**Predicting Cardiovascular Mortality Rates with “Just” Poverty Rates**

# Background

Around the world, rising healthcare costs are worrying policymakers. In 2015, compared to the average of other countries with comparable incomes, the United States’ healthcare costs exceed the average by a factor of three. The US spends more on healthcare than any other country, but with average health outcomes. According to the World Health Organization, the US is ranked 30th in the world in terms of life expectancy.

The American Health Association projects that the costs associated with this cardiovascular disease will exceed $1 Trillion annually by 2035.

## Project Purpose

The goal of the project is to answer one specific question. Can we predict cardiovascular mortality rates in the United States using the factors of poverty?

Of course, natural questions arise. Why would we only use only the poverty factors to predict the mortality rates? Should we include other factors like education, health insurance coverage, etc.?

All these questions are valid and bring up interesting points. However, we must understand the realities of the current political climate in the United States. Partisanship has never been higher and compromise is hard to come by.

If we were able to use the poverty factors to predict cardiovascular mortality rates in a satisfactory way, then we could direct policymakers to focus their efforts on reducing just these factors instead of a much larger combination of factors. Of course, this is based on the narrow assumption that their overall goal is to reduce the cardiovascular disease mortality rates.

This is not to say that the other factors are not important and do not impact the mortality rates in a significant way. Far from it, but the political climate has turned these factors into political third rails (see Obamacare as an example). There is not enough political will to influence these other factors. Thankfully, reducing poverty is something that still garners bipartisan support.

## Dataset

The master dataset is a combination of multiple datasets.

* US Census Poverty Rate 1980-2016
* Dataset for Information and Accountability Transparency
* US GDP 1980-2017 from the Bureau of Economic Analysis
* From the Bureau of Labor Statistics, Overall US Labour Productivity 1970-2016
* Average Annual Unemployment Rates by State from the US Bureau of Labor Statistics
* University of Washington’s Institute for Health Metrics and Evaluation Cardiovascular Mortality Rate dataset.

However, all the datasets were untidy. Therefore, they were not in a format that made analysis easy and required some cleaning.

The biggest problem we encountered was the fact that datasets had differing levels of detail. One dataset went down to the county level; the other dataset was limited to the overall Country level. One dataset has entries every 5 years; the other dataset had annual entries.

I had to figure out a way to combine the US level data and the State level data. So I took the overall country column value for a specific year and put that as the value for each state for that year. For example, the information transparency score for the United States was a 75 in 1980. So, for every entry where the year was set as 1980, a transparency score was entered as a 75.

This dataset cleansing process is documented under the Data Preprocessing portion of our code.

# Analysis

## Exploratory Data Analysis

### Outlier Detection

We needed to determine if there were outliers in our data and how we wanted to deal with them.

Given the fact that our data is considered high-dimensional, we came across a method called Local Outlier Factors. This is an anomaly detection method that uses measures how isolated a data point is compared to its neighbors. The points with few close neighbors are considered outliers. Data points with a score of -1 are considered outliers, 1 are considered fine.

After using the this Local Outlier Factor method, we found that we had 36 outliers. The breakdown was as follows.

|  |  |
| --- | --- |
| **State** | **# Of Outliers** |
| Alabama | 7 |
| California | 7 |
| Texas | 5 |
| New York | 4 |
| Arkansas | 2 |
| Wisconsin | 2 |
| Arizona | 2 |
| West Virginia | 2 |
| Alaska | 2 |
| Wyoming | 2 |
| Florida | 1 |

Remember that we only have 7 data points for each state. Therefore, the method is flagging all of the data point for Alabama and California as outliers, as well as a majority of Texas and New York’s data points.

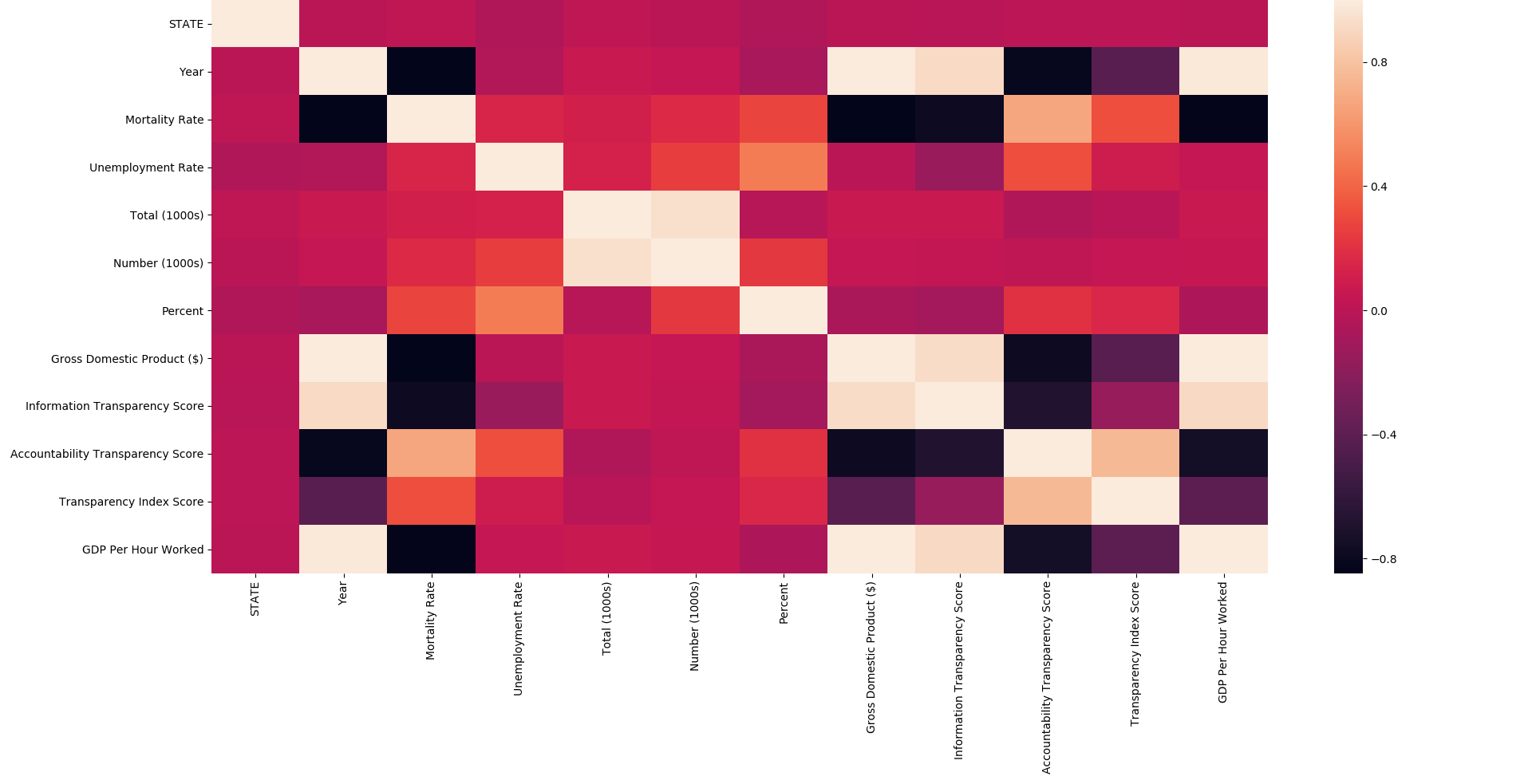
We will remove all flagged data points from out data. However, we will just put it aside for now and see how it affects our analyses later.

This method is documented under the Outlier Detection With LOF portion of our code.

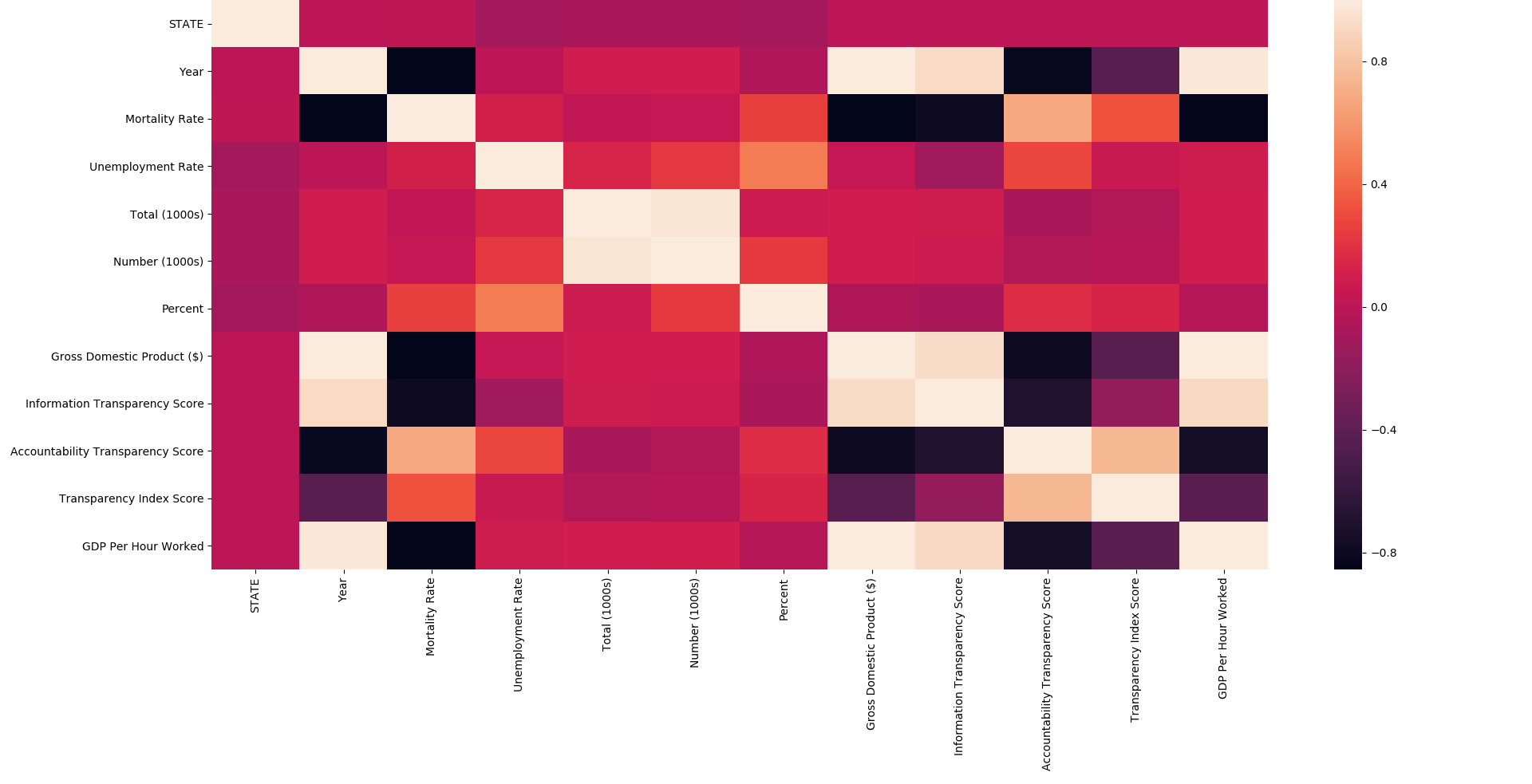
### Correlated Variables

The following is a heat map of pairwise correlation of all the variables when and when we don’t include the outliers.

Without Outliers



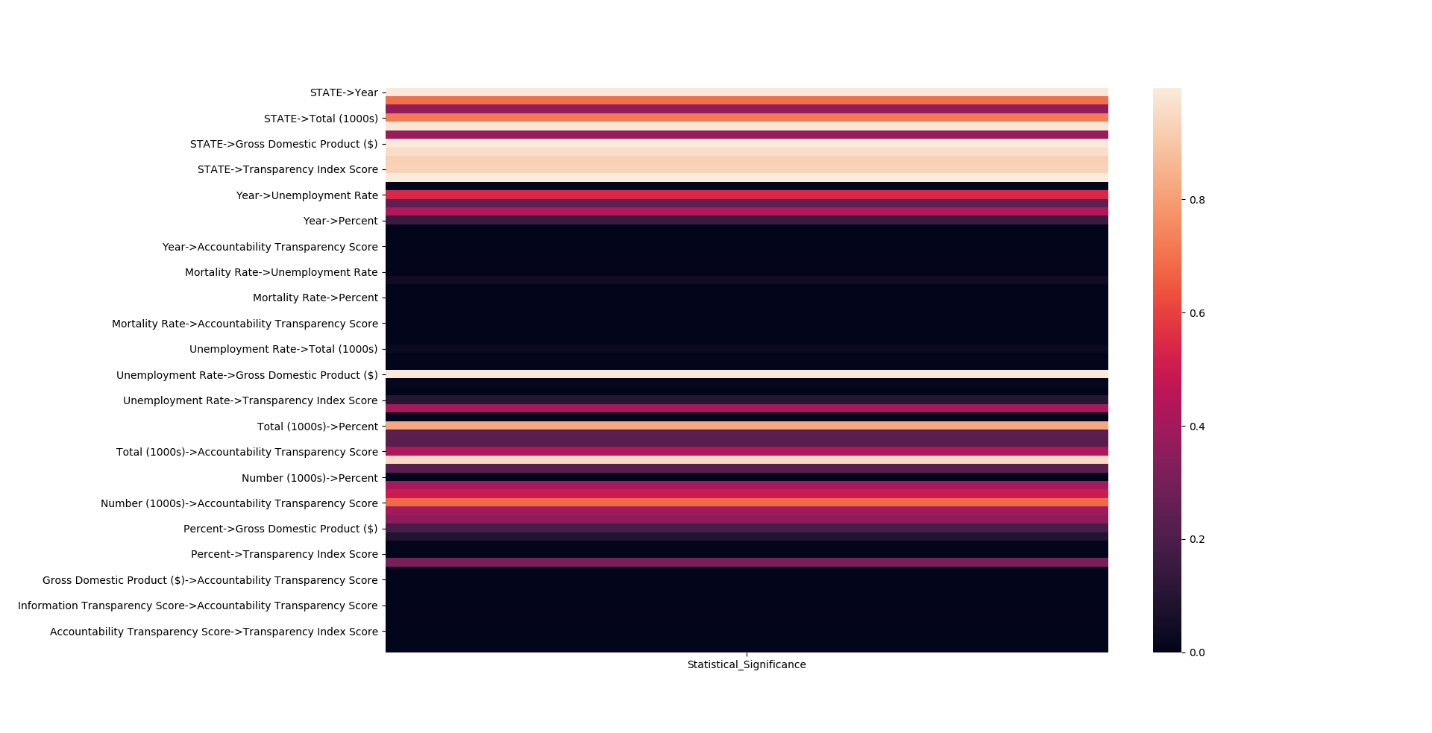
With Outliers



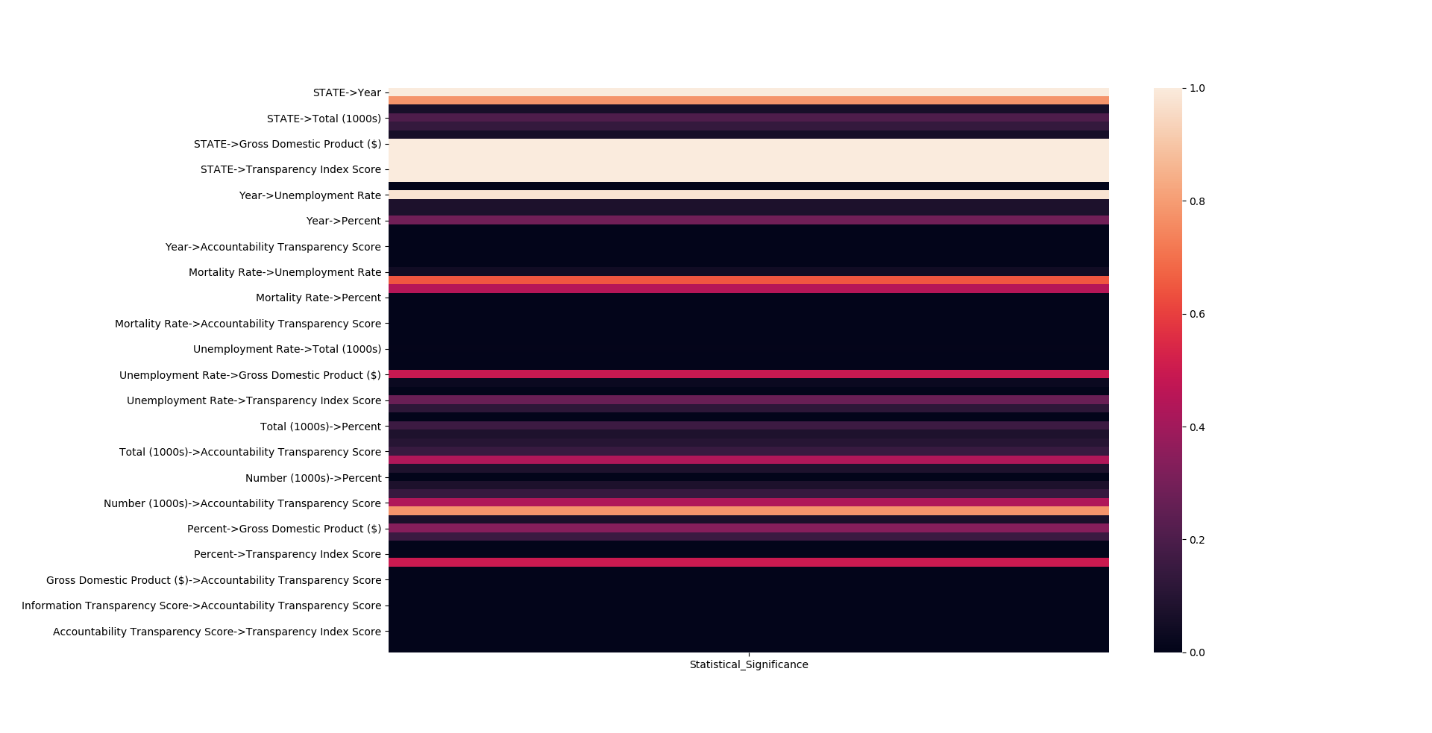
There does not seem to be much of a difference if we remove the outliers.

The following is a heatmap of the p\_values of the correlations above.

**Without Outliers**

****

**With Outliers**



Removing the outliers does result in some of the correlations becoming more statistically significant.

The next step we performed was to look at the specific values of these outliers to see if it were a data entry error. In reality, all their entries seemed plausible. Even though we believe data points are not outliers, we scaled the data using RobustScaler from Scikit-Learn anyway.

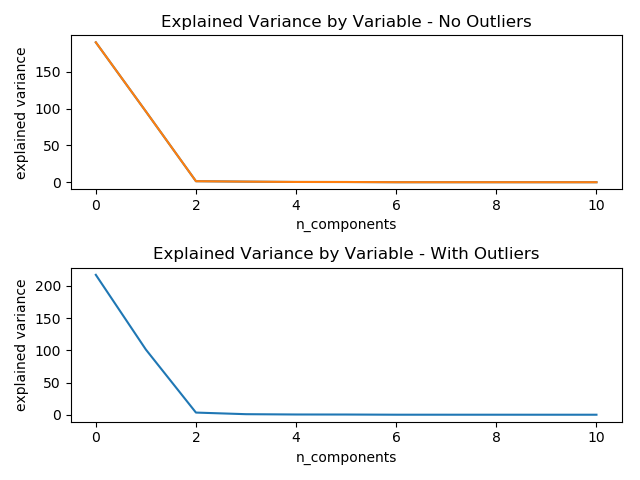
Because we were unable to come up with a way to justifiably remove these flagged outliers, we will keep them in our data for now.

The code for these correlation is under the Pairwise Correlation section of our code.

### Dimensionality Reduction

From our correlation heatmaps above, we see that certain variables are important and others are not. Does it mean all of them are useless when it comes to predicting the mortality rate? Of course not. Nevertheless, we will be using Singular Value Decomposition to reduce the dimensionality of the data.

Of course, we will do a quick test of the decomposed data with the non-simplified data using linear regression. If there is no significant accuracy increase after using the reduced data, then we will only use the non-simplified dataset.



|  |  |
| --- | --- |
| Variable | Explained Variance Ratio – Without Outliers |
| STATE | 65.74% |
| Year | 33.26% |
| Unemployment Rate | 0.48% |
| Total (1000s) | 0.28% |
| Number (1000s) | 0.13% |
| Percent | 0.10% |
| Gross Domestic Product ($) | 0.01% |
| Information Transparency Score | 0.01% |
| Accountability Transparency Score | 0.00% |
| Transparency Index Score | 0.00% |
| GDP Per Hour Worked | 0.00% |

|  |  |
| --- | --- |
| Variable | With Outliers  Explained Variance Ratio |
| STATE | 67.19% |
| Year | 31.29% |
| Unemployment Rate | 1.03% |
| Total (1000s) | 0.26% |
| Number (1000s) | 0.11% |
| Percent | 0.09% |
| Gross Domestic Product ($) | 0.01% |
| Information Transparency Score | 0.01% |
| Accountability Transparency Score | 0.00% |
| Transparency Index Score | 0.00% |
| GDP Per Hour Worked | 0.00% |

We see that the outliers do affect our analysis, albeit a tiny amount. Therefore, we will not be removing the outliers at all and all our future analysis will be based on the data that contains the outliers.

The code for this portion is under the Dimensionality Reduction portion of our code.

## Machine Learning

### Test/Train Split

Because our data is longitudinal grouped data, we could not rely on the normal k-fold method to come up with our test and train set. We had to perform a variation called GroupKFold, found in Scikit-Learn. This variation would ensure that the same group would not appear in two different folds.

Now the question raised is how we should group our data. Should we group it by state or by year? We decided to group by year so that we would have less hyperparamaters to tune later on. For the GroupKFold method to work, the number of distinct groups has to be at least equal to the number of folds. If we group by year, the max number of possible folds is 7.

### Linear Regression

Now that we had determined that we would use GroupKFold for our cross validation technique, we now had to determine which predictive techniques to use. The following are different type of linear regressions we tried.

* Ordinary Least Squares
* Lasso
* Elastic Net

Because we were not interested in tweaking hyperparameters, we used the default arguments for each type. The following are the resulting scores for each method. We compared using the Top 3 Variance Explaining Variables vs using all the variables.

|  |  |  |
| --- | --- | --- |
| Variation | Using All Variables Score | Using Top 3 Variance Explaining Variables Score |
| Least Squares | 0.2482 | 0.6185 |
| Lasso | 0.6811 | 0.6254 |
| Elastic Net | 0.6845 | 0.6434 |

### Decision Trees

The following are the different decision trees we used and the resulting score.

|  |  |  |
| --- | --- | --- |
| Variation | Using All Variables Score | Using Top 3 Variance Explaining Variables Score |
| Decision Tree Regression | 0.4374 | 0.0398 |
| Random Forest Regression | 0.6966 | 0.3446 |
| AdaBoost Regression | 0.6248 | 0.3658 |

Of course, if we were interested in tweaking the hyperparameters, then there are multiple knobs to turn. For example, we could tweak the number of folds in our cross-validation. For our dataset, the maximum possible number of folds is 7 due to requirement of the number of distinct groups being at least equal to the number of folds. The minimum number of folds is 2.

We could also tweak which variables to include in our regressions. We can clearly see from the tables above that different combinations of variables produce significantly different results.

The code for this section is in the Machine Learning portion of the code.

# Conclusions

Our main question is whether or not we can use the factors of poverty to predict cardiovascular mortality rates in the United States. Using our results charts above, the answer is a resounding yes. It is clear that we can explain more than half of the variability in our data. With some hyperparameter tweaking, we might be able to push it into the 70% range.

This does not mean that the factors of poverty are the best predictors of the mortality rates. Intuitively, we know that other factors must come into play.

However, even knowing nothing about how those other factors affect Cardiovascular mortality, we can recommend that focusing specifically on the factors of poverty will result in a positive outcome.