An Investigation of State Hate Crime Data

Statistics 202B Final Project

Wesley Cheng

Abstract

This project explores the topic of hate crimes in the United States. The original article on this topic finds that higher rates of hate crimes are closely associated with income inequality both pre and post election. For the purpose of this project, we investigate the factors that play into the the number of hate crimes post election. Hate crime data from 47 states was analyzed and techniques such as principal component analysis, independent component analysis, and factor analysis provide us visual aids in understanding the underlying structure of our hate crime data. Predictive modeling was also applied to the data set although we find that there are some factors limiting our full potential. Nonetheless, our predictve modeling results seem to be inline with the original author's findings that income inequality are associated with higher hate crimes post election.

Introduction

The original motivation for this project and choice of data set comes from an article on FiveThirtyEight, ESPN's well known blog that covers subjects ranging from sports to politics to pop culture. The article, *Higher Rates Of Hate Crimes Are Tied To Income Inequality* by Maimuna Majumder [1], investigates the potential factors and predictors of hate crimes and hate incidents across the fifty states (plus the District of Columbia). Hate crimes and incidents include Anti-Immigrant, Black, Muslim, LGBT, Women, Semitism, Trump, and White Nationalism. Using multiple regression and backward step-wise elimination, Majumder looked into the most significant determinants of hate crimes across the Untied States both before and after the controversial 2016 United States Election. Using existing literature on factors tied to neighborhood violence to choose the variables included in her data set, Majumder found that income inequality and percent population with a high school degree, were the two most significant variables in determining the number of hate crimes in both pre and post-election America. Of course, Majumder does not claim causation between these variables and hate crimes, as correlation does not imply causation, but her analysis strongly piqued my interest in the topic after reading it.

In a country that seems to become more politically polarized by the passing year, hate crimes have been on the rise over the past year [2]. The first purpose of this project is to implement some of the methods covered in class to gain more insight into the factors that might drive hate crimes. Instead of only looking at each variable as having an independent impact on the number of hate crimes as Majumder did, I investigate if there are combinations of variables that might explain a large percentage of the variation in the number hate crimes across the

states and if there might be logical groupings of these variables. If this is possible, it might give us a better visualization and intuitive grasp of the hate crime data.

The second purpose of this project is to see if it is possible to develop a model to predict the number of future hate crimes. There are many practical applications of a potential predictive model like. For one, states with a high number of predicted hate crimes may be able to better allocate state resources to law enforcement and policy making in detering future hate crimes. It is possible that some of the variables for the states might change over the course of the course of the Trump presidency. The author mentions the possiblility of increased economic disparity during the Trump administration, and if we create a predictive model, we might be able to numerically quantify the potential increase in hate crimes. It would also be interesting to see if it possible to statistically model the relationship between various demographic variables and the number of hate crimes.

Data and Exploratory Data Analysis

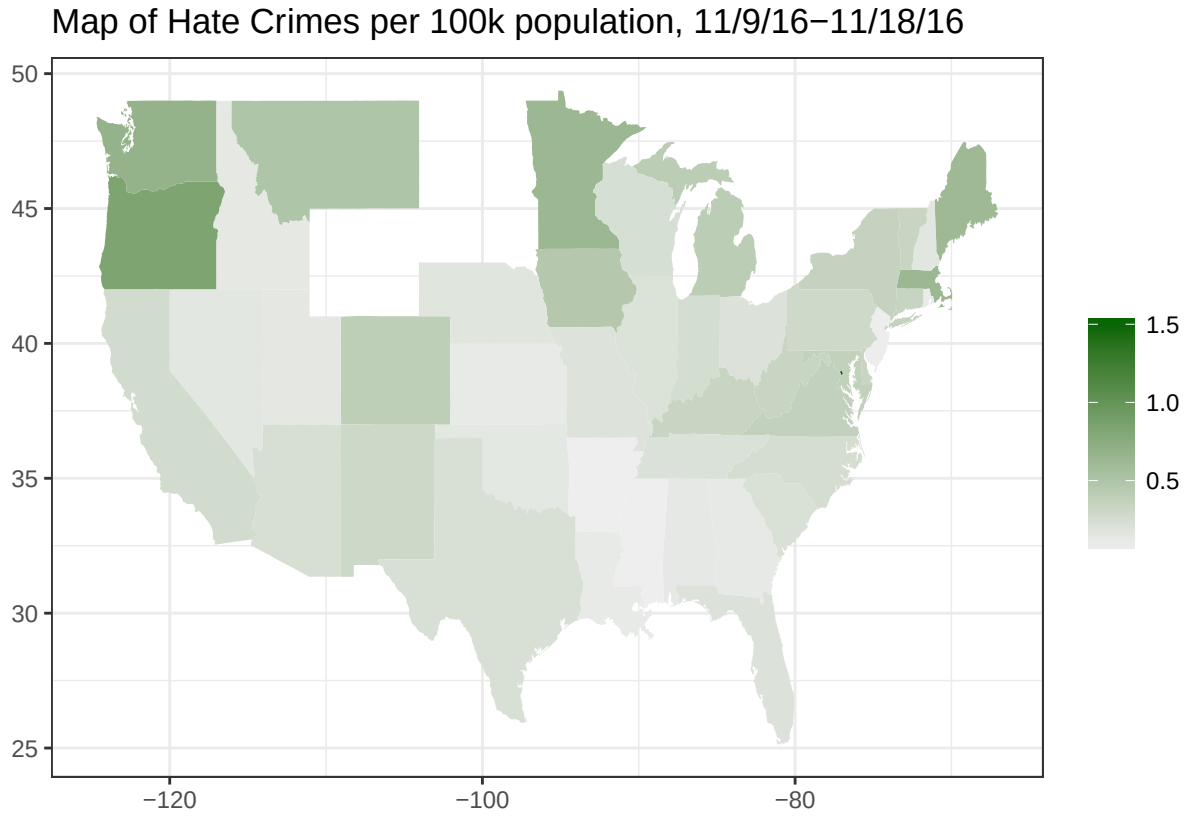
The data set is obtained from the data repository of FiveThirtyEight's GitHub [3]. Data was primarily sourced from the Kaiser Family Foundation site. The outcome variable of interest here is the number of hate crimes per 100,000 population the 10 days after the 2016 election (Nov. 9-18, 2016), and it comes from the Southern Poverty Law Center. The data set contained 51 observations (50 states plus the District of Columbia) and 11 columns (state name, 9 covariates, 1 outcome variable). A table of the variables of interest can be seen below.

Variable	Definition
state	State name
median_household_income	Median household income, 2016
share_unemployed_seasonal	Share of the population that is unemployed (seasonally adjusted), Sept. 2016
share_population_in_metro_areas	Share of the population that lives in metropolitan areas, 2015
share_population_with_high_school_degree	Share of adults 25 and older with a high-school degree, 2009
share_non_citizen	Share of the population that are not U.S. citizens, 2015
share_white_poverty	Share of white residents who are living in poverty, 2015
gini_index	Gini Index, 2015
share_non_white	Share of the population that is not white, 2015
share_voters_voted_trump	Share of 2016 U.S. presidential voters who voted for Donald Trump

Variable	Definition
hate_crimes_per_100k_splc	Hate crimes per 100,000 population, Southern Poverty Law Center, Nov. 9-18, 2016

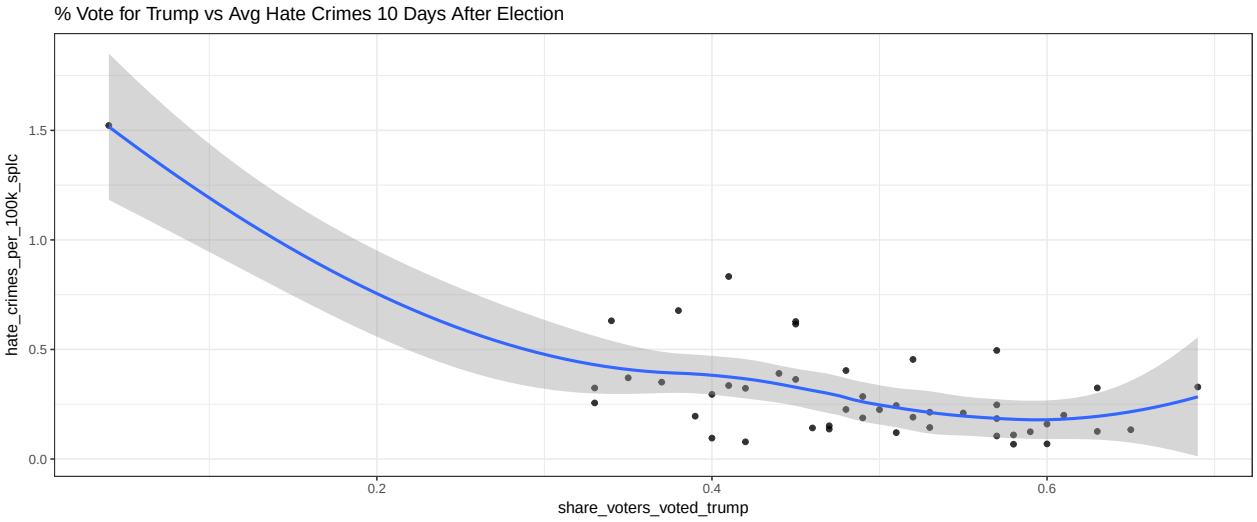
There was not much data cleansing or preperation needed. Hawaii, North Dakota, South Dakota, and Wyoming were removed from the data since there was no available outcome data for these states. Maine and Mississpi were missing non-citizen information, however upon visiting the Kaiser Family Foundation website directly [4], it seems that those values had uploaded since the data was pushed onto GitHub, so I simply imputed these values directly into the data set.

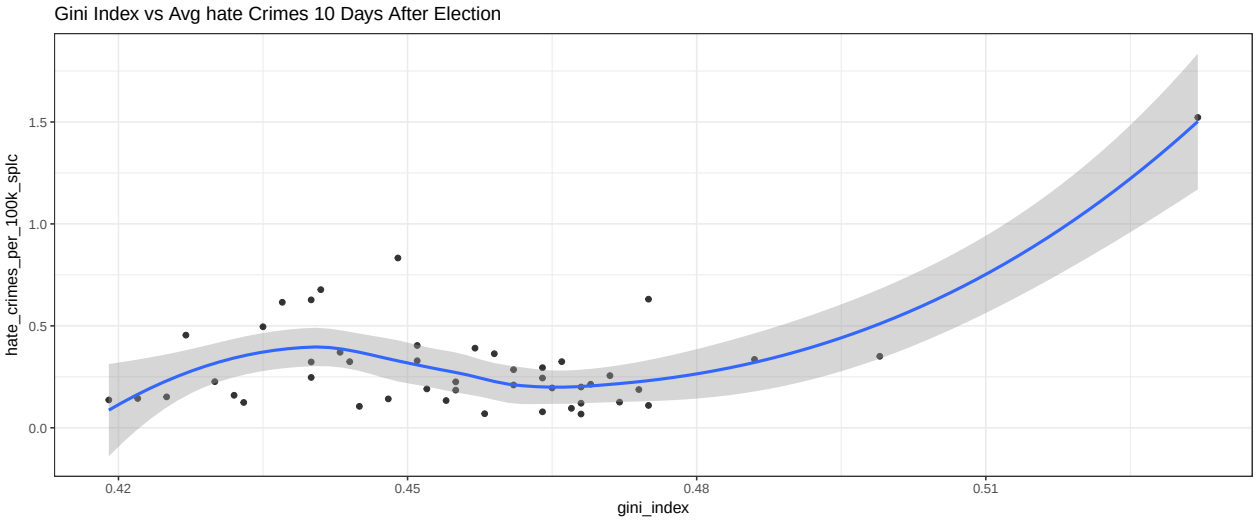
A plot of the hate crimes is shown below. It is evident that in the 10 days after the election, some states exhibited a larger number of hate crimes compared to other states. Oregon and Washington in particular had a higher number of hate crimes and interestingly enough they both had a low share of voters that voted for Trump at 41% and 38% respectively. The map does not make the District of Columbia of visible, but the District of Columbia had an astounding 1.52 hate crimes per 100k, accompanied by a 4% share of votes for Trump.



A more visual view of the relationship between the the percent voted for Trump and the average number of hate crime after the election is shown below. I also show a plot of the relationship between the Gini index and number of hates crimes post-election since author mentioned that she found income inquality to be the most important determinant of

population adjusted hate incident. The Gini coefficient measures inequality by assigning a value from 0 to 1, 0 being that everyone has the same income, and 1 representing maximal income inequality. We indeed a see a negative trend as the vote percentage for Trump increases and a positive trend as the Gini index increases. However, it is also clear that in both the plots there are high leverage points to keep in mind when evaluating the strength of the said relationship.





Data Analysis and Discussion

In this section I will apply various techniques to the hate crime data set to gain a deeper understanding of the varaibilty within our data set and ultimately develop a model to predict future hate crimes. I also describe the scientific and statistical issues that accompany the methods that I apply to the hate crime data set.

Dimension Reduction

Principal Component Analysis

Our discussion first begins with the classical dimension reduction technique, principal component analysis (PCA). Recall, PCA aims at finding new principal components directions that attempt to descirbe the most variation present in a data set. The principal components are defined to be orthogonal, thus uncorrelated and are linear combinations of the existing columns of the original matrix. Perhaps we would like to see if these new variables can differentiate between the different types data points in our data set. For the purpose of demonstrating PCA, let us see if we can visualize the difference between a state of high and low hate crime activity based off the data that we have. Here, "high" and "low" is determined by whether or not a particular state exhibited a higher than national average amount of hate crimes in the 10 days after the election.

	PC1	PC2	PC3
median_household_income	-0.216	0.4898	-0.04462
share_unemployed_seasonal	-0.2066	-0.384	-0.2962
share_population_in_metro_areas	-0.4305	0.01314	0.3528
share_population_with_high_school_degree	-0.1052	0.5243	-0.2041
share_non_citizen	-0.4549	-0.000855	0.3544
share_white_poverty	0.2922	-0.4016	-0.09208
gini_index	-0.327	-0.2811	-0.5116
share_non_white	-0.4091	-0.2161	0.2147
share_voters_voted_trump	0.3841	-0.225	0.5497

Ranked by Trump Vote Share	Rank by Non Citizen Percentage
West Virginia	Maine
Oklahoma	Mississippi
Alabama	Montana
Kentucky	Vermont
Tennessee	West Virginia
Arkansas	Alabama
Nebraska	Louisiana
Idaho	Missouri
Louisiana	Indiana
Mississippi	Iowa

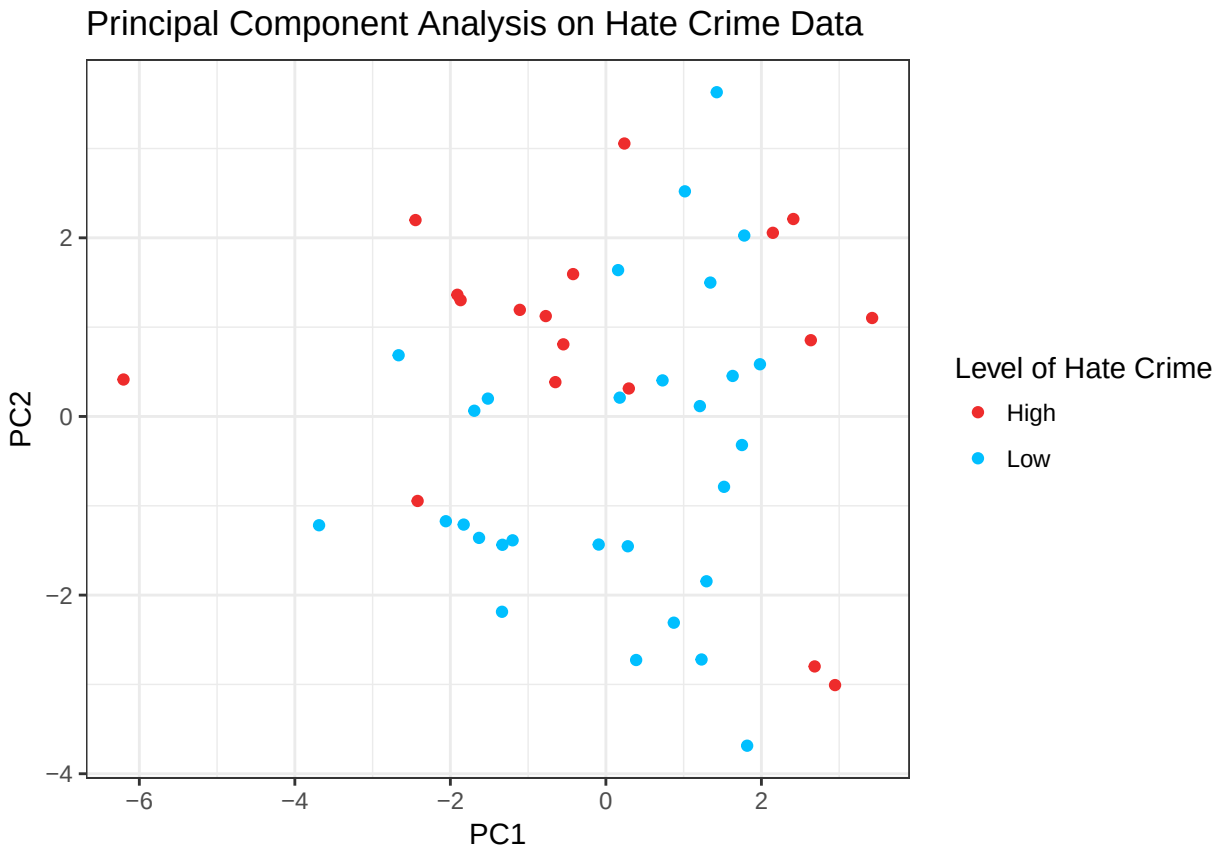
If we examine the principal components particularly the first two or three, since they explain a majority of the variation in the data, we can see that it is quite difficult to interpret each component. This is one disadvantage of PCA, when there is no clear identifiers of each principal components. Nonetheless it is still possible to extract some meaningful insights

from these principal components. The first table above shows the weights that we will use to transform our original variables into our first three principal components. For example points with a high PC1 value are associated with states with a higher share of Trump voters and lower percentage of non citizens. Intuitively this might make some sense. A quick glance at the second table shows that, West Virginia, Alabama, Lousiana, and Mississippi all rank in the top 10 in terms of Trump vote share and share of non citizens. From this example, we are able begin visualizing how PCA is extracting variance from the data set by finding these linear combinations of our original variables and creating new principal components.

The table below shows the percent of variation in the data that we are able to explain as we add more and more principal components. We can see that with just three principal components, we are able to explain 83% percent of variation in our original data! Therefore for the purpose of dimension reduction, instead of looking at 9 variables, one could instead look at the first 3-5 to principal components and know that these principal components are explaining a majority of the variation in the data and altogether much less cumbersome to deal with.

PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
0.4251	0.7625	0.8305	0.8892	0.9291	0.9611	0.9764	0.9902	1

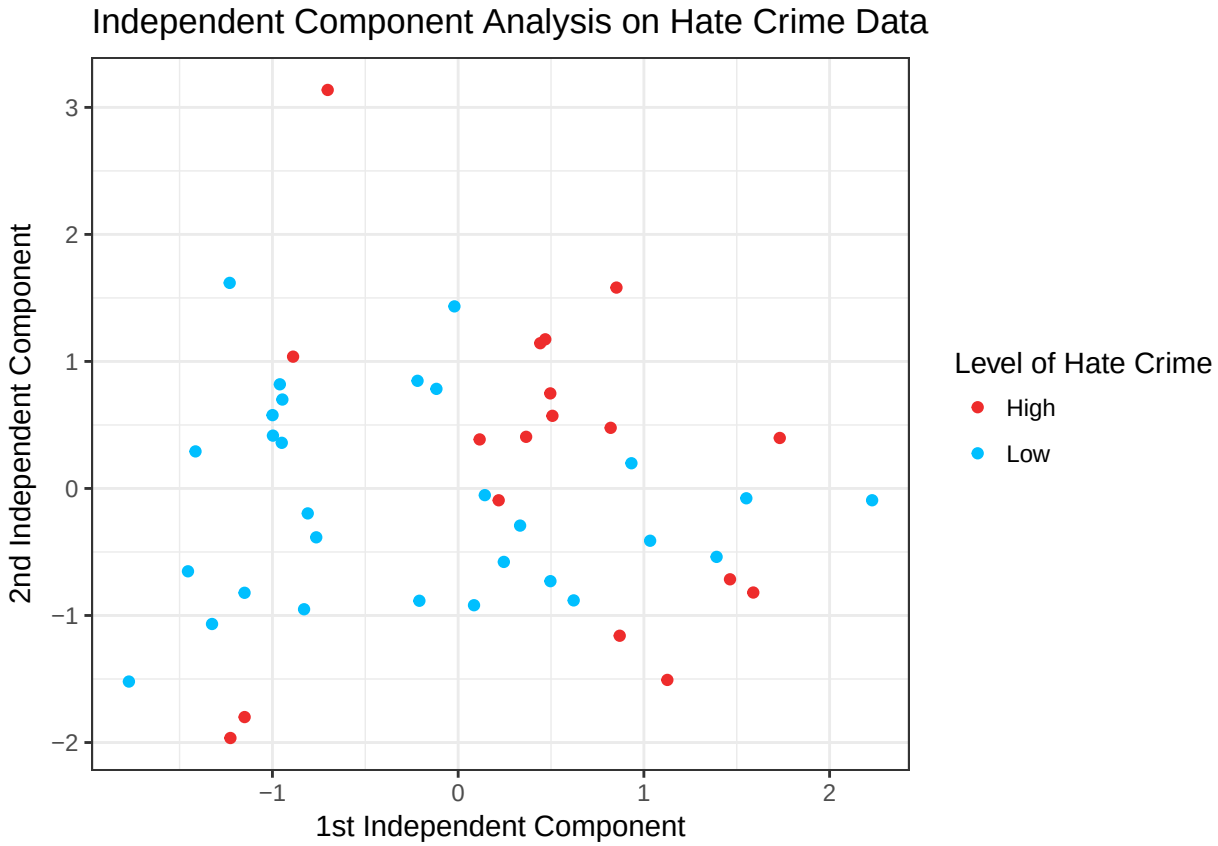
One of the greatest benefits of PCA is that ability to visualize the data in low dimensional subspace, without losing a proportional amount of information in the data. If we matrix multiply our scaled original matrix with the our princpal components and take the first two columns of the resulting matrix, we can plot this on the two-dimensional space that we are used to. The plots below shows our data represented with just two principal components. While we do not have perfect seperation in the data regarding the level of hate crimes, we do see that negative 2nd principal components values are a strong indicator of states with a low level of hate crimes after the 2016 election. Here, instead of looking at nine variables seperately, we are able to reduce this to just two dimensions and gain visual insight into some of the differences between states of high and low levels of hate crime.



Independent Component Analysis

One of the main disadvantages of PCA is that the each principal component must be orthogonal to all other components. By imposing this structural constraint, we limit ourselves in determining new informative dimensions. Independent component analysis (ICA) is similar to PCA in that it aims to describe the data a lower number of dimensions, but instead of assuming these new components are orthogonal to each other, we assume they are independent. It is often a good idea to see if ICA complements the results of PCA in order to confirm our PCA results.

Using the `fastICA` package in R, we can quickly implement ICA.



As we can see from the plot above, ICA seems to confirm our PCA findings. Similar to PCA, we can see that a negative 1st independent component values are associated primarily with states with low levels of hate crime. Additionally, it seems like data points in the top right quadrant are associated with primarily a high level of hate crimes. So while PCA and ICA do not do a perfect job in differentiating between states of high and level of hate crime, we see how by performing dimension reduction, we can visualize some differences between the two types of states that we defined.

Factor Analysis

We now move into our last dimension reduction technique, known as factor analysis. Factor analysis attempts to express the varaibility of the observed variables in terms of latent factors which are unobserved. The hope is that these latent factors might be a more broad concept that the observed variables describe. Statistically speaking, we attempt to estiamte the original correlation matrix of our observed variables.

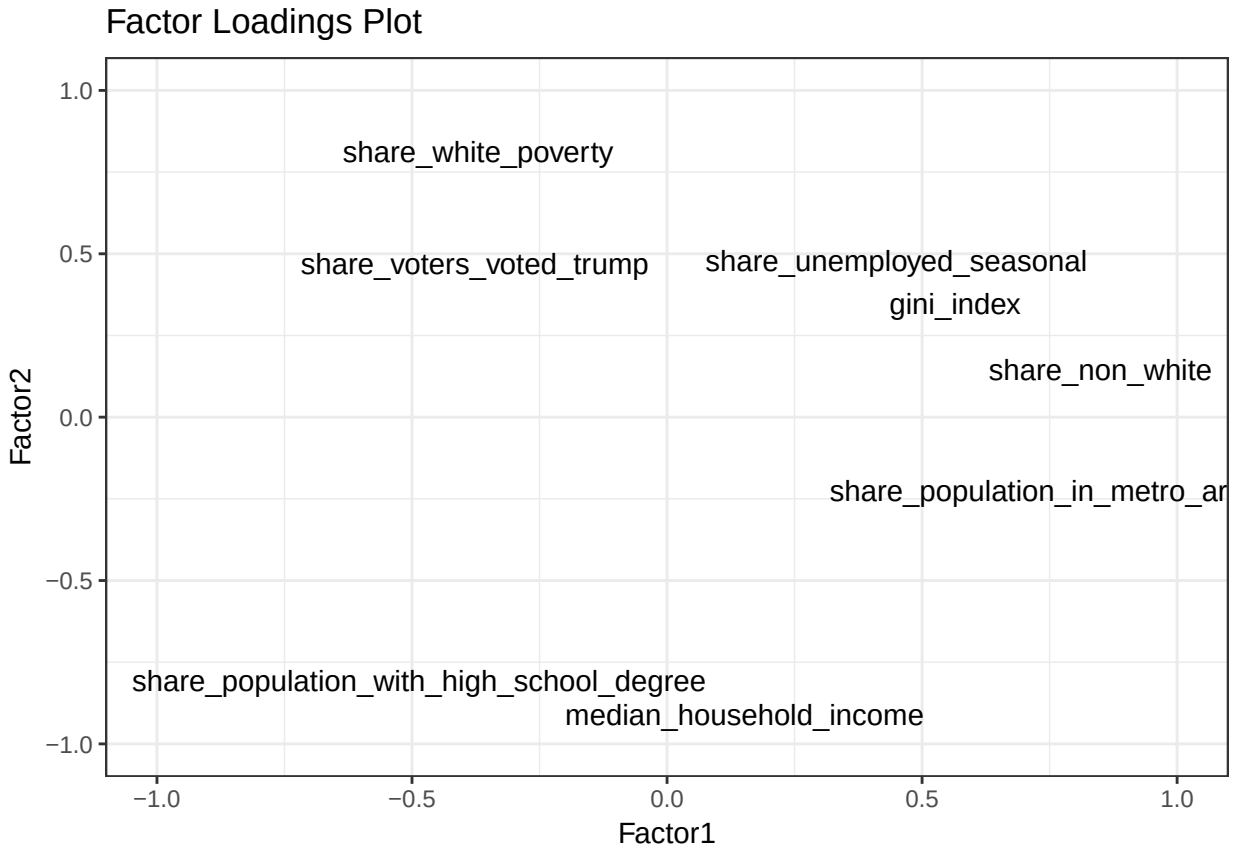
$$\hat{\Sigma} = \hat{\Lambda}\hat{\Lambda}^T + \hat{\Phi} \approx \Sigma$$

where $\hat{\Lambda}$ is the matrix of our factor loadings, with the number of factors being decided by the statistician. Luckily the function `factanal` includes a chi-squared hypothesis test for whether the specified amount of factors is sufficient which can help guide our decision in choosing an appropriate number of factors.

The motivation of running factor analysis other than the aforementioned dimension reduction motivation, is to see if we can create logical groups of certain variables in the hate crime data set. Suppose we hypothesize that there might be an underlying general factor such as 'restlessness' that drives the number of hate crimes after the election. It would be interesting to see if some of the variables reasonably fall into groups like these. If there is, then we might be able to say we have an effective indirect way of measuring something that is inherently difficult to measure in practice.

After running a factor analysis with with two factors, we find that your hypothesis test rejects the hypothesis that two factors are sufficient with a p-value of 0.00766. After a follow-up factor analysis with three factors, we fail to reject the hypothesis that the three factors are sufficient with a p-value of 0.0535. The table and plot below shows the factor loadings for the three factor model. High loadings for share_population_in_metro_areas, share_non_citizen, and share_non_white in the first factor suggest that this factor might describe some measure of "liberalness" within a state, because liberal states tend to be dominanted by large cities with a large contigent of non citzens and non whites. Take one of the most liberal states in California for example. California is largely characterized by its large ethnically diverse cities in Los Angeles, San Francisco, and San Diego. These three metropolitan areas largely dominate the state's liberal political leanings as a large conceneration of people live in these areas. Similarly, the second factor which as high negative loadings in median_household_income, share_population_with_high_school_degree, and a high positive loading in share_white_poverty, might be a indirect measure of "living affordability" as areas with higher median incomes tend to have a higher cost of living versus areas with high poverty rates.

Variable	**Factor1**	**Factor2**	**Factor3**
median_household_income		-0.91	
share_unemployed_seasonal	0.45	0.48	
share_population_in_metro_areas	0.76	-0.22	
share_population_with_high_school_degree	-0.49	-0.81	
share_non_citizen	0.82		0.24
share_white_poverty	-0.37	0.81	
gini_index	0.56	0.35	0.44
share_non_white	0.85		
share_voters_voted_trump	-0.38	0.47	-0.80



The second table below, shows how much variance is explained in the data by the three factors. We can see that the three factors combined explain about 76% of the variation in the original data, which may be high or low depending on whom you ask. Altogether, I would conclude that this is a fairly decent model that summarizes the observed variables in our data fairly well. We see that we are able to make some logical groupings with the variables in our data, and while the latent variables suggested are not nessecarily difficult to measure in practice, they provide insight into data that we might want to collect instead of the original varables, if data collection is expensive or time consuming. It is interesting to see how factor analysis is able to logically group certain variables together by examining the correlation structure. However, recall that throughout our factor analysis, many steps taken in the procedure were fairly subjective. Determining the number of latent factors, factor loading cutoffs (I used 0.2 above), and interpreting the latent factors can almost be considered an art, as various plausible interpretations could probably be made for this particular data set. Nonetheless, the hope of this particular analysis was to show one particular way that we might be able to analyze our particular hate crime data set and provide an supplementary analysis to the authors' original findings.

	Factor1	**Factor2**	**Factor3**
SS loadings	3.02	2.82	1.03
Proportion Var	0.34	0.31	0.11
Cumulative Var	0.34	0.65	0.76

Predictive Modeling

The second portion of the data analysis and discussion section deals deals with the topic of performing predictive modeling on the hate crime data set. As mentioned before, one strong motivation for this section is to be able to predict future hate crimes after an election. While this election may have been more polarizing and hotly contested than most elections, the idea is that we might be able to use our existing data set for future elections. Since various covariates in our data set may change over the course of the Trump presidency, it may not be safe to assume the number of hate crimes per state will stay constant from election to election. Preparing a statistical model might be very beneficial for state and federal governments in preventing hate crimes in high risk areas. This paper will implement four models discussed during class, ordinary least squares, ridge regression, lasso regression, and principal component regression, and evaluate the performance of these models.

For the modeling of our data, I take a random sample of 42 states to use as our training data, and leave 5 states as our testing set to measure the performance of each model. Mean squared error will used as the main metric and is defined as

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

Ordinary Least Squares

Recall that ordinary least squares attempts to fit a model in the form,

$$y = X\beta + \epsilon$$

In the context of our training data set, y is a 42×1 vector, X is our 42×10 design matrix (including a column of 1's for our intercept), β is the 10×1 vector of regression coefficients and ϵ is our 42×1 vector containing our unexplained error. Defining our loss function as sum of squares and minizizing this loss function for the optimal β yields the well-known closed form solution

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Fitting our OLS to our data set yields

	Estimate	Pr(>|t|)
(Intercept)	-7.373	0.009563
median_household_income	-0.000004581	0.509
share_unemployed_seasonal	8.139	0.04313
share_population_in_metro_areas	-0.1661	0.5498
share_population_with_high_school_degree	6.27	0.003871
share_non_citizen	1.41	0.4266
share_white_poverty	2.144	0.3297

	Estimate	Pr(>|t|)
gini_index	5.707	0.01193
share_non_white	-0.313	0.3927
share_voters_voted_trump	-1.159	0.01638

We immediately see that only share_voters_voted_trump, gini_index, share_unemployed_seasonal, and share_population_with_high_school_degree are the only two variables found to be statistically significant by this particular model. One additional note is that one of the statistically significant variables that we did find was the Gini index which recall is a measure of income inequality. This is consistent with the original findings that the original author of the article found.

Ridge and Lasso Regression

As we observed above, it seemed that many variables in our data set do not appear to be statistically significant predictors of the number of hate crimes the 10 days after an election. Peforming some type of regularization might improve the predictive performance by penalizing the coefficients and preventing overfitting. Mathematically, instead of minimizing the typical sum of squares $(y - X\beta)^T(y - X\beta)$ we now add a L2 penalty to yield

$$L(\beta) = (y - X\beta)^T(y - X\beta) + \lambda||\beta||_2$$

where λ is the shrinkage parameter and controls the amount of regularization. Adding this L2 penalty will perform shrinkage of the coefficient estimates, but will not perform zero estimates like Lasso's L1 penalty. Ridge regression has a closed-form solution similar to OLS and can be written as

$$\hat{\beta}_{ridge} = (X^T X + \lambda I)^{-1} X^T Y$$

Note the additional λI term in the solution has the nice property of dealing with multi-collinearity should it exist in our data set, as it prevents the inverse term from producing an numerically unstable estimate. Using leave one out cross validation from the `glmnet` package to select the optimal λ yields the following model.

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                                                 s0
## (Intercept)                         -4.4569264170146
## median_household_income             -0.0000006006359
## share_unemployed_seasonal            5.5456648484176
## share_population_in_metro_areas     -0.1371127879069
## share_population_with_high_school_degree  3.6652774857583
## share_non_citizen                    0.7283942628244
## share_white_poverty                  1.1418439390059
## gini_index                           4.2173870594046
## share_non_white                     -0.2925543960150
## share_voters_voted_trump            -1.0866186125902
```

Lasso regression is similar to ridge regression in that it introduces a penalty term to the loss function, however it instead uses an $L1$ penalty which produces sparse coefficient estimates as noted above. The loss function for Lasso is

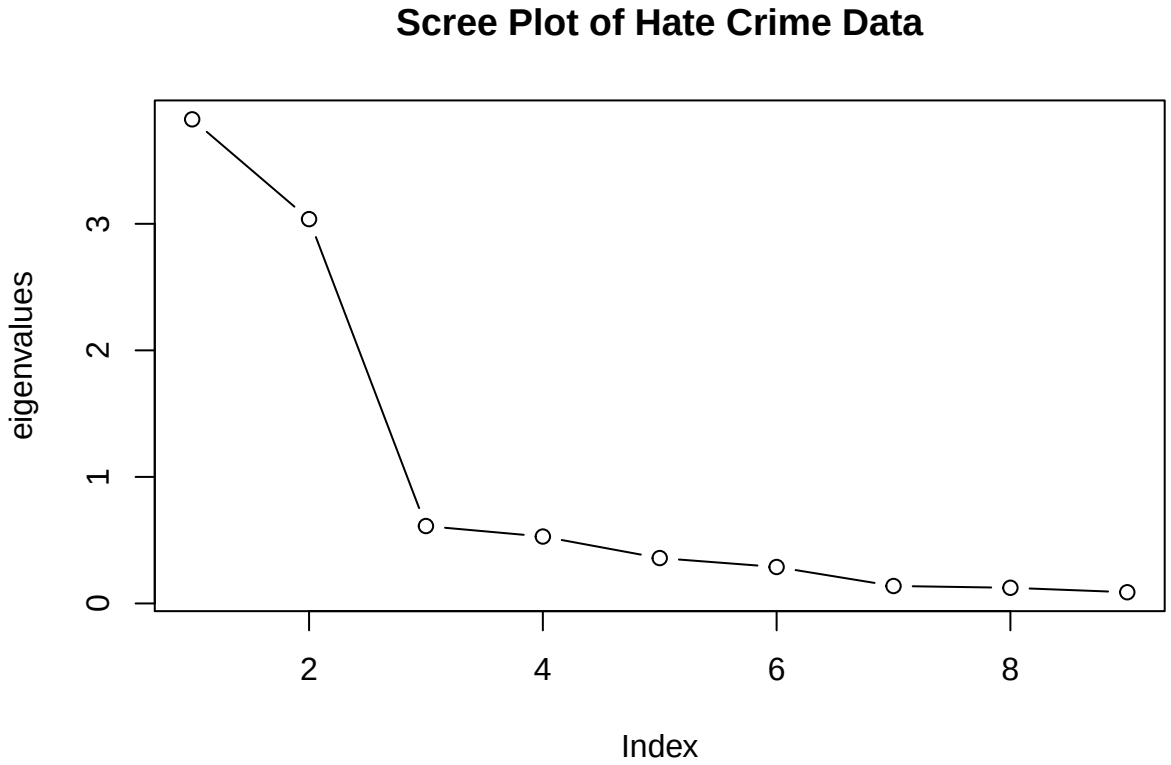
$$(y - X\beta)^T(y - X\beta) + \lambda||\beta||_1$$

which unlike ridge regression's closed form solution, is not as clean. We can fit a lasso model using the `glmnet` package as well, using alpha= 1 instead of 0 in ridge regression.

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                                                s0
## (Intercept)                            -0.8288322
## median_household_income                 .
## share_unemployed_seasonal               .
## share_population_in_metro_areas         .
## share_population_with_high_school_degree  1.2070180
## share_non_citizen                       .
## share_white_poverty                     .
## gini_index                              1.3500170
## share_non_white                         .
## share_voters_voted_trump               -1.0656761
```

Principal Component Regression

Principal compoonent regression is essentially the same thing as ordinary least squares, however instead of regression on the original design matrix, we now regress on the covariates projected onto the first k principal components. To choose the number of principle components, we can use a scree plot to determine a good place to cutoff. The scree plot below indicates that after three principal components, the additional variance explained by each principal component gives us marginal gains.



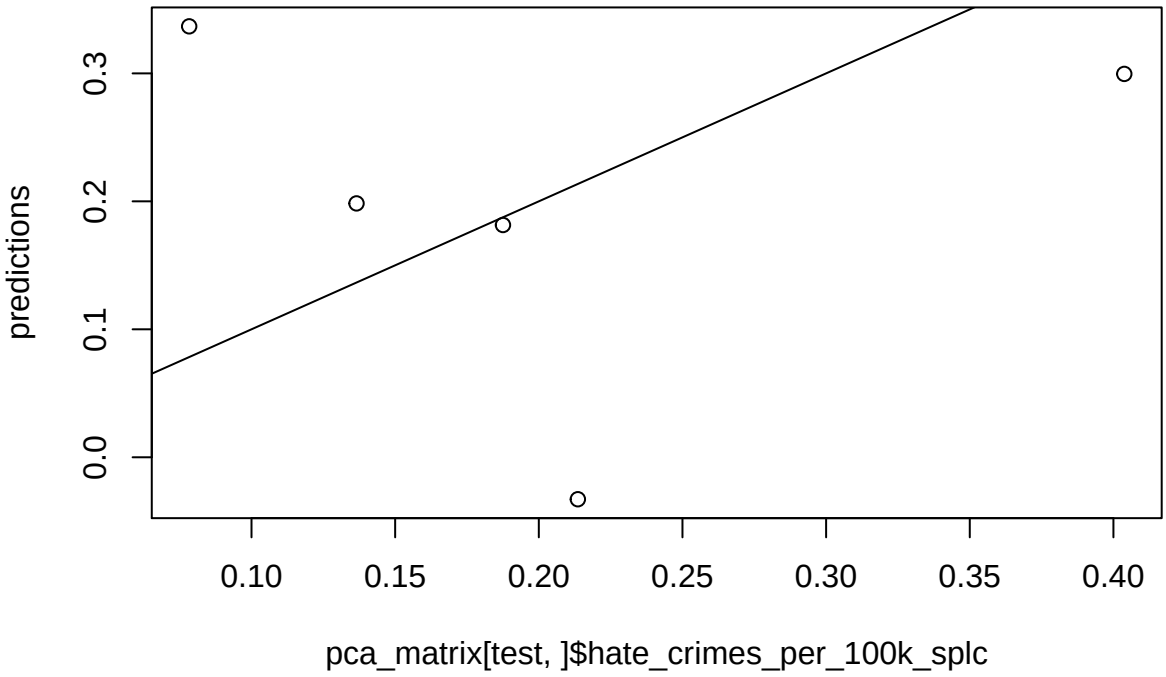
We then project our original scaled data matrix onto a three-dimensional subspace then perform OLS with these new transformed covariates.

	Estimate	Pr(>|t|)
(Intercept)	0.3033	0.0000000000002574
PC1	-0.04656	0.002049
PC2	0.04285	0.009263
PC3	-0.2005	0.000004357

Results and Discussion

The table below, shows the results from our predictive modeling. After fitting the model on our training set and comparing MSE's on the test set for the various models, the following results can be seen below.

method	lambda	MSE
OLS	NA	0.2611
Lasso	0.03024	0.1821
Ridge	0.03095	0.2046
PCR	NA	0.1687



Given these results, I do have some reservations regarding the predictve modeling of our data and coefficient estimates. Given the sample size of our data (47 observations) and nature of some of our predictors such as share_voters_voted_trump which contained some extreme outliers (e.g. District of Columbia's 4% vote share for Trump), our coefficients estimates particularly for our multiple regression were not very stable and were very dependent on the training rows that we sampled. For this particular seed, four predictors were found to be statistically significant, however on a subsequent run none of the predictors were found to be statistically significant! This produced high variablility in our coefficient estimates and gives me hesitation in using these models to predict future number of hate crimes. The most obvious solution to this problem would be to collect more data points on hate crimes, however, this particular dataset only contained hate crime data for the various states following the 2016 election. Performing box-cox transformations on our covariates would be another potential remedy for the issues that we faced in modeling the data. Since the main objective of this project was to demonstrate the core concepts covered in class and not producing the perfect model, not all potential solutions were explored.

Nonetheless our results found that PCR was the best method. If this proved to be the case with even more data available to us, one thing to keep in mind is that PCR is more difficult to interpret as each covariate is now a linear combination of the original variables. Thus if interpretability is an important factor in our predictive model, MSE should not be used as the sole determinant. Ridge regression and lasso regression provide limited benefit in our particular data set since we only have nine variables, although with lasso regression we were able to perform variable selection that is consistent with our findings from explaratory data analysis and the findings from the the article. If future data sets regarding hate crime, included more covariates, regulararized linear regression seems to be a very promising

technique particularly with interpretability.

Concluding Remarks

Throughout this project, we explored various methods of describing the variability in our data and attempted to create a predictive model for our hate crime data. By using principal component analysis, we were able to explain a large variation in our hate crime data, by using principal components to maximize the variance explained in our data. This allowed us to visualize the original data which had nine dimensions, and create some rough decision rules in determining the level of hate crimes across the states. Independent component analysis was used a additional method of analysis to PCA since PCA assumes that orthogonal components are the best way to explain the variance in the data. We found that ICA indeed complimented the results found in PCA. Factor analysis was used to see if we could find latent variables consisting of logical groupings of the original variables in the data. Using a three factor model to estimate the original correlation matrix of the hate crime data, 'liberalness' and 'living affordability' seemed to be reasonable latent factors underlying the level of hate crimes across states, although as always, factor analysis leaves some room for interpretation.

Predictive modeling proved to be of limited utility, due to our limited amount of data. However, we found that there were still some insights to be gleaned from our results. Given more data and additional work, it seems promising that we would be able to develop an effective model to predict the future number of hate crimes in the United States post election.

And lastly, a note should be made regarding the data that was used for this project. One large assumption that we made was that our measure of hate crime was accurate for each state. However, given that these hate crimes are self reported, it is important to keep in mind the types of people who report or do not report hate crimes most likely are not constant across over the states [1]. This brings up the possibility of over and underrepresentation in our data and we should remain cognizant of this potential problem in interpreting the results of our analysis.

References

[1] Majumder, Maimuna. "Higher Rates Of Hate Crimes Are Tied To Income Inequality." FiveThirtyEight, FiveThirtyEight, 21 Apr. 2017, fivethirtyeight.com/features/higher-rates-of-hate-crimes-are-tied-to-income-inequality/.

[2] Berman, Mark. "Hate Crimes in the United States Increased Last Year, the FBI Says." The Washington Post, WP Company, 13 Nov. 2017, www.washingtonpost.com/news/post-nation/wp/2017/11/13/hate-crimes-in-the-united-states-increased-last-year-the-fbi-says/?utm_term=.995c1779f8fa.

[3] "Fivethirtyeight/Data." GitHub, github.com/fivethirtyeight/data/tree/master/hate-crimes.

[4] "Population Distribution by Citizenship Status." The Henry J. Kaiser Family Foundation, 22 Sept. 2017, www.kff.org/other/state-indicator/distribution-by-citizenship-status/?currentTimeframe=0&sortModel=%7B%22colId%22%3A%22Location%22%2C%22sort%22%3A%22asc%22%7D.