**Prediction of Poverty Rate in US Counties**

Samuel U S Achibiri January, 2018

**Executive Summary**

This report deals with the analysis and modelling of data on the rate of poverty in counties in the United States. It investigates how factors such as economic typology, urban influence, health and other demographics affect poverty rate. The data is made available for the public by the United States Department of Agriculture Economic Research Services (USDA ERS) which is then hosted on the DRIVENDATA data science challenge website. The dataset has 3198 samples and 33 features excluding the label (poverty rate).

After computing descriptive/summary statistics and visualization of the data a number of relationships between certain features of the data and the target variable, poverty rate was identified. Using scikit-learn’s library a model of the data was built for making future predictions. Using the test data made available such prediction was made.

The following conclusions were arrived at after exploration, analysis and modelling of the data:

* Urban counties and those close to urban areas and with large population densities tend to have low rates of poverty, while those farther away have higher poverty rates.
* Age, Healthcare and lifestyle contributes to the rate of poverty of different counties.
* There is a very strong relationship between the population density of different racial groups and the poverty rate of the counties that they live in.
* The level of education of most members of a county affect the median rate of poverty of the county.

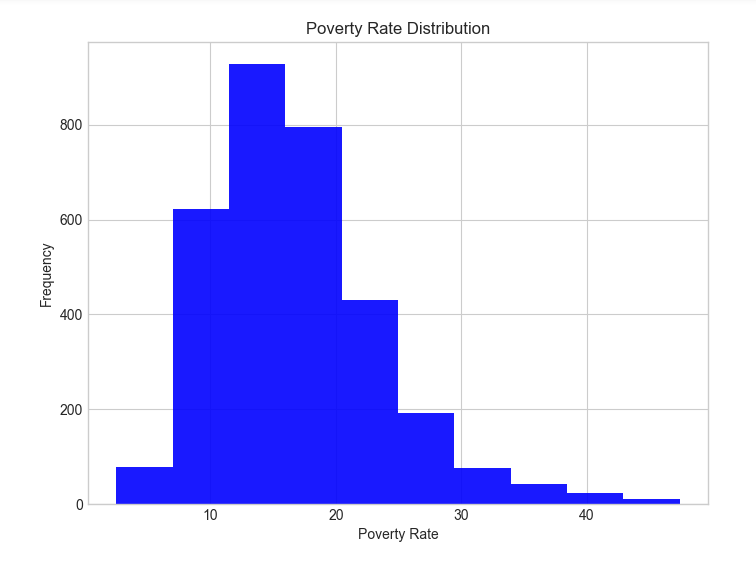
**Data Exploration**

Although the majority of the features in the dataset were mainly numeric, four were categorical. They include: area\_\_rucc, area\_\_urban\_influence, econ\_\_economic\_typology and yr. The first step in the data exploration was to look at the distribution of the rate of poverty for all the counties in the country. This was done using a histogram.

Sample size = 3198

Mean poverty rate = 16.82

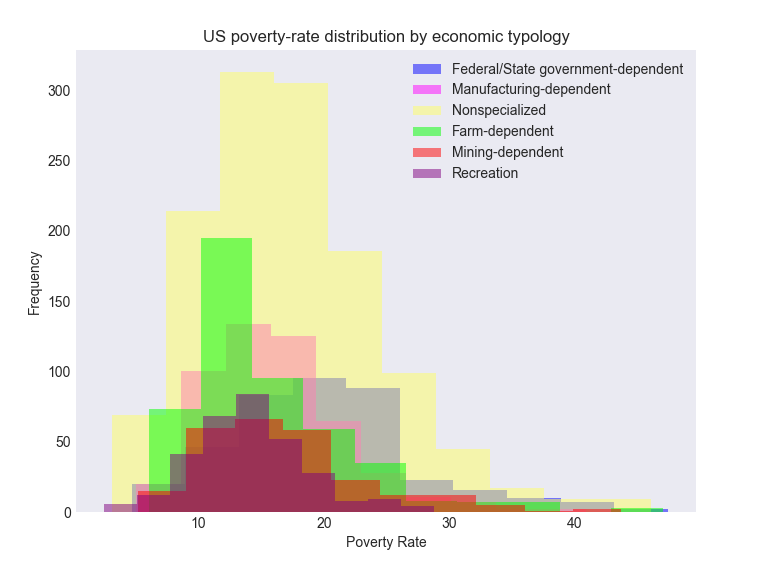
Median poverty rate = 15.8

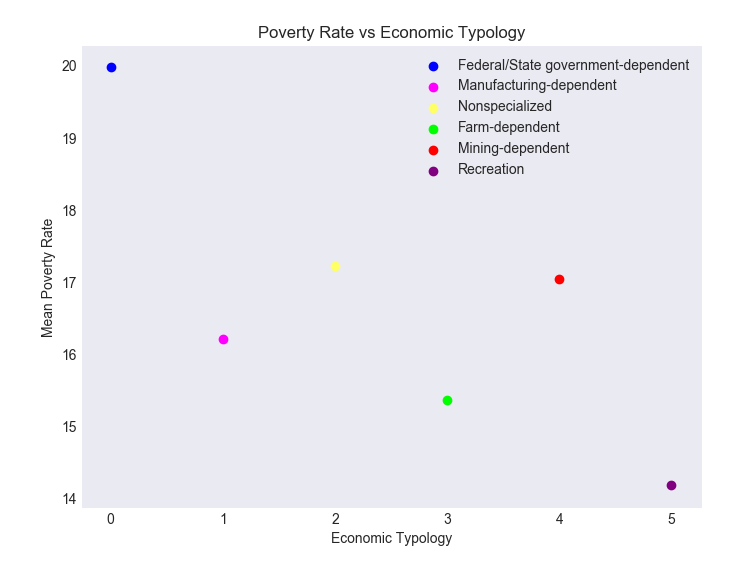


A histogram of the poverty rate for all the counties revealed a median poverty rate of 15.8 and a slightly higher mean poverty rate of 16.82. The histogram is also skewed to the right which indicates that the majority of the US population leave in counties with low poverty rates. However about 25 percent of the counties have a poverty rate above 20%. This represents a sizable chunk of the population leaving below the poverty threshold.

**Effect of economic typology on poverty rate**

The kind of jobs that are predominant in counties was found to influence the rate of poverty. A histogram of various economic typologies and their corresponding poverty rate distributions show that most people work in nonspecialized fields as such are the main contributors to the national mean of 16.82. The plot also shows that other than counties whose economic typology is recreational, the rest have a wide variance.



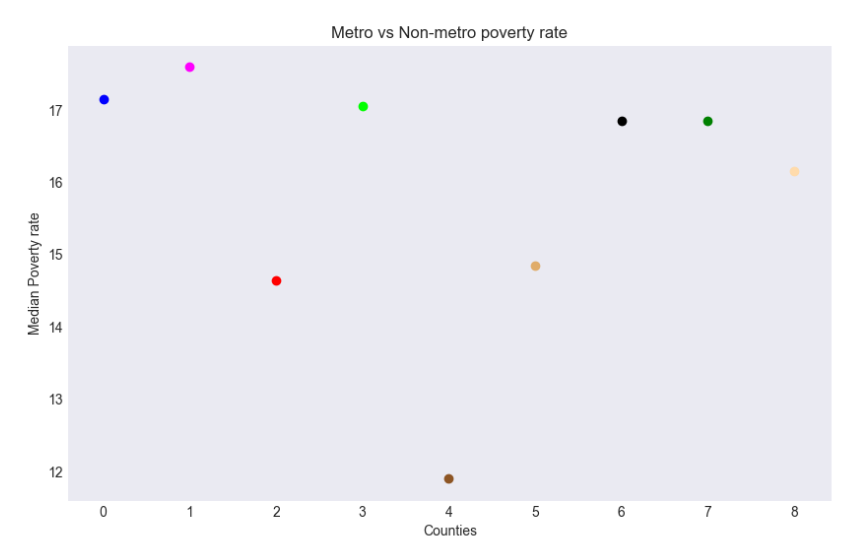


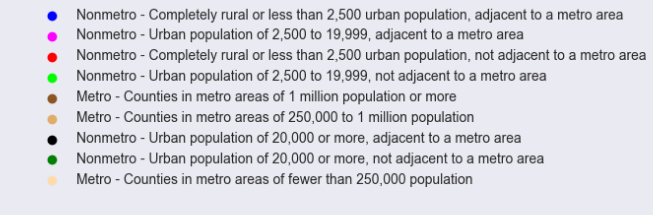
Of the specialized economic typologies, counties with higher mining and government dependence are also high on median poverty rate. A visible reflection of these is seen in mining towns becoming ghost towns. Counties that are mostly dependent on recreational activities have the lowest median poverty rate. One explanations for this phenomenon is the fact that recreational counties have a lot of tourists who are most often wealthy individuals traveling on vacation. In addition to the wealth brought into recreational counties by fun seekers, services in the hospitality industry carter to the needs of visitors while also providing good jobs for the local population.

Counties that are dependent on farming or agriculture were also discovered to have a median poverty rate that is below the median for the general population. This is not surprising as advances in mechanization, storage, pest/disease resistant crops, ease of transporting/exporting farm produce, has enabled the lowering of the poverty rate in these communities.

**Urban influence**

Analysis of the effect of urbanization of counties or their closeness to one, showed that counties that are metropolitan and with a large population have lower poverty rates.

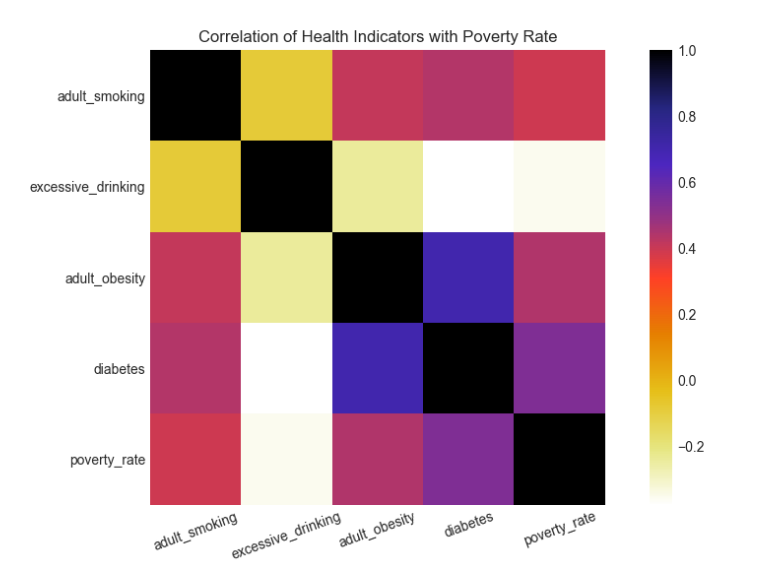




The low rate of poverty in metropolitan counties is not a shock as one of the main reasons people move into metro areas is to seek better job opportunities. One more advantage of metro counties over their non-metro counterparts is large population size, which often serve as incentive for starting businesses. This leads to the creation of more jobs and thus a reduction in the rate of poverty in these counties.

**Effect of health indicators on poverty rate**

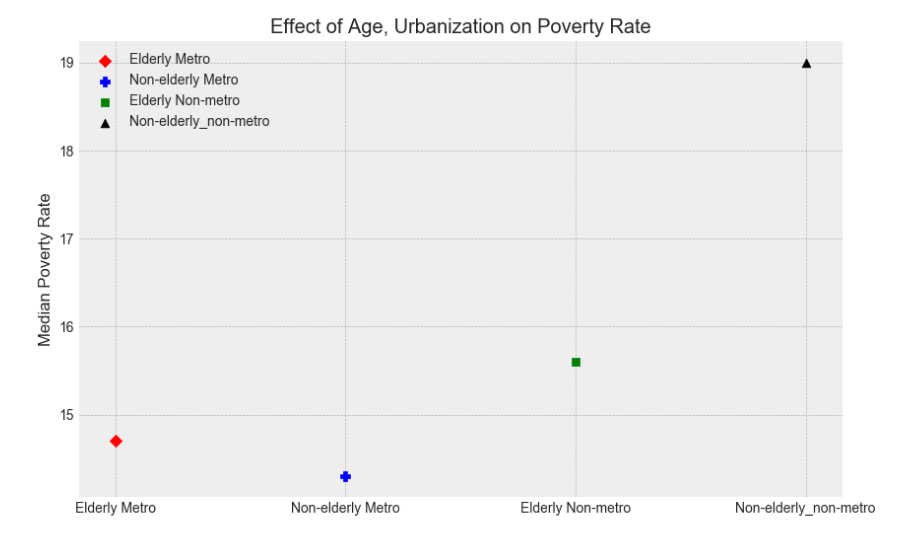
Health indicators such as smoking/drinking habits, obesity, diabetes is correlated to poverty rate. One interesting find is the negative relationship between excessive drinking and poverty rate. Counties that have a drinking problem also have lower poverty rates. This should not be mistaken as a recommendation for counties to develop an excessive drinking habit as a way of reducing poverty in their neighborhoods. Rather, this is an indication that most poor people are more worried about where to find the next meal, and not the fleshly gratification of alcohol.



Diabetes, adult obesity, and adult smoking are all positively correlated to the rate of poverty, with diabetes having the highest correlation of approximately 0.54. An explanation for this relationship is likely due to the high cost of managing this condition. Another reason why an increase in the number of people with diabetes in a county correlates to an increase in poverty rate might be due to reduced productivity of sufferers and other complications that are often associated with this ailment. Just like diabetes, the positive correlation between obesity and poverty rate is likely due to reduced productivity, obesity related complications and the healthcare cost of managing such. Also of note is the strong relationship between diabetes and obesity. Therefore, staying healthy and losing weight is a worthwhile investment.

**Effect of old age on poverty rate**

First, counties with percentage of their population age 65 and older greater than the national median are classified as elderly and the rest non-elderly. These are further subdivided in two categories, namely: metro and nonmetro – totaling four divisions. The median for each division is then obtained plotted.

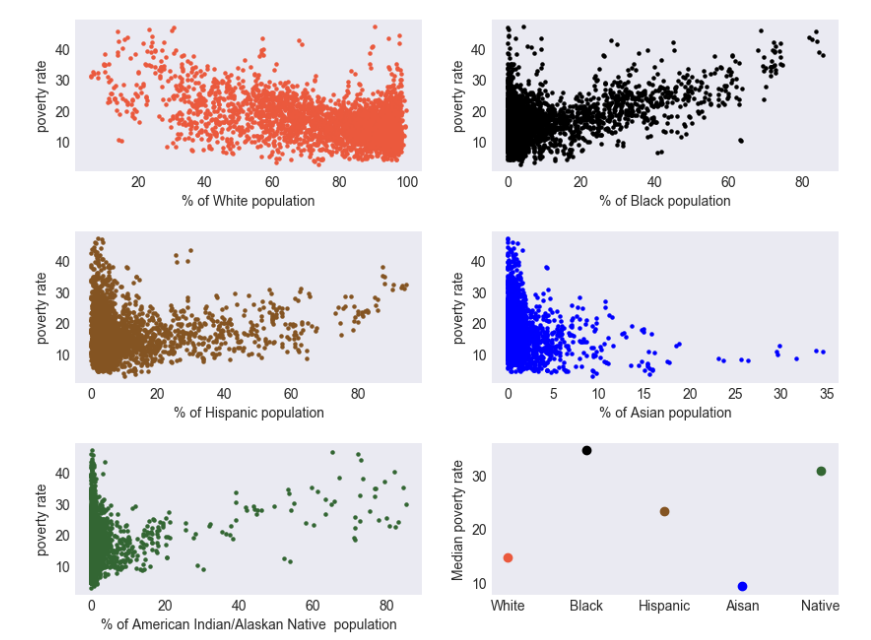


As can be seen from the plot, there does not exist a clearly defined relationship between counties classified as elderly or non-elderly and their median poverty rate. The one major factor that seemed to affect the poverty rate is if the county is metropolitan or not. Counties that are elderly and those that are non-elderly both had low rates of poverty provided they were also metro.

The counties with the highest poverty rates are those that are non-elderly and also non-metropolitan. Thus, education and exposure to opportunities are likely primary factors responsible for the poverty rate and not solely age.

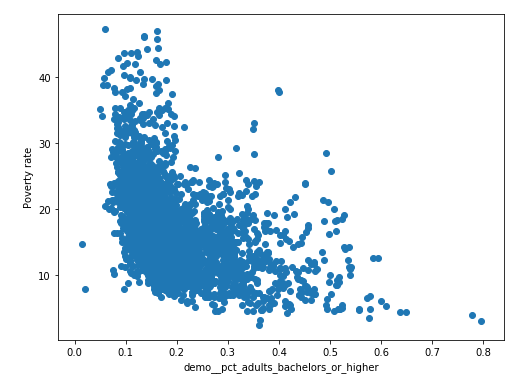
**Race vs poverty rate**

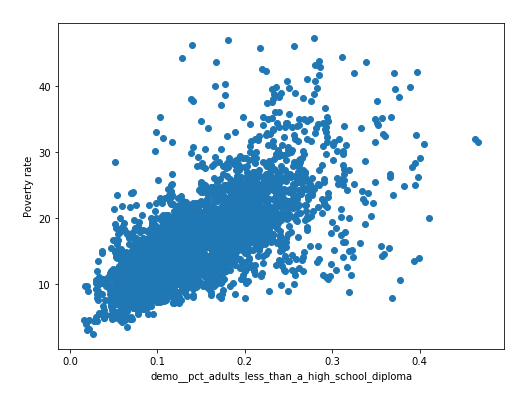
One controversial question a lot of people seem to be interested in, is the relationship between race and rate of poverty. Does the racial predominance of a county influence or affect its poverty rate? The answer can be found in the plots below.



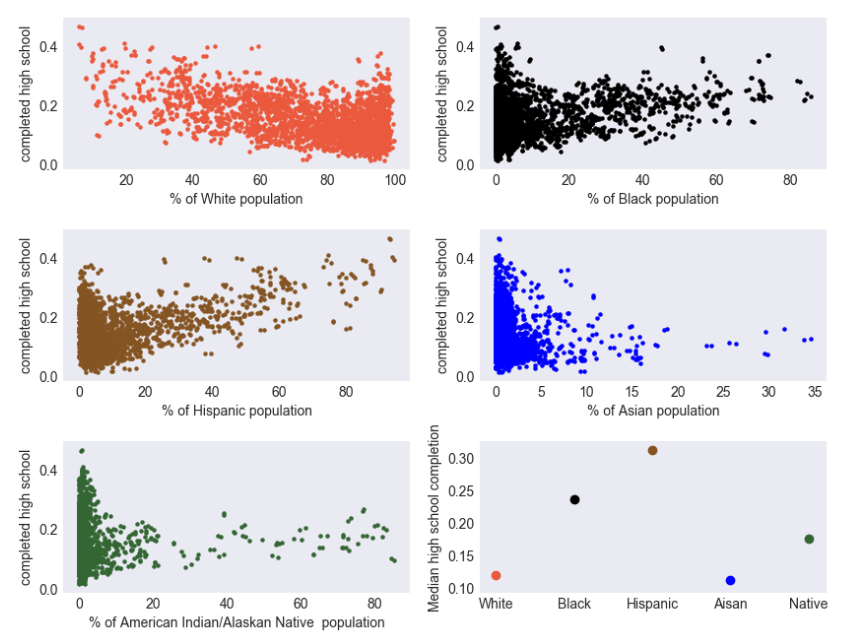
Evidently there seem to be a relationship between race and poverty rate. There weren’t any counties with and Asian population greater than 50 percent of the population. Thus, unlike other racial groups where >%50 is used to classify racial demographics, for people with Asian roots %25 is used. The plot shows that only two racial groups have a negative correlation with rate of poverty while the rest have a positive correlation. The Whiter or Asian the county, the lower the rate of poverty.

If this plot is treated in isolation, one might draw the wrong conclusion solely blaming racism for the increase in poverty rate with respect to increase in certain racial distributions. To investigate further, we compared other features that are positively correlated to rate of poverty to the various racial groups.





Notice how there is a strong positive relationship between the percentage of people in counties that did complete high school and the poverty rate for those counties. Counties with more people without a high school diploma tended to also have a high poverty rate. While those with a bachelor’s degree or higher had low poverty rates. The plot below compares the high school completion rate across the various racial distributions.



The similarity between plot above and the earlier plot relating race and rate of poverty shows that seeming relationship between race and poverty rate was actually a correlation between education and poverty rate. Counties with higher black, Latino and Native American populations had low high school completion rates. The reason for this trend is not covered in this report. However it is worthy of note that improving the quality of education and the completion rates in these communities might have a substantial effect in improving the rate of poverty in these counties.

**Modelling rate of poverty**

Before developing a model for predicting the rate of poverty in counties using the provided samples and features, the training data had to be prepared first. The categorical data in the dataset had to be transformed into a format compatible with machine learning. Also all null variables were replaced with -999999, which allows the algorithm to treat those specific features as outliers rather than delete complete rows and lose vital information.

Using scikit-learn’s linear regression algorithm a model for prediction was developed using the provided training data. After training, the model was used to make prediction using the provided test set. The resulting prediction was found to have and RMSE score of 2.9313.

**Conclusion and recommendation**

The rate of poverty in the various counties was found to be lower in metropolitan counties as compared to non-metropolitan. Counties that are mining based and government dependent also had high rates of poverty. Therefore, a more concerted effort should be put into retraining and developing new skills for individuals who used to be in one-time lucrative industries that are no more as attractive. A number of health indicators were also found to influence rate of poverty. More emphasis should be place on encouraging a healthy lifestyle while helping those with health challenges should be given more support. Education was found to be a key factor in the rate of poverty of different counties as such more emphasis should be place in helping counties that are struggling with low high school completion rates. There should be more incentive for completion.