modern, health conscious population.

Data Description:

- **Data required**
  - ❖ List of Neighborhoods in Austin, TX. This defines the scope of the project, which is confined to the city of Austin, TX.
  - ❖ Latitude and the Longitude of the Neighborhoods. This is required to plot the map and get the venues.
  - ❖ Venue data, specifically related to Gym. This data will be used to perform Clustering of the Neighborhoods.
- **Source and Method to Extract Data**

  The sources and methods of extracting the data are as follows:

  - ❖ From Wikipedia ('https://en.wikipedia.org/wiki/List_of_Austin_neighborhoods') we will extract and scrape Austin Neighborhood data using various Python commands.

  - ❖ We will get the geographical coordinates of the neighborhoods by using Python Geocoder library and that will give us data of the Latitude and the Longitude for the Neighborhoods.

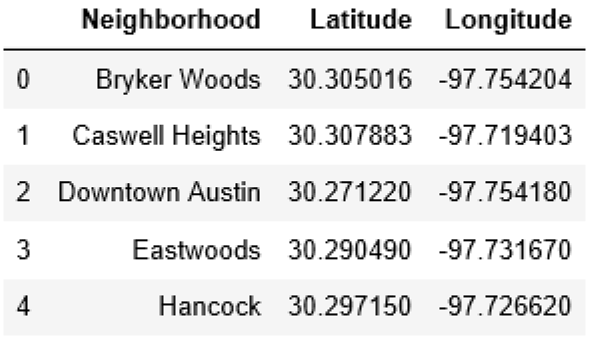
- With a list of Neighborhoods and their Latitudes and Longitudes, we will use Foursquare API to get venue information and we'll select the Gym category for further analysis.

- We will be using K-means Clustering (Machine Learning Technique) to determine suitable location(s) for our new business as well as Folium library to locate them in the Map. The processing of the data will help us identify which neighborhoods has less concentration of Gyms, therefore, indicating a suitable location to open a new one.

- Google Map to validate the neighborhoods.

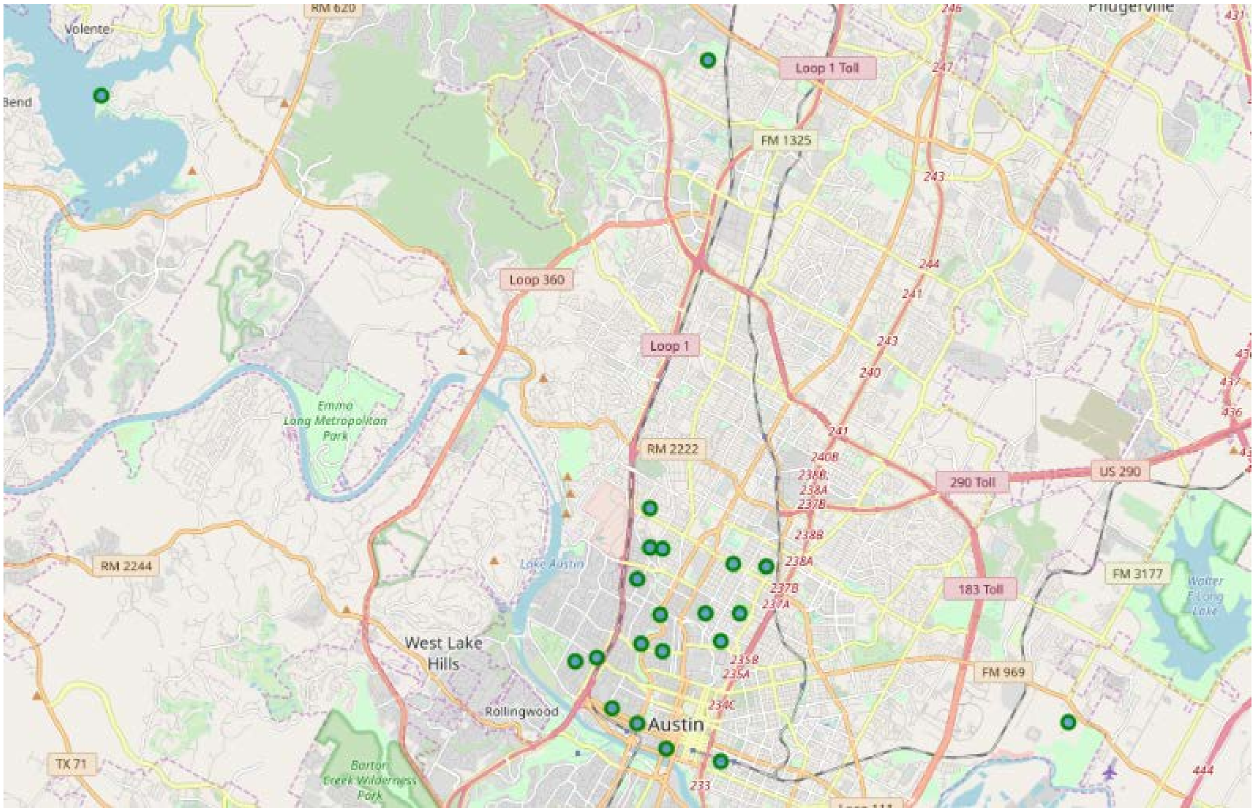
Methodology:

The first data we need is the list of neighborhoods and it is available at Wikipedia ('https://en.wikipedia.org/wiki/List_of_Austin_neighborhoods'). We extract and clean the data by web scraping method and using various python commands to get neighborhood data. Next, we need the geographical coordinates, data for the Latitude and the Longitude of the Neighborhoods to utilize Four square API to get venues and detail analysis. We get data for Latitude and the Longitude using Geocoder library. After gathering data for List of neighborhoods and the Latitude and the Longitude we create a table using pandas data frame.

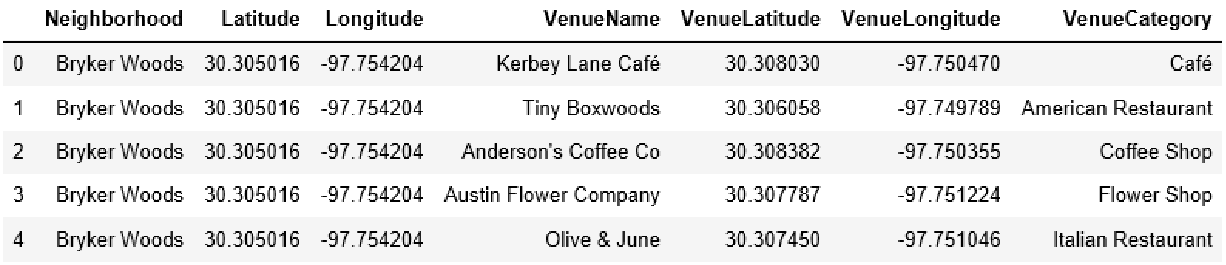


	Neighborhood	Latitude	Longitude
0	Bryker Woods	30.305016	-97.754204
1	Caswell Heights	30.307883	-97.719403
2	Downtown Austin	30.271220	-97.754180
3	Eastwoods	30.290490	-97.731670
4	Hancock	30.297150	-97.726620

We used python folium library to visualize geographic details of Austin and its neighborhoods and created a map of Austin with neighborhoods superimposed on top.



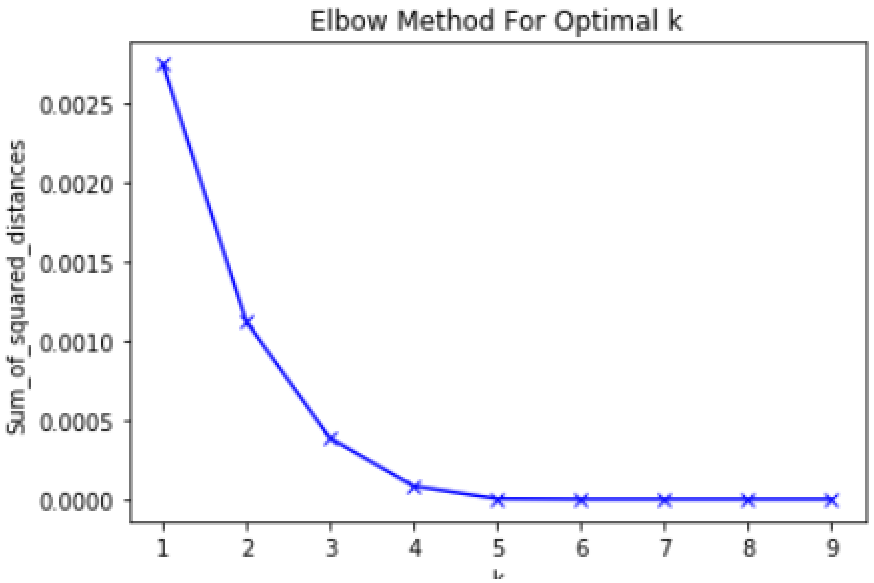
Next, we used Four square API to get the top 100 venues within a radius of 2000 meters. Making a call to Foursquare API we received venue name, venue category, venue Latitude and Longitude.
1,862 venues were returned by Foursquare. Here is a merged table of neighborhoods and venues.



	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	Bryker Woods	30.305016	-97.754204	Kerbey Lane Café	30.308030	-97.750470	Café
1	Bryker Woods	30.305016	-97.754204	Tiny Boxwoods	30.306058	-97.749789	American Restaurant
2	Bryker Woods	30.305016	-97.754204	Anderson's Coffee Co	30.308382	-97.750355	Coffee Shop
3	Bryker Woods	30.305016	-97.754204	Austin Flower Company	30.307787	-97.751224	Flower Shop
4	Bryker Woods	30.305016	-97.754204	Olive & June	30.307450	-97.751046	Italian Restaurant

We checked how many venues were returned and how many unique categories can be extracted from all the returned venues. We analyzed each neighborhood by grouping the rows by each neighborhood and taking the mean frequency of occurrence in each venue category. This also helps preparing the data for use in clustering. We filtered the data for 'gym' as a venue category for the neighborhood.

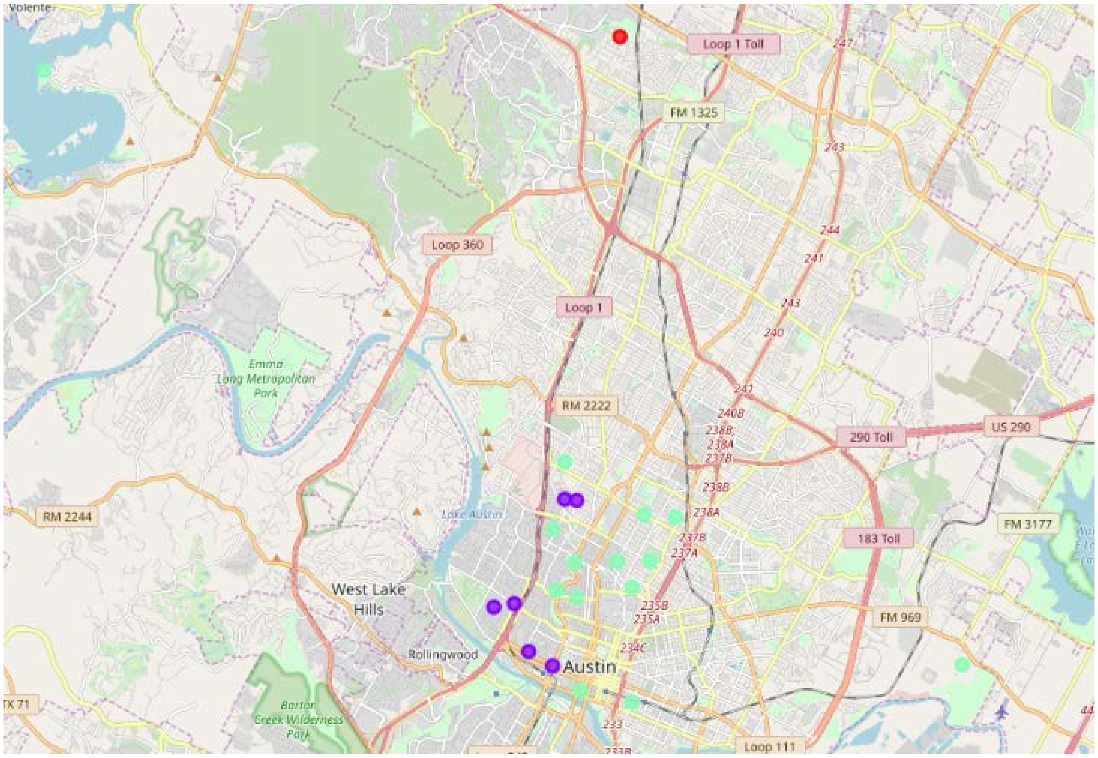
We used unsupervised learning **K-means algorithm** to cluster the neighborhoods. K-means clustering algorithm identifies k number of centroids and then allocates every data point to the nearest cluster, while keeping the centroid as small as possible. It is one of the simplest and popular unsupervised algorithms and suited for our project. We determined the optimal value of K of 3 for the K-means algorithm by elbow method. We clustered the neighborhoods into 3 clusters for the gym.



Results and Discussion:

The result from K-mean clustering shows that we can catagorize the neighborhood into 3 clusters based on the existance of the number of gyms.

- Cluster 0: Neighborhoods with a high concentration of gyms (red circle on the map)
- Cluster 1: Neighborhoods with a moderate number of gyms (purple circles on the map)
- Cluster 2: Neighborhoods with a few number to no gyms (mint circles on the map)



From the result it shows that most of the gym are located in cluster 0 (a total of only 1 neighborhood) which is mostly located in the northern neighborhood of Austin downtown. There are a moderate number of gyms in cluster 1 (a total of 6 neighborhoods) that are mostly located in the North West area of the city. A large number of Neighborhoods (a total of 14 neighborhoods) in cluster 2 are mostly located on the north east side of the city and have a very low concentrations or no gyms. This represents a great opportunity to open new gym in neighborhoods located in cluster 2 where there is very little to no competition.

Conclusion:

Purpose of this project is to identify Austin neighborhoods with a low number of gyms in order to aid stakeholders in narrowing down the search for an optimal location for a new gym. Clustering of those locations is performed to identify major areas of interest to be used as starting points for final exploration by stakeholders. Final decision on an optimal gym location will be made by stakeholders based on specific characteristics of neighborhoods and locations, taking into considerations that additional factors, such as, attractiveness of each location (near offices, proximity to park or water, etc.), proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

References:

[1] Austin-Wikipedia: ('https://en.wikipedia.org/wiki/List_of_Austin_neighborhoods')

[2] Foursquare API

[3] Google Map