STAT 515 VISUALIZATION FOR ANALYTICS

FINAL PROJECT REPORT
ON
HATE CRIMES IN USA

By
Tanmai Reddy Kothi
G01120670

DEPARTMENT OF DATA ANALYTICS AND ENGINEERING
GEORGE MASON UNIVERSITY
FAIRFAX, VA

Table of Contents

1. INTRODUCTION

The core idea of the project is to explore the Hate Crimes in USA and find the most affecting factors. The data exploration is done using a variety of plots and graphics, all using R Studio. The dataset has information about various factors that affect the average hate crime in every state of USA. A variety of plots and graphs are implemented to draw insights about patterns and trends in data concentrating on plotting the correlation between variables. A variety of visualizations included in this project walk us through the relationship between each attribute and their influence on the average annual hate crime.

2. DATASET OVERVIEW

2.1 Dataset Description

The dataset is taken from Github. The original dataset has 50 observations and 12 attributes. These variables belong to character, numeric datatype. The attributes of the dataset are:

- State – State name
- Income – Median household income
- Unemployed - % share of the population that is unemployed
- Metro Population - % share of the population that lives in metro areas
- HS Population - % share of adults 25 and older with a high school degree
- Non-Citizen - % share of the population that are not U.S. citizens
- White Poverty - % share of white residents who are living in poverty
- Gini Index – Gini Index
- Non-White - % share of the population that is not white
- Trump Voters - % share of voters who voted Trump
- Avg Hate Crimes – Average annual hate crimes per 100,000 population during 2010-2015

2.2 Data Preprocessing

Data preprocessing is a data mining technique where raw data is transformed into understandable format by deleting unnecessary data, imputing missing data, etc.

The dataset had few missing values. Preprocessing had been done before it was used for plotting in R. Missing values are found in the columns ‘Avg Hate Crimes’, ‘Non-Citizen’. All the missing values being numeric predictors, instead of deleting I have imputed the value 0 in place of NAs. Skewness of the data is ignored as it doesn’t affect the visualization.

A new column median which contains the median value of the average annual hate crimes is added to the dataset to facilitate plotting of a linked micromap.

2.3 Final Dataset

The dataset now has no NAs or missing values. The final dataset has 50 observations with 13 attributes. This is used for further visualization and find insights which help in identifying the factors affecting the average hate crimes.

State	Income	Unemployed	Metro Population	HS Population	Non Citizen	White Poverty	Gini Index	Non White	Trump Voters	Hate Crimes	Avg Hate Crimes	med
alabama	42278	0.06	0.64	0.821	0.02	0.12	0.472	0.35	0.63	0.12583893	1.806410489	1.923019
alaska	67629	0.064	0.63	0.914	0.04	0.06	0.422	0.42	0.53	0.14374012	1.656700109	1.923019
arizona	49254	0.063	0.9	0.842	0.1	0.09	0.455	0.49	0.5	0.22531995	3.413927994	1.923019
arkansas	44922	0.052	0.69	0.824	0.04	0.12	0.458	0.26	0.6	0.06906077	0.869208872	1.923019
california	60487	0.059	0.97	0.806	0.13	0.09	0.471	0.61	0.33	0.25580536	2.397985899	1.923019
colorado	60940	0.04	0.8	0.893	0.06	0.07	0.457	0.31	0.44	0.3905233	2.804688765	1.923019
connecticı	70161	0.052	0.94	0.886	0.06	0.06	0.486	0.3	0.41	0.33539227	3.772701469	1.923019
delaware	57522	0.049	0.9	0.874	0.05	0.08	0.44	0.37	0.42	0.32275417	1.469979563	1.923019
florida	46140	0.052	0.96	0.853	0.09	0.11	0.474	0.46	0.49	0.18752122	0.698070342	1.923019
georgia	49555	0.058	0.82	0.839	0.08	0.09	0.468	0.48	0.51	0.12042027	0.412011824	1.923019
hawaii	71223	0.034	0.76	0.904	0.08	0.07	0.433	0.81	0.3	0	0	1.923019
idaho	53438	0.042	0.7	0.884	0.04	0.11	0.433	0.16	0.59	0.12420817	1.891330529	1.923019
illinois	54916	0.054	0.9	0.864	0.07	0.07	0.465	0.37	0.39	0.19534455	1.044015798	1.923019
indiana	48060	0.044	0.79	0.866	0.03	0.12	0.44	0.2	0.57	0.24700888	1.757356567	1.923019
iowa	57810	0.036	0.6	0.914	0.03	0.09	0.427	0.15	0.52	0.45442742	0.561395565	1.923019
kansas	53444	0.044	0.64	0.897	0.04	0.11	0.445	0.25	0.57	0.10515247	2.143986672	1.923019

Figure 1: Final dataset

3. DATA EXPLORATION

Data exploration is about describing the data by means of statistical and visualization techniques. Data exploration is done to identify important aspects of data and draw insights from it for further analysis.

3.1 Dot Plots

The dot plots are created using the ggplot package of the R language. Dot plots for the median household income and gini index against each state are given below.

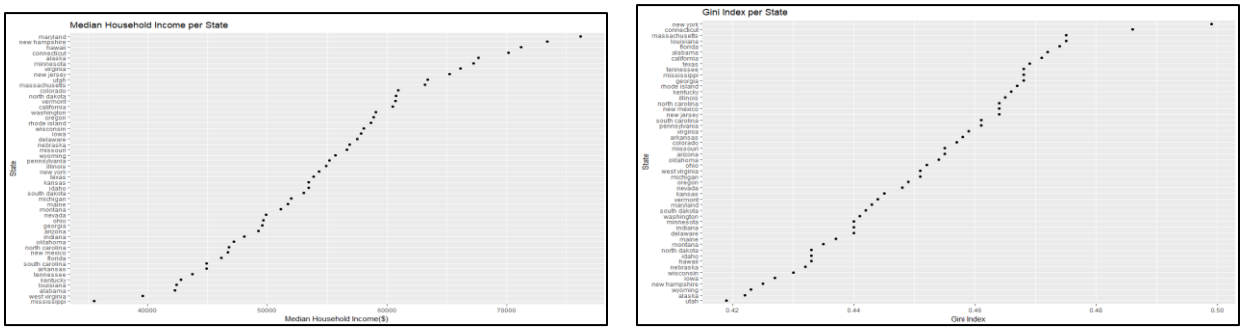


Figure 2: Dot plots using ggplot2

The dot plots above represent the median household income and gini index for each state in different graphs with the states being plotted in an increasing fashion. From the graph it can be analyzed that Mississippi and Maryland has the least and most household income whereas New York and Utah have highest and lowest gini index.

3.2 Bar Plot

A bar plot is used to show relation between categorical and numerical data in the form of bars.

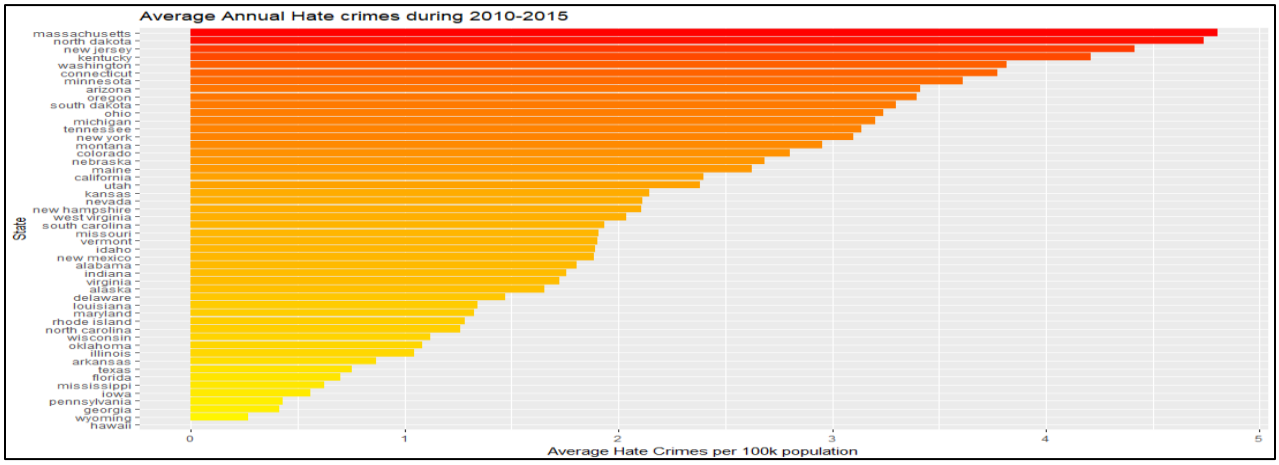


Figure 3: Bar plots using ggplot

The above graph represents the average annual hate crimes per 100,000 population for each state in an increasing fashion using the 'reorder' function of the ggplot. From the graph, Massachusetts stands top in the average annual hate crimes. Below is the code for the above bar plot.

```
ggplot(data=hcrime, aes(x = reorder(hcrime$State,hcrime$`Avg Hate Crimes`),
                        y = hcrime$`Avg Hate Crimes`,fill=hcrime$`Avg Hate Crimes`)) +
  geom_bar(stat="identity",show.legend = F)+
  scale_fill_gradient(low="yellow",high="red")+
  ggtitle("Average Annual Hate crimes during 2010-2015")+
  labs(x="State",y="Average Hate Crimes per 100k population")+
  coord_flip()
```

Figure 4: Code for bar plot

3.3 Histograms

A histogram is used to show the frequency distribution of continuous data.

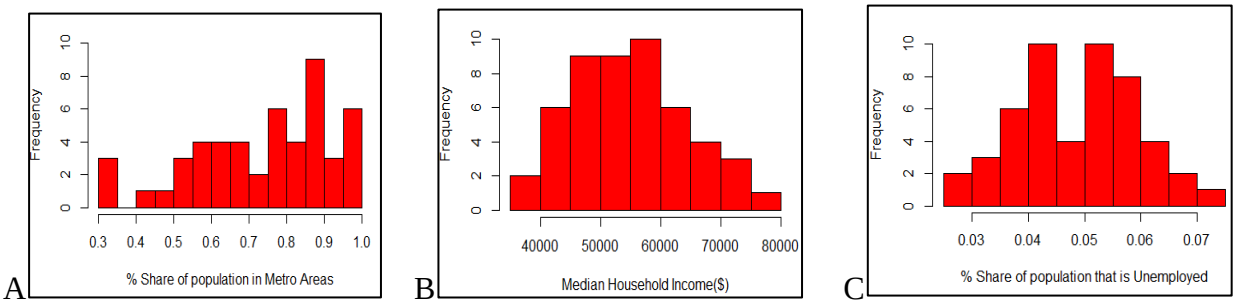


Figure 5: Histograms for various important variables

The above figure has three different histograms for the attributes that affect the average hate crimes in USA. These graphs are plotted using the hist() function in R.

Figure 5A shows the distribution of percentage share of population living in metro areas in different states.

Figure 5B shows the distribution of the median household income in USD in different states. From the graph it can be inferred that the median household income for most of the states lies in the range $45,000 to $50,000 with highest and lowest being $80,000 and $35,000.

Figure 5C shows the distribution of percentage share of the population in each state that is unemployed.

3.4 Density Plots

The density plots show the density of distribution of attributes. A density plot can be considered as a smoothed histogram. The peaks of the density plot help identify where the values are concentrated.

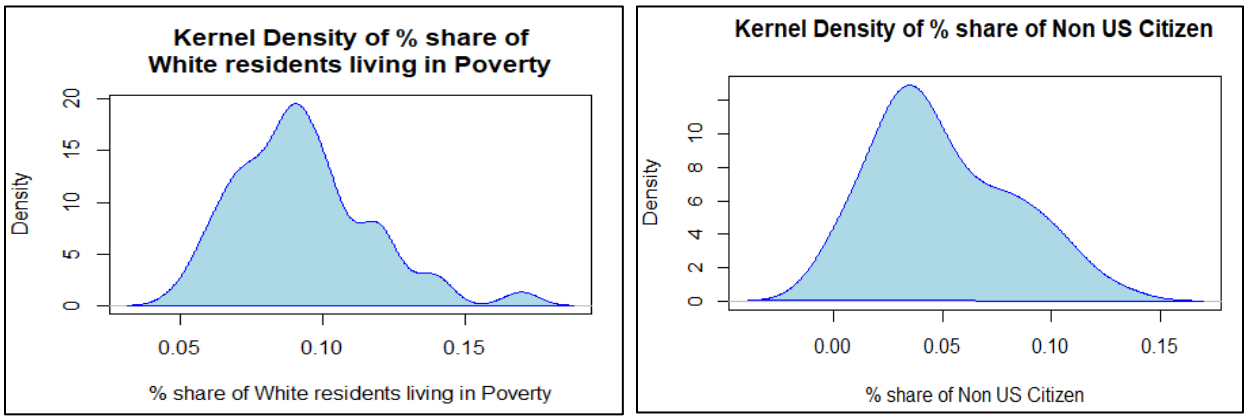
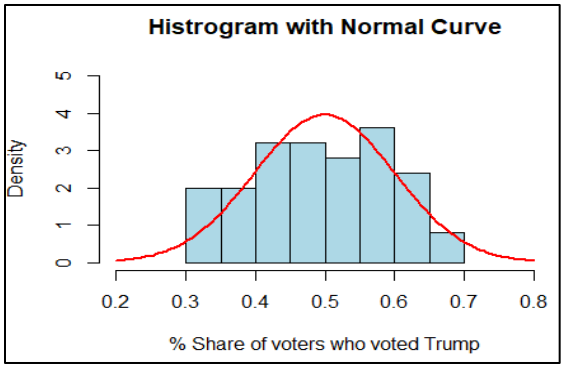


Figure 6: Density plots

The above plots visualize the distribution of percentage share of white residents living in poverty and non-US citizens in each state. From the graphs it can be inferred that 10% population of most states are Whites and 5% are non-US citizens.

The plot below compares the histogram and density curve for the percentage share of voters who voted Trump.




```
hist(hcrime$`Trump Voters`,prob=T,col="lightblue",
     main="Histrogram with Normal Curve",
     xlab="% Share of voters who voted Trump",
     xlim = c(0.2,0.8),ylim=c(0,5))
curve(dnorm(x, mean(hcrime$`Trump Voters`),
    sd(hcrime$`Trump Voters`)),col="red",lwd=2,add=TRUE)
```

Figure 7: Graph and Code for Histogram and Density plot combined

3.5 Regression Plots

Regression plots are used to identify the relation between the target and predictor variables. In this project, 'Avg hate crimes' is the target variable and others are the predictor variables.

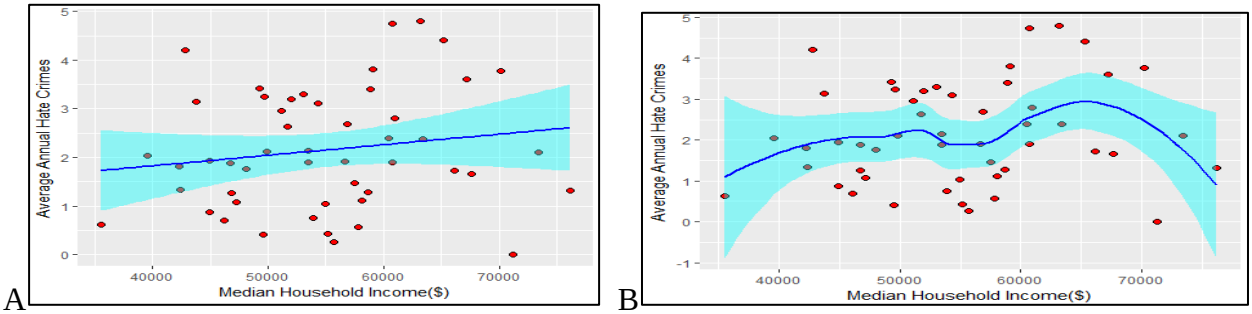


Figure 8: Regression plots for average hate crimes and median household income

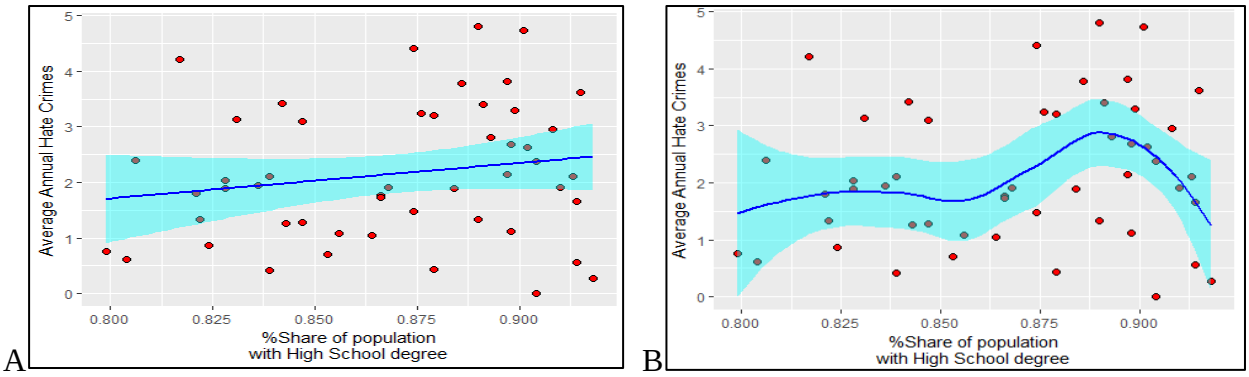


Figure 10: Regression plots for average hate crimes and %share of population with hs degree

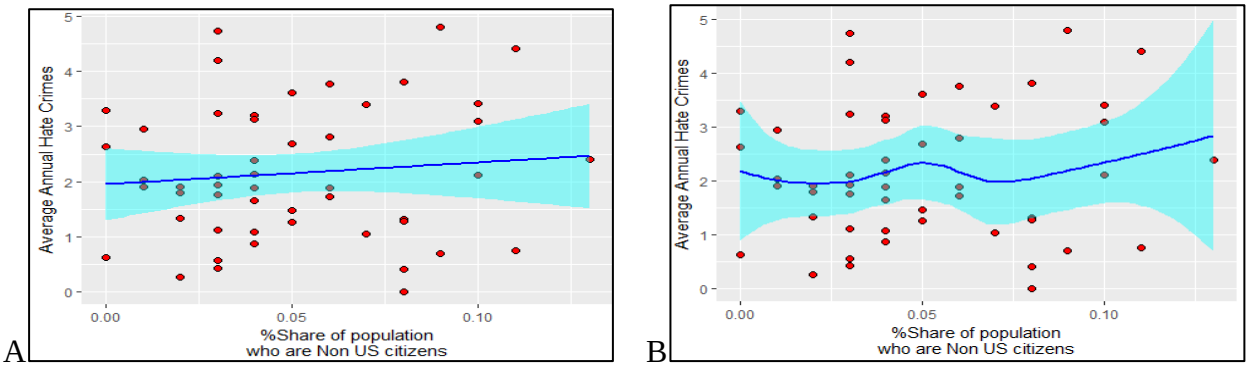


Figure 11: Regression plots for average hate crimes and %share of Non-US citizens

The above figures are plotted for same target variable 'Average Hate Crimes' against the predictor variables income, highschool, degree population and non-US citizen population using different regression methods namely 'lm' and 'loess'. A simple linear model fits the best possible straight line through a set of points whereas the loess model fits a complicated curve taking care of every point along the line.

3.6 Scatterplot Matrix

A scatterplot matrix is a collection of scatter plots arranged in the form of grid where each scatter plot shows the relation between a pair of variables. The diagonal contains the variables and each variable is plotted against each other.

The plot below is created using the ‘pairs’ feature of the ggplot. The lower triangle contains the scatter plot with a smooth added and the upper triangle contains the correlation coefficient foeeach pair of variables and the p-value. From the figure, the population with high school degree affects the average hate crimes the most.

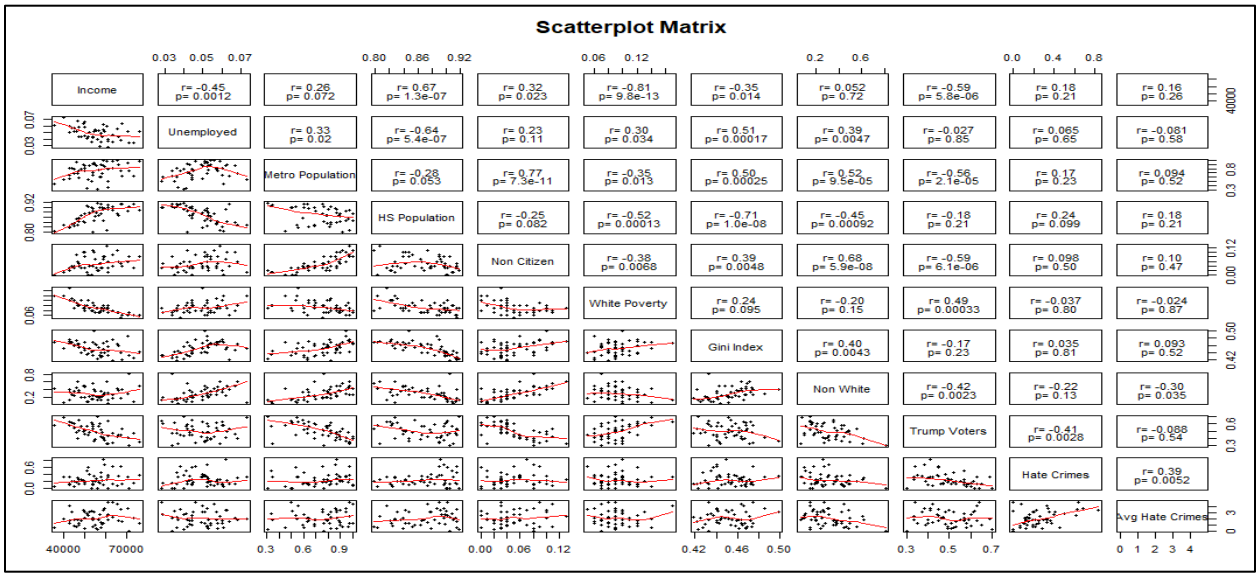


Figure 12: Scatter plot Matrix

3.7 Correlation Plot

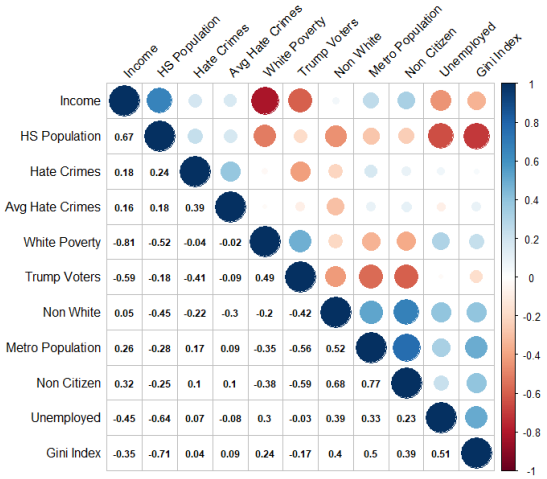


Figure 13: Correlation plot

A correlation plot represents the correlation between any two variables. The lower triangle contains the correlation coefficient of each pair of variables. The upper triangle pictorially represents the correlation coefficient in terms of size and color. From the plot it can be inferred that percentage of population with high school degree is highly correlated with the average hate crimes compared to other predictors.

3.8 Linear Regression Model

A linear regression model is used to study the relationship between two continuous variables. The diagnostic model consists of four plots as shown below.

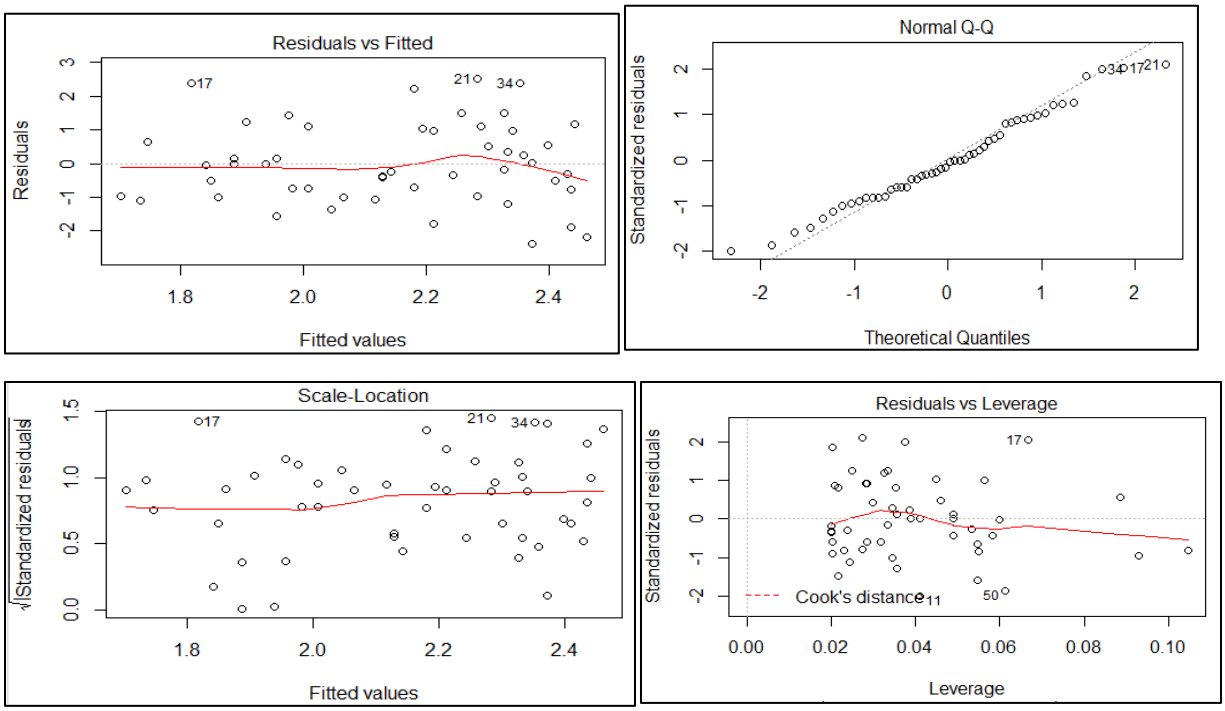


Figure 14: Diagnostic plots

The above figure represents the diagnostic plots of the linear regression model with the target variable being average hate crimes and predictor variable being percentage of population with high school degree. The Normal Q-Q plot has thin tail at both ends.

```
lm(formula = hcrime$`Avg Hate Crimes` ~ hcrime$`HS Population`,
    data = hcrime)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3713 -0.9061 -0.1090  0.9820  2.5199

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             -3.396      4.394  -0.773    0.443
hcrime$`HS Population`   6.380      5.052   1.263    0.213

Residual standard error: 1.217 on 48 degrees of freedom
Multiple R-squared:  0.03215,   Adjusted R-squared:  0.01199
F-statistic: 1.594 on 1 and 48 DF,  p-value: 0.2128
```

Figure 15: Summary of linear model

3.9 Random Forest Regression Model

A random forest model is used to measure the significance of predictor variables and measure of importance of internal structure of data.

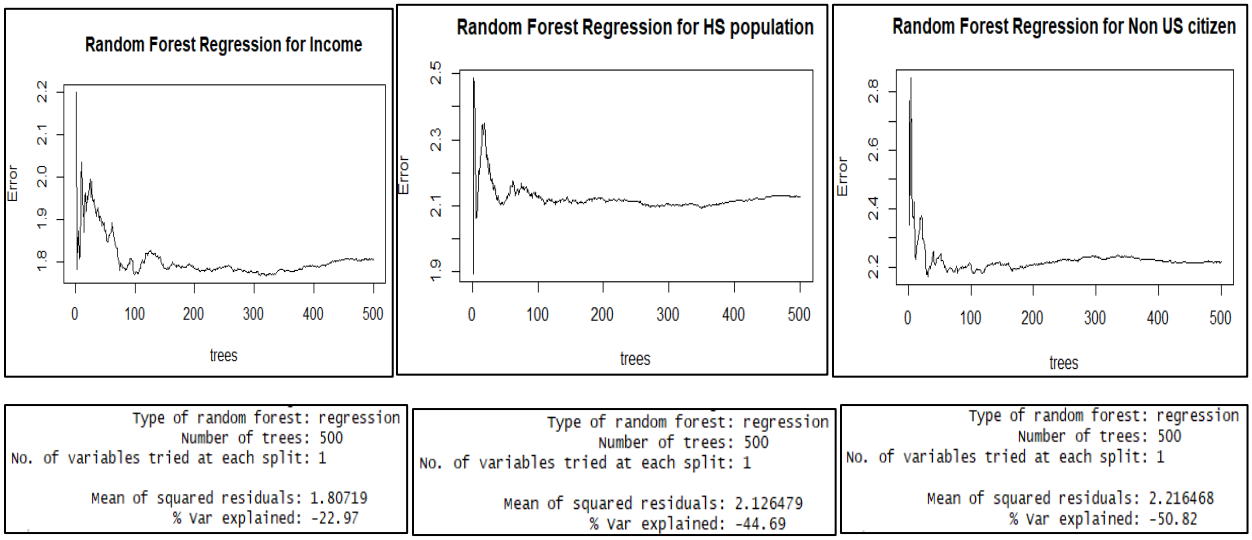


Figure 16 : Random Forest Model for different variables and its summary

3.10 Dendrogram

A dendrogram is a plot that represents hierarchical clustering in the form of tree. The y-axis represents the dissimilarity or distance between clusters and the x-axis contains the states that are clustered using the 'ward.D2' method of hclust().

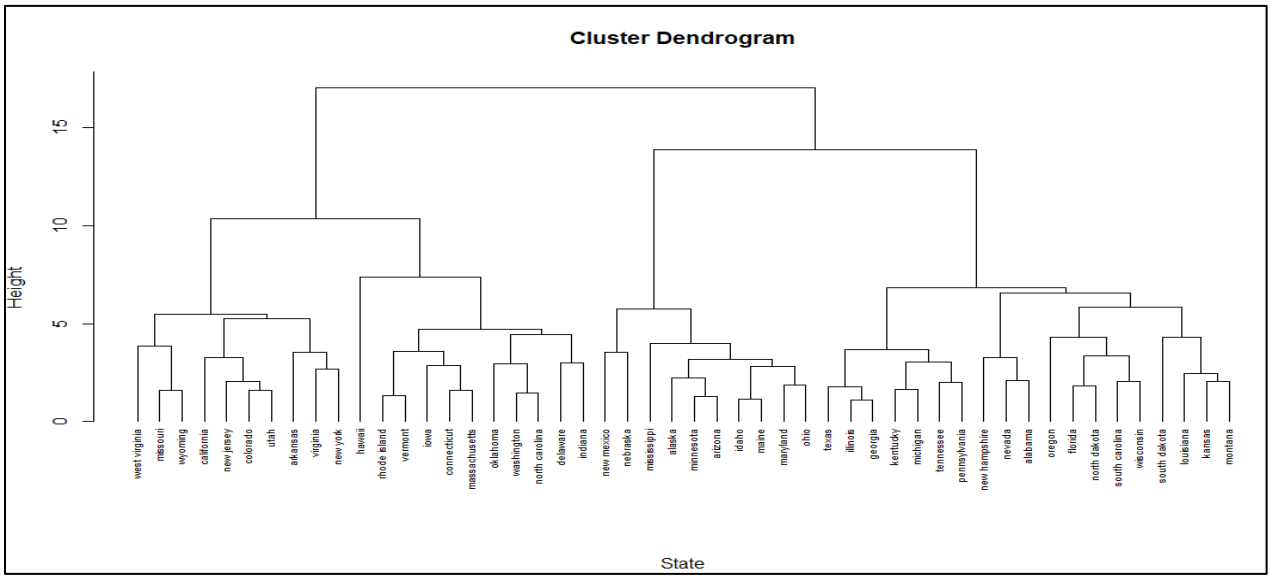


Figure 17: Dendrogram

3.11 Choropleth Maps

A choropleth map is a thematic map in which areas are shaded in proportion to the measurement of a continuous variable.

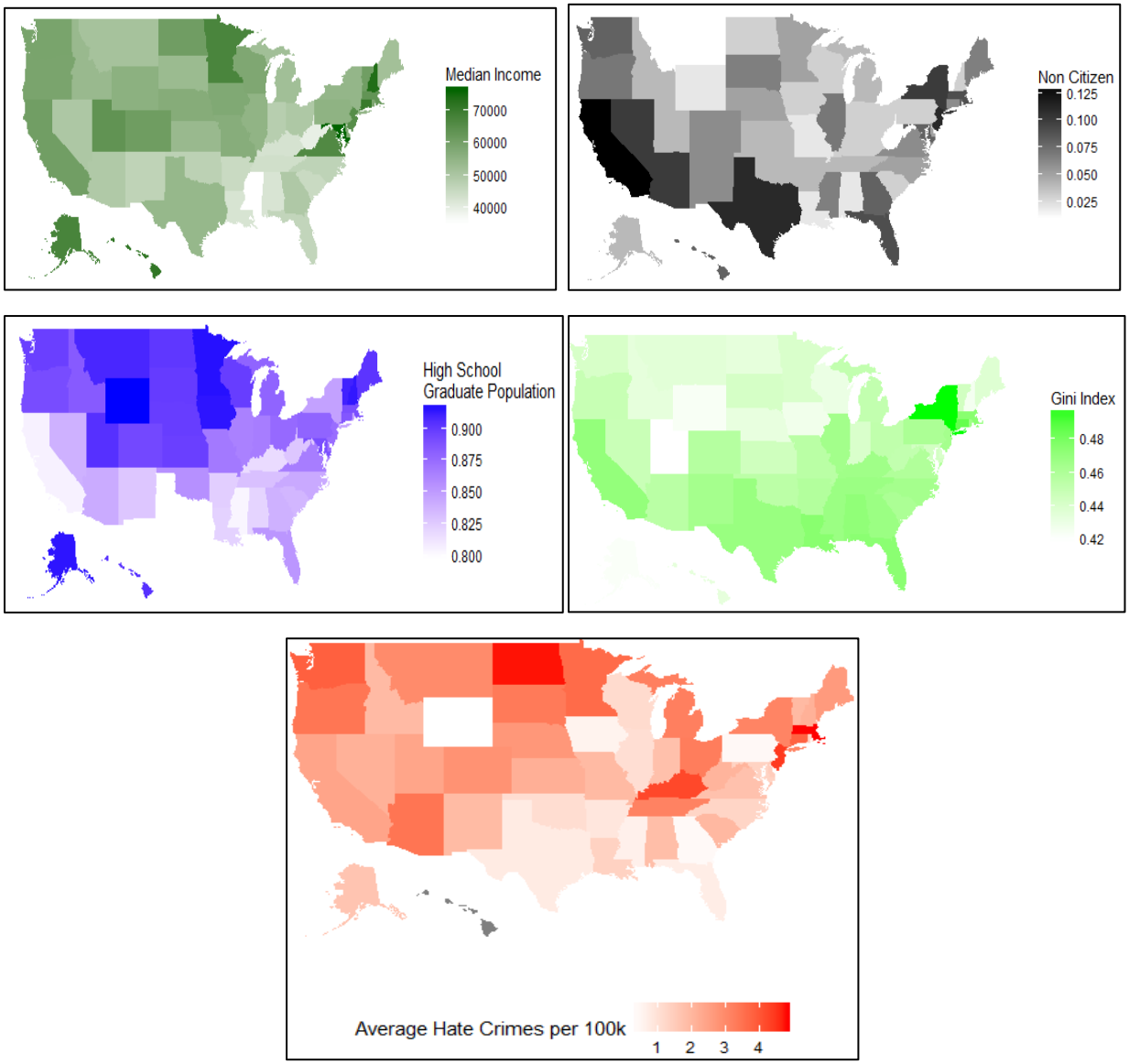


Figure 18: Choropleth maps for dome important attributes

The above figure represents the frequency of attributes income, non-US citizen, high school graduate population, gini index and average hate crimes. The increase in frequency is represented by the increased intensity in the color. The white color in the maps indicate the missing values.

It is clear from the last map that the hate crimes were high in the northern part of the country than the rest of the part during 2010-2015.

3.12 Linked Micromap

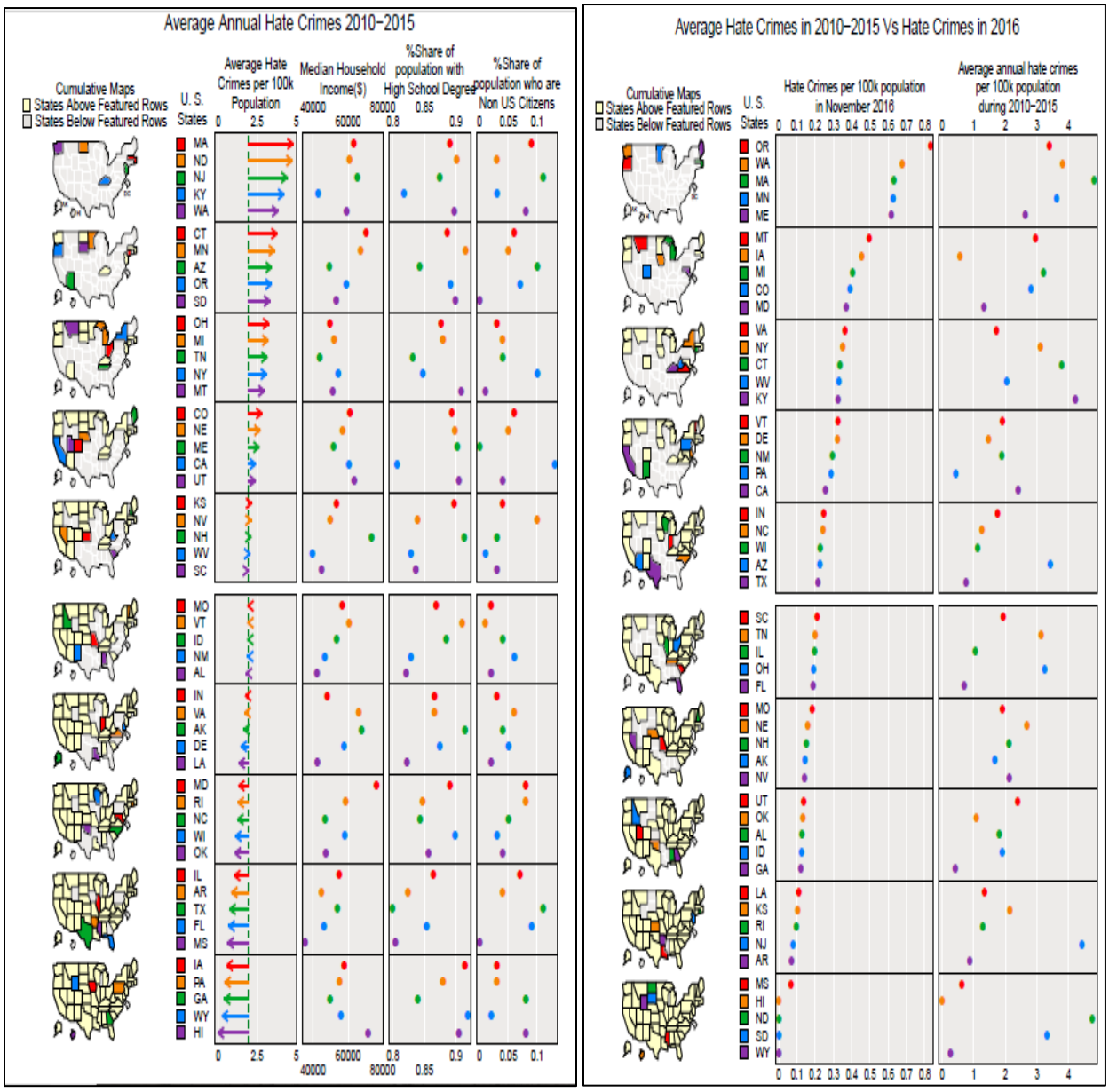


Figure 19: Linked Micromaps

The first linked micromap shows the average annual hate crimes per 100,000 population in each state in a decreasing fashion. The reference value considered for the arrow map is the median value of the average annual hate crimes. The arrows are pointed towards right in the top half and towards left in the bottom half of the plot representing the decreasing order of average hate crimes.

The three attributes considered for plotting other than average hate crimes are based on their correlation with the target variable i.e., average hate crimes. Although they have a better correlation with the average hate crimes in comparison to other variables it is not clearly seen in the map because of its low correlation coefficient value.

The second linked micromap compares the average annual hate crimes during 2010-2015 and the hate crimes in November,2016 in decreasing fashion with respect to the latter. Taking into consideration the correlation coefficient between the two variables, they are highly correlated in comparison to the rest of the attributes, but it cannot be clearly found in the above map.

Massachusetts which ranked top in hate crimes during 2010-2015 dropped to rank 3 in November 2016. Oregon has moved 8 places and reached top of the list in November 2016.

4. Conclusion

The order of predictors based on the correlation coefficients is %share of population with high school degree, median household income, %share of population that are non-US citizens, gini index, %share of residents living in metro areas, %share of white residents living in poverty, %share of population that is unemployed, %share of voters that voted Trump, %share of population that are not white.

From the linked micromap it is clear that the northern region of the US had the highest number of hate crimes than rest of the country during 2010-2015.

The highly correlated predictor is the %share of population with High school degree. It is clear that uneducated people tend to involve themselves in the hate crimes. Also, hate crimes are found to happen due to less household income.

So, increase in the number of high school graduates eventually increasing the household income might decrease the number of hate crimes in the US.

References:

- Dataset from https://github.com/fivethirtyeight/data/tree/master/hate-crimes