Coursera IBM Data Science Capstone

Battle of the Neighborhoods – Determining the Ideal Location for a New Gym in Austin, TX

May 5, 2020

Author: A.J. Murphy

Table of Contents

Introduction

Obesity is a major problem in America, resulting in major health, social, and economic disruptions that have a devastating effect on the country. According to the CDC, the annual medical cost of obesity was estimated at $147 billion in 2008 (1). This equates to an average additional annual medical cost of $1,429 for an obese person as compared to a person of normal weight. This does not even take into account the tertiary costs such as an increase in lost time at work due to higher rates of illness, injury or disability, or the higher insurance premiums that are incurred as a result of obesity (2).

Obesity rates have been on the rise in America for decades. In the last decade alone, the obesity rates have risen from 27.4% for adults in 2011 to 30.9% in 2018. The adolescent obesity rate (students in grades 9-12) has risen from 10.5% in 2001 to 14.8% in 2017 (3). In my home state of Texas, the prevalence of adult obesity was 34.8% in 2018.

Business Problem

The best method to combat the rising obesity rates in America is to promote prevention of obesity before its onset. According to the Center for Disease Control (CDC), the three primary strategies to prevent obesity include state and local programs, community efforts, and healthy living (4). Within the healthy living strategy, one of the primary goals is to ensure an adequate amount of physical activity each week. The current guideline for adults as of 2018 is 150 minutes of physical activity per week, or approximately 30 minutes per day, 5 days a week (5).

A great way to achieve the recommended 150 minutes of physical activity per day is to utilize a gym membership. Gyms offer a wide variety of options for physical activities, which can range from free weights and strength machines, to cardio machines such as treadmills and ellipticals, and many even offer group activities such as yoga and cardio classes. The popularity of gyms has been on a steady rise, with total number of fitness center/health club memberships increasing from 32.8 million in 2000 to 60.87 million in 2017 (6).

Given the severity of the obesity epidemic that plagues our nation, as well as the increasing popularity of gyms, I believe that opening a gym would provide a good return-on-investment both from a monetary perspective, and also from the societal benefit that would be provided by promoting a healthier lifestyle. Specifically, I believe that my home city of Austin is an ideal location to pursue opening a new gym, given Austin's increasing popularity and rapidly growing population. The population of Austin according to the 2010 census was 790,390, and the estimated population as of July 1, 2018 was 964,254, representing a 22% population increase over that timespan.

The target market for this gym would be the entire community in which it resides, as the gym would provide ample variety in activities that all age groups from adolescent to senior citizen would be able to partake and benefit from the physical activity, promoting healthier lifestyles and aiding in the prevention of obesity in Austin.

The Data

To evaluate this business proposition, several data sets will be utilized. The first data sets are the previously referenced CDC Obesity data and Austin Census data, as these will provide a good visualization of the obesity problem in America and reinforce the opportunity that opening a gym in Austin would provide.

The next data set will be a list of all US zip codes (7), from which we can determine all the zip codes within Austin, TX. We will use this as the means by which we will cluster the city, and use data regarding current gym locations obtained from the FourSquare API to determine the area that is currently most underserved in regards to current gym locations within Austin, as that will provide a good indication of where to consider for opening a new gym (8).

Methodology

To properly manipulate, visualize, and analyze the aforementioned datasets, multiple resources will be utilized. Jupyter Notebook, an open-source web application, will be used to utilize Python 3 programming code, which will be the primary means of collecting, analyzing, visualizing, and reviewing the data. The Pandas Library, which offers data structures and operations for manipulating data, will be used to import the datasets from CSV and XLS files downloaded from the internet with data pertaining to US zip codes and data associated with the zip codes, US adult and adolescent obesity rates, and data regarding US gym memberships. Pandas will then be used to extract the relevant data from each dataset and create new concise dataframes containing only the data that is required.

Matplotlib, a plotting library in Python, will be used to create graphs to provide a good visualization of historical US adult obesity rates and of historical US health club/fitness center memberships to provide a tool for justifying the business purpose of opening a new gym.

GeoPy, a client for geocoding web services, will be used to take advantage of Nomatim API package, which will be used to determine the coordinates of Austin, TX via it's search of OpenStreetMap based on the provided input of the city and state. These coordinates will then be used to generate a map with help of the Folium package in Python. Folium is a library that allows for a visualization of interactive spatial data. Once this map is created, markers will be added to the map to indicate the exact location of all the zip codes within Austin city limits, by drawing on the zip code dataframe that was created using Pandas.

Next the Foursquare API will be accessed to search the top 100 venues within each zip code. This list will be converted into a dataframe, and one-hot encoding will be used to determine the number of occurrences of each venue category within each zip code. The mean frequency of each category will then be taken and the updated dataframe

will group the rows of the dataframe by zip code and the resulting mean frequency of each venue category. This dataframe will be used to create a new dataframe which contains only the frequency of gyms in each zip code. There are two categories that could be classified as a gym in Foursquare ("Gym", and "Gym/ Fitness Center"), so frequency of both of these categories will be summed and the new dataframe will be grouped by zip code and the total frequency of the combined gym categories.

The zip codes will then be clustered into 5 clusters based on the mean frequency of occurrence of gyms in each zip code using k-means clustering from the Scikit-learn machine learning library in Python. The resulting cluster labels will be added to the dataframe containing the zip codes and the coordinates and population of each zip code.

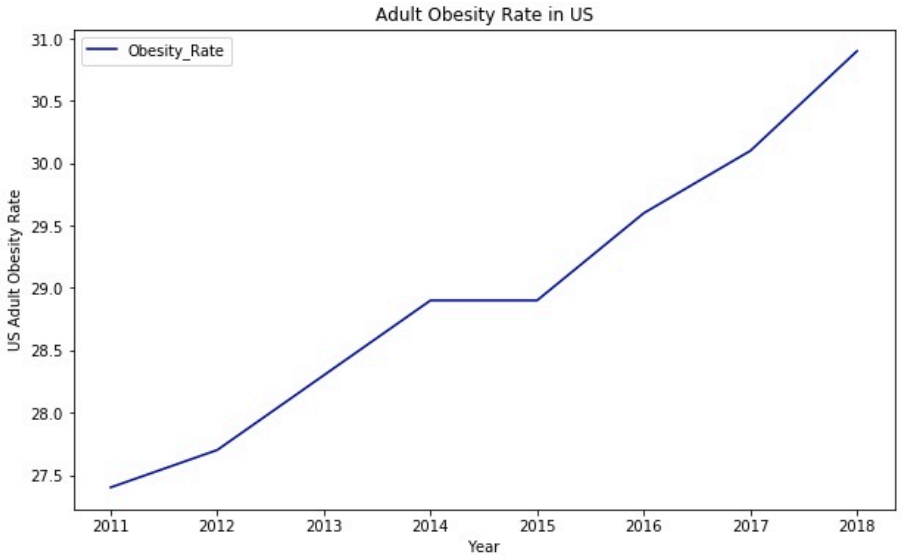
The previously created Folium map displaying the city of Austin, with markers representing each zip code, will then be updated such that the markers are color-coded to identify which cluster each zip code belongs to. Finally, each cluster will be analyzed and the results will be reviewed to make a recommendation of the ideal location for a new gym in Austin, TX.

Analysis

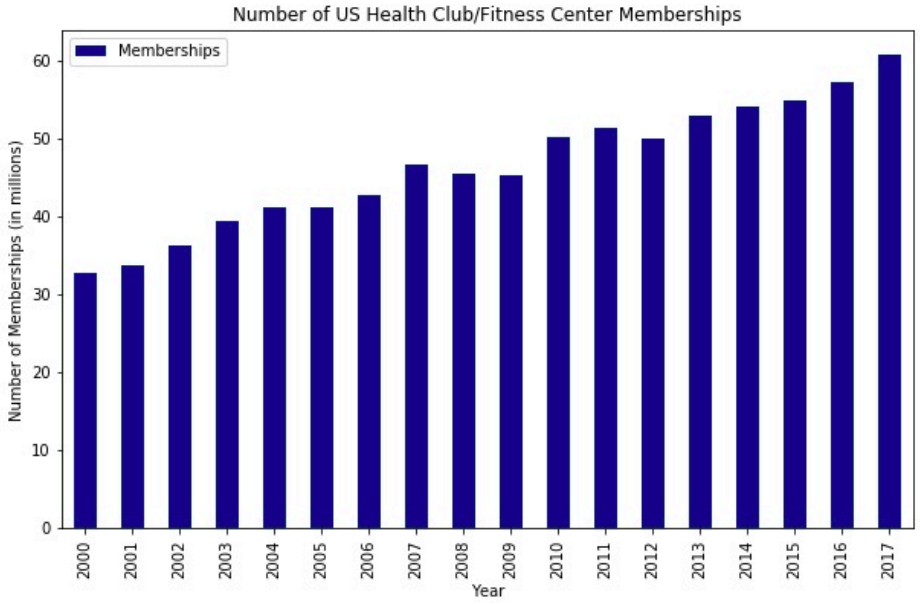
After installing and importing all the required dependencies, the dataset containing US Zip Codes was imported from a XLS file using pandas and then stored as a dataframe. The dataframe was filtered and cleaned to only retain the pertinent data regarding zip codes in Austin, TX . The following was the resulting dataframe (df_atx):

	Zip_Code	City	State	Latitude	Longitude	Population
0	78701	Austin	Texas	30.27049	-97.74235	9427
1	78702	Austin	Texas	30.26327	-97.71432	23389
2	78703	Austin	Texas	30.29409	-97.76571	20890
3	78704	Austin	Texas	30.24315	-97.76537	48486
4	78705	Austin	Texas	30.29437	-97.73855	33948
5	78712	Austin	Texas	30.28502	-97.73477	860
6	78717	Austin	Texas	30.48988	-97.75371	30218
7	78719	Austin	Texas	30.14483	-97.67083	1815
8	78721	Austin	Texas	30.27005	-97.68365	12492
9	78722	Austin	Texas	30.28997	-97.71465	7110
10	78723	Austin	Texas	30.30427	-97.68570	34569
11	78724	Austin	Texas	30.29440	-97.61415	24779
12	78725	Austin	Texas	30.23581	-97.60837	7886
13	78726	Austin	Texas	30.42949	-97.84207	13867
14	78727	Austin	Texas	30.42950	-97.71741	29509
15	78728	Austin	Texas	30.45655	-97.68986	21480
16	78729	Austin	Texas	30.45842	-97.75595	29315
17	78730	Austin	Texas	30.36489	-97.83731	9186
18	78731	Austin	Texas	30.34736	-97.76847	27175
19	78732	Austin	Texas	30.37912	-97.89310	17849
20	78733	Austin	Texas	30.32323	-97.87609	8611
21	78734	Austin	Texas	30.37853	-97.94961	18745
22	78735	Austin	Texas	30.26590	-97.86658	17923
23	78736	Austin	Texas	30.26110	-97.95944	9047
24	78737	Austin	Texas	30.18779	-97.95966	16160
25	78738	Austin	Texas	30.31942	-97.95838	15589
26	78739	Austin	Texas	30.17845	-97.88869	20312
27	78741	Austin	Texas	30.23049	-97.71401	52716
28	78742	Austin	Texas	30.24413	-97.65830	828
29	78744	Austin	Texas	30.18277	-97.72920	48969
30	78745	Austin	Texas	30.20685	-97.79738	62771
31	78746	Austin	Texas	30.29729	-97.81054	28495
32	78747	Austin	Texas	30.12653	-97.74017	20166
33	78748	Austin	Texas	30.16538	-97.82343	50997
34	78749	Austin	Texas	30.21376	-97.85821	37774
35	78750	Austin	Texas	30.41828	-97.80246	30847
36	78751	Austin	Texas	30.31082	-97.72274	15805
37	78752	Austin	Texas	30.33180	-97.70426	21324
38	78753	Austin	Texas	30.38204	-97.67361	59085
39	78754	Austin	Texas	30.35575	-97.64482	24408
40	78756	Austin	Texas	30.32227	-97.74017	8323
41	78757	Austin	Texas	30.35158	-97.73252	24823
42	78758	Austin	Texas	30.38799	-97.70684	47470
43	78759	Austin	Texas	30.40268	-97.76105	42524

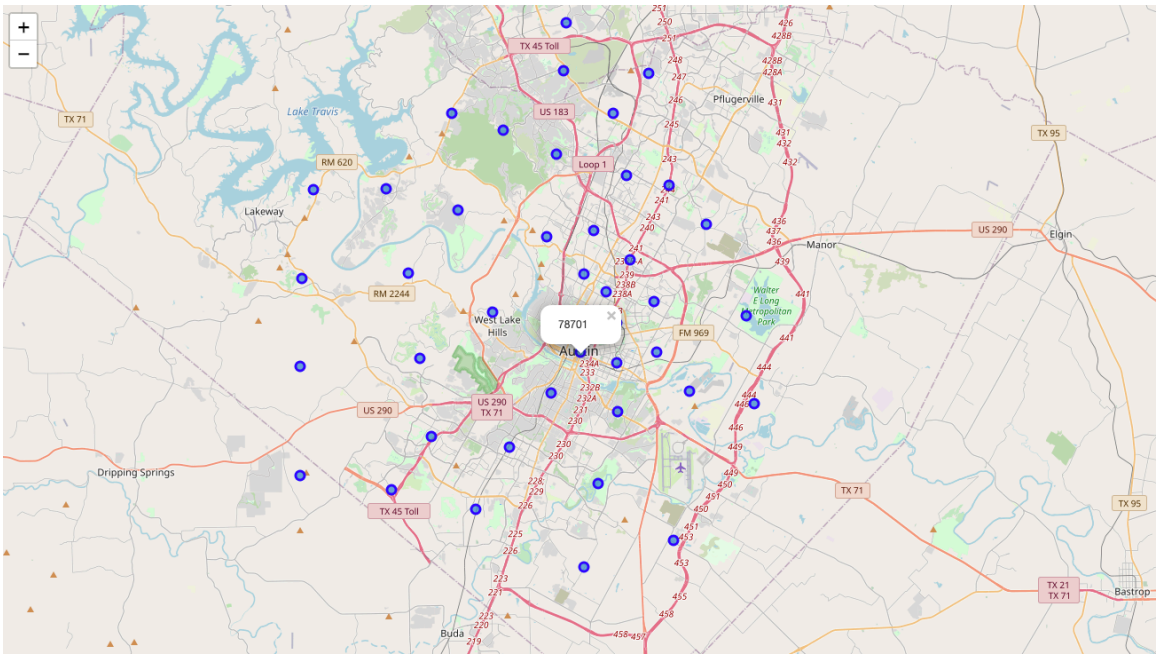
Pandas was then used to create a dataframe from a CSV file obtained from the Center for Disease Control and Prevention's website, containing historical data of adult obesity rates in America. Matplotlib was used to create the following plot. This plot portrays the growing obesity rate in the United States, and is a good visualization of the vital importance to curb this consistent rise in the rate of obesity in the country:



Pandas and Matplotlib were then utilized to perform a similar operation to create a graph displaying the historical trend of Health Club/Fitness Center Memberships in the US based on data obtained from the Statista website. This graph conveys the ever-increasing popularity of gyms in the United States:



The GeoPy library utilized Nomatim to determine the coordinates of Austin, TX as 30.2711286 Latitude and -97.7436995 Longitude. These coordinates were pushed to the Folium map creator to generate a map of the city of Austin. A loop was then used to create a marker for each of the individual zip codes within Austin by utilizing the df_atx dataframe, and these markers were added to the map:



At this point, the Foursquare API was called, using sandbox developer credentials, and a search was run to identify the top 100 venues within each zip code (with the radius set at1600 meters from the coordinates defined for each zip code), and the results were converted into a dataframe using Pandas. The following is what the first 5 results for the first zip code were displayed as using the "head" function:

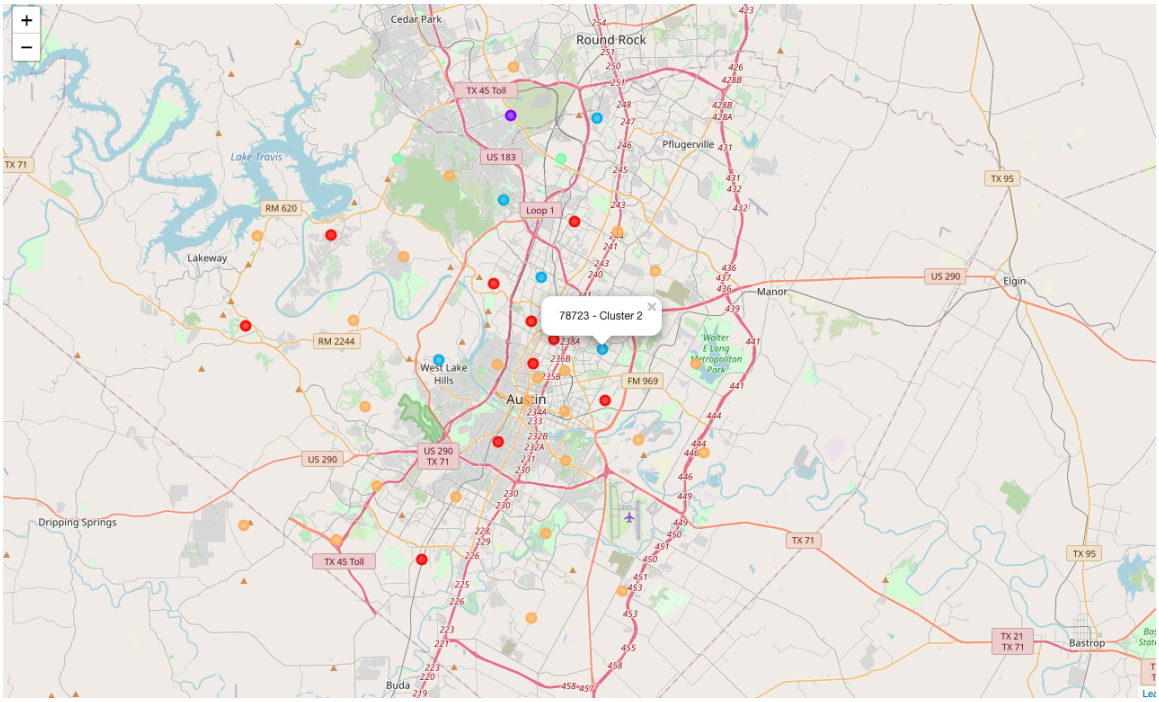
	Zip_Code	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	VenueCategory
0	78701	30.27049	-97.74235	Perry's Steakhouse	30.269374	-97.743676	Steakhouse
1	78701	30.27049	-97.74235	Paramount Theatre	30.269457	-97.742077	Movie Theater
2	78701	30.27049	-97.74235	Chi'lantro BBQ	30.270600	-97.741928	Food Truck
3	78701	30.27049	-97.74235	Caffé Medici	30.270119	-97.742154	Coffee Shop
4	78701	30.27049	-97.74235	The Townsend	30.269611	-97.742448	Lounge

To better analyze the data obtained from Foursquare, one-hot encoding was used to determine the count of each Venue Category occurrence for each zip code, and then the mean frequency of each occurrence of each Venue Category was taken and stored in a new dataframe, grouped by zip codes. Since the target venue category is gyms, the two category groups that classified a venue as a gym ("Gym" and "Gym/ Fitness Center") were summed together, and this total mean frequency was stored.

At this point, the Scikit learn library was used to run a k-means clustering on the dataframe containing the zip codes and the total mean frequency of the occurrence of a gym venue within the respective zip code, with 5 clusters being utilized. The cluster labels were then appended to the df_atx dataframe and sorted, utilizing the "merge" and "sort" functions from Pandas. The resulting dataframe contained the cluster labels for each zip code, along with the frequency of occurrence, coordinates, and population of each zip code in Austin, and the "head" function displays the following as the first five results of this merged and sorted dataframe (atx_merged):

	Zip_Code	Gyms	Cluster Labels	City	State	Latitude	Longitude	Population
39	78756	0.030000	0	Austin	Texas	30.32227	-97.74017	8323
3	78704	0.030000	0	Austin	Texas	30.24315	-97.76537	48486
4	78705	0.020000	0	Austin	Texas	30.29437	-97.73855	33948
24	78738	0.033333	0	Austin	Texas	30.31942	-97.95838	15589
41	78758	0.015385	0	Austin	Texas	30.38799	-97.70684	47470

Finally the Folium map was update with new markers, which were color-coded for each of the 5 clusters. The red markers represent zip codes belonging to cluster 0, the purple markers represent zip codes belonging to cluster 1, the blue markers represent zip codes belonging to cluster 2, the green markers represent zip codes belong to cluster 3, and the orange markers represent zip codes belonging to cluster 4:


78723 - Cluster 2
Round Rock
Cedar Park
Pflugerville
Lake Travis
Lakeway
Manor
Elgin
West Lake Hills
Dripping Springs
Buda
Bastrop
TX 45 Toll
US 183
Loop 1
RM 620
RM 2244
FM 969
US 290
TX 71
TX 95
TX 21

Results and Discussion

Of the 5 clusters that were generated utilizing k-means clustering with the Scikit learn library, the zip codes were clustered more heavily in two of the five cluster, with Cluster 4 containing 25 of the 43 total zip codes in Austin, and Cluster 0 containing 10 of the 43 total zip codes. The remaining clusters had much lower zip codes attributed to them at 1, 5, and 2 total zip codes for Cluster 1, Cluster 2, and Cluster 3, respectively. This was due to the very low level of mean frequency of occurrence of gyms in the zip codes contained within Clusters 4 and 0, and a relatively higher mean frequency of occurrence in Clusters 1, 2, and 3.

Given this, Clusters 4 and 0 would be the optimal clusters to isolate as potential location for a new gym. With the low overall average of the mean frequency of occurrence of a gym occurring in Cluster 4, this would be the target cluster proposing a new location for a gym in Austin. Based solely on the information presented herein, several zip codes that would potentially provide optimum return on investment by generating higher membership rates could be considered as zip codes 78745, 78753, and 78741, as these zip codes are all within Cluster 4, and they have the highest population per zip code within this cluster, providing a greater chance of exposure to the general public.

Additional research could be performed to incorporate average household income, demographic breakdown, and commercial real estate rental and purchase rates within each zip code to make a better assessment of the ideal location to open a gym by further identifying the target market.

Conclusion

Obesity in the United States is a nationwide pandemic, with the rate of adult obesity on a constant rise over the course of the last several decades, resulting in a high economic and societal burden for those affected by it. Obesity prevention is the best method by which to reverse the trend of this increasing rate of obesity. A convenient and economical option for obesity prevention is the utilization of a local gym.

This project used Python in Jupyter Notebook to access and manipulate datasets pertaining to US adult obesity rates, US gym memberships, US Zip Codes and associated census and geographic data. This data was manipulated and analyzed using various Python libraries, such as Pandas, Matplotlib, Scikit Learn, and GeoPy. Additionally, the Foursquare API was utilized to identify the frequency of occurrence of gyms within each zip code in Austin, TX. Using k-means clustering, 5 clusters were created from the 43 zip code within Austin. Using this information, it was identified that based on frequency of gym occurrence alone (negating additional data such as demographics of zip code, median income, etc.), the ideal location for a gym would be in one of the 25 zip codes attributed to Cluster 4. From there, it was suggested that one of the three highest populated zip codes (78745, 78753, 78741) within Cluster 4 be further explored as the potential site of a new gym.

References

1: Centers for Disease Control and Prevention. (2020, February 27). *Adult Obesity Facts.* https://www.cdc.gov/obesity/data/adult.html

2: Harvard T.H.Chan School of Public Health. (2020). *Obesity Prevention Source: Obesity Consequences – Economic Costs.* https://www.hsph.harvard.edu/obesity-prevention-source/obesity-consequences/economic/

3: Centers for Disease Control and Prevention. (2020, April 10). *Nutrition, Physical Activity, and Obesity: Data, Trends, and Maps.* https://nccd.cdc.gov/dnpao_dtm/rdPage.aspx?rdReport=DNPAO_DTM.ExploreByLocation&rdRequestForwarding=Form

4: Centers for Disease Control and Prevention. (2019, October 23). *Overweight & Obesity Strategies to Prevent Obesity.* https://www.cdc.gov/obesity/strategies/index.html

5: Centers for Disease Control and Prevention. (2020, April 10). *Physical Activity Basics - Adults.* https://www.cdc.gov/physicalactivity/basics/adults/index.htm

6: Statista. (2020). *Sports & Recreation – Sports & Fitness*. https://www.statista.com/statistics/236123/us-fitness-center--health-club-memberships/

7: SimpleMaps Interactive Maps & Data. (2020, February 27). *US Zip Codes Database.* https://simplemaps.com/data/us-zips

8: Foursquare Developers. (2020). *Foursquare Developers – My Apps.* https://foursquare.com/developers/apps