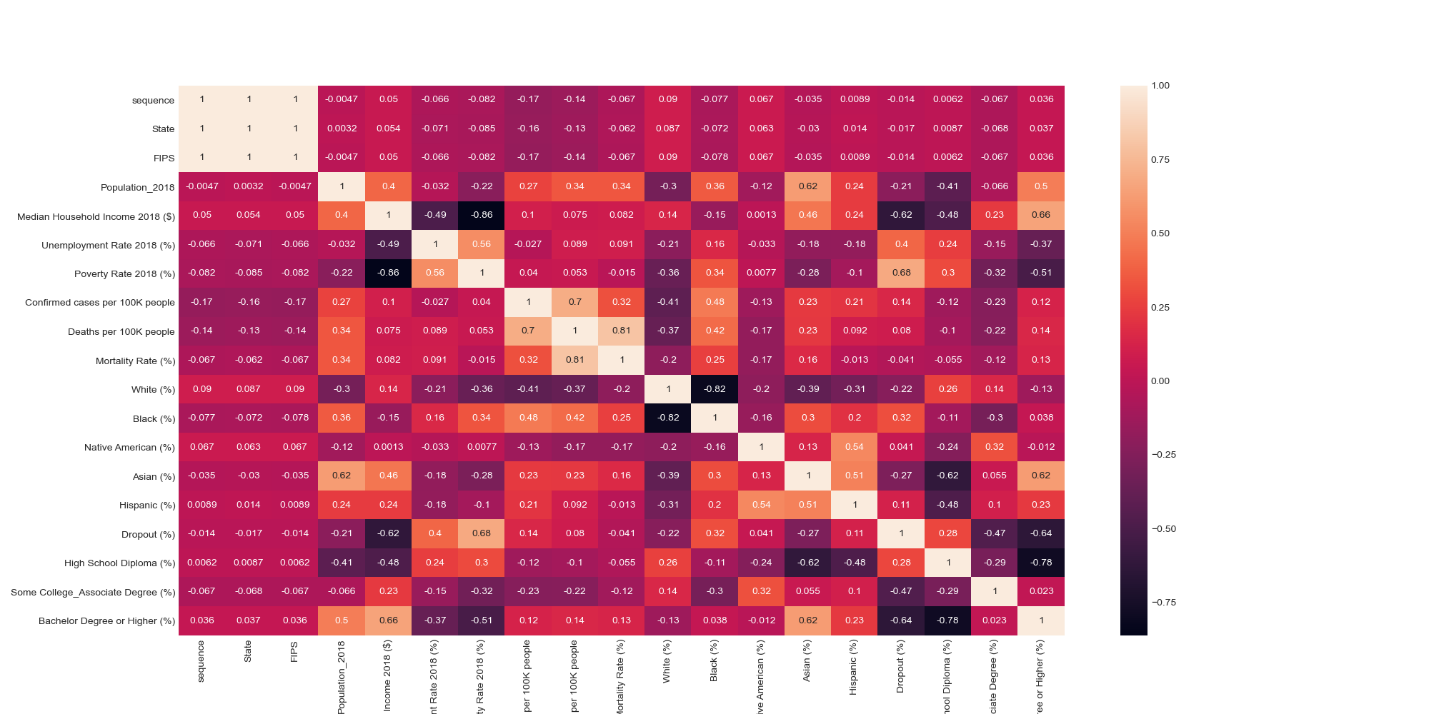
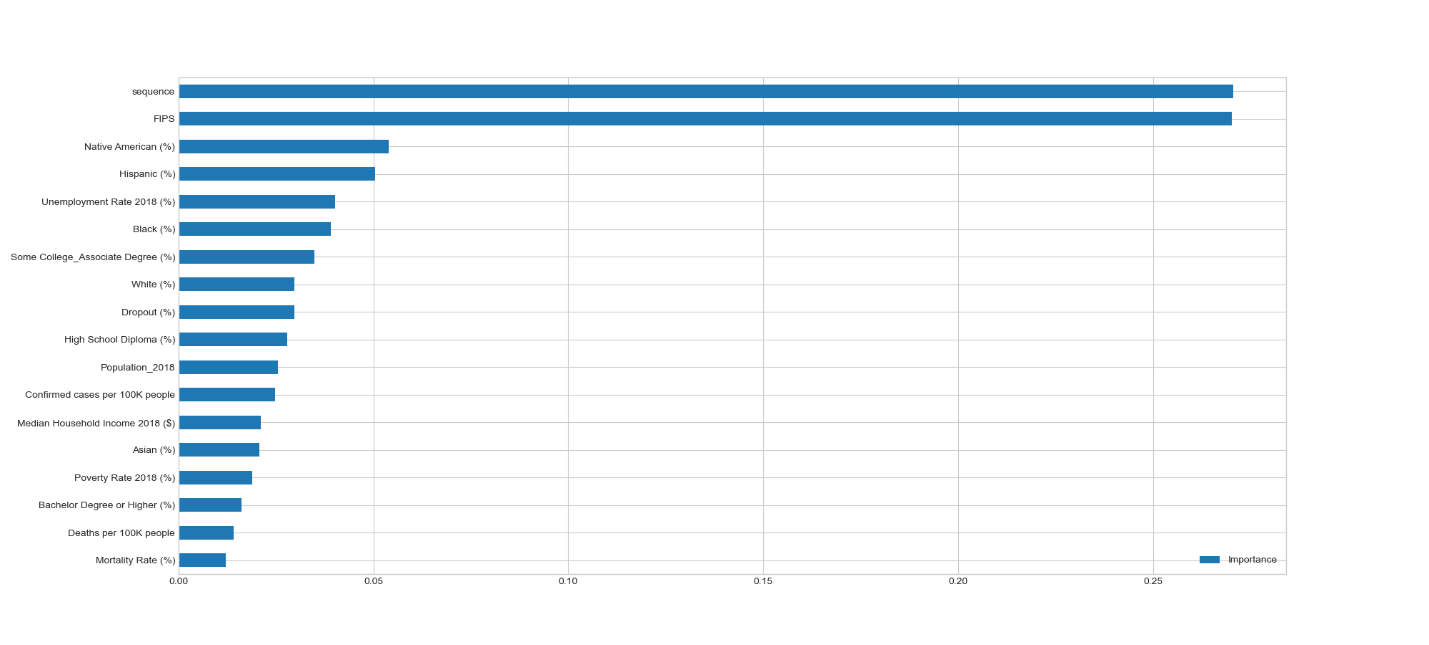


The above visualizations show the distribution of each feature or variable.

Feature Encoding: Machine Learning algorithms perform Linear Algebra on Matrices, which means all features need to have numeric values. The process of converting Categorical Features into values is called Encoding. Let's perform both One-Hot and Label encoding.

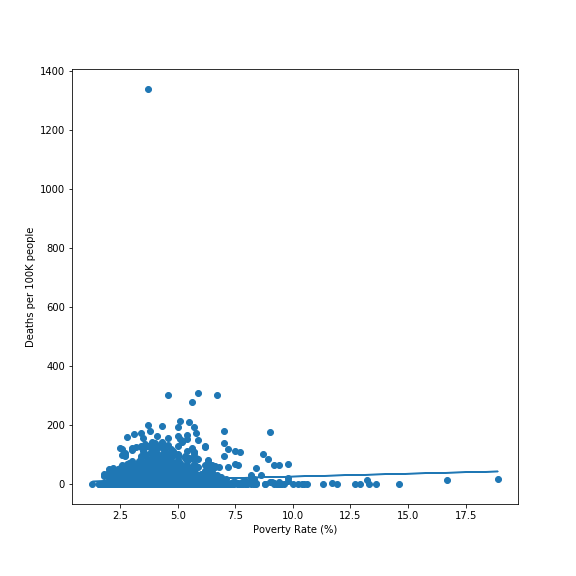


Feature Importance: Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single feature, designed to split the dataset into two so that similar response values end up in the same set. The measure based on which the (locally) optimal condition is chosen is called impurity. When training a tree, it can be computed by how much each feature decreases the weighted impurity in a tree. For a forest, the impurity decrease from each feature that can be averaged and the features that are ranked according to this measure. This is the feature importance measure exposed in sklearn’s Random Forest implementations.

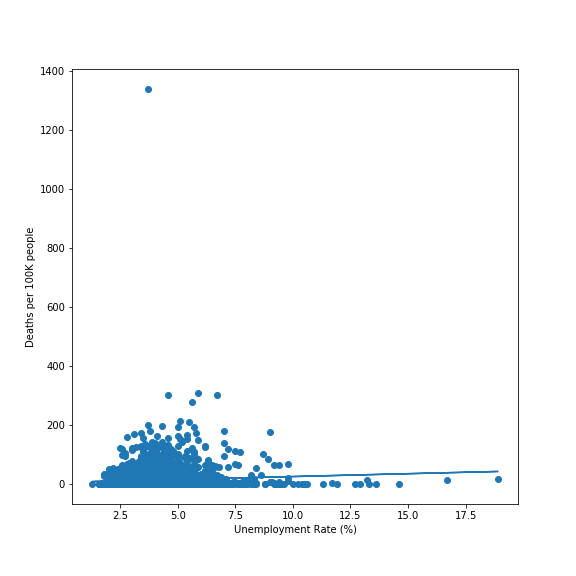


PCA: Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components.

We can use PCA to reduce the number of features to use in our ML algorithms, and graphing the variance gives us an idea of how many features we really need to represent our dataset fully.

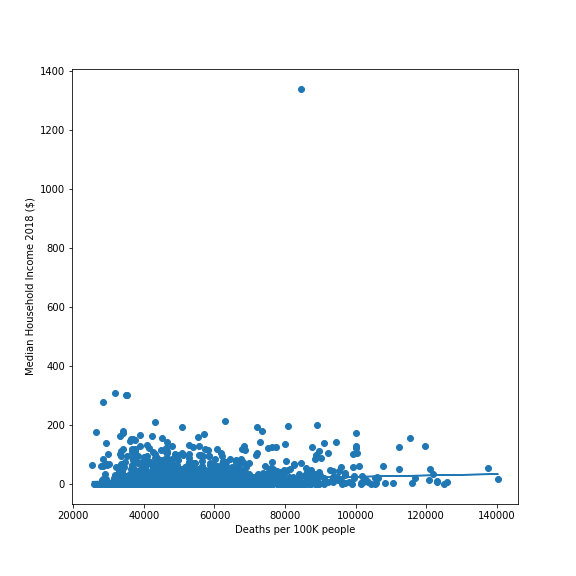


The graphs above show that there is positive correlation between poverty and deaths per 100K people. Because the relation is weak, the poverty level can be explained by the fact that people in poor counties can't afford healthcare or they can't afford to stay at home, as such, they are more exposed to the virus when they are at work.

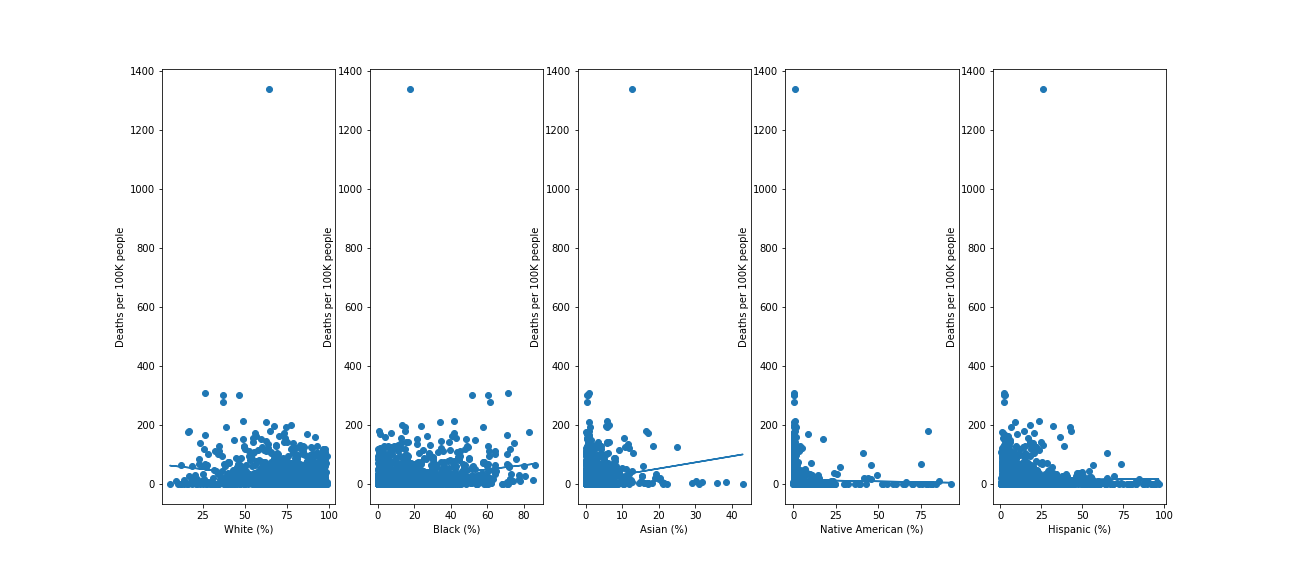


Based on the unemployment map above, there are three geographical regions from the deaths per 100K people map that coincide with this one - the South-West, the South-East, and the Midwest.

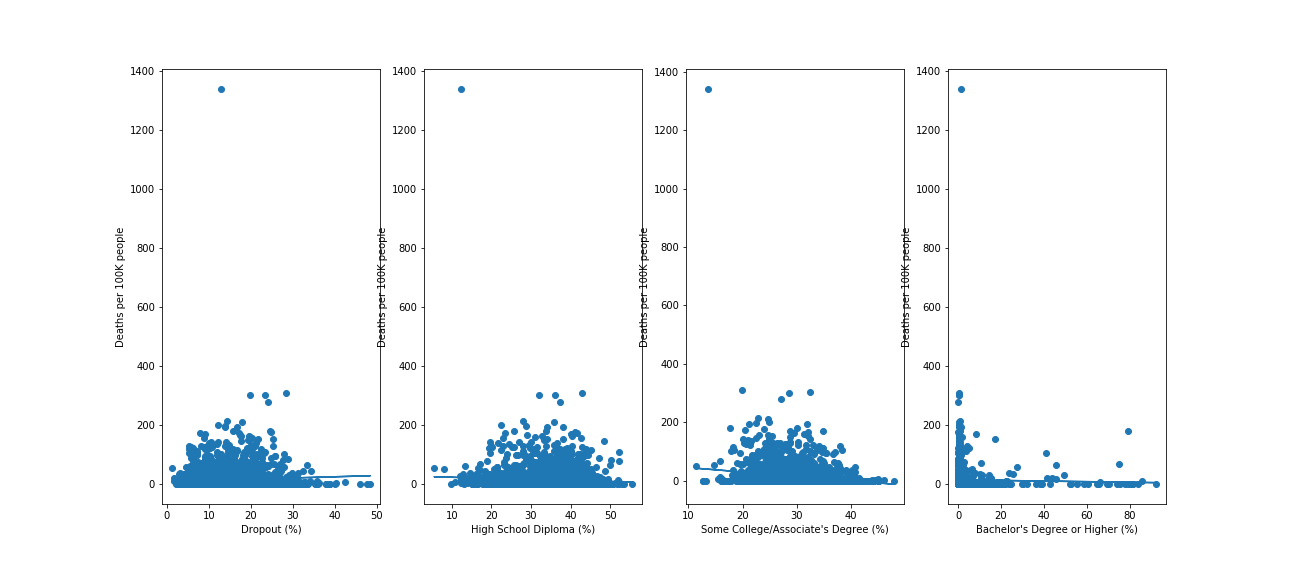
The South has both high poverty rate and high unemployment rate; the same goes for the South-West.



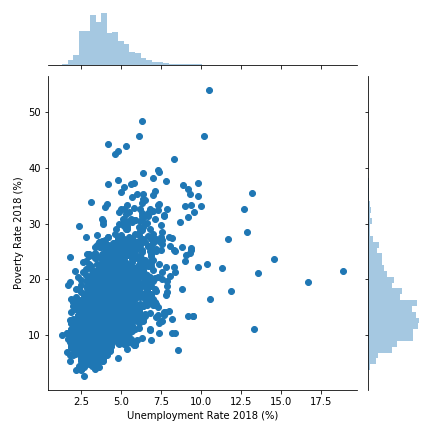
There is a positive correlation between median income and deaths per 100k people. The map shows that midwest and northeast have higher median income, however, the death per 100k people map also shows these areas to have higher deaths. The results might be due to higher population density in these regions.

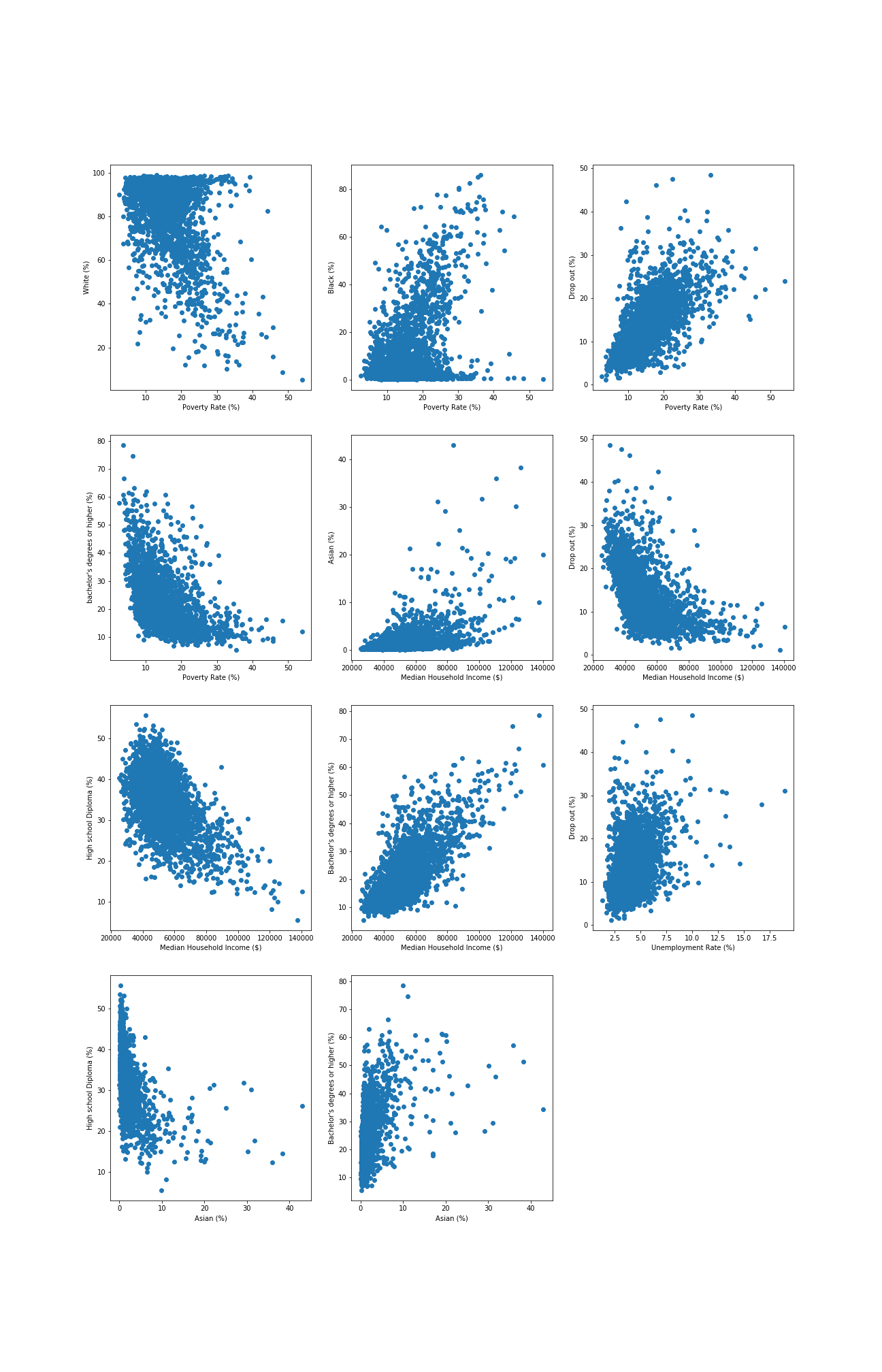


The racial analysis focuses on the racial makeup of a county's population.

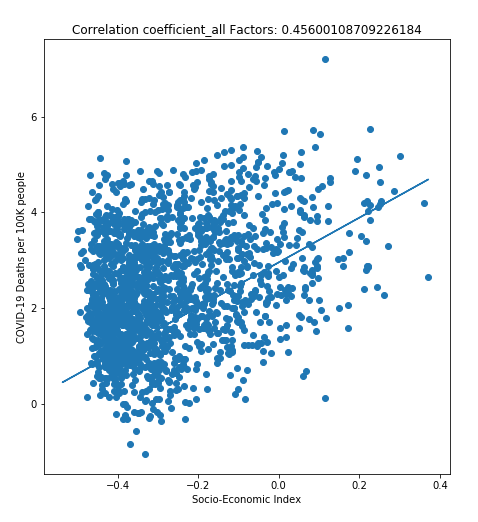


The educational analysis focuses on different types of educational backgrounds in a population.

The graph shows a positive correlation of unemployment and poverty, meaning the higher the unemployment rate, the more people living in poverty.



Final Analysis combines all the socio\_economic factors and correlates it with the deaths per 100K people to show the big picture.



It’s not just one factor that matters. As can be seen from the final analysis, the combination of all factors weighted according to their correlation coefficient yielded a much stronger correlation than any of the individual analyses. In general, there are two types of counties which are affected negatively more than others.

Counties with very high populations and population densities such as those in the Midwest. People in these counties are easily exposed to the virus and transmission occurs much quicker, often leading to overwhelmed hospitals and tired doctors.

Counties with high poverty/unemployment and low median incomes also tend to have higher proportions of Native Americans and African Americans such as those in the South-East and the reservations in the South-West. In addition, the lack of affordable quality healthcare also means that people don’t get adequate medical attention.

Sources - <https://github.com/CSSEGISandData/COVID-19>

<https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/>

<https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html>

Data Manipulation: To conduct machine learning, the remaining columns are taken to create important numerical values out of them.

Get\_dummies method of pandas is used to separate columns for each feature based on the unqiue values in the dataset.

Conclusion

The combination of the COVID-19 pandemic, a drop in unemployment, increasing number of cases and deaths have the potential to seriously impact stability state and county wide, and will strain public health and social security systems in the short to medium terms. This analysis uses machine learning models' accuracy scores to predict relationship betwen covid-19 cases/deaths and several socio-economic factors like poverty, household income, unemployment, education and race. The high accuracy scores of each model indicate that there is a strong relationship and all dependent variables can be highly impacted by changes in independent variables, such as covid-19 cases/deaths.