-------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- --------

# Gym Venue Optimal Location

IBM/Coursera Applied Data Science Capstone project

Oleksandr Bogach

May 31, 2020

-------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- -------- --------

Table of contents


Introduction....3
  Business goal....3
  Business objectives....3
  Target Audience....3
  In scope....3
  Out of scope....3
  Impact factors overview....3
Data....5
  Data sources....5
  Considered impact factors by data source....5
  Data processing....6
    Data retrieval....6
      Categories retrieval....6
      Venues retrieval....7
      Chicago demographics data....8
    Data consolidation....8
      Venue-community resolution....8
      Community scope reconciliation....9
      Put data together....9
Methodology....12
  Exploratory Data Analysis....12
  Clustering....16
Results....21
Discussion....23
Conclusion....23

## Introduction

This project is about finding an optimal location for a gym. Specifically, this report will be targeted to stakeholders interested in opening a gym in Chicago, IL.

There are a lot of sports facilities in Chicago, so the target is to detect optimal locations: accessible, not crowded with competitors, located closer to where people live, work or spend a lot of time, and proper adjacent tenants. More preferable locations will be identified among communities with higher income and greater population density.

Data science power will be utilized to generate the top five most promising neighborhoods based on these criteria. The advantages of each location will then be clearly expressed so that the best possible final location can be chosen by stakeholders.

### Business goal

Find an *optimal location* for opening a new gym in Chicago, IL.

### Business objectives

- identify positive and negative impact factors on a gym location;
- analyze impact factors across all communities in Chicago, IL;
- suggest the top five locations to open a new gym.

### Target Audience

- gym/fitness business owners looking for an expansion in Chicago, IL
- real estate agencies

### In scope

- Find an optimal location for an abstract gym

### Out of scope

- Publishing intermediate files produced by the code is not intended to comply with data sources Terms of Use and avoid any discrepancies. Nonetheless, this study is completely reproducible.
- Economical efficiency is not considered
- Facility availability is out of scope
- Obtaining foursquare credentials and switching to the Personal tier are not described here

### Impact factors overview

The following factors impact our business problem:

- **Demographics of the community** - Normally, the majority of gym members prefer to have workouts at a nearby facility. This makes demographics an essential factor either to find optimal location or plan gym facilities. Population and Population density can be used to assess foot traffic as it mainly depends on how many people live in a particular community. Foot traffic is also important for a business to grow - walking by people can be converted to regular clients. Income and Population can also be used to help in decision making when choosing from multiple communities or exploring suitable quality, facility specialization, or membership options. The nearby location of the gym also promotes overall attendance. Such demographics factors as education, employment status, industries, occupation, sex, and age can help better understand hidden correlations and therefore should also be explored.

- **Facility accessibility** - Location should be easily accessible by feet, bike, public transport or car. By feet is based on the closest location to such places as subway stations, public transport stops, shopping malls, or business centers. Clients who bike over to the gym facility will require additional area to secure it while on a workout. Such areas are often located nearby public commuting hubs, end of line transport stops or at car parking. Public transport in walking vicinity is also a significant factor, meaning clients can commute more faster and take a short walk to the gym. For clients driving a car, the nearby parking lot is a mandatory consideration.
- **Competitors** - The best strategy with competitors is to avoid or minimize their impact by distancing from them. Competitors are not only facilities that offer the same specialization, but those with similar specializations as well. For instance, yoga studio, fitness studio, pool, and gym are all competing in one way or another. Therefore, distance to the nearest competitors along with the rating will play a significant role in choosing an optimal location for the gym.
- **Adjacent Tenants** - Two sides of the same coin: facilities that promote the gym and facilities that demote it. For instance, located next to a bar, liquor or smoke store, or fast-food venue does not deliver a "be healthy" message, while having such neighbors as organic grocery or sporting gear stores can help to maintain a client base.

## Data

### Data sources

Four sources of data are considered to address the business problem: Foursquare Places API, Chicago GEO json, Chicago Demographics Data, and Reverse GEO coding service

- Chicago GEO json - Defines geospatial shape (boundaries) of Chicago communitites. This geo json file will be used for area calculation and visualization purposes;
- Foursquare Places API - API that helps to get venues at the specific location. Although there are some limitations of API access, this source is considered as a main provider of most up-to-date venues data;
- Chicago Demographic Data - Chicago demographic, income, and education data by community in CSV format and corresponding columns description file. Might be inaccessible out of USA;
- OpenCage Geocoder - API that will be used reverse geocoding, e.g. to get address by geospatial data.

### Considered impact factors by data source

The following table describes data source for the considered impact factors and further detalization.

| factor | data item | data source |
|---|---|---|
| Demographics of the community | communities boundaries | Chicago GEO json |
| Demographics of the community | population, population density | Chicago Demographics Data |
| Demographics of the community | income, education, employment status, industries, occupation, sex, and age | Chicago Demographics Data |
| Facility Accessibility | nearby parkings, transport stops | Foursquare Places API |
| Competitors | nearby gym facilities, rating | Foursquare Places API |
| Adjacent Tenants | "conflicting" venues | Foursquare Places API |
| Adjacent Tenants | relevant (supportive) venues | Foursquare Places API |
| Venue address | reverse geocoding | OpenCage Geocoder |

## Data processing

### Data retrieval

Chicago GEO json file contains shape of Chicago communities, ready to be used for the needs of visualization and area calculation. Number of retrieved Chicago communities is 77 which matches with corresponding Wikipedia page. This gives us geometrical shape (polygon) of all communities expressed as a series of longitude, latitude pairs.

Retrieved geospatial data is then used to calculate rough shape of Chicago area by finding min, max latitude and longitude of all Chicago communities and finally select an estimated viewpoint center of the city area, assuming it has a rectangular shape.

Foursquare Places API provides convenient way to explore specified area and retrieve various well-structured information about places: category, location, opening hours, social media, etc. Usage of the Foursquare Places will be split on two stages. First is to get all available categories and groups according to the Considered impact factors by data source. While second stage is to retrieve all relevant venues.

#### *Categories retrieval*

Regular expressions will serve the best to precisely identify needed categories. There are six categories of venues I was interested in:

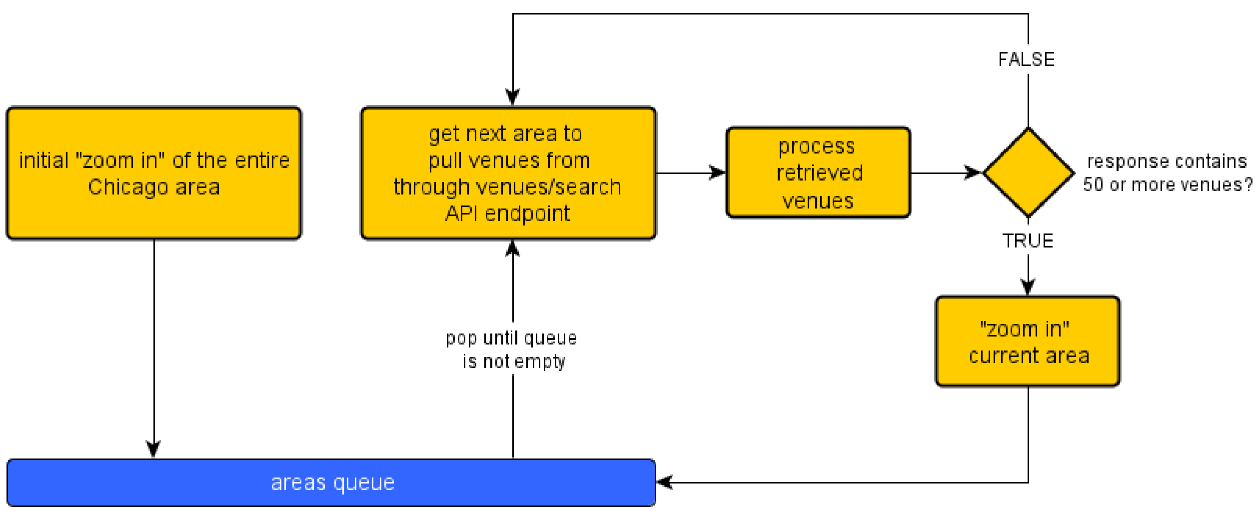
<table>
<tr><th rowspan="2">venue category</th><th colspan="2">category name tokens</th></tr>
<tr><th>included</th><th>excluded</th></tr>
<tr><td>transport</td><td>parking, auto garage, or combination of:<ul><li>bus, metro, train, rail, bike, sub way</li><li>stop, station, rental, lot, area, parking</li></ul></td><td></td></tr>
<tr><td>public places</td><td>movie, general entertainment, or combination of:<ul><li>shopping, business, office, corporate, outlet</li><li>mall, plaza, center, amenity, store</li></ul></td><td>paper</td></tr>
<tr><td>residence</td><td>residence</td><td>college</td></tr>
<tr><td>competitors</td><td>combination of:<ul><li>college</li><li>gym, stadium, baseball, cricket, football, hockey, soccer, tennis, track</li></ul>or combination of:<ul><li>athletics, gym</li><li>sports, center</li></ul></td><td></td></tr>
<tr><td>conflicting</td><td>fast food, nightlife or combination of<ul><li>smoke, vape, liquor, wine, beer</li></ul></td><td></td></tr>
</table>

| venue category | category name tokens | |
|---|---|---|
| | **included** | **excluded** |
| | • *store, shop, bar* | |
| supporting | combination of:<br>• *organic, supplement, health, sporting, bike*<br>• *store, grocery, shop, service* | |

With those categories I came to the second stage – long lasting process of pulling venues data from the Foursquare Places API.

### *Venues retrieval*

Main idea of venue data pulling is based on the city area scanning, simplified flow of which is shown below:




Roughly calculated Chicago area is initially "zoomed in" with factor 10 for each venue category. Areas are defined by four coordinates to form a rectangular shape. Initially this produces 100 sub-areas to be searched in through the Foursquare *venues/search* API endpoint. The former has one limitation which limits response to 50 venues maximum. This makes uncertain every response with 50 venues: whether particular sub-area has exactly 50 venues or more. To unveil that, another "zoom in" is performed whenever a response contains 50 venues, so that divides the area on 100 sub-areas. This continues recursively, until number of returned venues for the particular area becomes less than 50. This decision came at a cost: Because Chicago area was assumed to be rectangular, venues which do not belong to Chicago must be identified and filtered off later.

Another feature of venue data extraction is using a deque to manage the scope of areas to be retrieved. Handling of any exception, e.g. bad response, exceed hourly limit, server error is done through adding of problematic areas to the queue, so that re-iteration will occur later. Venue data retrieval is end when the queue is empty. Once venues data is pulled, the results are stored to disk to avoid API usage next run.

In the result I extracted 31733 venue records including geospatial coordinates, categories (one as per data source, another one according to previously composed venue groups) and two kind of identifiers – original and calculated one.

### *Chicago demographics data*

Chicago demographics data is available per community, available as a csv file, and complemented with columns description pdf file. Although it is being a little bit outdated (dataset update date is June 2019), ratio is assumed to be the same. Analysis of information available allowed to select following data (columns): Total population, age cohorts, race and ethnicity, employment status, mode of travel to work, vehicles available, educational attainment, household income, highly walkable percentage.

## Data consolidation

### *Venue-community resolution*

Venue-community resolution is necessary to filter out venues that do not belong to Chicago area, but were picked earlier due to the selected approach of venues retrieval. This is illustrated on the figure below where blue line polygons represent Chicago communities and green filled area corresponds to the considered Chicago area to retrieve venues data from. At this point every sing

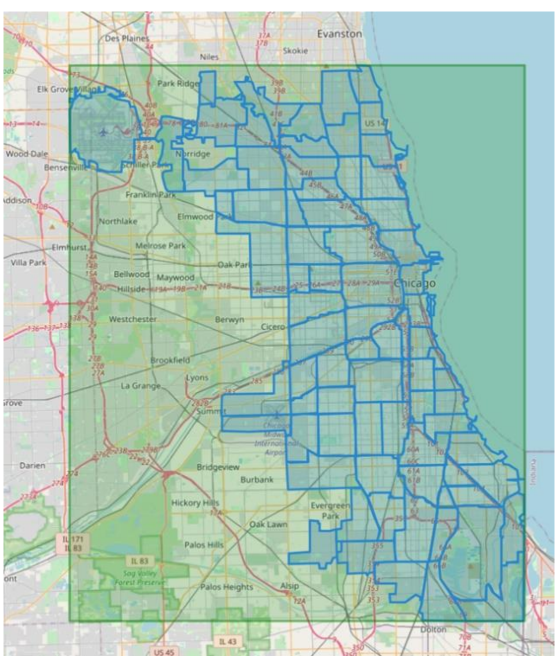



Fig. Considered Chicago area vs Communities

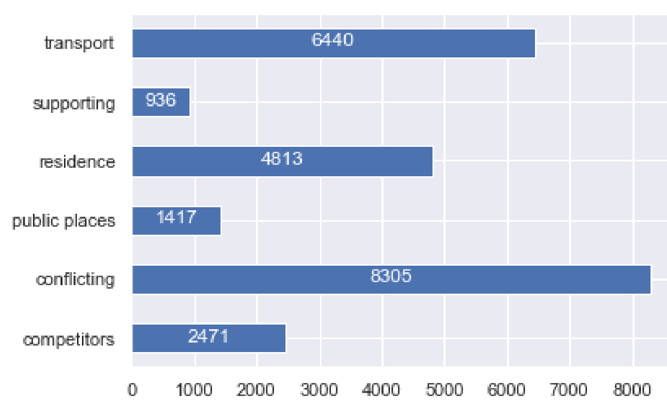



Fig. Retrieved venues breakdown by category

Total number of venues in scope was reduced from 31733 to 24382.

### *Community scope reconciliation*

Comparison of communities naming identified two discrepancies between Chicago GEO json and Chicago Demographics data: the former was fixed by renaming "LOOP" and "OHARE" to "THE LOOP" and "O'HARE" correspondingly.

### *Put data together*

At this point we have collected and pre-processed:

- Chicago venues by category groups and linked them to corresponding community
- Demographics data per community
- GEO json with fixed naming

What is a common scale for different communities of the same city in the context of business goal? Basically, there are two of them: average venue category density per community and venue category distance per community. To calculate the averages following steps were done for each Chicago community area to form an area grid:

- get community boundaries to obtain min, max latitude and longitude
- find community center and place first 300x300m rectangular polygon
- fill community area with adjacent 300x300m rectangular polygon areas in all directions from area center unless border is crossed
- extend coverage if at least one of polygon vertices belongs to the current community area

Latitude, longitude conversion to UTM coordinates was done for zone 16N. This corresponds to Chicago longitude.

Resulting area grid with 8547 areas looks like below

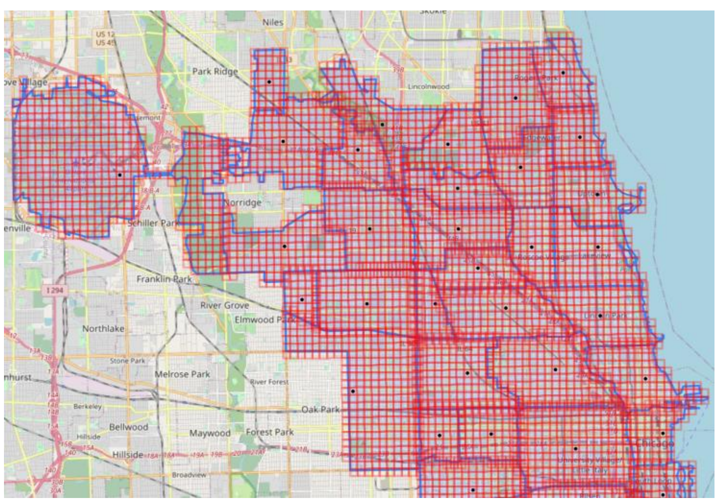



Fig. Grid areas are shown as red rectangles; community boundaries are shown with solid blue line; black dots correspond to community center. Extended areas are noticeable at north east.

Venues allocation by communities was obtained as cartesian product of venues and grid areas with consequent filtering: rough and precise. The former was required to reduce population of the precise filtering and was done on GPU in batches due to the high volumes (200M+). Precise check was necessary to avoid false positive allocations because of polygonal area shape.

As it was intended some venues were simultaneously allocated to more than one community which means that they are reachable from multiple communities simultaneously within 300x300m area.

Then I calculated average venue category density per community as a ratio of venue category count to number of grid cells within particular Chicago community. First five communities (alphabetical order) have following venue category density:

| category | competitors | conflicting | public places | residence | supporting | transport |
|---|---|---|---|---|---|---|
| community | | | | | | |
| ALBANY PARK | 0.439024 | 0.548780 | 0.134146 | 1.158537 | 0.097561 | 1.243902 |
| ARCHER HEIGHTS | 0.176471 | 0.470588 | 0.058824 | 0.073529 | 0.117647 | 0.308824 |
| ARMOUR SQUARE | 0.217391 | 0.717391 | 0.434783 | 0.239130 | 0.304348 | 0.565217 |
| ASHBURN | 0.076923 | 0.360947 | 0.082840 | 0.041420 | 0.029586 | 0.236686 |
| AUBURN GRESHAM | 0.069767 | 0.573643 | 0.077519 | 0.162791 | 0.069767 | 0.263566 |

Fig. Average venue category density per community

With ready grid and allocated venues, I calculated venue category distance for each community as Euclidian distance from area grid cell to closest venue of particular category, which belongs to this grid cell. Below is an illustration of WEST RIDGE Chicago community. Expectedly, some areas don’t have any venue category out of previously identified six groups.

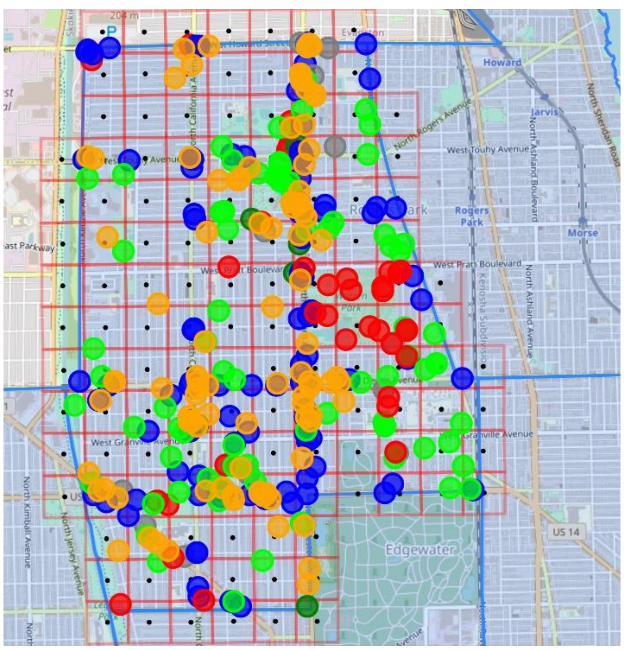



Fig. Grid areas of the WEST RIDGE community with areas centers as black circles and all allocated venue types (colors as per venue category – blue for transport, grey for public places, red for competitors, orange for conflicting, green for supporting, lime for residence)

| | area_id | competitors | conflicting | public places | residence | supporting | transport | community |
|---|---|---|---|---|---|---|---|---|
| 27 | 0180137c4b4b97ffc981fa6d807cd432 | NaN | 74.989431 | NaN | NaN | NaN | NaN | WEST RIDGE |
| 28 | 0181b08d83cbdad58560d84043e726dd | NaN | 114.527762 | NaN | 84.306120 | NaN | 60.471214 | WEST RIDGE |
| 61 | 03448309d537a7ea09b7b33eeafd3fe0 | NaN | 102.305196 | NaN | 33.285950 | NaN | 142.147547 | WEST RIDGE |
| 113 | 06c57fe2b6a33cb101b336068037c722 | 94.248909 | NaN | NaN | 100.852455 | NaN | 88.669067 | WEST RIDGE |
| 187 | 0a91e86d68075903848ecf3a6bd54b3e | 146.844806 | 168.753362 | NaN | 103.514808 | NaN | NaN | WEST RIDGE |
| 211 | 0b3c45a319c5457c2798496f9a1cf492 | NaN | NaN | NaN | NaN | NaN | 126.252841 | WEST RIDGE |
| 281 | 0f820c72135cee0da7fe5a2f3c7fd52d | NaN | 15.111498 | NaN | NaN | NaN | 21.421905 | WEST RIDGE |
| 295 | 1011775b900abd3be6dbe7084bf4df7c | NaN | NaN | NaN | 127.708653 | 128.349032 | 75.775071 | WEST RIDGE |
| 360 | 13aaa45e8254daaab2f1c7c163c75769 | 103.431131 | 109.584972 | NaN | NaN | NaN | 89.507939 | WEST RIDGE |
| 468 | 1a6e432860356192d9c5ba519978b8c6 | NaN | 88.067840 | NaN | NaN | NaN | 94.037285 | WEST RIDGE |

Fig. Calculation results of venue category distance to grid area cell center (in meters)

Since missed values cannot be kept as is and have to be imputed somehow, I decided that the best way to do it was to find distance beyond boundaries of such grid cell areas. So cartesian product of areas with missed venue category and all venues of the same category were obtained to find minimal possible distance per area grid cell per category. This calculation was done on GPU to speed up the process. Finally, two distance datasets were merged into one to form average distance per community per venue category table and assessed from data quality point of view to ensure all areas now have populated distance to a venue of particular category. Figure below shows results for three random Chicago communities:

| category | competitors | conflicting | public places | residence | supporting | transport |
|---|---|---|---|---|---|---|
| community | | | | | | |
| ALBANY PARK | 269.605149 | 317.099854 | 452.109839 | 166.124850 | 551.582775 | 184.619948 |
| THE LOOP | 117.839023 | 130.478823 | 265.548695 | 171.715404 | 205.577592 | 125.122677 |
| WEST RIDGE | 370.366112 | 265.272458 | 514.581730 | 241.431251 | 672.340682 | 202.503256 |

Fig. Complete result of venue category distance to grid area cell center (in meters)

## Methodology

### Exploratory Data Analysis

To identify optimal areas, we need to understand impact factors. So far average distance and average density were calculated, demographics data per community is also available

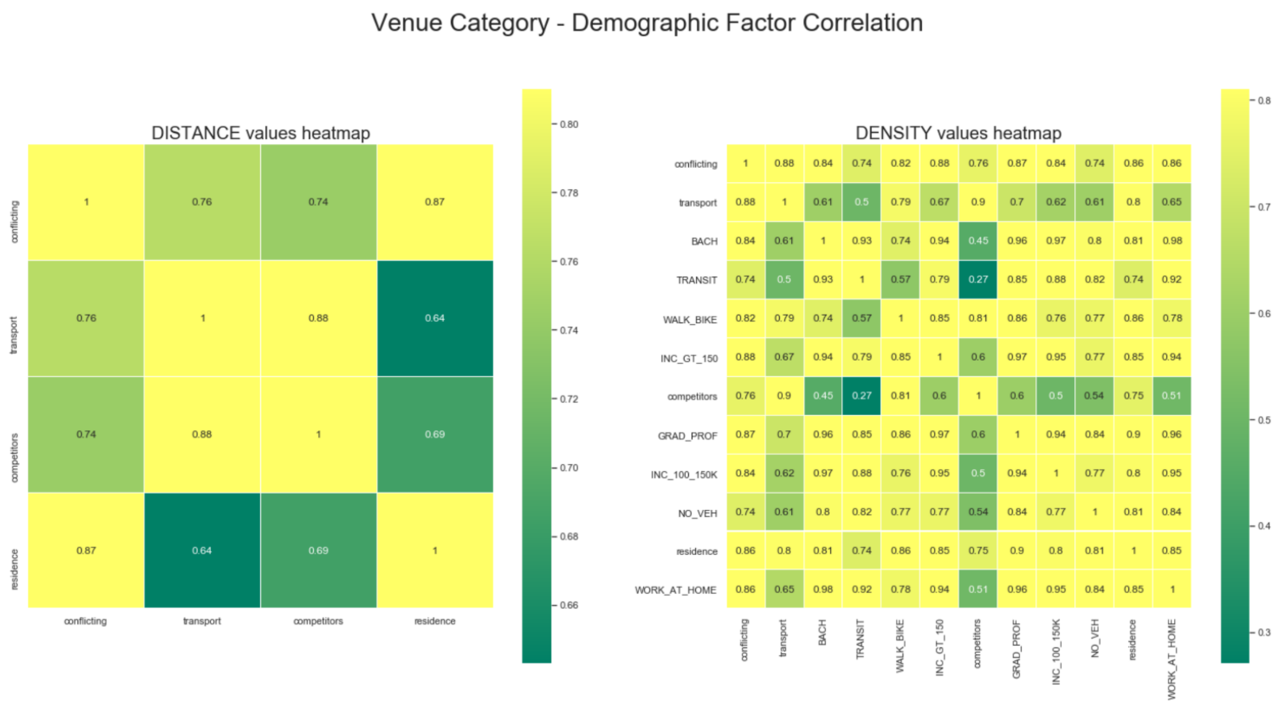



Fig. Venue Category - Demographic Factor Correlation Heatmap

Average density per community has greater number of strong correlations (>0.8), so it is reasonable to use average venue category density data set over average distance data set. Moreover, the latter one indicates a strong, but weaker than density data frame correlation between `conflicting`, `competitors`, `transport` and `residence` factors. Therefore, I focused efforts on the average density per community and demographics data set. Density heatmap in conjunction with the pairwise relationship plot (see below) allowed to bring following insights:

- *competitors* factor:
    - pair *competitors - transport* has strong correlation as it was initially thought. Venues of transport type (as defined at Venue category) act as a primary factor of proper location for a gym
    - areas where people travel to work either by walk or bike (*WALK_BIKE*) is also a strong factor for having a competitor in particular area. Same time correlation between *competitors* and *NO_VEH* (no vehicle in household) suggests previous one is more likely to be not by choice
    - good correlation also exists between *competitors* and *residence*, *conflicting*. To run business effectively, optimal location should be closer to *residence* rather than to *conflicting*
- *transport* factor:

  - correlation *transport - WALK_BIKE* suggests that the former looks like more bike than walk
  - expectedly, strong correlation between *transport* and *residence*, *transport* and *conflicting*
  - good correlation between *transport* and *GRAD_PROF* (Post Graduates)
- *conflicting* factor:
  - this factor has strong correlation with almost all selected demographics factors
  - *WORK_AT_HOME* factor has strong correlation with *conflicting*
  - strong correlation with *BACH* and *GRAD_PROF*. So it's more common within areas with higher number of Bachelors and Post Graduates
  - income *INC_100_150K*, *INC_GT_150* are also in a strong correlation with conflicting factor
- *supporting* and *public places* factors appeared to be uncorrelated with any other factor


Fig. average density – demographics data pairwise relationship plot

Feature set above looks reasonable for further analysis, however, to address the business problem, I will go with unsupervised learning method – clustering. Clustering algorithms are designed to properly reduce number of points in a dataset, so it should fit the best here as identified number of areas for location is 8547 and it must be reduced to 25…50 to select manually from. Dimensionality reduction, such as PCA will not work here as after dimensionality reduction, there usually is not a particular meaning assigned to each principal component because the new components are just the new main dimensions of variation.

As such, I considered average *competitors* density. At first, I calculated average density per community per venue category:

| | community | category | density_avg |
|---|---|---|---|
| **0** | ALBANY PARK | competitors | 0.439024 |
| **77** | ALBANY PARK | conflicting | 0.548780 |
| **154** | ALBANY PARK | public places | 0.134146 |
| **231** | ALBANY PARK | residence | 1.158537 |
| **308** | ALBANY PARK | supporting | 0.097561 |
| **...** | ... | ... | ... |
| **153** | WOODLAWN | conflicting | 0.477273 |
| **230** | WOODLAWN | public places | 0.068182 |
| **307** | WOODLAWN | residence | 0.363636 |
| **384** | WOODLAWN | supporting | 0.011364 |
| **461** | WOODLAWN | transport | 0.625000 |

Fig. Average density per community per venue category

Then I composed data frame with detailed breakdown for each area with at least one competitor venue. It included information about area centroid, count of competitor venues, weight and rank. Weight was calculated as average density divided by competitor venues count (the smaller number of competitors the higher weight) and then ranked area within communities. Rank was turned into negative to make a way for the areas without any competitor venue and corresponding representation on the geo heatmap (otherwise blank areas would be filled on the heatmap and it would make it unreadable). It contains 1342 records.

| | area_id | category | venue_count | community | centroid_lon | centroid_lat | density_avg | weight | rank |
|---|---|---|---|---|---|---|---|---|---|
| **340** | 001bea5d124376f9170b859ffcafbc44 | competitors | 8 | THE LOOP | -87.623985 | 41.878954 | 7.136364 | 0.892045 | -7.0 |
| **341** | 07cdad3da0e99283f6c2b31acbd3e47e | competitors | 17 | THE LOOP | -87.624037 | 41.884358 | 7.136364 | 0.419786 | -15.0 |
| **342** | 1178bf1cdbfe1667123746a8c09b109b | competitors | 12 | THE LOOP | -87.638500 | 41.884279 | 7.136364 | 0.594697 | -11.0 |
| **343** | 11e26605ce22cff3479190341fd7bd0e | competitors | 19 | THE LOOP | -87.631295 | 41.887021 | 7.136364 | 0.375598 | -16.0 |
| **344** | 139063e7110ab333d277ffaa2fbe668b | competitors | 16 | THE LOOP | -87.627653 | 41.884338 | 7.136364 | 0.446023 | -14.0 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| **10655** | e7b1afb4b45a4676c9131f00ea69784a | competitors | 1 | EDISON PARK | -87.810426 | 42.012608 | 0.400000 | 0.400000 | -1.0 |
| **10656** | eab255b9fad496dcf70afefb9101c194 | competitors | 1 | EDISON PARK | -87.806666 | 42.001826 | 0.400000 | 0.400000 | -1.0 |
| **10657** | ed6facfcded3e8631687ae1d9ee69579 | competitors | 5 | EDISON PARK | -87.810220 | 41.996397 | 0.400000 | 0.080000 | -4.0 |
| **10658** | ef7dd0459798d110de169a65227b0090 | competitors | 4 | EDISON PARK | -87.817567 | 42.004451 | 0.400000 | 0.100000 | -3.0 |
| **10659** | fcf8c100a0f6c390d3a00e6a2429b0fc | competitors | 3 | EDISON PARK | -87.806769 | 42.009931 | 0.400000 | 0.133333 | -2.0 |

Fig. Chicago 300x300m areas with competitors

Obtained data frame allowed to exclude those 1342 areas from 8547 grid areas to act upon remaining 7205 areas. To reduce it further, I considered to keep only those areas where distance to closest competitor venue is greater than average for corresponding community. This resulted in 2638 optimal location candidate areas:

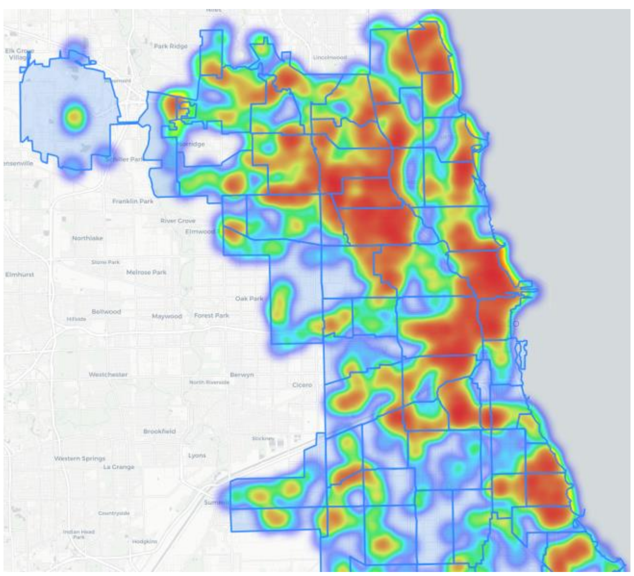


Fig. Competitors heatmap (1342 areas)

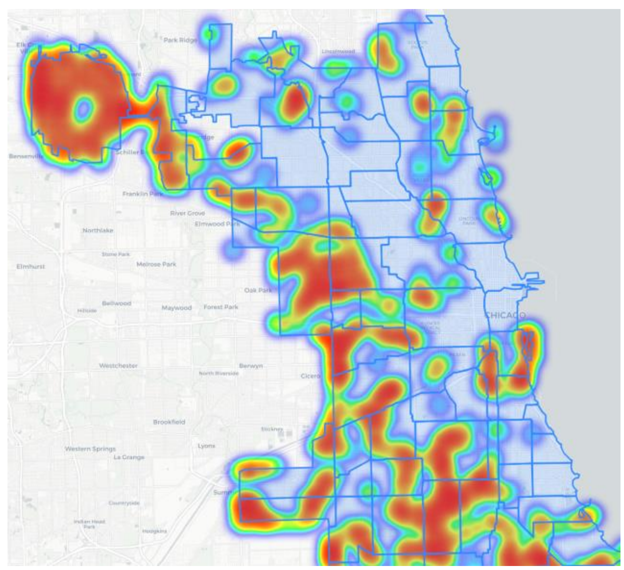


Fig. Optimal location candidate areas without competitors and with distance to the closest competitor venue greater than average for current community

Exploratory analysis indicated transport as a factor strongly correlated with competitors, therefore I considered candidate areas with at least one transport venue. This resulted in candidate venues reduction from 2638 to 429 areas.

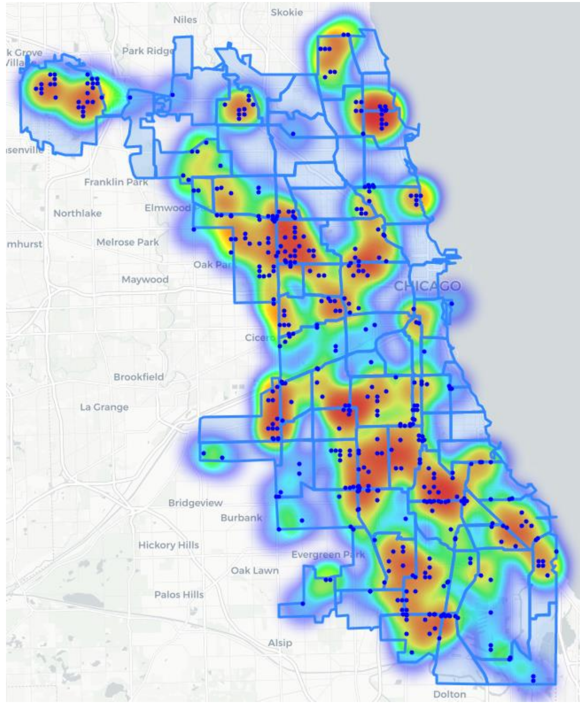



Fig. areas centers where there is at least one venue of transport category (429 areas)

429 areas is still ten times greater than target number (25… 50), so with these I will use clustering to reduce this number even more.

## Clustering

Clustering is done by community where previously identified 429 area candidates are located. Initialization method is `k-means++`. Results of clustering are visualized over the heatmap with candidate areas.

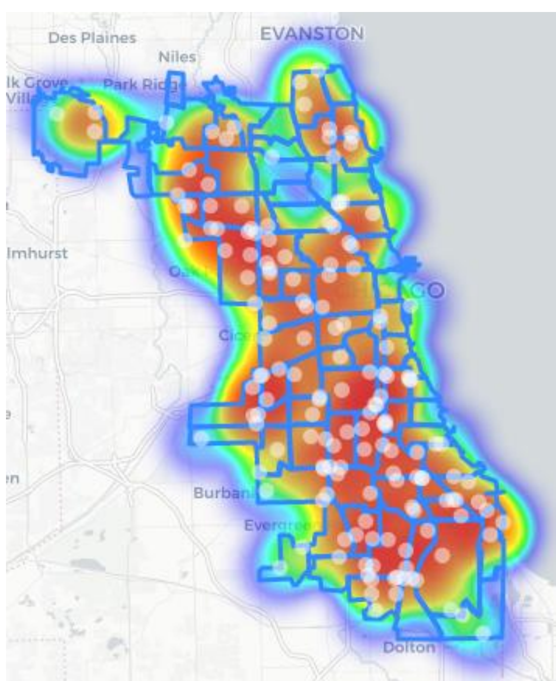



Fig. Clustering 429 area candidates (shown in white)

Special handling (clustering imitation) was done for cases when a community contained either one or two area candidates. Such cases should have been dropped when acted in isolated environment, however I decided to keep them as they may become useful during next stage.

Finding optimal clusters number is based on silhouette score and illustrated below.

Fig. Silhouette score of Clustering

To sum up clustering process, it produced 158 clusters from 429 area candidates.

| | id | community | cluster_number | area_count | lon | lat | x | y |
|---|---|---|---|---|---|---|---|---|
| **0** | LIN-PAR$0 | LINCOLN PARK | 0 | 6 | -87.639458 | 41.925573 | 446979.810667 | 4.641710e+06 |
| **1** | LIN-PAR$1 | LINCOLN PARK | 1 | 4 | -87.669676 | 41.931932 | 444479.802213 | 4.642436e+06 |
| **2** | BEL-CRA$0 | BELMONT CRAGIN | 0 | 4 | -87.746927 | 41.915715 | 438059.503561 | 4.640688e+06 |
| **3** | BEL-CRA$1 | BELMONT CRAGIN | 1 | 3 | -87.782986 | 41.930786 | 435084.496418 | 4.642388e+06 |
| **4** | BEL-CRA$2 | BELMONT CRAGIN | 2 | 2 | -87.755242 | 41.930522 | 437384.503245 | 4.642338e+06 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **153** | NEA-SOU-SID$0 | NEAR SOUTH SIDE | 0 | 2 | -87.633043 | 41.857883 | 447456.239119 | 4.634191e+06 |
| **154** | NEA-SOU-SID$1 | NEAR SOUTH SIDE | 1 | 1 | -87.606001 | 41.864784 | 449706.248905 | 4.634941e+06 |
| **155** | BUR$0 | BURNSIDE | 0 | 1 | -87.602824 | 41.729373 | 449864.475494 | 4.619905e+06 |
| **156** | PUL$0 | PULLMAN | 0 | 2 | -87.590930 | 41.716079 | 450843.559197 | 4.618422e+06 |
| **157** | PUL$1 | PULLMAN | 1 | 3 | -87.609929 | 41.691210 | 449243.567434 | 4.615672e+06 |

158 rows × 8 columns

Fig. Produced clusters

It's good but still insufficient, especially when it comes to review them one by one. Fortunately, this can be automated even more by using graph and longest path. This also explains the reason to keep one or two areas per community and make "imputed" clusters. Cluster radius is decided to be 600m. In order to build cluster group (a cluster of clusters) where they all have overlapped areas is solved by selecting pairs of previously obtained clusters with distance between their centers less than 1200m (radius x2). Resulting vertices matrix will look like below:

| | id_x | id_y | cluster_graph_id_x | cluster_graph_id_y |
|---|---|---|---|---|
| **0** | LIN-PAR$1 | LAK-VIE$0 | 89 | 1 |
| **1** | LIN-PAR$1 | LAK-VIE$1 | 89 | 16 |
| **2** | LAK-VIE$0 | LAK-VIE$1 | 1 | 16 |
| **3** | BEL-CRA$0 | HER$0 | 11 | 65 |
| **4** | HUM-PAR$0 | HER$0 | 17 | 65 |
| **...** | ... | ... | ... | ... |
| **93** | MON$0 | MON$1 | 35 | 39 |
| **94** | LOW-WES-SID$0 | LOW-WES-SID$1 | 76 | 99 |
| **95** | WES-PUL$1 | WES-PUL$3 | 106 | 42 |
| **96** | WES-PUL$2 | WES-PUL$4 | 59 | 49 |
| **97** | ARM-SQU$0 | NEA-SOU-SID$0 | 14 | 10 |

98 rows × 4 columns

Fig. Vertices of clusters graph

Corresponding graph is shown below. For instance, edges with labels 89, 1, 16 form a cluster of clusters with some underlying 300x300m areas.

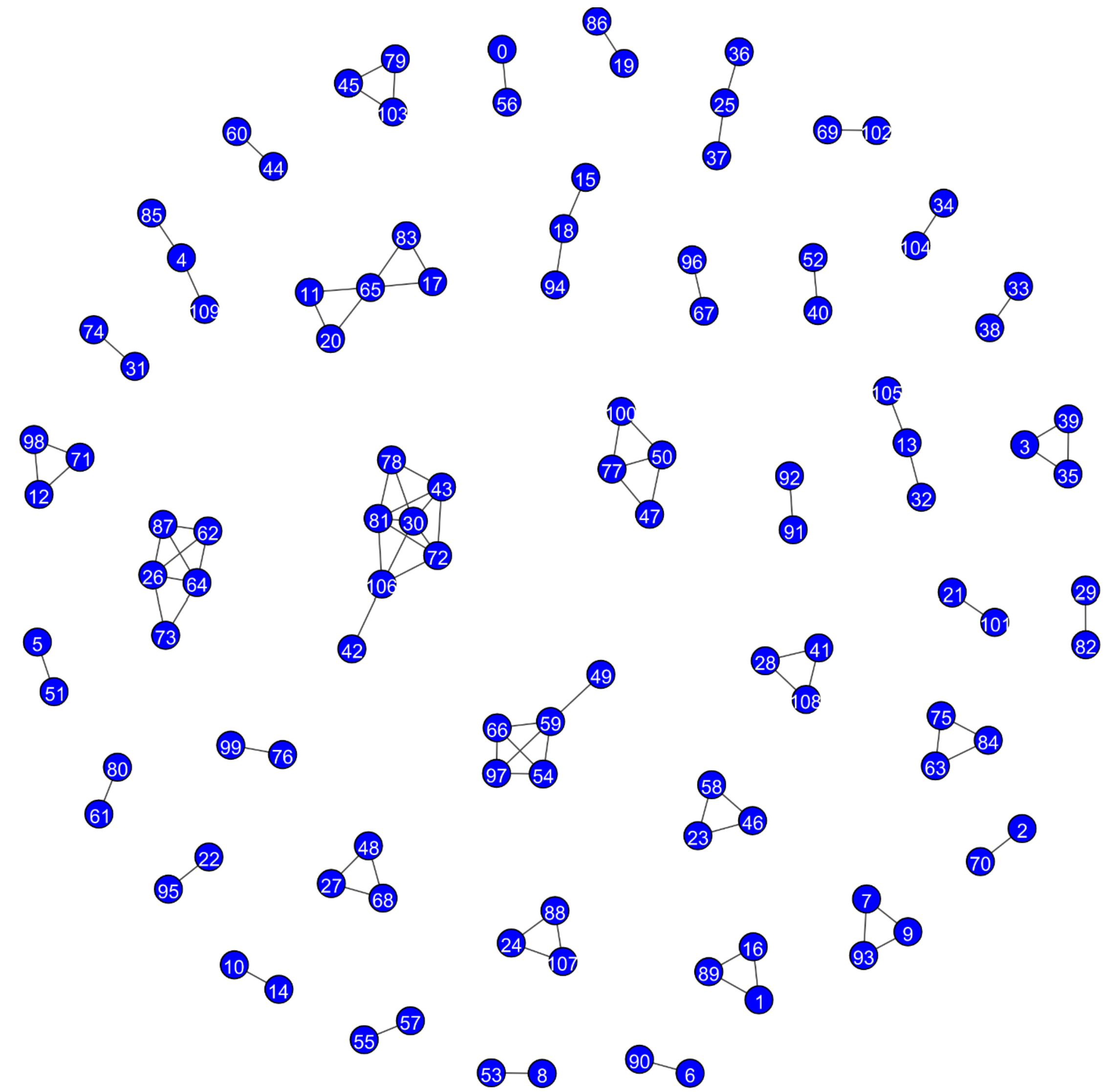



Fig. Cluster of clusters graph

All characteristics are composed into single data frame, including cluster group size and number of areas included

| | id | community | cluster_number | area_count | lon | lat | x | y | cluster_graph_id | cluster_group | group_size |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LIN-PAR$1 | LINCOLN PARK | 1 | 4 | -87.669676 | 41.931932 | 444479.802213 | 4.642436e+06 | 89 | 26 | 3 |
| 1 | BEL-CRA$0 | BELMONT CRAGIN | 0 | 4 | -87.746927 | 41.915715 | 438059.503561 | 4.640688e+06 | 11 | 36 | 5 |
| 2 | BEL-CRA$3 | BELMONT CRAGIN | 3 | 1 | -87.776770 | 41.915517 | 435584.503290 | 4.640688e+06 | 61 | 9 | 2 |
| 3 | CAL-HEI$0 | CALUMET HEIGHTS | 0 | 2 | -87.568766 | 41.736660 | 452702.386875 | 4.620695e+06 | 88 | 31 | 3 |
| 4 | HUM-PAR$0 | HUMBOLDT PARK | 0 | 6 | -87.731039 | 41.908439 | 439370.210588 | 4.639869e+06 | 17 | 36 | 5 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 102 | MON$1 | MONTCLARE | 1 | 1 | -87.805349 | 41.930526 | 433230.131747 | 4.642376e+06 | 39 | 29 | 3 |
| 103 | LOW-WES-SID$1 | LOWER WEST SIDE | 1 | 1 | -87.665280 | 41.852796 | 444776.156886 | 4.633646e+06 | 99 | 16 | 2 |
| 104 | WES-PUL$3 | WEST PULLMAN | 3 | 3 | -87.618941 | 41.674193 | 448480.022077 | 4.613788e+06 | 42 | 39 | 7 |
| 105 | WES-PUL$4 | WEST PULLMAN | 4 | 1 | -87.640564 | 41.674074 | 446680.026289 | 4.613788e+06 | 49 | 37 | 5 |
| 106 | NEA-SOU-SID$0 | NEAR SOUTH SIDE | 0 | 2 | -87.633043 | 41.857883 | 447456.239119 | 4.634191e+06 | 10 | 23 | 2 |

107 rows × 11 columns

Fig. Cluster groups characteristics

Top 3 cluster groups by number of clusters are considered further as more beneficial from area coverage and a strong indicator of an optimal location. OpenCage API was used to retrieve more human readable address by geospatial coordinates. 26 areas instead of 429.

| | id | community | cluster_number | area_count | lon | lat | x | y | cluster_graph_id | cluster_group | group_size | address |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | WES-PUL$1 | WEST PULLMAN | 1 | 4 | -87.619939 | 41.684320 | 448405.031975 | 4.614913e+06 | 57 | 39 | 7 | 120 East Kensington Avenue, Chicago, IL 60628 |
| 1 | ROS$2 | ROSELAND | 2 | 3 | -87.617933 | 41.685286 | 448572.758391 | 4.615019e+06 | 98 | 39 | 7 | St. John M.B. Church, 205-221 East 115th Stree... |
| 2 | PUL$1 | PULLMAN | 1 | 3 | -87.609929 | 41.691210 | 449243.567434 | 4.615672e+06 | 31 | 39 | 7 | Pullman National Monument Visitors Center, 111... |
| 3 | WES-PUL$3 | WEST PULLMAN | 3 | 3 | -87.618941 | 41.674193 | 448480.022077 | 4.613788e+06 | 78 | 39 | 7 | 12105 South Edbrooke Avenue, Chicago, IL 60628 |
| 4 | RIV$0 | RIVERDALE | 0 | 2 | -87.612131 | 41.684248 | 449054.805200 | 4.614901e+06 | 21 | 39 | 7 | 470 East Kensington Avenue, Chicago, IL 60628 |
| 5 | RIV$1 | RIVERDALE | 1 | 2 | -87.604922 | 41.684286 | 449654.805199 | 4.614901e+06 | 70 | 39 | 7 | 690 East Kensington Avenue, Chicago, IL 60628 |
| 6 | SOU-DEE$2 | SOUTH DEERING | 2 | 1 | -87.600809 | 41.683276 | 449996.352685 | 4.614786e+06 | 13 | 39 | 7 | South Doty Avenue, Chicago, IL 60633 |
| 7 | ROS$3 | ROSELAND | 3 | 5 | -87.639394 | 41.692735 | 446792.768713 | 4.615859e+06 | 14 | 36 | 5 | 644 West 111th Street, Chicago, IL 60643 |
| 8 | MOR-PAR$0 | MORGAN PARK | 0 | 2 | -87.645654 | 41.693344 | 446272.418692 | 4.615931e+06 | 25 | 36 | 5 | 11023 South Sangamon Street, Chicago, IL 60643 |
| 9 | MOR-PAR$2 | MORGAN PARK | 2 | 2 | -87.643784 | 41.686599 | 446422.418626 | 4.615181e+06 | 41 | 36 | 5 | 800-900 West 115th Street, Chicago, IL 60643 |
| 10 | WES-PUL$2 | WEST PULLMAN | 2 | 2 | -87.640658 | 41.683531 | 446680.026251 | 4.614838e+06 | 28 | 36 | 5 | 727 West 116th Street, Chicago, IL 60628 |
| 11 | WES-PUL$4 | WEST PULLMAN | 4 | 1 | -87.640564 | 41.674074 | 446680.026289 | 4.613788e+06 | 42 | 36 | 5 | 12101 South Emerald Avenue, Chicago, IL 60628 |
| 12 | AUB-GRE$0 | AUBURN GRESHAM | 0 | 5 | -87.670693 | 41.753575 | 444240.835554 | 4.622634e+06 | 10 | 37 | 5 | 7711 South Wolcott Avenue, Chicago, IL 60620 |
| 13 | ASH$2 | ASHBURN | 2 | 4 | -87.682875 | 41.756930 | 443231.014361 | 4.623014e+06 | 67 | 37 | 5 | Coc Water South District Headquarters, 7501-75... |
| 14 | WES-ENG$2 | WEST ENGLEWOOD | 2 | 3 | -87.667814 | 41.757883 | 444483.926107 | 4.623110e+06 | 20 | 37 | 5 | 1815 West 74th Street, Chicago, IL 60636 |
| 15 | CHI-LAW$1 | CHICAGO LAWN | 1 | 3 | -87.683370 | 41.758413 | 443191.200460 | 4.623179e+06 | 11 | 37 | 5 | 7410-7426 South Western Avenue, Chicago, IL 60643 |
| 16 | WES-ENG$3 | WEST ENGLEWOOD | 3 | 2 | -87.677441 | 41.758277 | 443683.928800 | 4.623160e+06 | 44 | 37 | 5 | 2148 West 75th Place, Chicago, IL 60620 |
| 17 | HUM-PAR$0 | HUMBOLDT PARK | 0 | 6 | -87.731039 | 41.908439 | 439370.210588 | 4.639869e+06 | 66 | 38 | 5 | 1513 North Keeler Avenue, Chicago, IL 60651 |
| 18 | BEL-CRA$0 | BELMONT CRAGIN | 0 | 4 | -87.746927 | 41.915715 | 438059.503561 | 4.640688e+06 | 58 | 38 | 5 | 1923 North La Crosse Avenue, Chicago, IL 60639 |
| 19 | HER$1 | HERMOSA | 1 | 4 | -87.732159 | 41.917445 | 439285.848666 | 4.640870e+06 | 16 | 38 | 5 | 4218 West Armitage Avenue, Chicago, IL 60639 |
| 20 | AUS$6 | AUSTIN | 6 | 2 | -87.749511 | 41.913487 | 437843.078345 | 4.640442e+06 | 100 | 38 | 5 | 4900 West Bloomingdale Avenue, Chicago, IL 60639 |
| 21 | HER$0 | HERMOSA | 0 | 1 | -87.742956 | 41.912648 | 438385.851680 | 4.640345e+06 | 86 | 38 | 5 | 4700 West Grand Avenue, Chicago, IL 60651 |
| 22 | NEW-CIT$1 | NEW CITY | 1 | 5 | -87.637454 | 41.799669 | 447042.201722 | 4.627730e+06 | 15 | 35 | 4 | Conrail 51st Street Freight House, West 51st S... |
| 23 | ENG$3 | ENGLEWOOD | 3 | 3 | -87.641745 | 41.794253 | 446681.201571 | 4.627132e+06 | 9 | 35 | 4 | 638-640 West Garfield Boulevard, Chicago, IL 6069 |
| 24 | FUL-PAR$1 | FULLER PARK | 1 | 3 | -87.635980 | 41.799149 | 447164.218123 | 4.627672e+06 | 3 | 35 | 4 | Conrail 51st Street Freight House, West 51st S... |
| 25 | FUL-PAR$0 | FULLER PARK | 0 | 2 | -87.632427 | 41.805023 | 447464.218322 | 4.628322e+06 | 82 | 35 | 4 | 4914 South Wells Street, Chicago, IL 6069 |

Fig. 26 optimal location areas with address given for area center

Fig. optimal location areas visualized

One last step we can do here is to try find corresponding venues to open a new gym. For this purpose I use centers of already identified 26 clusters and look for buildings, coworking spaces, business centers, and rehab centers in 600m radius with help of Foursquare Places API. Obtained result is filtered of religious and some other venues. It counts 65 optimal venues within optimal locations to open a new gym.

| | id | lat | lng | formatted_address | cluster_id | cluster_group |
|---|---|---|---|---|---|---|
| 0 | 4f7c5dc6e4b032c26c42aad2 | 41.681278 | -87.617058 | 11634 South Prairie Avenue, Chicago, IL 60628 | WES-PUL$1 | 39 |
| 1 | 4f7c98a4e4b086fa1f31b394 | 41.689148 | -87.620651 | 108-112 East 113th Street, Chicago, IL 60628 | WES-PUL$1 | 39 |
| 2 | 4f8c2f21e4b0c71a74483d18 | 41.684486 | -87.617729 | 218-224 East Kensington Avenue, Chicago, IL 60628 | WES-PUL$1 | 39 |
| 3 | 4c21fc299390c9b649c0c9cd | 41.690792 | -87.604322 | Comcast, 11201-11203 South Ellis Avenue, Chica... | PUL$1 | 39 |
| 4 | 5197a4ee498e820998bda70d | 41.694714 | -87.606482 | The University of Chicago Press, 11030 South L... | PUL$1 | 39 |
| 5 | 4de177ef7d8b2547eaed0c5d | 41.689927 | -87.604198 | Raffin, 734-744 East 113th Street, Chicago, IL... | PUL$1 | 39 |
| 6 | 4c094055a1b32d7f7c5897f0 | 41.677610 | -87.620831 | 11912 South Michigan Avenue, Chicago, IL 60628 | WES-PUL$3 | 39 |
| 7 | 4e475e47fa76a07fde675d62 | 41.673782 | -87.616482 | 224-228 East 121st Place, Chicago, IL 60628 | WES-PUL$3 | 39 |
| 8 | 4c0940c66071a593fc95dd32 | 41.676058 | -87.620742 | 12006 South Michigan Avenue, Chicago, IL 60628 | WES-PUL$3 | 39 |
| 9 | 4bd6d905cfa7b7131d2028da | 41.680777 | -87.609952 | Chicago, IL 60627 | RIV$0 | 39 |

Fig. First ten optimal venues within optimal locations to open a gym

## Results

- 24K venues of six different categories were analyzed
- Chicago area was split on 8.5K 300x300m areas for detailed analysis
- Impact factors are analyzed: gym venue location is mostly influenced by nearby transport-related venues
- Answers to address business problem are:
  - 26 optimal locations were identified to open a gym in 600m radius:

| *community* | *optimal location address (600m radius)* |
|---|---|
| ASHBURN | Coc Water South District Headquarters, 7501-7521 South Western Avenue, Chicago, IL 60643 |
| AUBURN GRESHAM | 7711 South Wolcott Avenue, Chicago, IL 60620 |
| AUSTIN | 4900 West Bloomingdale Avenue, Chicago, IL 60639 |
| BELMONT CRAGIN | 1923 North La Crosse Avenue, Chicago, IL 60639 |
| CHICAGO LAWN | 7410-7426 South Western Avenue, Chicago, IL 60643 |
| ENGLEWOOD | 638-640 West Garfield Boulevard, Chicago, IL 6069 |
| FULLER PARK | 4914 South Wells Street, Chicago, IL 6069 |
| | Conrail 51st Street Freight House, West 51st Street, Chicago, IL 60632 |
| HERMOSA | 4218 West Armitage Avenue, Chicago, IL 60639 |
| | 4700 West Grand Avenue, Chicago, IL 60651 |
| HUMBOLDT PARK | 1513 North Keeler Avenue, Chicago, IL 60651 |
| MORGAN PARK | 11023 South Sangamon Street, Chicago, IL 60643 |
| | 800-900 West 115th Street, Chicago, IL 60643 |
| NEW CITY | Conrail 51st Street Freight House, West 51st Street, Chicago, IL 60632 |
| PULLMAN | Pullman National Monument Visitors Center, 11139-11141 South Cottage Grove Avenue, Chicago, IL 60628 |
| RIVERDALE | 470 East Kensington Avenue, Chicago, IL 60628 |
| | 690 East Kensington Avenue, Chicago, IL 60628 |
| ROSELAND | 644 West 111th Street, Chicago, IL 60643 |
| | St. John M.B. Church, 205-221 East 115th Street, Chicago, IL 60628 |
| SOUTH DEERING | South Doty Avenue, Chicago, IL 60633 |
| WEST ENGLEWOOD | 1815 West 74th Street, Chicago, IL 60636 |
| | 2148 West 75th Place, Chicago, IL 60620 |
| WEST PULLMAN | 120 East Kensington Avenue, Chicago, IL 60628 |
| | 12101 South Emerald Avenue, Chicago, IL 60628 |
| | 12105 South Edbrooke Avenue, Chicago, IL 60628 |
| | 727 West 116th Street, Chicago, IL 60628 |

  - 65 optimal venues were identified to open a gym in there:

| *community* | *optimal venue address* |
|---|---|
| *AUBURN GRESHAM* | 7511 South Damen Avenue, Chicago, IL 60620 |
| | 7906-7910 South Hermitage Avenue, Chicago, IL 60620 |
| *AUSTIN* | 5012 West Concord Place, Chicago, IL 60639 |
| | BMO Harris Bank, 4959 West North Avenue, Chicago, IL 60639 |
| *BELMONT CRAGIN* | 4545-4555 West Armitage Avenue, Chicago, IL 60639 |
| | 4712 West Armitage Avenue, Chicago, IL 60639 |
| | 4901-4915 West Armitage Avenue, Chicago, IL 60639 |
| | 5000-5008 West Bloomingdale Avenue, Chicago, IL 60639 |
| | 5008-5010 West Dickens Avenue, Chicago, IL 60639 |
| *FULLER PARK* | 4749 South Wentworth Avenue, Chicago, IL 60621 |

| | |
|---|---|
| | Dan Ryan Expressway, Chicago, IL 60616 |
| | Dan Ryan Expressway, Chicago, IL 6069 |
| | Department of Fleet Management - Bureau of Police Motor Maintenance, 5219 South Wentworth Avenue, Chicago, IL 60621 |
| | Harold Washington Professional Building, 5341 South Wentworth Avenue, Chicago, IL 60609 |
| *HERMOSA* | 1755 North Karlov Avenue, Chicago, IL 60639 |
| | 1924 North Pulaski Road, Chicago, IL 60639 |
| | 2045 North Kenneth Avenue, Chicago, IL 60639 |
| | 4149 West Armitage Avenue, Chicago, IL 60639 |
| | 4335-4359 West Armitage Avenue, Chicago, IL 60639 |
| | 4335-4359 West Armitage Avenue, Chicago, IL 60639 |
| | 4447-4459 West Cortland Street, Chicago, IL 60639 |
| | 4556 West Grand Avenue, Chicago, IL 60651 |
| | Beat 2525, North Hamlin Avenue, Chicago, IL 60618 |
| | Beat 2525, West McLean Avenue, Chicago, IL 60647 |
| | Beat 2525, West McLean Avenue, Chicago, IL 60647 |
| *HUMBOLDT PARK* | 1250-1256 North Kildare Avenue, Chicago, IL 60641 |
| | 1409 North Pulaski Road, Chicago, IL 60651 |
| | 1620 North Karlov Avenue, Chicago, IL 60639 |
| | 1621-1657 North Kostner Avenue, Chicago, IL 60641 |
| | 1753 North Tripp Avenue, Chicago, IL 60639 |
| | 3925-3929 West Grand Avenue, Chicago, IL 60651 |
| | 3950-3952 West Grand Avenue, Chicago, IL 60651 |
| | 4059 West North Avenue, Chicago, IL 60302 |
| | 4113-4115 West Kamerling Avenue, Chicago, IL 60651 |
| | 4123-4125 West North Avenue, Chicago, IL 60302 |
| | 4216 West Potomac Avenue, Chicago, IL 60651 |
| | 4259 West Kamerling Avenue, Chicago, IL 60651 |
| | 4301 West Grand Avenue, Chicago, IL 60651 |
| | North & Pulaski Apartments, 3949 West North Avenue, Chicago, IL 60647 |
| *MORGAN PARK* | 11355-11359 South Halsted Street, Chicago, IL 60827 |
| | 11435 South Halsted Street, Chicago, IL 60827 |
| | Mobil Mart, 11501-11507 South Halsted Street, Chicago, IL 60827 |
| *NEW CITY* | Parkman School, 245 West 51st Street, Chicago, IL 60632 |
| *PULLMAN* | Comcast, 11201-11203 South Ellis Avenue, Chicago, IL 60628 |
| | Raffin, 734-744 East 113th Street, Chicago, IL 60628 |
| | The University of Chicago Press, 11030 South Langley Avenue, Chicago, IL 60628 |
| *RIVERDALE* | Chicago, IL 60627 |
| | Chicago, IL 60627 |
| *ROSELAND* | 10844-10848 South Halsted Street, Chicago, IL 60628 |
| | 11130-11142 South Halsted Street, Chicago, IL 60628 |
| | 11300-11306 South Halsted Street, Chicago, IL 60628 |
| *WEST ENGLEWOOD* | 1919 West 74th Street, Chicago, IL 60636 |
| | 7206 South Seeley Avenue, Chicago, IL 60636 |

| | |
|---|---|
| | 7400 South Damen Avenue, Chicago, IL 60620 |
| *WEST PULLMAN* | 108-112 East 113th Street, Chicago, IL 60628 |
| | 11634 South Prairie Avenue, Chicago, IL 60628 |
| | 11839 South Lowe Avenue, Chicago, IL 60628 |
| | 11912 South Michigan Avenue, Chicago, IL 60628 |
| | 12006 South Michigan Avenue, Chicago, IL 60628 |
| | 12143 South Normal Avenue, Chicago, IL 60628 |
| | 218-224 East Kensington Avenue, Chicago, IL 60628 |
| | 224-228 East 121st Place, Chicago, IL 60628 |
| | 829-837 West 119th Street, Chicago, IL 60827-6427 |
| | Chase, 11721 South Halsted Street, Chicago, IL 60628 |
| | West Pullman Elementary School, South Parnell Avenue, Chicago, IL 60628 |

## Discussion

Although business problem was addressed, following areas can be improved

- area center calculation
- consequent area coverage by the grid
- grid to exclude irrelevant areas where a venue cannot be located
- reduce number of venues in the "areas, venues" cartesian product, by filtering out all venues located further than 5km either by x or y coordinate
- analyze which venue categories are currently occupied by competitors and use those for finding optimal location
- make transport category more granular, probably by reducing it to the following venues: private and public transport; embarkation, dis-embarkation train stations/hubs.
- apply PCA, make use of demographics data
- notebook code refactoring

## Conclusion

In this study I identified optimal location for opening a gym in Chicago, IL. I analyzed impact factors on having a gym venue and their correlation with demographics data. Optimal locations can be very useful for anyone real estate agencies or gym/fitness business owners, allowing them to save a lot of time and choose from optimal locations prepared with the power of machine learning. Also, existing approach is highly customizable and can be applied for different venue type or even different location with minimal changes.