Estimating saturation level of Food Venues in NYC

Shaimerdenov Sanzhar

August 9, 2019

Contents

1. Introduction

1.1. 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.

1.2. Problem

The problem is how do you choose a location to open 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?

1.3. 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.

2. Data acquisition and cleaning

2.1. Data sources (jupiternotebook part 1-2)

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/

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

2.2. Data cleaning (jupiternotebook part 1)

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, 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 (10 categories) had been attached to each Venue, in order to aggregate them.

2.3. Feature selection (jupiternotebook part 2-3)

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.

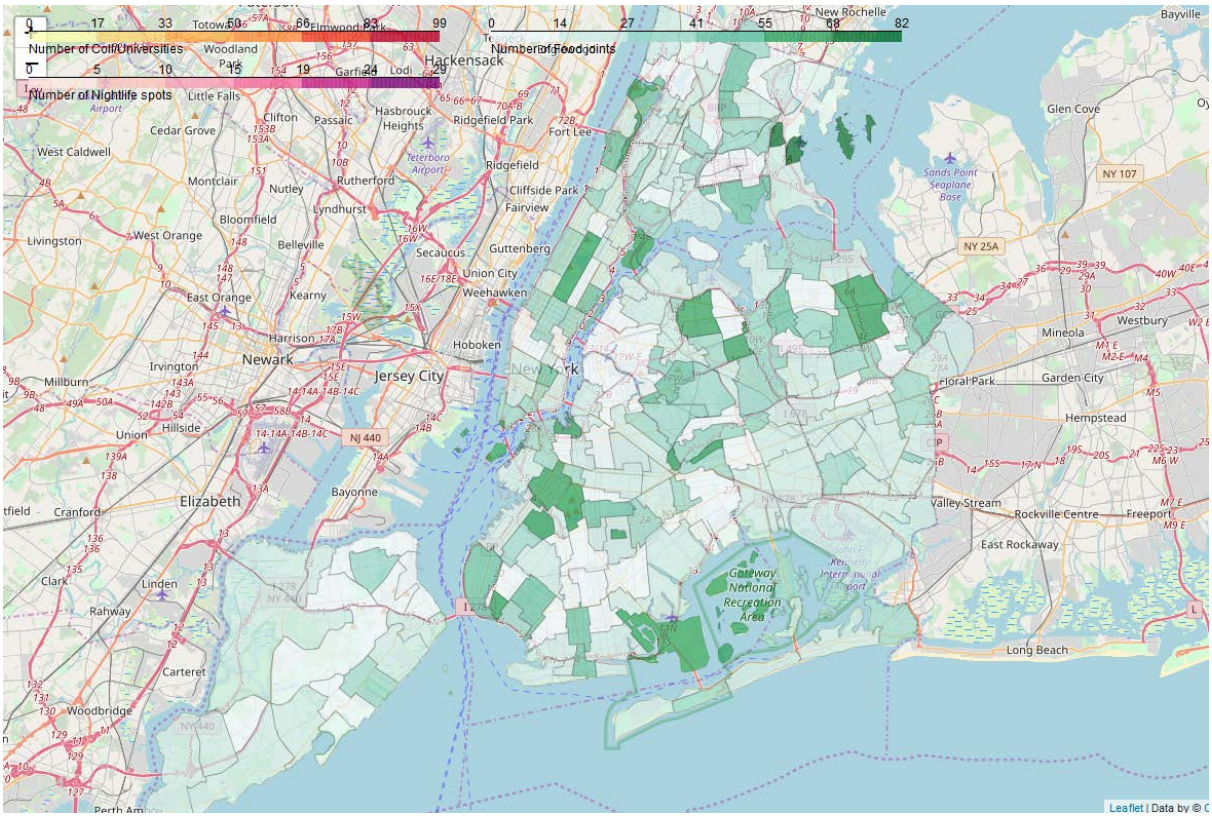
At the end, 8 features had been remaining.

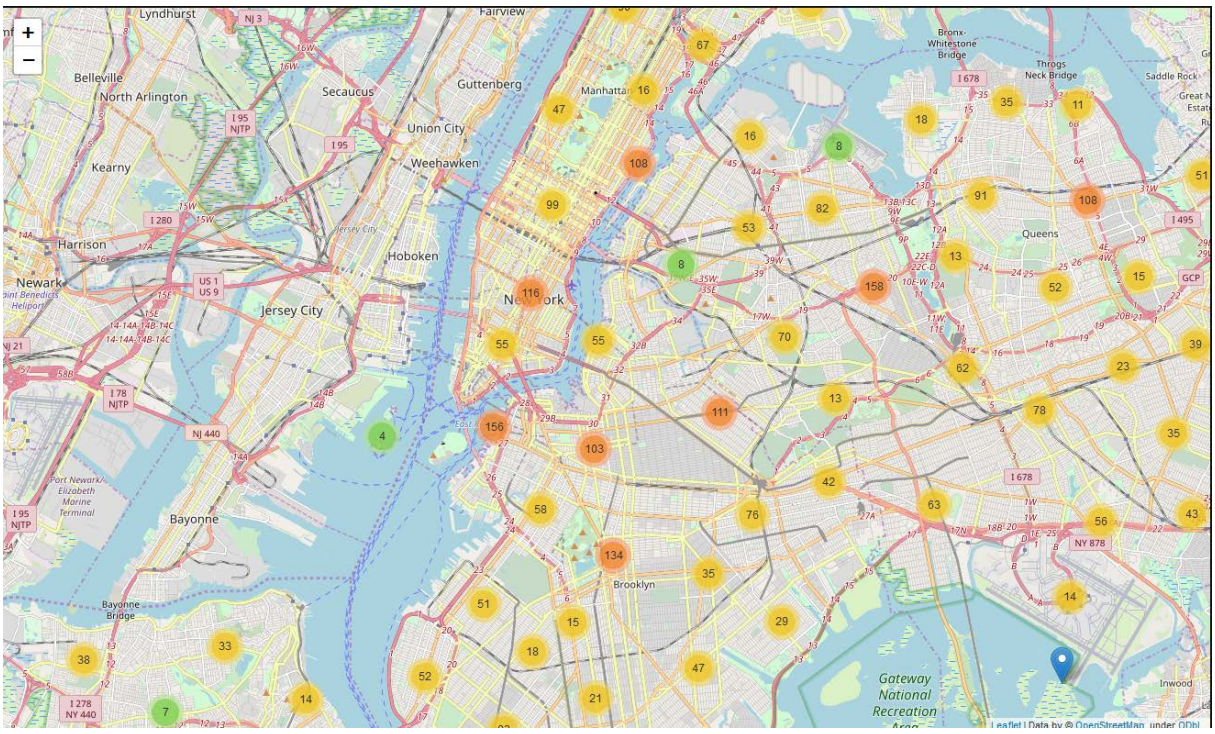
Kept features	**Dropped features**	**Reason for dropping features**
'popinlabou'	'Population', 'Area_normalized, sq km','medianinco', 'medianage', 'Population density'	No correlation with the target, created a linear regression, R squared for those estimations were less than 0.05
'popinlabou'	'labour_coef'	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'

3. Exploratory Data Analysis

3.1. Calculation of target variable (jupiternotebook part 2)

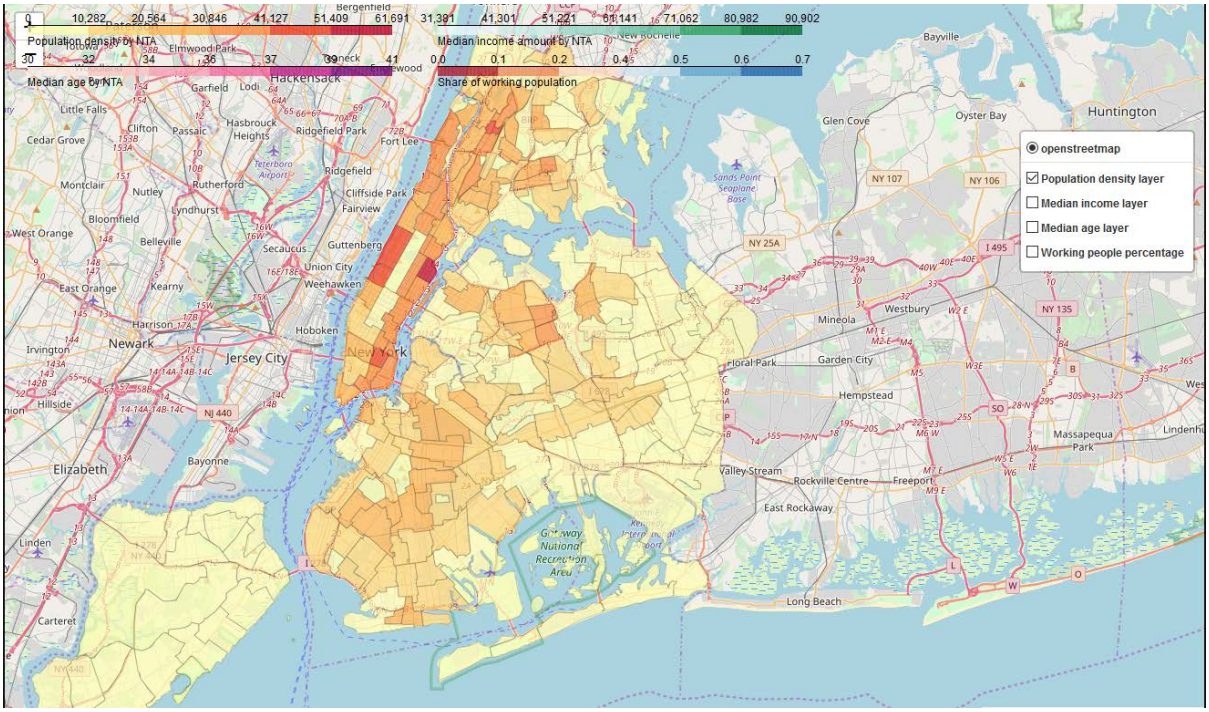
Count of Food Venues, aggregated information from Foursquare by Neighborhood tabulation areas for New York. It will be more understandable to see (please check the dynamic map in jupiternotebook):



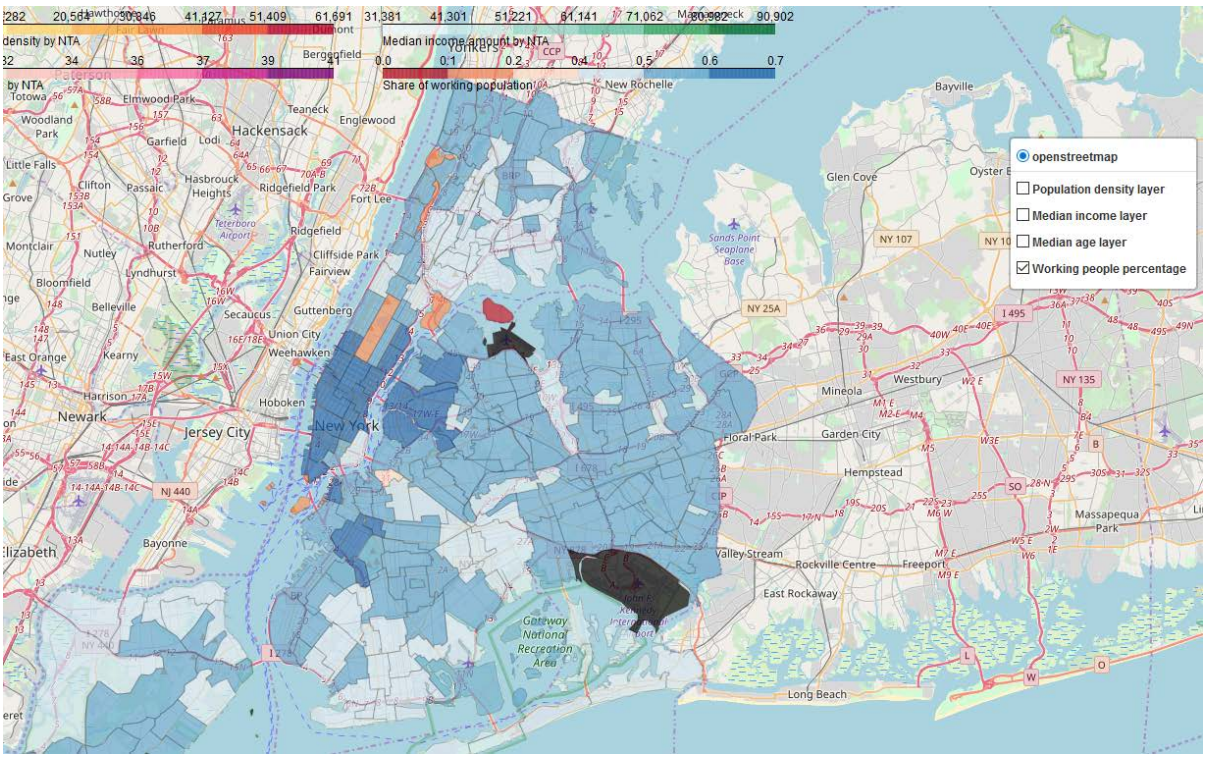


3.2. Population statistics by NTA (jupiternotebook part 1)

The information regarding different population statistics, which had been extracted can be shown in jupiternotebook as a multi-layer map, for example – population density shows the increase in Manhattan districts (except for central park):



The next picture is the share of the working population (black is the NTA's with missing information):

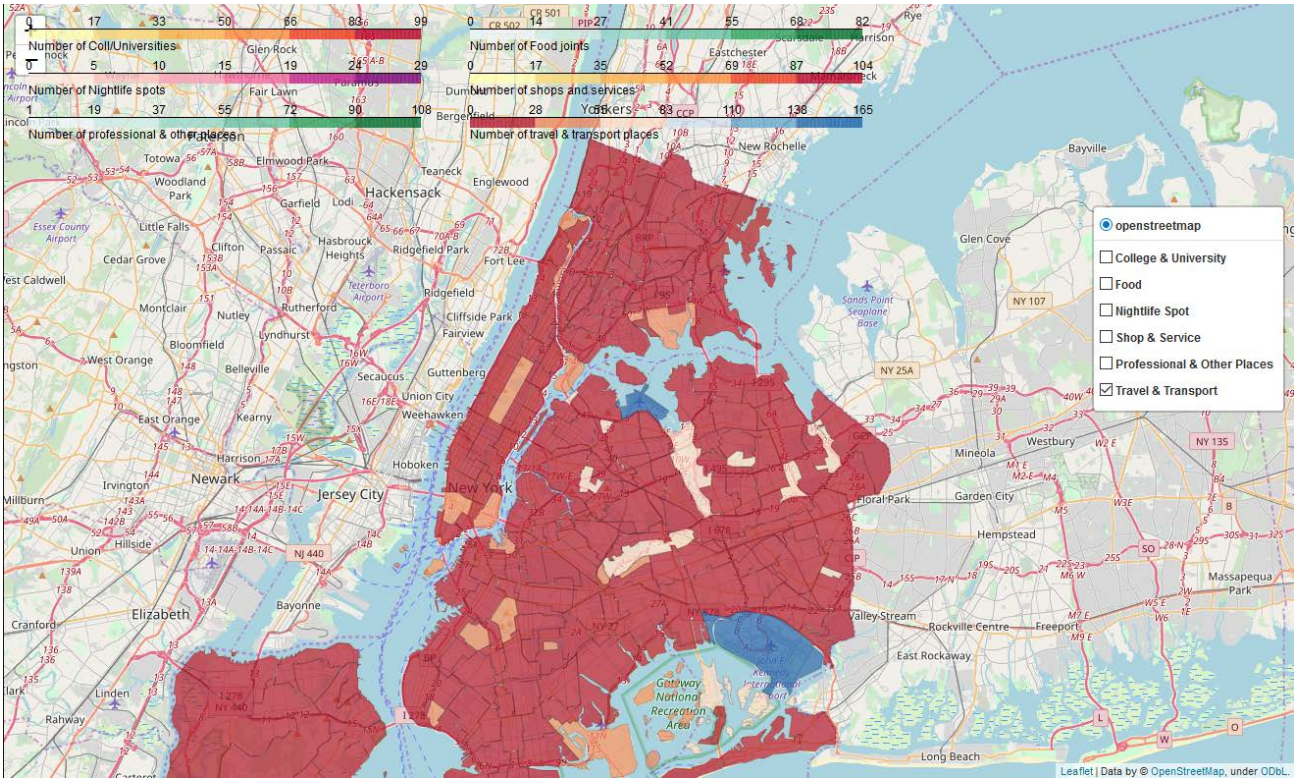


3.3. Venue statistics by NTA (jupiternotebook part 2)

So up next, is the information regarding the venues, that we extracted from Foursquare. Initially, we extracted 47365 Venues, however this information had Venues without the category inserted, contained duplicated Venues and Venues that is not inside any of the NTA due to the nature of the Foursquare request – so we used **shapely contain** method in order to see only the venues that is inside the Polygon (or multipolygon) of the NTA. After cleansing, 23465 Venues remained, that had been aggregated by 10 generalized categories (Foursquare has a 4-layer hierarchy of categories, which contains 865 category endpoints overall). Here is the list of venues:

Arts & Entertainment
College & University
Event
Food
Nightlife Spot
Outdoors & Recreation
Professional & Other Places
Residence
Shop & Service
Travel & Transport

Here is the screenshot of the multi-layer map showing count of certain venue categories by NTA on a map – currently there is opened Travel & Transport, and the most intense areas (blue) is near the airport area.

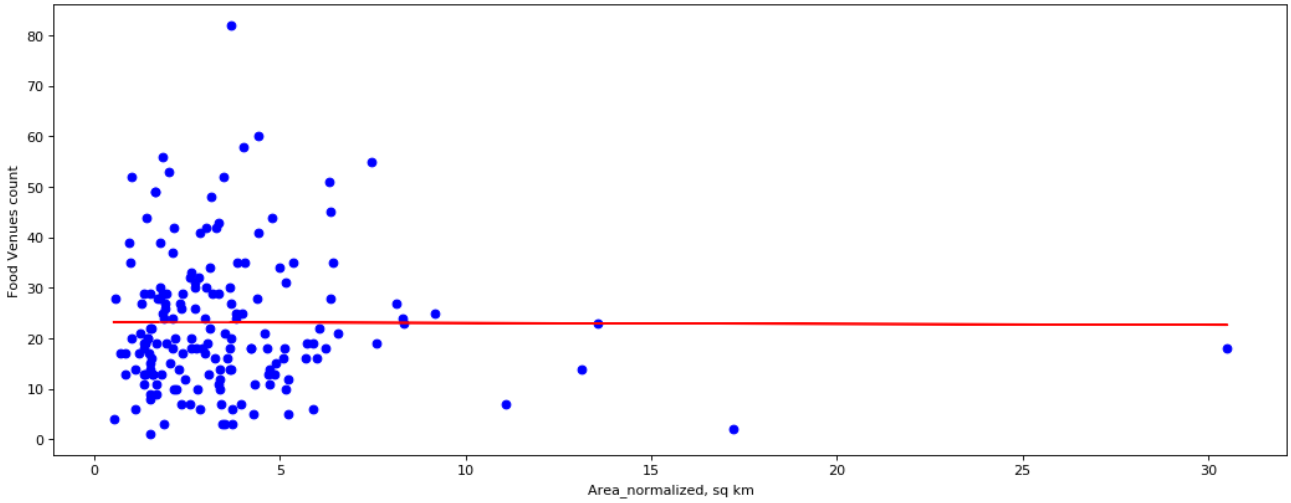


3.4. Relationship between target variable and other features (jupiternotebook part 3)

We've performed analysis of relationship between some singular features by creating a linear regression and analyzing a scatterplot:

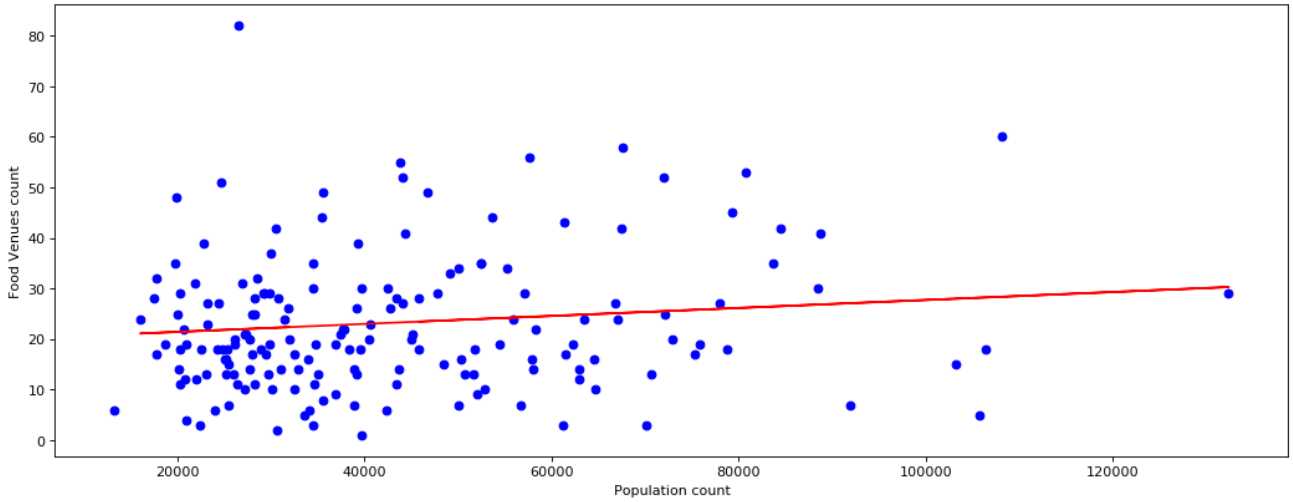
Area size normalized, sq km, linear regression coefficient=-0.019, R squared=-0.03

Doesn't show anything, and doesn't look like any non-linear dependency also



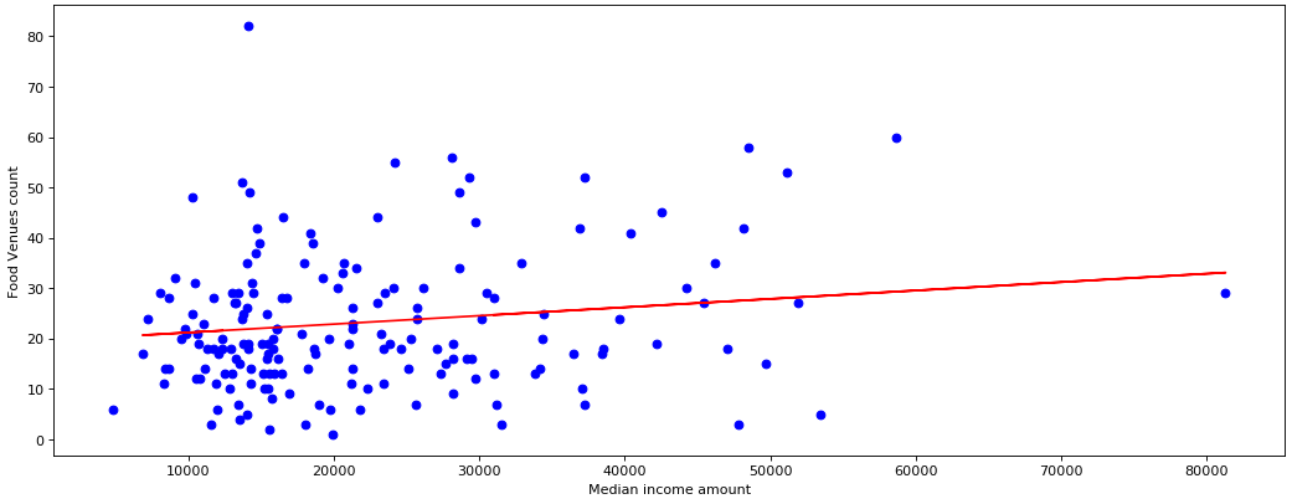
Population count, linear regression coefficient=0.0001, R squared=0.04

Doesn't show anything, and doesn't look like any non-linear dependency also



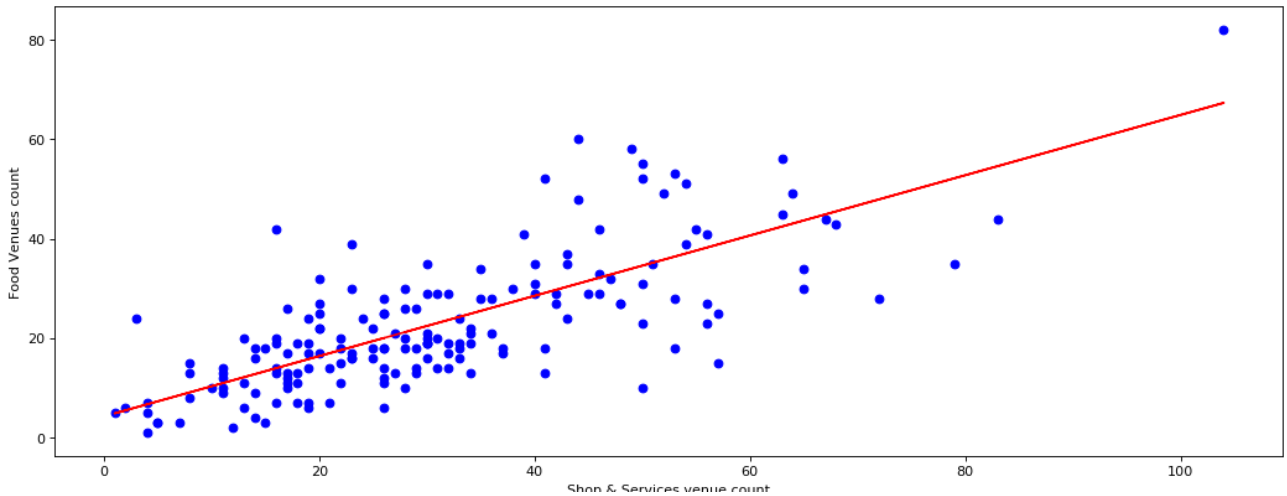
Median income amount, linear regression coefficient=0.0002, R squared=0.08

The angle is too flat to be effective used in algorithm, and R squared shows that prediction is very inconsistent.



Shop & Services venue count, linear regression coefficient=0.6065, R squared=0.45

This is an example of a good parameter.



4. Predictive modeling

4.1. Regression models (jupiternotebook part 3)

We've tried a bunch of various regression techniques, for most of them (except k-fold) we've separated the population into a train and a test splits by 80/20 size ratio. The resulting performance of the models on a test sample were:

Linear model 'Nightlife spot venues' - **R squared** is equal to 0.16 mean squared error (**MSE**) is equal to 152.42

Linear model 'Shop & Service venues' - **R squared** is equal to 0.36 mean squared error (**MSE**) is equal to 115.48

Multilinear model - **R squared** is equal to 0.42 mean squared error (**MSE**) is equal to 105.92

Ridge regression (best with poly=2, alpha=200,000,000) - **R squared** is equal to 0.44 mean squared error (**MSE**) is equal to 101.18

K-fold multilinear regression (k=4) - **R squared** is equal to 0.52 mean squared error (**MSE**) is equal to 84.98

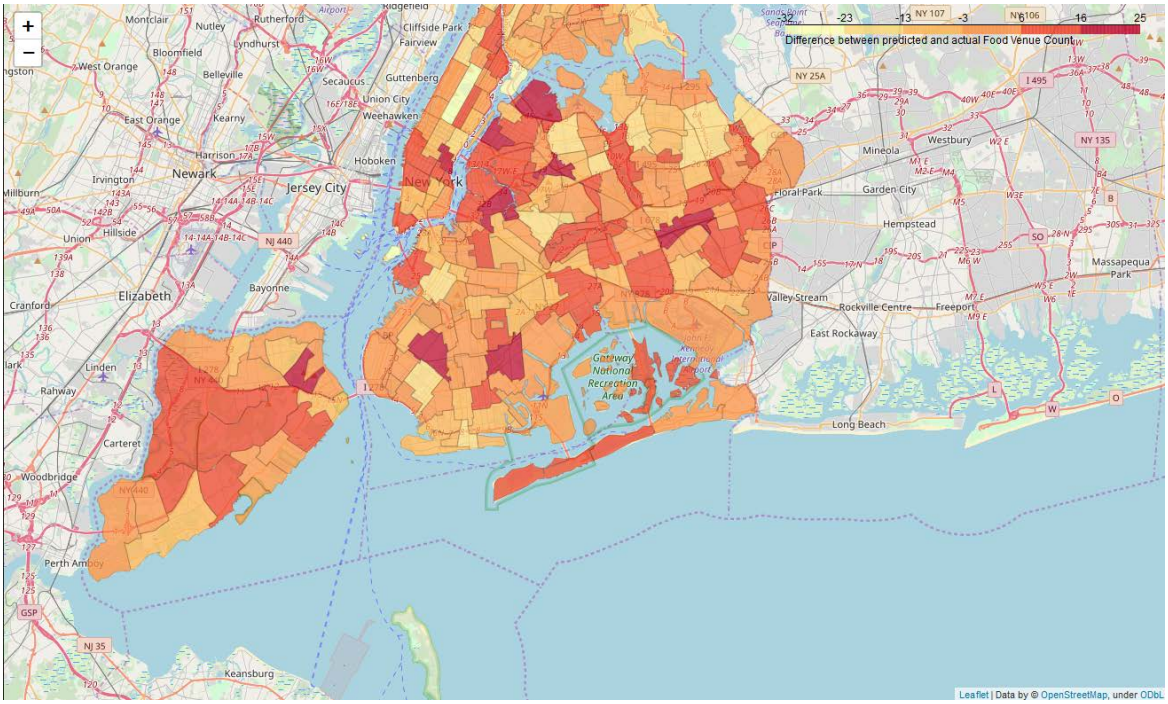
Based on the presented data – we will be using K-fold multilinear regression for our prognosis creation.

5. Conclusions

The main idea here, is to use the difference between the the actual and predicted value not as error, but as some kind of missed opportunity, that we can research specified neighborhoods, and decide that we can add another Food Venue there, with high demand and low competitive situation. Let's see top-10 Neighborhoods with the biggest differences between actual and predicted values:

NTA Code	NTA Name	Borough	Food Venue Count	Food venues estimate	Food Venue count diff
QN72	Steinway	Queens	14	39.48	25.48
BK58	Flatlands	Brooklyn	10	34.20	24.20
SI08	Grymes Hill-Clifton-Fox Hills	Staten Island	3	25.42	22.42
BK88	Borough Park	Brooklyn	18	39.90	21.90
BK90	East Williamsburg	Brooklyn	6	25.87	19.87
MN20	Murray Hill-Kips Bay	Manhattan	13	31.55	18.55
QN68	Queensbridge-Ravenswood-Long Island City	Queens	5	22.29	17.29
QN61	Jamaica	Queens	16	32.89	16.89
QN50	Elmhurst-Maspeth	Queens	15	31.03	16.03
MN21	Gramercy	Manhattan	17	32.73	15.73

Also, a map of NYC NTAs with the differences:



6. Future directions

We should include the information about population more like – how many people pass through that neighborhood, or street on a daily basis, because how many people actually live in this neighborhood turned out to be useless information.

We can point out, what exact location is the best to place a Food venue in, depending on the distance, count and category of the nearby Venues.

We can also add information on other cities to enrich the data and increase the number of elements.

The final comment for development is on data modelling – we can create an Ensemble model – first to segment similar neighborhoods, and then create specific predictive algorithm tailored for that specific segment.