Gym 💪 Venue Optimal Location

Applied Data Science Capstone by IBM/Coursera
Capstone Project
Oleksandr Bogach
May 2020

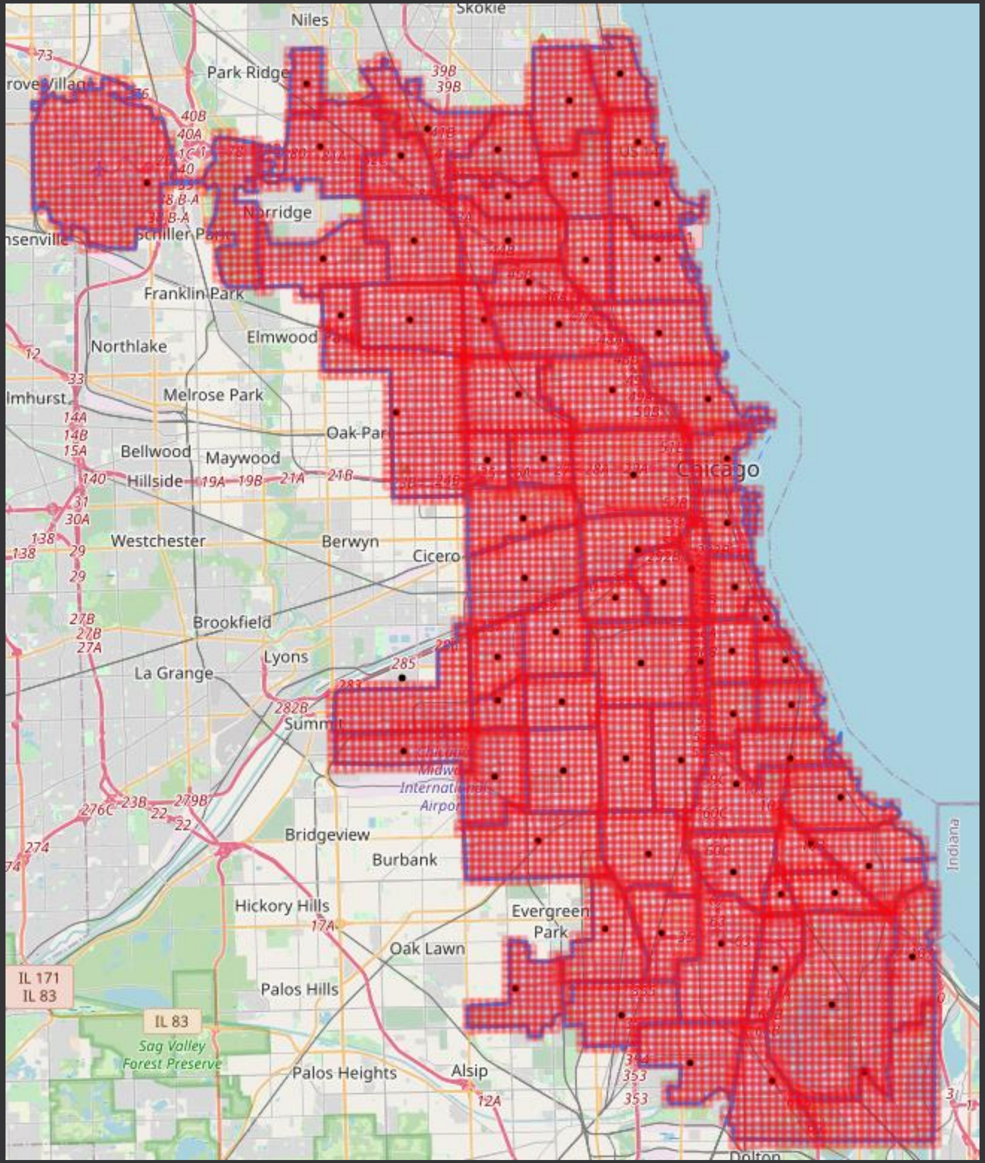
Finding an optimal location for opening a gym in Chicago, IL

- Usually finding an optimal location across whole city is time consuming operation as it heavily relies on detailed analysis of neighborhood areas and should take into consideration a lot of factors (such as adjacent venues, competitors, transport, etc.)
- Useful for gym/fitness business owners, real estate agencies
- To avoid routine work automated approach should output optimal addresses and suggests optimal venues to consider

Data acquisition and processing

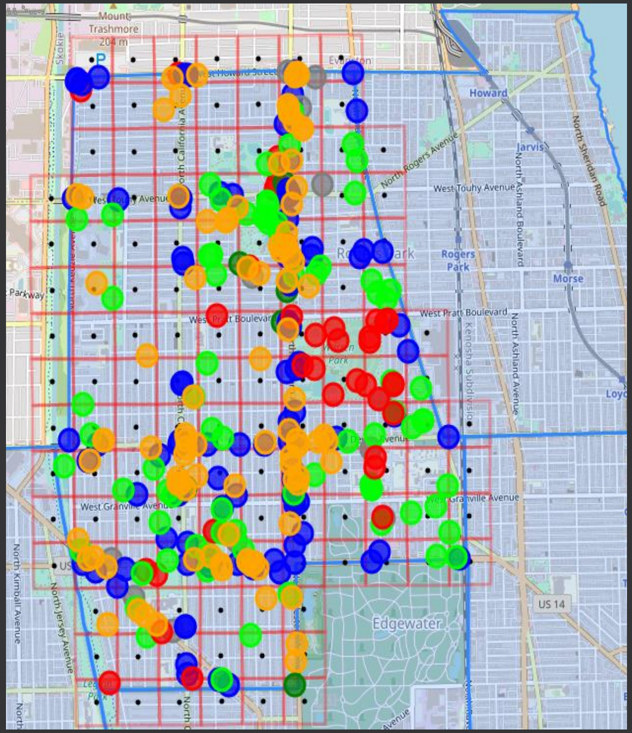
- Four data sources used: Foursquare Places API, Chicago GEO json, Chicago Demographics data, OpenCage API
- 24K+ venues of six categories retrieved from Foursquare Places API for all 77 Chicago communities
- Cleaned, consolidated, and cross-checked data used for analyzing 8.5K+ small area candidates for optimal location

Optimal location area candidates

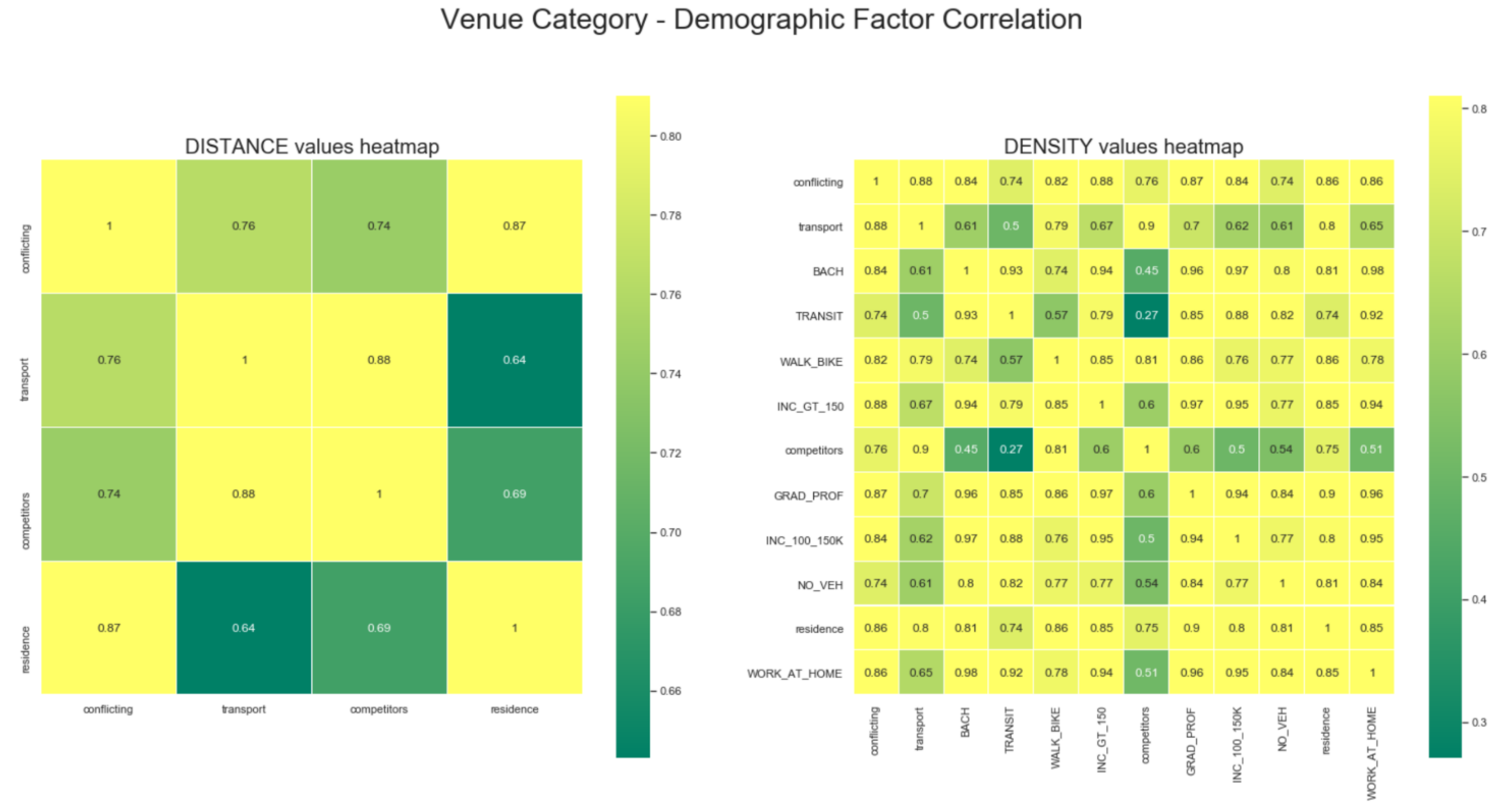


- There are 8547 optimal area candidates (300x300m)

- Six different venue categories were considered for each area candidate: transport, public places, competitors, conflicting, supporting, residence



Venue Category – Demographic Factor Correlation Heatmap



Venue Category - Demographic Factor Correlation

DISTANCE values heatmap

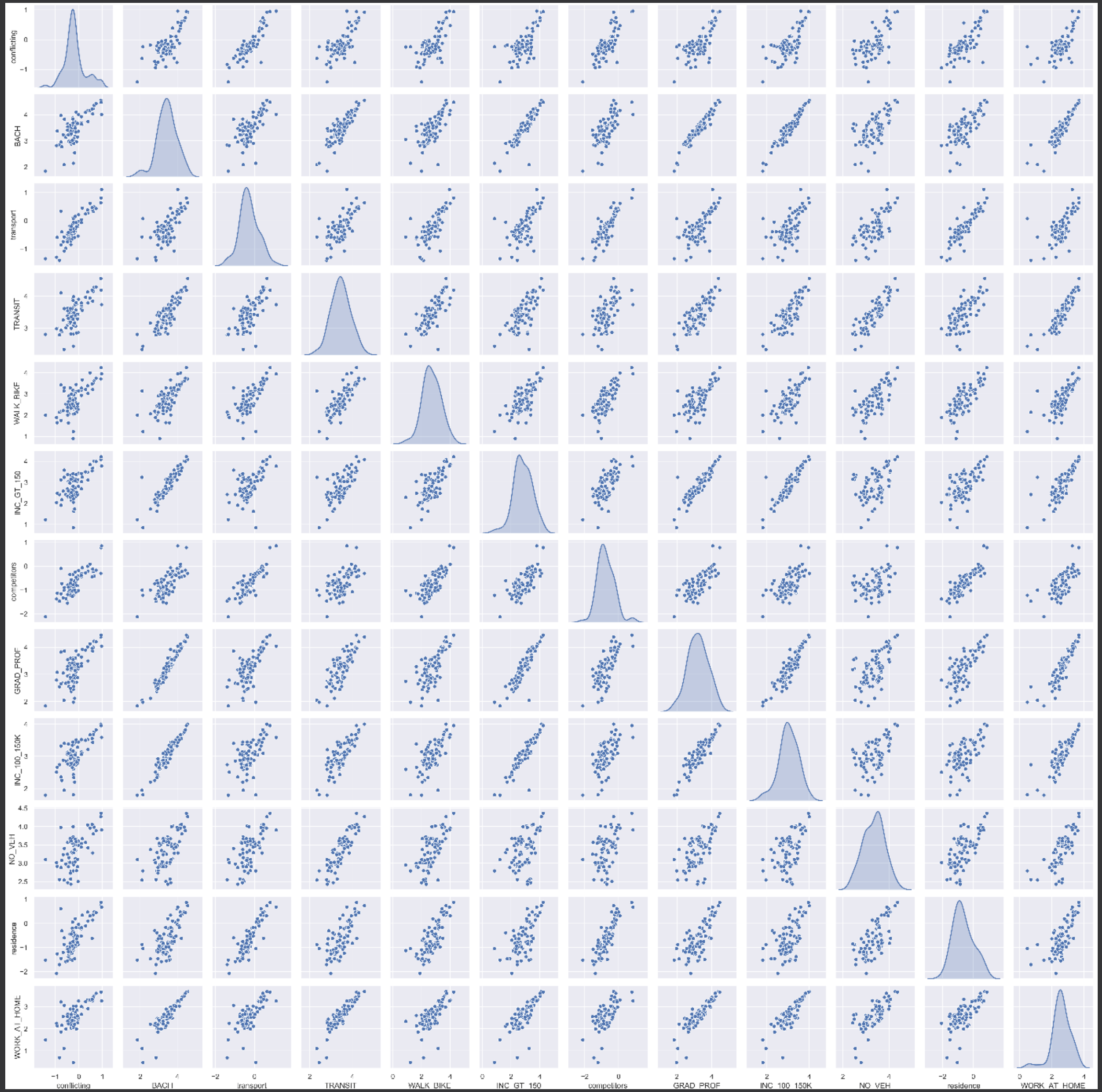
	conflicting	transport	competitors	residence
conflicting	1	0.76	0.74	0.87
transport	0.76	1	0.88	0.64
competitors	0.74	0.88	1	0.69
residence	0.87	0.64	0.69	1

DENSITY values heatmap

	conflicting	transport	BACH	TRANSIT	WALK_BIKE	INC_GT_150	competitors	GRAD_PROF	INC_100_150K	NO_VEH	residence	WORK_AT_HOME
conflicting	1	0.88	0.84	0.74	0.82	0.88	0.76	0.87	0.84	0.74	0.86	0.86
transport	0.88	1	0.61	0.5	0.79	0.67	0.9	0.7	0.62	0.61	0.8	0.65
BACH	0.84	0.61	1	0.93	0.74	0.94	0.45	0.96	0.97	0.8	0.81	0.98
TRANSIT	0.74	0.5	0.93	1	0.57	0.79	0.27	0.85	0.88	0.82	0.74	0.92
WALK_BIKE	0.82	0.79	0.74	0.57	1	0.85	0.81	0.86	0.76	0.77	0.86	0.78
INC_GT_150	0.88	0.67	0.94	0.79	0.85	1	0.6	0.97	0.95	0.77	0.85	0.94
competitors	0.76	0.9	0.45	0.27	0.81	0.6	1	0.6	0.5	0.54	0.75	0.51
GRAD_PROF	0.87	0.7	0.96	0.85	0.86	0.97	0.6	1	0.94	0.84	0.9	0.96
INC_100_150K	0.84	0.62	0.97	0.88	0.76	0.95	0.5	0.94	1	0.77	0.8	0.95
NO_VEH	0.74	0.61	0.8	0.82	0.77	0.77	0.54	0.84	0.77	1	0.81	0.84
residence	0.86	0.8	0.81	0.74	0.86	0.85	0.75	0.9	0.8	0.81	1	0.85
WORK_AT_HOME	0.86	0.65	0.98	0.92	0.78	0.94	0.51	0.96	0.95	0.84	0.85	1

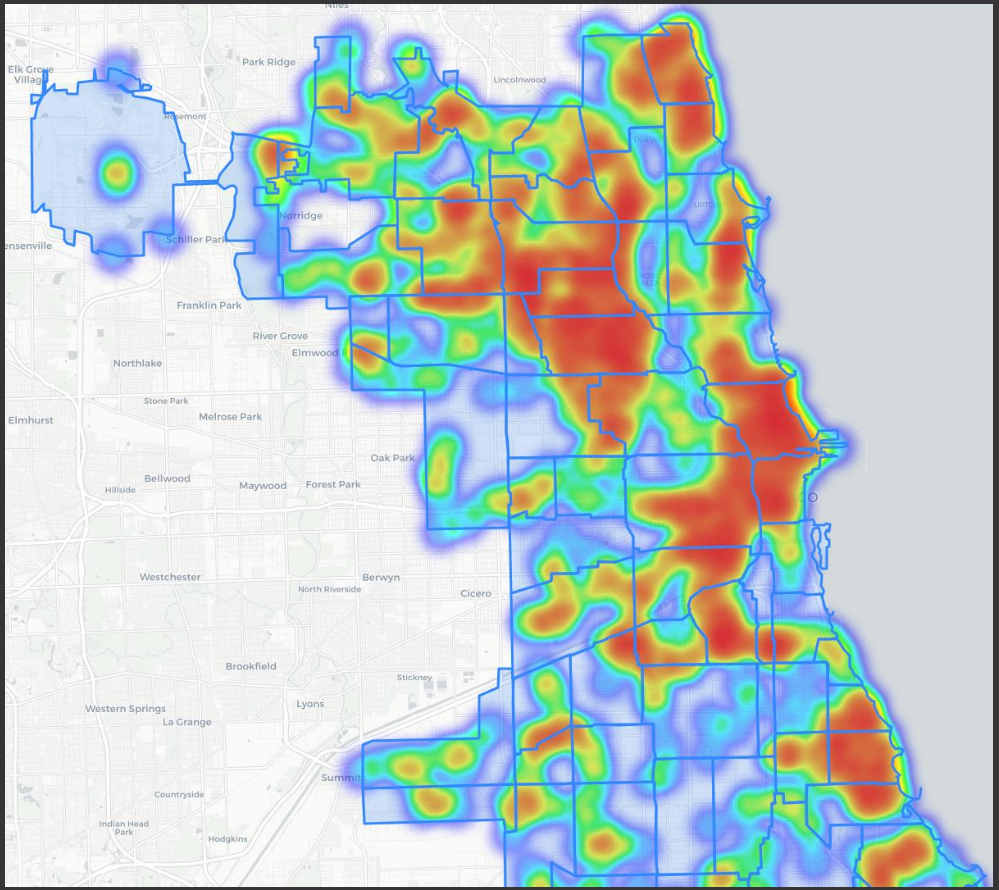
Venues density demonstrates good correlation with demographic data

Exploratory Analysis Insights

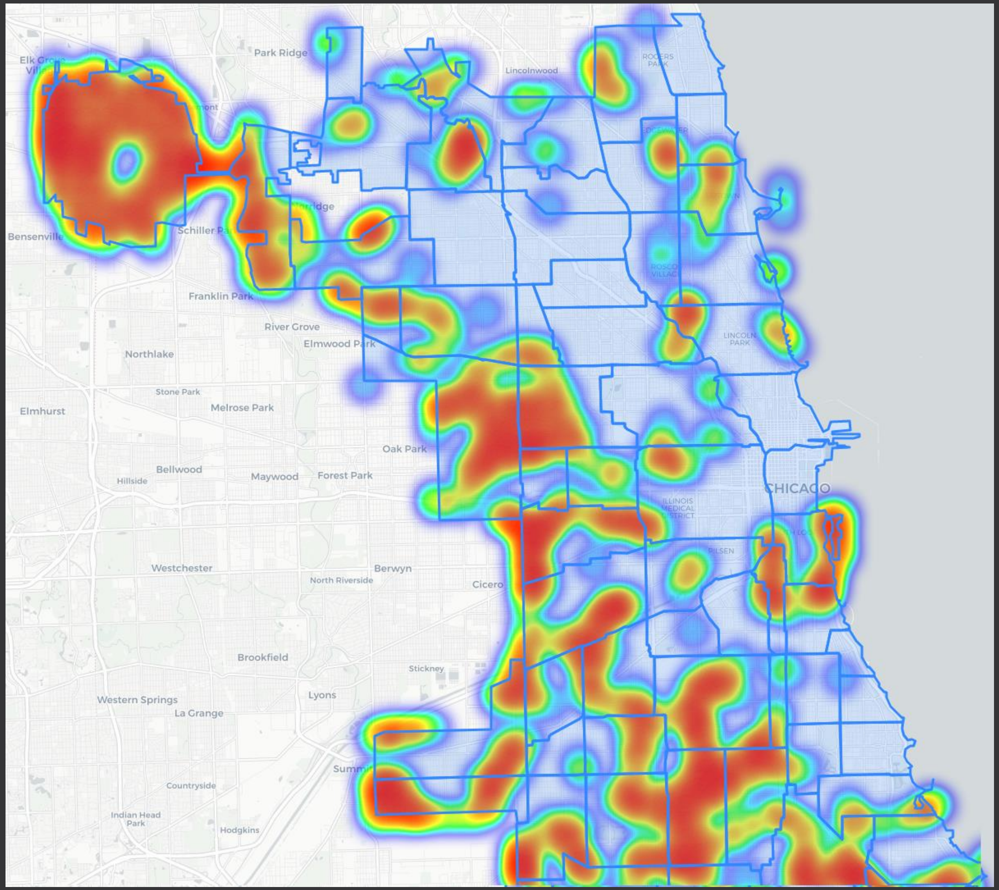


- *competitors* factor:
  - pair *competitors* - *transport* has strong correlation as it was initially thought. Venues of transport type (as defined at Venue category) act as a primary factor of proper location for a gym
  - areas where people travel to work either by walk or bike (*WALK_BIKE*) is also 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

Reduce candidate areas



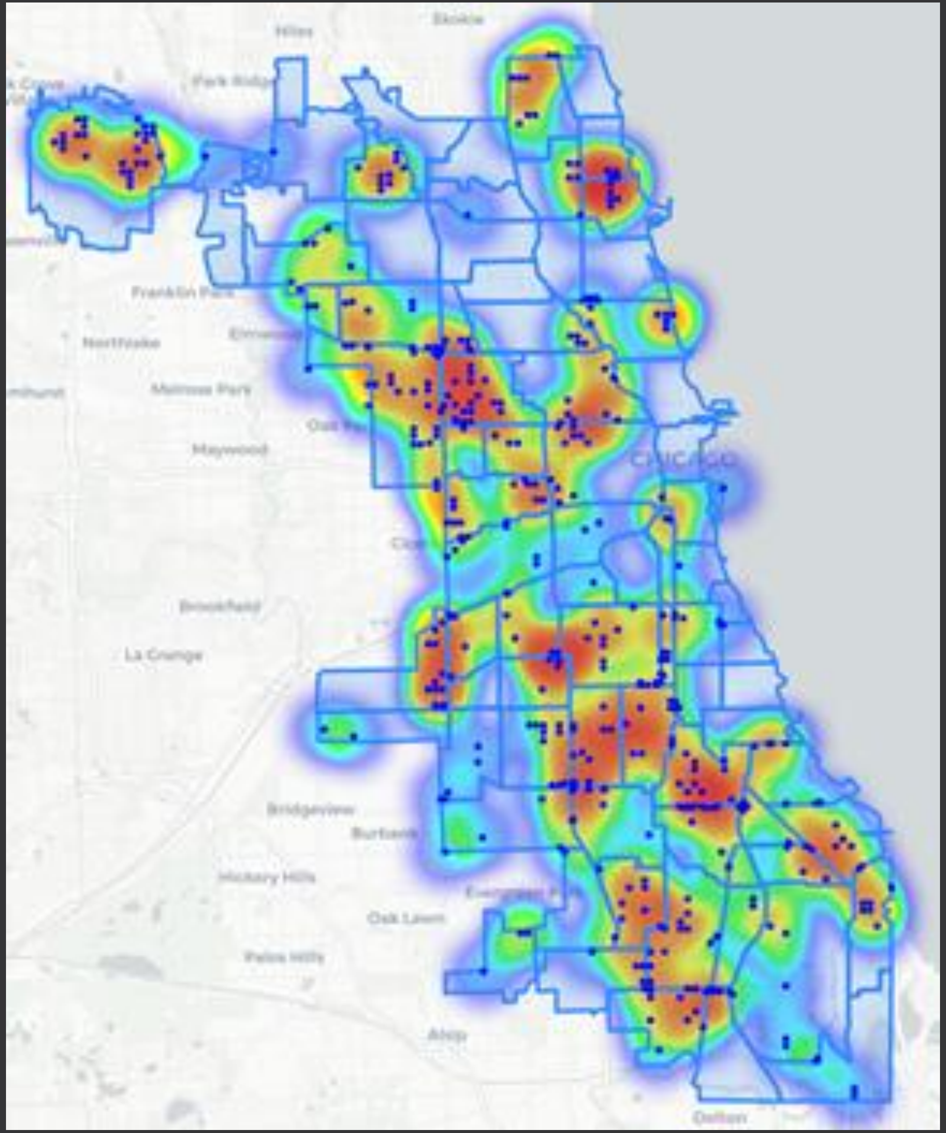
Competitors heatmap (1342 areas)



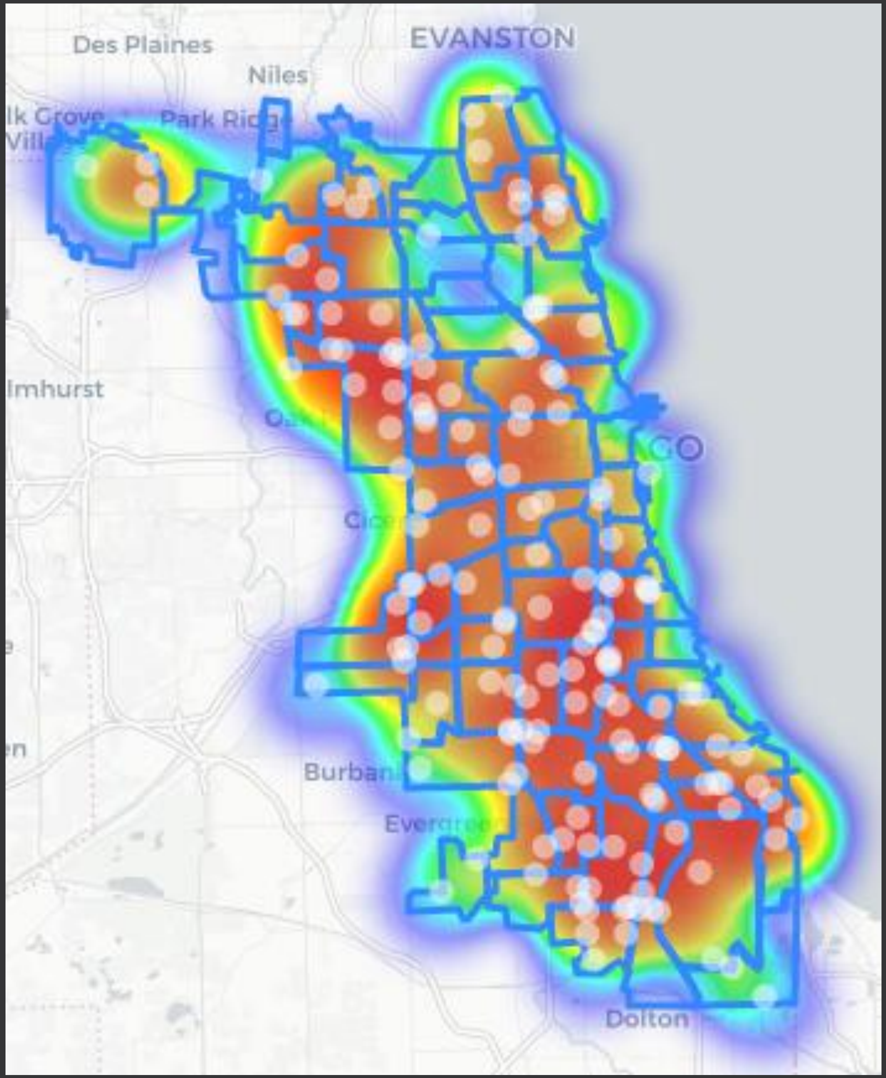
Optimal location candidate areas without competitors and with distance to the closest competitor venue greater than average for current community

Reduce candidate areas

Even more reduce of candidate areas by considering only areas (blue dots denote area center) with at least one venue of transport category. Thus number was reduced from 7K to 429.



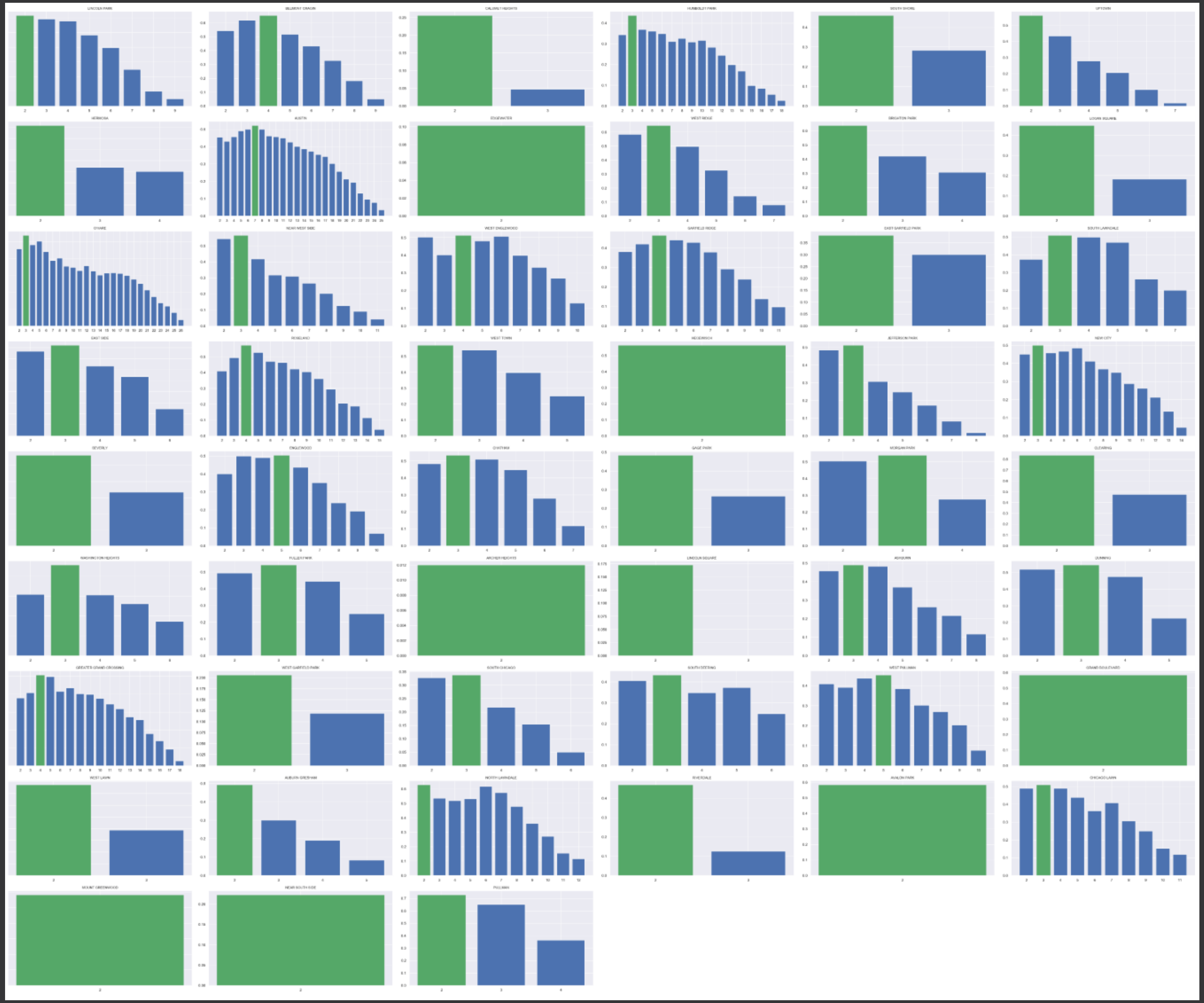
Clustering



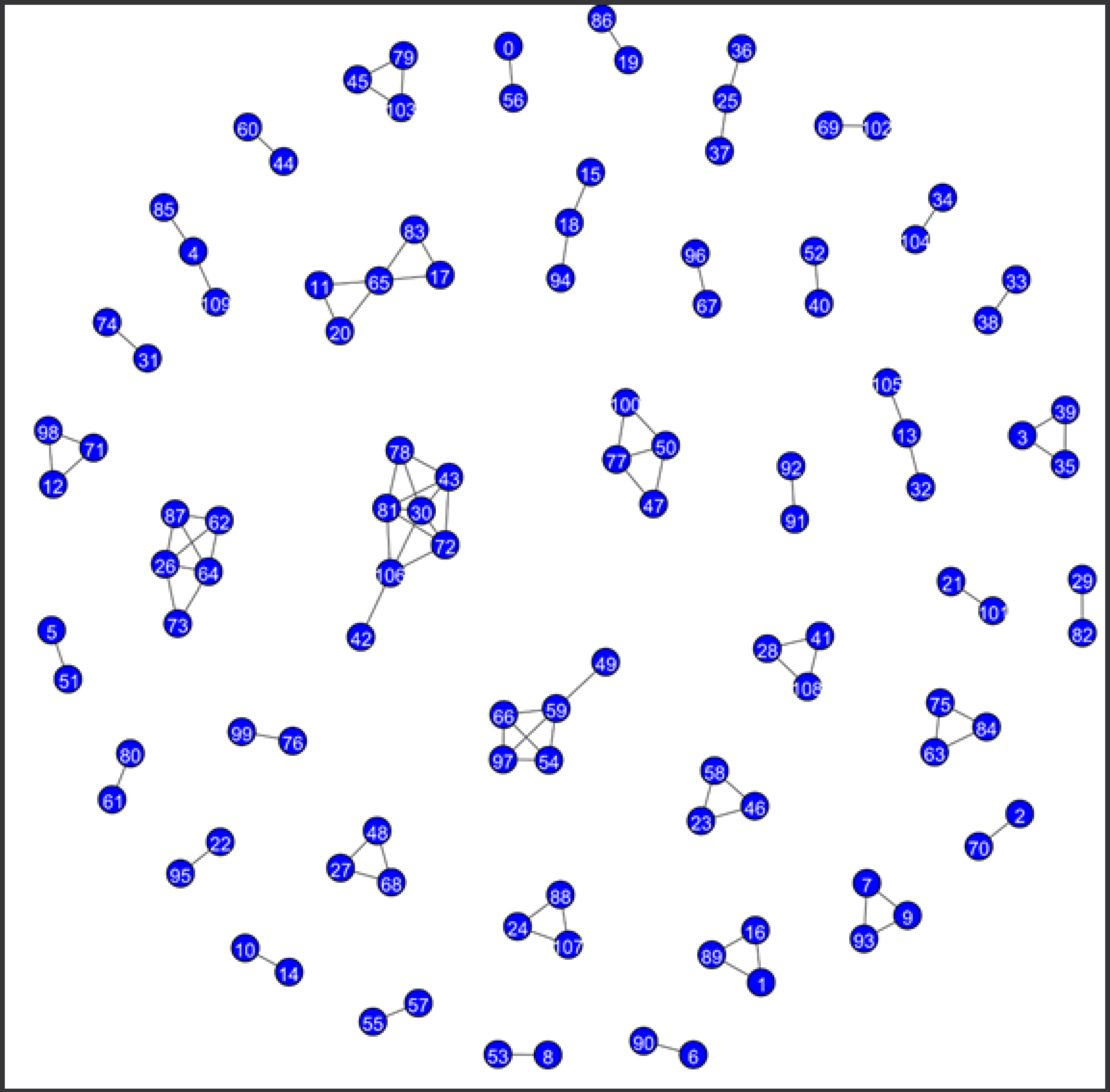
- Clustering is done by community where previously identified 429 area candidates are located. Initialization method is \`k-means++\`. Results of clustering are visualized over the heatmap with candidate areas.
- Special handling (clustering imitation) was done for cases when a community contained either one or two area candidates.
- 158 clusters replaced 429 area candidates

Silhouette score of Clustering

Finding optimal clusters number is based on silhouette score and illustrated below. Green bar corresponds to number of clusters with max Silhouette score.

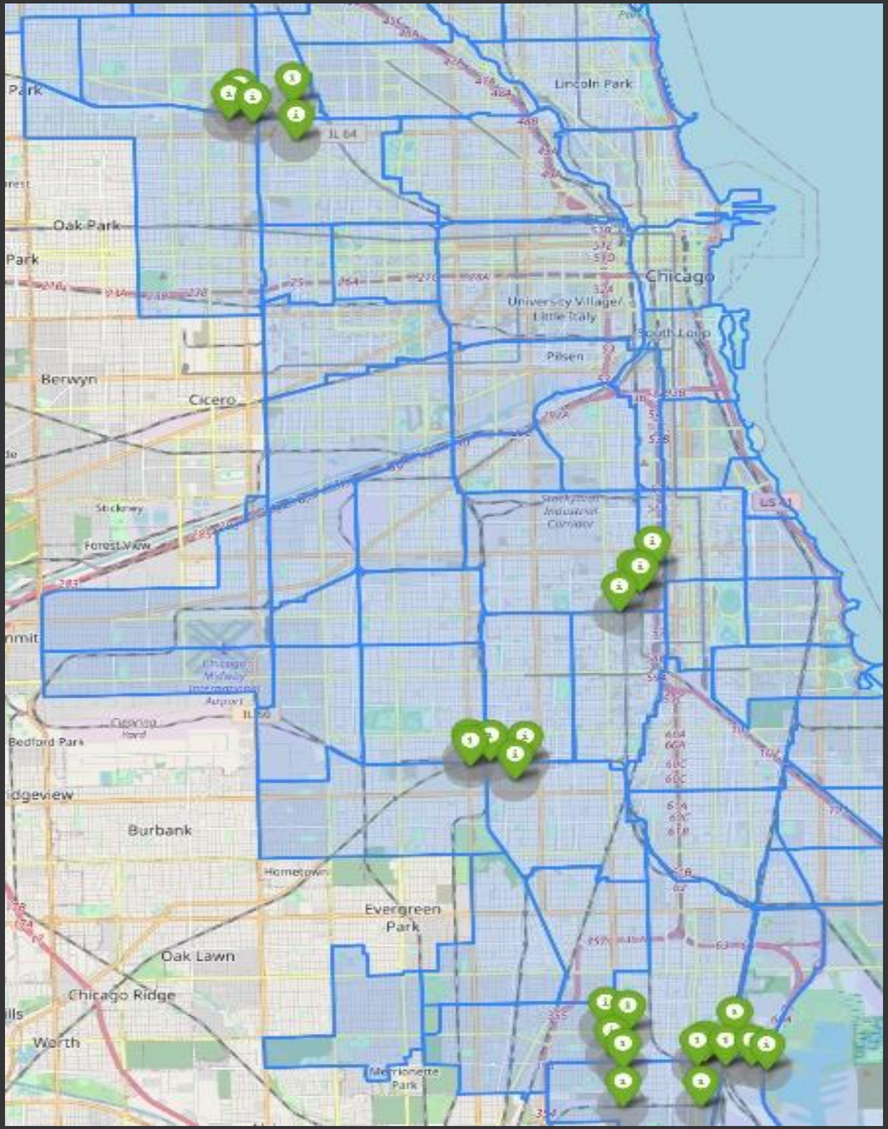


Cluster of clusters graph



- Cluster of clusters is built from clusters with distance between their centers less than 1200m.
- Bigger group (longest path) is more beneficial for an optimal location from area coverage perspective

Top 5 cluster groups (26 clusters)



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

Conclusion and future directions

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.

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
- code refactoring