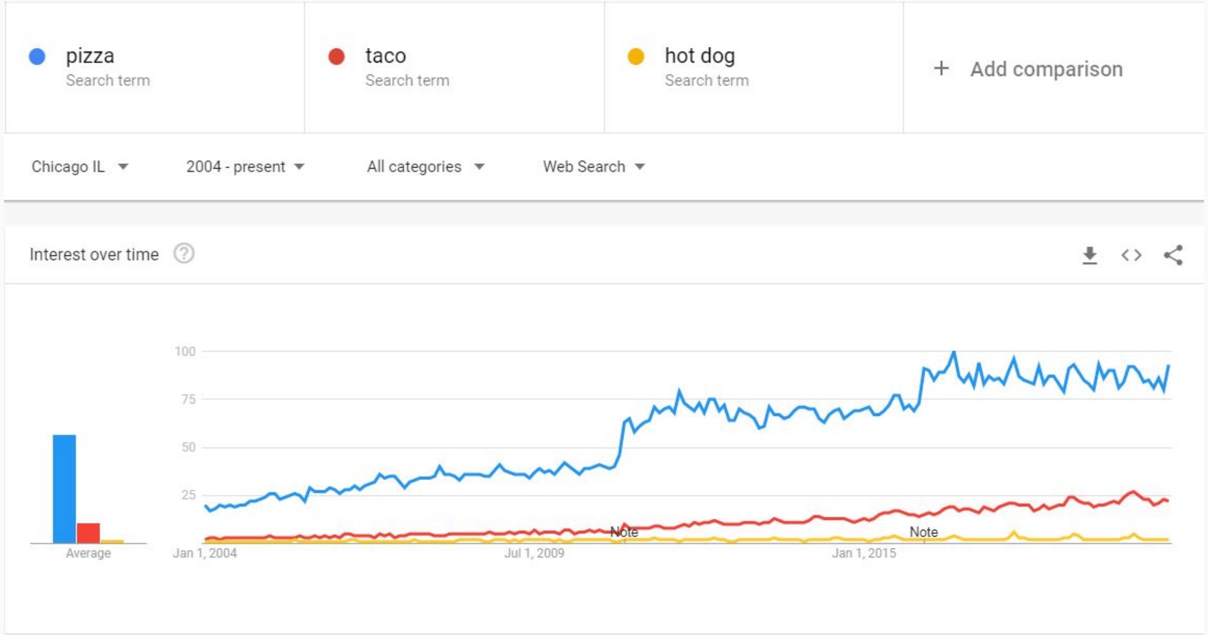
IBM Data Science Capstone Project

Using Foursquare Data to Determine Best Neighborhoods in Chicago to Open New Pizza Place

Background

Chicago is the #5 city in the USA for search popularity of the term "pizza".

Pizza searches far outweigh other popular dishes in the city.



● pizza
Search term

● taco
Search term

● hot dog
Search term

Chicago IL ▾ 2004 - present ▾ All categories ▾ Web Search ▾

Interest over time

Are There Neighborhoods in Chicago That Need Additional Pizza Places?

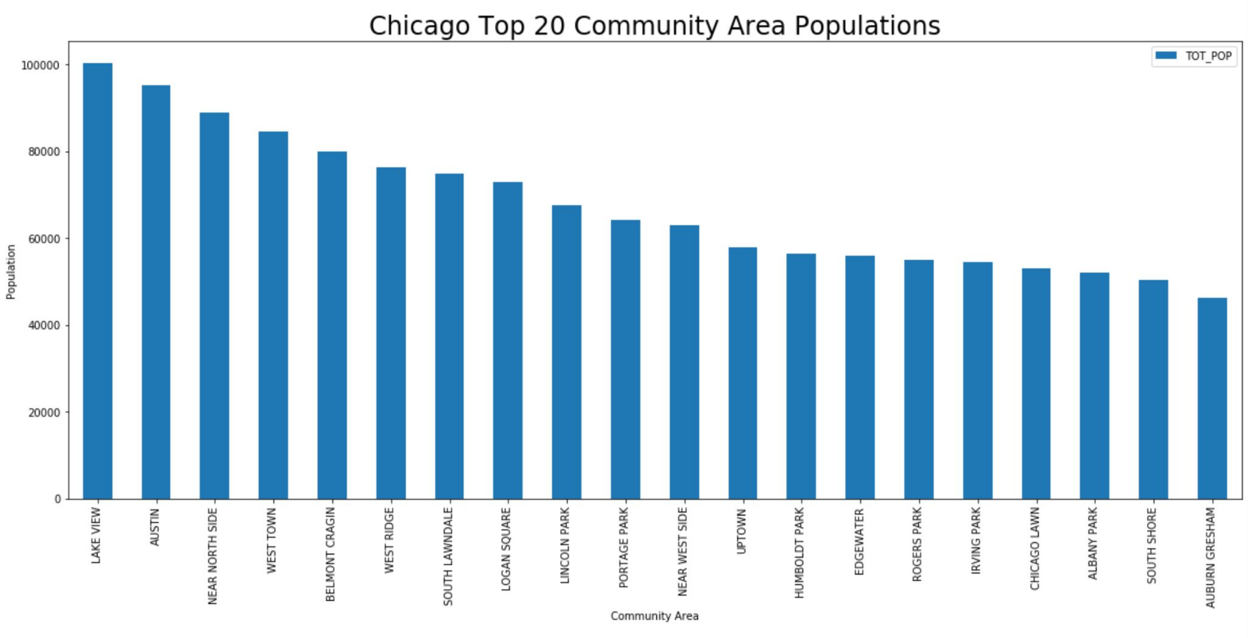
Data Sources

1. Chicago Community Area Demographic Data
    a. https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data
2. Chicago Neighborhoods Geojson Data
    a. https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9
3. Foursquare API

230 demographic features across all 77 designated community areas in the city of Chicago

Examples include

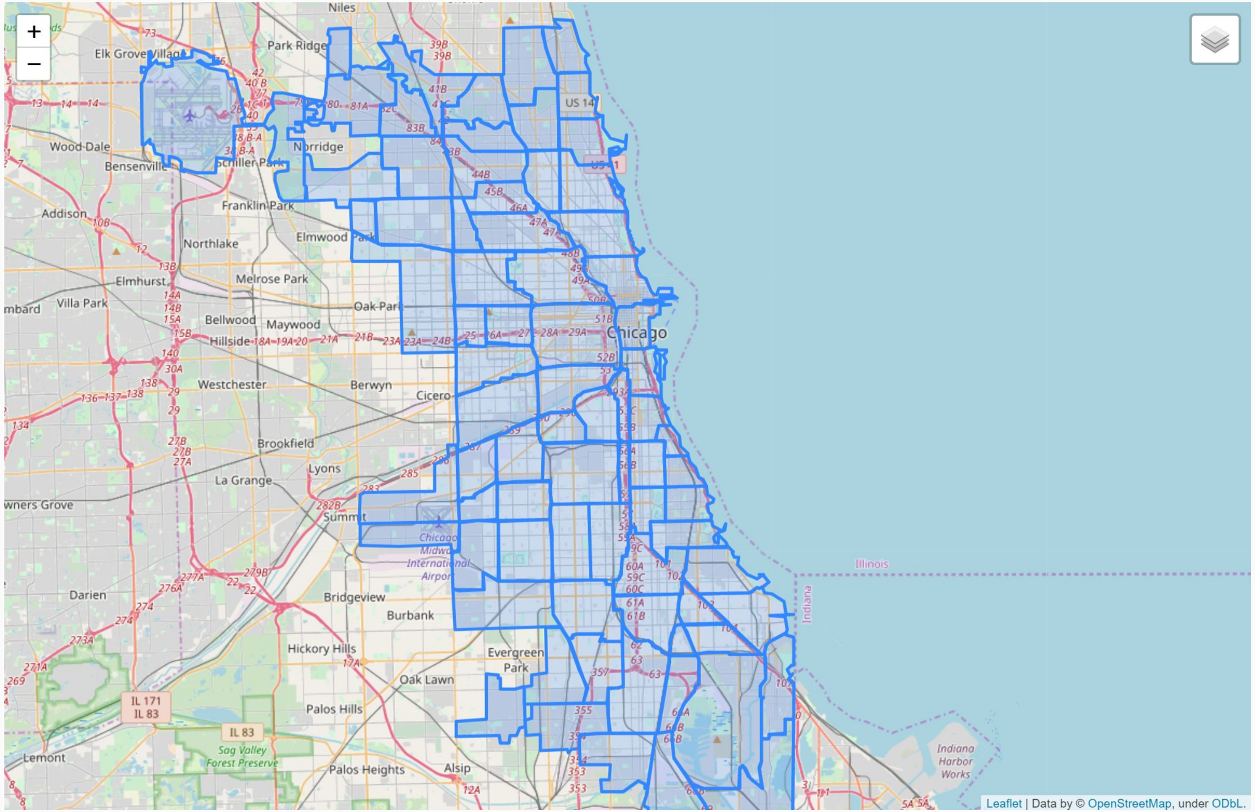
- Total Population
- Median Income
- Labor Force Participation Rate



Chicago Demographic Data

Neighborhood boundaries provided by the City of Chicago on its data portal

Latitude/Longitude boundary data for all 77 designated community areas in the city



Leaflet | Data by © OpenStreetMap, under ODbL.

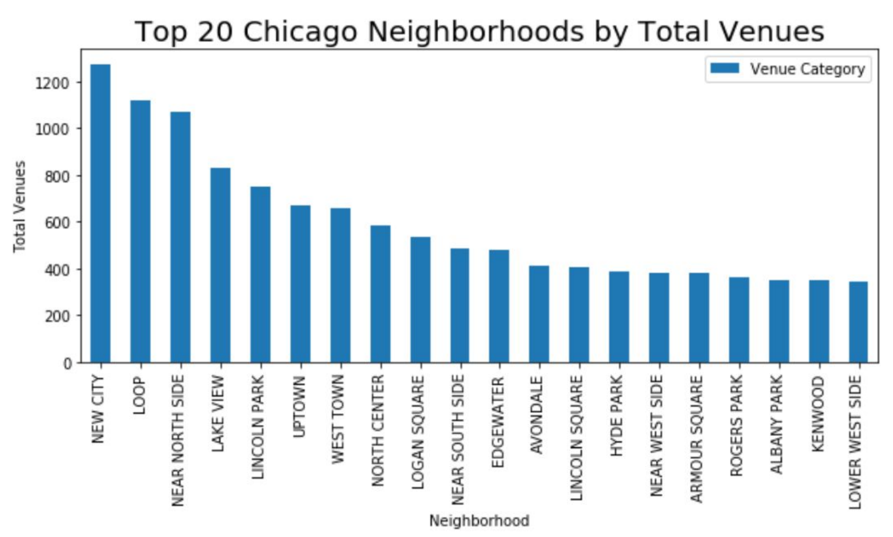
Chicago Geojson Data

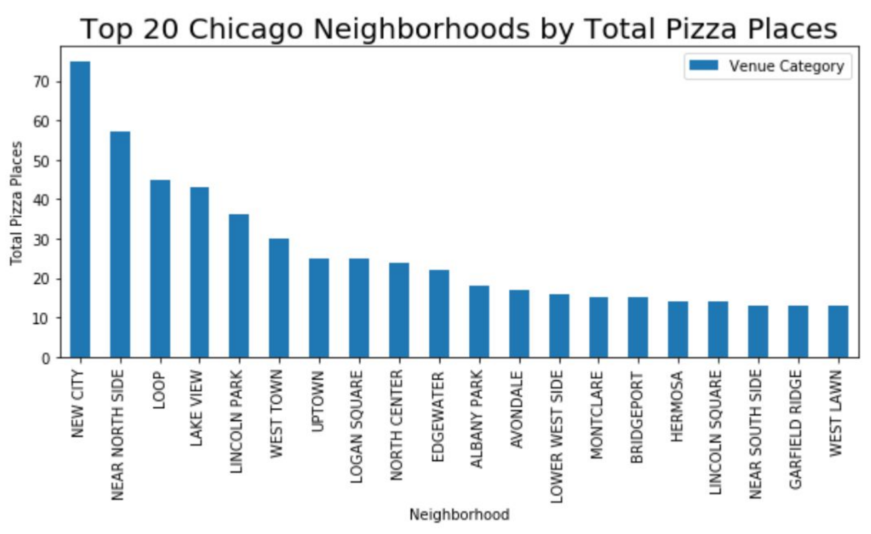
Details for all venues were collected for a square mile radius of all 77 community areas from the other data sources

- 23,933 total venues collected
- 484 unique venue categories
- 880 total pizza places

Data was transformed using one hot encoding, and then grouped by neighborhood using mean values of each venue category

Resulting data was 77 neighborhood rows with 484 columns

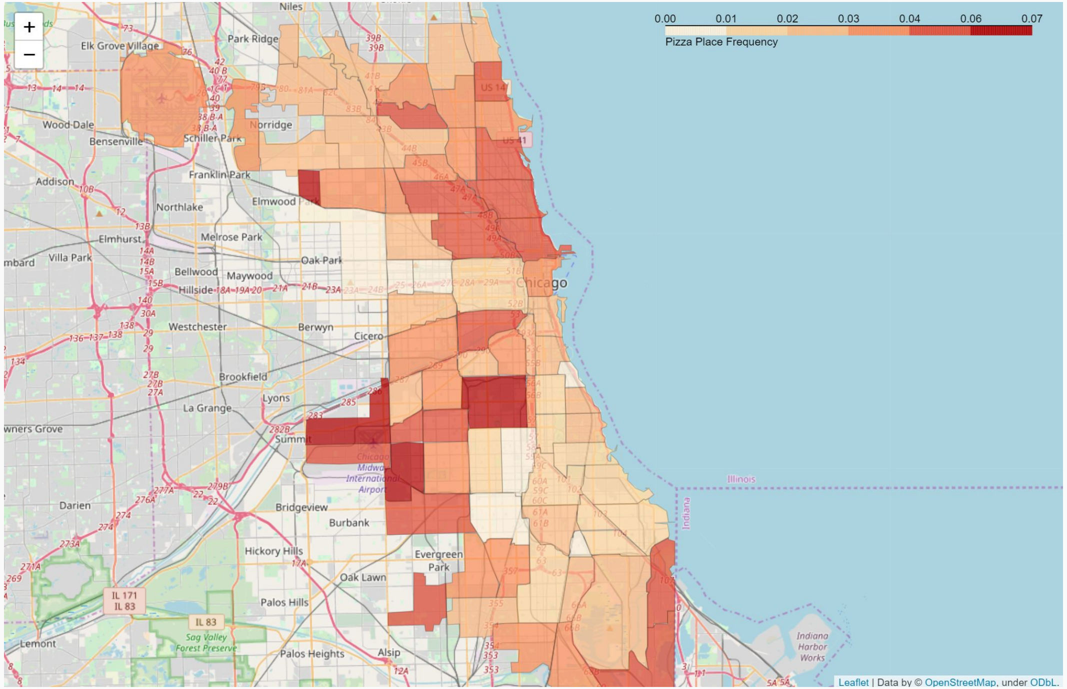




Foursquare Data

Target Variable

Percent of venues in a neighborhood that are Pizza Places



Neighborhood	% Venues Pizza Place
Garfield Ridge	6.7%
West Lawn	6.3%
Montclare	6.2%
New City	5.9%
Ashburn	5.5%

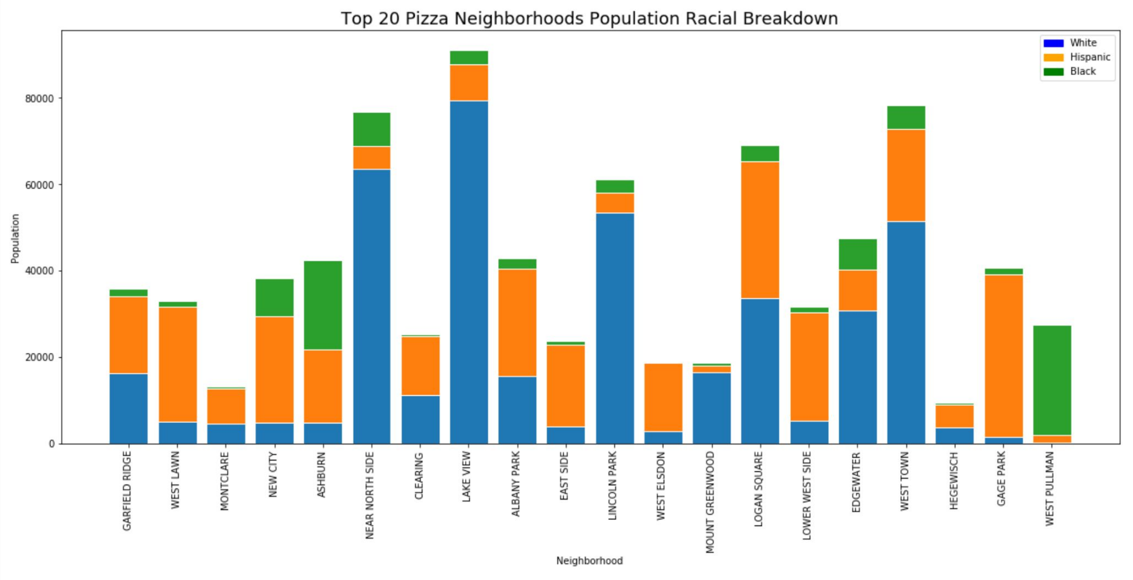
Feature Set

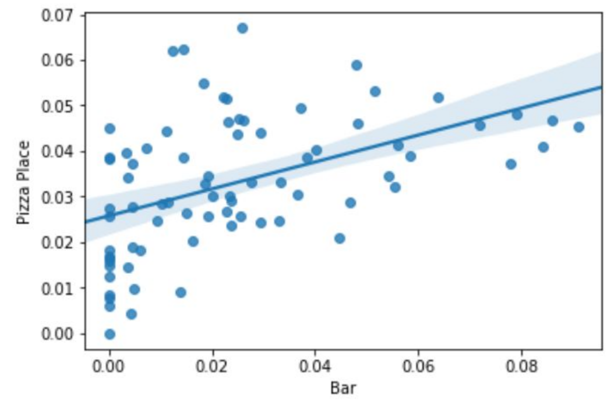
Originally there were 230 demographic and 484 Foursquare venue category variables.

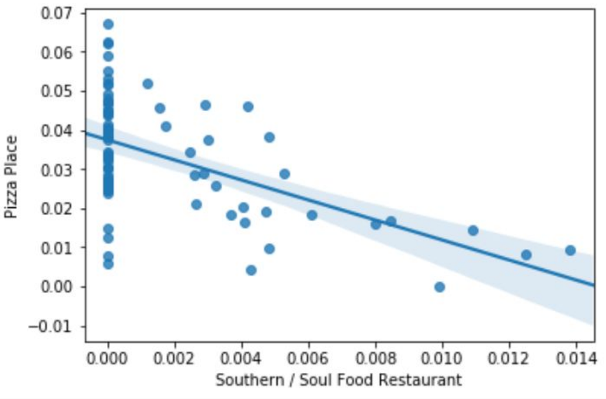
Used univariate regression testing to find top 5% of variables

Evaluated those variables for correlation and eliminated redundant features

Final set: 12 demographic features, 15 venue category features

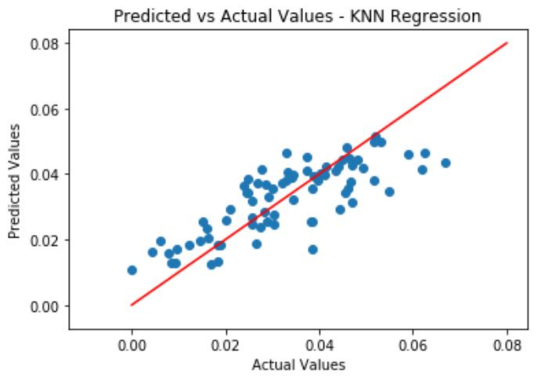


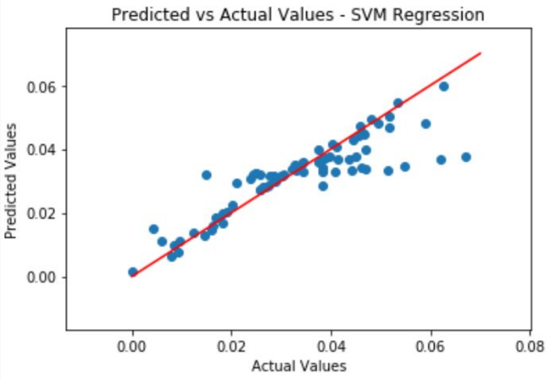


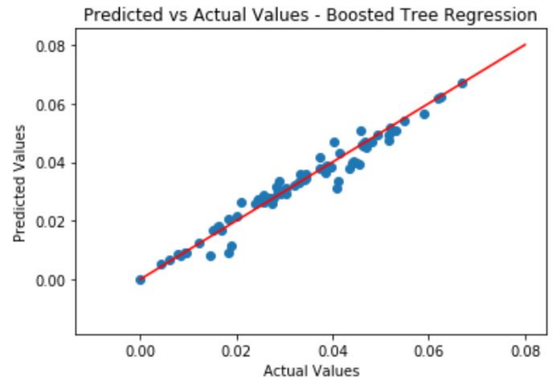


Regression Modeling Results

Method	Train Score	Test Score	Comments
KNN Regression	0.62	0.41	Best fit, low bias in results
SVM Regression	0.76	0.39	Excellent fit on low values, but unable to predict higher values
Boosted Tree Regression	0.97	0.37	Heavily overfit to train data
Ridge Regression	0.56	0.37	Light regularization improved performance compared with OLS
Multiple Linear Regression	0.64	0.24	Poor performance overall

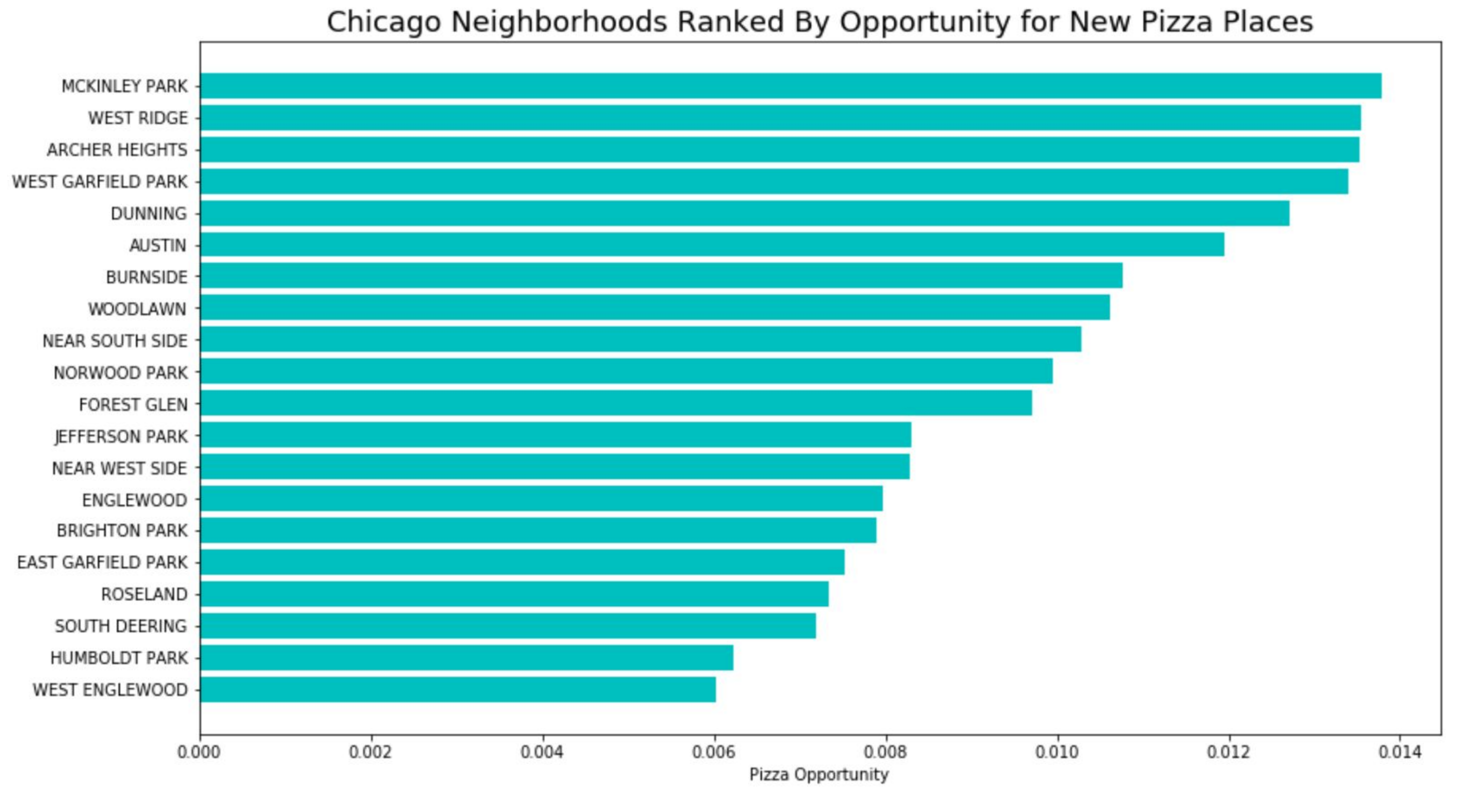






Using KNN Regression to Find Neighborhoods with Highest Pizza Place “Opportunity”

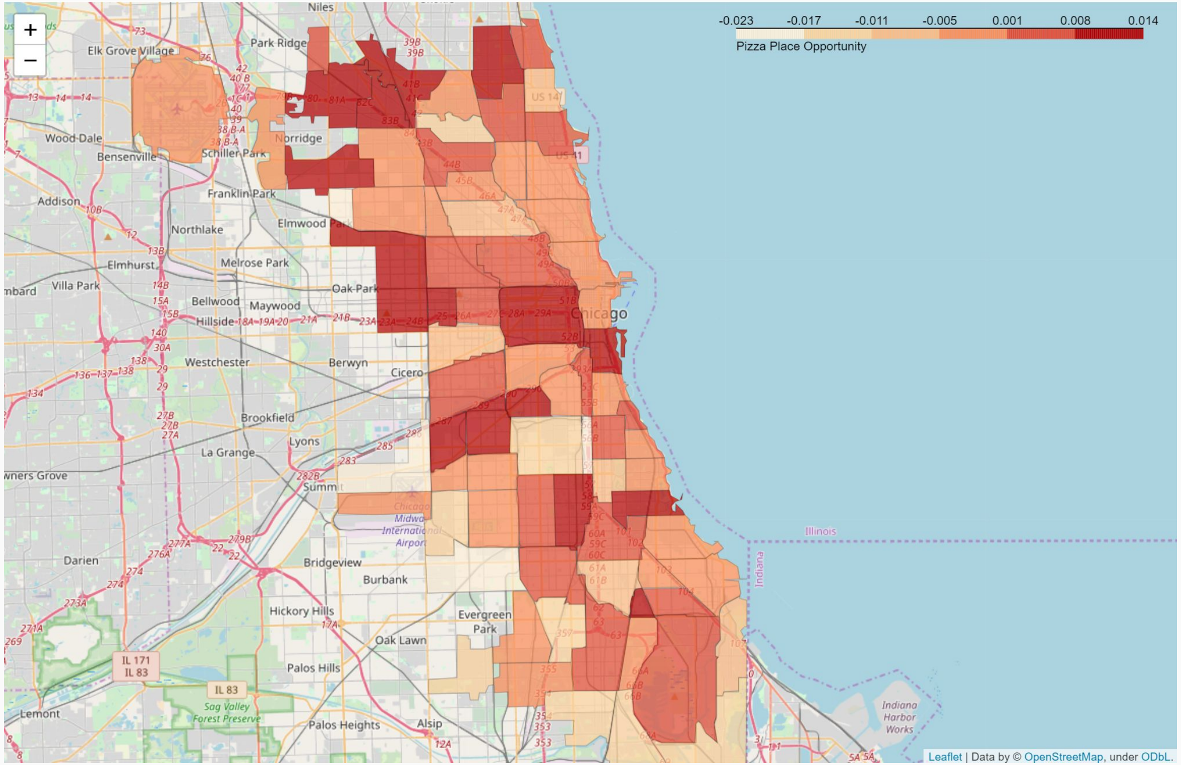
“Opportunity” is determined by the difference between predicted Pizza Place propensity and actual Pizza Place propensity



Spatially Visualized Heat Map

Much opportunity appears in Northwest and Southwest neighborhoods.

Additionally, neighborhoods directly to the West and South of the downtown Loop neighborhood, although rent prices will likely be high in those neighborhoods



Conclusions

Based off of the KNN regression modeling, the top 5 neighborhoods by pizza place opportunity are presented to the right.

McKinley Park and Archer Heights neighbor each other and are located along a public transit train line, suggesting that the greater “near Southwest” area could use additional pizza places.

Neighborhood	Opportunity
McKinley Park	+1.37%
West Ridge	+1.35%
Archer Heights	+1.35%
West Garfield Park	+1.34%
Dunning	+1.28%