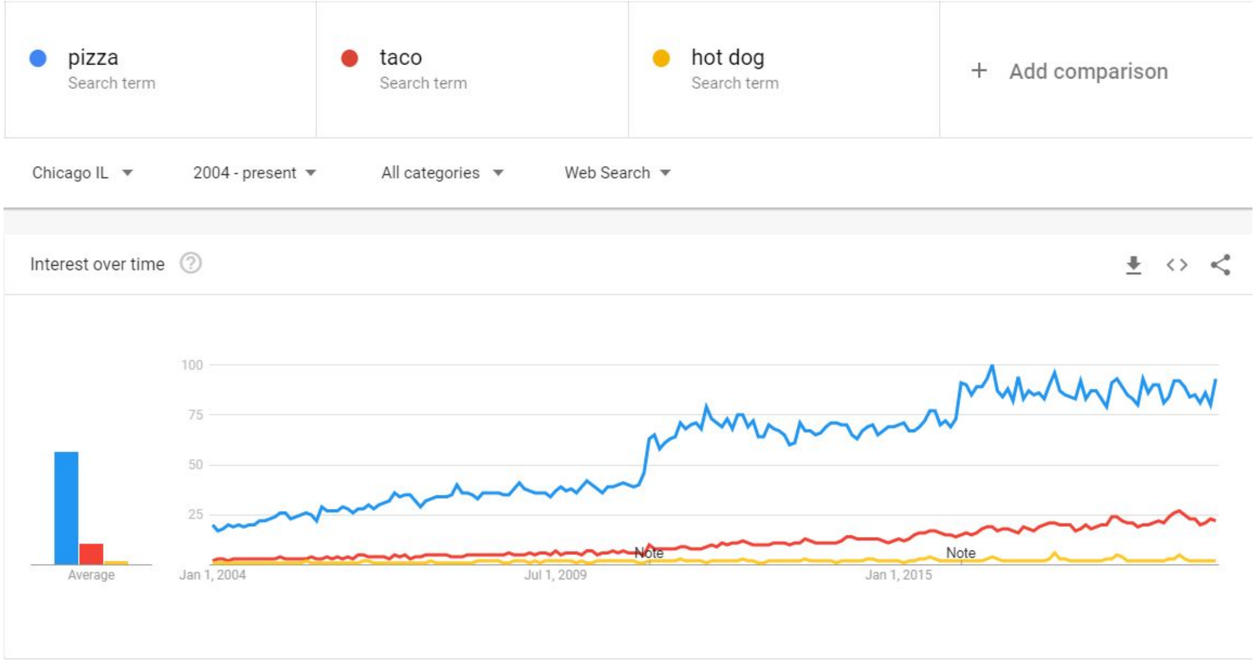
IBM Data Science Capstone Project

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

Introduction

Background

According to Google Trends, Chicago is the #5 city when it comes to search popularity for the term “pizza” dating back to 2004. In fact, Chicago’s interest in pizza far outweighs other popular dishes such as tacos and hot dogs during that same time frame:



Chicago pizza even made national news in 2019 when the local police department made a Twitter post that provoked a response from the New York Police Department. This debate has been ongoing for years between residents of the two cities, neither of which are willing to give in.

So, it’s fair to say that Chicagoans are passionate about pizza. Tourists flock to try one of the large chain deep dish joints, while locals tend to order tavern style thin crust. Regardless of style, it’s important that all neighborhoods in Chicago have appropriate access to the dish.

NYPD NEWS
@NYPDnews

We recognize the slice of pizza on the right, but what's the one on the left? Is there pasta in it? twitter.com/Chicago_Police…

Chicago Police @Chicago_Police
Saturday is #NationalPizzaDay. How will you celebrate, Chicago? Deep dish, or ol' fashioned thin crust? Either choice beats New York-style slices. Like/RT if you agree.

5,502 11:05 AM - Feb 6, 2019

Problem Statement & Research Question

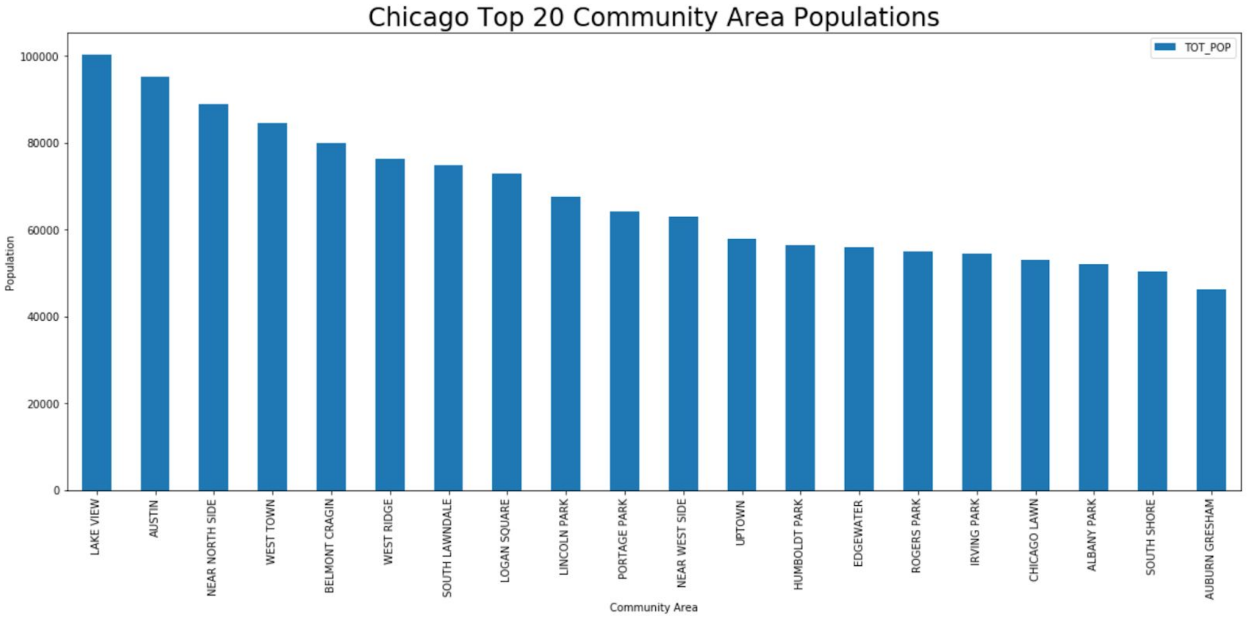
Are there neighborhoods in Chicago that are currently underserved by pizza places? It's unlikely that a "pizza desert" exists, but there may be neighborhoods with relatively few options when compared with others of similar composition. This knowledge would present an opportunity for small business owners to locate into a market that almost certainly would welcome additional options.

Data

Chicago Community Area Demographics

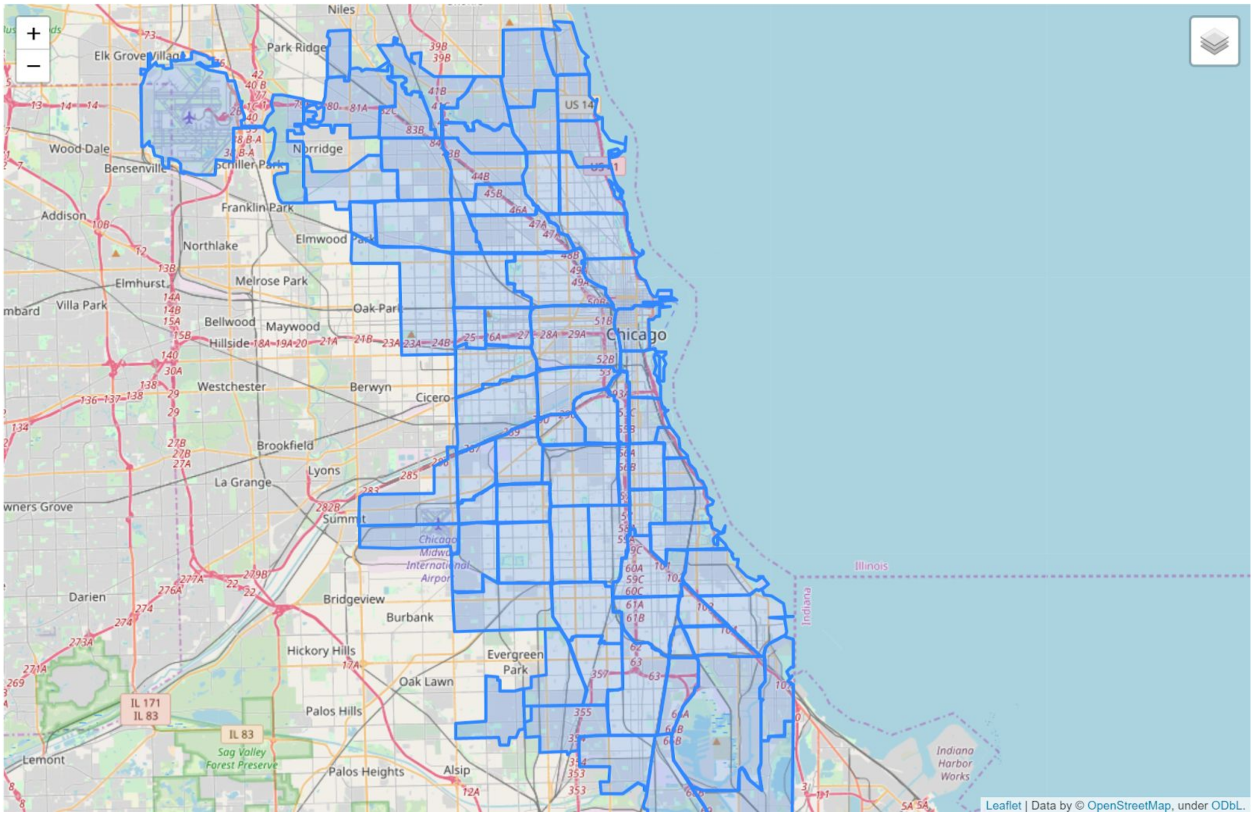
The state of Illinois maintains data on the demographic characteristics of the 77 official community areas (neighborhoods) in Chicago. This data spans 230 different features across the 77 neighborhoods, and can be found at the following URL:
https://datahub.cmap.illinois.gov/dataset/community-data-snapshots-raw-data

As an example of the type of data available, here are the top 20 neighborhoods in Chicago ranked by total population:



Chicago Data Portal - Neighborhood Boundaries

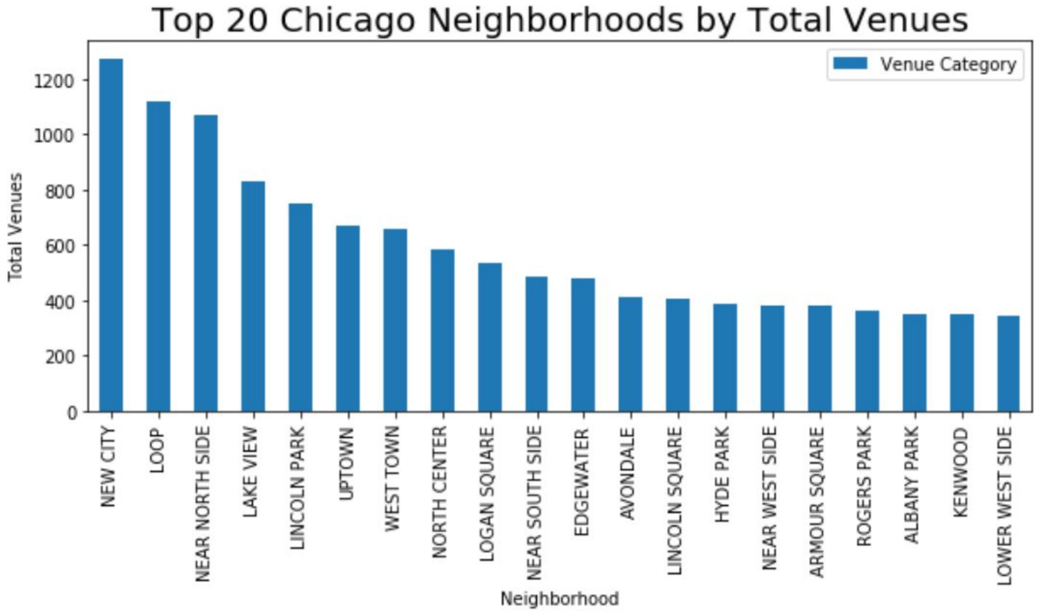
The city's data portal provides a neighborhood boundary geojson file that I'll be using to visually represent neighborhood level data, and can be found at the following URL: https://data.cityofchicago.org/Facilities-Geographic-Boundaries/Boundaries-Neighborhoods/bbvz-uum9



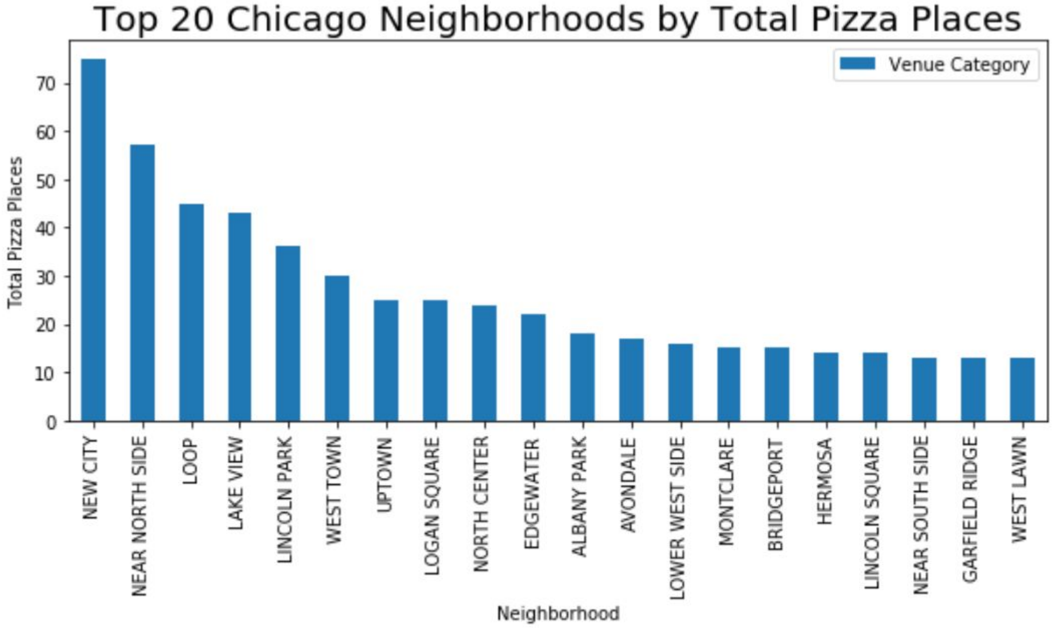
Foursquare API

Finally, I collected venue information for a 1 square mile radius in all of the neighborhoods. Ultimately, I obtained details for 23933 venues in Chicago. 880 of the venues were pizza places, which was the 3rd highest frequency of any venue category in the data collected.

Here are the top 20 neighborhoods by total venues collected:



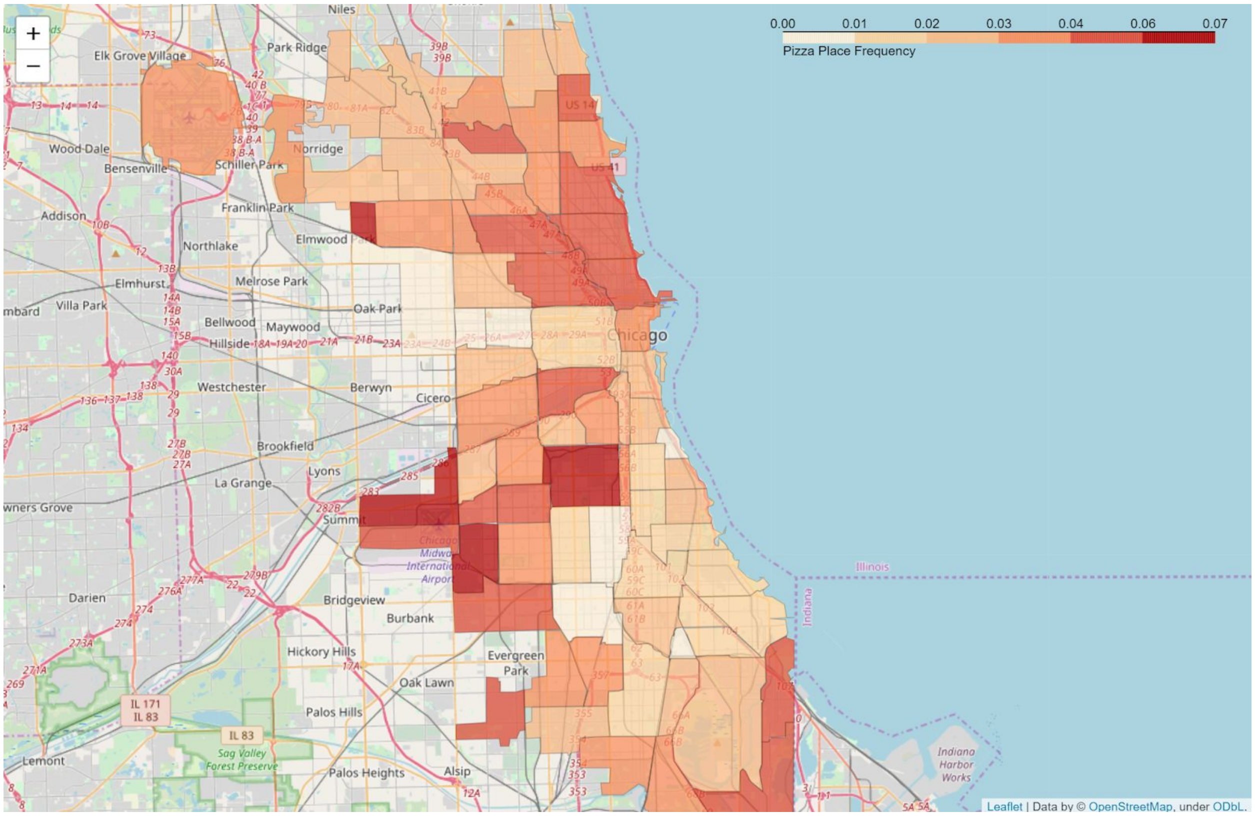
Here are the top 20 neighborhoods by total pizza places. You'll notice that a lot of the same neighborhoods are appearing here. Because of the discrepancy in total venues collected by neighborhood, I'm ultimately going to be using Pizza Places as a percentage of total venues as my target variable.



To acquire propensity by venue type, I one hot encoded the neighborhood data, and then grouped by neighborhood and found the mean for each venue category. That gave me a venue data set of 484 features across the 77 neighborhoods.

Target Variable

As mentioned, I am using the Pizza Place propensity by neighborhood as the target variable in this project. We can visualize each neighborhood's pizza propensity on a choropleth maps:



It appears that some of the southwest neighborhoods have the highest propensity of pizza places in the city, while the near north neighborhoods all appear to have between 4% and 6% of their total venues being pizza places. It's interesting that there are some southern neighborhoods with very low pizza propensity neighboring those with high propensity. These neighborhoods may potentially be an opportunity area, or there could be a good reason for the disparity.

Feature Set

Having 230 demographic features and 484 venue propensity features, it would be a fool's errand to attempt to model without doing some initial feature selection. I used univariate regression testing to automatically eliminate 95% of the features (using sklearn's SelectPercentile operation), and only keep those which scored best when regressed against the target variable. This method can lead to selecting many correlated variables, so I reviewed correlations and eliminated features that were highly correlated. This left me with the following set. During modeling I may only use a subset of these variables:

Demographic Variables	Venue Propensity Variables
WHITE	ATM
HISP	Bakery
BLACK	Bar
CARPOOL	Cosmetics Shop
INC_100_150K	Elementary School
MEDINC	Fast Food Restaurant
OWN_OCC_HU	Fried Chicken Joint
HV_150_300K	Gas Station
AVG_VMT	Gym
VACperc	High School
HCOV75K_LT20PCT	Hotel
in_lbr_frc_pct	Italian Restaurant
	Mexican Restaurant
	Park
	Southern / Soul Food Restaurant

Source Code

Data acquisition:
https://github.com/mkrivus/Coursera_Capstone/blob/master/FinalProject/Data_Acquisition.ipynb

Data processing and feature selection:
https://github.com/mkrivus/Coursera_Capstone/blob/master/FinalProject/Data_Processing_Pizza.ipynb