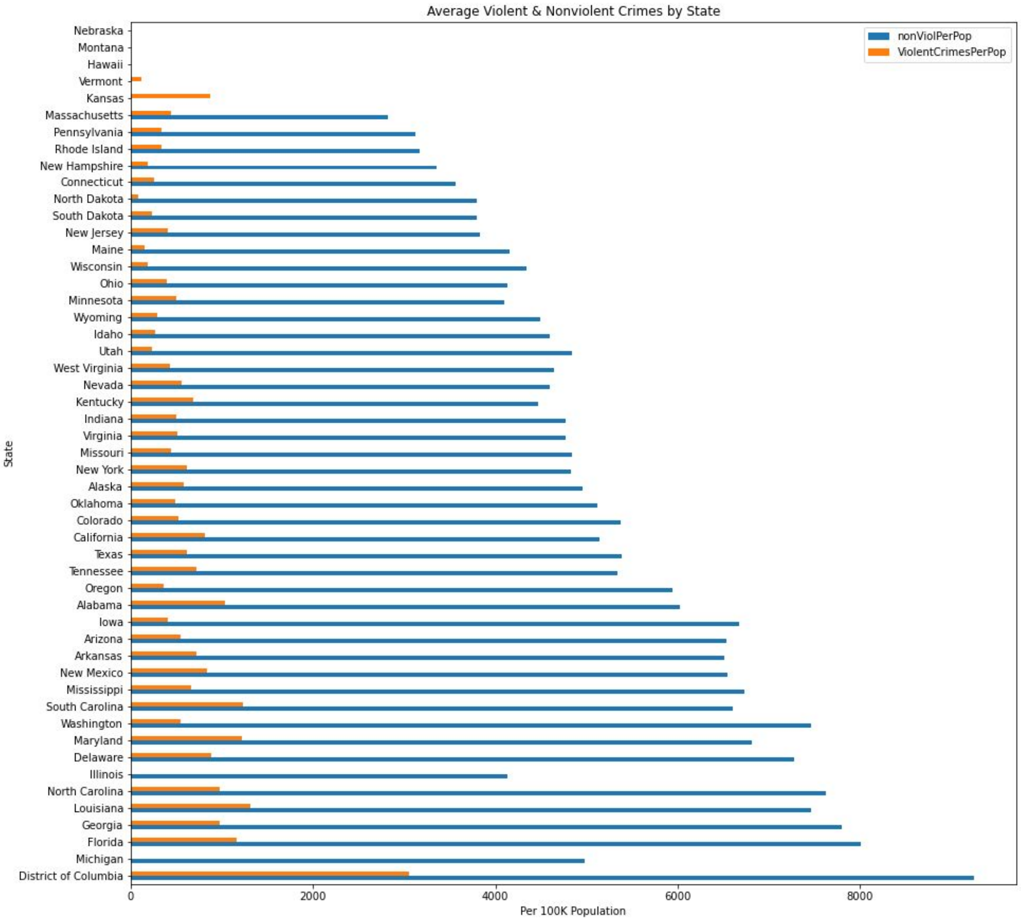
Final Report: US Crime Analysis

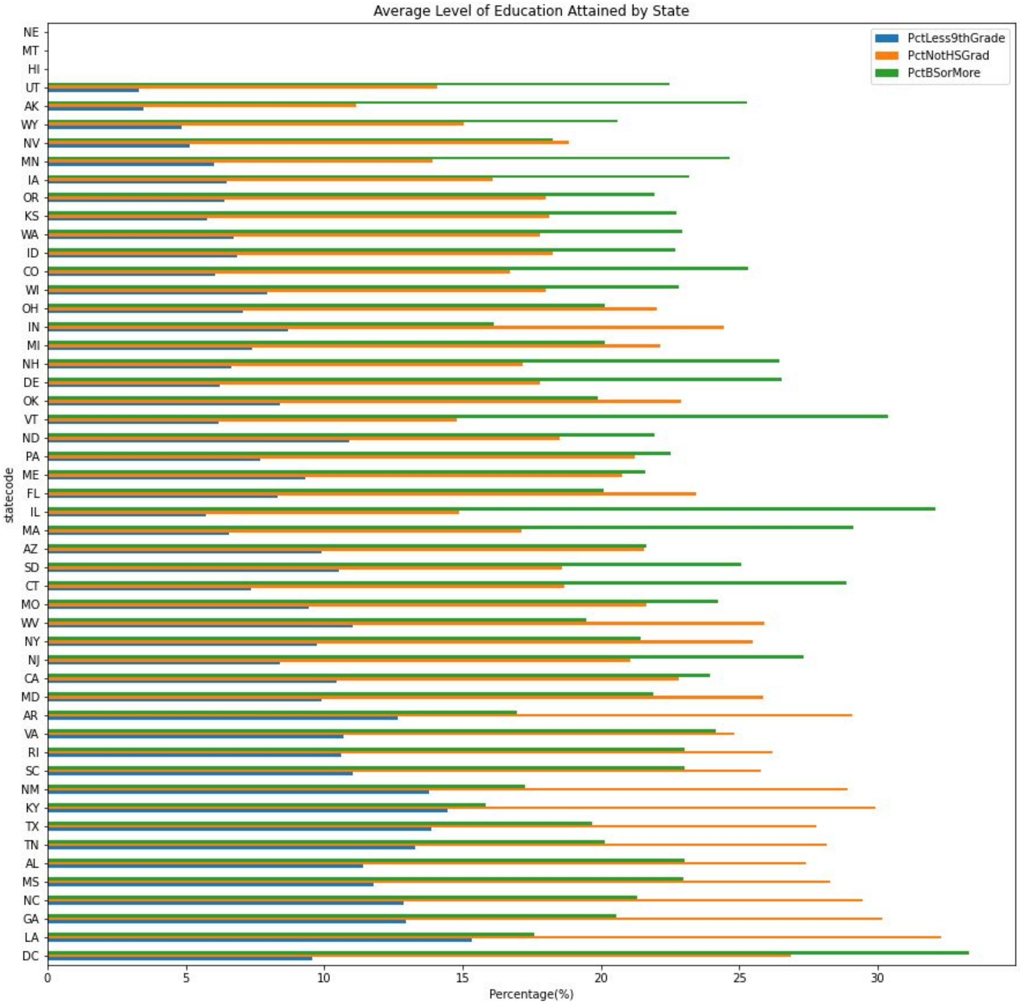
Problem Identification:

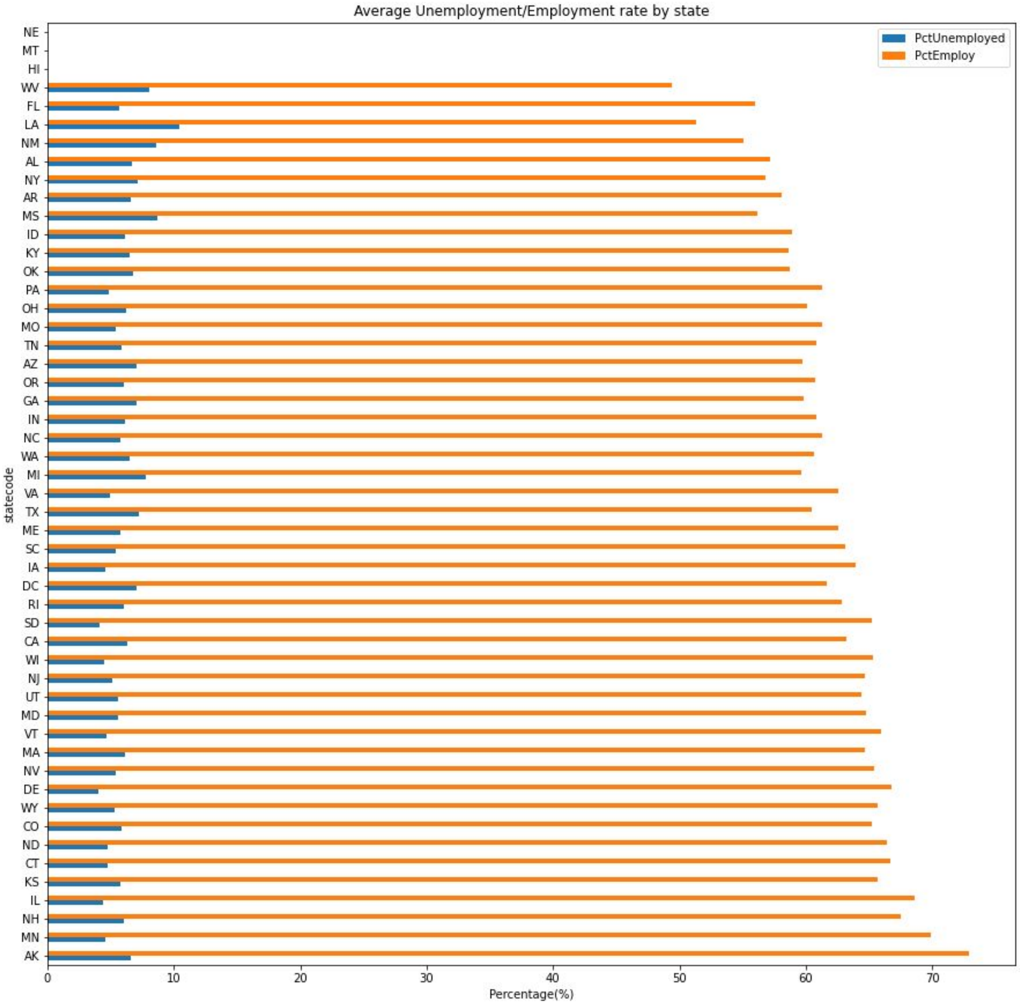
- Hypothesis: There can be many reasons as to why crime occurs in US communities. It can be due to eviction, unemployment, ethnic background, age group, and/or education.
- Context: Crime occurrences can be validated because of a number of reasons. Many people are subject to hardships in the form of eviction and/or unemployment which makes them resort to crime. This study will analyze which variable is strongly correlated with crime occurrences.
- Criteria for Success: Identify which variable is strongly correlated with crime and if education brings down crime rates.
- Scope of solution space: Identify the variables that are strongly correlated with crime occurrences.
- Constraints within solution space: The data file which will be used for the analysis of this study is missing values which can lead to the skepticism of the accuracy of the study.
- Data Sources: UCI Machine Learning Repository and Kaggle.
- By using the data I was able to test Regression models on the dataset to help verify which variable is the best predictor of crime rates in the US.

Data Wrangling:

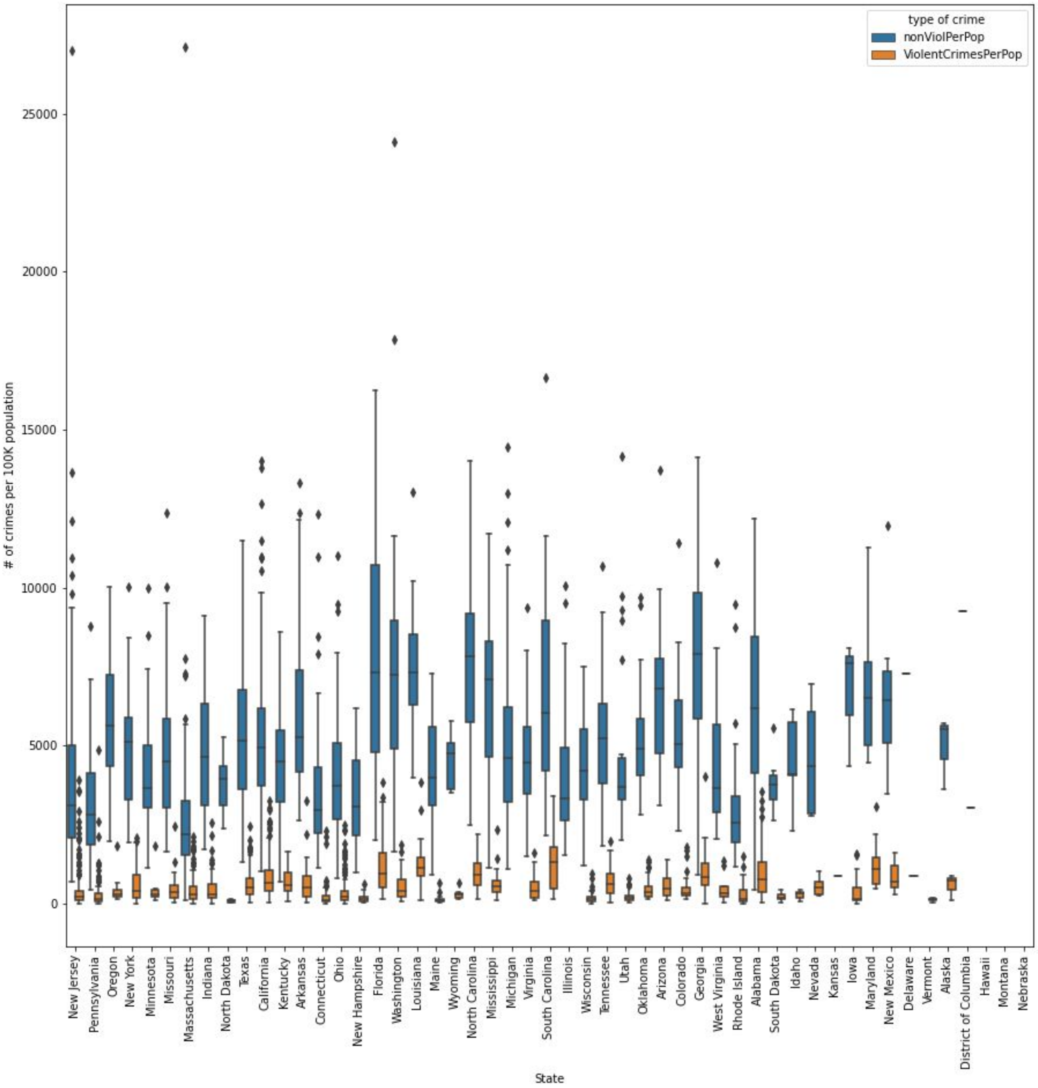
- The dataset is originally from the UCI Machine Learning Repository and was prepared using real data from socio-economic data from 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR [13]. UCI Machine Learning Repository contains a range of free datasets which can be used by anyone trying to hone their Machine Learning skills.
- To get a feel for the data, I started to explore the dataset to see the shape of the dataset, amount of missing values, and the column names. This dataset contains a total number of 147 attributes and 2216 instances. After thoroughly exploring the dataset by seeing the amount of missing values per column, 15 columns were chosen to be appropriate for the study. The variables were statistically observed by checking their mean, min, max, percentile, and std. The variables were missing values from as little as 3 rows to 227 rows. Instead of dropping the entire row, only the missing values were dropped.
- Horizontal Bar Graphs were used to visualize the distribution of the variables being worked with.


Average Violent & Nonviolent Crimes by State
nonViolPerPop
ViolentCrimesPerPop
State
Nebraska
Montana
Hawaii
Vermont
Kansas
Massachusetts
Pennsylvania
Rhode Island
New Hampshire
Connecticut
North Dakota
South Dakota
New Jersey
Maine
Wisconsin
Ohio
Minnesota
Wyoming
Idaho
Utah
West Virginia
Nevada
Kentucky
Indiana
Virginia
Missouri
New York
Alaska
Oklahoma
Colorado
California
Texas
Tennessee
Oregon
Alabama
Iowa
Arizona
Arkansas
New Mexico
Mississippi
South Carolina
Washington
Maryland
Delaware
Illinois
North Carolina
Louisiana
Georgia
Florida
Michigan
District of Columbia
0
2000
4000
6000
8000
Per 100K Population


Average Level of Education Attained by State
PctLess9thGrade
PctNotHSGrad
PctBSorMore
statecode
NE
MT
HI
UT
AK
WY
NV
MN
IA
OR
KS
WA
ID
CO
WI
OH
IN
MI
NH
DE
OK
VT
ND
PA
ME
FL
IL
MA
AZ
SD
CT
MO
WV
NY
NJ
CA
MD
AR
VA
RI
SC
NM
KY
TX
TN
AL
MS
NC
GA
LA
DC
0
5
10
15
20
25
30
Percentage(%)



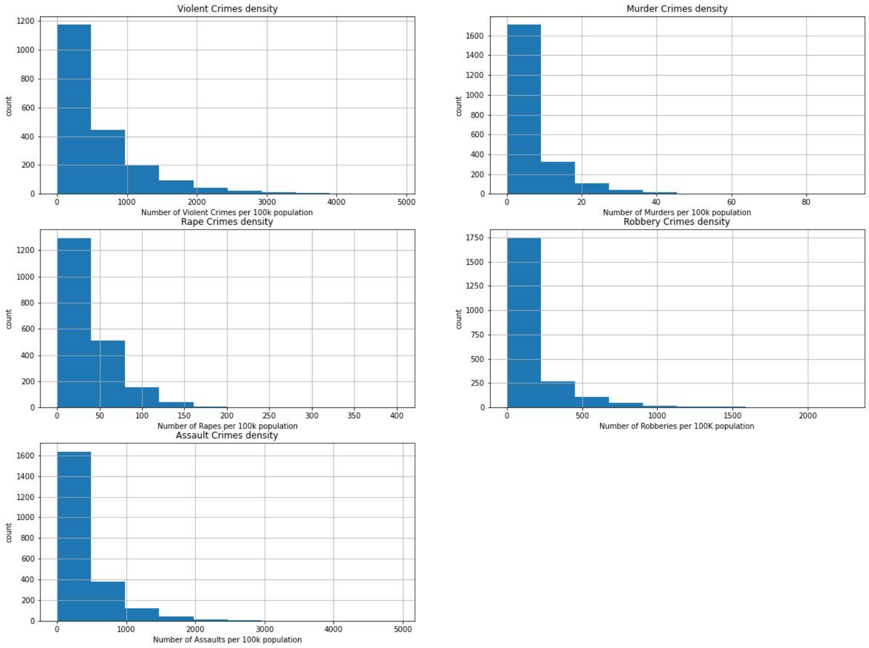
- The dataset's categorical attributes were checked for unique values and any duplicate values under 'communityname'. Communities with the same names were checked which states they are from and were validated that the duplicates belonged to different states. Then the attributes of interest were averaged and grouped by state to see how each attribute differed from state to state. The distributions were also visualized through horizontal bargraphs. Boxplots were also used to visualize the distribution of Violent and NonViolent Crimes for each state.

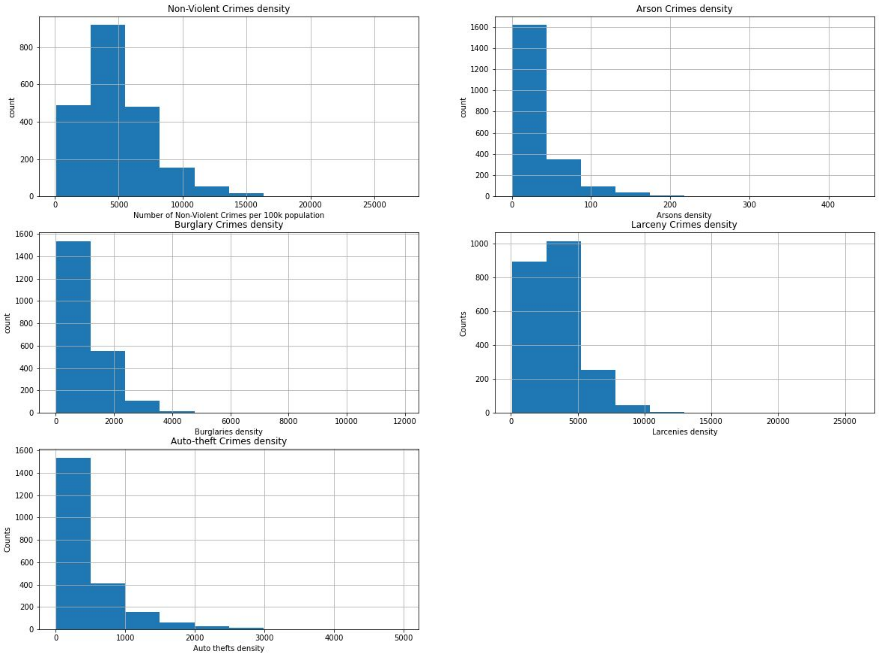


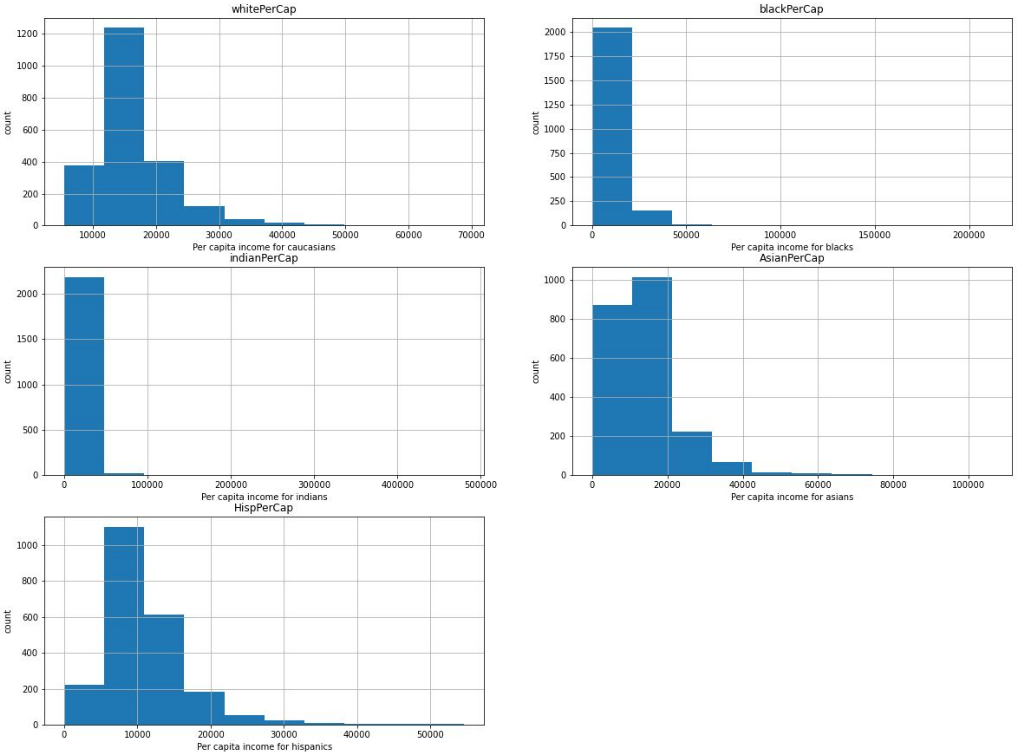
- Scatter plots were created for the target features to analyze the correlation between the socioeconomic variables and crime variables, in order to gain a premature understanding of which attribute is highly correlated with the occurrence of crime. I am really interested in seeing how all the variables are correlated with crime, especially education.

EDA:

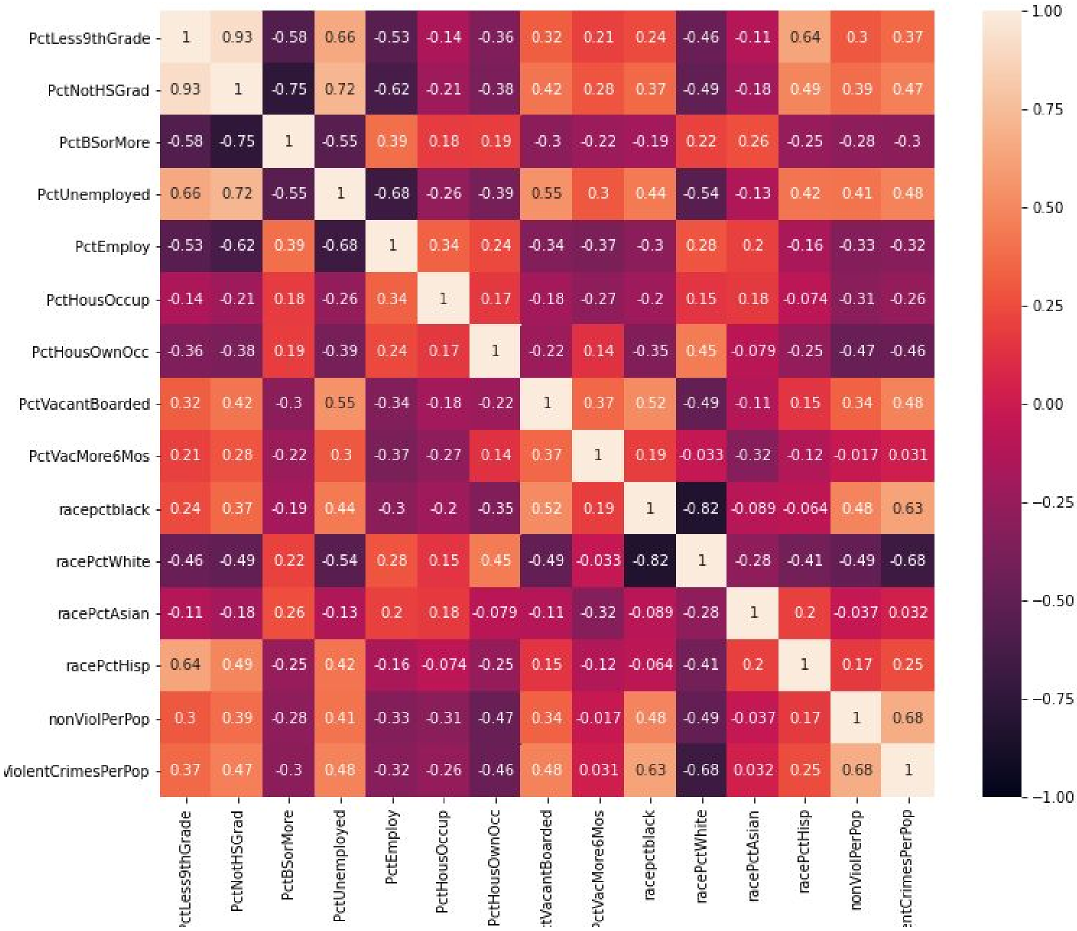
- Step 1: Created bar graphs plotting the distribution for violent crimes, non-violent crimes, and income levels of each race.


Violent Crimes density
count
0 200 400 600 800 1000 1200
0 1000 2000 3000 4000 5000
Number of Violent Crimes per 100k population
Murder Crimes density
count
0 200 400 600 800 1000 1200 1400 1600
0 20 40 60 80
Number of Murders per 100k population
Rape Crimes density
count
0 200 400 600 800 1000 1200
0 50 100 150 200 250 300 350 400
Number of Rapes per 100k population
Robbery Crimes density
count
0 250 500 750 1000 1250 1500 1750
0 500 1000 1500 2000
Number of Robberies per 100K population
Assault Crimes density
count
0 200 400 600 800 1000 1200 1400 1600
0 1000 2000 3000 4000 5000
Number of Assaults per 100k population


Non-Violent Crimes density
count
0 200 400 600 800
0 5000 10000 15000 20000 25000
Number of Non-Violent Crimes per 100k population
Arson Crimes density
count
0 200 400 600 800 1000 1200 1400 1600
0 100 200 300 400
Arsons density
Burglary Crimes density
count
0 200 400 600 800 1000 1200 1400 1600
0 2000 4000 6000 8000 10000 12000
Burglaries density
Larceny Crimes density
Counts
0 200 400 600 800 1000
0 5000 10000 15000 20000 25000
Larcenies density
Auto-theft Crimes density
Counts
0 200 400 600 800 1000 1200 1400 1600
0 1000 2000 3000 4000 5000
Auto thefts density



- Step 2: Regression Plots were used to visualize the correlation between target variables and crime variables.
- Step 3: A heatmap was used to provide a concise way of visualizing the correlation coefficient between all variables.



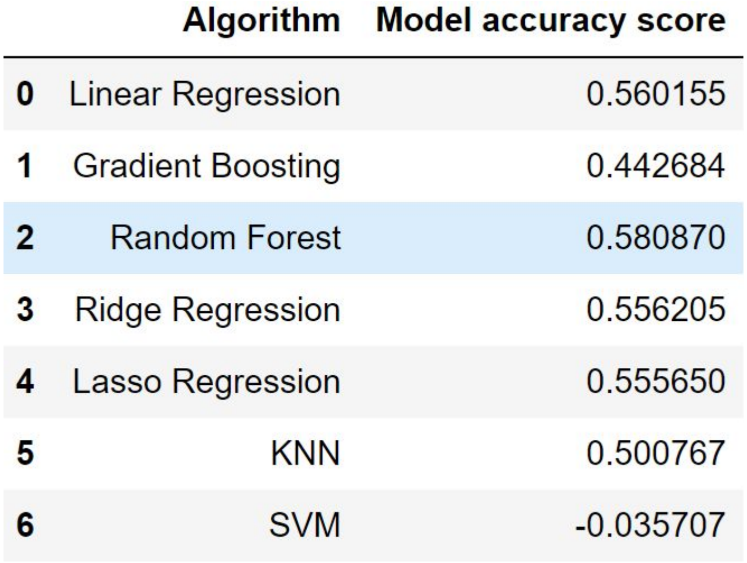
Preprocessing:

- The Preprocessing step focuses on cleaning the dataset to be used for the Modelling portion of the Capstone. Dummy variables were created for the ‘State’ variable but were not used for the modelling portion. All missing values were dropped rather than filled. A few regression models were tried on the dataset to see whether the models were working or not. The cleaned dataset was uploaded into a new csv file for modelling.

Modelling:

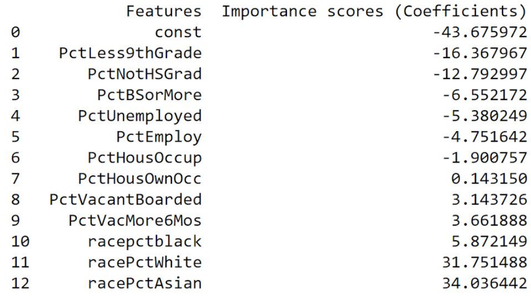
- The Modelling step purely focuses on Machine learning. For this process, different Regression models were used on the dataset. The models consist of: Linear Regression, Gradient Boosting, Random Forest, Lasso Regression, Ridge Regression, K Nearest Neighbors, and SVM. The models were fit and tested. Each model was tested twice, once for Violent Crimes and Non Violent Crimes. The scores were compared to see which model did the best. The models were also fit with a cross-validation score to avoid overfitting and estimate the skill of the model on the new data.

- Violent Crimes: Linear Regression and Random Forest are the top performers with Violent Crimes as the dependent variable.

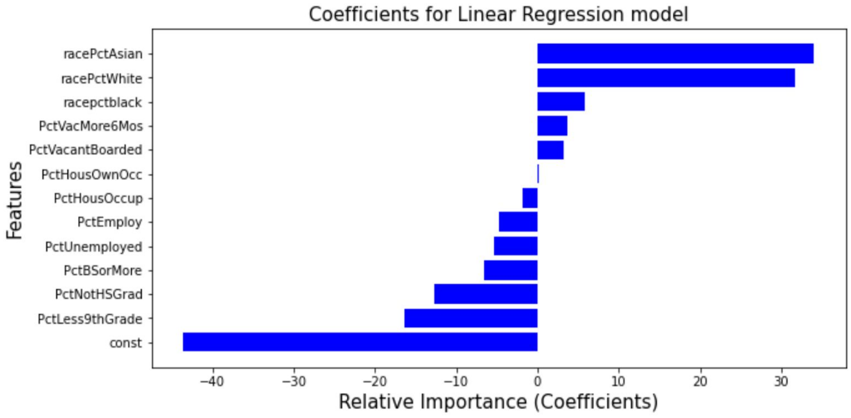


	Algorithm	Model accuracy score
0	Linear Regression	0.560155
1	Gradient Boosting	0.442684
2	Random Forest	0.580870
3	Ridge Regression	0.556205
4	Lasso Regression	0.555650
5	KNN	0.500767
6	SVM	-0.035707

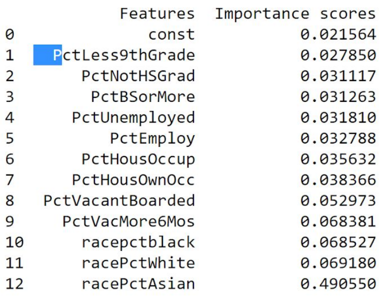
- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Linear Regression is 0.587 and 0.545, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 34.03.



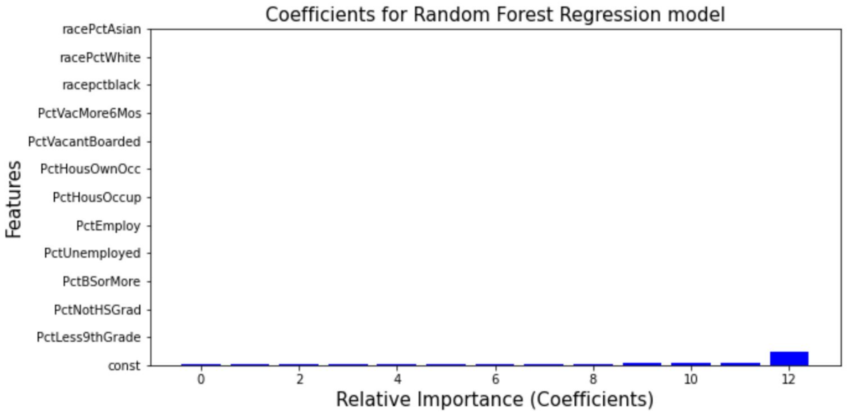
```
             Features  Importance scores (Coefficients)
0               const                        -43.675972
1     PctLess9thGrade                        -16.367967
2        PctNotHSGrad                        -12.792997
3         PctBSorMore                         -6.552172
4       PctUnemployed                         -5.380249
5           PctEmploy                         -4.751642
6        PctHousOccup                         -1.900757
7       PctHousOwnOcc                          0.143150
8    PctVacantBoarded                          3.143726
9      PctVacMore6Mos                          3.661888
10       racepctblack                          5.872149
11       racePctWhite                         31.751488
12       racePctAsian                         34.036442
```



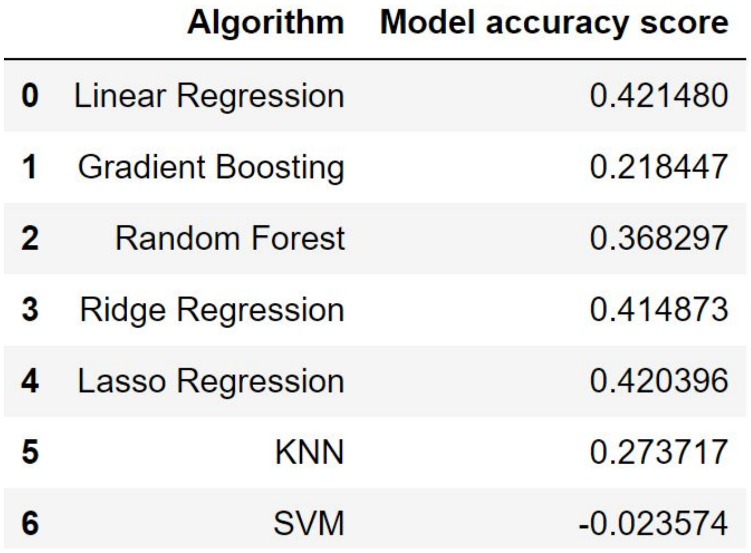
- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Random Forest Regression is 0.582 and 0.539, respectively.The feature importance is formatted as a bar graph below. The feature importance, again shows that racePctAsian is the best predictor with an importance score of 0.49.



```
              Features  Importance scores
0                const           0.021564
1     PctLess9thGrade           0.027850
2        PctNotHSGrad           0.031117
3         PctBSorMore           0.031263
4       PctUnemployed           0.031810
5           PctEmploy           0.032788
6        PctHousOccup           0.035632
7       PctHousOwnOcc           0.038366
8    PctVacantBoarded           0.052973
9     PctVacMore6Mos            0.068381
10       racepctblack           0.068527
11       racePctWhite           0.069180
12       racePctAsian           0.490550
```

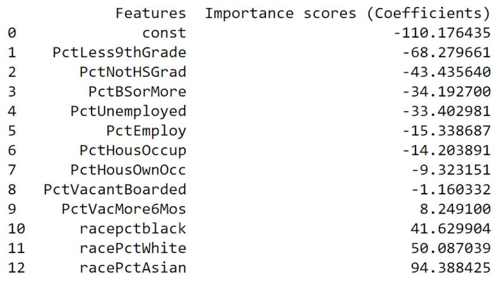


- Non Violent Crimes: Linear Regression and Lasso Regression are the top performers with Non-Violent Crimes as the dependent variable.

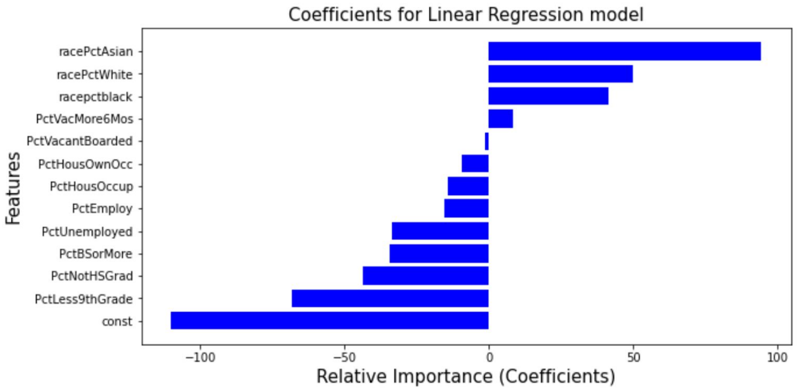


	Algorithm	Model accuracy score
0	Linear Regression	0.421480
1	Gradient Boosting	0.218447
2	Random Forest	0.368297
3	Ridge Regression	0.414873
4	Lasso Regression	0.420396
5	KNN	0.273717
6	SVM	-0.023574

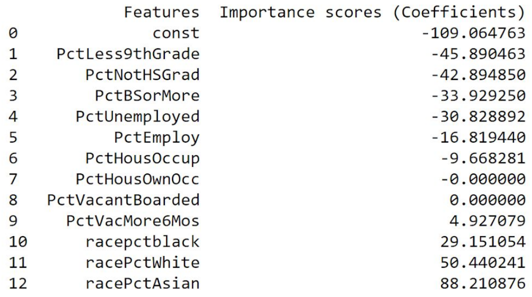
- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Linear Regression is 0.399 and 0.417, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 94.39.



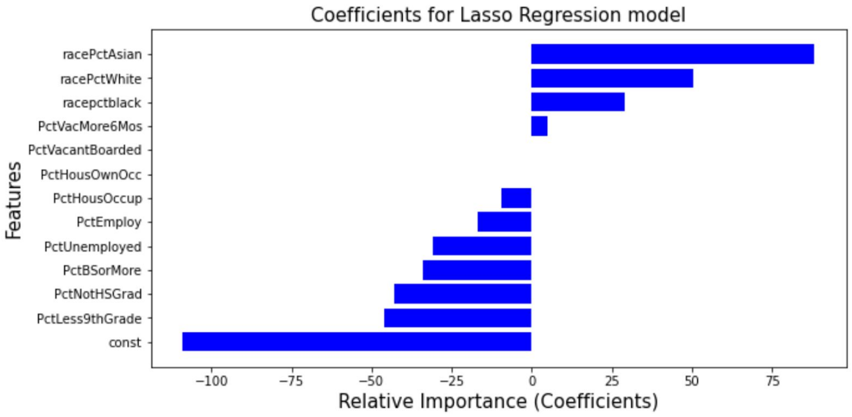
```
              Features  Importance scores (Coefficients)
0                const                        -110.176435
1     PctLess9thGrade                          -68.279661
2        PctNotHSGrad                          -43.435640
3         PctBSorMore                          -34.192700
4       PctUnemployed                          -33.402981
5           PctEmploy                          -15.338687
6        PctHousOccup                          -14.203891
7       PctHousOwnOcc                           -9.323151
8    PctVacantBoarded                           -1.160332
9      PctVacMore6Mos                            8.249100
10       racepctblack                           41.629904
11       racePctWhite                           50.087039
12       racePctAsian                           94.388425
```



- After being fit with optimal hyperparameters, the cross validation score and r^2 score for Lasso Regression is 0.402 and 0.416, respectively. The feature importance is formatted as a horizontal bar graph below. The feature importance shows that racePctAsian is the best predictor with an importance score of 88.21.



```
              Features  Importance scores (Coefficients)
0                const                        -109.064763
1     PctLess9thGrade                          -45.890463
2        PctNotHSGrad                          -42.894850
3         PctBSorMore                          -33.929250
4       PctUnemployed                          -30.828892
5           PctEmploy                          -16.819440
6        PctHousOccup                           -9.668281
7       PctHousOwnOcc                           -0.000000
8    PctVacantBoarded                            0.000000
9      PctVacMore6Mos                            4.927079
10       racepctblack                           29.151054
11       racePctWhite                           50.440241
12       racePctAsian                           88.210876
```



Conclusion:

- After fitting the best performing models with optimal hyperparameters and by calculating the feature importance of each model, it seems that the top three predictors are racePctAsian, racePctWhite, and racepctblack.
- The percentage of Asian and White population could be the best predictor of crime rates since when it comes to income distribution, they have higher incomes than other races. This could indicate that non-violent crimes (burglaries, auto thefts, larcenies, etc.) occur at communities/neighborhoods which are predominantly white or asian, since they are more affluent.
- Such results should not be correlated with the current consensus, since this data is from the US Census of 1990. A better approach would be to update the data every year for more accurate results.