GitHub link: https://github.com/jbreiger/Exploring-Dermatology-Hate-Crime-Datasets

Dermatology Dataset

Background

The differential diagnosis of erythemato-squamous diseases is a real problem in dermatology. They all share the clinical features of erythema and scaling, with very little differences. The diseases in this group are psoriasis, seboreic dermatitis, lichen planus, pityriasis rosea, cronic dermatitis, and pityriasis rubra pilaris. Usually a biopsy is necessary for the diagnosis but unfortunately these diseases share many histopathological features as well. Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The values of the histopathological features are determined by an analysis of the samples under a microscope.

In the dataset constructed for this domain, the family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. Every other feature (clinical and histopathological) was given a degree in the range of 0 to 3. Here, 0 indicates that the feature was not present, 3 indicates the largest amount possible, and 1, 2 indicate the relative intermediate values.

Use gradient descent (GD) to build your regression model (model1). Start by writing the GD algorithm and then implement it using a programming language of your choice.

I used gradient descent to find the optimal slope and intercept. I used Age as the predictor variable, and I used disease as the output variable.

Before I created the model, I began by dropping the Age columns that had “?” in the field. To keep consistency throughout the entire modeling process, I dropped the age columns for the whole dataset. Because it was only a few rows, my assumption was that this would lead to a cleaner and more accurate dataset.

I then created a gradient descent model and when I ran it initially with an alpha of 0.01 and the cost was increasing. I therefore lowered the alpha to .0001. Once I did this, the gradient descent model went through 384 iterations.

I got the output of [[2.78212142] [-0.40355736]]. This means that the slope was -0.40 and the intercept was 2.78.

Doing an eyeball check of the data, this slope and intercept seems like it fists the data.

2. Use random forest on the clinical as well as histopathological attributes to classify the disease type (model2).

I created a random forest model. I then ran the model with 2 different sets of attributes. Both set of attributes used the same predictor value of disease because we were trying to predict the disease type. I normalized both sets of X values to begin.

One set of attributes was the clinical set of data. This included the 13 clinical attributes including age.

I ran this random forest algorithm 200 times and took the average accuracy. The average accuracy was 78.1%.

I then ran the algorithm with the other sets of attributes, the Histopathological attributes. There were 21 attributes in the X variable.

I ran this algorithm 200 times as well and took the average. The average accuracy was 85.2%.

3. Use kNN on the clinical attributes and histopathological attributes to classify the disease type and report your accuracy (model3).

For the KNN model, I used the same attribute split as I did for the random forest. One of the attributes lists was the 13 attributes in the clinical side. The other set of attributes was the Histopathological attributes. The predictor value was still disease.

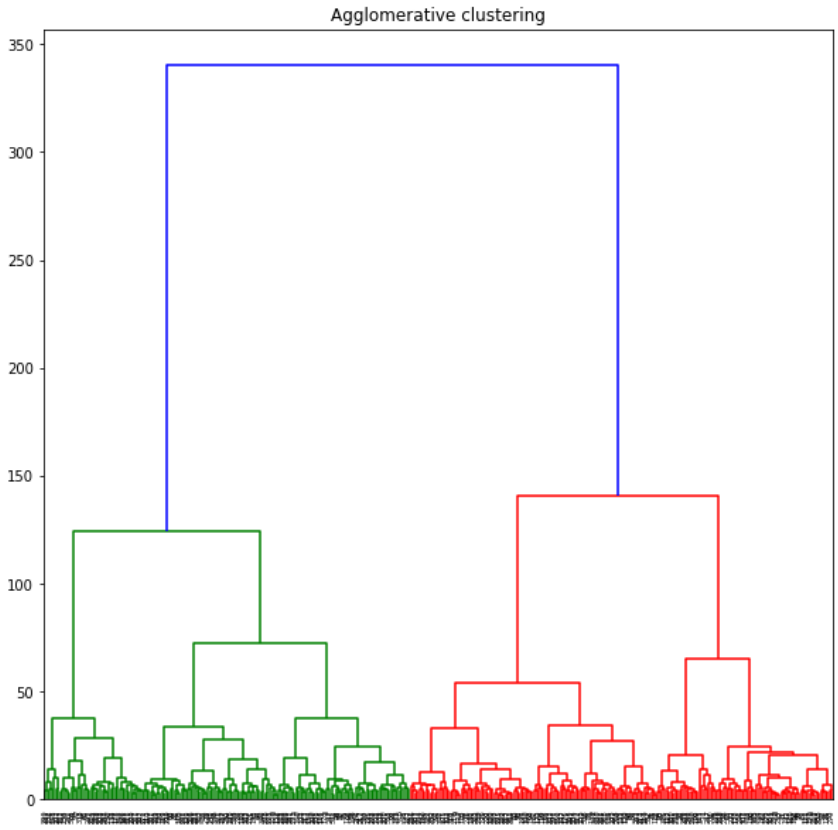
For the KNN, I ran the KNN model with K values ranging between 2 and 10. Accuracy was between 80.8% and 81.1% with the best accuracies appearing when k was either 4 or 5.

I then did the same thing with the histopathological data. The accuracy ranged from 91.0% when k was equal to 2 to 93.1% when K was equal to 6. On average it was 92.1% accurate across the K values.

4. Finally, use two different clustering algorithms and see how well these attributes can determine the disease type (model4 and model5).

For the clustering algorithm, I did not separate out the attributes into histopathological and clinical sections. I included every attribute including disease in my model. I included disease in my model to try and see how the other attributes clustered around each specific disease.

I began by running an Agglomerative clustering model. I used ward's model to do this. When I ran the model, I got 3 clusters. One of these clusters consisted of just one bar. The other two clusters seemed to have relatively similar number of records inside of them.

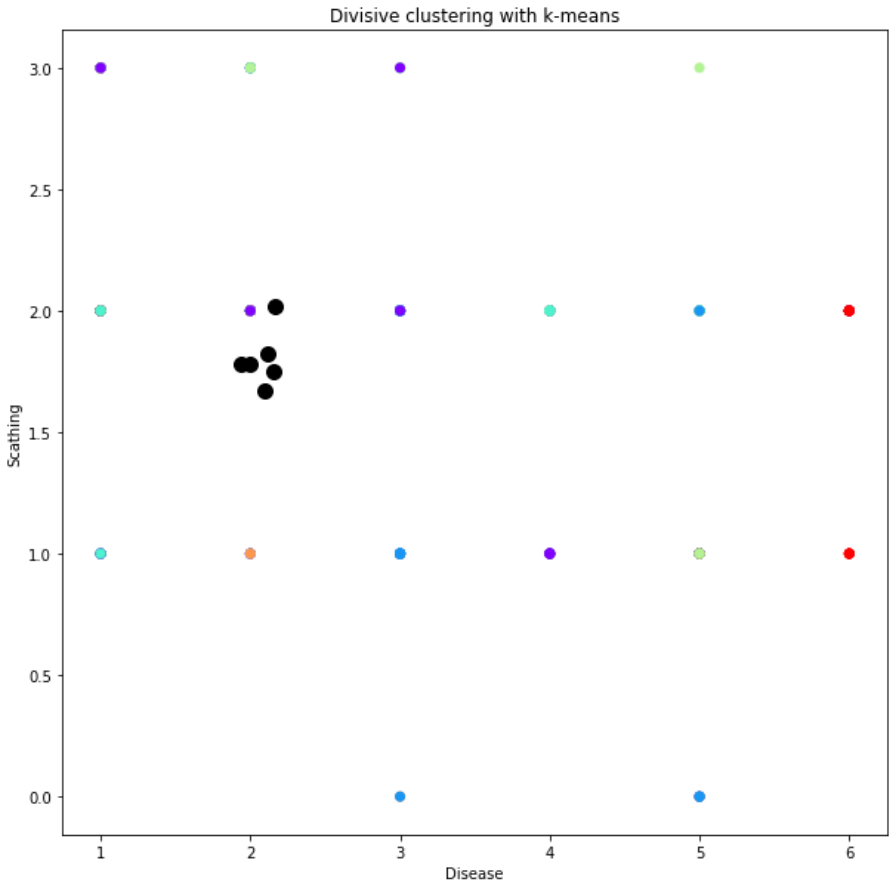


I then ran a divisive clustering model. I ran this with 6 clusters because there are 6 disease types, so I was curious if there would be 6 distinct clusters that we would see around each disease type.

For this visual, I included disease on the X-axis and scathing on the Y axis although different attributes could have been selected. I selected disease on the X-axis to try and see if there were patterns to the clustering around disease type.

The centroids can be seen in the visual in black. They are clustered very close together.

Although there are some colored clusters patterns, overall, there are not very many distinct patterns to the clusters. Although there are some clusters for disease 6 and disease 1, they are not very profound.



Now, compare and contrast the five models you built. Having done both classification and clustering on the same dataset, what can you say about this data and/or the techniques you used? Write your thoughts in 2-3 paragraphs. [10 points]

Because disease type is not a continuous variable running gradient descent probably does not make sense to use. This is because a higher disease number does not mean anything except about how to classify the disease. Therefore, I do not think we should look to much into this model although in general the model did not seem extremely inaccurate.

Both the classification models of the random forest and the KNN model were in general pretty accurate about predicting the disease type. Each of the models created an accuracy with at least 78% accuracy.

Overall, the KNN models were more accurate than the random forest. The histopathological attributes were also more accurate than the clinical data. The most accurate model was the histopathological KNN which was accurate over 90% of the time when K was 6.

The clustering algorithm was much less reliable than the classification models. The agglomerative clustering algorithm clustered into 3 groups, with one group only having one record inside it. This is not as useful as the other models in predicting disease type. The divisive model did not show many distinct patterns for clusters by disease type.

If I were to choose one strategy for determining disease type, I would use the KNN model and focus on using histopathological attributes.

Hate Crimes Dataset

Background

Attached is a dataset downloadcontaining statistics of hate crimes that happened within 10 days of 2016 US elections. In that period, nearly 900 hate incidents were reported to the Southern Poverty Law Center, averaging out to 90 per day. By comparison, about 36,000 hate crimes were reported to the FBI from 2010 through 2015 — an average of 16 per day.

The numbers we have here are tricky; the data is limited by how it’s collected and can’t definitively tell us whether there were more hate incidents in the days after the election than is typical. What we can do, however, is to look for trends within the numbers, such as how hate crimes vary by state, as well as what factors within those states might be tied to hate crime rates.

How does income inequality relate to the number of hate crimes and hate incidents?

I began by looking through the data and found that there were some blanks. For the blanks that were in both the SPLC and the FBI column I decided to delete the entire row of data. This is because I ultimately wanted to be able to compare the 2 different hate crime stats and it would not be an even comparison if some did not have all their data points. There were also blanks in the share unemployed seasonal field. I decided to take the average of the remaining records in the column and include them for this field. Because there was only 3 of these data points it would not impact the data that much and would still allow me to use the other parts of the row.

For the first two questions I chose to use linear regression. This is because it is a continuous variable as the response variable. I began by choosing variables which are related to the research question.

For the first question, these are variables are relating to income inequality. I excluded all columns that did not have to do with wealth. I chose the variables 'share unemployed seasonal', 'share white poverty', ‘Gini index', and “median household income".

Instead of using gradient descent, I decided to run a regression model with all the variables and then incrementally remove the ones that are not significant based on the p values.

I ran a regression model with the previously mentioned attributes and SPLC hate crimes as the response variable.

When I analyzed the regression output, I noticed that the p-values were very high for median household income. I therefore removed it from the model and ran the regression again. I repeated this process again and saw that the unemployment numbers were also not significant.

I was therefore left with the two variables, share white poverty and gini index.

The SPLC regression line was -2.8 * white poverty + gini index * 4.03 + -1.26.

I then followed a similar process with the difference being I used the hate crimes by the FBI as the response variable instead of the SPLC field. As I removed variables incrementally based on the p values, I ended up with the same two variables as with the SPLC data.

FBI regression line was -13.07 * white poverty + gini_ ndex * 37.5+ -13.07.

For both the FBI and the SPLC data white poverty was negatively correlated with hate crime incidents. This means as there is a larger share of white poverty then there are less hate crimes. The Gini index is positively correlated with both the SPLC and the FBI which means that the higher the Gini index (which means more inequality) then the more hate crimes can expect.

2.How can we predict the number of hate crimes and hate incidents from race/nature of the population?

To be able to predict the number of hate crimes based on the population, I chose attributes relating to race and background including share non-citizen and share nonwhite. I also included variables which I thought would explain about the nature of the population such as their political leaning and education. This included share population with high school degree and share voters voted trump.

I then ran a regression model with SPLC hate crimes as my response variable. Like question 1, I ran an iterative process of removing variables until I only had variables with significant p values left. For the SPLC regression model I ended up with share non-citizen and share voters voted trump as my two significant variables.

For the SPLC data, the regression line was share_non_citizen * -2.04 + share_voters_voted_trump * -1.8.

I then repeated a similar process with the FBI field as the predictor output and eliminating outputs with different P values.

I got that the regression line for the FBI data was share population with high school degree * 1.1284 + share voters voted trump * -1.3883 + -0.001.

Now that we have these regression models, we are able to make predictions about the number of hate crimes.

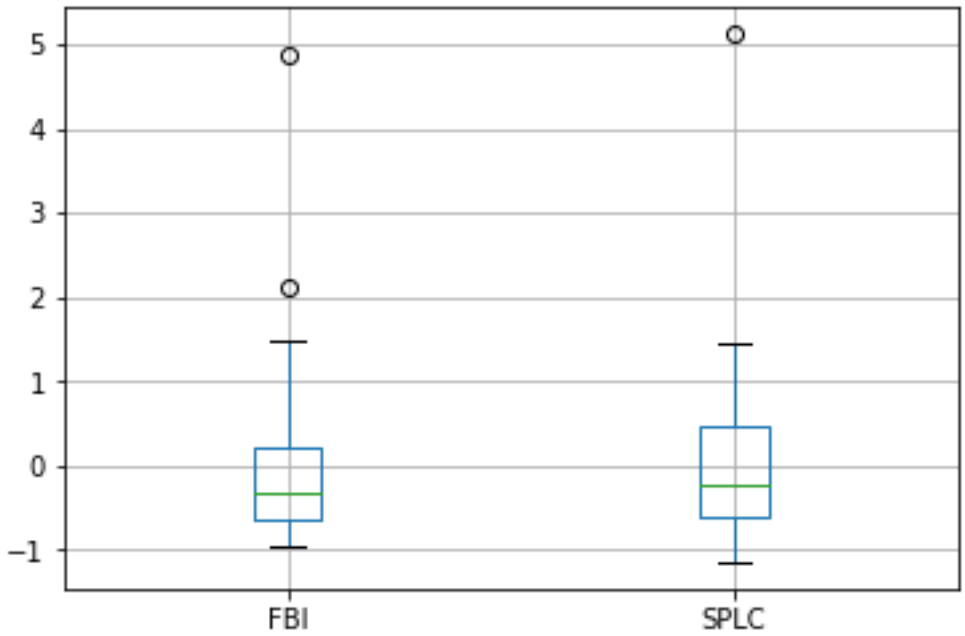
For example, if you want to predict the number of hate crimes according to the FBI based on the race and nature of the population, you can input the values for population with high school degree and the percentage of voters who voted for trump into the regression equation. You will then get a prediction for the number of hate crimes for that particular state.

How does the number of hate crimes vary across states? Is there any similarity in number of hate incidents (per 100,000 people) between some states than in others — both according to the SPLC after the election and the FBI before it? [10 points]

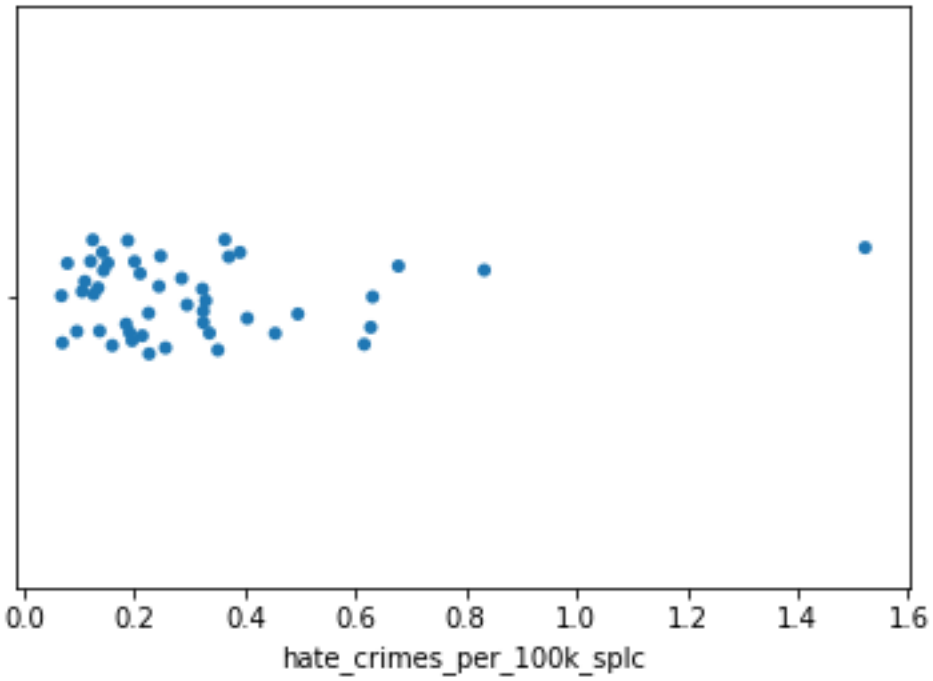
To answer this question, I decided to use data visualization to try and find groupings for the number of incidents between the FBI and the SPLC.

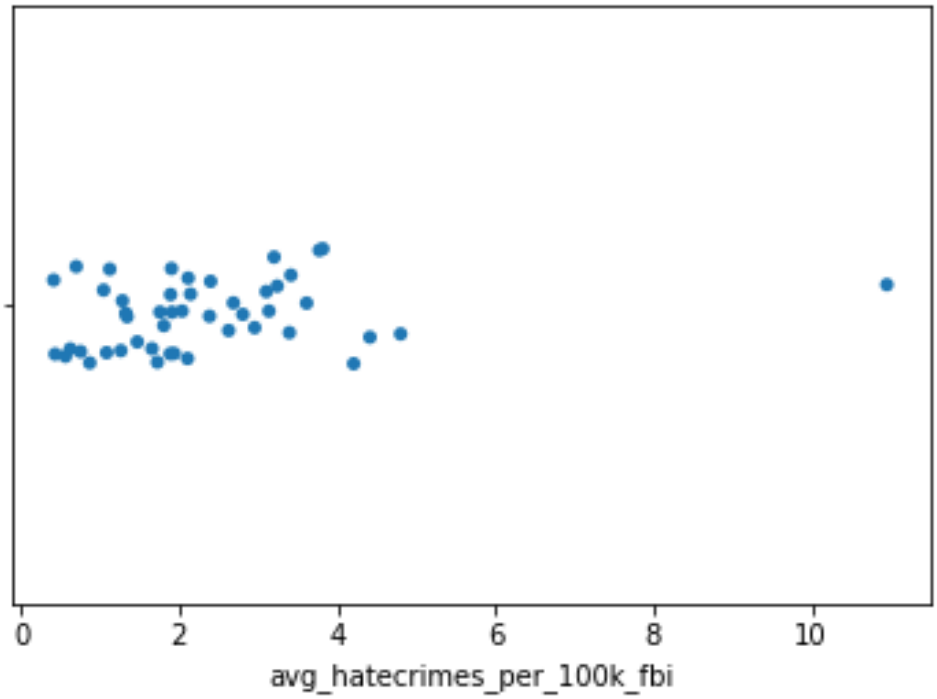
I began by looking at the variance of the number of hate crimes for both SPLC and the FBI. Because they are on different scales, I normalized the scales to compare the results better.

I created a boxplot to look at this information. It looks like the FBI has slightly more variance in terms of the amount of hate crimes reported compared to the SPLC.

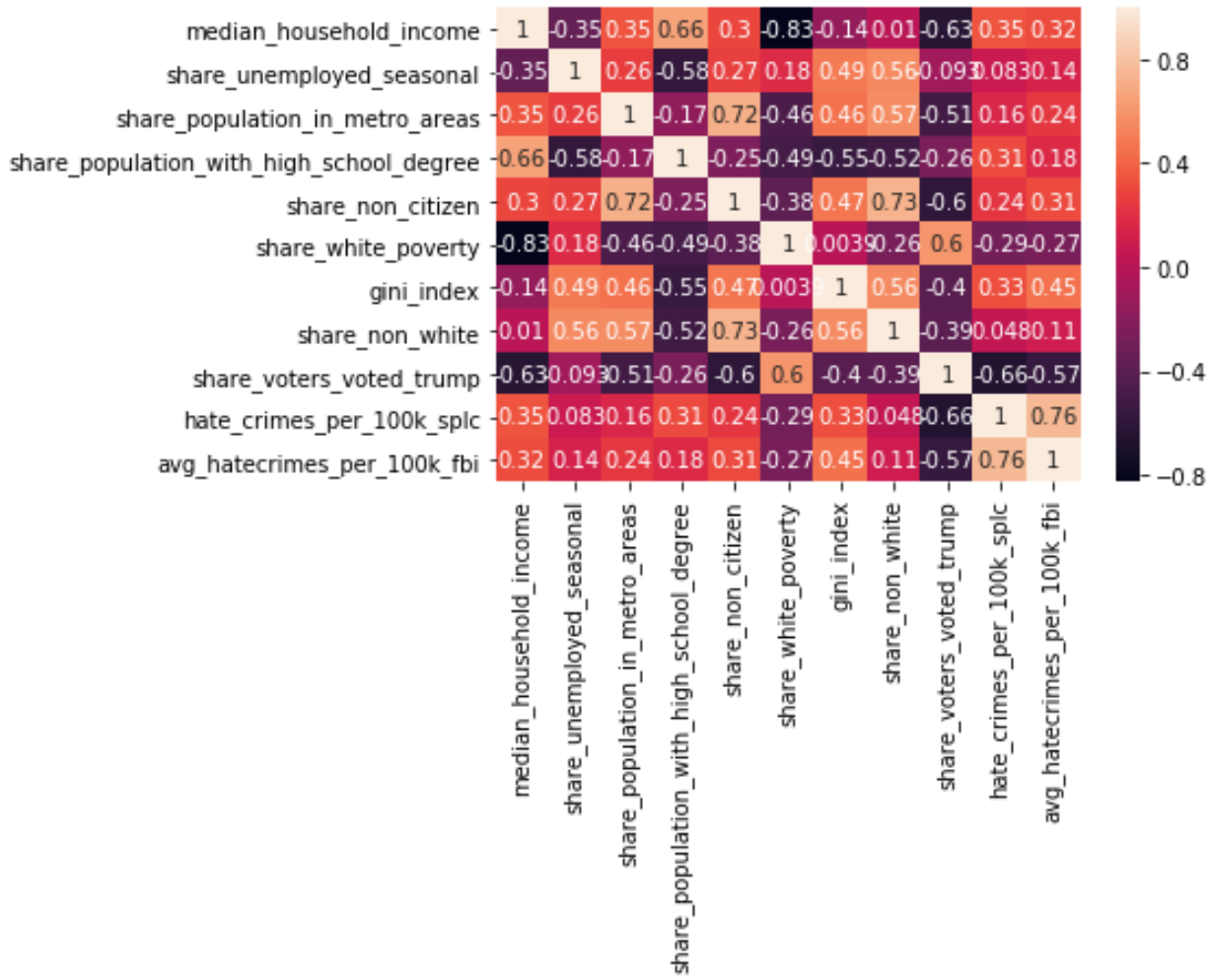


To look at individual points of how the data is spread I next created two different strip plots looking at both the SPLC and FBI data. These allow for some jitter on the Y axis so you can better see the number of results of the predictor variable. I created one plot for both FBI data and one for SPLC.





Based on those visualizations, it looks like to me there was 3 distinct clustering around the number of hate crimes. One group is closer to 0, one group is clustered closer to the middle, and then there is one group which is just an outlier that is far off to the right of the graph.



Finally, I created a correlation matrix to see how different features were correlated with both hate crimes for FBI and for the SPLC. I did this to try and find relationships between the states.

The darker colors in the visual above mean that there is a large postive correlation and the lighter colors mean there is a negative correlation.

One of the big takeaways from this correlation matrix was similar to what I found with the regression equation in problem 2, the share of voters who voted for Trump was strongly negatively correlated for both FBI and SPLC.

This means it is likely there is a grouping of states who voted for Trump and a grouping of states who dind't vote for Trump.