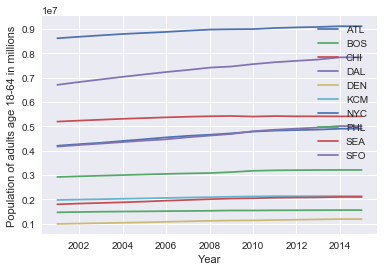
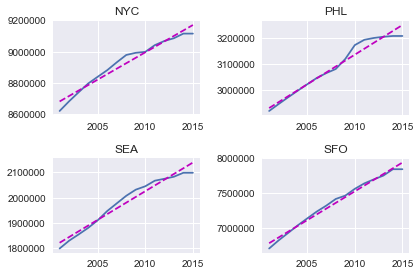
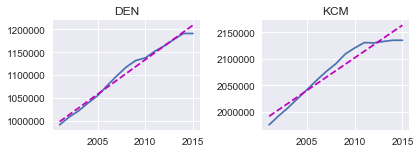
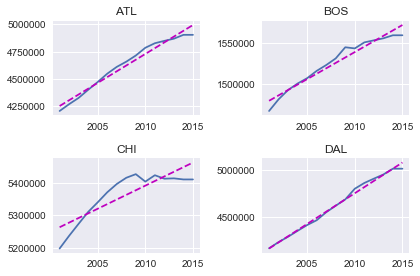
**Social Security Administration Dataset**

The Social Security Administration (SSA) releases data every so often about the state of the disability system in the United States. In addition to helping informing voters, give numbers for various media outlets to talk about, etc. Various businesses can use the data to understand the way populations behave in different parts of the country and use it to predict what future costs will look like.

In order to understand a set of data, it is important to see trends and interactions between different variables. The most common way to do this is through plotting out the data. Both adult population and the adult disability rate were plotted and grouped by region to see how some of these variables behaved and give us insight into what model(s) we should use.

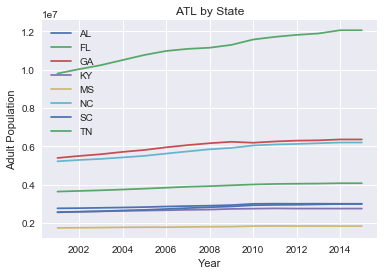


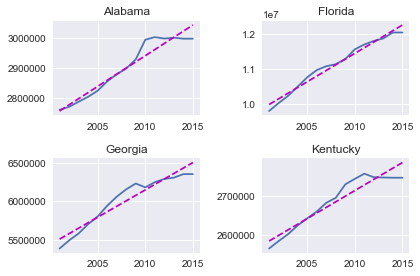
Populations by Region:

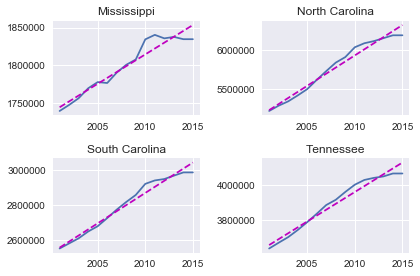


Initially looking at the population, all regions appear to be very highly correlated with each other. In fact, the lowest correlation between regions was between Dallas and Chicago at 0.86. A breakdown of how individual state population fluctuate within regions is plotted on the following charts. The effects of major events are evident.

Atlanta Region

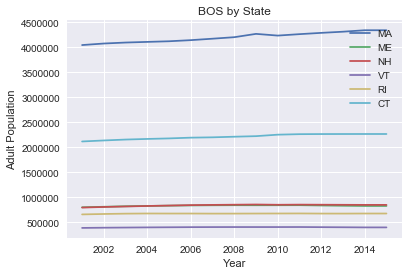


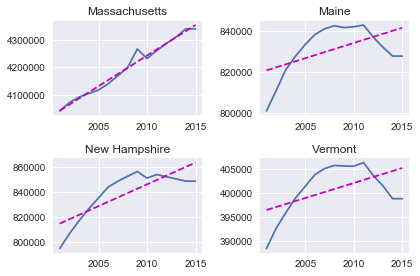


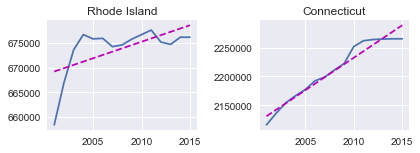


Although Florida may seem like an outlier, the data is in fact accurate and matches that in other data sets on the web.

Boston Region

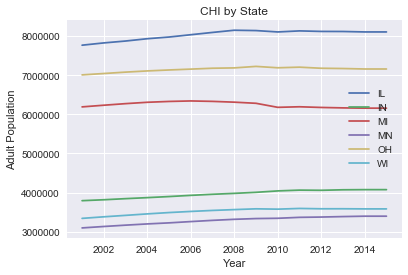


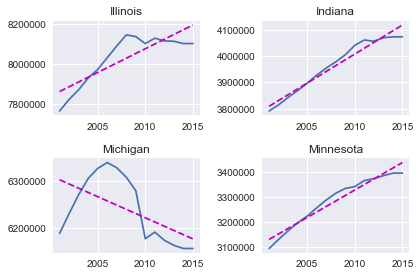


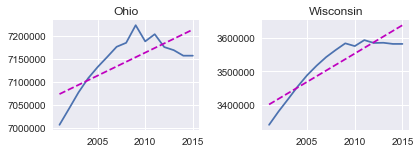


New Hampshire, Vermont, and Maine all have very similar trends in population since they all share the same forest and mountain geography. Rhode Island has been struggling to build industry since losing manufacturing to overseas.

Chicago Region

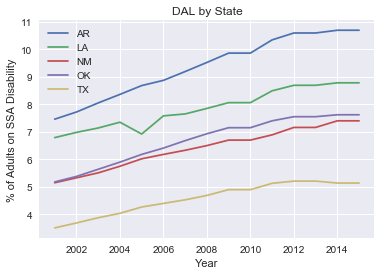


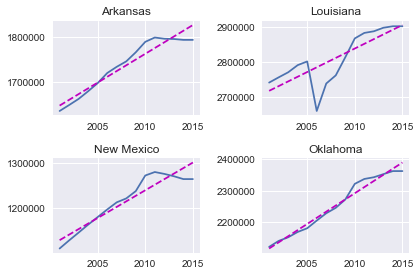


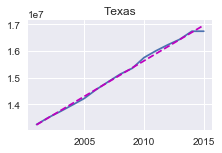


Michigan was hit especially hard by the financial crisis and lost many jobs related to the auto industry, which it relied heavily on. Ohio is also losing jobs, especially in manufacturing which could be related to the decline in its youth population. Illinois has had a steady decline for some time and consensus on the web seems to be tax related. Their severe financial burden is chasing residents out in addition to the state's stagnant growth.

Dallas Region

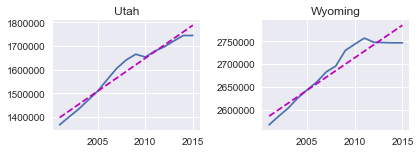
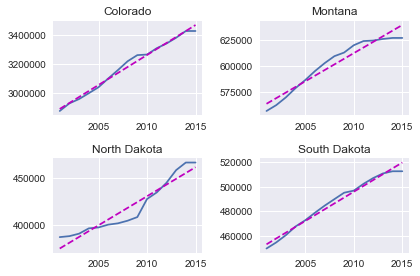
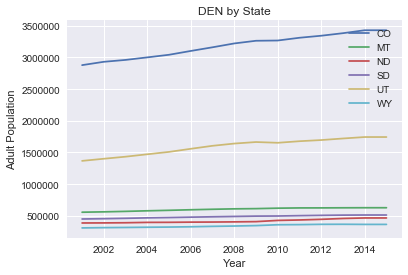




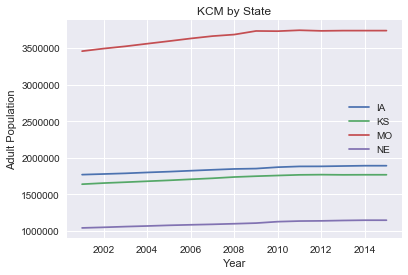


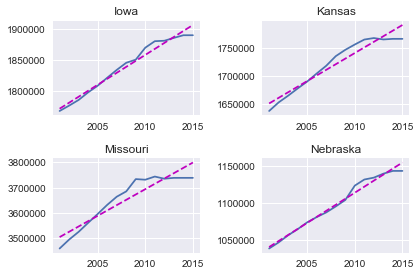
New Mexico is also accurate, despite appearing as wrong data. Louisiana looks especially hard hit from Hurricane Katrina, but has since rebounded quickly.

Denver Region

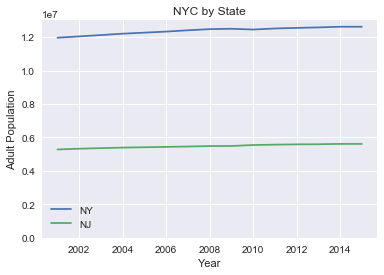


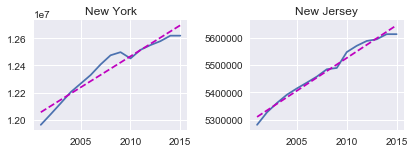
Kansas City Region



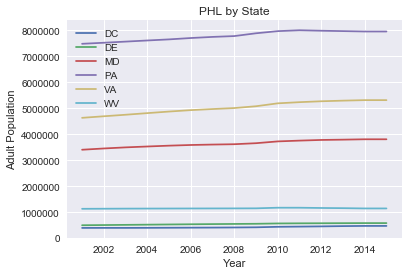


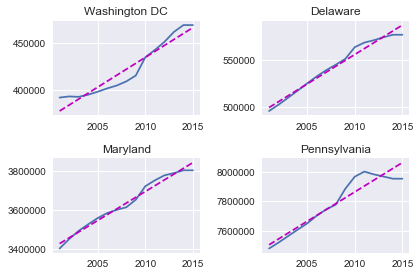
New York City Region

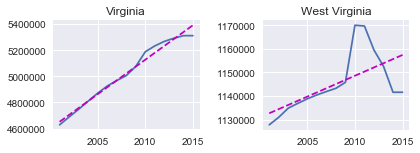




Philadelphia Region

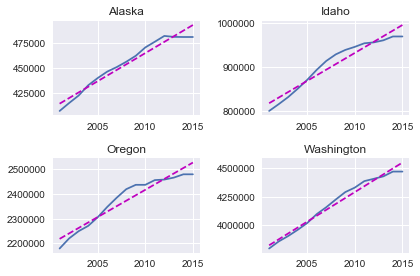
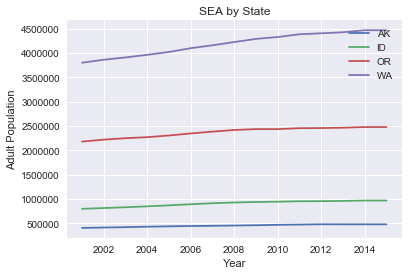




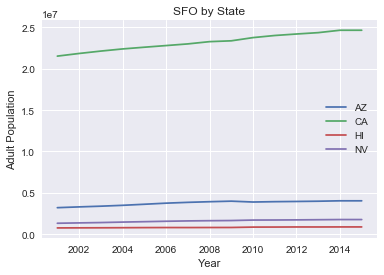


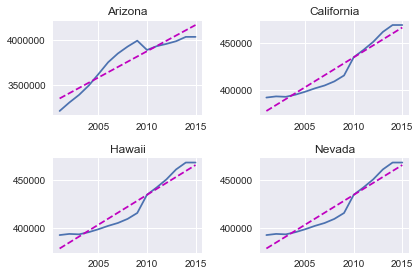
The phenomenon in West Virginia showing a sharp increase in population then back down was confirmed in other datasets as true data.

Seattle Region



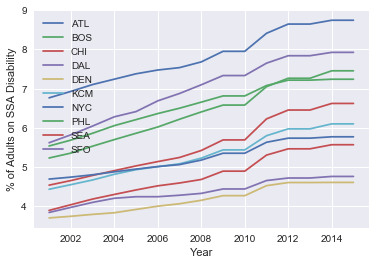
San Francisco Region

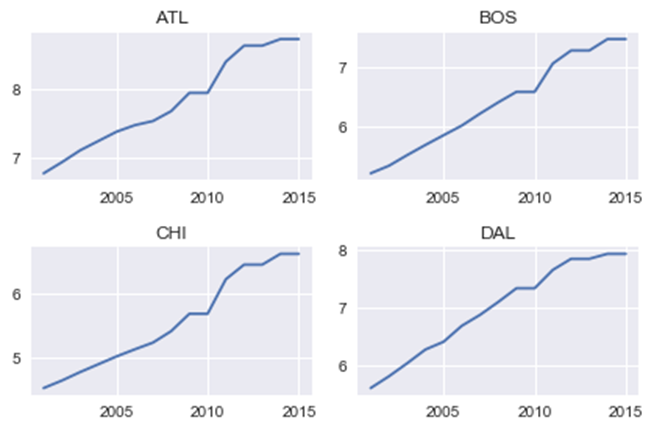


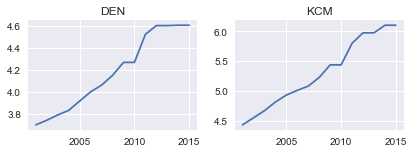


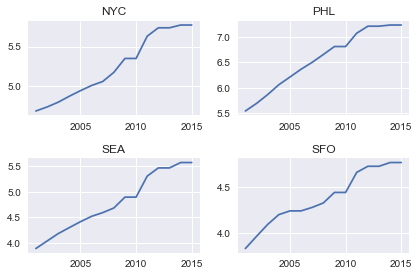
Disability Rates

Next column explored is adult disability rate. This is the percentage of people age eighteen through sixty-four that are currently on disability. Later on, we will use this feature as target variable in our model. The rates were plotted individually and then again by state.

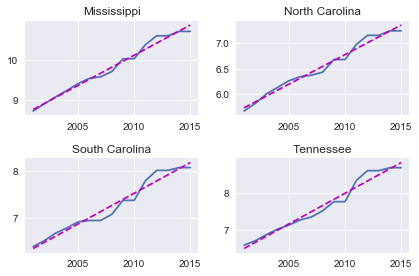
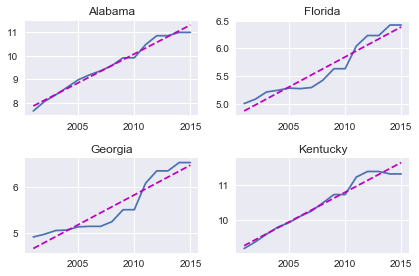
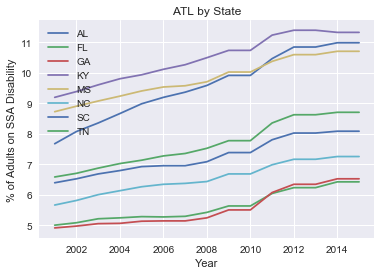




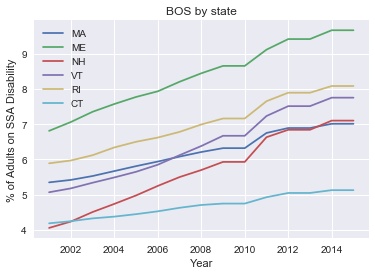


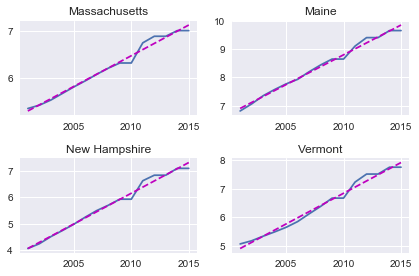


Atlanta Region



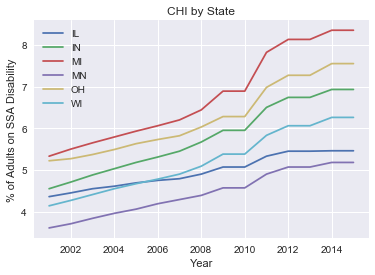
Boston Region

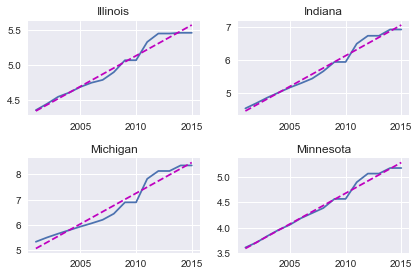


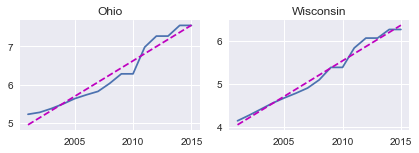




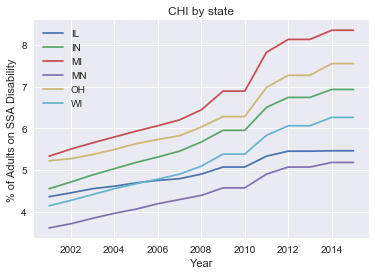
Chicago Region

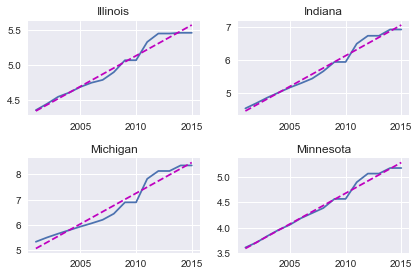


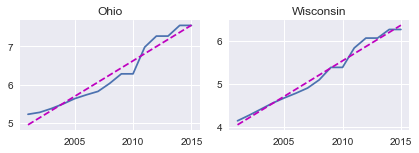




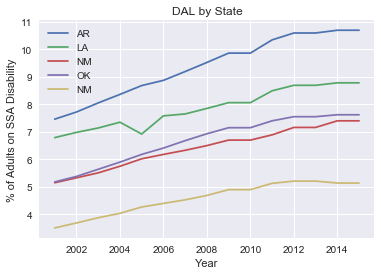
Chicago Region

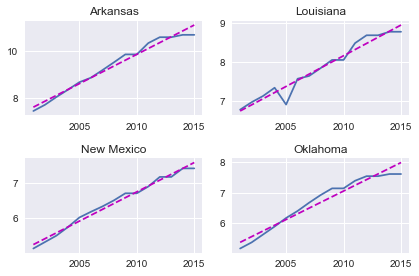


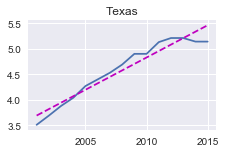




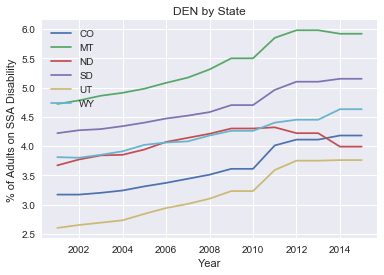
Dallas Region

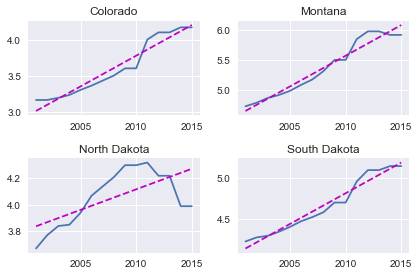


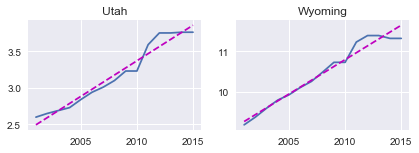




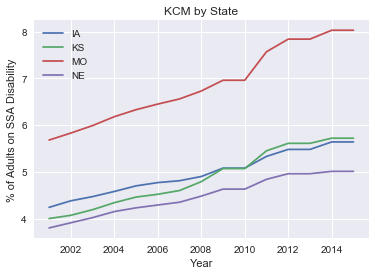
Denver Region

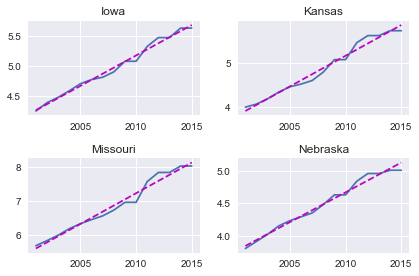




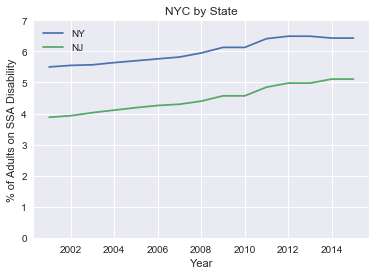


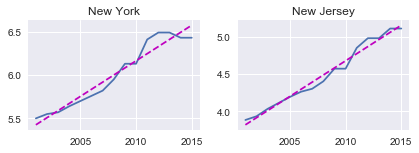
Kansas City Region



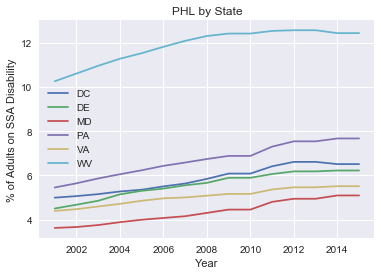


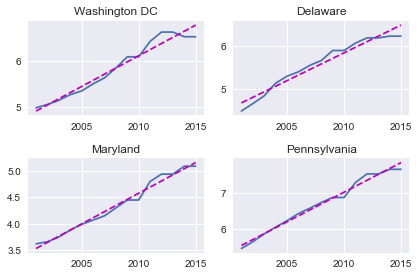
New York City Region

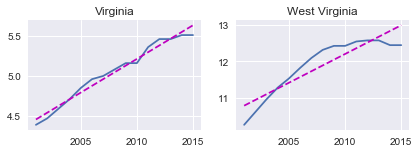




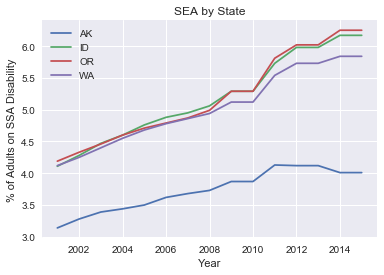
Philadelphia Region

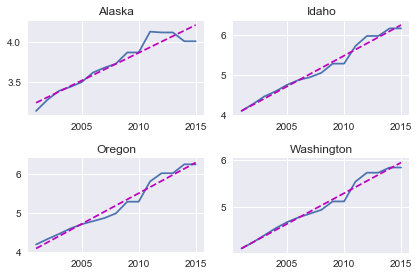




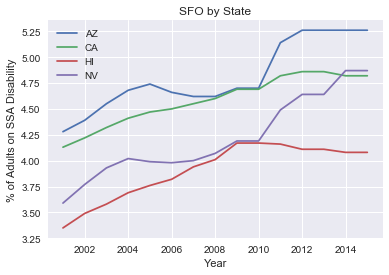


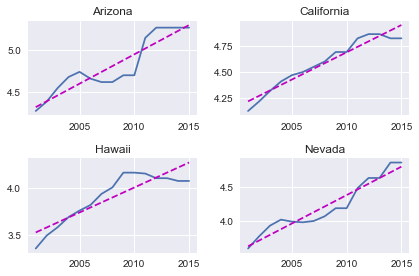
Seattle Region





San Francisco Region





Correlations

Many of the population and percentages of adult disability on SSA disability appear to have some correlation with each other. A correlation Pearson correlation test was done between population and the adult disability rate to see if this relationship was indeed true. The following tables have correlations broken down by both region and within states.

|  |  |  |
| --- | --- | --- |
| **Region** | **Correlation** | **p-value** |
| Overall | -0.092548924 | 0.010434449 |
| ATL | -0.694608449 | 1.38419E-18 |
| BOS | -0.271833772 | 0.009547709 |
| CHI | 0.293177632 | 0.005041785 |
| DAL | -0.641318153 | 5.6913E-10 |
| DEN | -0.481353878 | 1.56513E-06 |
| KCM | 0.870691433 | 1.59024E-19 |
| NYC | 0.905963057 | 5.78157E-12 |
| PHL | -0.220776939 | 0.036521464 |
| SEA | 0.430472128 | 0.000596645 |
| SFO | 0.358349498 | 0.004932691 |

\*Correlation between population and % of adults on SSA disability

|  |  |  |
| --- | --- | --- |
| **State** | **Correlation** | **p-value** |
| Alabama | 0.968425231 | 3.20732E-09 |
| Alaska | 0.986253664 | 1.50407E-11 |
| Arizona | 0.775428211 | 0.000681921 |
| Arkansas | 0.98585493 | 1.80955E-11 |
| California | 0.977188497 | 3.95914E-10 |
| Colorado | 0.949497324 | 6.48721E-08 |
| Connecticut | 0.97267784 | 1.26541E-09 |
| Delaware | 0.991209903 | 8.32079E-13 |
| Florida | 0.922542655 | 9.79027E-07 |
| Georgia | 0.822304881 | 0.000168076 |
| Hawaii | 0.858100765 | 4.26607E-05 |
| Idaho | 0.943296418 | 1.35652E-07 |
| Illinois | 0.825692311 | 0.000149603 |
| Indiana | 0.967193343 | 4.10102E-09 |
| Iowa | 0.977041477 | 4.12651E-10 |
| Kansas | 0.956469273 | 2.51213E-08 |
| Kentucky | 0.974331269 | 8.46726E-10 |

\*Correlation between population and % of adults on SSA disability

|  |  |  |
| --- | --- | --- |
| **State** | **Correlation** | **p-value** |
| Louisiana | 0.779835574 | 0.00060643 |
| Maine | 0.572185694 | 0.02582112 |
| Maryland | 0.98372462 | 4.48096E-11 |
| Massachusetts | 0.982475304 | 7.22512E-11 |
| Michigan | -0.672829392 | 0.005982788 |
| Minnesota | 0.968776376 | 2.98497E-09 |
| Mississippi | 0.967425078 | 3.91854E-09 |
| Missouri | 0.901944375 | 4.30894E-06 |
| Montana | 0.948347477 | 7.4885E-08 |
| Nebraska | 0.992902151 | 2.08107E-13 |
| Nevada | 0.896313699 | 6.10826E-06 |
| New Hampshire | 0.826840797 | 0.000143732 |
| New Jersey | 0.976828525 | 4.37949E-10 |
| New Mexico | 0.961593387 | 1.12693E-08 |
| New York | 0.923586915 | 8.98655E-07 |
| North Carolina | 0.974554129 | 8.00497E-10 |
| North Dakota | 0.453281826 | 0.089716362 |

\*Correlation between population and % of adults on SSA disability

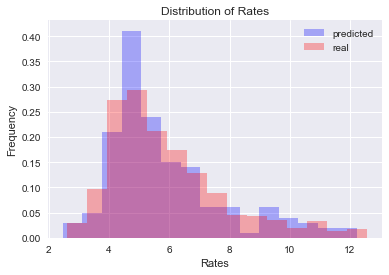
|  |  |  |
| --- | --- | --- |
| **State** | **Correlation** | **p-value** |
| Louisiana | 0.779835574 | 0.00060643 |
| Maine | 0.572185694 | 0.02582112 |
| Maryland | 0.98372462 | 4.48096E-11 |
| Massachusetts | 0.982475304 | 7.22512E-11 |
| Michigan | -0.672829392 | 0.005982788 |
| Minnesota | 0.968776376 | 2.98497E-09 |
| Mississippi | 0.967425078 | 3.91854E-09 |
| Missouri | 0.901944375 | 4.30894E-06 |
| Montana | 0.948347477 | 7.4885E-08 |
| Nebraska | 0.992902151 | 2.08107E-13 |
| Nevada | 0.896313699 | 6.10826E-06 |
| New Hampshire | 0.826840797 | 0.000143732 |
| New Jersey | 0.976828525 | 4.37949E-10 |
| New Mexico | 0.961593387 | 1.12693E-08 |
| New York | 0.923586915 | 8.98655E-07 |
| North Carolina | 0.974554129 | 8.00497E-10 |
| North Dakota | 0.453281826 | 0.089716362 |

\*Correlation between population and % of adults on SSA disability

Hypothesis that the adult population was correlated with the rate of adults on disability returned with mixed results. Although most of the states came back with very high correlations, most of the regions came back with negative and weaker correlations. This was surprising seen as the regions exhibited high correlations when only considering population or rate of disability. There is a chance that both of these variables happen to be correlated due to outside forces. Evidence of this could be seen in Michigan since with high correlations elsewhere, we would expect the rate to go down if it were true. As the sole state with a negative population growth, this may be a telling sign.

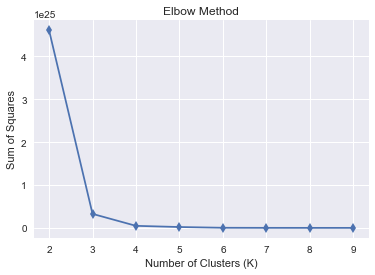
Creating a model

Now that we have explored all 765 rows of our data, it is time to try to apply a model. A linear regression model was applied using all 22 numerical features in the data set to use as a baseline to compare other models. Surprisingly, this model worked better than the other model using the fast correlation method.



|  |  |  |
| --- | --- | --- |
| **Predicted Rate** | **Standard Deviation** | **R2** |
| 5.85396924765 | 1.93172537739 | 0.906105057316 |

Now that we have a model to predict using the whole dataset. Let’s try to break down these numbers into smaller groups. We begin by attempting to cluster the regions together using clustering to see if we build a better model. The ‘elbow graph’ on the following page shows the result of the mean squared error as we increase the number of clustered used. After the point K=3 clusters, the graph flattens off as adding more clusters does not explain much more of the variance.



Clustering on region resulted in three clusters grouped as follows:

* Cluster 0: Dallas
* Cluster 1: San Francisco
* Cluster 2: Atlanta, Boston, Chicago, Denver, Kansas City, New York, Philadelphia, Seattle

We will now apply linear regression to these clusters separately and see if we explain more of the variance.

|  |  |
| --- | --- |
|  | **R2** |
| Cluster 0 | 0.89557477004 |
| Cluster 1 | 0.803402067542 |
| Cluster 2 | 0.936030588144 |

Using the linear regression model by clustering did not have a large impact on the first cluster with just DAL. The SFO region underperformed the model with only 80.3% of the variance explained by the model. The third cluster with all the other values did better by a few percentage points, a decrease of 3% from the original model. Overall, this model could be considered an improvement since we have some regions that we can predict with a higher degree of certainty.

Conclusion

Although feature selection did not improve the result over the model with all features, clustering was able to allow us to build a better model by grouping states together that may not have been intuitive otherwise. Combining these two approaches allows us to use the new models on clusters 0 and 2 and revert to the original model for cluster 1.