# Estimating saturation level of Food Venues in NYC

Shaimerdenov Sanzhar

August 9, 2019

## Introduction

### Background

New York City is one of the most famous cities of the world, located in the so-called tri-state area (between three different states), its population is 8.6 million, and every year, about 60 million tourists visit this city. To be honest, my choice of city based on the following logic – “what city I can choose, and have the most detailed and updated information?” In the means of detailed – existing Neighborhood tabulation areas (NTA) really helped, because there is 195 of them for NYC, each and every NTA has it’s own special features, and unique places that you’ll have to research.

### Problem

The problem is how do you choose a location to create a new restaurant, or buy an existing one? Let’s try to describe each and every Neighborhood tabulation area of NYC by the means, how good of a place is there to build a Food Venue. How we’ll take into account existing café/restaurant count, in order to see if the Neighborhood is over/under saturated in means of the Food Venue competitive market?

### Interest

Well, this task is as important for the startup companies just planning to open the very first food Venue, big companies which has a chain restaurants, and need to determine a perfect timing and location or a new place to be opened, and the created approach allows us to help them with the research of the most favourable NTA.

## Data acquisition and cleaning

### Data sources

From the first source, I’ve taken GeoJSON file containing polygons for every NTA in NYC:

<https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page>

The second source of information I used in order to get the population information, dated 2010:

<https://data.cityofnewyork.us/City-Government/New-York-City-Population-By-Neighborhood-Tabulatio/swpk-hqdp/data>

The third source had the data for NTAs regarding the median income amount, median age, working population amount, average for 2008-2012.

[https://geodacenter.github.io/data-and-lab//NYC-Nhood-ACS-2008-12/](https://geodacenter.github.io/data-and-lab/NYC-Nhood-ACS-2008-12/)

I also used a few Wikipedia pages in order to check the quality and reasonableness of the provided data, by summing up the population amount, area size (extracted from links above) for 5 Boroughs of NYC.

And, as a last but not least, I’ve calculated Polygon centroids for NTAs and the farthest point of polygon to center and requested the data from Foursquare for each and every NTA (search - browse):

<https://api.foursquare.com/v2/venues/search>

After that, I had to see the structures of the Venue categories, in order to make them more generalized, and this can be done with the following request to Foursquare:

<https://api.foursquare.com/v2/venues/categories>

### Data cleaning

As previously written, a lot of the data had been checked with alternative sourses of information:

We checked alternative GeoJSON source with NYC NTAs in order to compare the polygon area sizes – no major differences identified.

We’ve extracted the population amount for 5 NYC Boroughs from Wikipedia in order to check the precision of current NTA population information. There were some differences that I’ve eliminated through transformation of existing data.

Out of 195 NTAs, 23 of them didn’t have the information regarding the median income and age, so initially I’ve deleted them from dataset initially, however the initial research showed no real correlation between this parameters and the Food venue count, so these elements had been restored for the final modelling.

Regarding the foursquare data, I’ve deleted every Venue without category (about 2000 from 47000), and also I’ve deleted the Venues that was provided for certain NTA – in a circle around its polygon centroid, but wasn’t actually inside the polygon. Also the generalized category of venue (about 12 categories) had been attached to each Venue, in order to aggregate them.

### Feature selection

After the data was cleansed, there had been 195 elements with 17 features (including target). Before creating the model, I wanted to see, what parameters is not bringing any value to the model, and can be safely deleted. Also some of the features can be correlated to each other, which is also very bad.

At the end, 8 features had been remaining.

|  |  |  |
| --- | --- | --- |
| **Kept features** | **Dropped features** | **Reason for dropping features** |
| 'popinlabou' | 'Population', 'Area\_normalized, sq km','medianinco', 'medianage', 'Population density' | No correlation with the target, created a linear regression, R squared for those estimations were less than 0.05 |
| 'popinlabou' | 'labour\_coef' | popinlabou - meaning population count that is working, and labour\_coef, meaning percentage of working population is correlated to each other, and additing two of them simultaneously wouldn't give any information gain |
| 'Arts & Entertainment','College & University','Nightlife spot','Outdoors & Recreation', 'Professional & Other places', 'Residence', 'Shop & Service', 'Travel & Transport' | 'Event' | Not enough venues under the 'Event' |