An Analysis to Identify Ideal Location for a New Fitness Center in Chicago

Alan Xin

July 9, 2020

Executive Summary

This report utilizes data science and related tools to help choose the ideal location for a new fitness center in Chicago. It uses Python as the primary tool to fetch, analyze and visualize different datasets. There are three datasets used for the analysis in this report, including Foursquare's location database, 2013-2017 American Community Survey Data and Wikipedia.

With visualization and correlation analysis, multiple linear regression algorithm and its model are selected to predict the number of fitness venues needed for the target community in Chicago. Two communities Rogers Park and Albany Park are shown as the most suitable locations to open a new fitness center. In the meantime, additional questions need to be addressed, in order to further decide the detail location.

Part I: Introduction

According to National Center for Health Statistics (note), 39.7% of adults aged 20 and over in the US have obesity. As the obesity issue becomes worse over the last decade among Americans, there is a strong demand for gym and fitness facilities, especially in the cities with large population.

Chicago is the third largest city in the US and well-known for its beautiful downtown skyline as well as it long windy winter. Its cold weather starts in December and stretches all the way to March or even April. In cold and windy days, people are more likely to stay indoor when doing exercise, and therefore gym and fitness facility is a popular business in Chicago. This report will analyze and compare different datasets in order to help business owner find the best possible location for opening a new fitness center in the city of Chicago. The ideal location should have strong customer demand and less competition, i.e. less gyms and fitness centers.

Part II: About the Data

1. Data Source

This report will mainly use three datasets for analysis and discussion, including venue data in Chicago area from Foursquare; the demographics and community data from the Illinois State's data portal website; as well as the geospatial data of Chicago communities from Wikipedia and City of Chicago.

The venue data obtained via Foursquare API will be used to show the current available gym and fitness facilities with their coordinates, i.e. the 'market supply' in Chicago. The geospatial data from Wikipedia mainly provides the coordinates of the community centroids. These coordinates offer the key information when using the explore function to request venue data from Foursquare API based on a location point. The geographic boundaries data from City of Chicago is used to decide which community a venue resides in, as well as for the visualization of communities on the map.

The demographic data from 2013-2017 American Community Survey contains the population, age, employment and other information by communities in Chicago. The venue and demographic data together will be used to analyze which area of Chicago may need more gym and fitness facility services, i.e. the market demand.

2. Data Preparation

a. Venue data

For the venue data, Foursquare provides various types of venue information. The priority of current analysis is to find out the market demand and the market need for gym and fitness service market, thus we will not focus on venues other than gym and fitness related services.

When requesting data from Foursquare, the searching radius of each community's centroid is set to 2500 meters in order to reach to all the communities' boundaries. Since the radius causes adjacent communities having overlapped venues, the duplicates are identified and dropped from the search results. The original result has 2237 venues. With duplicates removed, there are 1086 venues left.

Foursquare sets up a specific catalog 'Gym / Fitness Center' category with 12 sub categories, including boxing gym, climbing gym, cycle studio, gym pool, gymnastics gym, gym, martial arts dojo, outdoor gym, pilates studio, track, weight loss center and yoga studio. Since this report's main purpose to provide suggestion for opening fitness center with large space and more functions, the specialized gym, studio and outdoor venue are not direct competitors for fitness centers. This report will divide the venue data into two groups for further analysis. The venues will be treated as fitness center if they fall in 'gym' category or if they are not categorized as specialized venue above. After renamed and sorted, the venue data (Table 1) include each venue's name, venue category, coordinates and the community it belongs to.

Table 1 Gym and Fitness Venue Samples

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
30	Albany Park	41.9700	-87.7200	Yoga Studio	41.978258	-87.718332	Yoga Studio
85	Ashburn	41.7500	-87.7100	WW (Weight Watchers)	41.735279	-87.704613	Gym / Fitness Center
86	Ashburn	41.7500	-87.7100	Bally Total Fitness	41.736340	-87.705978	Gym / Fitness Center
89	Ashburn	41.7500	-87.7100	Women's Workout World	41.733466	-87.725417	Gym / Fitness Center
90	Ashburn	41.7500	-87.7100	The Path	41.732928	-87.702677	Track

b. Geospatial data

To prepare the geospatial data for Chicago's communities, Beautiful Soup function is used to retrieve the data from the Wikipedia web page. The page has Chicago's all the 77 communities' names and overview, with a link that redirects to each community's own page which contains the coordinates of the community centroid. After cleaning and sorting, remaining dataset has community name, centroid coordinates, as well as area and total population.

The geographic boundary data of all the 77 communities is downloaded as GEOJSON file from City of Chicago's website and does not need to be cleaned (Table 2). It will be used to filter out the venues that are not belonging to that community when calculating the total venue numbers.

Table 2 Chicago Community Geospatial Data Samples

	Community	Density	Population	Area	Latitude	Longitude	Active_Density
0	ALBANY PARK	10455.33	51992	4.97	41.9700	-87.7200	6558.551303
1	ARCHER HEIGHTS	2524.46	13142	5.21	41.8100	-87.7300	1390.125275
2	ARMOUR SQUARE	5195.00	13455	2.59	41.8333	-87.6333	2936.753487
3	ASHBURN	3479.05	43792	12.59	41.7500	-87.7100	2080.063542
4	AUBURN GRESHAM	4739.53	46278	9.76	41.7400	-87.6600	2679.405738

c. Demographic data

There are 83 main features with 233 labels in in the original dataset. In order to make the analysis efficient and consistent, the following steps are performed on the dataset.

Step 1. Main features that are unrelated to people in community are dropped, such as community water, retail sales, park access and etc.

Step 2. After the first step, main features that are dependent to each other are grouped, making the remaining data into 10 main features.

1) Population (Grouped), including Age Cohorts, Household Size;
2) Household Type;
3) Economic Status (Grouped), including Employment Status, Household Income, Monthly Housing & Transportation Cost as a Percentage of Household Income;
4) Commute (Grouped), including Mode of Travel to Work, Vehicles Available;
5) Educational Attainment;
6) Housing & Tenure;
7) Housing (Grouped), including Housing Type, Housing Size, Housing Age;
8) Race and Ethnicity;
9) Nativity;
10) Language Spoken at Home.

Step 3. Remove main features that are not suitable for analysis, including Race and Ethnicity, Nativity, Language Spoken at Home. After this step, there are 7 main features left.

Step 4. Select the sub features that represents the grouped feature.

1) Population (Grouped): keeps only Age Cohorts;
2) Household type;
3) Economic status (Grouped), keeps only Household Income;
4) Commute (Grouped), keeps only Mode of Travel to Work;
5) Educational Attainment;
6) Housing & Tenure;
7) Housing (Grouped), keeps only Housing Type;

Step 5. Group and select one label that can represent for each feature.

1) Population (Age Cohorts): This original feature has 6 labels, including Under 19, 20-34, 35-49, 50-64, 65-74 and Over 85. Because the active groups will be kept, including 20-34, 35-49 and 50-64, which will be combined as 'Active Group'. The 'Active Group' will be used to represent Population feature.
2) Household type: This original feature has Family, Single parent with Child, Non-Family. The Family and Single parent with Child will be grouped into 'Family'. The percentage of 'Family' in total of Family and Non-Family will be used to represent this feature.
3) Economic status (Household Income): The median Household Income for each community will be used for the further analysis.
4) Commute (Mode of Travel to Work): this original feature has 7 labels, including Work at Home, Total Commuters, Drove Alone, Carpool, Transit, Walk or Bike, Other. Since the major difference is about close to work or not. 'Work at Home' and 'Walk or Bike' will be grouped together as 'Close to Work'. The percentage of 'Close to Work' in total group number will be used to represent this feature.
5) Educational Attainment: this original feature has 6 labels, including Less than High School, High School Diploma or Higher, Some College, Associate's Degree, Bachelor's Degree or Higher and Graduate or Professional Degree. The higher education will be grouped together as 'Higher Education', including Some College, Associate's Degree, Bachelor's Degree or Higher and Graduate or Professional Degree. The percentage of 'Higher Education' in the total of higher Education and non-higher education will be used to represent this feature.
6) Housing & Tenure: this feature has 4 labels, including Occupied Housing Units, Owner-Occupied, Renter-Occupied, Vacant Housing Units. The percentage of Owner-Occupied in total of Occupied Housing Units will be used to represent this feature.
7) Housing (Housing Type): Housing type has 5 labels, including Single Family, Detached, Single Family, Attached, 2 Units, 3 or 4 Units, 5 or more Units. Because the most distinguished housing type from others is 5 or more Units, this type will be calculated as the percentage of total number to represent this feature.

After the above 5 steps, the demographic data is ready to be analyzed as shown in Table 3.

Table 3 Sample Demographic Data for Chicago Community

	Community	Median Income	Active Group	Close to Work	Non Family	Higher Edu	Renter	More Units
0	Albany Park	59883.04094	32595.999976	0.051012	0.331214	0.542158	0.279288	0.394359
1	Archer Heights	44108.84007	7242.552685	0.033225	0.230125	0.342306	0.204073	0.039665
2	Armour Square	27463.83031	7606.191531	0.114301	0.356499	0.350517	0.302179	0.457145
3	Ashburn	68463.73500	26187.999992	0.011158	0.197522	0.532218	0.083445	0.042905
4	Auburn Gresham	34661.22650	26151.000004	0.026618	0.346228	0.529572	0.248322	0.217234

Now all the datasets are cleaned by removing unnecessary columns and information from the data. In order to merge the data for comparison, community names are all capitalized to keep the same format between different datasets.

Part III: Methodology

1. Overview

This part will go through details about the progress of analyzing the three datasets: the venue data, the geospatial data and the demographic data that are prepared in the previous part. The analysis will start with visualization of all the venues on the map, then select the preliminary results for the new fitness center from the 77 communities in Chicago. At the end, proper regression algorithm from machine learning will be selected and used to justify which communities should be kept on the final list. Now let's go step by step to analyze the data.

2. Visualization of Venues

Since the goal of this report is to choose a location, it is important to start with understanding Chicago's communities on the map. Using the centroid coordinates, all 77 communities of Chicago are shown as blue circles on the map below (Figure 1). The shape of Chicago is long vertical shape stretching from north to south, the east side of city is connected to Lake Michigan. The downtown of Chicago is in the middle to the right, just next to Lake Michigan. Most of the communities are square-like shapes.

Figure 1 Chicago Community

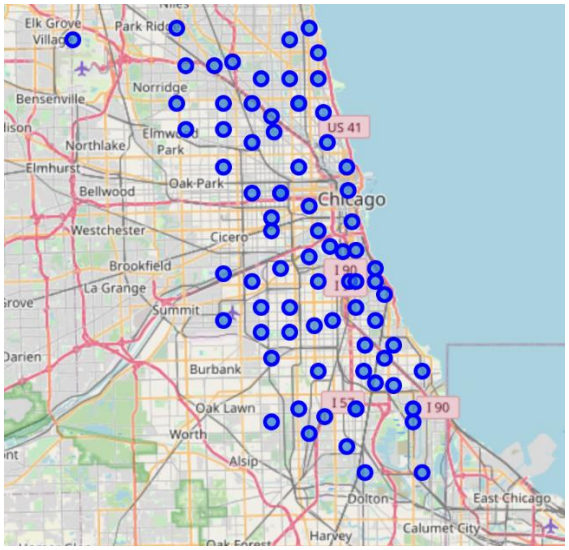
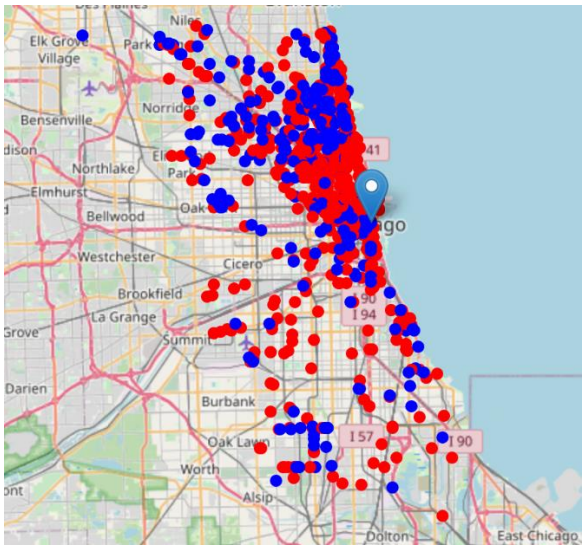


Figure 2 Gym and Fitness Facilities in Chicago

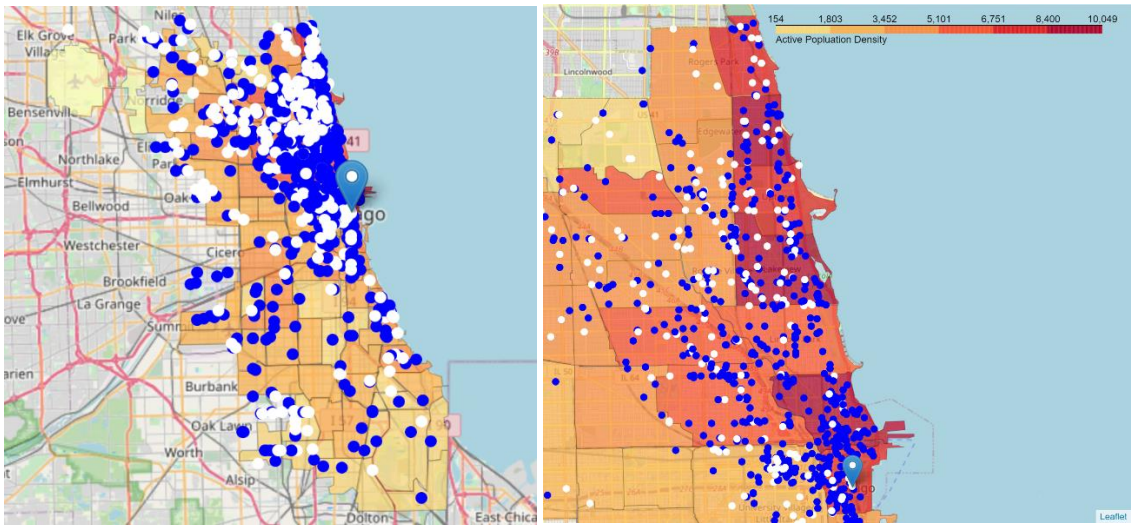


The next venue map (Figure 2), shows all the venues of gym and fitness facilities in Chicago. The red color dots represent the direct competitors: the venues of gyms and fitness center. The blue color dots represent the venue of other types: specialized gyms and studios, outdoor tracks and etc. From the map, the venues are not distributed evenly. It appears that the downtown area, northern/northwestern Chicago and part of the south side along the lake have higher density than the rest of the city.

Next, the population density layer (Figure 3) is applied to the map, in order to compare with the venue distribution. As mentioned in the data preparation section, the active population is grouped and calculated, which includes age between 20 to 64 years old. This is based on the common sense and experience on the major customers of gym and fitness facilities. The population density layer is in orange/red color, with light orange representing lowest density and dark red representing highest density. To better distinguish the venue data on the map, the venues of gyms and fitness centers use blue dots instead of red on the previous map, and other type of gyms use white dots instead of blue.

Based on active population and area of each community, active population density is calculated to represent the population density on this map. Figure 3 on the left shows the venue density mostly matches the population density in each community. The right part of Figure 3 shows the communities with the most population density, which is over 5,000 people/km^2. It appears that some areas in those high populated communities are not fully filled with blue and white dots, which will be analyzed further.

Figure 3 Map with Venue Density and Population Density



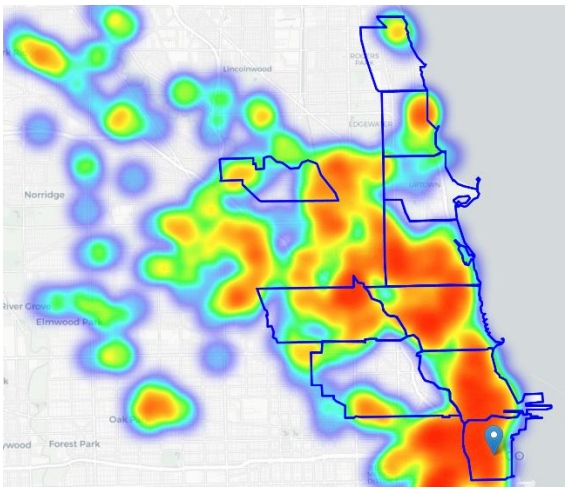
3. Preliminary Selection of Best Locations

The top 10 communities with highest population density are shown in Table 4. From Top 1 to 10, they are Lake View, Near North Side, Edgewater, Rogers Park, Uptown, Loop, Albany Park, Lincoln Park, Logan Square and West Town. The heat map of Figure 4 shows the venue density of the ten communities, while the blue frames outlines the ten communities mentioned above.

Table 4 Top 10 Population Densities List (ppl/km^2)

	Community	Active_Density
0	LAKE VIEW	9952.103960
1	NEAR NORTH SIDE	9542.253520
2	EDGEWATER	8954.323725
3	ROGERS PARK	8008.176103
4	UPTOWN	7192.678869
5	LOOP	6759.953165
6	ALBANY PARK	6558.551303
7	LINCOLN PARK	6196.088019
8	LOGAN SQUARE	5740.215055
9	WEST TOWN	5573.355818

Figure 4 Heat Map of Venue Density



In the heat map among the top 10 most populated communities, there are five communities that have areas with low venue density, which are Edgewater, Rogers Park, Uptown, Albany Park and West Town. Before doing any regression calculation to predict the market supply of fitness facilities, all these blank areas in the communities need to be further investigated on whether

they are suitable for opening a new fitness venue. This can help decrease the unnecessary work load for the regression process.

After internet search and satellite map verification, it appears that the blank areas of Edgewater and Uptown are occupied by cemeteries. Therefore, these two communities are removed from the candidate list. Till now, Rogers Park, Albany Park and West Town are the preliminary candidates ready for the next analysis.

4. Regression Algorithm and Model Selection

In the regression process, the possible correlation is investigated and proper regression algorithm is selected and used to train the model. The following steps are performed.

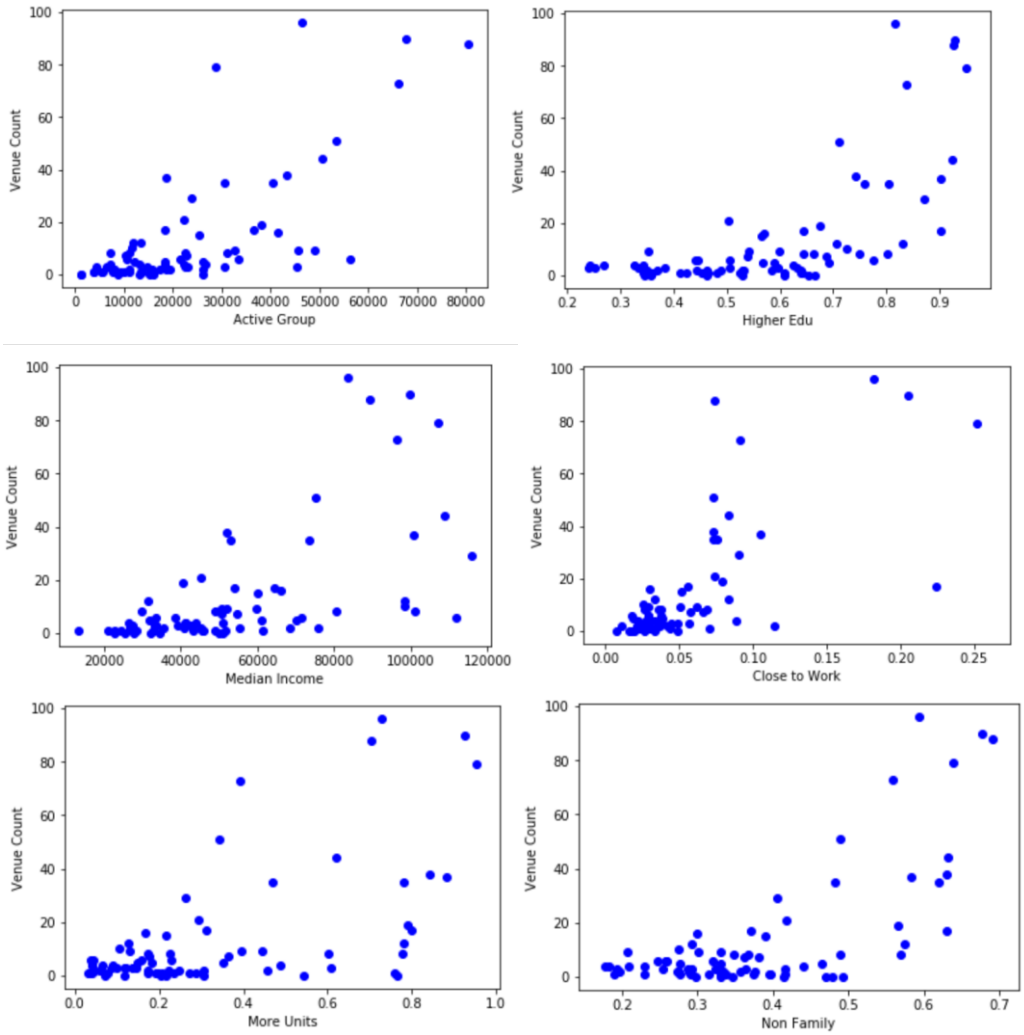
1) The correlation (Pearson Method) between the demographic data features and the number of venue are checked. As Table 5 shows, the venue number has strong correlation with Median Income, Active Group, Close to Work, Non Family, Higher Edu, and More Units. Because of the lack of correlation with Renter is indicated, the Renter Feature is removed.

Table 5 Correlations between Demographic Data and Venue Number (Excluding 3 Candidate Communities)

	Median Income	Active Group	Close to Work	Non Family	Higher Edu	Renter	More Units	Count
Median Income	1.000000	0.348840	0.365791	0.278445	0.689543	-0.503170	0.189910	0.593451
Active Group	0.348840	1.000000	0.342904	0.340964	0.278296	0.164698	0.314803	0.685223
Close to Work	0.365791	0.342904	1.000000	0.671680	0.576835	0.345570	0.724985	0.736766
Non Family	0.278445	0.340964	0.671680	1.000000	0.740838	0.434646	0.851170	0.692396
Higher Edu	0.689543	0.278296	0.576835	0.740838	1.000000	0.002366	0.650604	0.640805
Renter	-0.503170	0.164698	0.345570	0.434646	0.002366	1.000000	0.579020	0.179369
More Units	0.189910	0.314803	0.724985	0.851170	0.650604	0.579020	1.000000	0.616756
Count	0.593451	0.685223	0.736766	0.692396	0.640805	0.179369	0.616756	1.000000

2) The correlation distributions of the remaining 6 features are visualized and shown in the Figure 5. It is obvious that the Close to Work feature are not consistent in the chart, which breaks up between 0.1 and 0.2. In addition. Most of the communities are accumulated between 0 and 0.05, which indicates that most of the population are not ‘Close to Work’. Though this feature is strong correlated with venue number, it is not consistent and has insignificant impact on venue number. Thus, the feature will be dropped. The Median Income, Active Group, Non Family, Higher Edu and More Units are kept for the next step.

Figure 5 Distribution of Demographic Data and Venue Number



3) The five remaining features all show strong correlations with the venue number. Multi linear regression algorithm is selected to train the model with the demographic data as X and the venue number as Y. The random split parameter is set as less than 0.8, with the data split into training set and test set. The model is further fitted using the training set data, followed by applying of the test set data to the trained model. The calculated variance score is 0.72, which is considered as a good score.

   As of now the predicted number is for the total number of venues including both gym and fitness center and other type of fitness facilities. Further filtering for only gym and fitness center is needed for a more accurate result.

4) To predict the venue number for gym and fitness center, the gym and fitness center group will be only used when train and test the model. After filtering out non gym and fitness center venues, there are 330 venues left. The correlation check is performed again and has no material change compared with the previous numbers (Table 6).

Table 6 Correlations between Demographic Data and Venue Number (Excluding 3 Candidate Communities, Only Gym and Fitness Centers)

	Median Income	Active Group	Close to Work	Non Family	Higher Edu	Renter	More Units	Count
Median Income	1.000000	0.377615	0.370751	0.285801	0.692059	-0.479654	0.177699	0.599268
Active Group	0.377615	1.000000	0.358825	0.376885	0.313220	0.192357	0.326218	0.663557
Close to Work	0.370751	0.358825	1.000000	0.675814	0.584201	0.354445	0.720833	0.777078
Non Family	0.285801	0.376885	0.675814	1.000000	0.746269	0.449275	0.847255	0.694509
Higher Edu	0.692059	0.313220	0.584201	0.746269	1.000000	0.020593	0.643480	0.642867
Renter	-0.479654	0.192357	0.354445	0.449275	0.020593	1.000000	0.592517	0.180757
More Units	0.177699	0.326218	0.720833	0.847255	0.643480	0.592517	1.000000	0.609006
Count	0.599268	0.663557	0.777078	0.694509	0.642867	0.180757	0.609006	1.000000

5) The data split, fit and test process are processed again. The variance score is 0.72, which is consistent with the previous analysis. At this point, multi linear algorithm and its trained model has been proven to be suitable. The test set data is put back to the dataset. Then, the entire dataset is used to train the multi linear model again to calculate the coefficients. The coefficients of Median Income, Active Group, Non Family, Higher Edu and More Units are 2.34138277e-04, 5.06717086e-04, 6.88969721e+01, -6.21441736e+00 and 1.39846732e-01, respectively.

Part IV: Result

Based on the algorithm and model chosen in the Methodology, the predicted venue number of all types of gym and fitness facilities are shown in Table 7, and the predicted number of gym and fitness center only are shown in Table 8. From the results, West Town appeared to be a saturated market for gym and fitness facilities based on both predictions. On the other hand, Albany Park and Rogers Park have demands that almost double their current facilities. These two communities will be the final choices to open a new fitness center.

Table 7 Predicted Venue Number vs Current Venue Number (All types)

Community	Predicted Venue Number	Current Venue Number	New Venue Needed
Albany Park	16	9	7
Rogers Park	29	19	10
West Town	55	73	-18

Table 8 Predicted Venue Number vs Current Venue Number (Gym and Fitness Center Only)

Community	Predicted Venue Number	Current Venue Number	New Venue Needed
Albany Park	7	4	3
Rogers Park	12	6	6
West Town	19	26	-7

Part V: Discussion

To make the final call on selecting the ideal location to open new gym, opportunity and challenge from the target community also needed to considered.

1. Challenge

First, the possible challenge may come from specialized gyms and studios. Though these types of facilities don't have all the bells and whistles as fitness center, they may still have capacity to attract some customers. Secondly, when choosing a location, space needs to be available purchase or lease for the venue. In addition, getting a permit for a new business from locals may take unreasonably long time for some areas. All these challenges need to be taken into consideration when deciding the ideal location. Lastly, the effect of COVID-19 pandemic cannot be ignored. How soon will the society be ready for fully open still remains uncertain, and social distancing is not a friend of gym and fitness centers.

2. Opportunity

Opportunity may lie in when there is a good deal for acquiring or leasing the venue location, which is considered one major cost to open a new fitness center. There could also be other opportunity, such as local tax cut for certain type of new business.

To better address all the challenges and opportunities above, field trips and further due diligence need to be planned and conducted.

Part VI: Conclusion

With strong correlations between the number of fitness facilities and the demographic data of communities, the proper venue number to meet the demand can be well predicted. The result based on the multi linear algorithm and model can give the stakeholders a good idea about where to open a new fitness center business in Chicago. The result will be even more valuable with further due diligence investigations.

Reference

1. https://www.cdc.gov/nchs/hus/contents2018.htm#Figure_008
2. https://datahub.cmap.illinois.gov/dataset/2010-census-data-summarized-to-chicago-community-areas
3. https://en.wikipedia.org/wiki/Community_areas_in_Chicago#List_of_community_areas
4. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Community-Areas-current-/cauq-8yn6