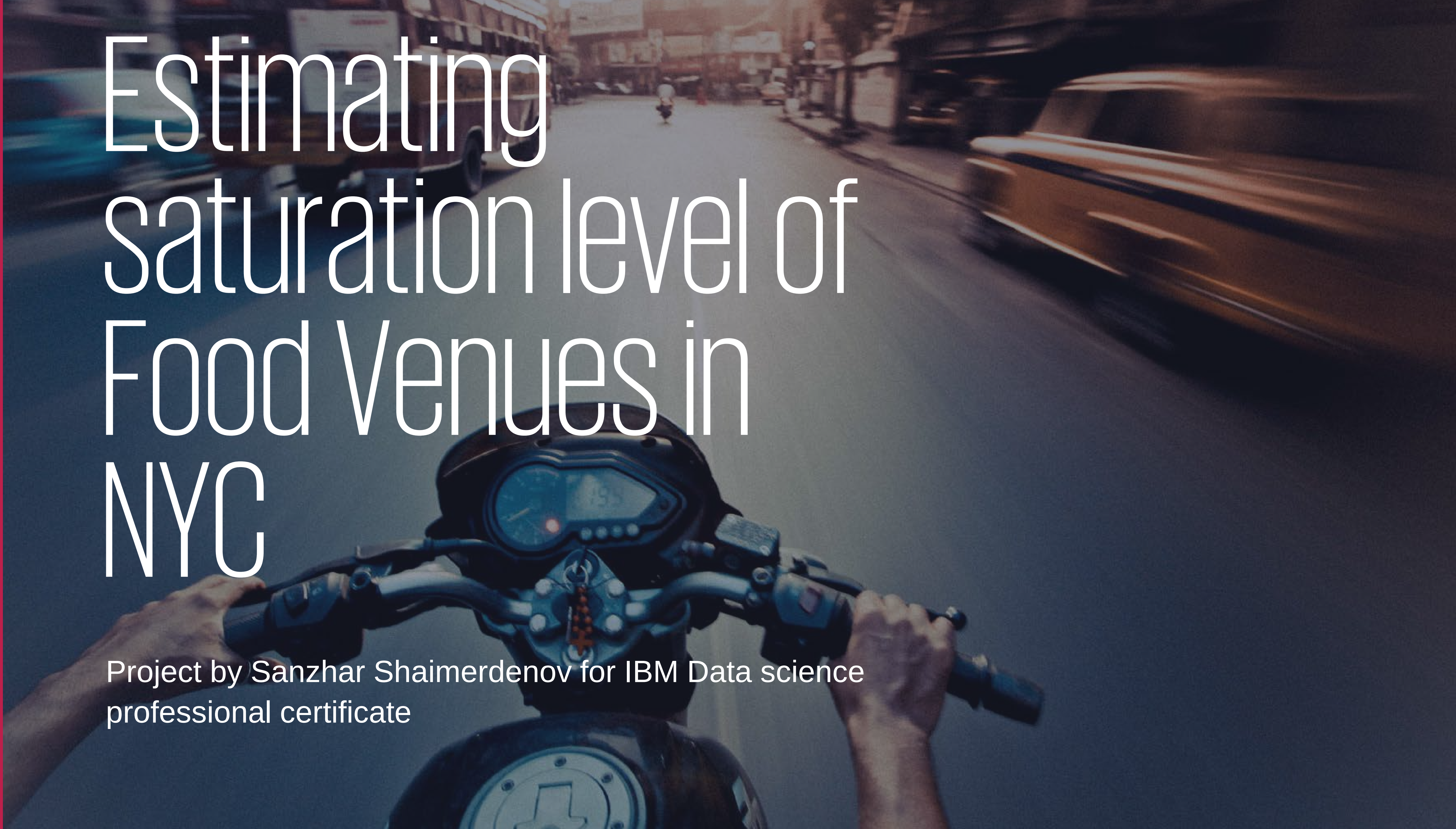


Estimating saturation level of Food Venues in NYC

Project by Sanzhar Shaimerdenov for IBM Data science professional certificate

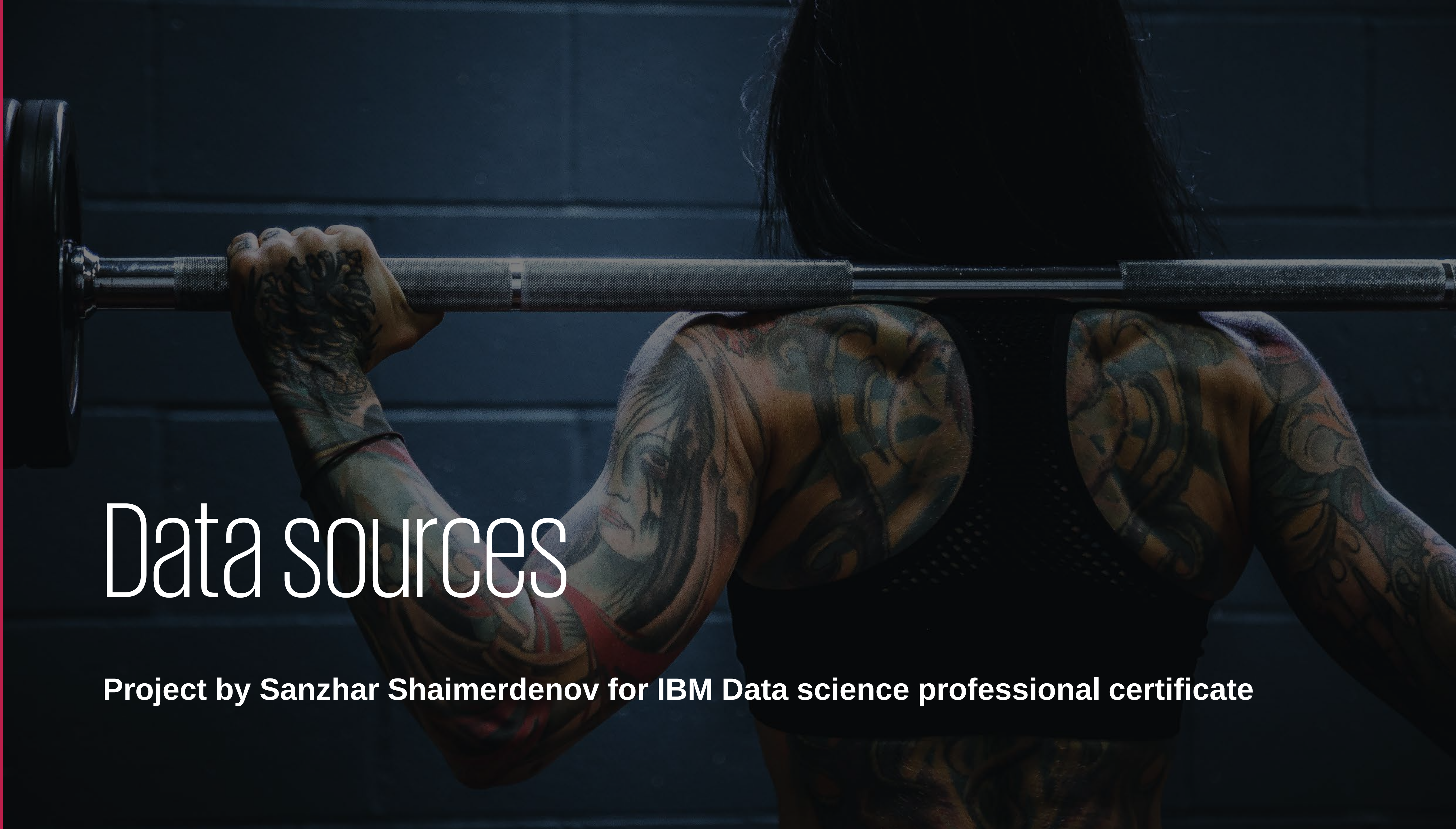
A problem appears

How do I choose the best spot for my new Point of sale?

We can you Data science instruments to suggest, what kind of spots we should be closely looking at, because they don't have enough similar kind of Venues nearby to the suggested spot.

Specify the task with assumptions

1. New-York city is a big and well-known all across the world, it must have a lot of data sources describing it, that we can use to create our model.
2. Let's focus on the specific type of Point of Sale – Food Venues (restaurants, cafes, etc.).
3. Divide New-York city into some manageable amount of areas that we can aggregate our parameters into – and then suggest the potential client to look into some specific area.



Data sources

Project by Sanzhar Shaimerdenov for IBM Data science professional certificate

What kind of data we will be using and merging together?

GOVERMENT
WWW1.NYC.GOV
Neighborhood tabulation Areas (NTA) shapes, latitude longitude coords
GeoJSON file

195 NTAs – with the notion of polygon and their location on the map

NYC OPEN DATA
data.cityofnewyork.us
Population statistics

2010 Population amount and calculate the population density to each NTA

GITHUB
geodacenter.github.io
Statistical information On income, Population age Working population

2008-2012 aggregated Various statistical information regarding the people/ households to NTA (about 20 out of 195 had missing stats)

Foursquare
api.foursquare.com
Extraction of venues Categorization of venue types

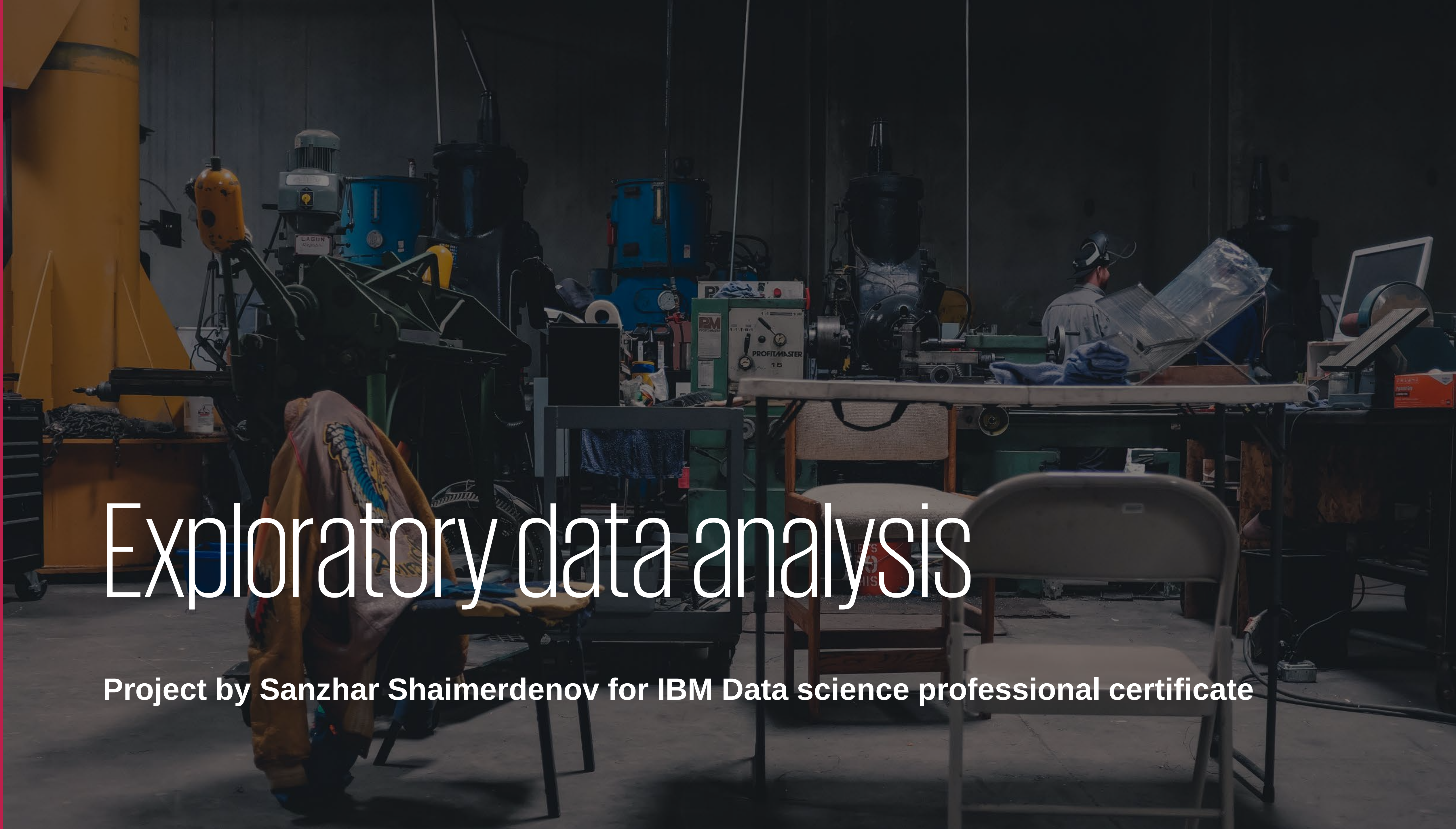
47k Venues – cleansed into 43k Venues.

865 categories aggregated into 10 general types

Wikipedia
Double-check calculated Area sizes, population & Other information

Feature selection table

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'

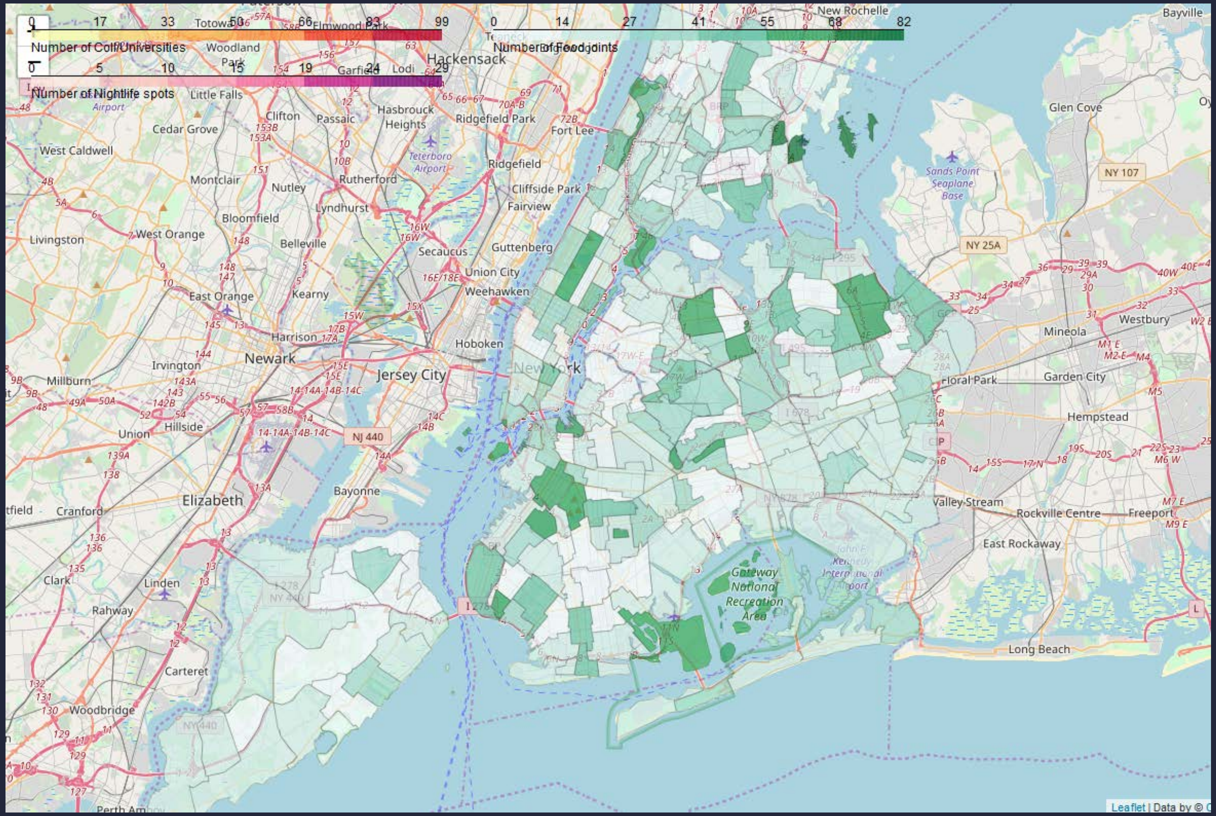


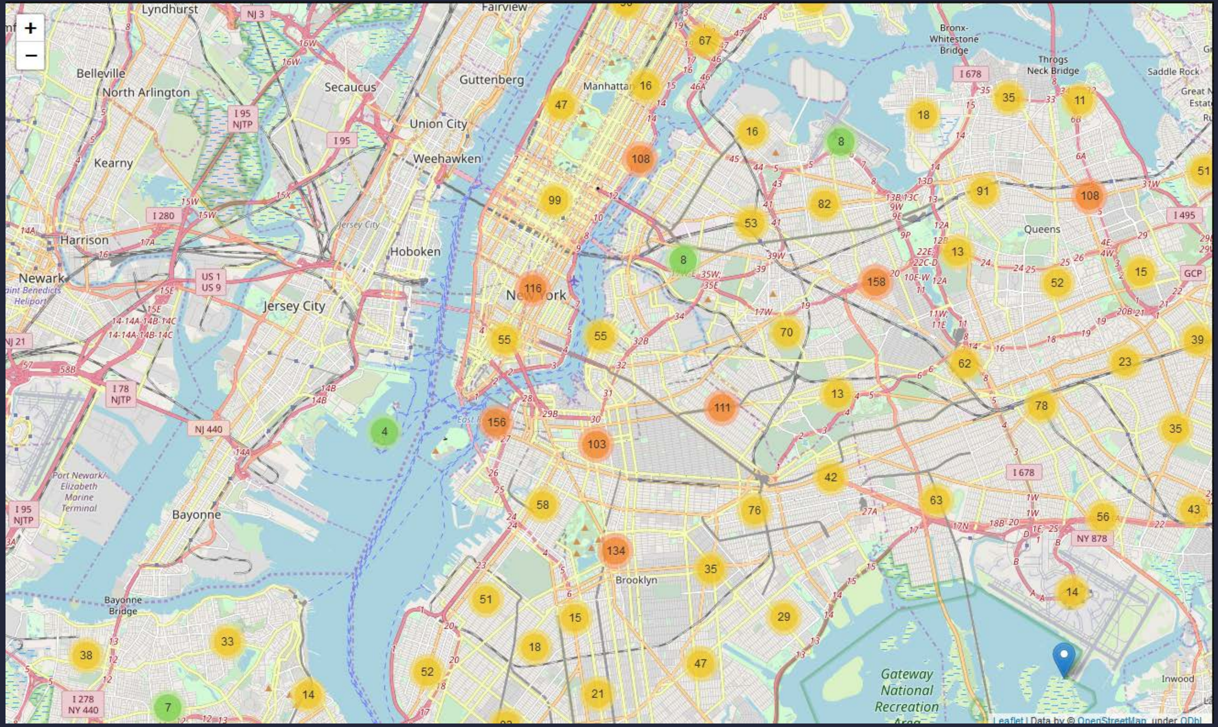
Exploratory data analysis

Project by Sanzhar Shaimerdenov for IBM Data science professional certificate

How do we define a target variable?

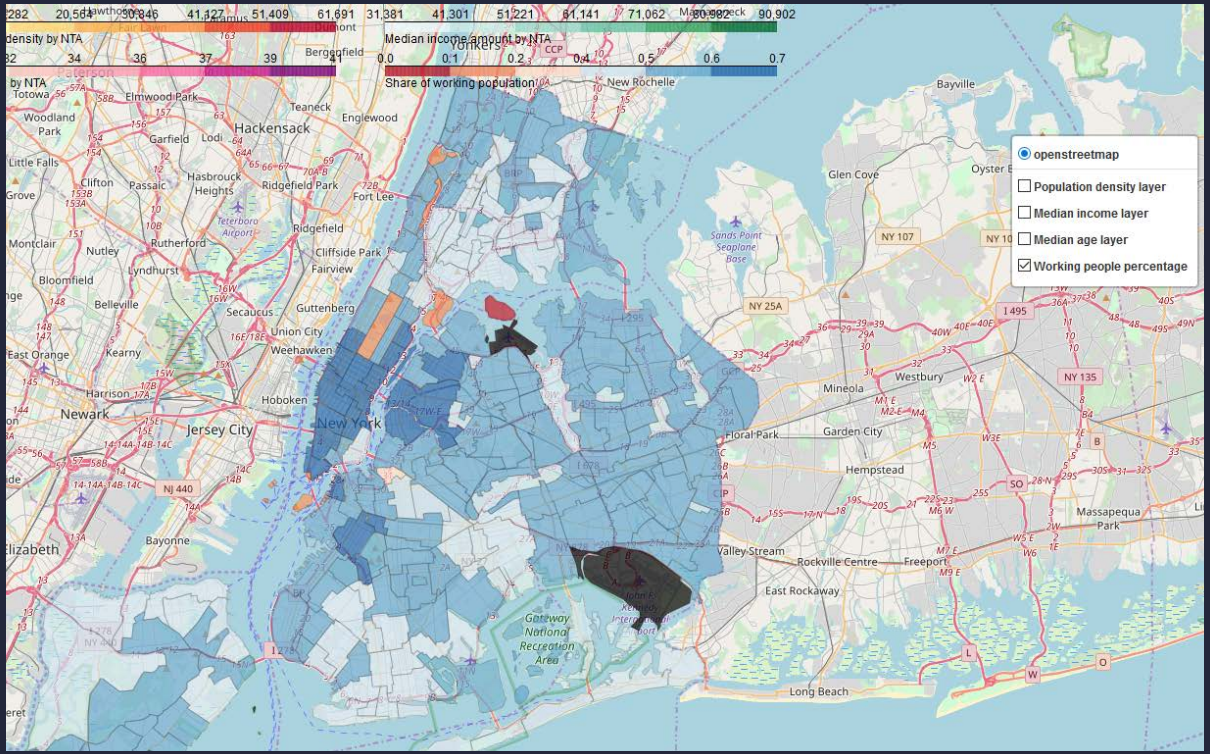
Venue count aggregated by NYC NTA's, with the general category 'Food'

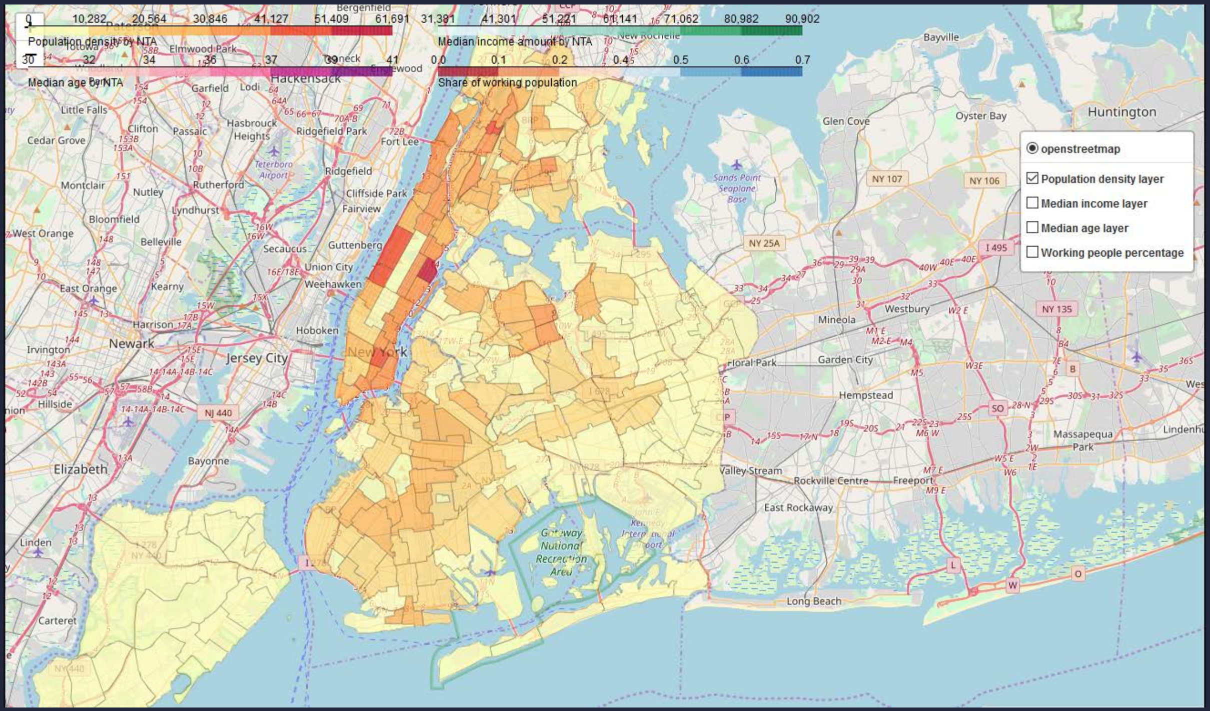




Population statistics by NTA

The information regarding different population statistics, which had been extracted can be shown in jupiternotebook as a multi-layer map

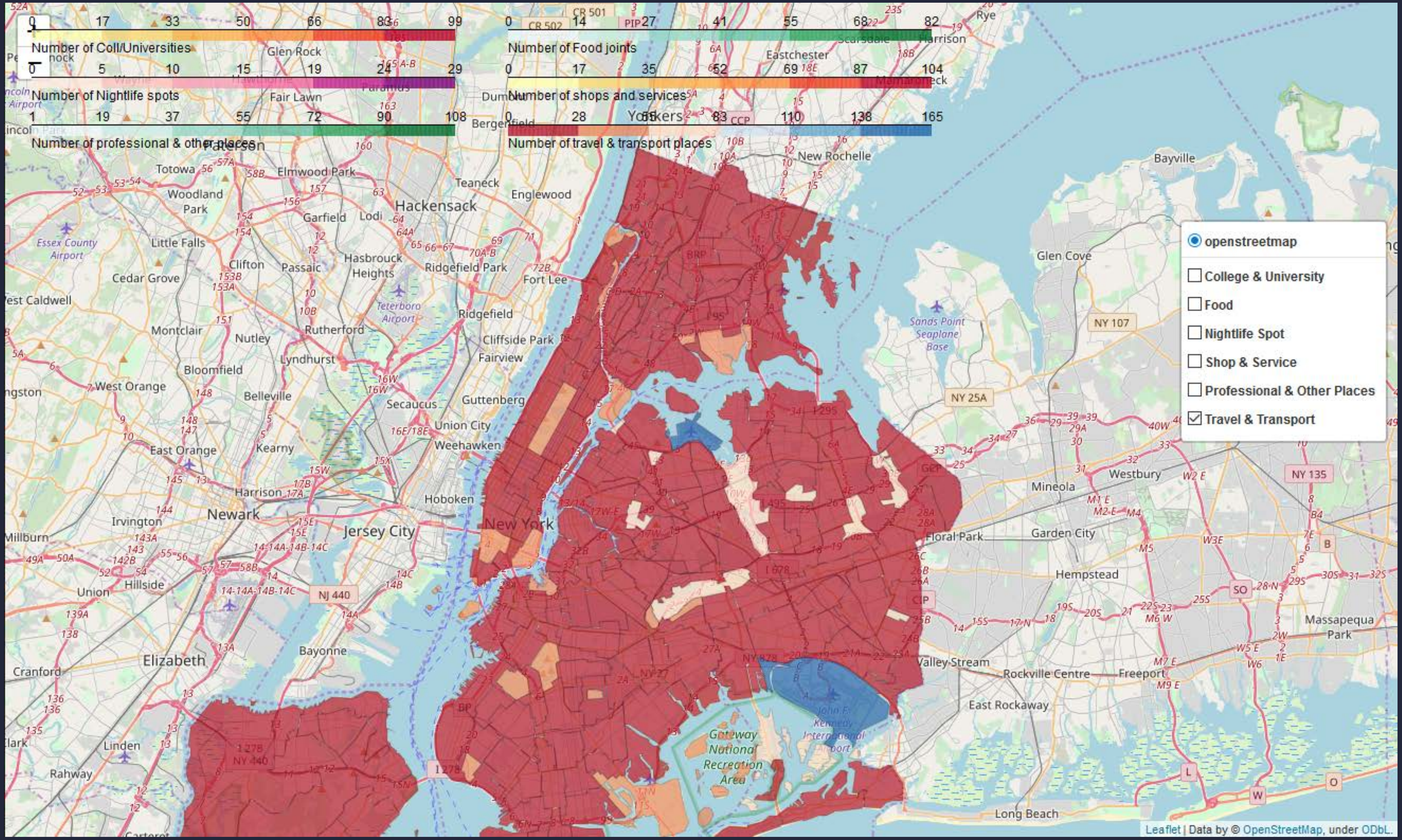




Venue statistics by NTA

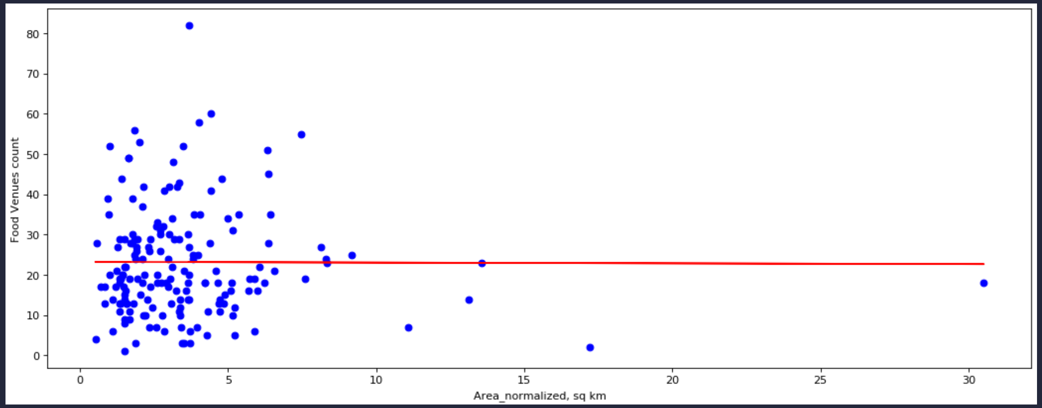
The information regarding various general categories assigned to each neighborhood tabulation area (NTA)

You can check out that fully dynamic multi-layer map in attached jupiternotebook - https://github.com/Lovecraft-hp/Data_science_pile/blob/master/Final_capstone_IBM_DSP%20-%20part2.ipynb

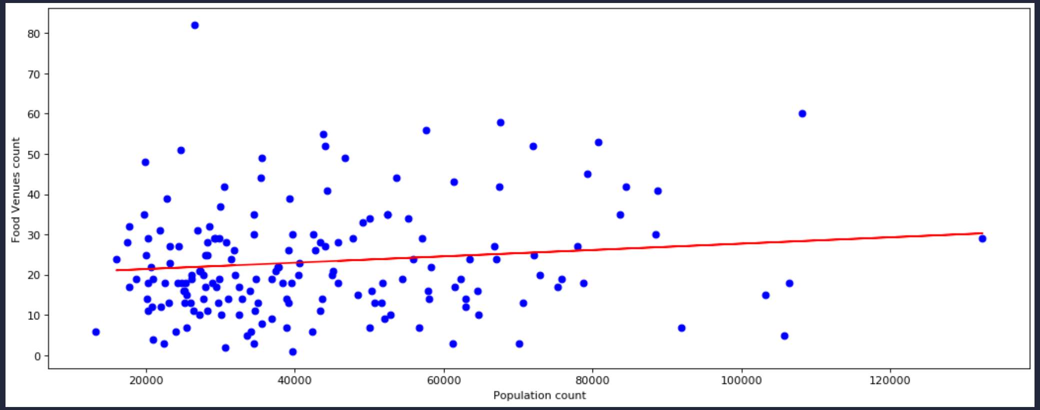


Relationship between target and parameters

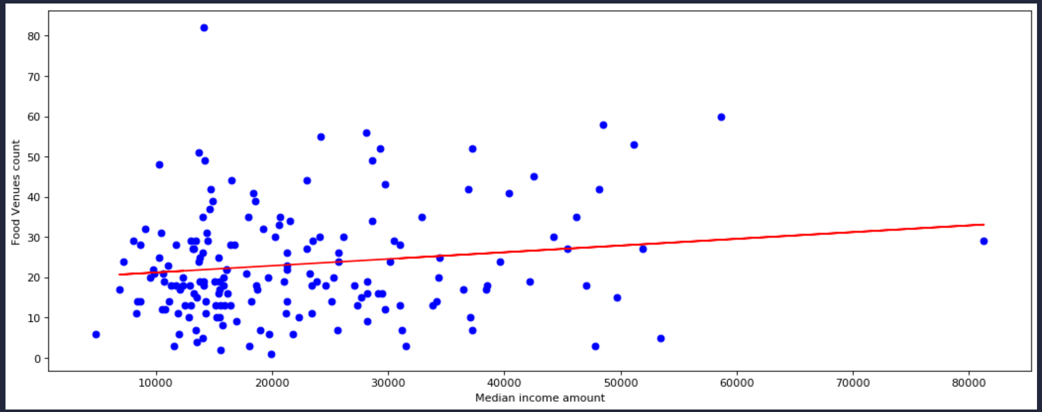
Area size normalized, sq km, linear regression coefficient=-0.019, R squared=-0.03 (dismissed)



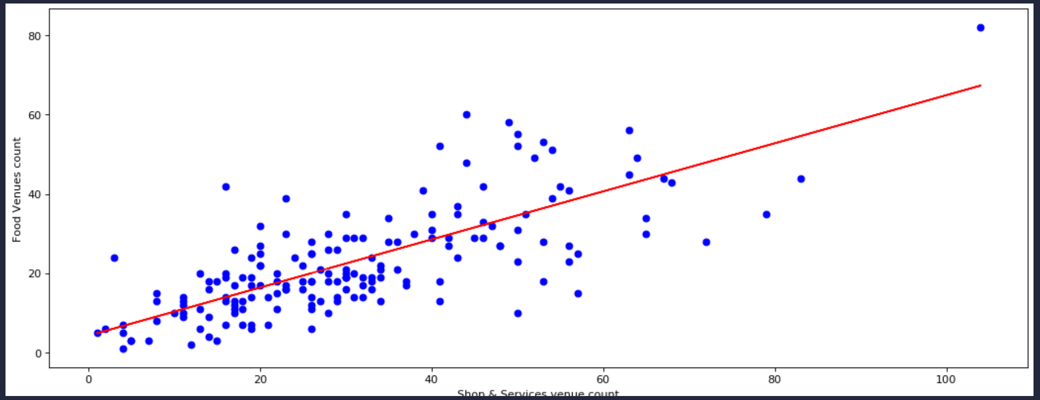
Population count, linear regression coefficient=0.0001, R squared=0.04 (dismissed)



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



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



Predictive modeling

Project by Sanzhar Shaimerdenov for IBM Data science professional certificate

Regression models comparison

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.

Conclusions & Future directions

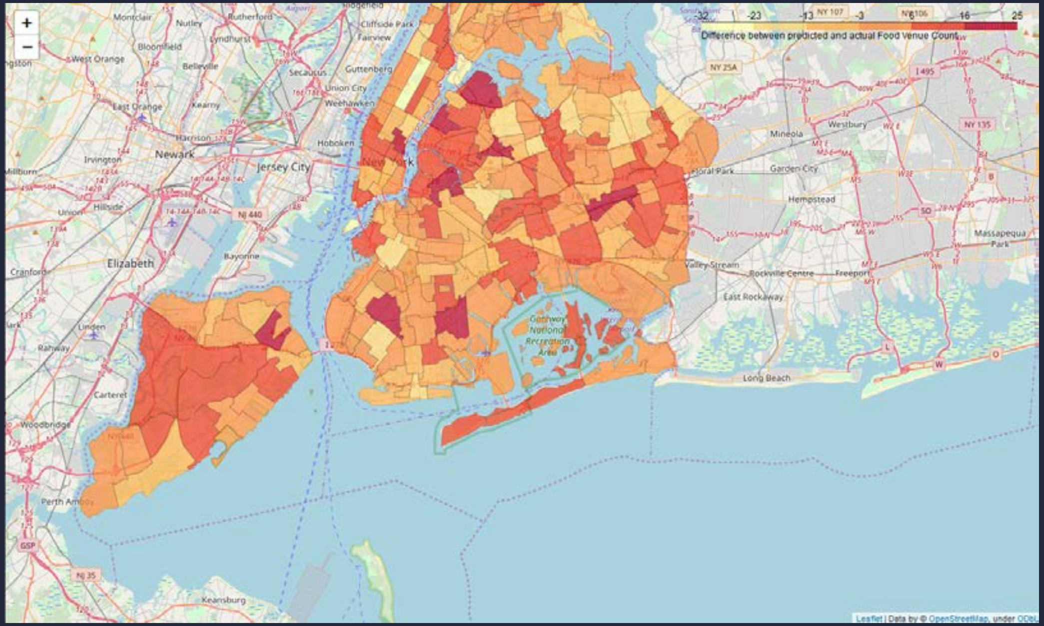
Project by Sanzhar Shaimerdenov for IBM Data science professional certificate

The main idea - use the difference between the **actual** and **predicted** food Venue count not as error, but as missed opportunity, probably very **under saturated** neighborhood we should look into

TOP 10 under saturated NTAs discovered

NTA Code	NTA Name	Borough	Food Venue Count	Food venues estimate	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:



Future directions - something to think about

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.

Thank you!