Course Name: Regression and Time Series

Course Number and Section: 16:954:596:01

Project: State-wise analysis of criminal rate (ratio of perpetrators to the population for the year reported) and its correlation with factors like Demographics and Age, Sex, and Race of the perpetrator.

Instructor: Yaqing Chen

Group Members:

- Sarthak Singh – SS4767
- Jack Kogut – JK2114
- Bhavna Ajit Sharma – BS1167
- Anika Madhusudhan – AM2622

Date Submitted: 12.15.2023

Submitted by: Sarthak Singh, Jack Kogut, Bhavna Ajit Sharma, and Anika Madhusudhan

Index

Introduction

To gain deeper insights into the complex landscape of crime rates in the United States, this project undertakes a meticulous examination of the Uniform Crime Reporting Program Data: Arrests by Age, Sex, and Race, Summarized Yearly, United States, 2019 (ICPSR 38780). Leveraging this extensive dataset, we delve into the multifaceted dimensions of criminal activities across states, seeking correlations with factors such as demographics, age, sex, race of perpetrators, and external variables like area, median income, and political affiliation.

Through rigorous statistical analyses, this project aims to unravel patterns and correlations within crime data, providing valuable insights for policymakers, law enforcement agencies, and researchers. By exploring the interplay of demographic factors, external variables, and geographical divisions, the findings contribute to a nuanced understanding of crime dynamics in the United States, fostering informed discussions on effective crime prevention strategies and resource allocation.

Literature Survey

Our project and [5] share common ground in dissecting the intricate relationships between demographics and crime rates in the United States. While our exploration delves deeply into suburban and non-suburban areas, investigating crime rates based on factors like political affiliation and race, the referenced study focuses on three core objectives: understanding why males commit more crimes than females, examining the correlation between social class and crime rates, and exploring the impact of race and ethnicity on criminal behaviors. Despite these nuanced differences, both studies converge on certain aspects, such as scrutinizing racial demographics in crime rates. Our project's hypothesis testing, which considers the prevalence of white criminals as an indicator of severe crime rates, aligns with the referenced study's confirmation that white individuals constitute a substantial portion of violent crimes.

Although we could not find any direct publications on the data we are utilizing, there have been several interesting research done on the same type of data reported for a different year. In the Uniform Crime Reported Data for the year 2018 [6], Puzzanchera, Charles [7] talked about how arrests of Juveniles in 2018 reached the lowest level in nearly 4 decades. Wang, Yiwen’s thesis [8], researched the effect of Marijuana Legalization on Juvenile delinquency. Sheehan et al.[9] talked about the racial disparity of cannabis possession arrests among adults and youths with state-wise cannabis decriminalization and legalization.

Additionally, our study contrasts with another investigation exploring the racial invariance hypothesis, focusing on the Asian community in the United States [10]. The racial invariance hypothesis posits that, when placed in similar circumstances, all race/ethnicity groups will exhibit similar crime rates. While our project includes an analysis of the Indian population and considers forty-nine different crimes, the referenced study concentrates solely on White, Black, and Asian populations, particularly examining violent crimes. Despite acknowledging that previous research on the racial invariance hypothesis predominantly centers on White and Black populations, we contend that a more inclusive approach, encompassing various races/ethnicities, could strengthen evidence for or against the racial invariance hypothesis. Our commitment to a comprehensive examination of crime rates seeks to contribute to a more holistic understanding of the dynamics at play across diverse demographic groups.

Data Wrangling

Dataset

We have used the Uniform Crime Reporting Program Data: Arrests by Age, Sex, and Race, Summarized Yearly, United States, 2019 (ICPSR 38780) available as part of the National Crime Victimization Survey (NCVS) data are available at https://bjs.ojp.gov/data-collection/ncvs.
These data provide information on the number of arrests reported to the Federal Bureau of Investigation's Uniform Crime Reporting (UCR) Program each year by police agencies in the United States. These arrest reports provide data on 49 offenses including violent crime, drug use, gambling, and larceny. The data received by ICPSR were structured as a hierarchical file containing, per reporting police agency: an agency header record, and 1 to 49 detail offense records containing the counts of arrests by age, sex, and race for a particular offense. [1]

Data Scrubbing and Selection

We removed several rows from the raw dataset which had 'N/A' entries. We also removed several columns which were statistically insignificant like 'ASR_ID', 'YEAR', 'MSA', and 'SEQ_NO' which had the same value for all the rows of the dataset or had no explanatory characteristics like 'CORE' and 'CONTENTS'. We also dropped certain variables like 'GROUP' (containing population ranges) whose properties were mirrored by other more exhaustive variables like 'POP'(which had current population numbers under an agency for the period during which the crimes were reported). After that, some more rows were dropped where the population under the jurisdiction of an agency was 0.

Feature Engineering

In the original dataset, Age, and Sex of a perpetrator were given in the format of "X_Y" where X= 'M' or X= 'F' depending on the sex of the person and Y= "either an age value like 18 or an age range like 60_64". We used these variables provided in the raw dataset and came up with some new variables like-

1. TAM - Total Adult Males which had the total number of Male Perpetrators over the age of 17.
2. TAF - Total Adult Females which had the total number of Female Perpetrators over 17.
3. TJF - Total Juvenile Females which had the total number of Female Juveniles Perpetrators i.e. Females under 18.

4. TJM - Total Juvenile Males which had the total number of Male Juveniles Perpetrators i.e. Males under 18.
5. Total_perps_per_agency_per_offense - Containing the total number of arrests made by an agency for a particular offense type.

Data Summarization

We have also summarized the cleansed dataset into two sub-datasets namely, Agency_Criminal_Counts and State_Criminal_Counts for further analysis at different levels.
In Agency_Criminal_Counts we have grouped the dataset into a single row if they had the same Agency and State name. In State_Criminal_Counts we have further grouped the data given they had the same State name. This was done so that Agency and state-level analysis could be done more efficiently.

Integrating External Features

We added the following variables to try to get more predictor variables that can better explain the variance of the response. The predictors we are adding are -

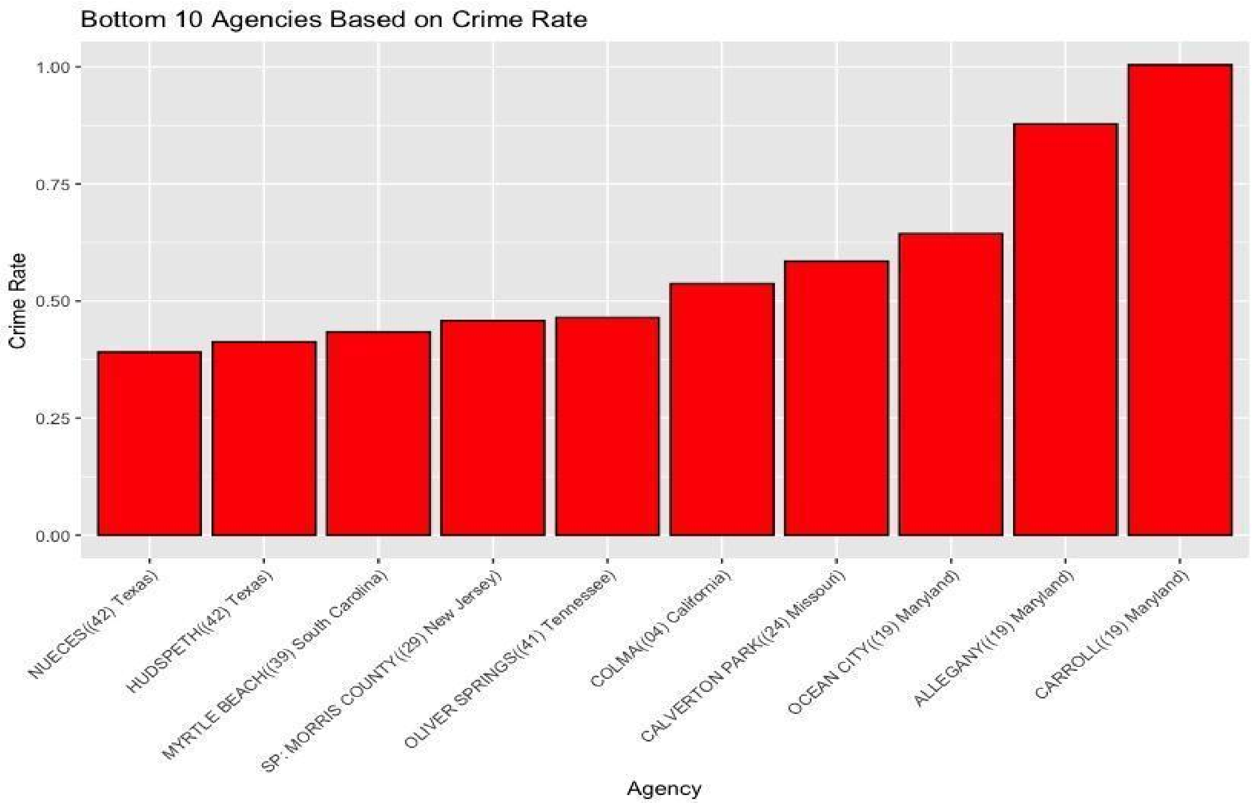
- Area- The area in square miles of each state is added to the data as a potential predictor.[2]
- Median Income- We do not have the exact income brackets for each state for when the crimes were reported, we have added the median household income for each state as a predictor.[3]
- Political Affiliation- This is added to the data to check for an everpresent analytical question that political leaning could affect criminal rates.[4]

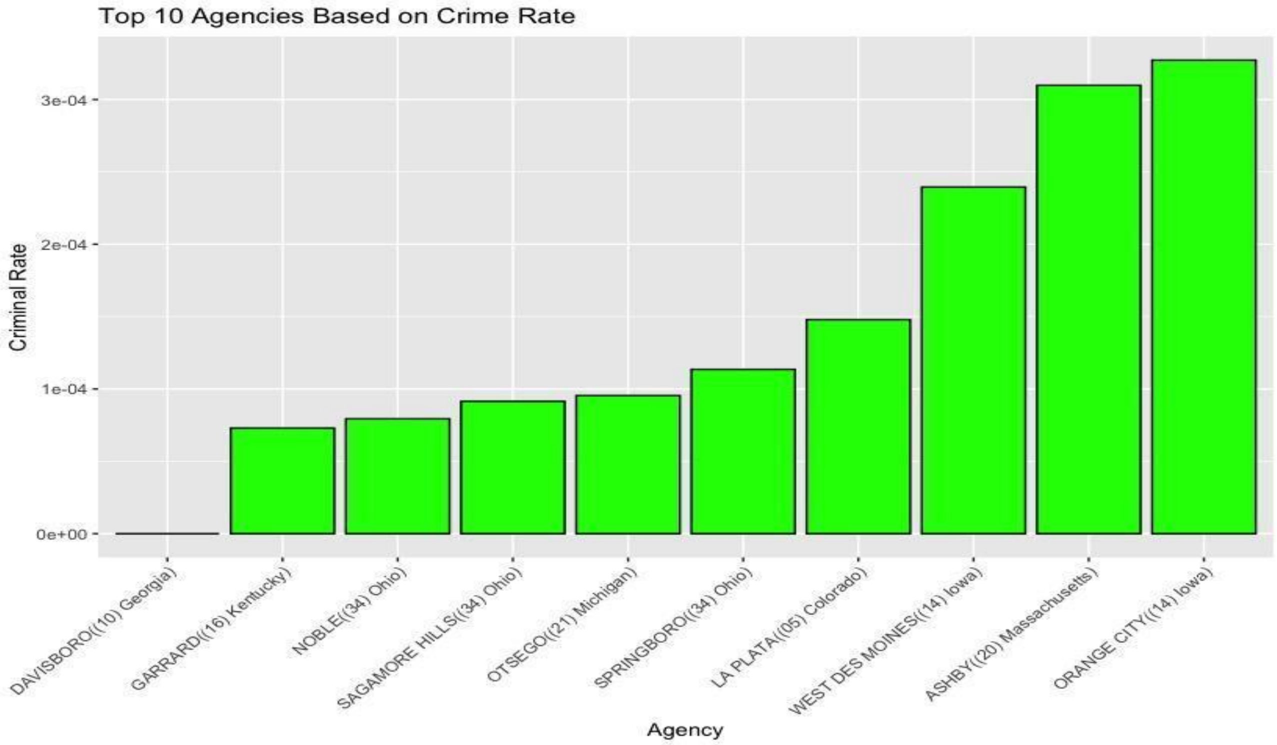
Exploratory Data Analysis

Agency Wise Analysis

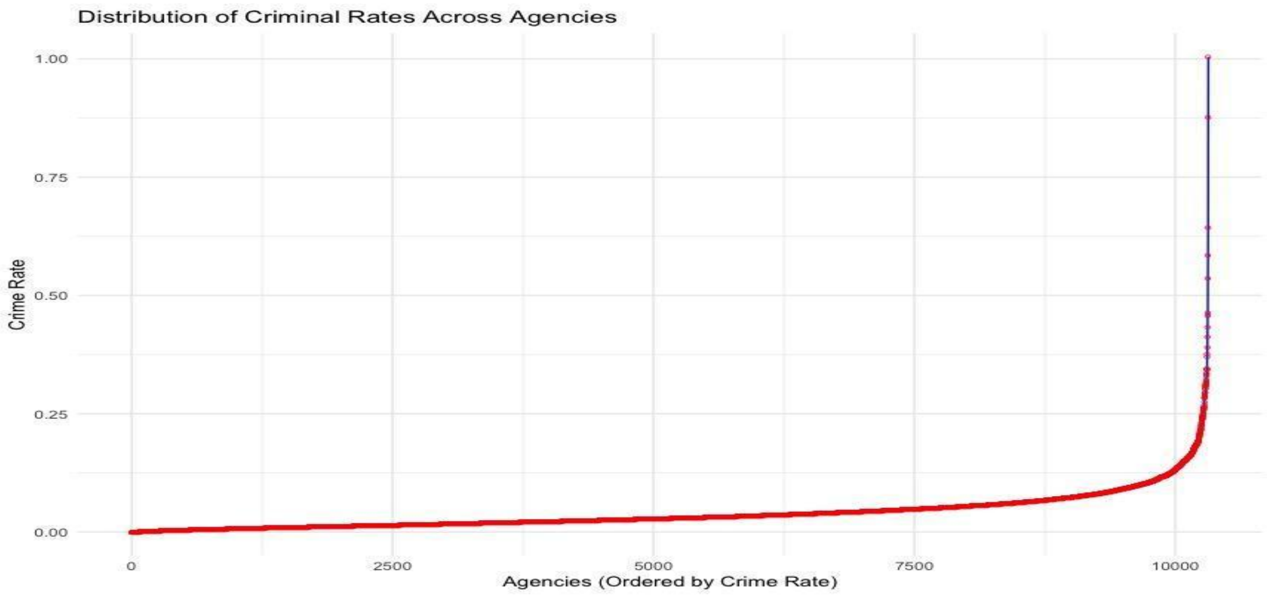
As the number of Agencies is huge, we first focused on the 10 most effective and 10 least effective agencies.

We assumed an agency to be more effective if the criminal rate was the lowest and the reverse criteria for it to be among the least effective agencies. Here are some of the least effective agencies (with their corresponding States) followed by some of the most effective agencies-





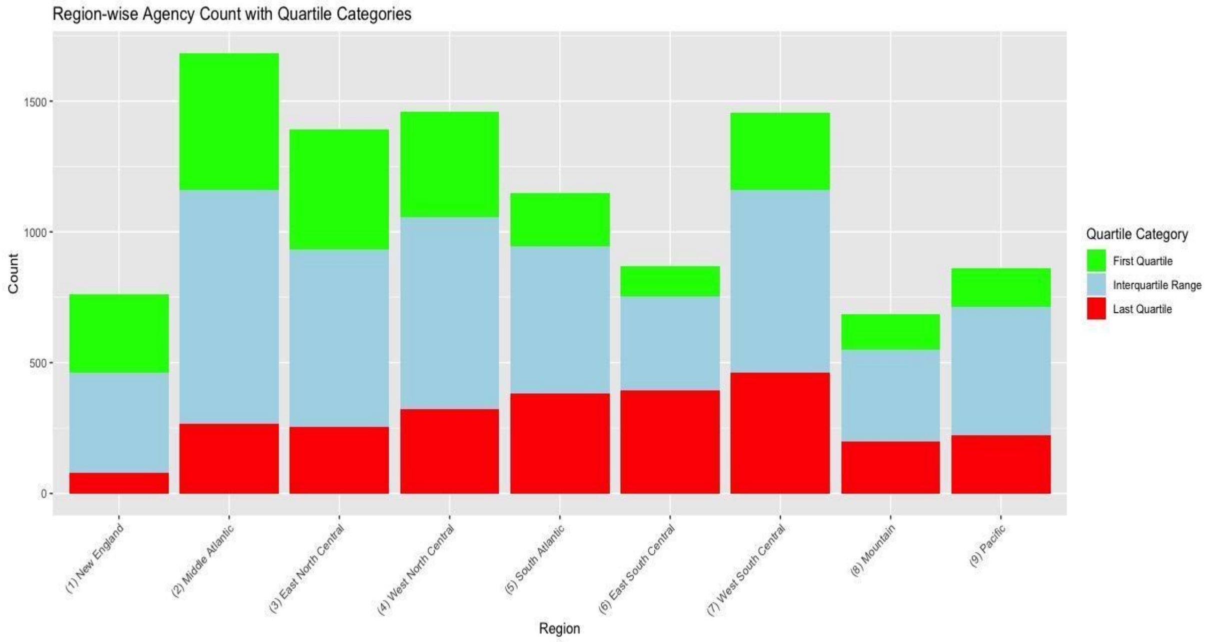
We then visualized the plot of ordered criminal rates for different agencies to better get an intuition about the distribution of criminal rates across agencies.

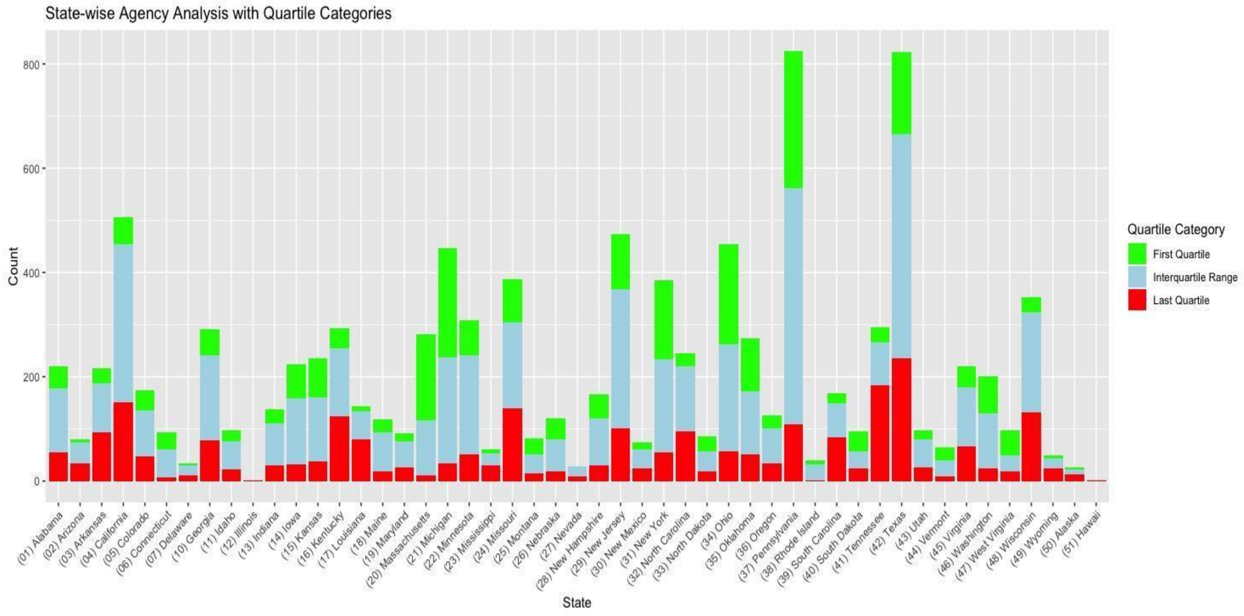


As we can see from the plot above, we can see that there are some extreme values for criminal rates as can be inferred by the sudden jump after 10000 agencies. So, we divided Agency performance using the interquartile range as follows-

1. Below First Quartile Criminal Rates- Good
2. Interquartile Criminal Rates- Moderate
3. Above Third Quartile Criminal Rates- Bad

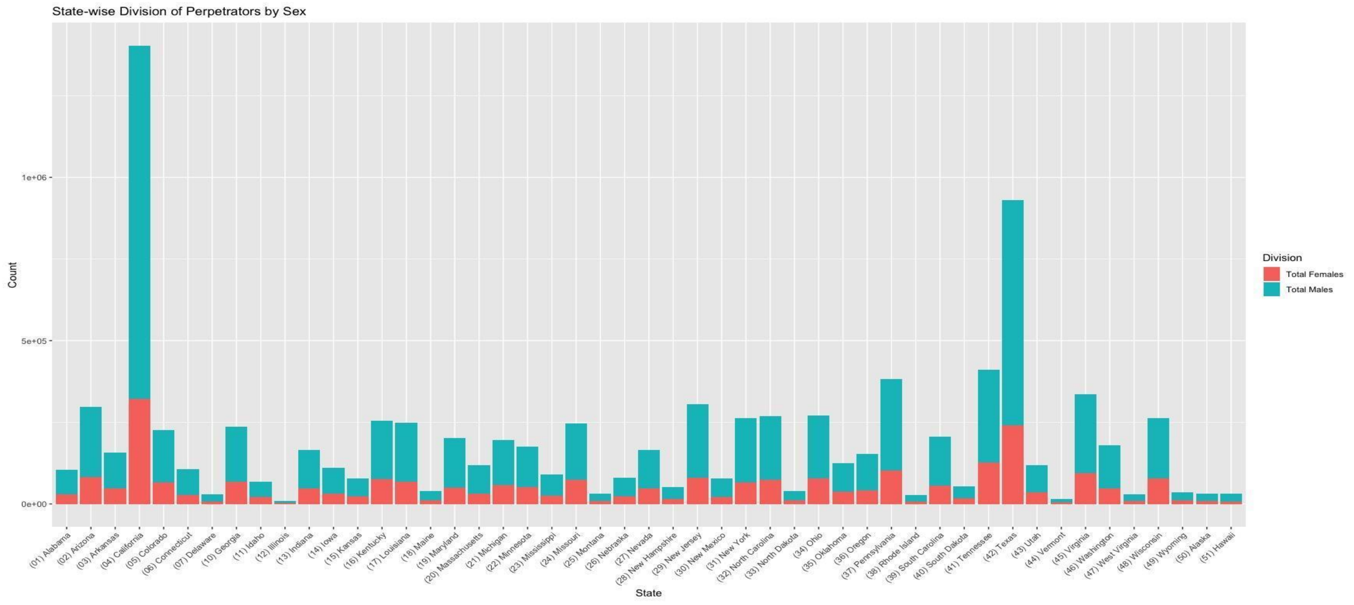
Then we visualized the State and Region level Agency performance based on the above metric.

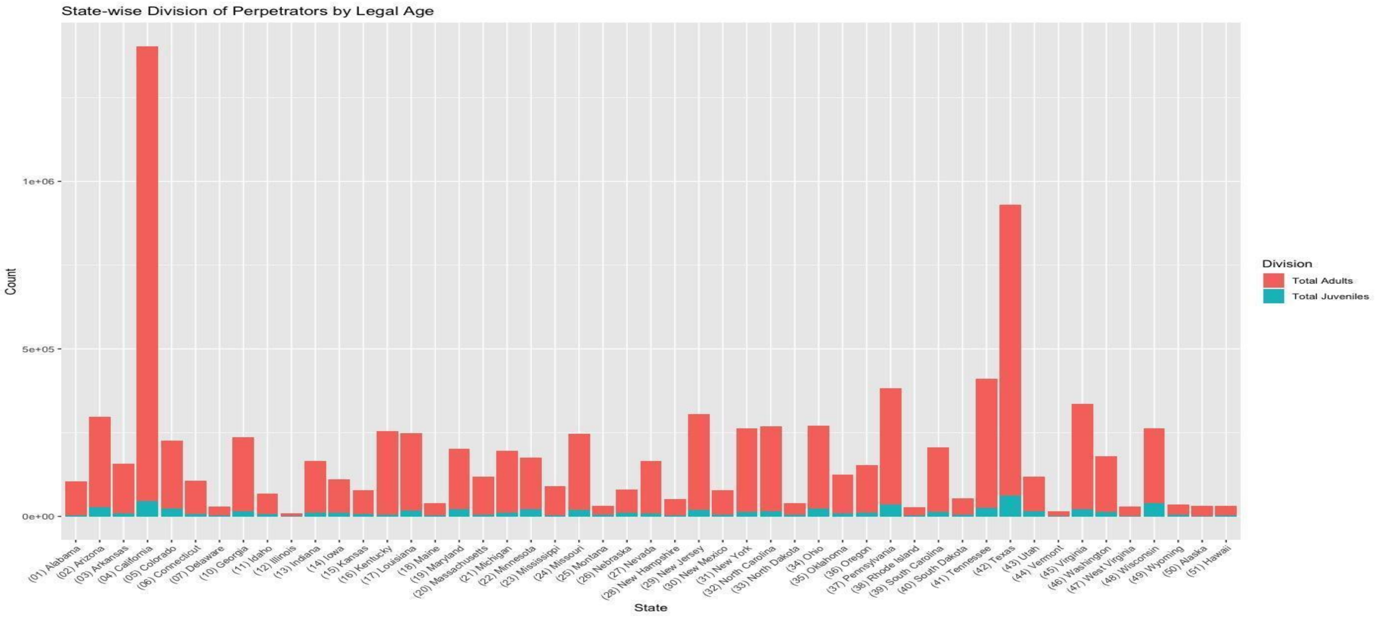


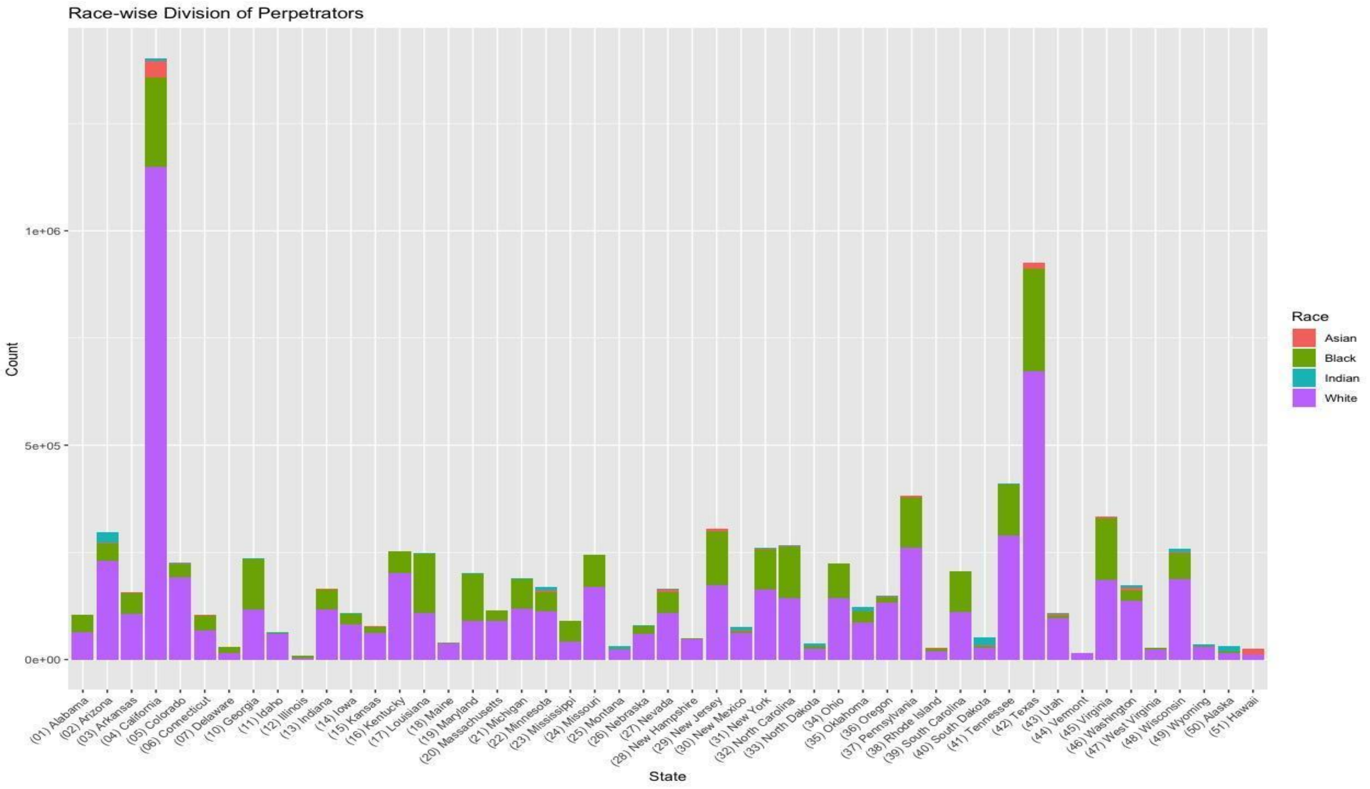


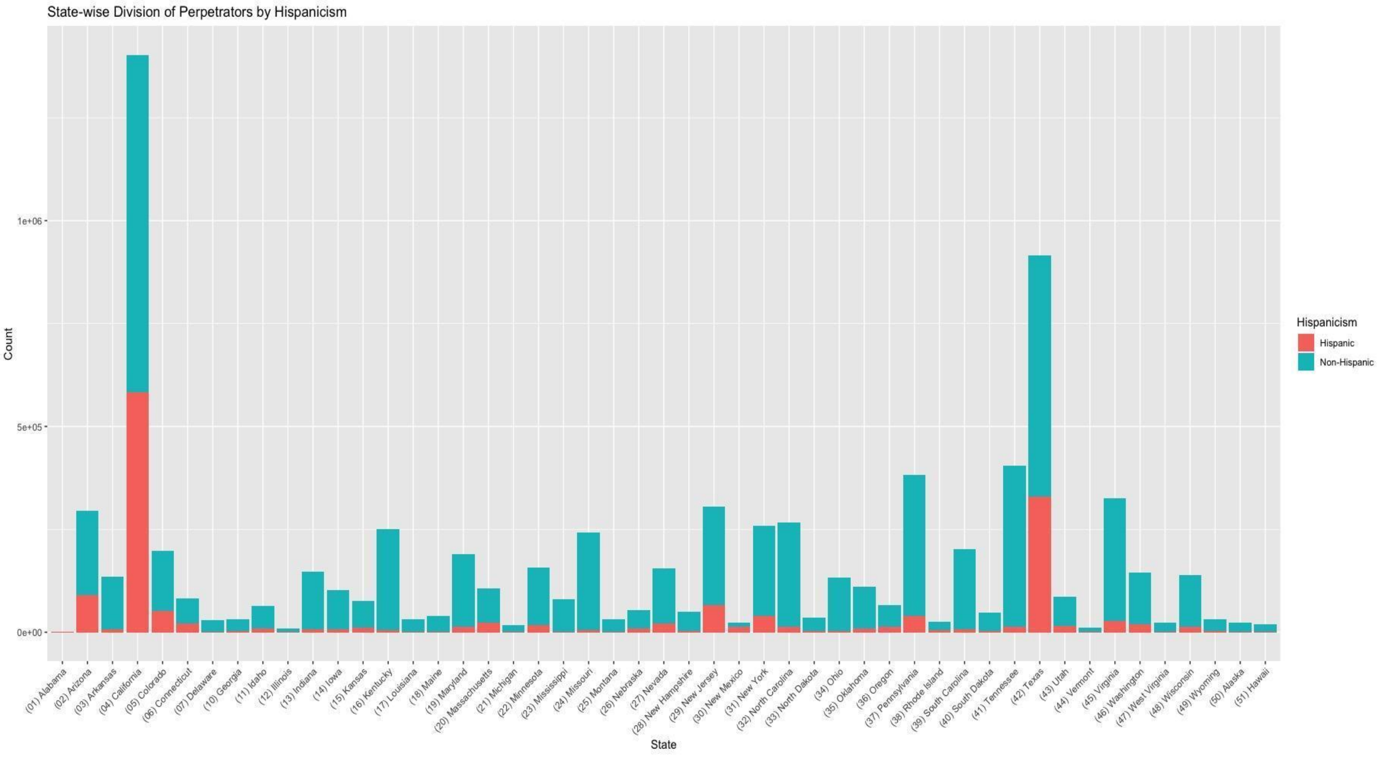
State Wise Analysis

Different factors were visualized to get a grasp of the data better. Firstly we analyzed the division of perpetrators by sex, age (minor or not), and race(White/Black/Asian/Indian and Hispanicism) across the states.

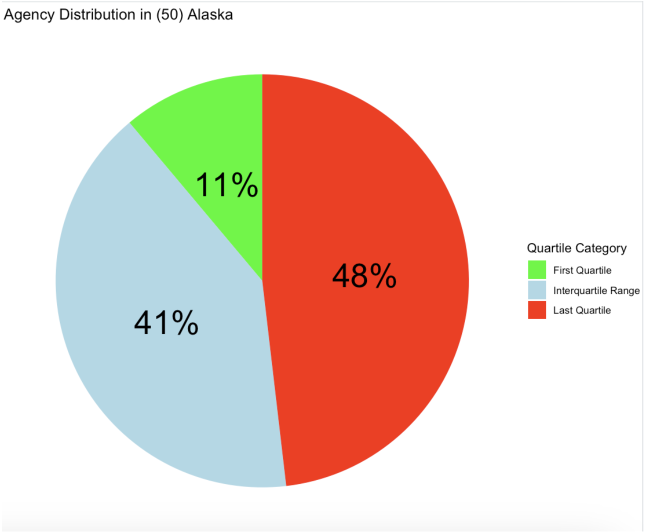


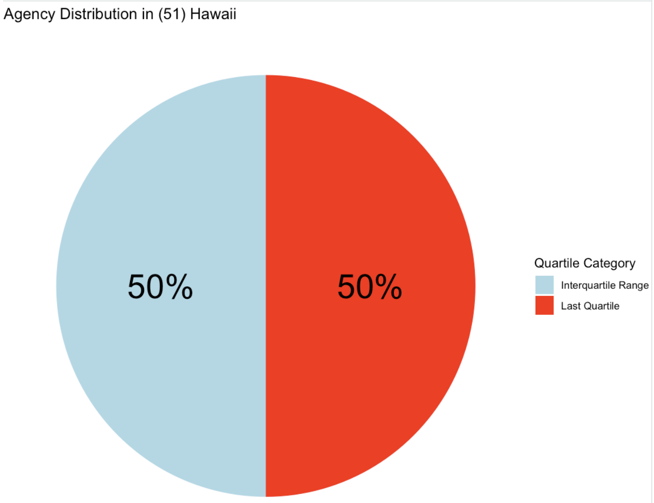


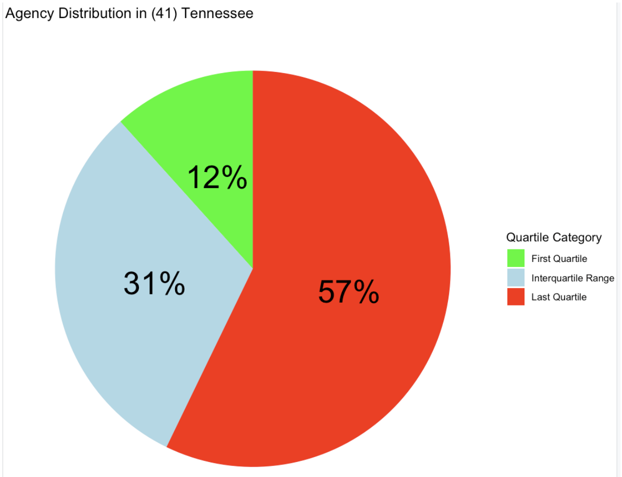

Race-wise Division of Perpetrators


State-wise Division of Perpetrators by Hispanicism

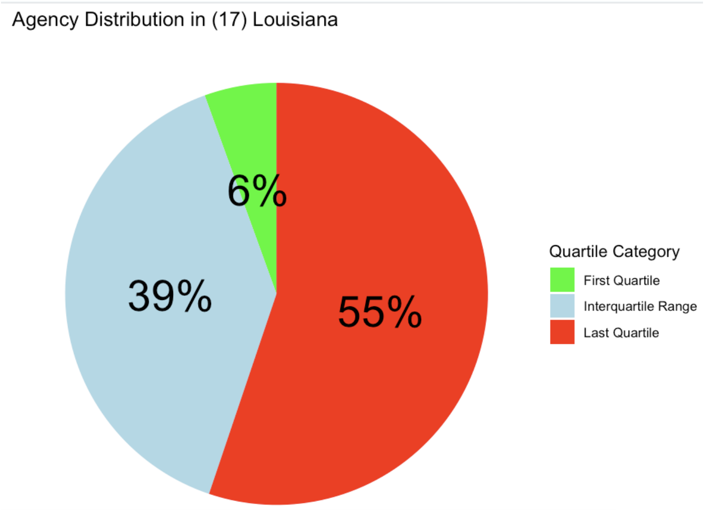
Worst Performing States In Suburban Areas

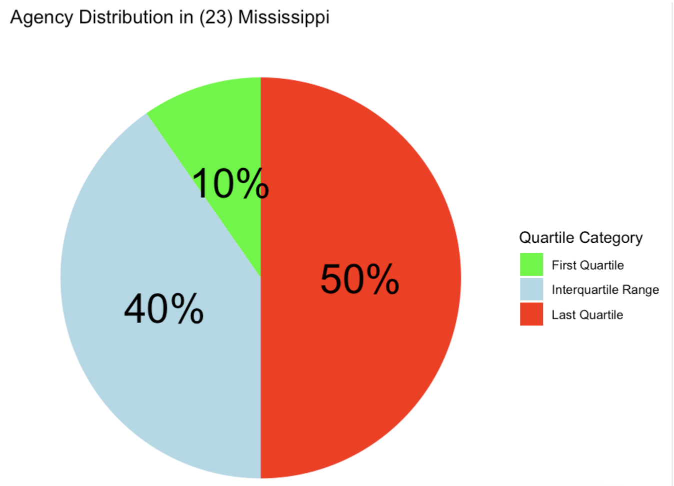


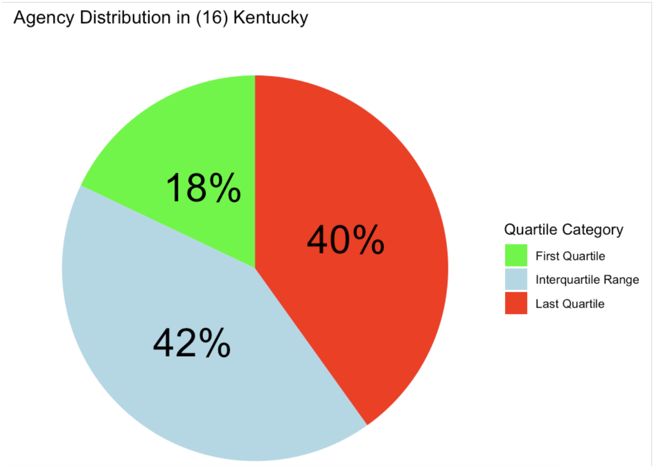




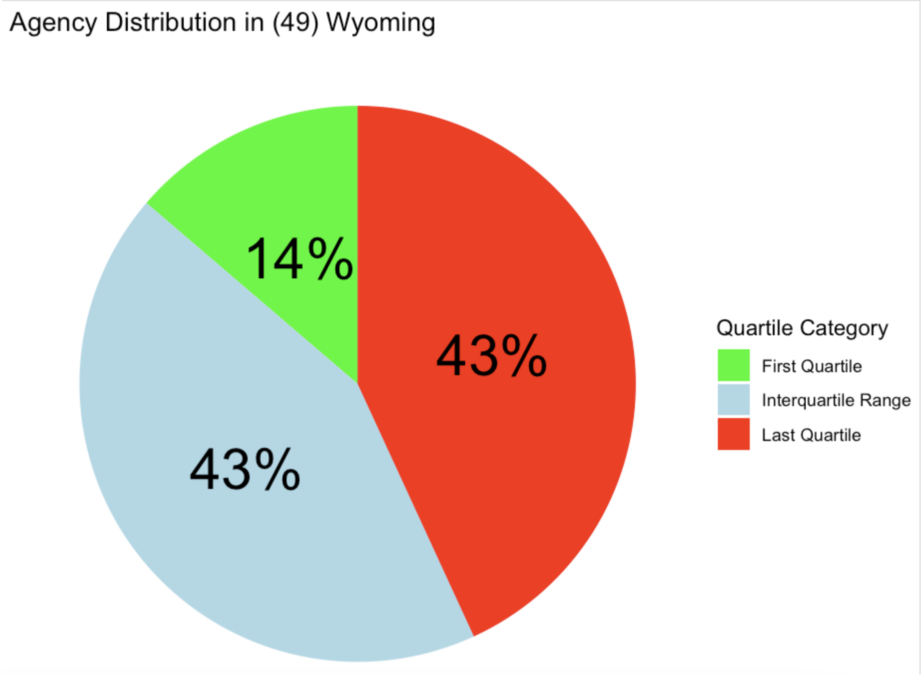
Worst Performing States In Non-Suburban Areas

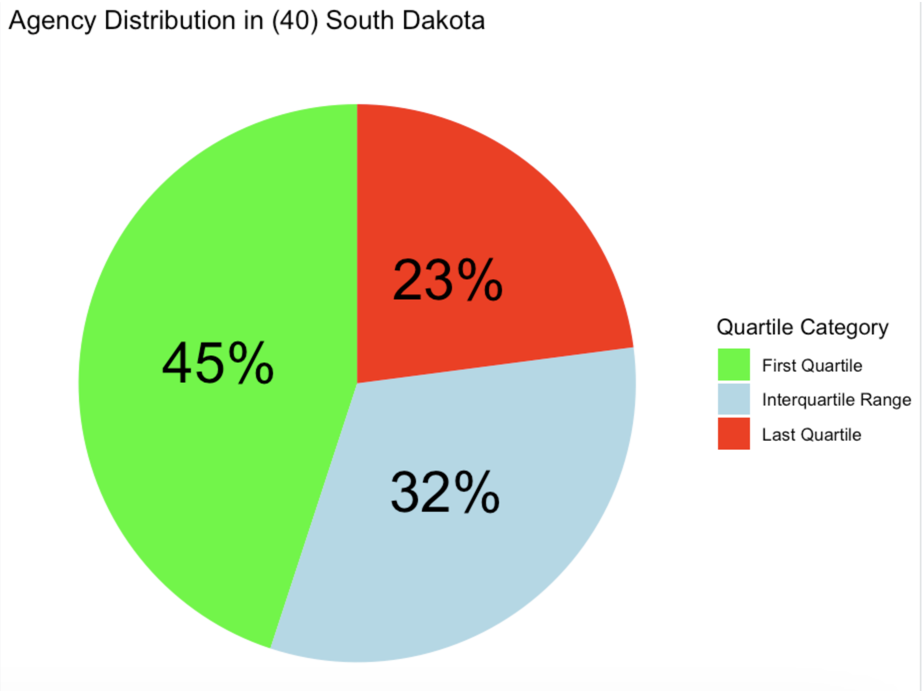

Agency Distribution in (17) Louisiana


Agency Distribution in (23) Mississippi


Agency Distribution in (16) Kentucky

Worst Performing States Overall



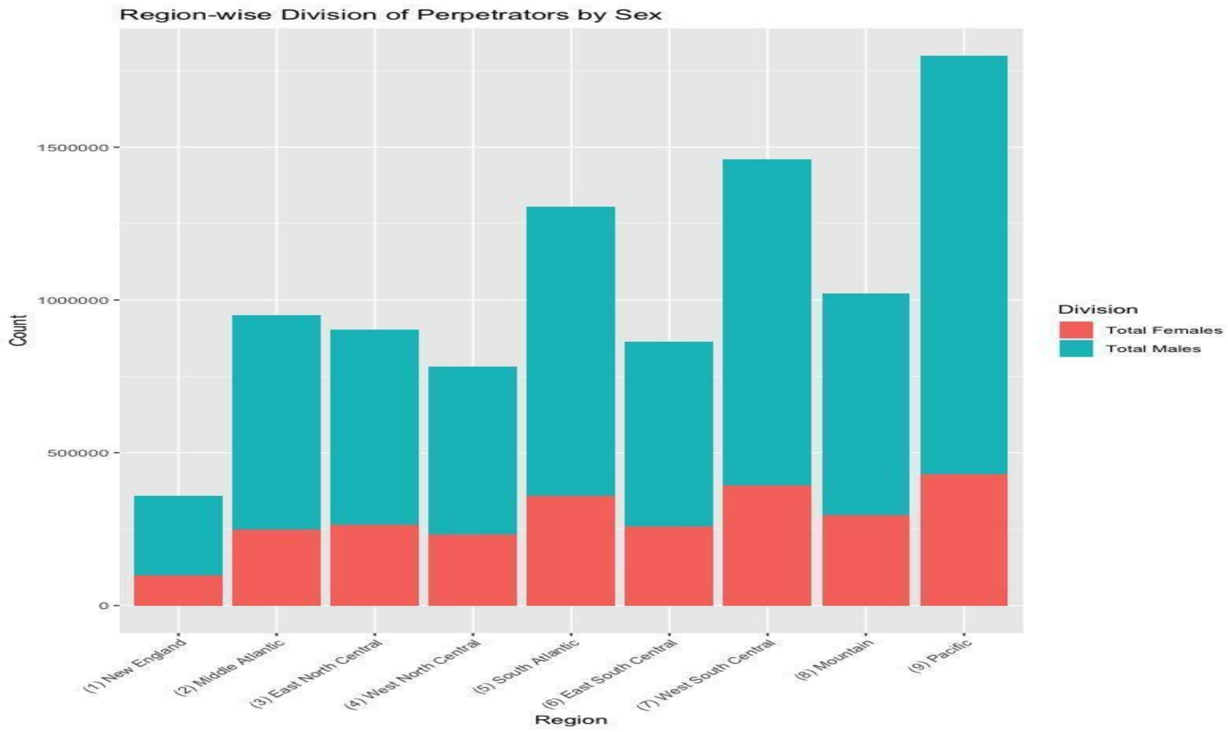


Region Wise Analysis

According to the dataset compiled by the FBI of The United States is divided into the following regions-

1. New England
2. Middle Atlantic
3. East North Central
4. West North Central
5. South Atlantic
6. East South Central
7. West South Central

We visualized the perpetrators by the usual metrics region-wise too.



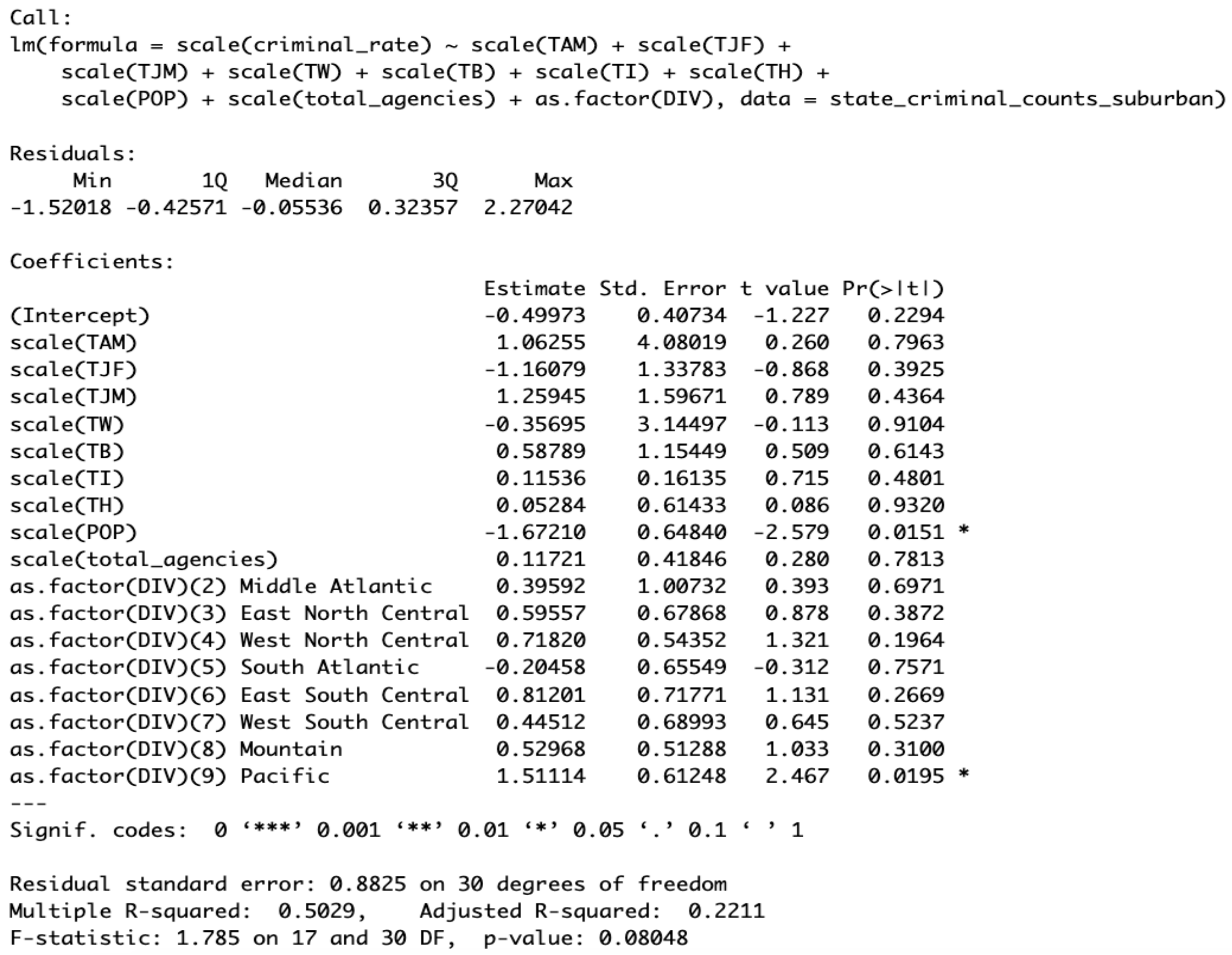
Methodology

Linear Regression Analysis

Initial Model using all available predictors.

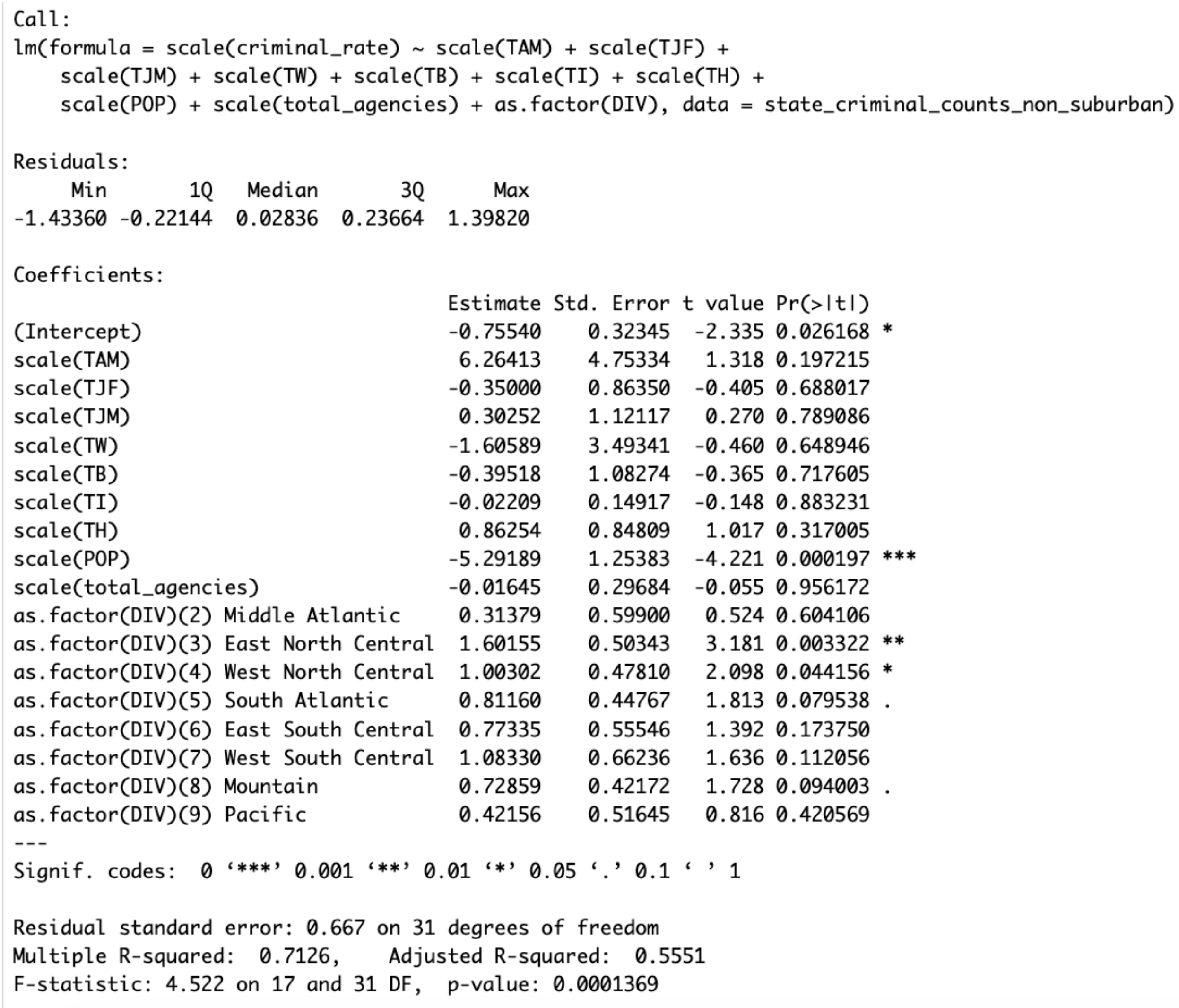
We first fitted a Linear Regression Model with Criminal Rate as the response and all the predictors present in the summarized state-wise dataset. The state-wise dataset is further divided into two suburban and non-suburban state-wise datasets. Both the response variable and the predictors were standardized to help us with our statistical analysis.

Suburban_Data:



```
Call:
lm(formula = scale(criminal_rate) ~ scale(TAM) + scale(TJF) +
    scale(TJM) + scale(TW) + scale(TB) + scale(TI) + scale(TH) +
    scale(POP) + scale(total_agencies) + as.factor(DIV), data = state_criminal_counts_suburban)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.52018 -0.42571 -0.05536  0.32357  2.27042 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)  
(Intercept)                          -0.49973    0.40734  -1.227   0.2294  
scale(TAM)                            1.06255    4.08019   0.260   0.7963  
scale(TJF)                           -1.16079    1.33783  -0.868   0.3925  
scale(TJM)                            1.25945    1.59671   0.789   0.4364  
scale(TW)                            -0.35695    3.14497  -0.113   0.9104  
scale(TB)                             0.58789    1.15449   0.509   0.6143  
scale(TI)                             0.11536    0.16135   0.715   0.4801  
scale(TH)                             0.05284    0.61433   0.086   0.9320  
scale(POP)                           -1.67210    0.64840  -2.579   0.0151 *
scale(total_agencies)                 0.11721    0.41846   0.280   0.7813  
as.factor(DIV)(2) Middle Atlantic     0.39592    1.00732   0.393   0.6971  
as.factor(DIV)(3) East North Central  0.59557    0.67868   0.878   0.3872  
as.factor(DIV)(4) West North Central  0.71820    0.54352   1.321   0.1964  
as.factor(DIV)(5) South Atlantic     -0.20458    0.65549  -0.312   0.7571  
as.factor(DIV)(6) East South Central  0.81201    0.71771   1.131   0.2669  
as.factor(DIV)(7) West South Central  0.44512    0.68993   0.645   0.5237  
as.factor(DIV)(8) Mountain            0.52968    0.51288   1.033   0.3100  
as.factor(DIV)(9) Pacific             1.51114    0.61248   2.467   0.0195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8825 on 30 degrees of freedom
Multiple R-squared:  0.5029,	Adjusted R-squared:  0.2211 
F-statistic: 1.785 on 17 and 30 DF,  p-value: 0.08048
```

Non_Suburban_Data:



```
Call:
lm(formula = scale(criminal_rate) ~ scale(TAM) + scale(TJF) +
    scale(TJM) + scale(TW) + scale(TB) + scale(TI) + scale(TH) +
    scale(POP) + scale(total_agencies) + as.factor(DIV), data = state_criminal_counts_non_suburban)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43360 -0.22144  0.02836  0.23664  1.39820

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          -0.75540    0.32345  -2.335 0.026168 *
scale(TAM)                            6.26413    4.75334   1.318 0.197215
scale(TJF)                           -0.35000    0.86350  -0.405 0.688017
scale(TJM)                            0.30252    1.12117   0.270 0.789086
scale(TW)                            -1.60589    3.49341  -0.460 0.648946
scale(TB)                            -0.39518    1.08274  -0.365 0.717605
scale(TI)                            -0.02209    0.14917  -0.148 0.883231
scale(TH)                             0.86254    0.84809   1.017 0.317005
scale(POP)                           -5.29189    1.25383  -4.221 0.000197 ***
scale(total_agencies)                -0.01645    0.29684  -0.055 0.956172
as.factor(DIV)(2) Middle Atlantic     0.31379    0.59900   0.524 0.604106
as.factor(DIV)(3) East North Central  1.60155    0.50343   3.181 0.003322 **
as.factor(DIV)(4) West North Central  1.00302    0.47810   2.098 0.044156 *
as.factor(DIV)(5) South Atlantic      0.81160    0.44767   1.813 0.079538 .
as.factor(DIV)(6) East South Central  0.77335    0.55546   1.392 0.173750
as.factor(DIV)(7) West South Central  1.08330    0.66236   1.636 0.112056
as.factor(DIV)(8) Mountain            0.72859    0.42172   1.728 0.094003 .
as.factor(DIV)(9) Pacific             0.42156    0.51645   0.816 0.420569
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.667 on 31 degrees of freedom
Multiple R-squared:  0.7126,    Adjusted R-squared:  0.5551
F-statistic: 4.522 on 17 and 31 DF,  p-value: 0.0001369
```

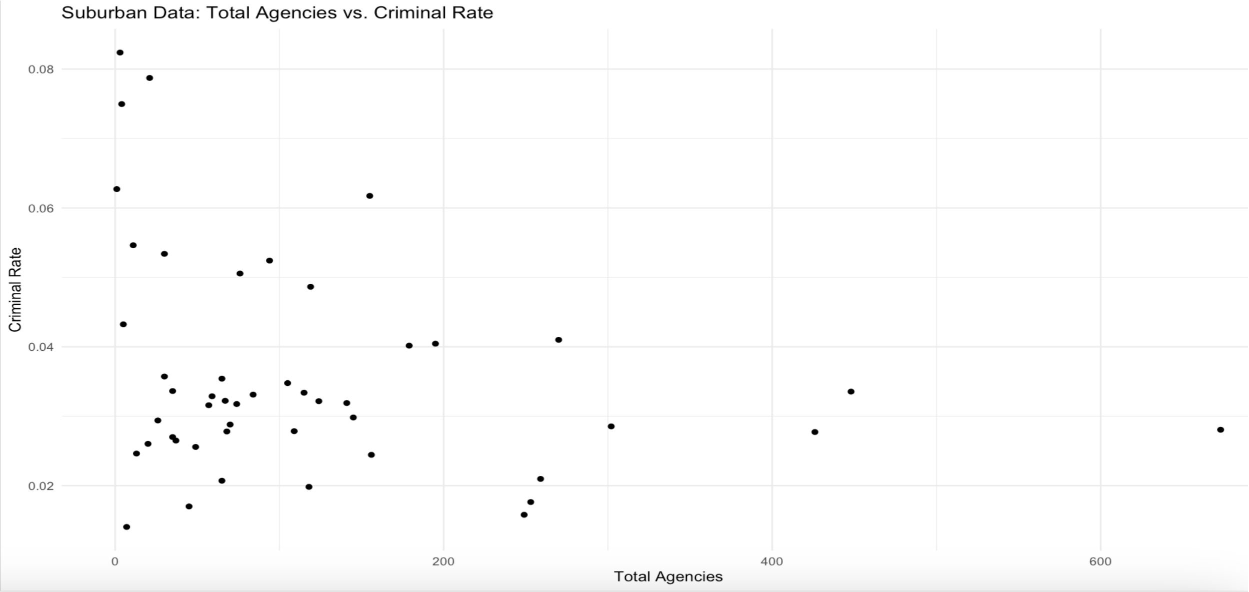
Preliminary Inferences-

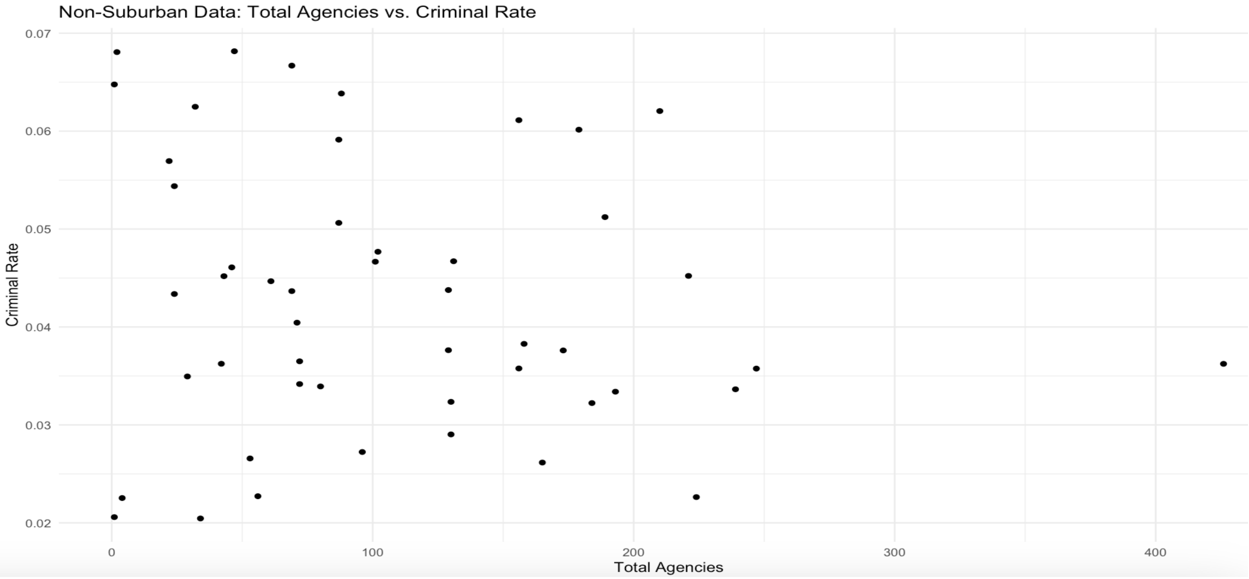
- POP and some DIV factors (East North Central, West North Central, and Pacific) seem statistically significant predictors with p-values<0.05.
- The intercept’s p-value is close to 0.05 for non_suburban_data suggesting that the value of response is non-zero when all the other predictors are zero.
- Adjusted-R squared value of 0.5551 for non-suburban data suggests that the model explains 55.5% percent of the variability in the response, adjusted for the number of predictors. This suggests a moderate level of explanatory power.
- The model requires further analysis of the relationship between the predictors and the response variable.

Correlation Analysis

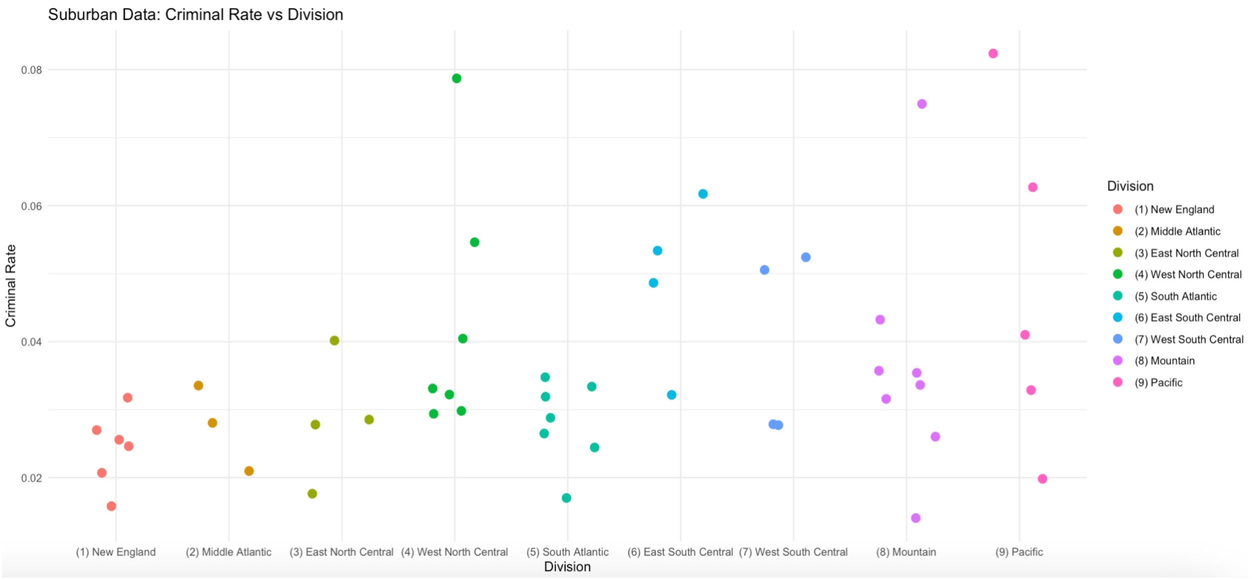
We wanted to analyze the relationship between some variables, and we made use of scatter plots to get an idea about the relationship.

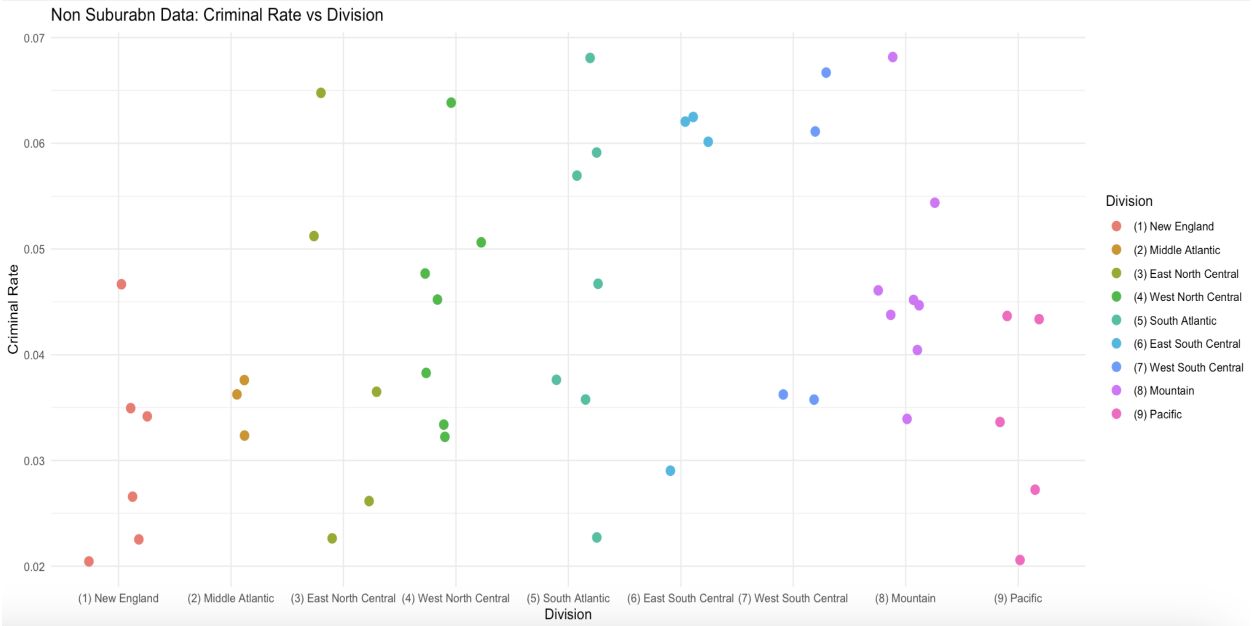
- Criminal_Rate_vs_Agency_Count
  We can infer from the plot below that there is minimal correlation if any between total agencies in a state vs criminal rate. This is an expected result as we would hope that there is some positive correlation between population under and agency vs criminal rate.



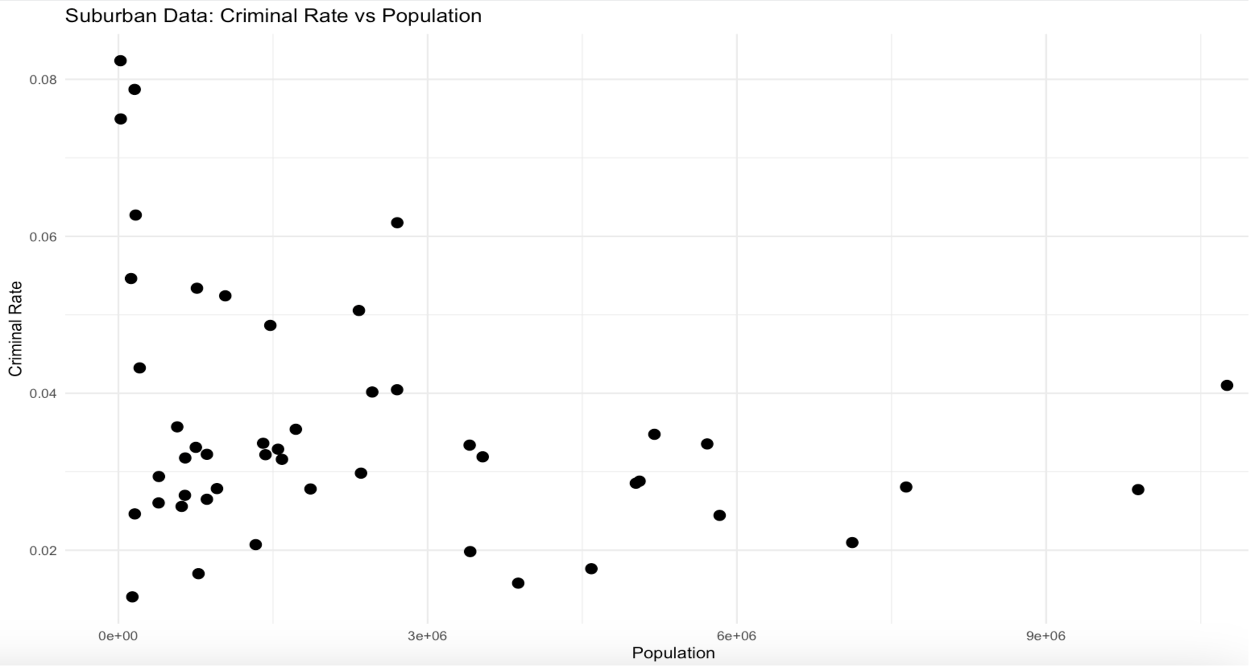


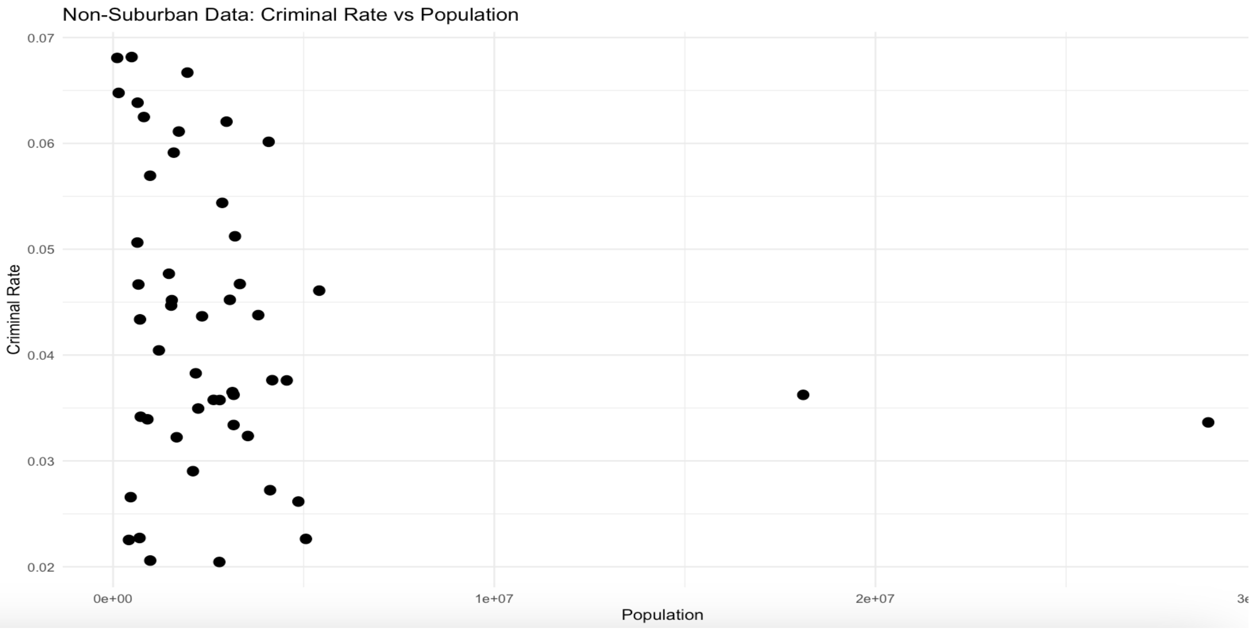
- Criminal_Rate_vs_Suburban
  In the plot below we have the scatter plot between criminal rate and the categorical variable division. From the plots below we can see that some regions especially New England, South Atlantic, and Pacific areas have lower crime rates for suburban areas. Also, we can observe that there is more variation in the non-suburban criminal rates for all the regions except for the Pacific.





- Criminal_Rate_vs_Population
  In the plot below we have the scatter plot between criminal rate and summation of population under all the agencies being considered for a state. From the plots below there is no clear relationship between population and criminal rate despite it being an important predictor variable.

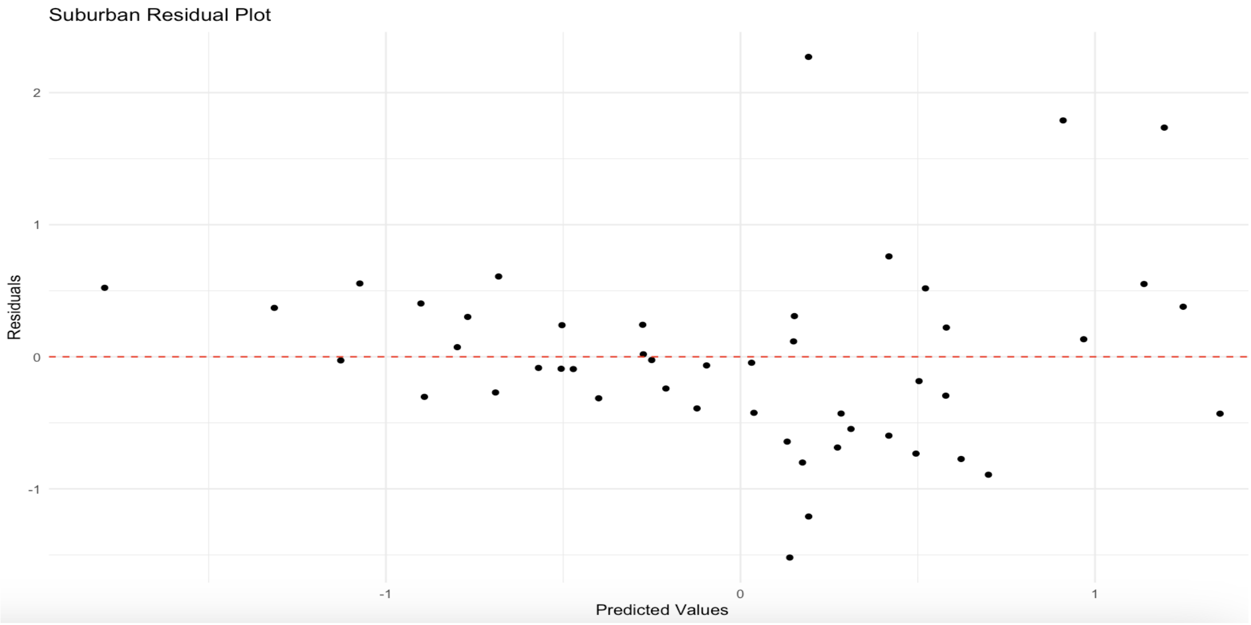


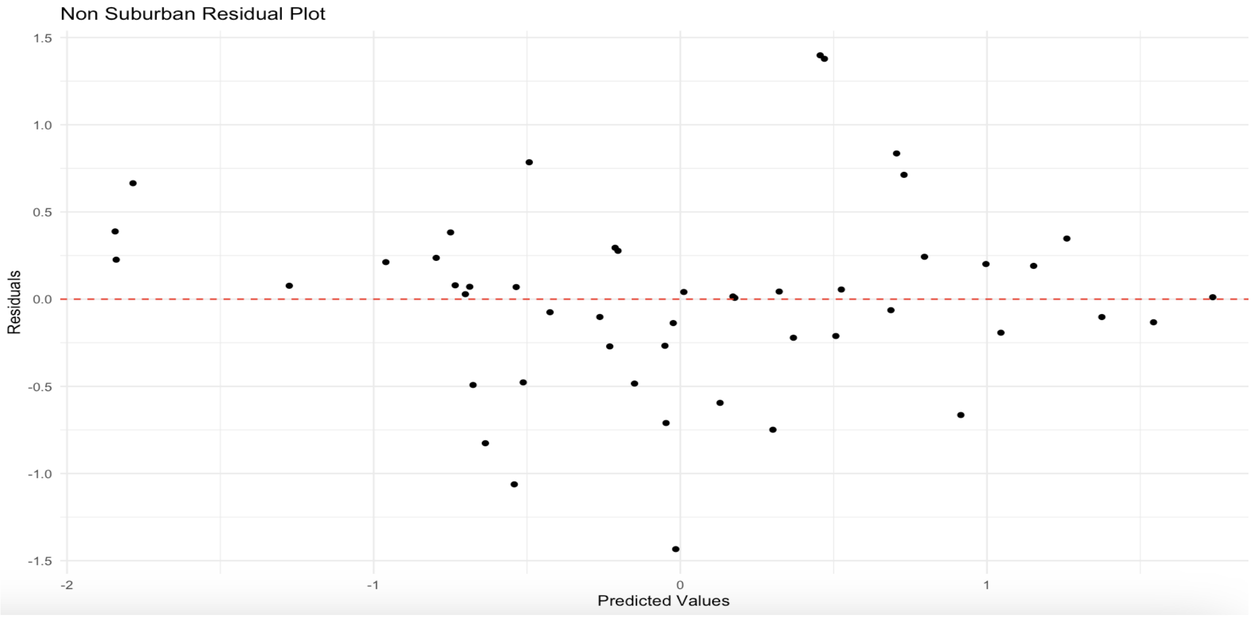


Model Diagnostics

Residual Plot

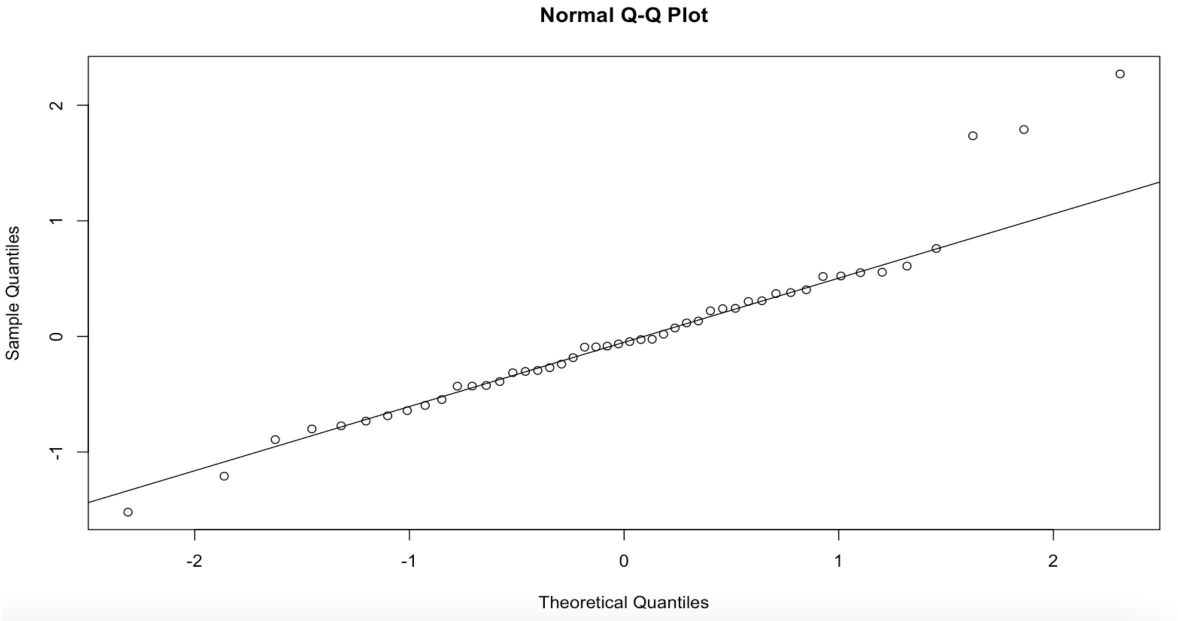
The residual plots below seem to have a constant mean around zero and most of the residuals seem to be within a constant band around the y=0 red line. Some values are outside the band, but they can be treated as anomalies/outliers.



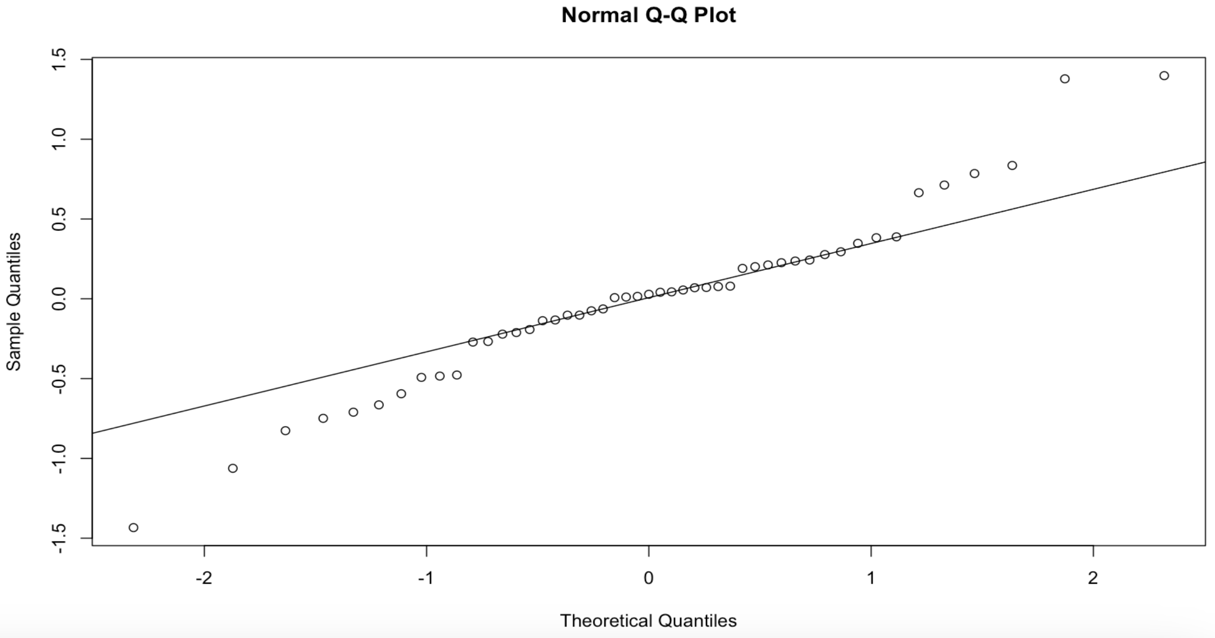


Normality Check (Quantile-Quantile Plot and Shapiro Wilk Test)

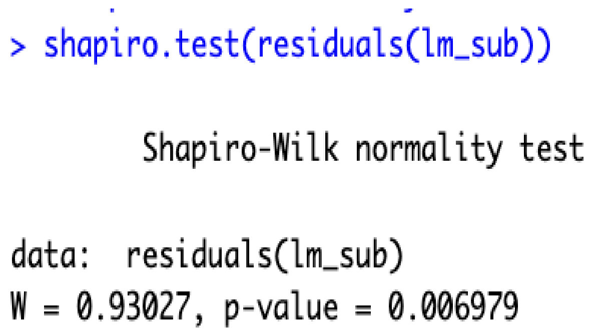
All the dots seem to generally form a line as depicted in the Q-Q plot for suburban data below. Therefore, we can safely assume that our residuals come from a normal distribution.



Q-Q plot looks similar to an inverted S, indicating the possibility of a heavy-tailed distribution.

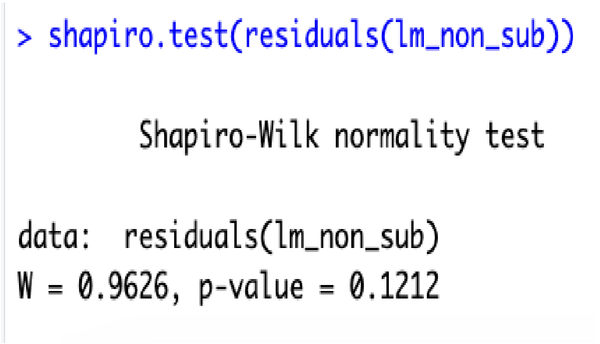


We have a p-value= 0.006979 (<0.05), therefore we reject the null hypothesis which assumes the residuals to be normally distributed for suburban data.



```
> shapiro.test(residuals(lm_sub))

	Shapiro-Wilk normality test

data:  residuals(lm_sub)
W = 0.93027, p-value = 0.006979
```

For non-suburban data, however, p-value=0.121(>0.05), therefore we cannot reject the null hypothesis.

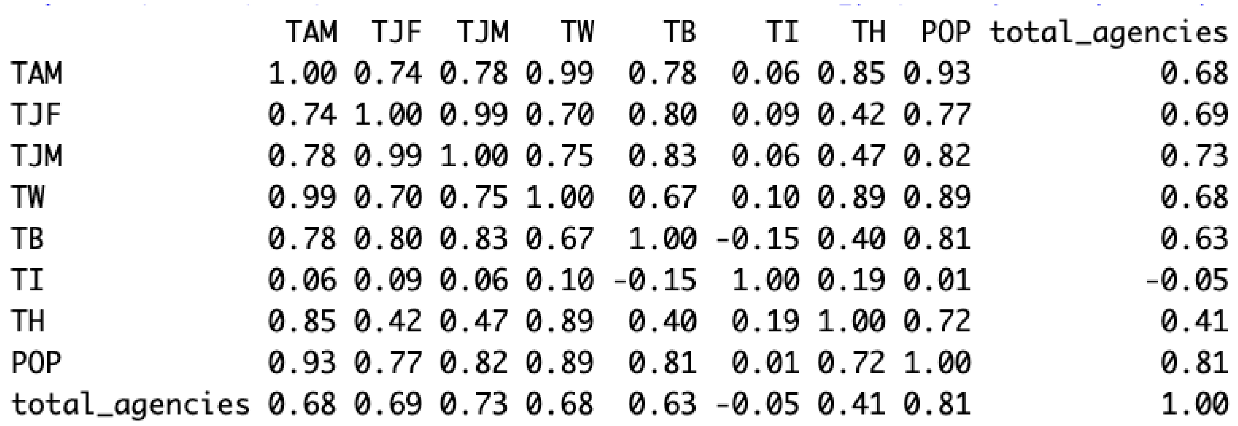


```
> shapiro.test(residuals(lm_non_sub))

	Shapiro-Wilk normality test

data:  residuals(lm_non_sub)
W = 0.9626, p-value = 0.1212
```

Correlation between non-categorical predictors

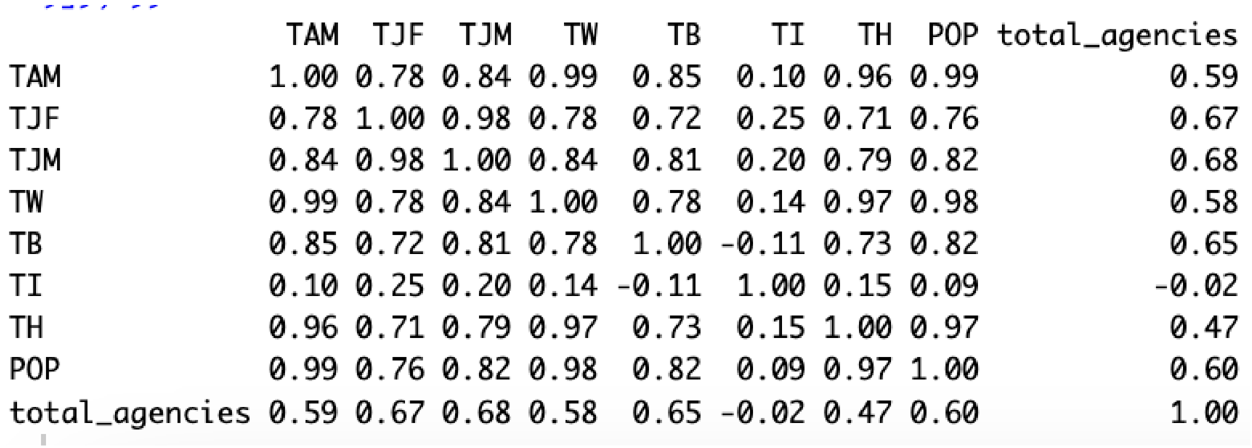
From the image below we can see that the several predictors are highly correlated with other predictors in both the suburban and non-suburban data. TI representing the total Indian perpetrators seems to be the only predictor uncorrelated with the other predictors. POP representing the population (one of the most statistically significant predictor variables) is strongly correlated with predictors like TAM and TW.

Suburban Data-



```
                TAM  TJF  TJM   TW    TB    TI   TH  POP total_agencies
TAM            1.00 0.74 0.78 0.99  0.78  0.06 0.85 0.93           0.68
TJF            0.74 1.00 0.99 0.70  0.80  0.09 0.42 0.77           0.69
TJM            0.78 0.99 1.00 0.75  0.83  0.06 0.47 0.82           0.73
TW             0.99 0.70 0.75 1.00  0.67  0.10 0.89 0.89           0.68
TB             0.78 0.80 0.83 0.67  1.00 -0.15 0.40 0.81           0.63
TI             0.06 0.09 0.06 0.10 -0.15  1.00 0.19 0.01          -0.05
TH             0.85 0.42 0.47 0.89  0.40  0.19 1.00 0.72           0.41
POP            0.93 0.77 0.82 0.89  0.81  0.01 0.72 1.00           0.81
total_agencies 0.68 0.69 0.73 0.68  0.63 -0.05 0.41 0.81           1.00
```

Non-Suburban Data-

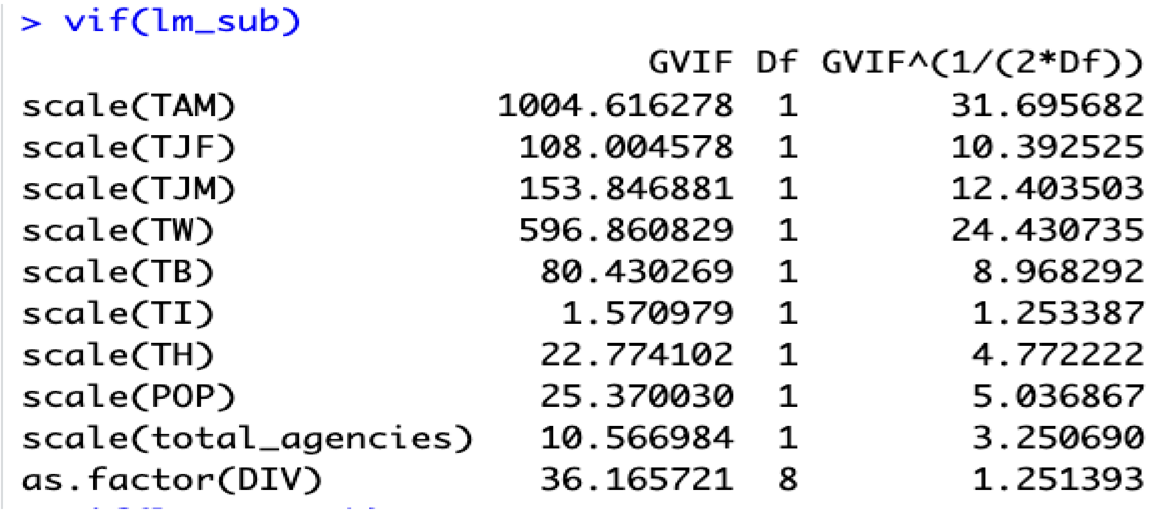


```
                TAM  TJF  TJM   TW    TB    TI   TH  POP total_agencies
TAM            1.00 0.78 0.84 0.99  0.85  0.10 0.96 0.99           0.59
TJF            0.78 1.00 0.98 0.78  0.72  0.25 0.71 0.76           0.67
TJM            0.84 0.98 1.00 0.84  0.81  0.20 0.79 0.82           0.68
TW             0.99 0.78 0.84 1.00  0.78  0.14 0.97 0.98           0.58
TB             0.85 0.72 0.81 0.78  1.00 -0.11 0.73 0.82           0.65
TI             0.10 0.25 0.20 0.14 -0.11  1.00 0.15 0.09          -0.02
TH             0.96 0.71 0.79 0.97  0.73  0.15 1.00 0.97           0.47
POP            0.99 0.76 0.82 0.98  0.82  0.09 0.97 1.00           0.60
total_agencies 0.59 0.67 0.68 0.58  0.65 -0.02 0.47 0.60           1.00
```

Multicollinearity

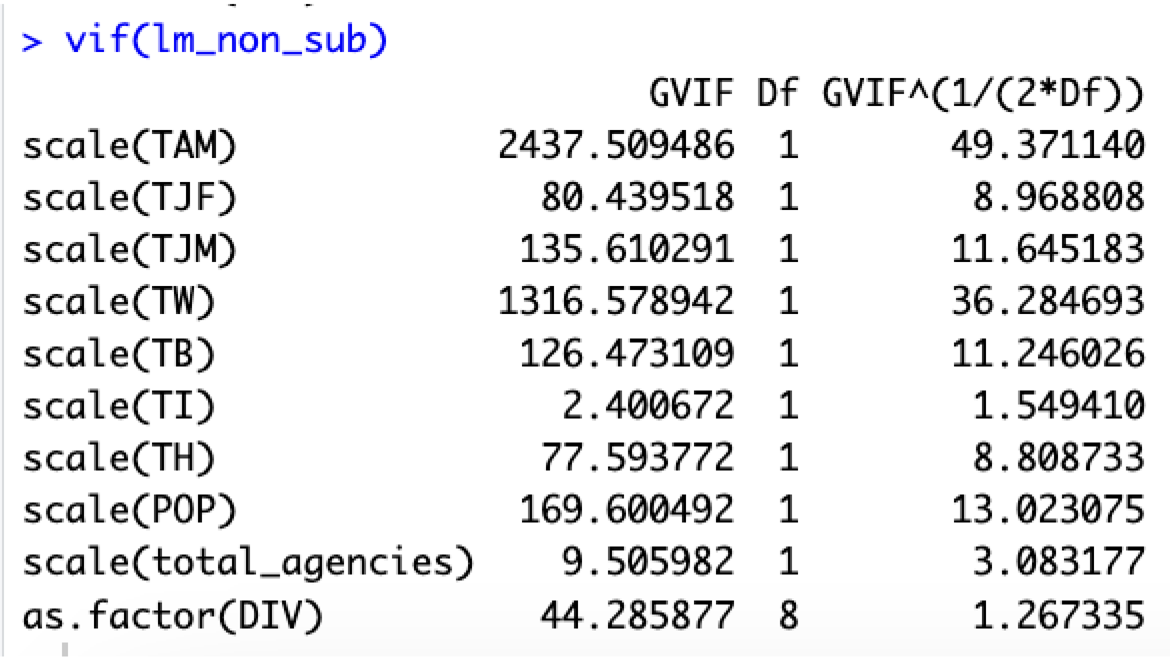
We have already scaled the variables to address the potential collinearity between variables and make the variation more manageable. Even after that, we are testing the adequacy of the model through a technique called Variance Inflation Factor (VIF) for each of the predictors, before the next steps in our analysis.

Suburban Data:



```
> vif(lm_sub)
                            GVIF Df GVIF^(1/(2*Df))
scale(TAM)           1004.616278  1       31.695682
scale(TJF)            108.004578  1       10.392525
scale(TJM)            153.846881  1       12.403503
scale(TW)             596.860829  1       24.430735
scale(TB)              80.430269  1        8.968292
scale(TI)               1.570979  1        1.253387
scale(TH)              22.774102  1        4.772222
scale(POP)             25.370030  1        5.036867
scale(total_agencies)  10.566984  1        3.250690
as.factor(DIV)         36.165721  8        1.251393
```

Non-Suburban Data:



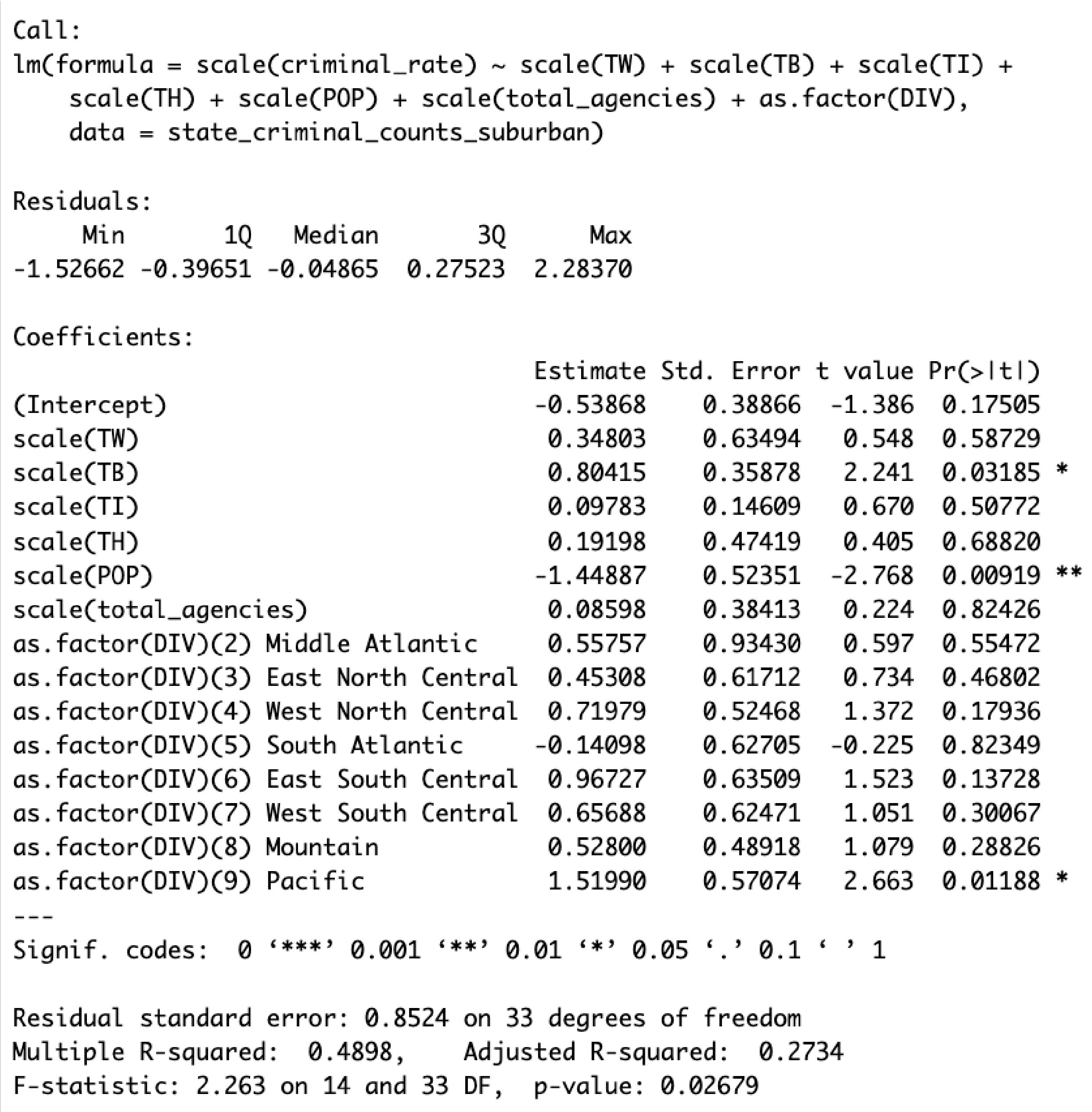
```
> vif(lm_non_sub)
                            GVIF Df GVIF^(1/(2*Df))
scale(TAM)           2437.509486  1       49.371140
scale(TJF)             80.439518  1        8.968808
scale(TJM)            135.610291  1       11.645183
scale(TW)            1316.578942  1       36.284693
scale(TB)             126.473109  1       11.246026
scale(TI)               2.400672  1        1.549410
scale(TH)              77.593772  1        8.808733
scale(POP)            169.600492  1       13.023075
scale(total_agencies)   9.505982  1        3.083177
as.factor(DIV)         44.285877  8        1.267335
```

The extremely high GVIF values (equivalent to VIF for one-coefficient terms) suggest high multicollinearity in our model.

Updated Model:

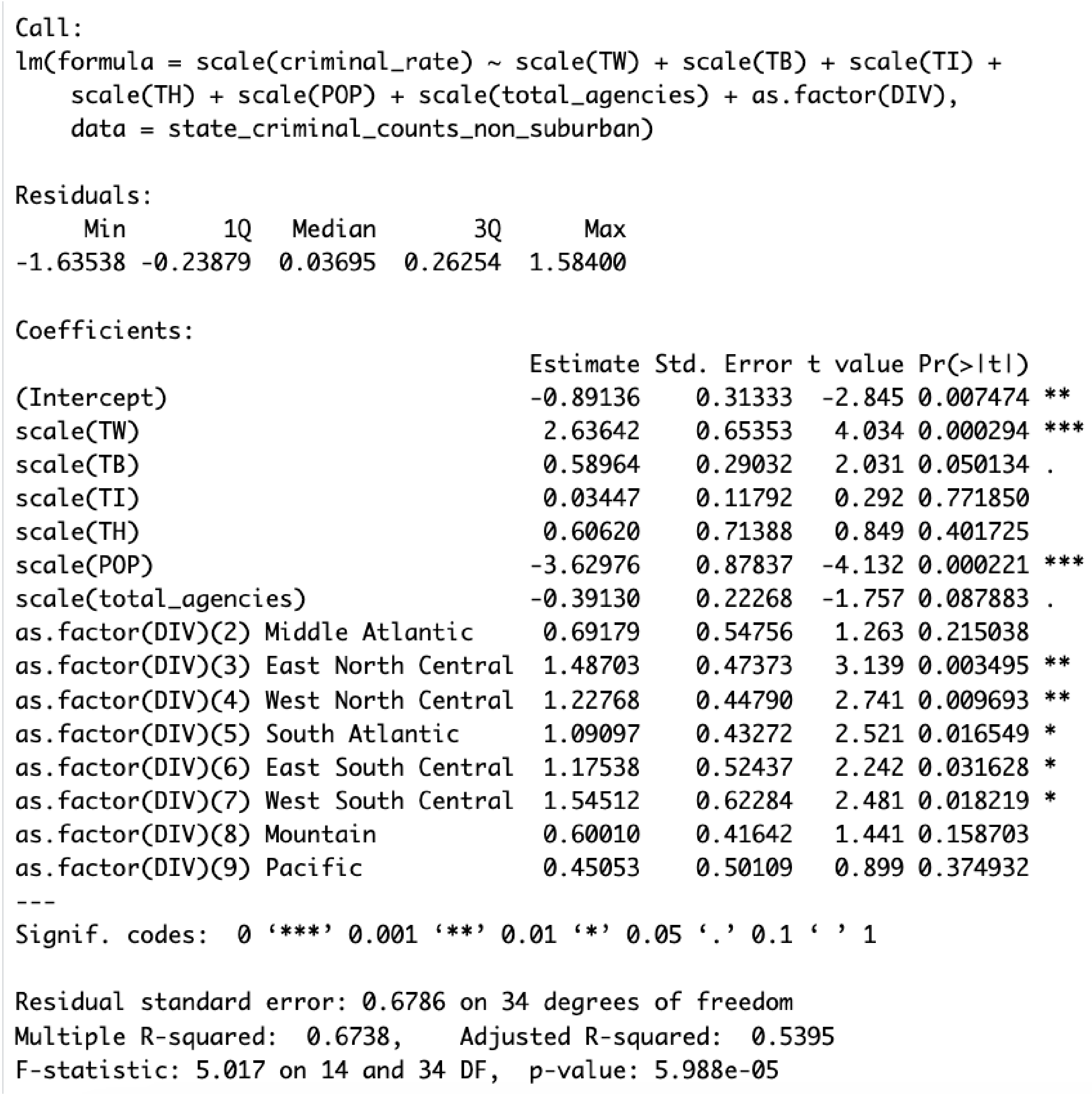
We consequently removed some of the predictors manually (selective replacement), which didn't affect our adjusted R-squared value. Our new models had the following specifications-

Suburban-Data Model:



```
Call:
lm(formula = scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(TI) +
    scale(TH) + scale(POP) + scale(total_agencies) + as.factor(DIV),
    data = state_criminal_counts_suburban)

Residuals:
     Min       1Q   Median       3Q      Max
-1.52662 -0.39651 -0.04865  0.27523  2.28370

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                          -0.53868    0.38866  -1.386  0.17505
scale(TW)                             0.34803    0.63494   0.548  0.58729
scale(TB)                             0.80415    0.35878   2.241  0.03185 *
scale(TI)                             0.09783    0.14609   0.670  0.50772
scale(TH)                             0.19198    0.47419   0.405  0.68820
scale(POP)                           -1.44887    0.52351  -2.768  0.00919 **
scale(total_agencies)                 0.08598    0.38413   0.224  0.82426
as.factor(DIV)(2) Middle Atlantic     0.55757    0.93430   0.597  0.55472
as.factor(DIV)(3) East North Central  0.45308    0.61712   0.734  0.46802
as.factor(DIV)(4) West North Central  0.71979    0.52468   1.372  0.17936
as.factor(DIV)(5) South Atlantic     -0.14098    0.62705  -0.225  0.82349
as.factor(DIV)(6) East South Central  0.96727    0.63509   1.523  0.13728
as.factor(DIV)(7) West South Central  0.65688    0.62471   1.051  0.30067
as.factor(DIV)(8) Mountain            0.52800    0.48918   1.079  0.28826
as.factor(DIV)(9) Pacific             1.51990    0.57074   2.663  0.01188 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8524 on 33 degrees of freedom
Multiple R-squared:  0.4898,    Adjusted R-squared:  0.2734
F-statistic: 2.263 on 14 and 33 DF,  p-value: 0.02679
```

Non-Suburban Data Model:



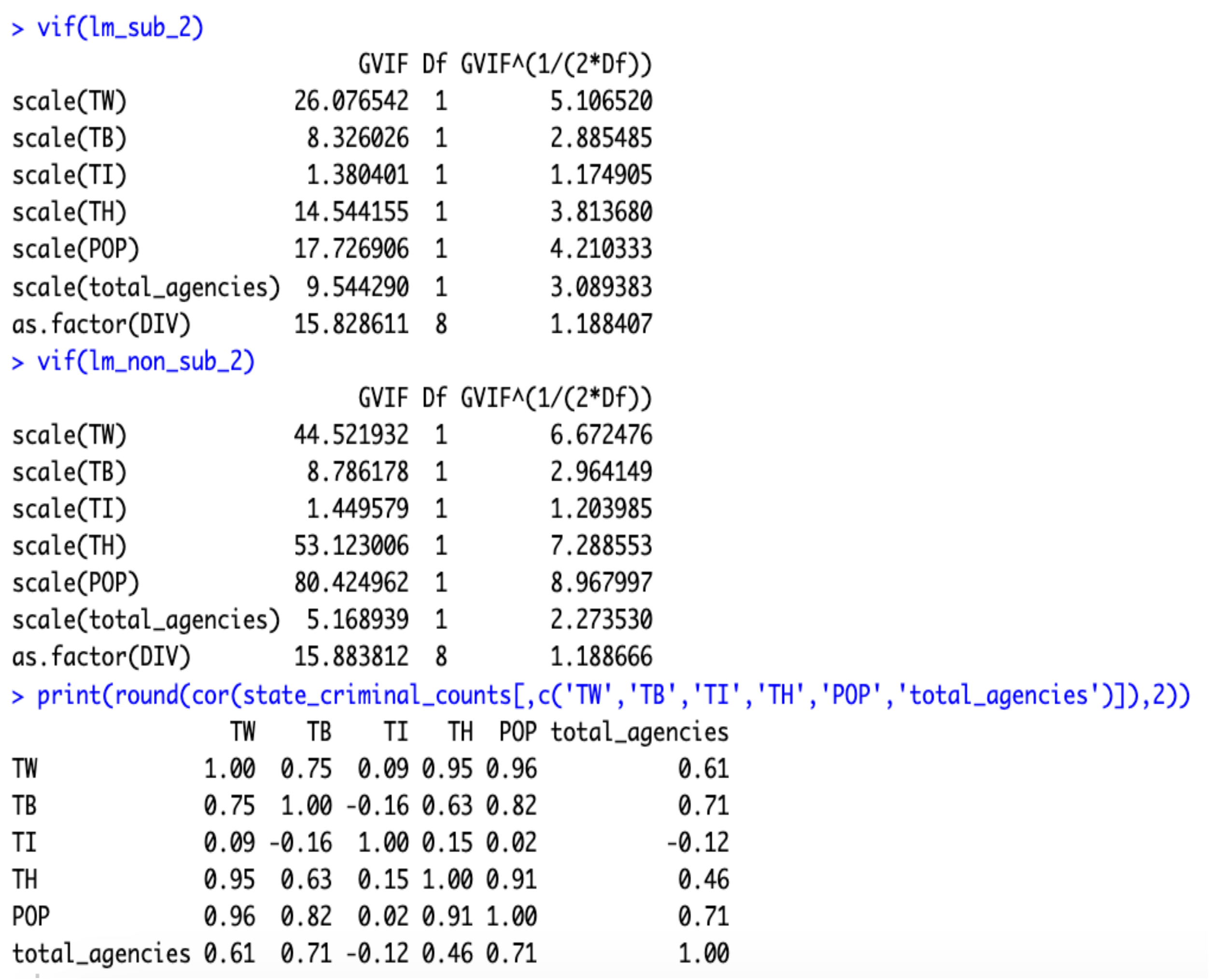
```
Call:
lm(formula = scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(TI) + 
    scale(TH) + scale(POP) + scale(total_agencies) + as.factor(DIV), 
    data = state_criminal_counts_non_suburban)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.63538 -0.23879  0.03695  0.26254  1.58400 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -0.89136    0.31333  -2.845 0.007474 ** 
scale(TW)                             2.63642    0.65353   4.034 0.000294 ***
scale(TB)                             0.58964    0.29032   2.031 0.050134 .  
scale(TI)                             0.03447    0.11792   0.292 0.771850    
scale(TH)                             0.60620    0.71388   0.849 0.401725    
scale(POP)                           -3.62976    0.87837  -4.132 0.000221 ***
scale(total_agencies)                -0.39130    0.22268  -1.757 0.087883 .  
as.factor(DIV)(2) Middle Atlantic     0.69179    0.54756   1.263 0.215038    
as.factor(DIV)(3) East North Central  1.48703    0.47373   3.139 0.003495 ** 
as.factor(DIV)(4) West North Central  1.22768    0.44790   2.741 0.009693 ** 
as.factor(DIV)(5) South Atlantic      1.09097    0.43272   2.521 0.016549 *  
as.factor(DIV)(6) East South Central  1.17538    0.52437   2.242 0.031628 *  
as.factor(DIV)(7) West South Central  1.54512    0.62284   2.481 0.018219 *  
as.factor(DIV)(8) Mountain            0.60010    0.41642   1.441 0.158703    
as.factor(DIV)(9) Pacific             0.45053    0.50109   0.899 0.374932    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6786 on 34 degrees of freedom
Multiple R-squared:  0.6738,    Adjusted R-squared:  0.5395 
F-statistic: 5.017 on 14 and 34 DF,  p-value: 5.988e-05
```

Variance Inflation Factor Analysis On Updated Model:

From the attached image below, we can see that the VIF, even though it is higher than ideal, has decreased drastically by removing some of the predictors that were highly correlated with other more statistically significant predictors.
As we will see in the Hypothesis Testing part of our report, we have tested on the following variables of interest-

- TW- Total White Criminals
- DIV- Region of the state.
- POP- Combined Population under all agencies reported for a state.

From the VIF values below, we have to be cautious about our results for the TW variable due to its extremely high VIF value. Also, even though DIV's VIF value is still high, when corrected for the degrees of freedom (GVIF^(1/(2*Df)) it is close to 1, so there should not be much variation in it when the response changes slightly.



```
> vif(lm_sub_2)
                          GVIF Df GVIF^(1/(2*Df))
scale(TW)            26.076542  1        5.106520
scale(TB)             8.326026  1        2.885485
scale(TI)             1.380401  1        1.174905
scale(TH)            14.544155  1        3.813680
scale(POP)           17.726906  1        4.210333
scale(total_agencies)  9.544290  1        3.089383
as.factor(DIV)       15.828611  8        1.188407
> vif(lm_non_sub_2)
                          GVIF Df GVIF^(1/(2*Df))
scale(TW)            44.521932  1        6.672476
scale(TB)             8.786178  1        2.964149
scale(TI)             1.449579  1        1.203985
scale(TH)            53.123006  1        7.288553
scale(POP)           80.424962  1        8.967997
scale(total_agencies)  5.168939  1        2.273530
as.factor(DIV)       15.883812  8        1.188666
> print(round(cor(state_criminal_counts[,c('TW','TB','TI','TH','POP','total_agencies')]),2))
                 TW    TB    TI   TH  POP total_agencies
TW             1.00  0.75  0.09 0.95 0.96           0.61
TB             0.75  1.00 -0.16 0.63 0.82           0.71
TI             0.09 -0.16  1.00 0.15 0.02          -0.12
TH             0.95  0.63  0.15 1.00 0.91           0.46
POP            0.96  0.82  0.02 0.91 1.00           0.71
total_agencies 0.61  0.71 -0.12 0.46 0.71           1.00
```

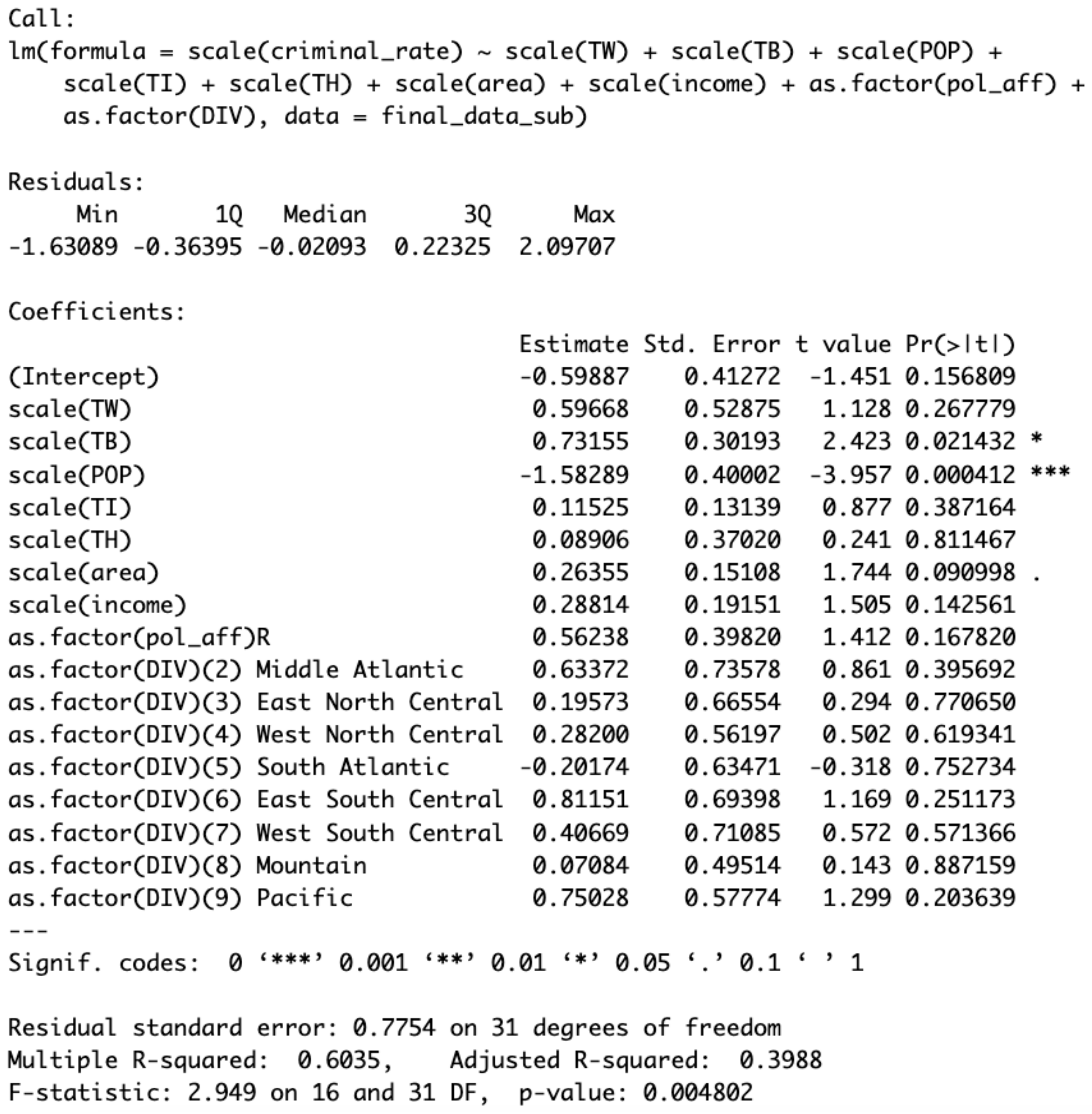
More Data!

As highlighted in Integrating External Features we added some more potentially significant before doing our final hypothesis testing.

Model Summaries After Adding the New Variables-

Suburban Data:

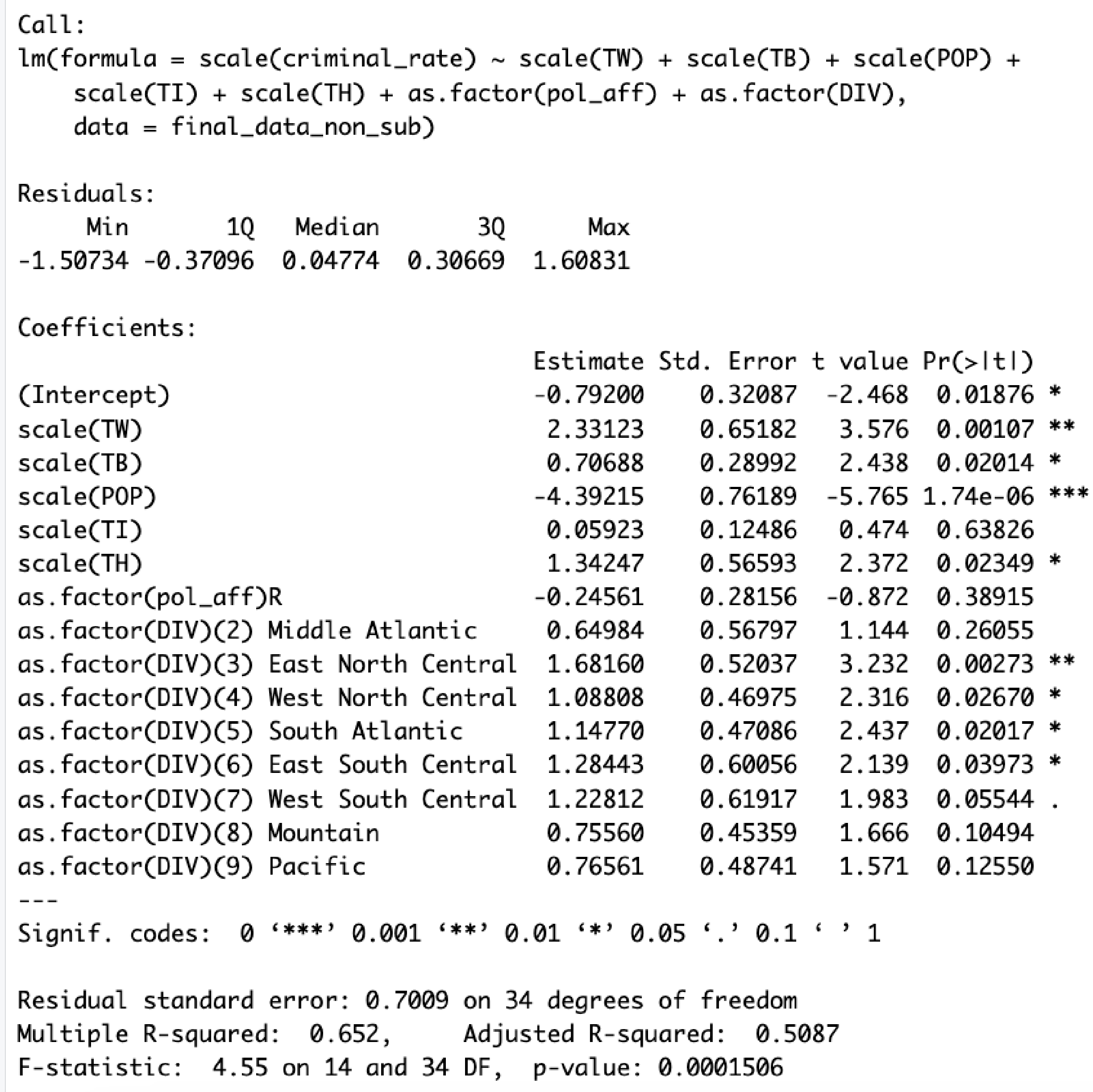
We see a significant rise in the adjusted R-squared value indicating the significance of the predictors added in explaining the variance of the response.



```
Call:
lm(formula = scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + 
    scale(TI) + scale(TH) + scale(area) + scale(income) + as.factor(pol_aff) + 
    as.factor(DIV), data = final_data_sub)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.63089 -0.36395 -0.02093  0.22325  2.09707 

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)                         -0.59887    0.41272  -1.451 0.156809    
scale(TW)                            0.59668    0.52875   1.128 0.267779    
scale(TB)                            0.73155    0.30193   2.423 0.021432 *  
scale(POP)                          -1.58289    0.40002  -3.957 0.000412 ***
scale(TI)                            0.11525    0.13139   0.877 0.387164    
scale(TH)                            0.08906    0.37020   0.241 0.811467    
scale(area)                          0.26355    0.15108   1.744 0.090998 .  
scale(income)                        0.28814    0.19151   1.505 0.142561    
as.factor(pol_aff)R                  0.56238    0.39820   1.412 0.167820    
as.factor(DIV)(2) Middle Atlantic    0.63372    0.73578   0.861 0.395692    
as.factor(DIV)(3) East North Central 0.19573    0.66554   0.294 0.770650    
as.factor(DIV)(4) West North Central 0.28200    0.56197   0.502 0.619341    
as.factor(DIV)(5) South Atlantic    -0.20174    0.63471  -0.318 0.752734    
as.factor(DIV)(6) East South Central 0.81151    0.69398   1.169 0.251173    
as.factor(DIV)(7) West South Central 0.40669    0.71085   0.572 0.571366    
as.factor(DIV)(8) Mountain           0.07084    0.49514   0.143 0.887159    
as.factor(DIV)(9) Pacific            0.75028    0.57774   1.299 0.203639    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7754 on 31 degrees of freedom
Multiple R-squared:  0.6035,	Adjusted R-squared:  0.3988 
F-statistic: 2.949 on 16 and 31 DF,  p-value: 0.004802
```

Non-Suburban Data:

The adjusted R-squared value dropped after adding the above predictors. So, after selective replacement (removing area and income as predictors), we arrived at the following model. Even though political affiliation seems to be a statistically insignificant predictor we are keeping it in the model because it is of interest to our analysis.



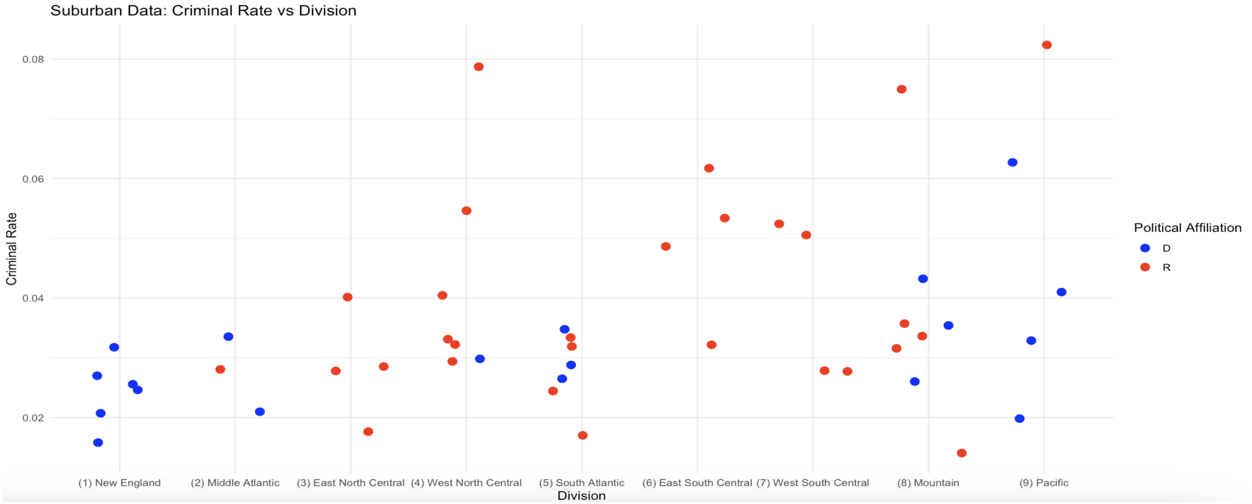
```
Call:
lm(formula = scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + 
    scale(TI) + scale(TH) + as.factor(pol_aff) + as.factor(DIV), 
    data = final_data_non_sub)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.50734 -0.37096  0.04774  0.30669  1.60831 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -0.79200    0.32087  -2.468  0.01876 *  
scale(TW)                             2.33123    0.65182   3.576  0.00107 ** 
scale(TB)                             0.70688    0.28992   2.438  0.02014 *  
scale(POP)                           -4.39215    0.76189  -5.765 1.74e-06 ***
scale(TI)                             0.05923    0.12486   0.474  0.63826    
scale(TH)                             1.34247    0.56593   2.372  0.02349 *  
as.factor(pol_aff)R                  -0.24561    0.28156  -0.872  0.38915    
as.factor(DIV)(2) Middle Atlantic     0.64984    0.56797   1.144  0.26055    
as.factor(DIV)(3) East North Central  1.68160    0.52037   3.232  0.00273 ** 
as.factor(DIV)(4) West North Central  1.08808    0.46975   2.316  0.02670 *  
as.factor(DIV)(5) South Atlantic      1.14770    0.47086   2.437  0.02017 *  
as.factor(DIV)(6) East South Central  1.28443    0.60056   2.139  0.03973 *  
as.factor(DIV)(7) West South Central  1.22812    0.61917   1.983  0.05544 .  
as.factor(DIV)(8) Mountain            0.75560    0.45359   1.666  0.10494    
as.factor(DIV)(9) Pacific             0.76561    0.48741   1.571  0.12550    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7009 on 34 degrees of freedom
Multiple R-squared:  0.652,	Adjusted R-squared:  0.5087 
F-statistic:  4.55 on 14 and 34 DF,  p-value: 0.0001506
```

Correlation Analysis for the added predictors:

Political Affiliation:

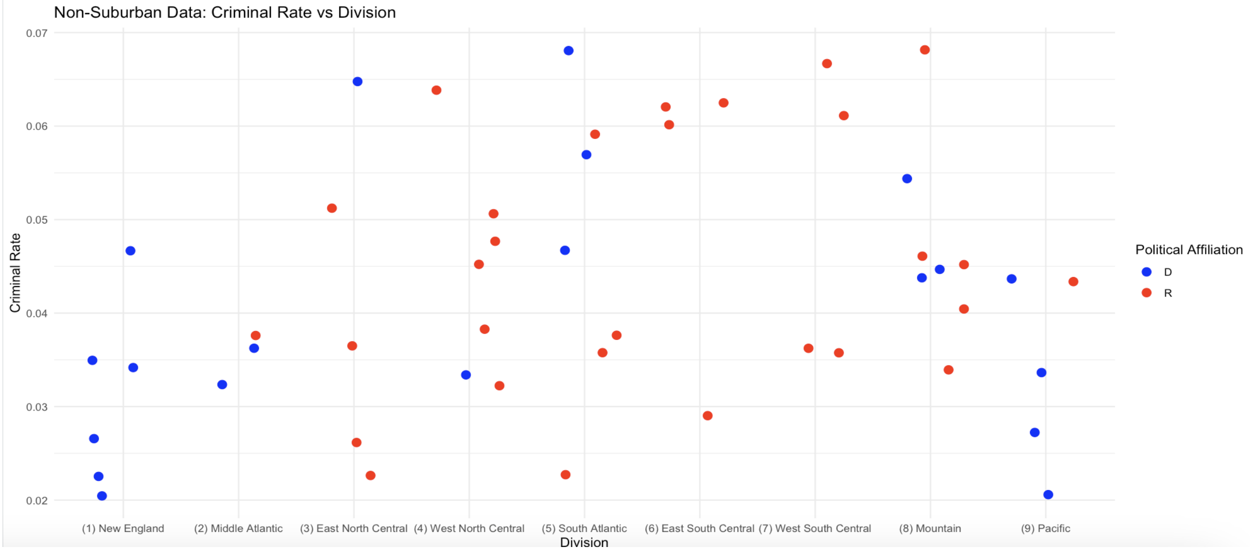
Suburban Data:

From the data below we can see that, for suburban areas of states of all regions, there are higher criminal rates in Republican-leaning states compared to the Democratic-leaning states.



Non-Suburban Data:

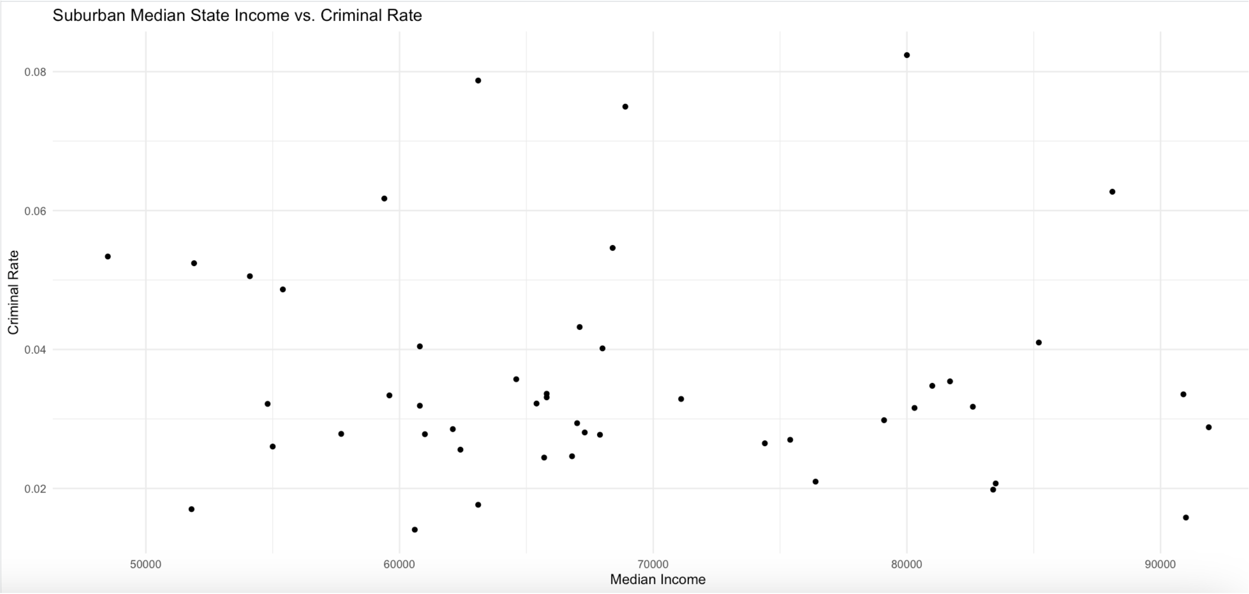
In contrast to the suburban areas of the states, there is no clear correlation between political affiliation and criminal rates in the non-suburban areas. Nevertheless, on the whole, Democratic-dominated Regions seem to have a lower criminal rate than the Republican-dominated ones (with a few exceptions).



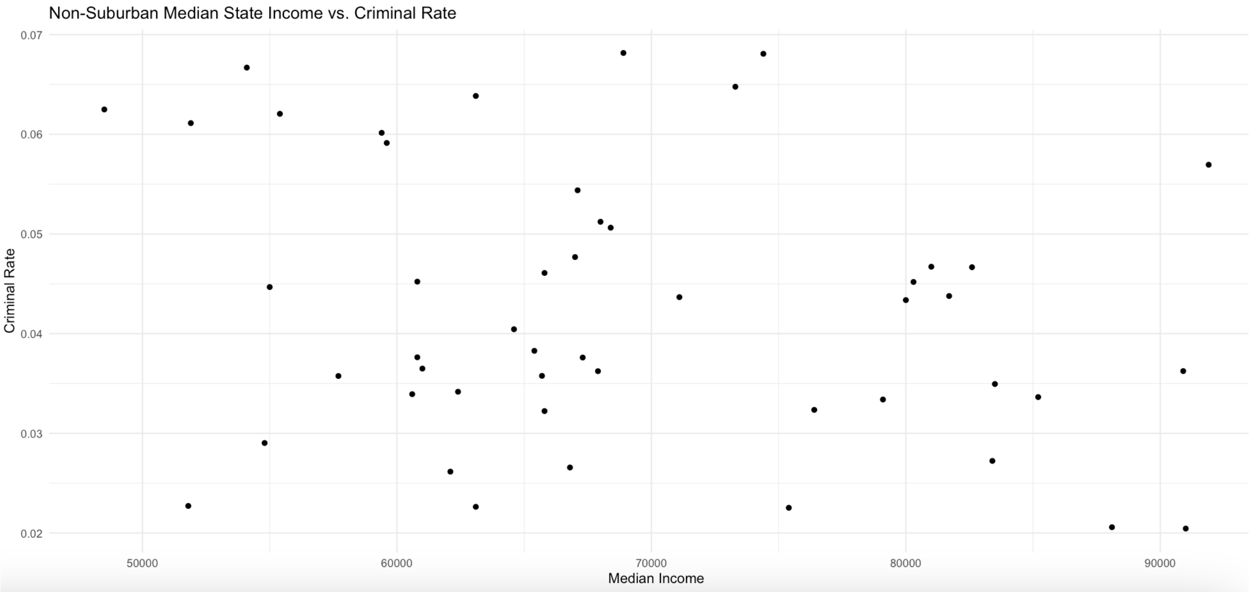
Median Income:

For both the suburban and non-suburban data, there does not appear to be any straightforward relationship between median income and criminal rate.

Suburban Data:



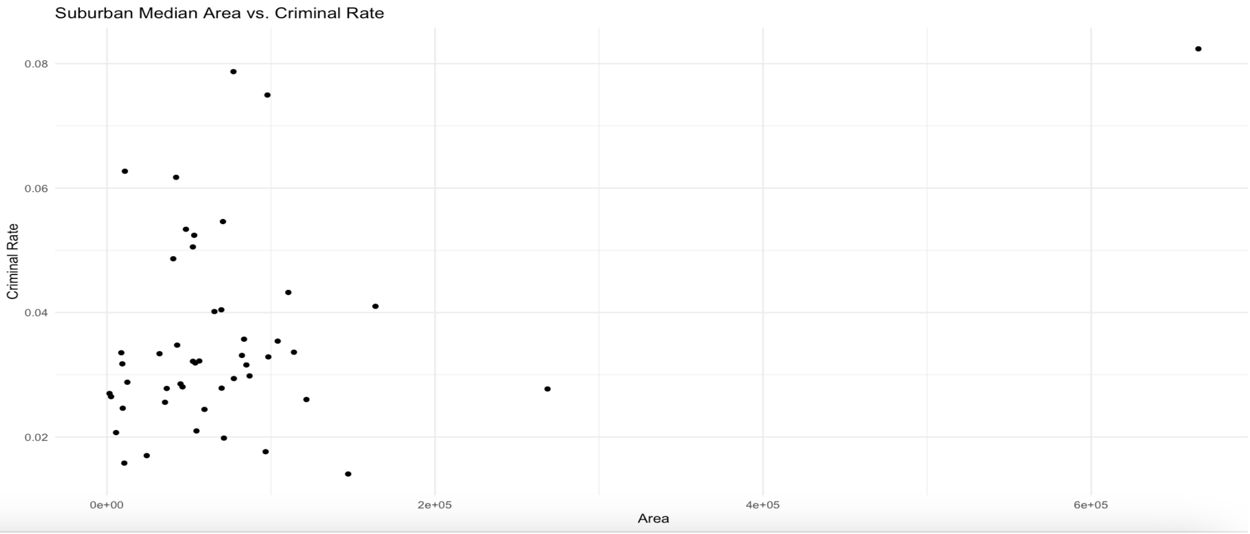
Non-Suburban Data:



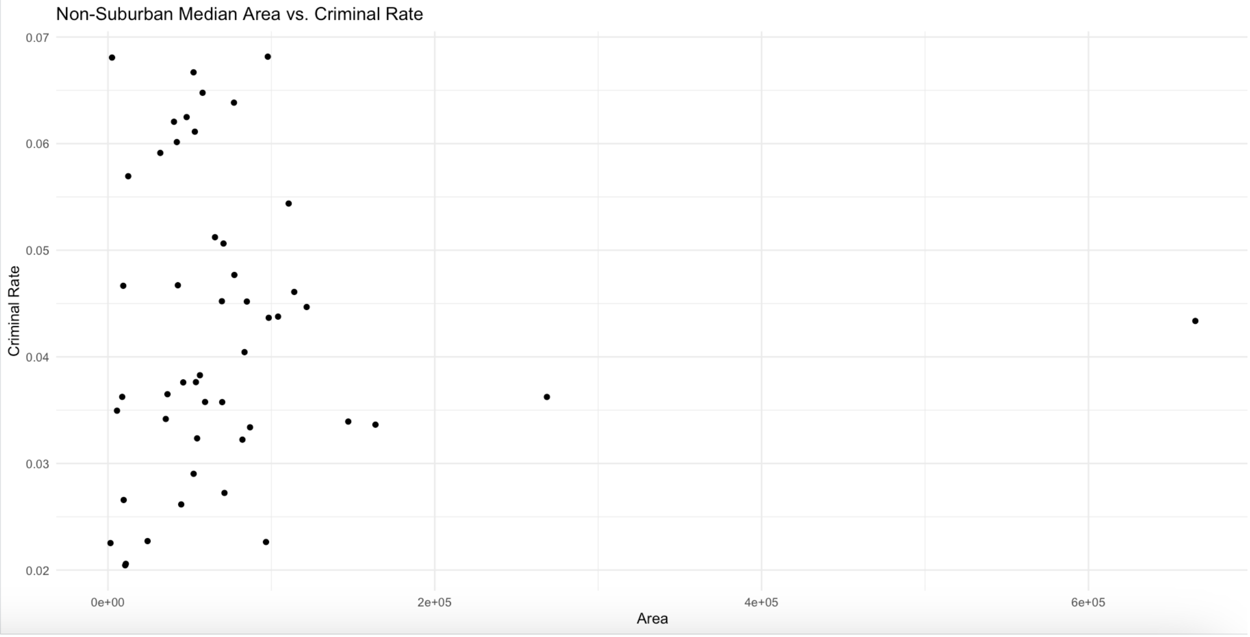
Area:

Again, like median income, there doesn't seem to be any correlation between the Area (in sq. miles) of the state as a whole against the criminal rate for both suburban and non-suburban areas of the state. Nevertheless, from these plots, it becomes clear that, broadly, the suburban areas of the states have a lower criminal rate compared to non-suburban areas.

Suburban Data:



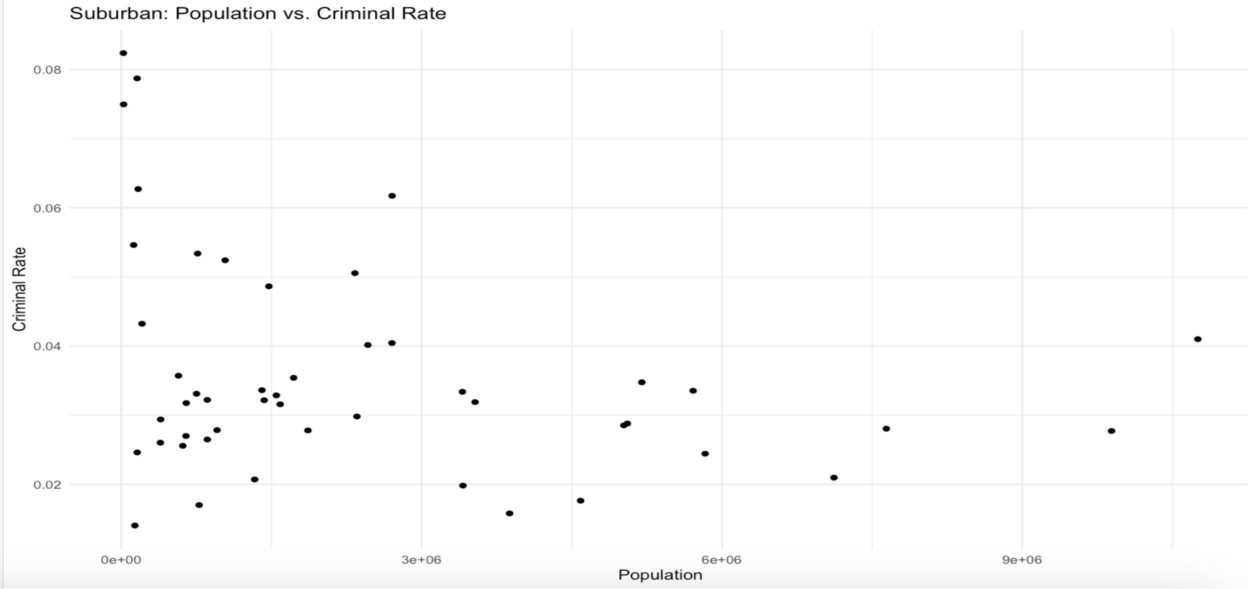
Non-Suburban Data:



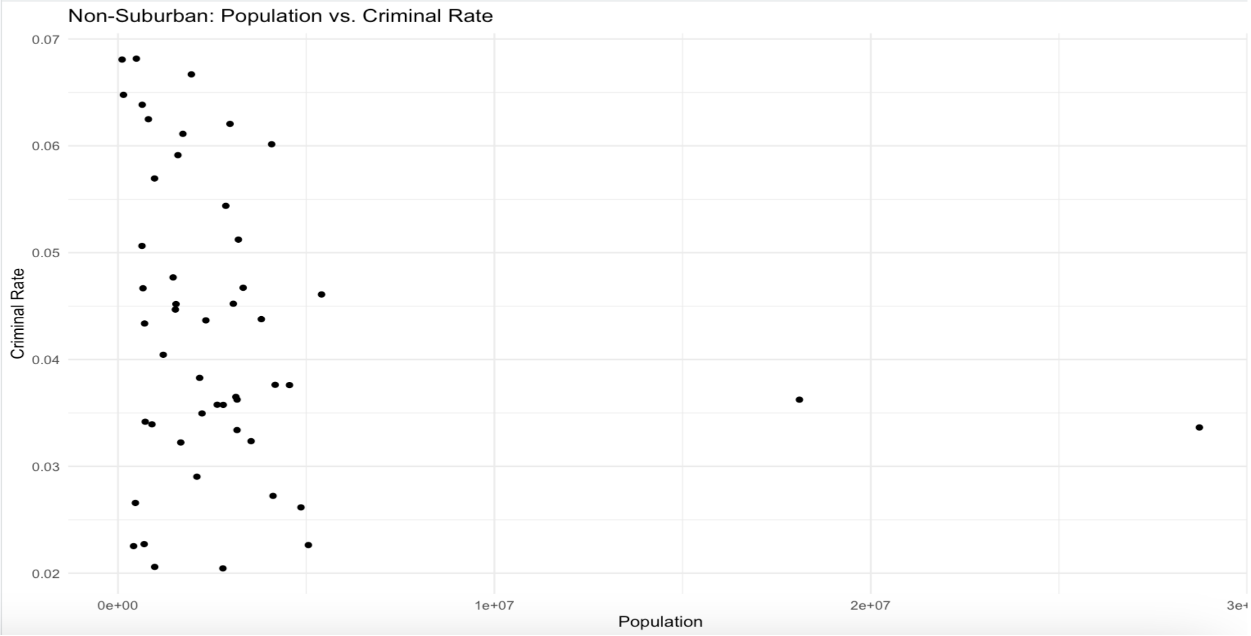
Population:

One would normally assume that having a higher population under an agency would lead to more crimes. But our models predict an inverse relationship. This is clear through the plots below. There can be many reasons for this, which are highlighted in the Discussion section.

Suburban Data:



Non-Suburban Data:



Hypothesis Testing

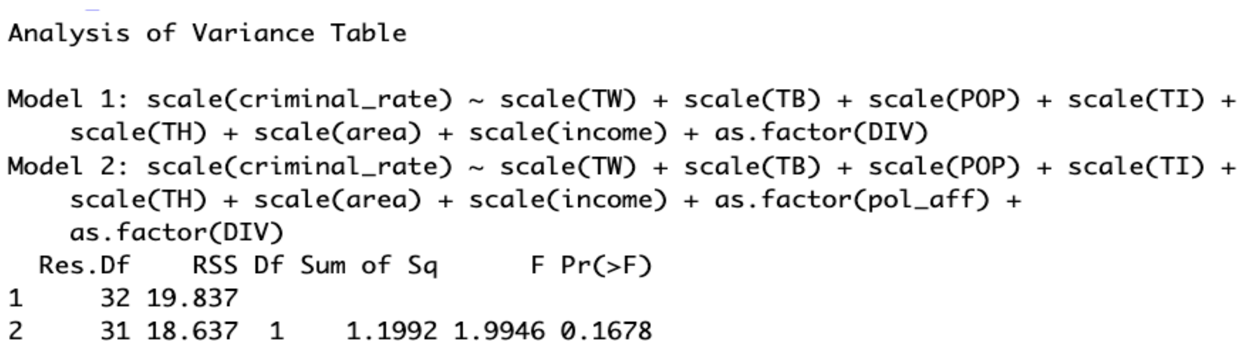
Here we are performing hypothesis tests on some of the statistically/analytically significant predictors. We are assuming the significance level to be 0.05 for all our analyses. We will be using the ANOVA method of hypothesis testing in our analysis.

Political Affiliation:

Assuming B_p represents the coefficient of political affiliation. For our model, being republican leaning has been coded as 1.

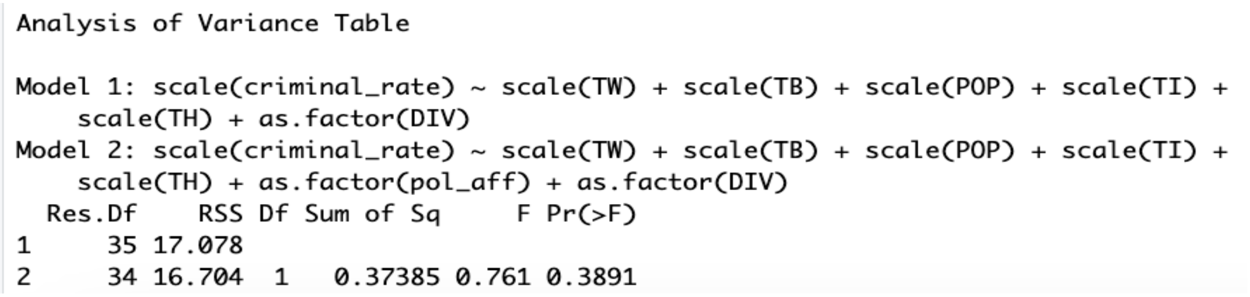
Statement - Political Affiliation affects criminal rates.
Null (H_o): $B_p = 0$
Alternate (H_a): $B_p != 0$

Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + scale(area) + scale(income) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + scale(area) + scale(income) + as.factor(pol_aff) +
    as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     32 19.837
2     31 18.637  1    1.1992 1.9946 0.1678
```

Non-Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + as.factor(pol_aff) + as.factor(DIV)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     35 17.078
2     34 16.704  1   0.37385 0.761 0.3891
```

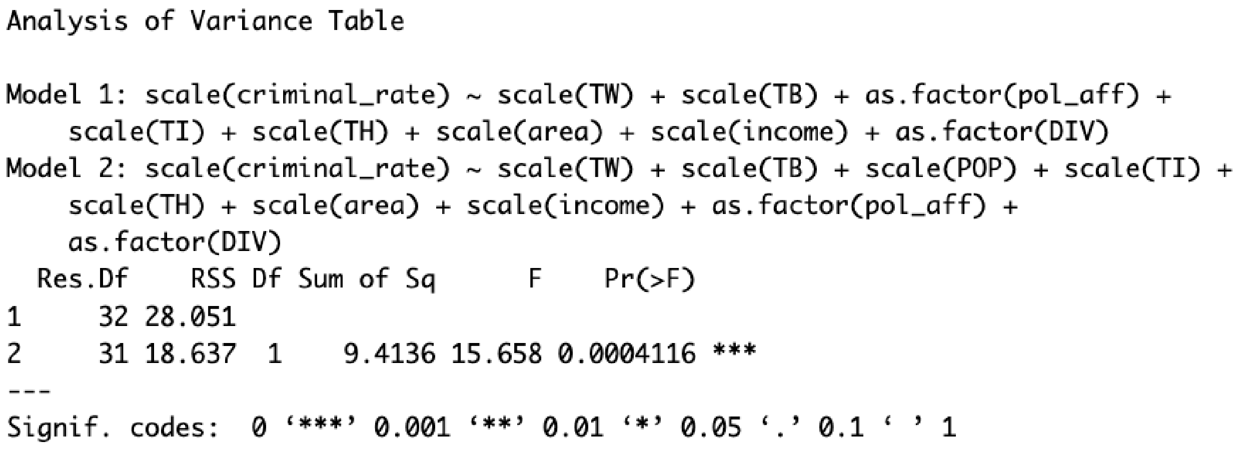
Result: We hypothesized through the plots above that Republican-leaning states have more criminal rates than Democratic-leaning states (especially in the suburban areas of the states) and tested to see if there is any effect. Through the ANOVA tests, we conclude that we have insufficient evidence to conclude that the observed effect is statistically significant.

Population:

Assuming B_p represents the coefficient of political affiliation.

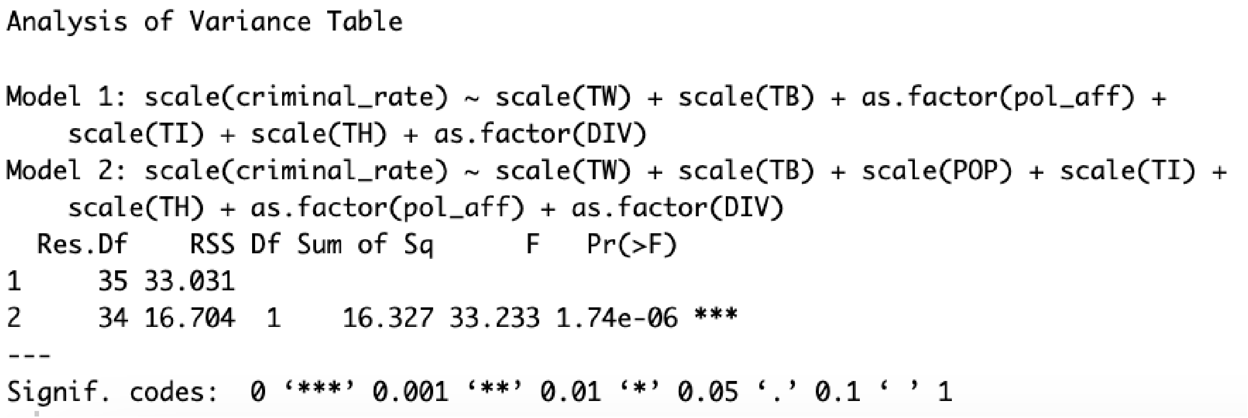
Statement - Population (combined population under all the agency data of the state available to us) has an inverse effect on criminal rates.
Null (H_o): $B_p >= 0$
Alternate (H_a): $B_p < 0$

Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(TW) + scale(TB) + as.factor(pol_aff) +
    scale(TI) + scale(TH) + scale(area) + scale(income) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + scale(area) + scale(income) + as.factor(pol_aff) +
    as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     32 28.051
2     31 18.637  1    9.4136 15.658 0.0004116 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Non-Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(TW) + scale(TB) + as.factor(pol_aff) +
    scale(TI) + scale(TH) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + as.factor(pol_aff) + as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     35 33.031
2     34 16.704  1    16.327 33.233 1.74e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

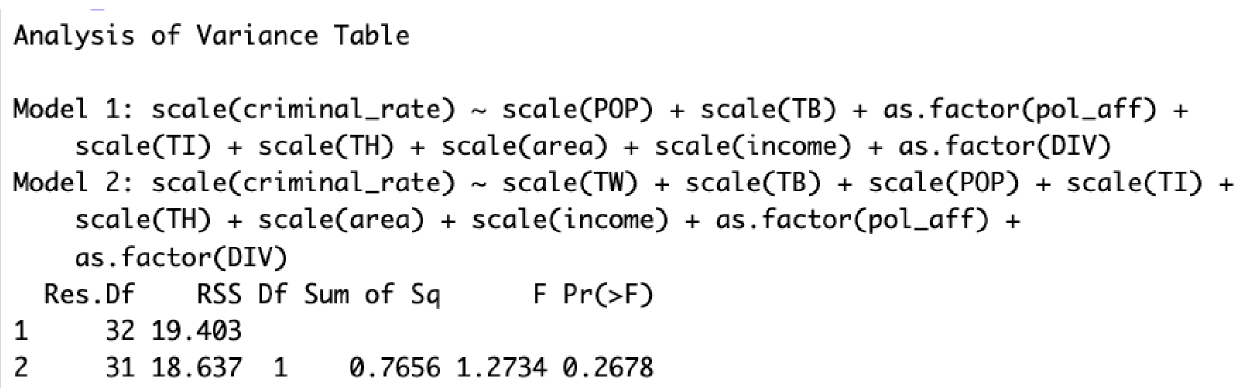
Result: As the estimate of POP lies in the alternate region for both the suburban and non-suburban data, the p-value should be halved to get the one-sided test p-value. Consequently, the final p-value < 0.05 for both cases. Hence, we reject the null hypothesis and safely say that our hypothesis is statistically significant.

Total White Criminals:

Assuming B_w represents the coefficient of political affiliation.

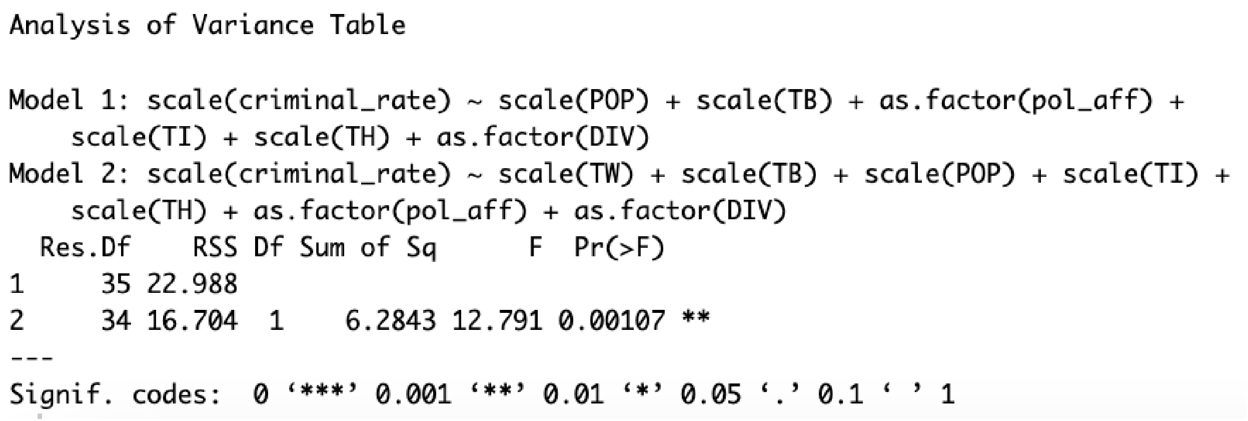
Statement - Having more white criminals is an indicator of having a severe criminal rate.
Null (H_o): $B_w <= 0$
Alternate (H_a): $B_w > 0$

Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(POP) + scale(TB) + as.factor(pol_aff) +
    scale(TI) + scale(TH) + scale(area) + scale(income) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + scale(area) + scale(income) + as.factor(pol_aff) +
    as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     32 19.403
2     31 18.637  1    0.7656 1.2734 0.2678
```

Non-Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(POP) + scale(TB) + as.factor(pol_aff) +
    scale(TI) + scale(TH) + as.factor(DIV)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + as.factor(pol_aff) + as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1     35 22.988
2     34 16.704  1    6.2843 12.791 0.00107 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

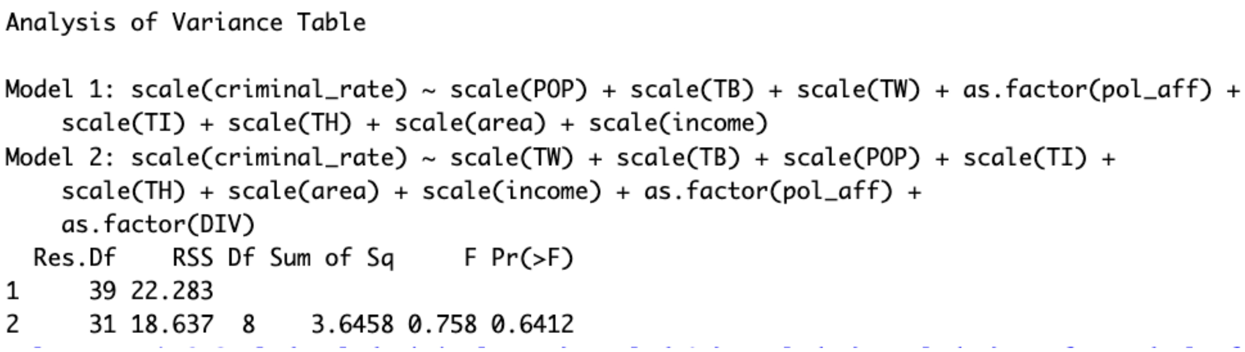
Result: As the estimate of TW lies in the alternate region for both the suburban and non-suburban data, the p-value should be halved to get the one-sided test p-value. From above, we fail to reject the null hypothesis for suburban data, but we can reject the null hypothesis for non-suburban data. Hence, our hypothesis is statistically significant for non-suburban areas of the states.

Division:

Assuming B_d represents the coefficient of political affiliation.

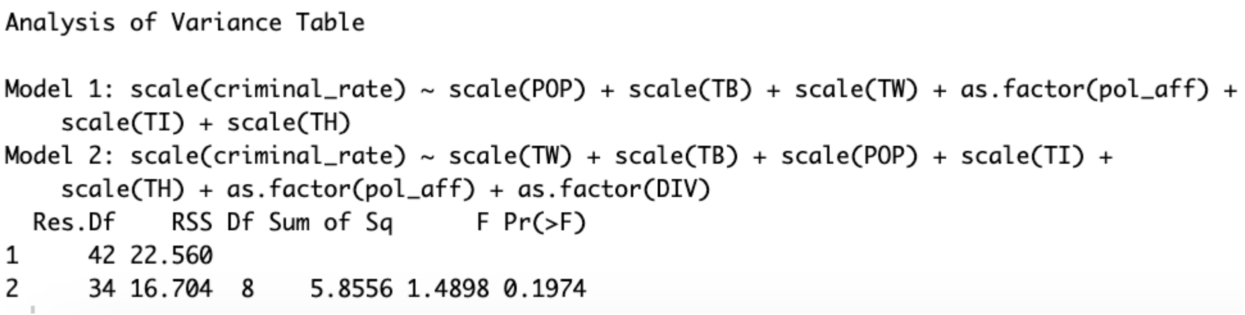
Statement - The region in which a state is in affects the criminal rate of that state.
Null (H_o): $B_d = 0$
Alternate (H_a): B_d != 0

Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(POP) + scale(TB) + scale(TW) + as.factor(pol_aff) +
    scale(TI) + scale(TH) + scale(area) + scale(income)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + scale(area) + scale(income) + as.factor(pol_aff) +
    as.factor(DIV)
  Res.Df    RSS Df Sum of Sq     F Pr(>F)
1     39 22.283
2     31 18.637  8    3.6458 0.758 0.6412
```

Non-Suburban Data:



```
Analysis of Variance Table

Model 1: scale(criminal_rate) ~ scale(POP) + scale(TB) + scale(TW) + as.factor(pol_aff) +
    scale(TI) + scale(TH)
Model 2: scale(criminal_rate) ~ scale(TW) + scale(TB) + scale(POP) + scale(TI) +
    scale(TH) + as.factor(pol_aff) + as.factor(DIV)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     42 22.560
2     34 16.704  8    5.8556 1.4898 0.1974
```

Result: We hypothesized through the plots above that Division has an effect on the overall criminal rate for a state. Through the ANOVA test, we cannot reject the null hypothesis at a significance level of 0.05 and we conclude that the observed effect is statistically insignificant and might just be by chance.

Discussion

Through our analysis, we have discovered many insightful things about the criminal rates and their dependency on some of the available and scraped variables. However, we acknowledge that there are additional variables that could be analytically or explanatorily beneficial for a holistic analysis. Some aspects that we believe can further enhance our descriptive analysis are as follows:

1. We need comprehensive information about the income brackets of states. Merely relying on the median income for a state as a whole might provide a superficial understanding.

2. Detailed information about agencies is missing. Factors such as their budget, number of active agents, and other relevant data could prove insightful for in-depth analysis.

3. The exact areas covered by agencies are not available. Generalizing the area of a state cannot fully describe the behavior of the data.

4. Data regarding repeat offenders, migrating offenders, and mob crimes is absent, which could provide further insights for analysis.

5. Examining the number of crimes committed by juveniles alone versus with adults could be an interesting point to consider.

6. Some agencies had a higher number of criminals than the population they were responsible for. We did not include those agencies in our analysis, but understanding the underlying reasons, such as repeat offenders or migrant criminals, could be insightful.

7. The data was missing rows for certain agencies. For example, states like Illinois had just one agency with complete crime data reported in the required format.

8. A crime-type analysis (violent, non-violent, financial, etc.) is needed for a more in-depth understanding of the subject.

9. Additionally, we would have liked to measure the impact of current events on crime rates. For instance, analyzing how the overturning of Roe v. Wade affected crime rates in states with pre-Roe bans or how the decriminalization of marijuana influenced crime rates since it alters the definition of what constitutes a crime.

Considering these factors would contribute to a more comprehensive and nuanced analysis of the criminal rates and their underlying dynamics.

References

1. United States. Federal Bureau of Investigation. Uniform Crime Reporting Program Data: Arrests by Age, Sex, and Race, Summarized Yearly, United States, 2019. Inter-university Consortium for Political and Social Research [distributor], 2023-09-28. https://doi.org/10.3886/ICPSR38780.v1
2. https://nces.ed.gov/programs/digest/d22/tables/dt22_102.30.asp
3. https://beef2live.com/story-ranking-states-area-89-118259
4. https://www.nytimes.com/elections/2016/results/president
5. "8.3 Who Commits Crime?" Social Problems, University of Minnesota Libraries Publishing edition, 2015. This edition is adapted from a work originally produced in 2010 by a publisher who has requested that it not receive attribution., 25 Mar. 2016. open.lib.umn.edu/socialproblems/chapter/8-3-who-commits-crime/
6. United States. Federal Bureau of Investigation. Uniform Crime Reporting Program Data: Arrests by Age, Sex, and Race, United States, 2018. Inter-university Consortium for Political and Social Research [distributor], 2023-09-27. https://doi.org/10.3886/ICPSR38743.v1
7. Puzzanchera, Charles Arrests of Juveniles in 2018 Reached the Lowest Level in Nearly 4 Decades. Office of Juvenile Justice and Delinquency Prevention.https://ojjdp.ojp.gov/library/publications/arrests-juveniles-2018-reached-lowest-level-nearly-4-decades
8. Wang, Yiwen The Effect of Marijuana Legalization on Juvenile Delinquency. Thesis, Georgetown University. https://repository.library.georgetown.edu/handle/10822/1082815
9. Sheehan BE, Grucza RA, Plunk AD. Association of Racial Disparity of Cannabis Possession Arrests Among Adults and Youths With Statewide Cannabis Decriminalization and Legalization. JAMA Health Forum. 2021;2(10):e213435. doi:10.1001/jamahealthforum.2021.3435 https://jamanetwork.com/journals/jama-health-forum/fullarticle/2785582
10. Sun, D., Feldmeyer, B. Racial Invariance or Asian Advantage: Comparing the Macro-Level Predictors of Violence Across Asian, White, and Black Populations. Race Soc Probl 14, 114–130 (2022). https://doi.org/10.1007/s12552-021-09344-1