The relationship between Foursquare activities clusters, socio-economic factors and COVID-19 health data for US counties

ALEXEY USOLTSEV

05/05/2020

Goal of the study

- Use Foursquare activity patterns on US county level to classify counties in to clusters
- Explain cluster by countiy's socio-economic statistical data
- Use county clusters information and socio-economic variables to predict COVID-19 characteristics such as the growth rate
- Use data science methodology and Python to perform the analysis

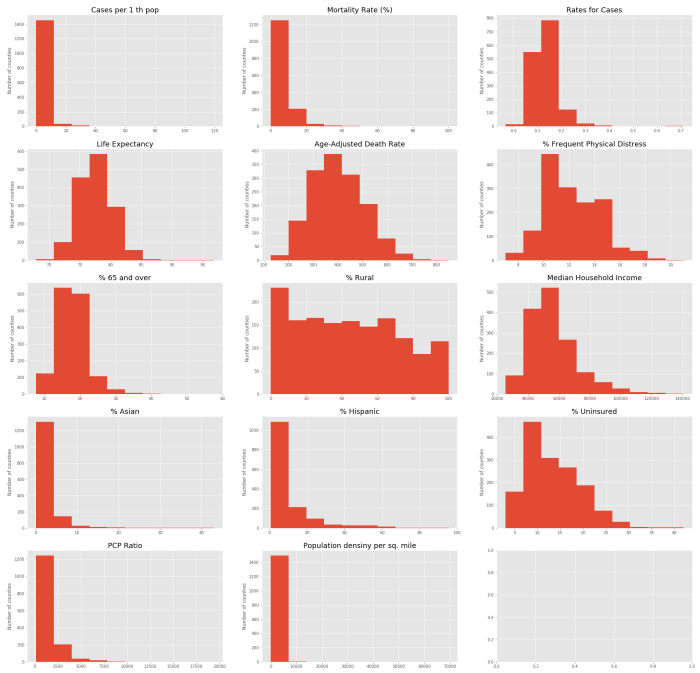
Data description

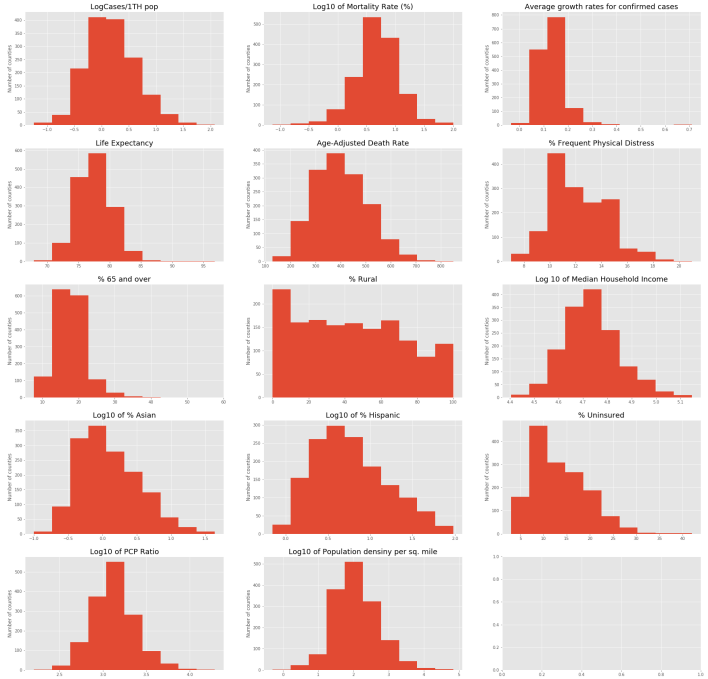
- COVID-19 data: *https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/*
- 2020 County Health Rankings National Data: https://www.countyhealthrankings.org
- Main features
  - COVID-19 data: mortality rate, average growth rate, number of cases per 1 thousand population
  - Socio-economic data: life expectancy, percentage of frequent physical distress, percentage of uninsured population, median household income, percentage of people over 65, percentage of Asian and Hispanic population, percentage of rural population
- Foursquare API datac

countyFIPS	object
County Name	object
State	object
stateFIPS	object
totalCases	int64
RateCases	float64
totalDeaths	int64
population	int64
Density	float64
cases/1TH pop	float64
Mortality Rate	float64
Latitude	float64
Longitude	float64
Life Expectancy	float64
Age-Adjusted Death Rate	float64
% Frequent Physical Distress	int64
% Uninsured	float64
PCP Ratio	float64
Median Household Income	float64
% 65 and over	float64
% Asian	float64
% Hispanic	float64
% Rural	float64

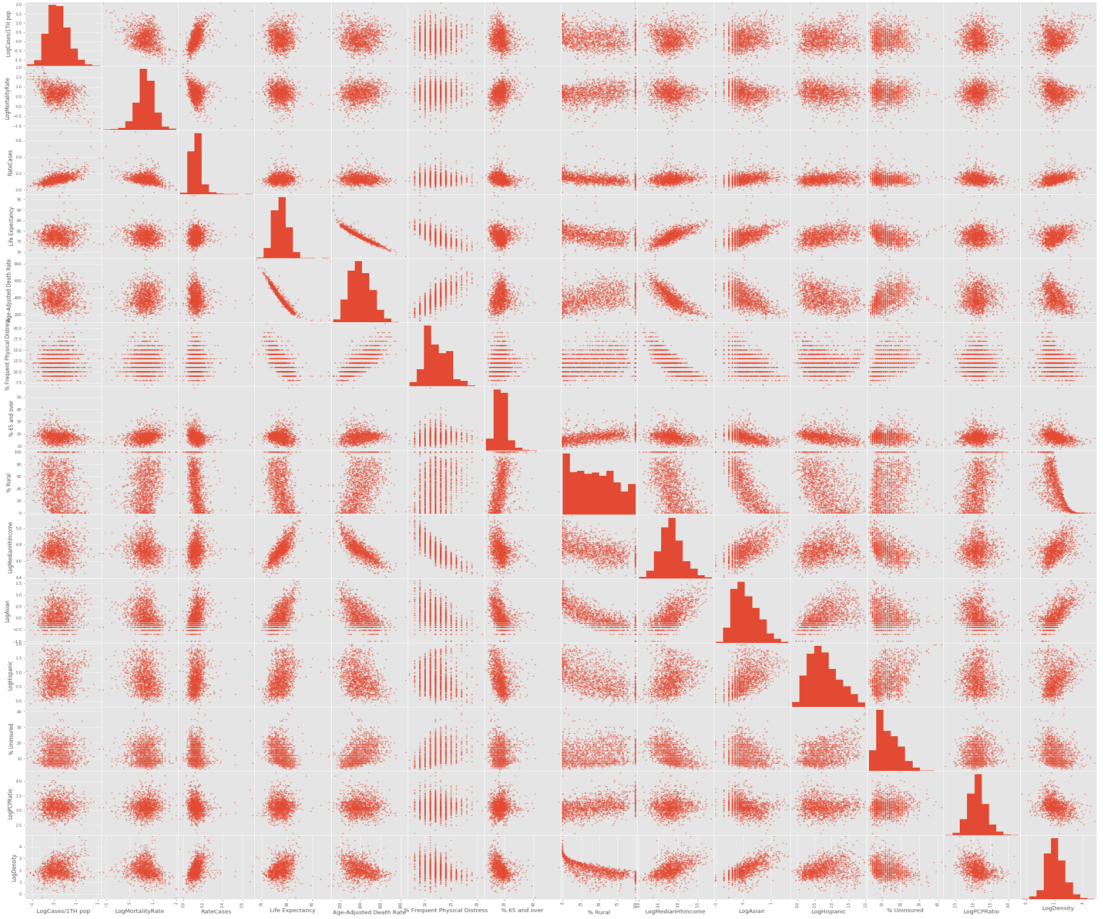
Exploratory data analysis

- Some non-normal variables were log-transformed for following linear model analysis





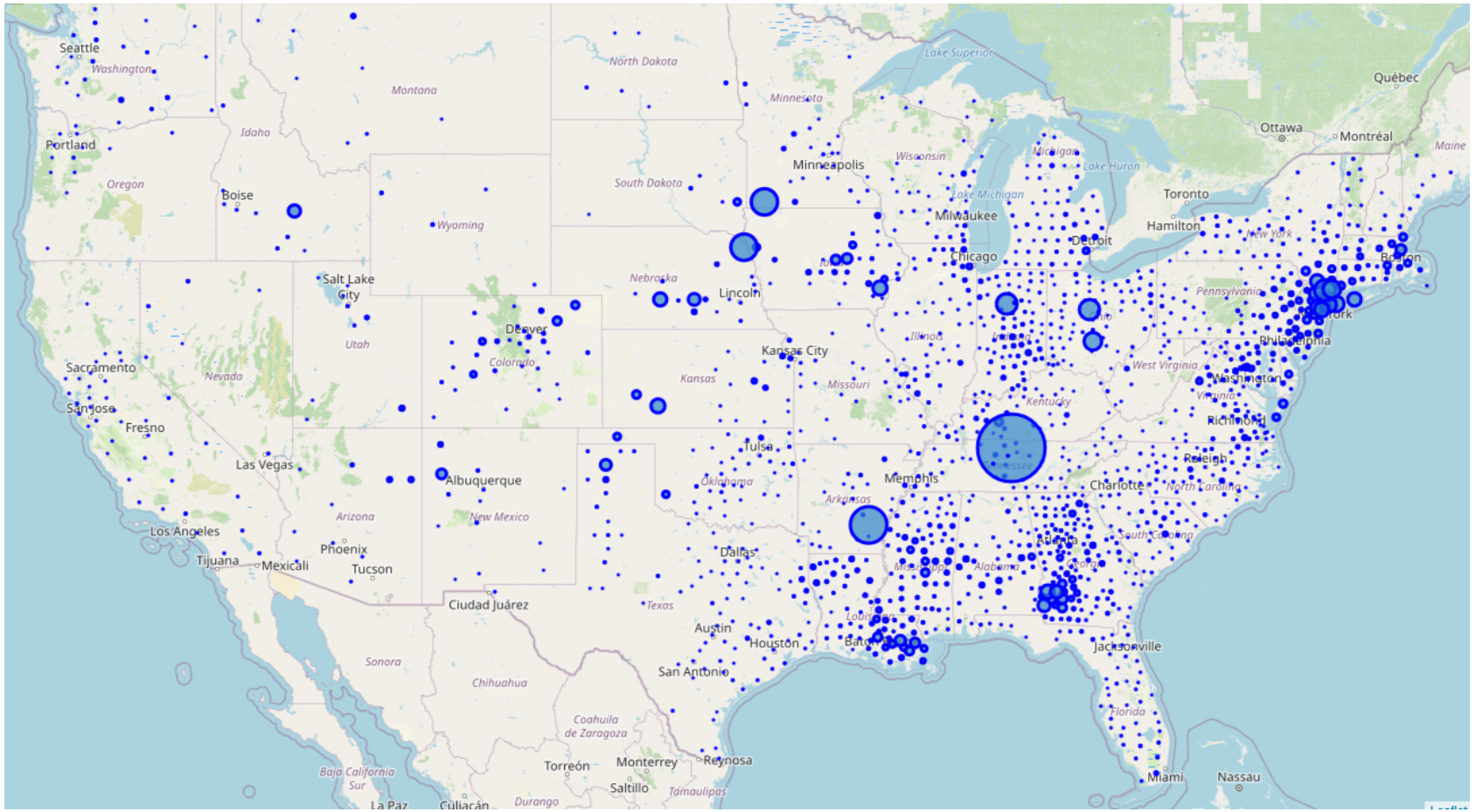
Exploratory data analysis



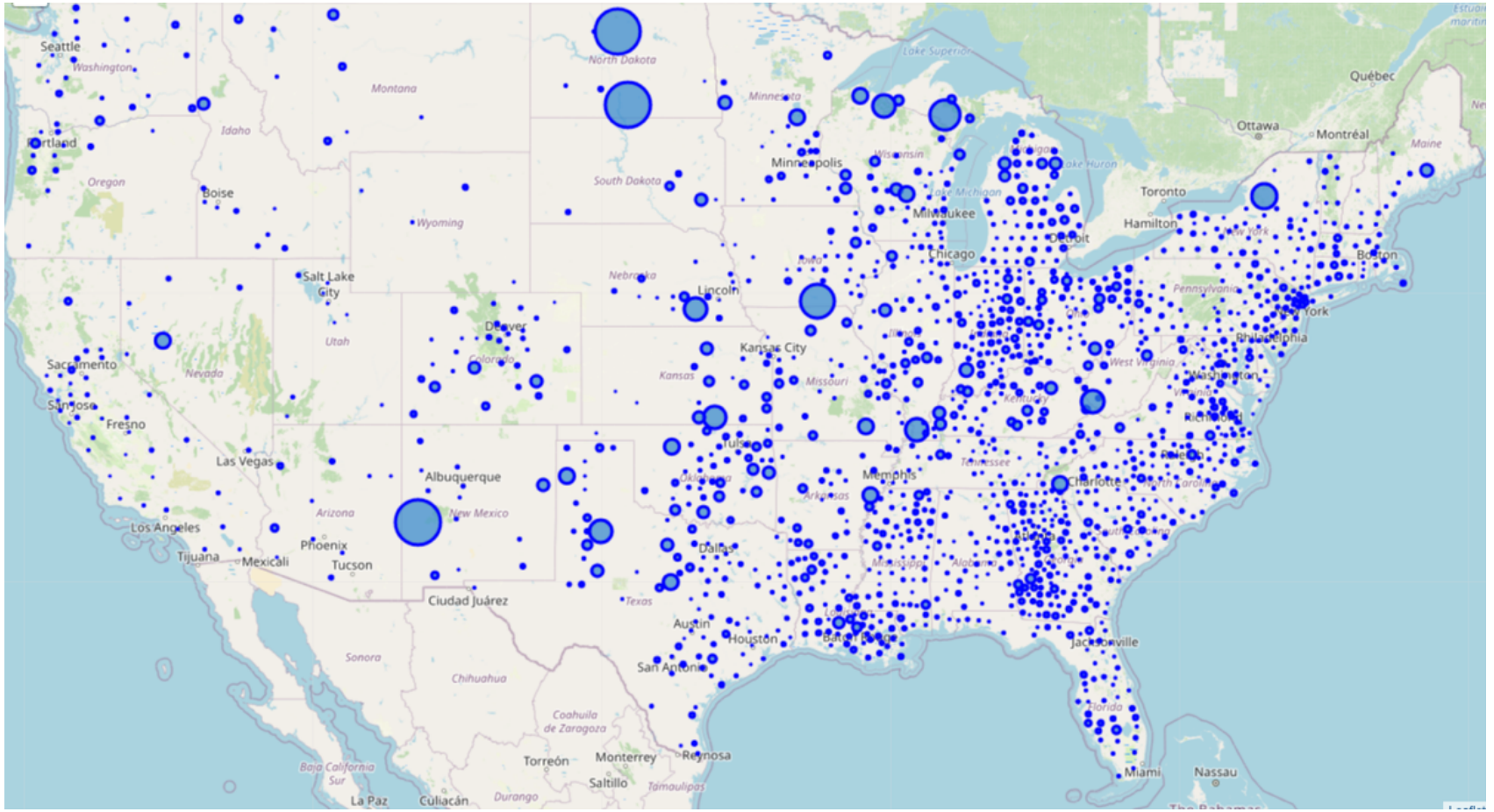
Most notable correlations:

- Life expectancy and Age-adjusted death rate
- Positive correlation between Life Expectancy and the Median Household Income.
- Positive correlation between the percentage of elderly people (65 and over) and the percentage of rural population.
- Negative relationship between population density and the percentage of rural population.

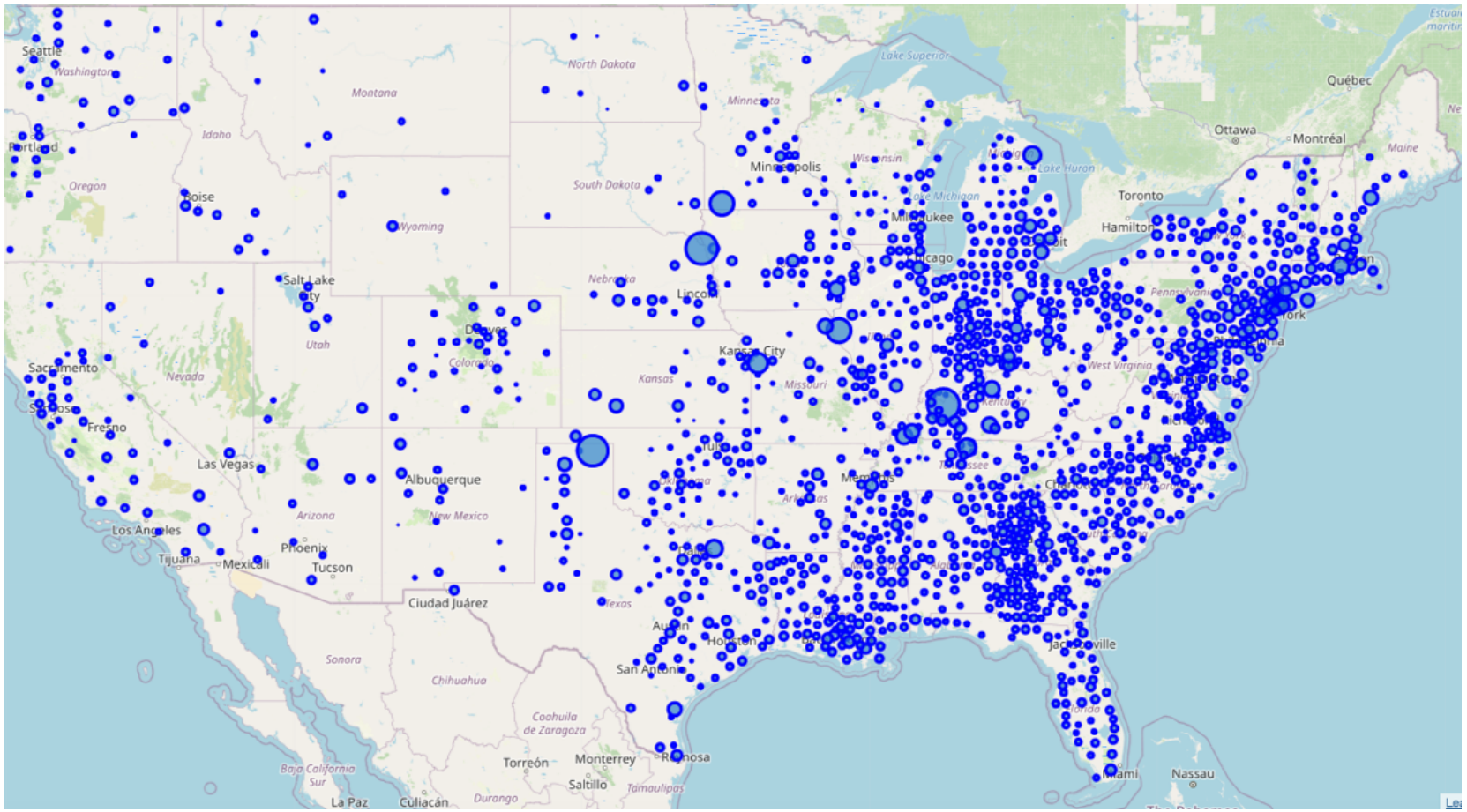
COVID-19 Cases per 1000 population



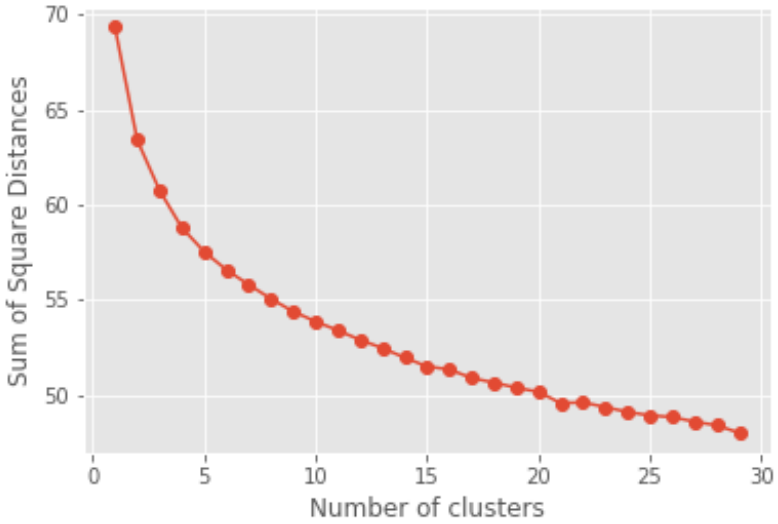
COVID-19 Mortality rate map

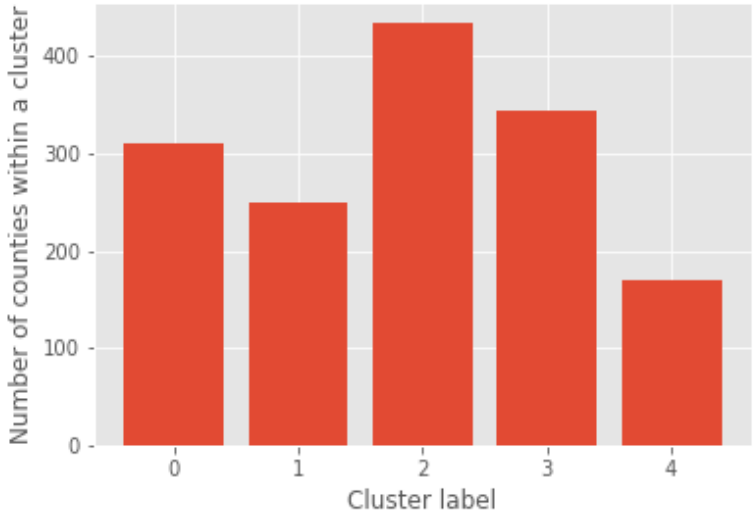


COVID-19 average growth rate



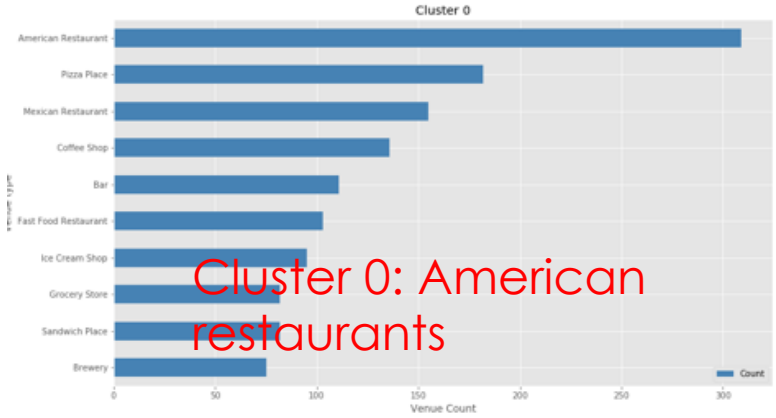
Foursquare API data clustering for US counties



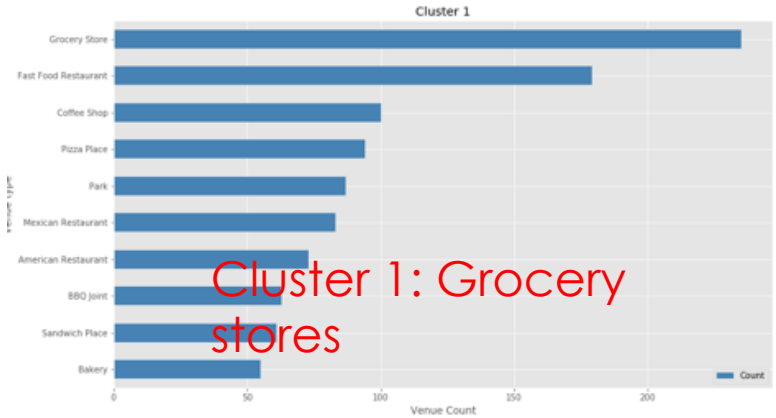


1. Total of 44,300 different venues accros the USA, 472 venues categories
2. Only top 20 venues are used for clustering
3. Scree plot is used to determine optimal number of cluster for K-Means algorithm
4. 5 clusters for classification

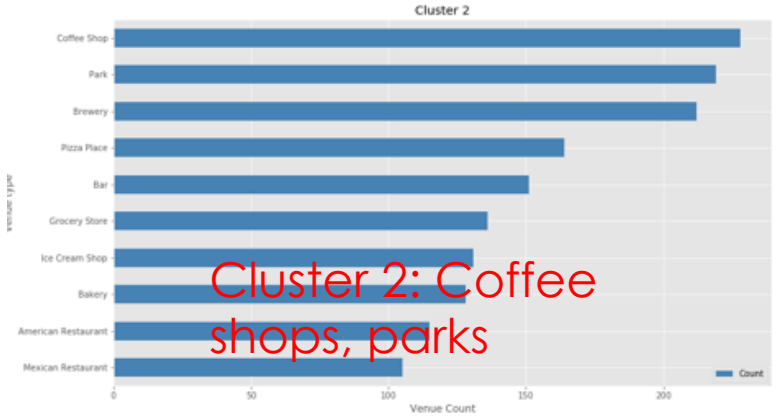
Clusters description based on popular venues



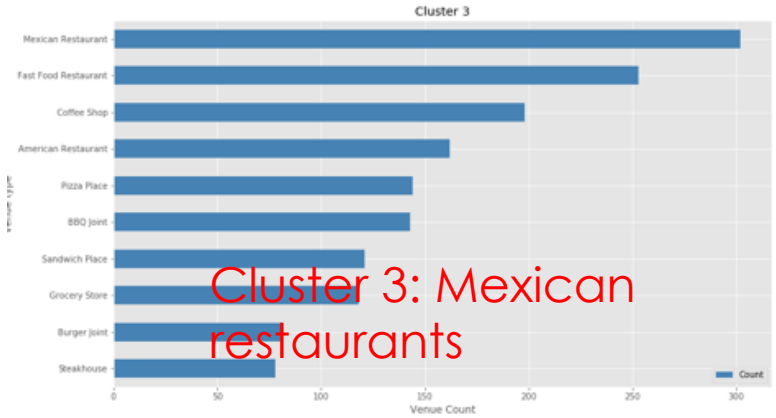
Cluster 0: American restaurants



Cluster 1: Grocery stores



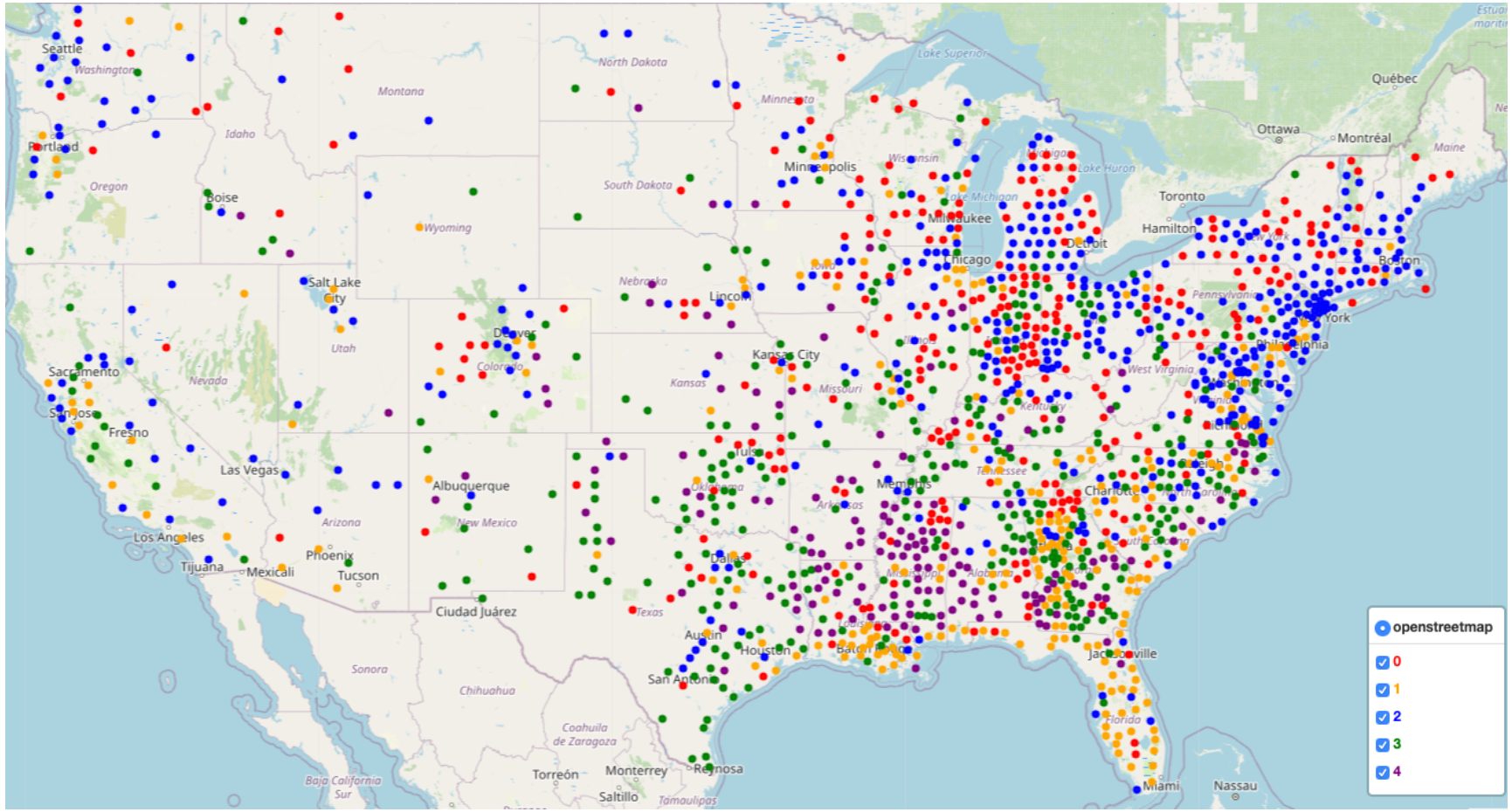
Cluster 2: Coffee shops, parks



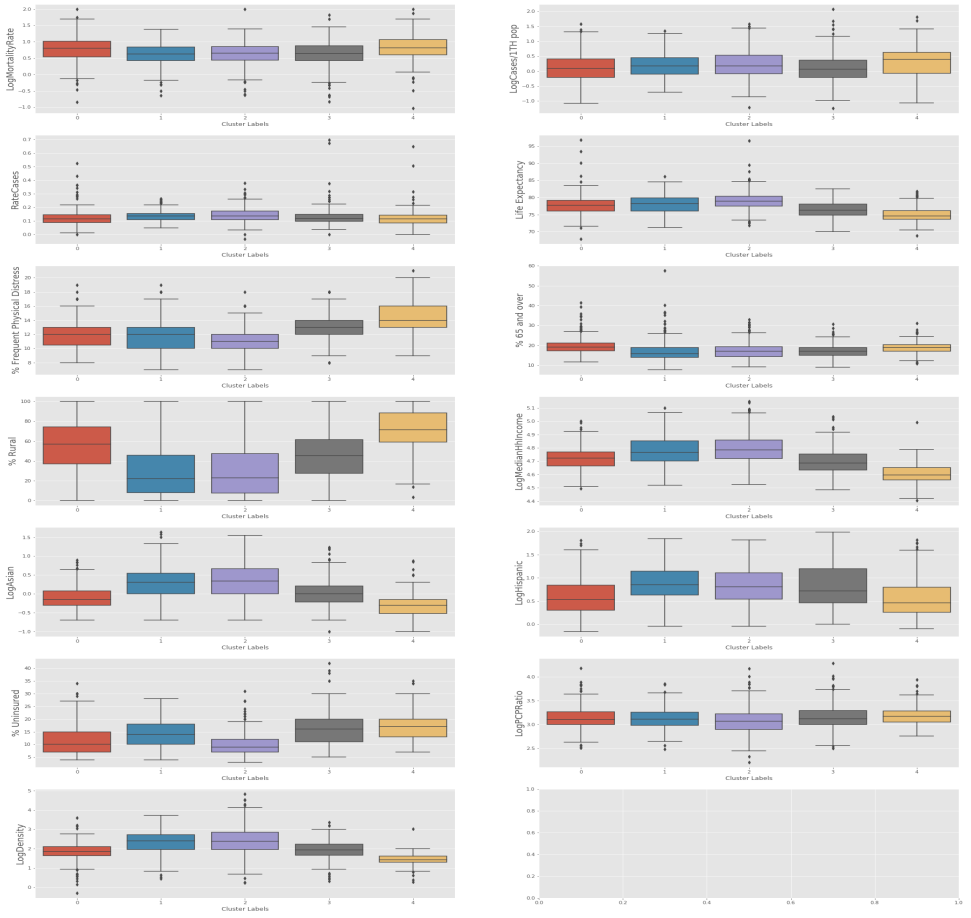
Cluster 3: Mexican restaurants

Cluster 4: Discount stores

Clusters on the map



Clusters and socio-economic factors



Observations:

- Cluster 4: rural, smallest life expectancy, high percentage of people over 65, the highest percentage of frequent physical distress, the lowest median household income, the smallest percentage of Asian population and the smallest population density. The median logarithm of Mortality Rate for COVID-19 is slightly greater
- Clusters 1 and 2: the least rural clusters, the highest median household income, the least percentage of physical distress, the highest life expectancy, the most percentage of Asian population and the population density. Both have significantly higher virus average growth rate

Multiple linear regerssion

OLS Regression Results

Dep. Variable:	y	R-squared:	0.163
Model:	OLS	Adj. R-squared:	0.156
Method:	Least Squares	F-statistic:	20.82
Date:	Wed, 06 May 2020	Prob (F-statistic):	7.61e-49
Time:	19:23:34	Log-Likelihood:	2341.9
No. Observations:	1507	AIC:	-4654.
Df Residuals:	1492	BIC:	-4574.
Df Model:	14		
Covariance Type:	nonrobust		

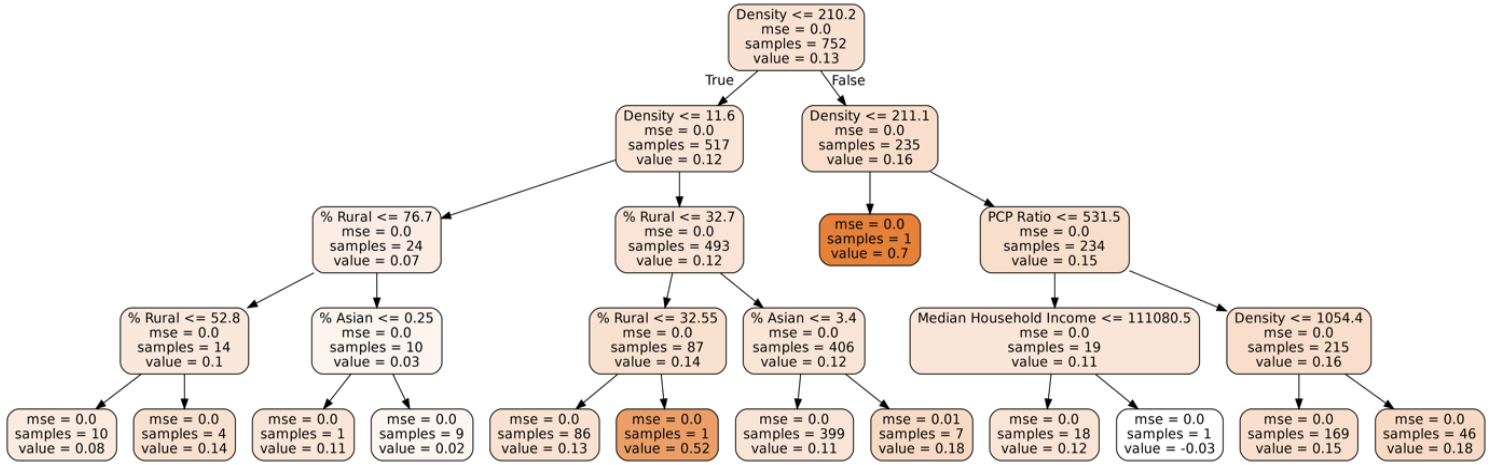
	coef	std err	t	P>|t|	[0.025	0.975]
Const	0.2105	0.107	1.968	0.049	0.001	0.420
Life Expectancy	-0.0004	0.001	-0.472	0.637	-0.002	0.001
% Frequent Physical Distress	0.0015	0.001	1.296	0.195	-0.001	0.004
% 65 and over	-0.0014	0.000	-3.608	0.000	-0.002	-0.001
% Rural	-0.0002	8.69e-05	-2.096	0.036	-0.000	-1.17e-05
LogMedianHhIncome	-0.0331	0.027	-1.247	0.213	-0.085	0.019
% Uninsured	-0.0005	0.000	-1.694	0.090	-0.001	8.54e-05
LogAsian	0.0029	0.006	0.488	0.626	-0.009	0.014
LogHispanic	0.0119	0.005	2.453	0.014	0.002	0.021
LogPCPRatio	0.0080	0.006	1.302	0.193	-0.004	0.020
LogDensity	0.0256	0.004	7.282	0.000	0.019	0.032
Cluster Labels_1	0.0365	0.022	1.656	0.098	-0.007	0.080
Cluster Labels_4	0.0500	0.021	2.337	0.020	0.008	0.092
Cluster Labels_3	0.0382	0.021	1.793	0.073	-0.004	0.080
Cluster Labels_0	0.0442	0.021	2.058	0.040	0.002	0.086
Cluster Labels_2	0.0416	0.022	1.908	0.057	-0.001	0.084

Omnibus:	1256.895	Durbin-Watson:	1.899
Prob(Omnibus):	0.000	Jarque-Bera (JB):	56133.243
Skew:	3.588	Prob(JB):	0.00
Kurtosis:	32.025	Cond. No.	9.53e+16

Observations:

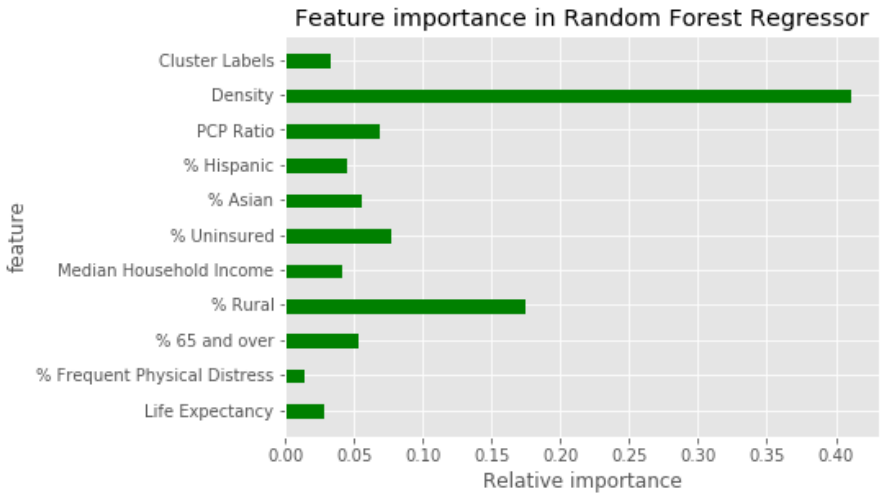
- Adjusted R-squared: 0.156
- Significant factors:
  - logarithm of population density (positive effect)
  - percentage of population over 65 (negative effect could be aliased with density),
  - log-percentage of Hispanic population
  - clusters 4 and 0

Random Forest Regression



Observations:

- Adjusted R-squared: 0.205
- Significant factors:
  - Population density
  - Percentage of rural population



Conclusion

- Significant relationship between COVID-19 average growth rate and two main factors: population density and a percentage of a rural population
- Social distancing makes sense for a disease control
- Nonparametric classification algorithms based on decision trees perform better that linear models for complex datasets
- Cluster classification showed to be significant for COVID-19 average growth rate prediction.