Applied Data Science Capstone Project: The relationship between Foursquare activities clusters, socio-economic factors and COVID-19 health data for US counties

Alexey Usoltsev

05/05/2020

1. Introduction

Currently, a lot of effort is being put in researching different factors affecting the spread of COVID-19 around certain areas. In this project I would use Foursquare API to find clusters of US counties based on their top popular venues and then attempt to relate those clusters with COVID-19 growth factor. I would also use county-level socio-economic data to understand clustering and building a prediction model for COVID-19 growth factor.

The results may be of interest to any party related to disease control.

The part of the analysis related to classification of social activity type of clusters and their relationships with socio-economic factors may be used for business development purposes.

The project should not be considered as a comprehensive research of any kind. The analysis performed only for demonstration of data science principles and workflow using Python.

2. Data

The project uses several datasets:

1. COVID-19 data sets are available on *https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/* under Creative Commons license. *covid_confirmed_usafacts.csv* contains daily cumulative coronavirus cases on counties level.

   covid_deaths_usafacts.csv contains daily cumulative deaths due to the coronavirus on counties level.

   covid_county_population_usafacts.csv contains 2019 Census counties population estimates with added population density feature (people per square mile of land area) All county files contain FIPS codes, a standard geographic identifier, to make it easier to combine data with other data sets. Following features from these datasets will be computed and used as main response variables for prediction:

   - Cases per 1 thousand population: number of confirmed COVID-19 cases per 1 thousand people as of 04.05.2020
   - Mortality rate (%): the number of deaths due to coronavirus divided by the number of confirmed COVID-19 cases, expressed in percentages as of 04.05.2020
   - Overall average growth rate for confirmed COVID-19 cases: calculated as an average percentage change for cumulative daily totals of confirmed case.

2. 2020 County Health Rankings National Data (available on https://www.countyhealthrankings.org/) Socio-economic features from this county-level data set will be used for explanation of county clusters based on most popular Foursquare venues. Also, at the end of the project I will use these features in an attempt

to build a prediction model for COVID-19 growth factor. Features from this dataset will be used as explanatory variables for prediction and classification. For simplicity, only following set of factors from County health ranking are used:

- Life Expectancy: Average number of years a person can expect to live;
- Age-Adjusted Death Rate: Number of deaths among residents under age 75 per 100,000 population (age-adjusted).
- % Frequent Physical Distress: Percentage of adults reporting 14 or more days of poor physical health per month;
- % Uninsured: Percentage of adults under age 65 without health insurance;
- PCP Ratio: Ratio of population to primary care providers other than physicians;
- Median Household Income: The income where half of households in a county earn more and half of households earn less;
- % 65 and over: Percentage of population ages 65 and older;
- % Asian: Percentage of population that is Asian;
- % Hispanic: Percentage of population that is Hispanic;
- % Rural: Percentage of population living in a rural area.

3. Centroid coordinates for every FIPS code were fetched from datasets available at Kaggle and were merged with other data
4. Foursquare data. With county coordinates available, counties were analysed using Foursquare data. Only request for most popular venues were used for the purpose of classification.

3. Methodology

3.1 Exploratory Data Analysis

The complete dataset contains 1508 individual counties (rows) and variables (columns) shown on a picture below. At this point all response and explanatory features are numerical. Only counties with non-zero mortality rate were used for the study. There are relatively few missing values, therefore they are dropped from the following analysis.

```
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
```

Figure 1: Dataset variables

Since all variables are numerical it is useful to visualize and understand distribution of every feature. Histograms are shown at Figure 2 and it is clear that some variables have extremely skewed distribution. It was decided to log-transform following variables: cases per 1 thousand population, Mortality rate, Median household income, % of Asian population, % of Hispanic population, PCP ration, population density.

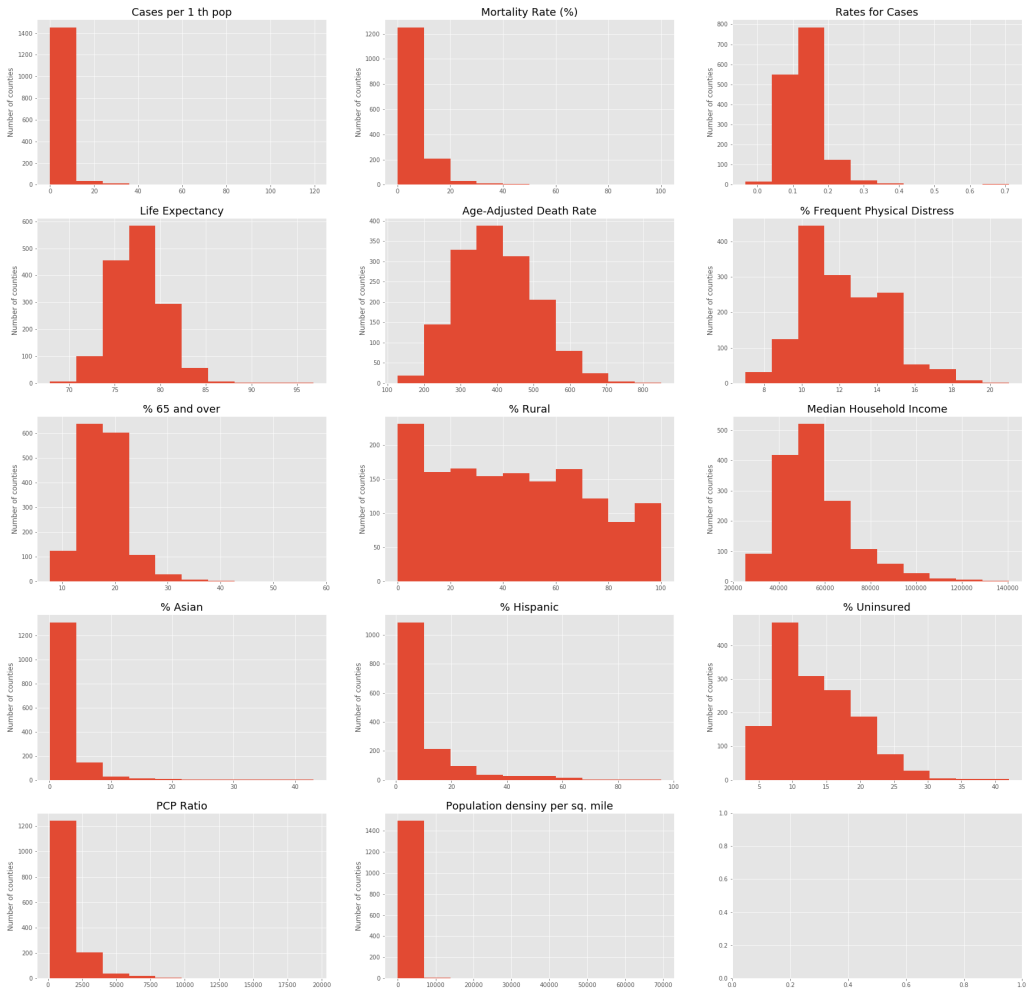

Figure 2: Histograms for original variables

Histograms for transformed variables are shown on Figure 3. It's clear that data normality assumption, required for some types of analyses, can be satisfied after log-transformation.

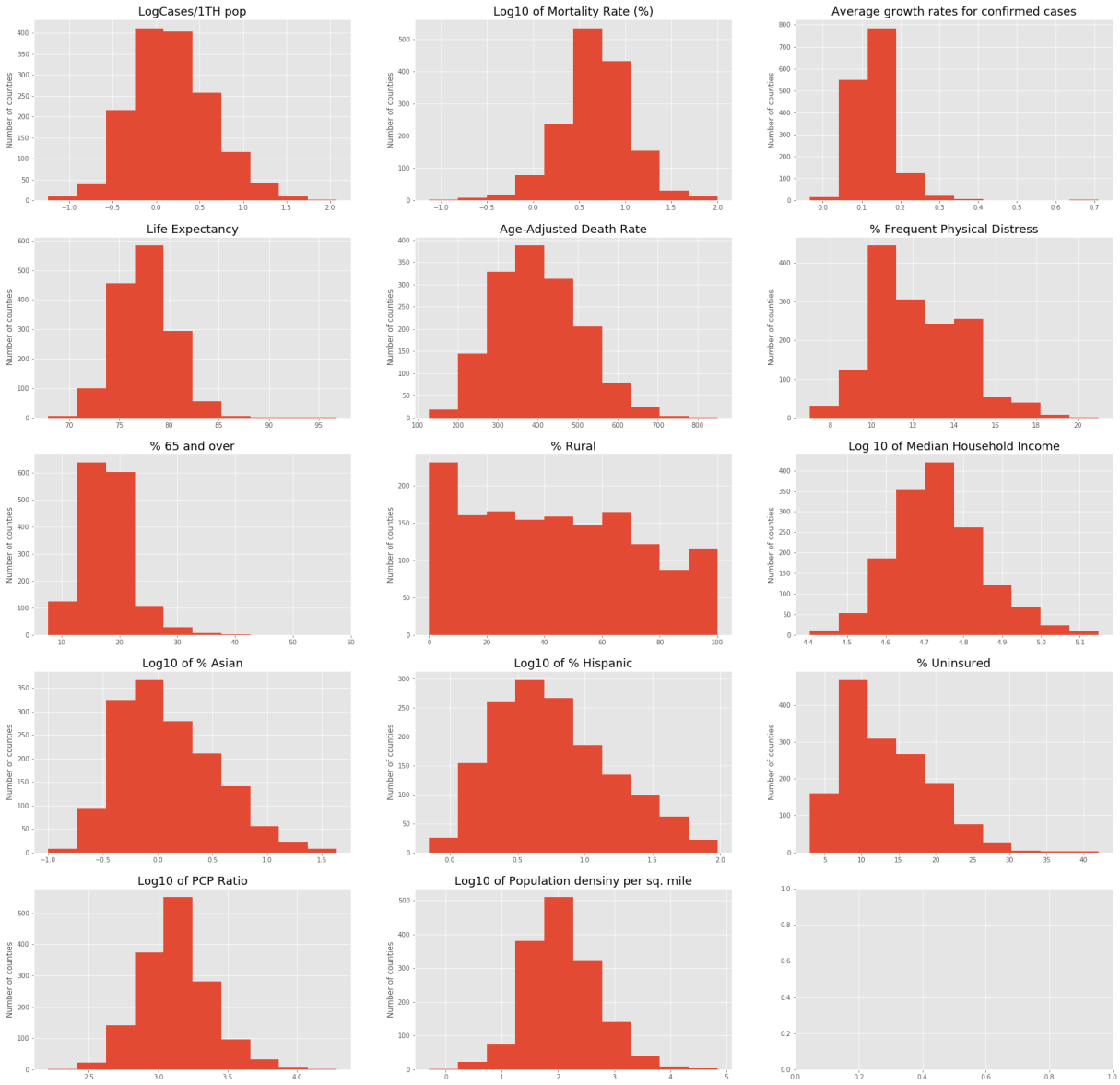

Figure 3: Histograms for transformed data

In order to see if any of variables are related, scatterplot matrix was produced. The result is shown on Figure 4. There are no obvious relationships between dependent and explanatory variables. Some positive correlation is present between the mortality rate and the percentage of rural population, slight negative correlation is evident between epidemic growth rate and the percentage of rural population; whereas growth rate is positively correlated with the median household income, the percentage of Asian population, the percentage of Hispanic population and the overall population density.

Also, there are many relationships between exploratory variables. The most notable are correlations are:

– Life expectancy and Age-adjusted death rate. I will use only the former one for the analysis because of high correlation.

– 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.

All these observations suggest that we may expect some multicollinearity when linear model is fitted.

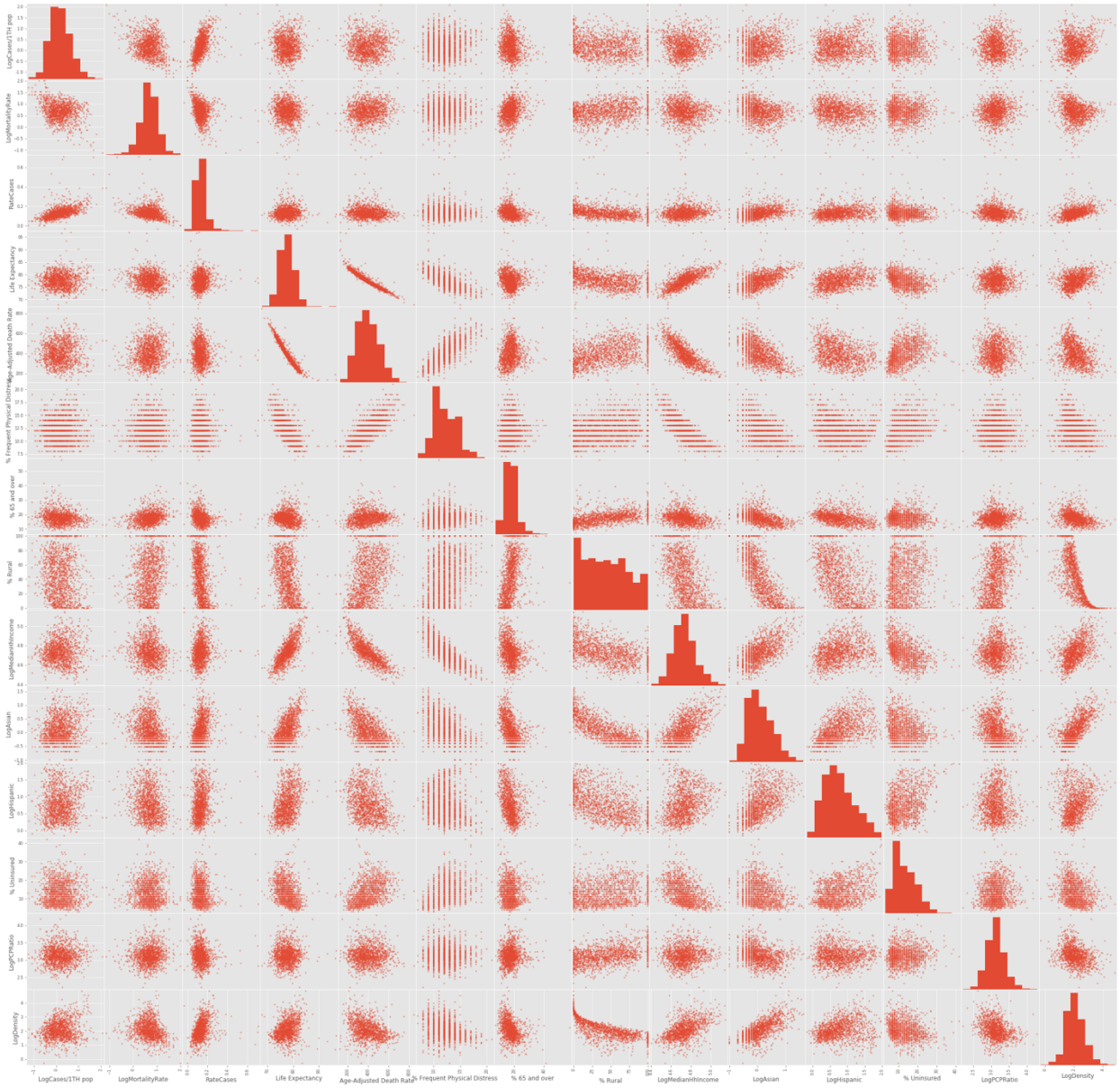

Figure 4: Scatter Matrix Plot

3.2 Response variables visualization

Because our dataset instances have geospatial nature, it is possible to visualize response variable on the map using Folium library in Python.

Fugures 5-7 show bubble plots for three response variables superimposed on the geographical map.

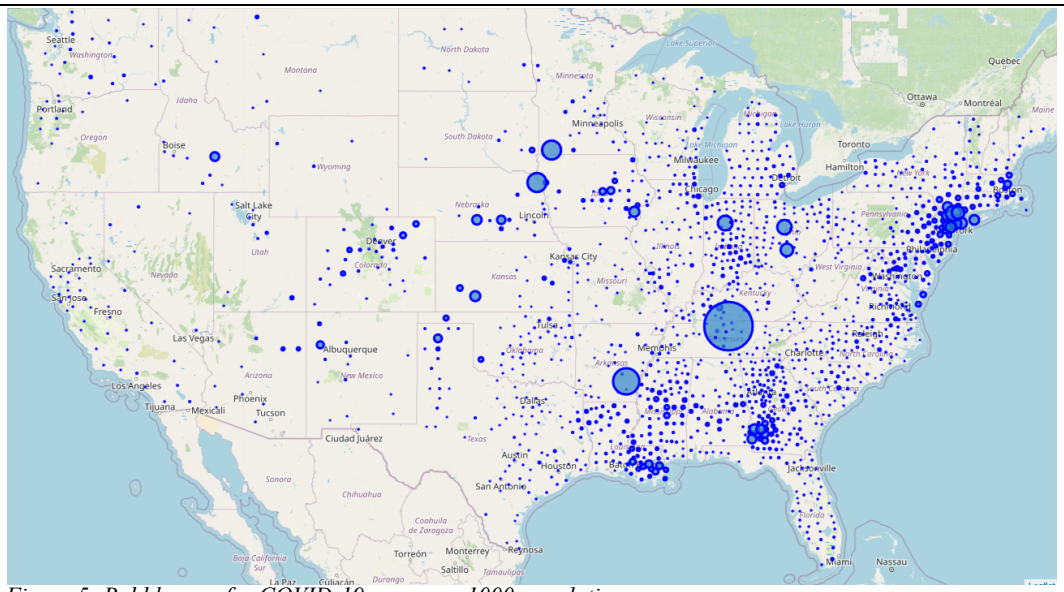

Figure 5: Bubble map for COVID-19 cases per 1000 population

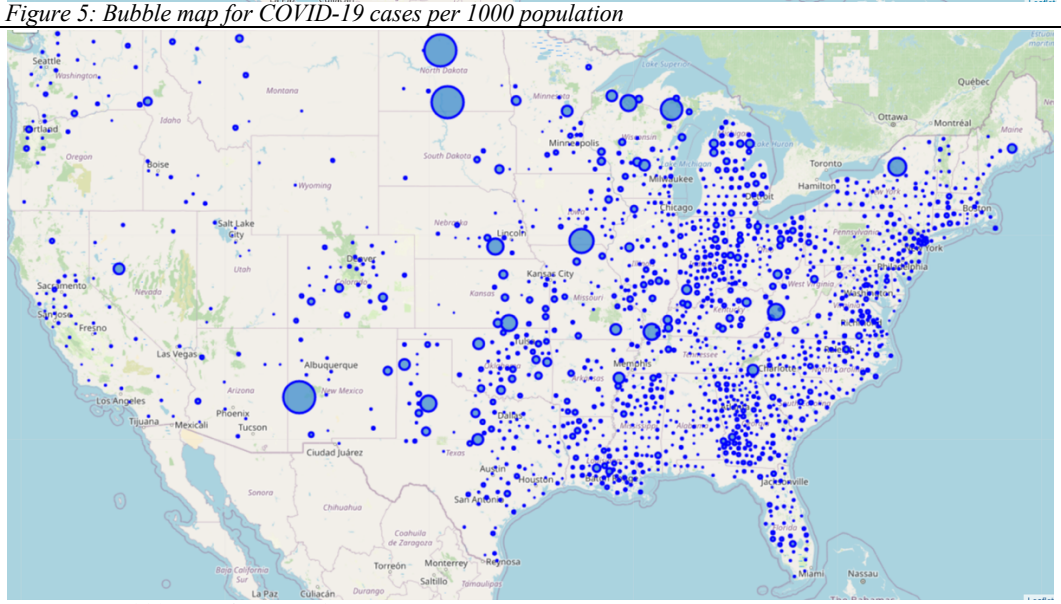

Figure 6: Bubble map for Mortality rate

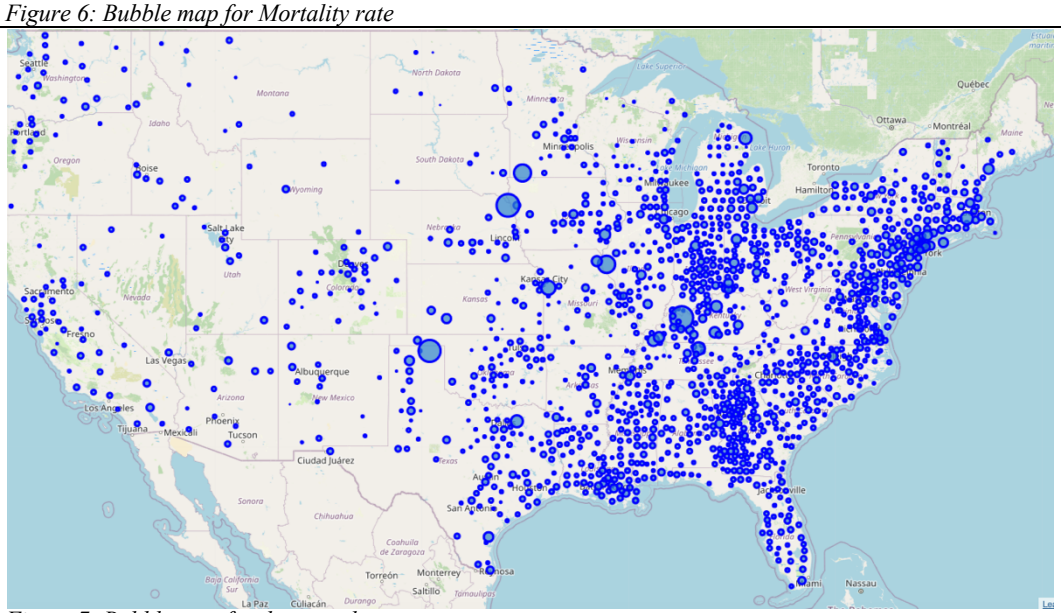

Figure 7: Bubble map for the growth rate

3.3 Foursquare analysis and counties clustering

We use the following method for counties clustering:

1. Find at most 30 venues within 40,000 meters from county's centroid coordinates based on Foursquare usage information. The limitation is due to personal Foursquare API. As a result we have 44,300 different venues. Venues coverage for all counties is presented on Figure 8.
2. Assign each venue to a venue category and then calculate the mean number of venues within each category for every county. All 44,300 venues are grouped to 472 venue categories.
3. Pick top 20 venue categories for each county for classification and use **K-means clustering** algorithm to divide counties to clusters. Counties within one cluster will have similar types of venues.
4. In order to provide the number of clusters to the algorithm, we use a scree-plot and "elbow method". The method calculates sum of square distances between cluster centroids and observations for increasing of clusters. It suggests to use such a number of clusters where a plot has its breaking point. Figure 9 shows a scree-plot for counties clustering. There is no obvious breaking point in the graph but at around five clusters sum-of-squared distances is significantly reduced. I will use ***k=5*** for K-means clustering algorithm. As it can be seen on Figure 10, there is no big disparity between cluster counts when 5 clusters are used. Cluster 2 contains 434 counties, while Cluster 4 has 169 counties.

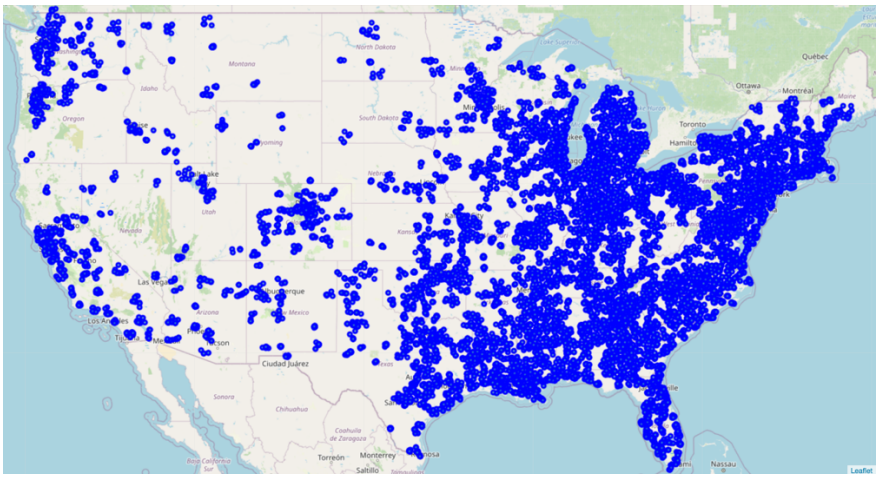

Figure 8: Venues distribution

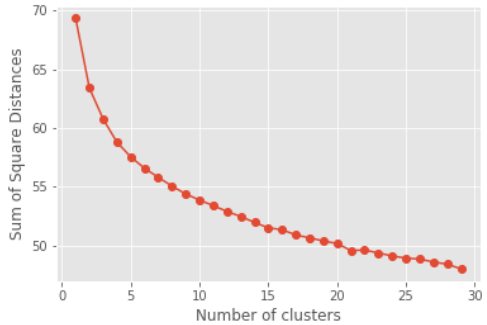

Figure 9: Scree-plot

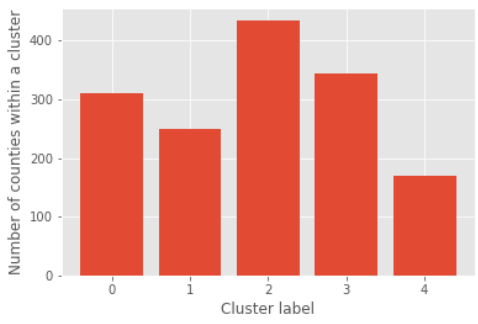

Figure 10: Cluster counts

Let's try to interpret counties clusters solely based on most popular venue categories. Later on, in the results section, we will also describe clusters based on socio-economical characteristics. Figure 11 shows frequencies of occurrence of a venue type in a list of top 20 venues. Figure 12 positions each county and their cluster attributes on the map.

Cluster 0: almost all counties within this cluster contain **American restaurants** in top-20 venues list, pizza places, Mexican restaurants, coffee shops and bars follow the list. Counties within the cluster are scattered away from large cities and are concentrated mostly in the northeastern region of the country.

Cluster 1: **grocery store** is the most frequent venue type in this cluster. Fast food places, coffee shops, pizza places and parks are mentioned less frequently among Foursquare users. Counties within this cluster are concentrated in Southern region in Louisiana, Georgia and Florida

Cluster 2: **Coffee shops, parks, breweries** are equally popular in this cluster. Also pizza places and bars are included in top-5 most frequent venues. The map on Figure 12 reveals that counties in this cluster are concentrated closer to the center of large metropolitan areas in Northeastern states.

Cluster 3: Among 344 counties belonging to this cluster, around 300 of them have a **Mexican restaurant** in top-20 venues list. Another popular places are fast food restaurants, coffee shops, American restaurants and pizza places. Cluster 3 is very similar to Cluster 0 but with Mexican restaurants being more popular that American restaurants. Looking at the map on Figure 12 it can be seen that counties in this cluster are concentrated more to the south relative to counties in Cluster 0

Cluster 4: out of 168 counties 166 of them contain **discount store** as a most frequently mentioned venue. Top-5 also includes fast food restaurants, sandwich places, pizza places and even gas stations. It can be seen from the map on Figure 12 that counties in this cluster are concentrated almost exclusively in the rural areas of southern states of Georgia, Alabama, Mississippi and Louisiana.

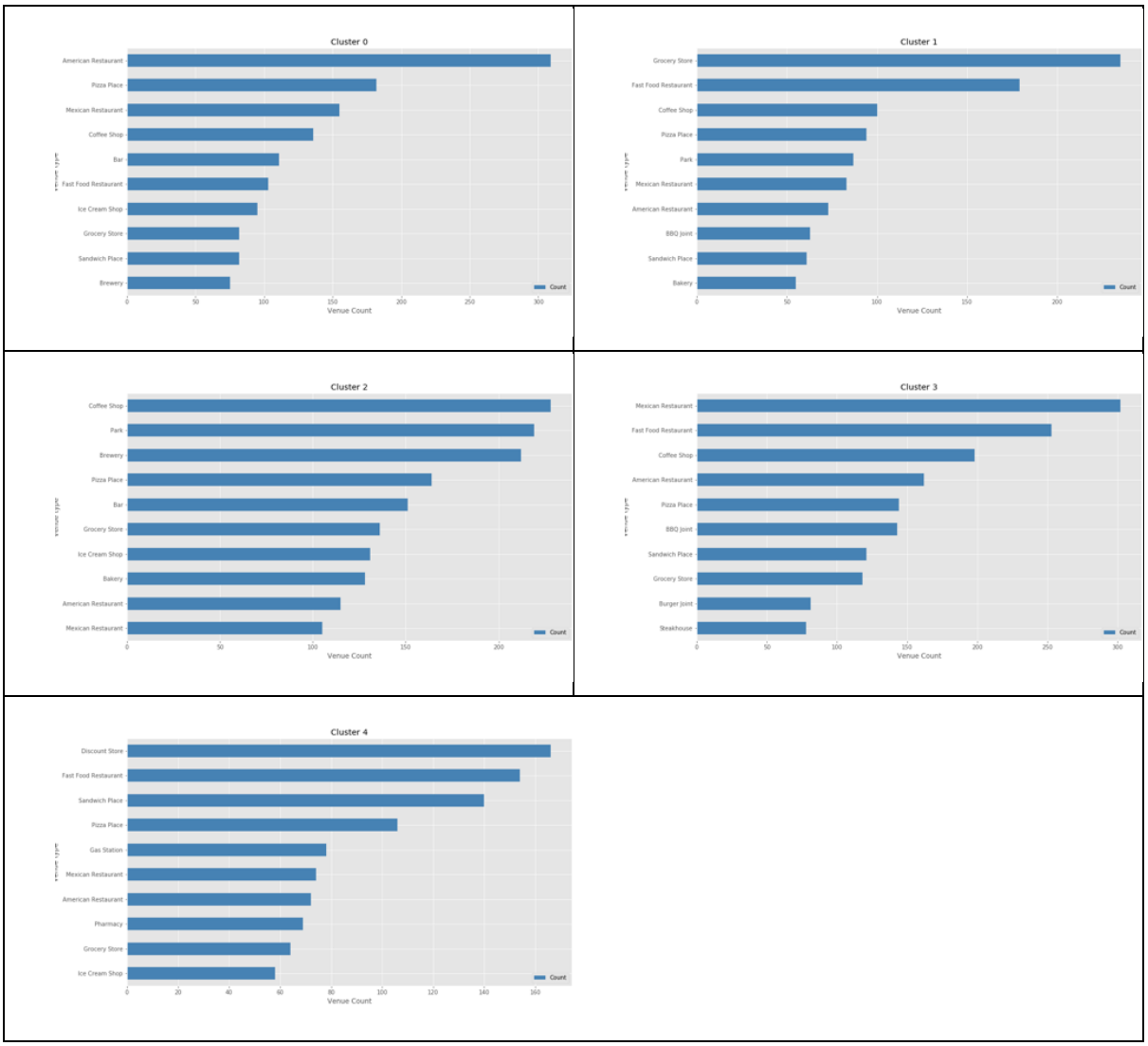

Figure 11: Bar plot of top venue types for each cluster

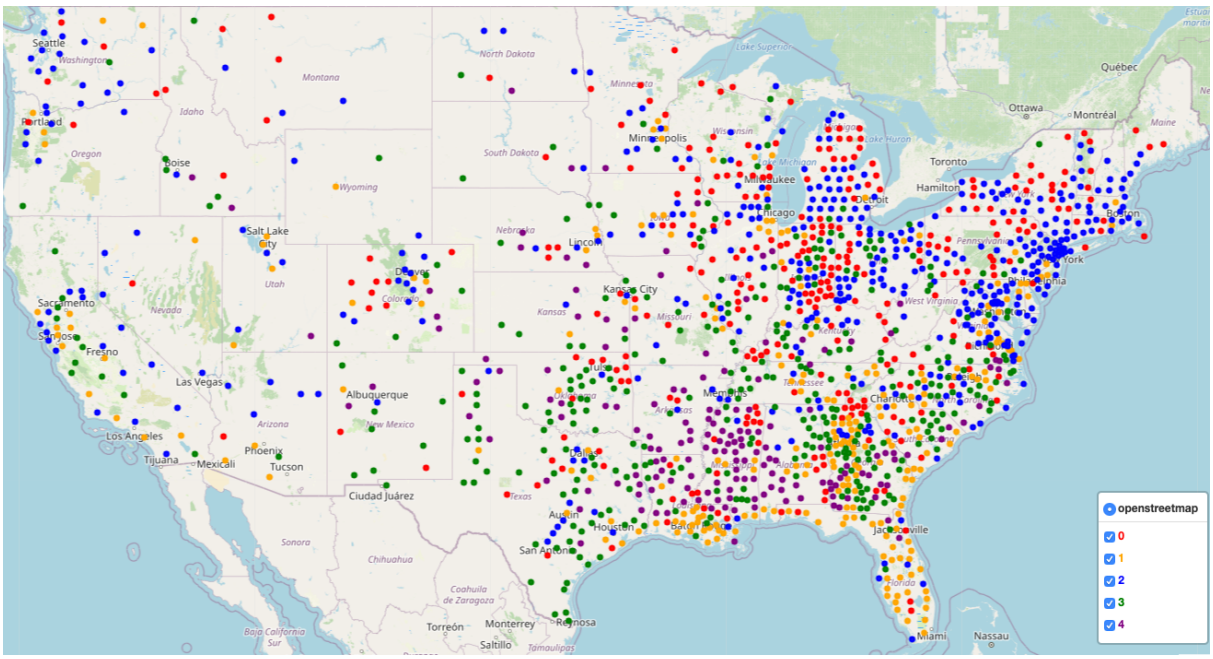

Figure 12: Counties with their cluster attribute coded by color

4. Results

In the results section we further examine counties clusters for relationships with socio-economic factors and COVID-19 using simple boxplots visualizations and one-way analysis of variance analysis and multiple comparison method, then a linear regression model for predicting COVID-19 variables based on the rest of the features will be built and, finally, we compare it with random forest regression model for COVID-19 metrics.

4.1 Analysis of county clusters based on socio-economic features

Figure 13 shows boxplots for all transformed continuous variables separated based on county cluster number. Figures 14-16 gives ANOVA pairwise comparison tables for average growth rates, log-mortality rate, and log-number of cases per 1 thousand population, respectively. Following observations can be made based on graphics and analysis of variance pairwise comparison:

1. Cluster 4 is the most rural cluster among all. At the same time it has the 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 than the same statistics for clusters 1, 2 and 3. The significance is confirmed by pairwise comparison on Figure 15.
2. Cluster 2 together with Cluster 1 are the least rural clusters. They have 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. According to socio-economic data, clusters 1 and 2 are very similar, but the only difference is between percentage of uninsured population. In cluster 1 this percentage is significantly larger than in cluster 2. Figures 14-16 show that with respect to COVID-19 metrics, Clusters 1 and 2 cannot be considered different between each other, however they both have significantly higher average growth rate than other clusters. This may be explained by the population density.

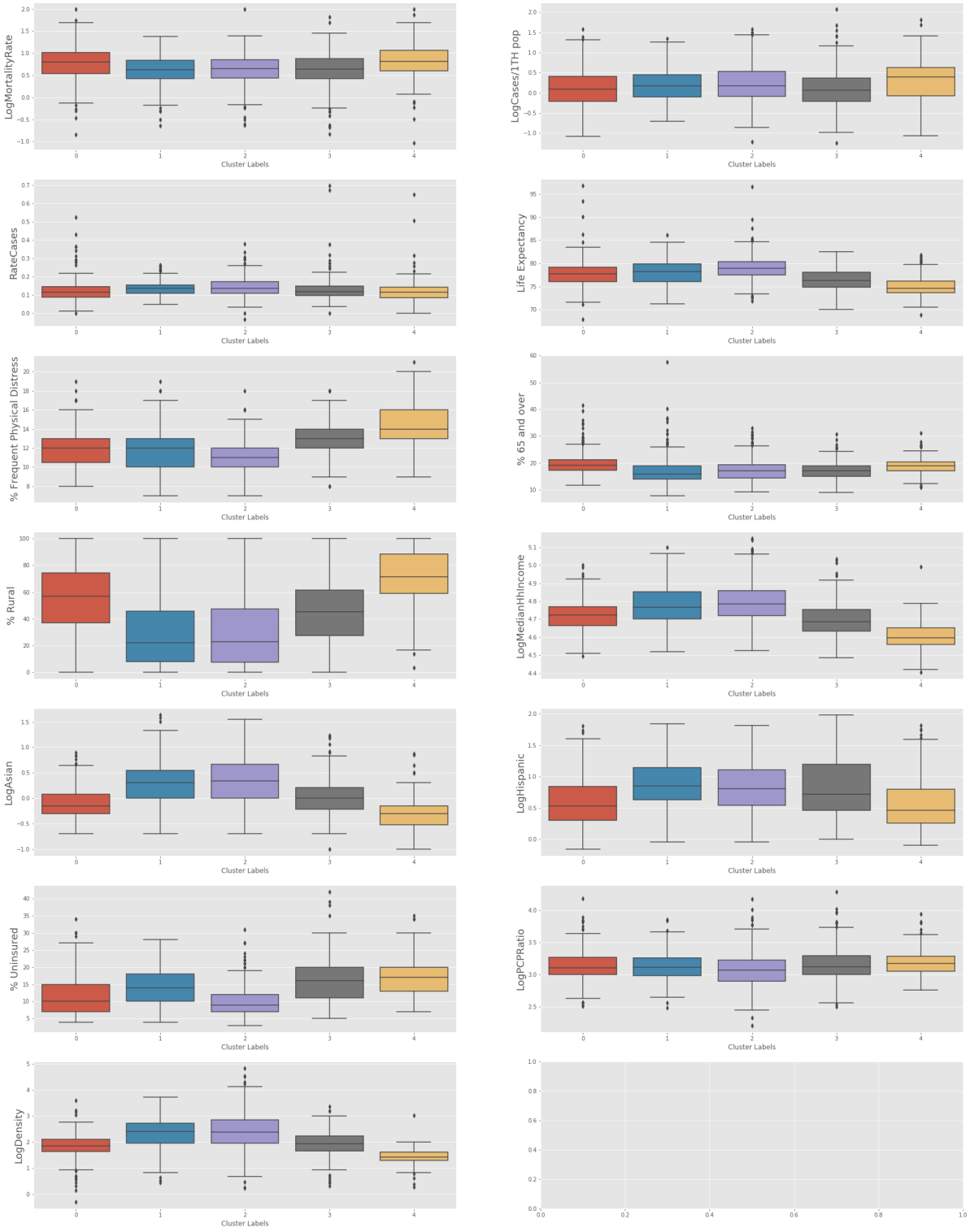

Figure 13: Boxplots of continuous variables and counties clusters

	coef	std err	t	P>|t|	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
1-0	0.014505	0.004702	3.084780	2.074192e-03	0.005281	0.023728	0.014429	True
2-0	0.020519	0.004112	4.989435	6.760554e-07	0.012452	0.028586	0.000007	True
3-0	0.005641	0.004331	1.302384	1.929850e-01	-0.002855	0.014137	0.528567	False
4-0	-0.001147	0.005300	-0.216380	8.287210e-01	-0.011543	0.009250	0.828721	False
2-1	0.006014	0.004395	1.368435	1.713807e-01	-0.002607	0.014635	0.528567	False
3-1	-0.008864	0.004600	-1.926723	5.420260e-02	-0.017888	0.000160	0.243184	False
4-1	-0.015651	0.005522	-2.834264	4.654595e-03	-0.026483	-0.004819	0.027605	True
3-2	-0.014878	0.003996	-3.723306	2.038963e-04	-0.022716	-0.007040	0.001630	True
4-2	-0.021666	0.005030	-4.307497	1.758240e-05	-0.031532	-0.011800	0.000158	True
4-3	-0.006788	0.005210	-1.302792	1.928455e-01	-0.017008	0.003432	0.528567	False

Figure 14: Multiple comparison for average COVID-19 growth rate between clusters

	coef	std err	t	P>|t|	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
1-0	-0.170404	0.031061	-5.486118	4.814685e-08	-0.231332	-0.109477	4.814684e-07	True
2-0	-0.143487	0.027167	-5.281734	1.467593e-07	-0.196776	-0.090199	9.699801e-07	True
3-0	-0.154181	0.028612	-5.388735	8.227464e-08	-0.210305	-0.098058	6.581970e-07	True
4-0	0.027974	0.035012	0.798970	4.244344e-01	-0.040704	0.096651	8.258718e-01	False
2-1	0.026917	0.029033	0.927102	3.540226e-01	-0.030033	0.083867	8.258718e-01	False
3-1	0.016223	0.030390	0.533823	5.935427e-01	-0.043388	0.075834	8.347925e-01	False
4-1	0.198378	0.036479	5.438078	6.278120e-08	0.126822	0.269934	5.650307e-07	True
3-2	-0.010694	0.026397	-0.405129	6.854401e-01	-0.062472	0.041084	8.347925e-01	False
4-2	0.171461	0.033226	5.160387	2.793406e-07	0.106286	0.236636	1.396702e-06	True
4-3	0.182155	0.034418	5.292439	1.385686e-07	0.114643	0.249667	9.699801e-07	True

Figure 15: Multiple comparison for log-mortality rate between clusters

	coef	std err	t	P>|t|	Conf. Int. Low	Conf. Int. Upp.	pvalue-hs	reject-hs
1-0	0.113841	0.039003	2.918765	3.566540e-03	0.037335	0.190348	0.017706	True
2-0	0.138579	0.034113	4.062319	5.109680e-05	0.071664	0.205494	0.000358	True
3-0	-0.003995	0.035928	-0.111204	9.114695e-01	-0.074469	0.066479	0.911469	False
4-0	0.221809	0.043965	5.045171	5.082930e-07	0.135570	0.308047	0.000005	True
2-1	0.024737	0.036457	0.678534	4.975375e-01	-0.046775	0.096250	0.747531	False
3-1	-0.117837	0.038161	-3.087925	2.052503e-03	-0.192690	-0.042983	0.012252	True
4-1	0.107967	0.045807	2.356998	1.855110e-02	0.018115	0.197820	0.072165	False
3-2	-0.142574	0.033146	-4.301354	1.807090e-05	-0.207592	-0.077556	0.000145	True
4-2	0.083230	0.041722	1.994853	4.623935e-02	0.001390	0.165070	0.132403	False
4-3	0.225804	0.043219	5.224693	1.989442e-07	0.141029	0.310580	0.000002	True

Figure 16: Multiple comparison for log-number of cases between clusters

4.2 Multiple linear regression for average COVID-19 growth rate

In this section we will build a linear model (multiple linear regression) for predicting average COVID-19 growth rate based on social activity cluster from the section 3.3 and other socio-economic features.

In order to properly deal with a categorical variable "Cluster labels", one-hot-encoding was performed. Since linear models require data to be normally distributed, we used log-transformed variables, identified during exploratory data analysis.

LinearRegression function of *"sklearn"* package was used to find the best fit for the average COVID-19 growth rate. Model summary is presented in Figure 17.

To judge the model performance, we will use adjusted R-squared metric which is **0.15** in case of multiple linear regression.

Analysis also show that only following variables appear to be significant with p-value < 0.05: logarithm of population density (positive effect), percentage of population over 65 (negative effect could be aliased with density), log-percentage of Hispanic population, and clusters 4 and 0.

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

Figure 17: Multiple linear regression model for the average COVID-19 growth rate

4.3 Random Forest Regression model for the average COVID-19 growth rate

The RandomForestRegressor function of "sklearn" package was used for this task. The dataset was divided into the training set and the test set in proportion 80% to 20%, respectively. The maximal depth of the tree was chosen to be 4 in order to simplify interpretation and for visualization purposes.

Random forest regression model with these parameters fits data with the adjusted R-squared value of **0.205** which is about 30% better than ordinary least squares model.

Figure 18 shows relative importance of features used for the model. The population density is the most important factor for the growth rate, with the percentage of rural population being second important. One of the trees in a forest, shown on Figure 19, confirms the importance of factors.

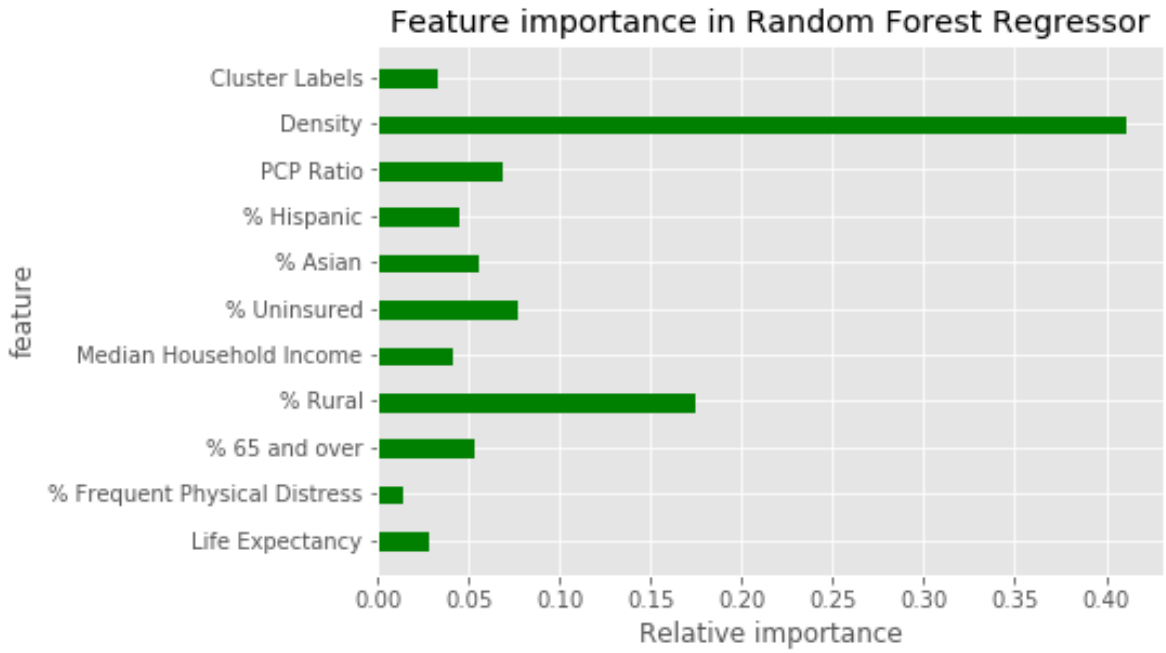

Figure 18: Features importance for random forest regression model for average COVID-19 growth rate

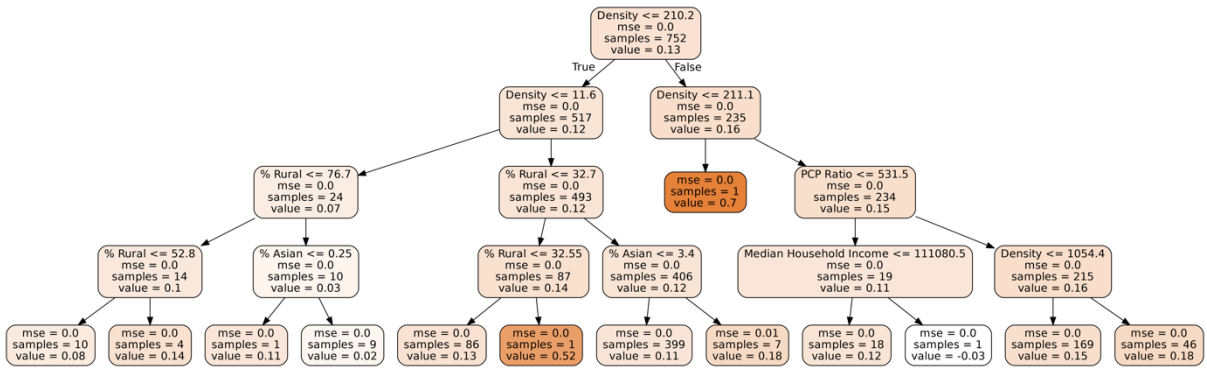

Figure 19: One of the trees of the random forest regression model for the average COVID-19 growth rate

5. Discussion

As a result of different analyses in can be concluded that two most important factors for the average COVID-19 growth rate are population density and percentage of rural population, with the former factor positively correlated to the growth rate, and the latter factor negatively correlated. Therefore, activities related to social distancing seem reasonable for slowing down the virus.

The slightly higher mortality rate in the rural cluster was not explained by other socio-economic factors based on simple boxplots. As a further analysis it may be necessary to perform a random forest regression for the mortality rate to get a better insight.

Also, it may be interesting to get a better insight into classification of Foursquare clusters based of socio-economic data.

Another improvement in the future analysis will be made for model refinement by random forest hyperparameters selection.

6. Results

The main result gained from the study is a relationship between the COVID-19 average growth rate and two main factors: population density and a percentage of a rural population by different types of data analysis models. The relationship is summarized on Figure 20.

Another important observation is that nonparametric classification algorithms based on decision trees perform better that linear models for complex datasets.

As an intermediate result a reasonable classification of US counties based on Foursquare activities patterns was performed. The classification may be used for business development purposes. Also cluster classification showed to be significant for COVID-19 average growth rate prediction.

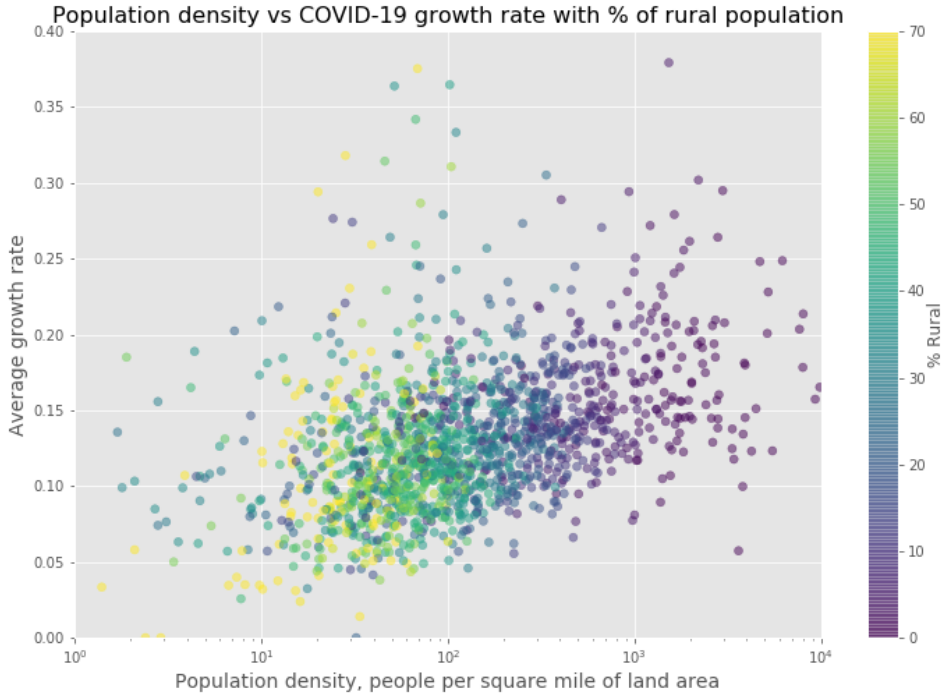

Figure 20: Scatter plot of population density and average COVID-19 growth rate with percentage of rural population color-coded