Northeastern University
College *of* Professional Studies

DATA MINING APPLICATIONS

ALY 6040, CRN 80440

PROFESSOR JUSTIN GROSZ

MODULE 6 ASSIGNMENT

FINAL PROJECT

SUBMITTED BY:

Vamshi Krishnamurthy

Richa Umesh Rambhia

Jyothsna Akula

Sai Madhav Avaku

Aishwarya Jangam

Abstract

Gun control activists describe the restricting availability of guns making it more difficult for criminals to acquire and use them to breach the law. This makes the illegal activity more complex, leading to a dissuasion effect, in which the stricter gun prohibitions reduce crime. With 2,40,593 prisoners arrested by state or federal authorities in 1975, the United States set a new high. During the next 34 years, the United States set imposed new records in every category. In the United States today, there are about 1,500,000 people incarcerated. The United States houses hold more than a quarter of the world's total jail population.

The crimes and incarceration in the United States contains data on crimes that are committed and the prisoner count in every 50 states, covering from 2001-2016 where the most interest crime types in the data are violent crime total, murder manslaughter, and aggravated assaults. The important and significant features would help in determining the crime estimated within the state and help predict the whether or not the state changed reporting system had an affect with comparison to the previous years in United States based on the classification and prediction of the prisoner count in each state.

The data would be analyzed using exploratory data analysis and data visualization to gain further insights about the dataset and understand the relationship between the various independent variables that can be considered in the implementation of the machine learning model. This analysis would be helpful in understanding the crimes that are estimated and the prisoner count in each state based on the various crimes that are committed and to recommend the alertness and strictness in those particular crimes that are increasing the prisoner count.

Table of Contents

Introduction

The United States had set a record with 240,593 prisoners incarcerated by state and federal agencies in 1975 and in the next 34 years it achieved new record totals each of next 34 years where today there are over 1,500,000 prisoners in the United States. The crimes and incarceration in the United States contains information about each of the crimes taken place within the states in US from 2001 to 2016. This information will help in analyzing and understanding the relationship between the crimes and imprisonment which would help researchers to determine if the significant investment into imprisonment improves public safety.

The objective here is to analyze the relationship between crime incarceration rates and crime rates in order to analyze and answer the question of whether or not it improved public safety. The dataset would be analyzed, and implementation of machine learning models would help in the prediction of the prisoner count in each state. The dataset thus consists of data on crimes committed and prisoner count in every 50 state and the interested crime types to analyze are violent crime total, murder manslaughter, and aggravated assaults. The important and significant features would help in determining the crime estimated within the state and help predict the prisoner count in each state based on the various crimes that are committed for each state population.

In order to understand and analyze the data, exploratory data analysis along with data visualization is performed which gives a clear understanding of the data and the relationship between the different parameters of the crime and incarceration data which will help in determining the important features to be considered for the implementation and training of the model. The relationship between each of the variables would help in avoiding the multicollinearity for which we can remove the features to avoid overfitting of the model.

Exploratory Data Analysis

The exploratory data analysis is performed on the crime and incarceration data of United States for 50 states from year 2001 to 2016 which contains **17 data field values** and **816 rows of data** having categorical, numerical, and boolean type data. The different parameters of the dataset are state, year, prisoner count, crimes estimated, violent crime total, rape legacy, and other various types of crime data.

```
Column Names:
 Index(['jurisdiction', 'includes_jails', 'year', 'prisoner_count',
       'crime_reporting_change', 'crimes_estimated', 'state_population',
       'violent_crime_total', 'murder_manslaughter', 'rape_legacy',
       'rape_revised', 'robbery', 'agg_assault', 'property_crime_total',
       'burglary', 'larceny', 'vehicle_theft'],
      dtype='object')
```

Figure 1. Field values of the dataset

	year	prisoner_count	state_population	violent_crime_total	murder_manslaughter	rape_legacy	rape_revised	robbery	agg_assault	property_crime_total	burglary	larceny	vehicle_theft
count	816.0	816.0	799.0	799.0	799.0	749.0	199.0	799.0	799.0	799.0	799.0	799.0	799.0
mean	2008.5	28606.0	6072322.2	26228.5	313.7	1788.3	2406.2	7696.8	16256.3	187800.6	40870.4	127912.3	19017.9
std	4.6	39556.9	6725499.8	33866.8	386.0	1865.4	2550.5	11107.5	20849.5	213850.3	47829.9	139434.6	30780.4
min	2001.0	1088.0	493754.0	496.0	5.0	99.0	110.0	43.0	270.0	8806.0	1689.0	6660.0	178.0
25%	2004.8	5698.0	1790025.5	5213.0	48.5	571.0	780.0	1106.0	3529.0	47497.5	9406.0	32765.5	4191.0
50%	2008.5	16915.0	4314113.0	15744.0	179.0	1238.0	1723.0	3933.0	10083.0	132773.0	27698.0	95079.0	10583.0
75%	2012.2	30920.5	6808844.5	31843.0	429.0	2092.0	2680.0	8702.0	20308.0	225957.5	47941.0	155688.0	20872.5
max	2016.0	216915.0	39296476.0	212867.0	2503.0	10198.0	13702.0	71142.0	136087.0	1227194.0	250521.0	731486.0	257543.0

Figure 2. Statistical Analysis

As observed in the statistical analysis, the average count of the various types of crime in the dataset are computed which is helpful in understanding the total count of the particular crimes in the United States. If we consider the earlier mentioned three interested features, violent crime total, murder manslaughter, and aggravated assaults, the average count is 26228.5, 313.7, and 16256.3 respectively. Also, the **prisoner count** average is **28606** considering all the 50 states mentioned and thus this statistics would further help in the analysis of the data.

Data Profiling

The data profiling report gives a detailed overview of the dataset that we are trying to analyze. The dataset statistics shown in the below figure helps understand the different statistics of the data with respect to missing data, duplicate data values, variable types, and number of observations.

Dataset statistics

Number of variables	17
Number of observations	816
Missing cells	871
Missing cells (%)	6.3%
Duplicate rows	0
Duplicate rows (%)	0.0%
Total size in memory	187.0 KiB
Average record size in memory	234.7 B

Variable types

Categorical	1
Boolean	3
Numeric	13

Figure 3. Dataset Statistics

Here, we can observe that there are **871 missing values** which is **6.3% missing cells** in the dataset having **no duplicate rows**, which has 13 numerical data, 1 categorical data, and 3 boolean data values.

Data Cleaning

The data cleaning is an important step to be performed in the analysis process and from the above dataset statistics we observe that there are **871 missing values** which we need to consider for cleaning. In order to determine which field values consists of the missing values, the count of missing values for each column is displayed. Apart from this, the data profiling report generates the missing values plot which gives a better understanding of the missing data.

It is observed that the rape revised column has maximum of null values of the entire dataset and thus the column can be dropped out as it is not an important feature that needs to be considered while the analysis of the data. Apart from that, the remaining columns have **17 null values** which when explored was observed that the same row values are missing for all those field values. Thus, the following methods are performed to clean the data.

jurisdiction	0
includes_jails	0
year	0
prisoner_count	0
crime_reporting_change	17
crimes_estimated	17
state_population	17
violent_crime_total	17
murder_manslaughter	17
rape_legacy	67
rape_revised	617
robbery	17
agg_assault	17
property_crime_total	17
burglary	17
larceny	17
vehicle_theft	17

Figure 4. Count of Missing Values

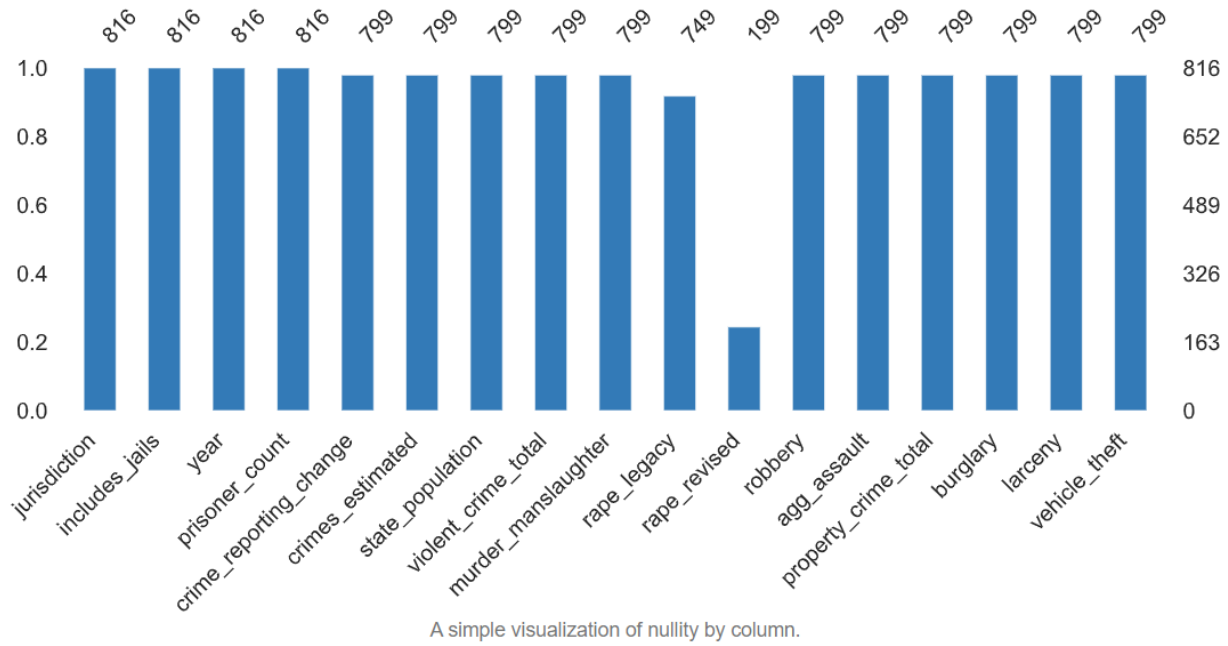

Figure 5. Missing Values Plot

As mentioned, the rape revised column has maximum null values and is not an important feature for consideration, hence the field is dropped. Apart from this, for the 17 null values, since each of those fields have same null values for each row, all the 17 values are dropped which results to 799 rows of data and 16 field values. But as observed below, the rape legacy still has some missing values despite performing the mentioned cleaning steps.

jurisdiction	0
includes_jails	0
year	0
prisoner_count	0
crime_reporting_change	0
crimes_estimated	0
state_population	0
violent_crime_total	0
murder_manslaughter	0
rape_legacy	50
robbery	0
agg_assault	0
property_crime_total	0
burglary	0
larceny	0
vehicle_theft	0

Figure 6. Missing Values after Data Cleaning

In order to check for the rape legacy column of missing values, since the missing data in the field is approximately 20% or less, mean, median or mode can be applied to the data column to impute the missing values. Now, the distribution plot for rape legacy is taken into consideration to check if the data is normally distributed or the data is skewed which would help to determine the imputation method.

And as observed, the data is right skewed and hence the median operation would be the best to perform for imputation of the missing data. If the data was normally distributed, the mean or median could have been used but since data is skewed, the median approach is used to fill in for the missing values. Hence, the data cleaning step is now completed.

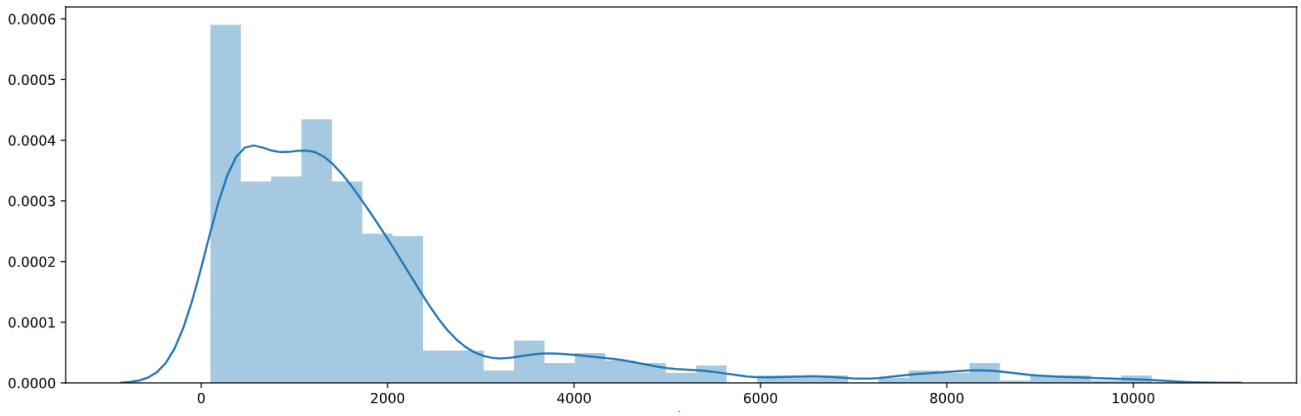

Figure 7. Distribution Plot for rape legacy column

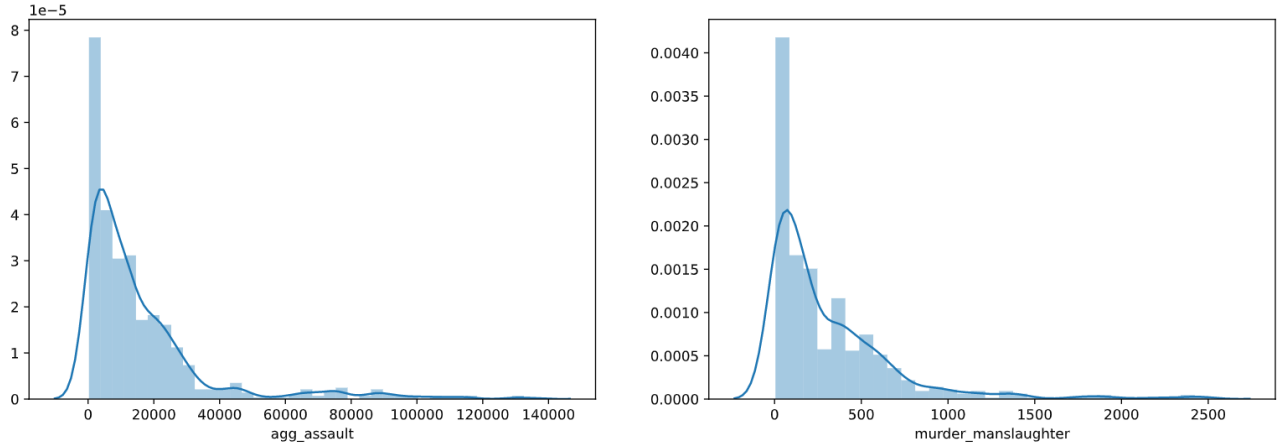

Figure 8. Distribution Plot for aggravated assaults and murder manslaughter

The distribution plot for aggravated assaults and murder manslaughter show that the data for both the field values is skewed. The boxplot gives an understanding of the outliers present in the data values for prisoner count and violent crime field values as shown in the below figure. As observed from the boxplot, we can see that there are outliers present in the field values of prisoner count and violent crime total data.

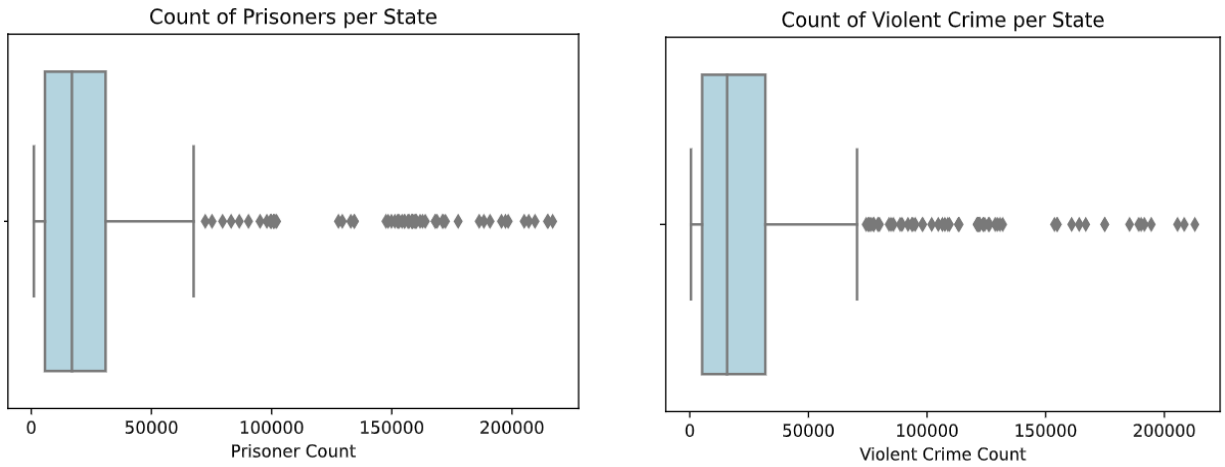

Figure 9. Boxplot of prisoner count and violent crime per state

Data Visualization

Data visualization helps in analyzing and presenting the data in an effective way through the graphical representations. The crime and incarceration dataset contains various parameters for which the visualizations would help in gaining some insights from the data and also helping to analyzing the relationship between the parameters.

The below visual represents the prisoner count in each state and as it is observed the count of prisoners is highest in the state of California and Texas. This gives an overview of the distribution of prisoners in each states which is helpful in analyzing the count of prisoners in each state based on crime types. Through this analysis we get an understanding for the crimes estimated and the imprisonment obtained for the crimes within the state.

The prisoner count is an important feature to be considered while the analysis of the data as for the highest count of prisoners in the state the crimes committed need to be noted through the years based on which the comparison of the report change for the future can be predicted. Hence, this would be one of the important features to be considered for the prediction of the analysis.

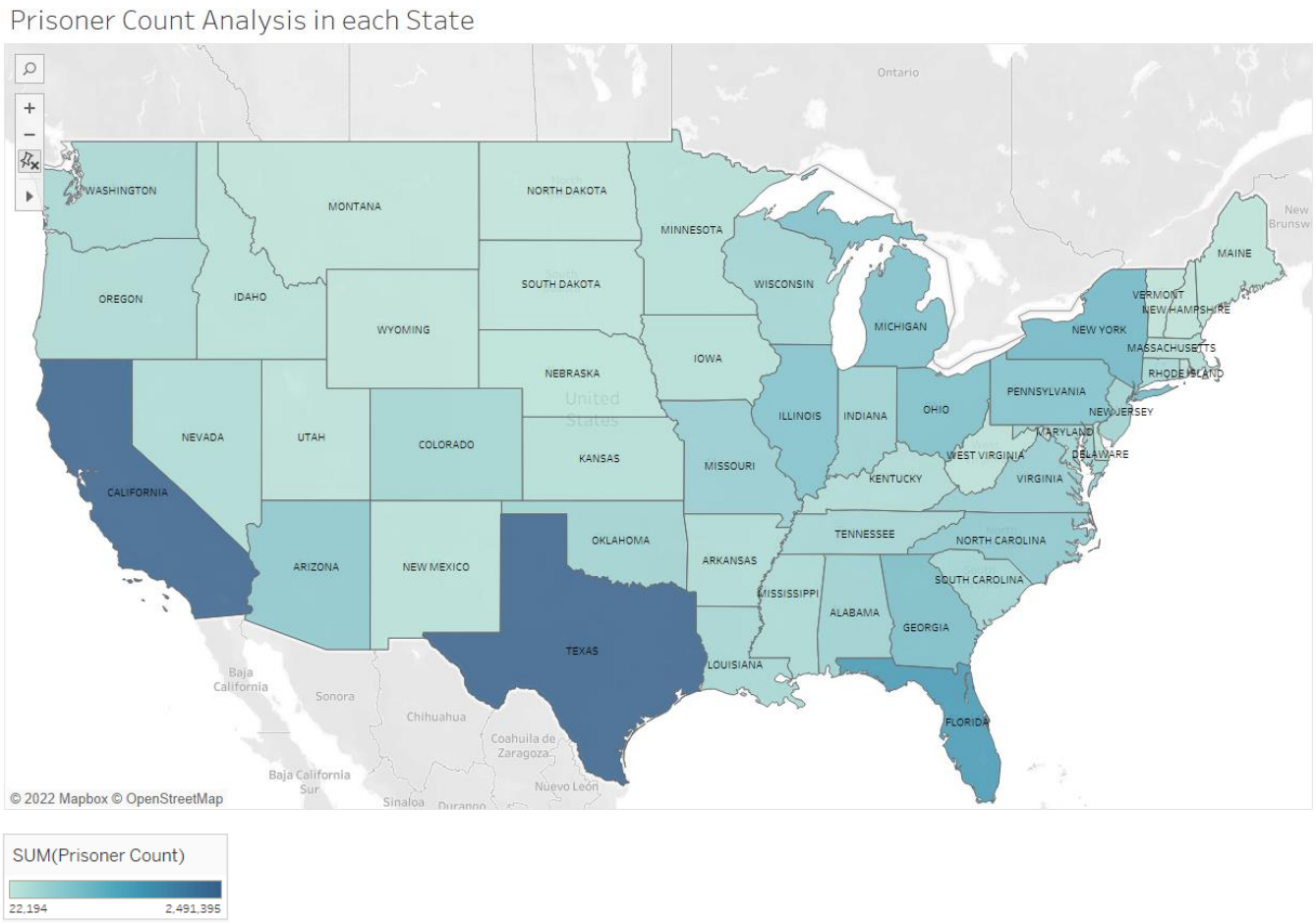

Figure 10. Prisoner count analysis in each State

The next visual is the average prisoner count in the year which helps in understanding the prisoner count during the years from 2001 to 2016 and the previous year analysis with respect to the prisoner count would be important to analyze as this would help understand the data and information for the previous years in order to compare it with the future reports change. As observed, the highest average prisoner count of 29,895 is in the year 2009 and the lowest average prisoner count of 26,075 is in the year 2001. Also, after the year 2009 the average count of prisoners decreased until 2016.

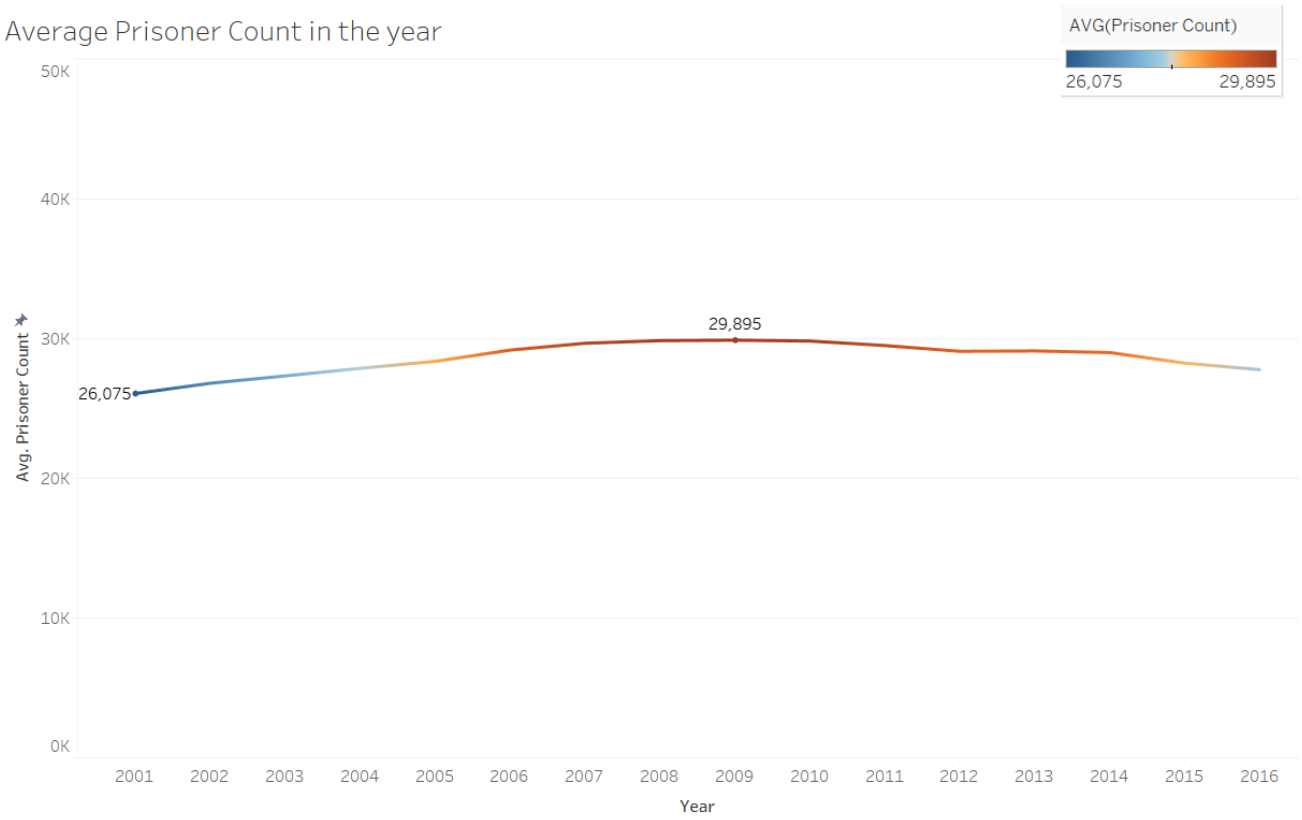

Figure 11. Average Prisoner Count in the year

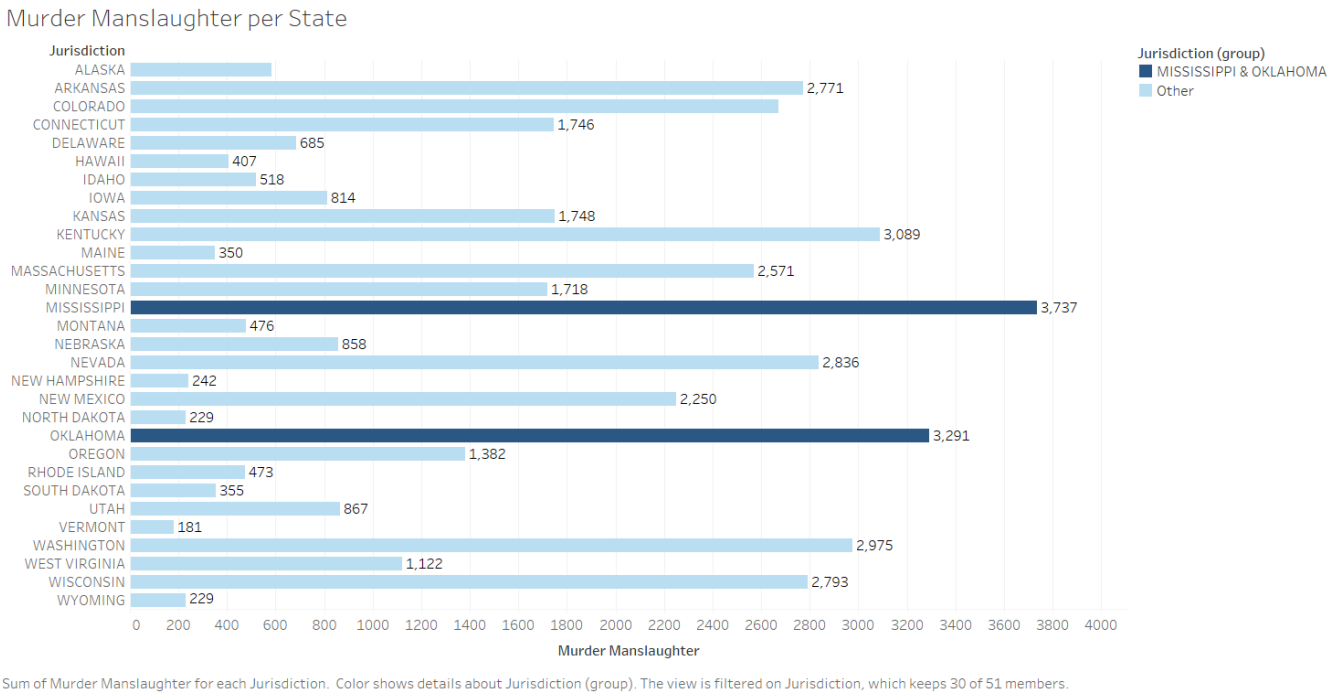

Figure 12. Murder Manslaughter per State

The above visual represents the murder manslaughter per state which shows that Mississippi and Oklahoma have the highest count of murder manslaughter as compared to the other states and this analysis will be helpful in pointing to the researchers based on a particular crime type. As mentioned earlier, violent crime, murder manslaughter, and aggravated assault are the interested crime types which need to be considered and hence this is one of the analysis that would be further useful in the comparison of the crime stating reports.

Next visual of whether the crime include jail or not is based on the two crime types, namely, average violent crime total and average aggravated assault crime type. It is observed that despite the average count being higher for violent crime type with an average of 29,203, it does not include jail as compared to the one which includes jail. Also, it can be observed that the average violent crime count is more that the aggravated assault crime type which is one analysis to note as this could help researchers understand the count of different crime types for future reports generation.

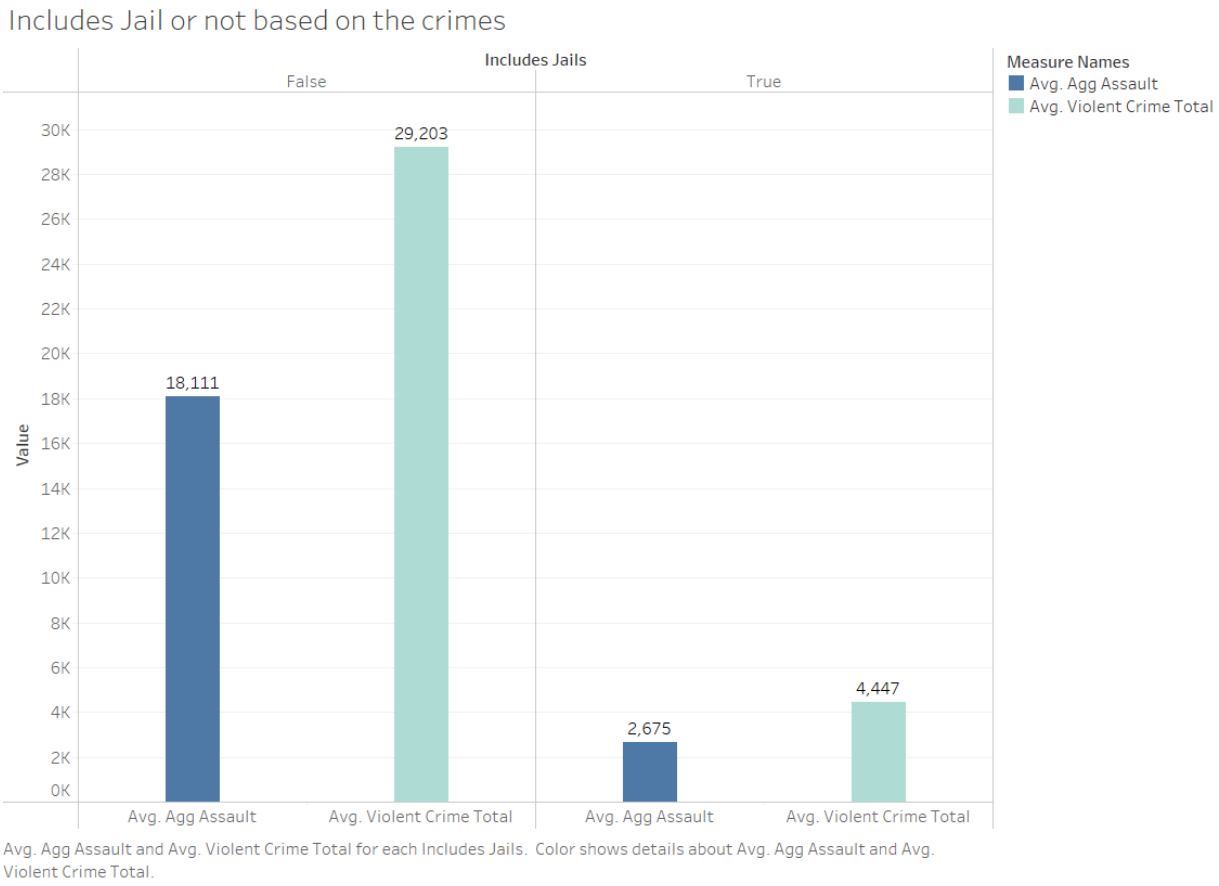

Figure 13. Includes jail or not based on the crimes

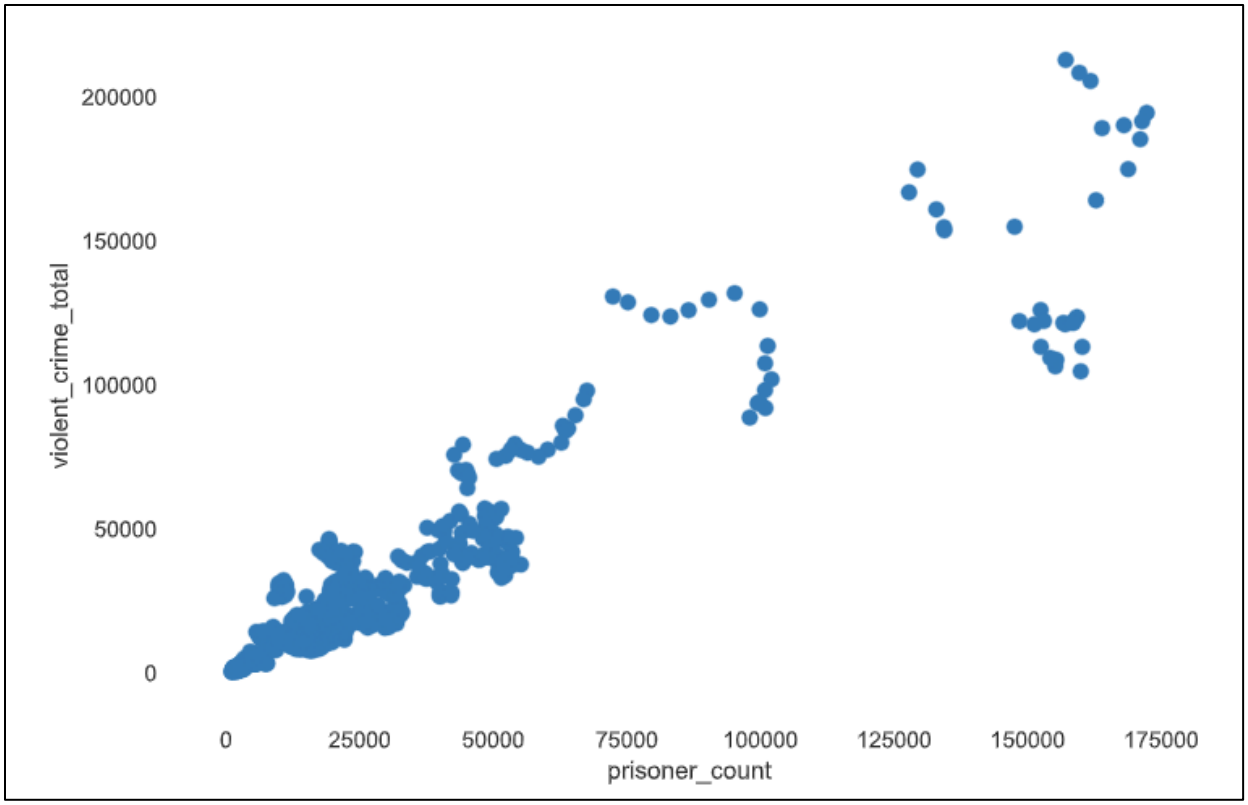

Figure 14. Interaction between Prisoner Count & Violent Crime Total

The interaction graph of the prisoner count and violent crime total gives an overview of the total count of prisoners for a total of violent crime committed in each state from 2001 to 2016. As observed, there is a strong interaction between the prisoner count and the crime at a level of **25000 prisoner count and 50000 violent crime**. This gives an understanding of the prisoner count for the violent crime committed which would further help in the analysis while predicting the prisoner count.

Similar to the above visual, the interaction of prisoner count with murder manslaughter and aggravated assault is also plotted to understand each of these interactions to monitor the prisoner count for each of the crimes mentioned which is as shown below.

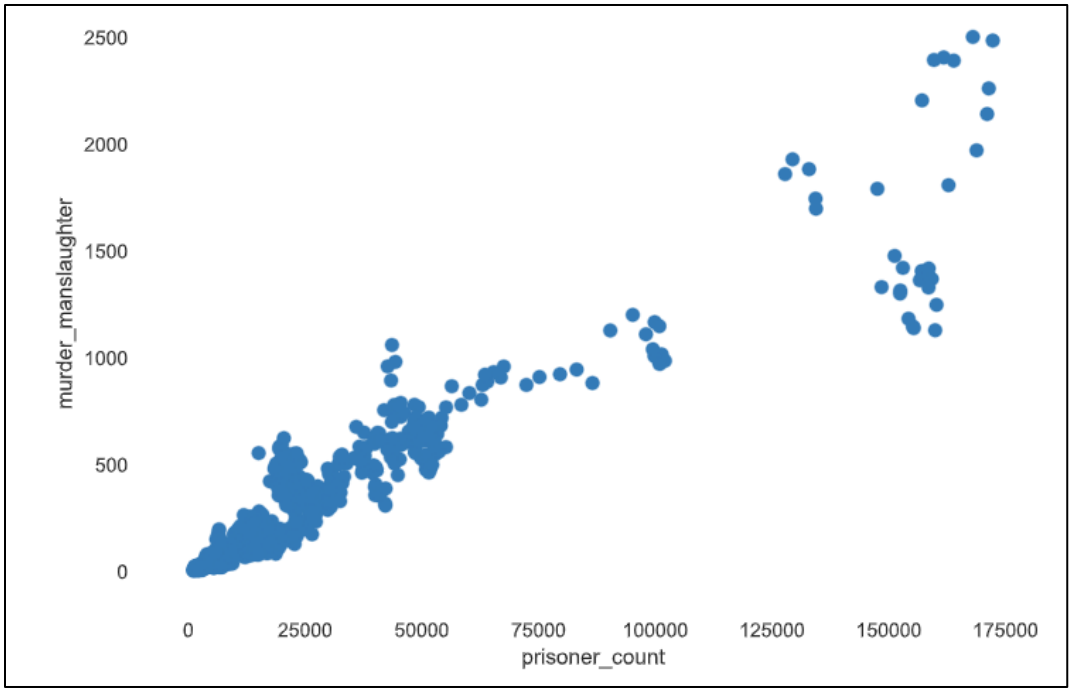

Figure 15. Interaction between Prisoner Count & Murder Manslaughter

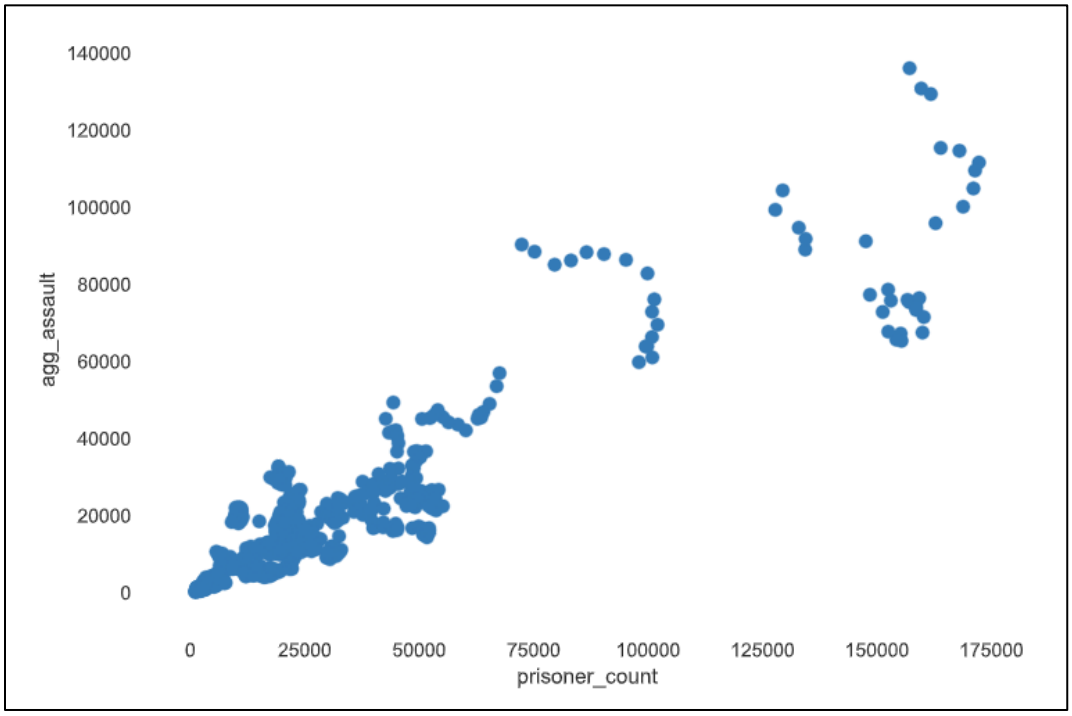

Figure 16. Interaction between Prisoner Count & Aggravated Assault

Pre Modeling Steps

The pre modeling steps include various process and implementation of steps in order to understand the independent variables and the relationship between each of these parameters and features in order to build the machine learning model where these features would be given for the training of the model. The pre modeling steps consists of feature selection and extraction, correlation plot, and label encoding. The automatic feature selection and extraction will automatically identify the important and significant parameters and variables which will help to predict the dependent variable. The correlation plot determines the correlation between each of the variables and if there is a high collinearity that exists between the variables, either of the variables are dropped to avoid multicollinearity and overfitting of the model. Lastly, label encoding is performed on the categorical variables which are considered for the feature extraction.

Feature Selection & Extraction

The automatic feature selection and extraction is performed in order to extract the important and significant features from the dataset which would help in the prediction of the prisoner count. The features that were selected are **'state_population', 'jurisdiction', 'includes_jails', 'crime_reporting_change', and 'crimes_estimated'**. These features would help in the prediction of prisoner count in each state and to also understand the various crimes that are affecting the prisoner count in each state.

Correlation Plot

The correlation plot helps to understand the correlation between each of the independent variables of the dataset and to plot the correlation values of the parameters to know the collinearity between the variables such that the variables with high collinearity value can be dropped in order to avoid multicollinearity and overfitting of the model. From the correlation plot it is observed that since there are different crime variables having the same value, there is a **high collinearity** that is existing between the variables with a correlation value of **0.9**. This would result in multicollinearity and the model would outperform resulting in inaccurate model and results. Thus, some of the features having high collinearity are dropped before considering them for training of the model.

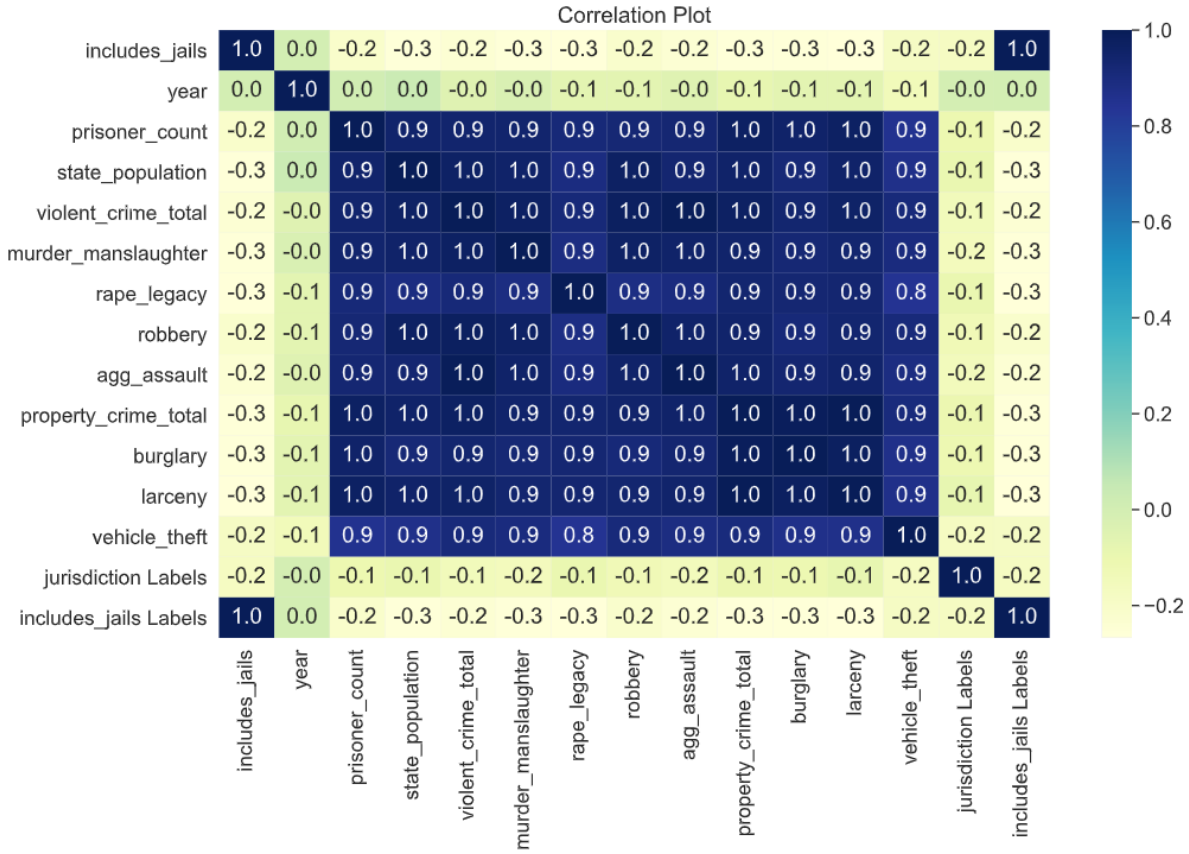

Figure 17. Correlation Plot

Label Encoding

Since the dataset consists of categorical data which needs to be considered as the features for model building, the independent variables need to be label encoded in order to have numerical data passed to the model. Thus, the features such as ‘includes_jails’, and ‘jurisdiction’ are label encoded to numerical form which can be considered as features to training of the model.

crimes_estimated	state_population	violent_crime_total	murder_manslaughter	rape_legacy	robbery	agg_assault	property_crime_total	burglary	larceny	vehicle_theft	jurisdiction Labels	includes_jails Labels
False	4468912.0	19582.0	379.0	1369.0	5584.0	12250.0	173253.0	40642.0	119992.0	12619.0	0	0
False	633630.0	3735.0	39.0	501.0	514.0	2681.0	23160.0	3847.0	16695.0	2618.0	1	1
False	5306966.0	28675.0	400.0	1518.0	8868.0	17889.0	293874.0	54821.0	186850.0	52203.0	2	0
False	2694698.0	12190.0	148.0	892.0	2181.0	8969.0	99106.0	22196.0	69590.0	7320.0	3	0
False	34600463.0	212867.0	2206.0	9960.0	64614.0	136087.0	1134189.0	232273.0	697739.0	204177.0	4	0
False	4430989.0	15492.0	158.0	1930.0	3555.0	9849.0	170887.0	28533.0	121360.0	20994.0	5	0
False	3434602.0	11492.0	105.0	639.0	4183.0	6565.0	95299.0	17159.0	65762.0	12378.0	6	1
False	796599.0	4868.0	23.0	420.0	1156.0	3269.0	27399.0	5144.0	19476.0	2779.0	7	1
False	16373330.0	130713.0	874.0	6641.0	32867.0	90331.0	782517.0	176052.0	516548.0	89917.0	8	0
False	8405677.0	41671.0	598.0	2180.0	14402.0	24491.0	347872.0	71799.0	238484.0	37589.0	9	0

Figure 18. Label Encoding

Model Building

The model building is performed after the implementation of the pre modeling steps in order to predict the prisoner count in each state. Since the objective of the project is to predict the prisoner count in each state and determine the features affecting the imprisonment, classification machine learning models cannot be implemented and thus regressor models are trained and implemented which would predict the prisoner count of the state. The different regressor models implemented for the prediction of prisoner count are **Linear Regressor Model, Decision Tree Regressor, and Random Forest Regressor** models. For the implementation of these machine learning models and based on the feature extraction and correlation plot, the significant features that are considered are *'jurisdiction', 'includes_jails', 'state_population', 'violent_crime_total', 'murder_manslaughter', and 'agg_assault'*. Since the different types of crimes variables gave a high collinearity, only the three important crime features are considered along with the state, state population, and includes jail or not.

Linear Regressor Model

Linear regression is a basic predictive analytics approach that predicts an output variable using historical data. The core concept is that if we can fit a linear regression model to observed data, we can use it to predict future values. The implementation of the Linear Regressor Model consist of various features such as jurisdiction, includes_jails, state_population, violent_crime_total, murder_manslaughter, and agg_assault for the prediction of the prisoner count in each state. The accuracy obtained for both the training and testing set of the Linear Regression model is **92% and 91.01%** respectively. Since this is a regressor type of model, the model evaluation is based on the **MAE, MSE, RMSE and R-Squared** values in order to determine the prediction error of the model implemented.

```
Accuracy of Linear Regressor model on training set: 0.92
Accuracy of Linear Regressor model on test set:     0.91

0.9101
```

```
Model Evaluation of Linear Regression.
Mean Absolute Error: 5857.4
Mean Squared Error: 94183008.6
Root Mean Squared Error: 9704.8
R-Squared value: 0.9100756689449834
```

Figure 19. Accuracy & Model Evaluation

Decision Tree Regressor Model

Decision trees are frequently used in operations research, particularly in decision analysis, to aid in the identification of the most likely method to achieve a goal. One of the advantages of a decision tree model is that the predictor and target variables are straightforward to grasp.

The independent variables considered for the implementation of the Decision Tree Regressor are *'jurisdiction', 'includes_jails', 'state_population', 'violent_crime_total', 'murder_manslaughter', and 'agg_assault'*. The decision tree model is implemented with a train and test data **split of 80-20** with a **maximum branch depth of 5**. The accuracy of the Decision Tree model obtained for the prediction of the prisoner count is **99%** for both the **training and testing set**, which is a good accuracy overall, but machine learning models having a 99% accuracy is not effective. This is because the model is overfitted due to the multicollinearity of the parameters and passing less data values for the training of the model. The feature importance, and model evaluation for the Decision Tree Regressor model is as follows.

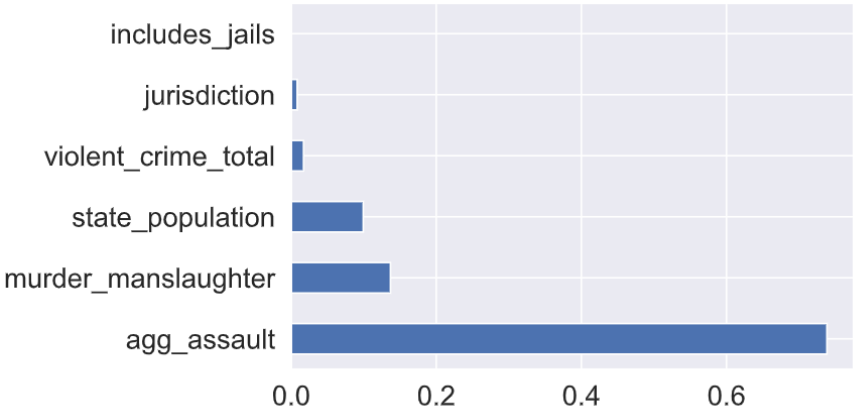

Figure 20. Feature Importance

```
Model Evaluation of Decision Tree Regressor.
Mean Absolute Error: 2371.9
Mean Squared Error: 14738704.8
Root Mean Squared Error: 3839.1
R-Squared value: 0.9859277358751884
```

Figure 21. Model Evaluation

The feature importance graph represents the important features that are considered during the prediction of the prisoner count by the decision tree model. This implies that certain features were repeated several times which helped in the prediction of the outcome variable. The model evaluation gives the Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, and R-Squared value which would help in comparing between the various models implemented and understanding which model accurately predicted the prisoner count.

Random Forest Regressor Model

To generate a more precise and reliable forecast, Random Forest creates many decision trees and blends them together. It has the benefit of being able to solve both classification and regression issues. The random forest is a classification system that uses numerous decision trees to make judgments. When creating each individual tree, it employs bagging and feature randomization to generate an uncorrelated forest of trees whose committee prediction is more accurate than that of any one tree.

For the prediction of the prisoner count, the features selected for the training of the model is same as that selected for the Linear Regression model and Decision Tree model. The data is split into **80-20 ratio** for training and testing of the model where the Random Forest Regressor is implemented with a minimum of **5000 tress** and **maximum depth branch of 5.** The accuracy of the model obtained is **99% for training data** and **98.8% for test dataset**. The feature importance and model evaluation for the random forest regressor model is as follows.

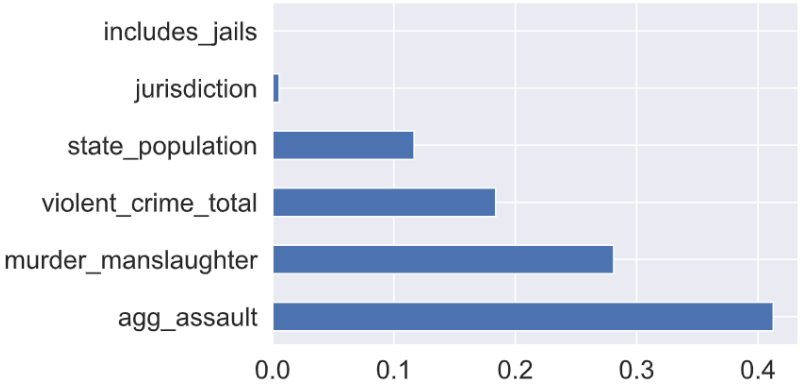

Figure 22. Feature Importance

```
Model Evaluation of Random Forest Regressor.
Mean Absolute Error: 2310.6
Mean Squared Error: 12862887.4
Root Mean Squared Error: 3586.5
R-Squared value: 0.9877187343522992
```

Figure 23. Model Evaluation

Recommendation

The crimes and incarceration data contains information about the various crimes that have been recorded along with the prisoner count in each state from the year 2001 to 2016. The objective here was to predict the crimes estimated and crime reporting change for the various crimes reported in the year. The correlation matrix as plotted above shows the correlation between the various parameters of the dataset and it is observed that the variables are highly correlated with each other, and prediction of crimes estimated is not effective. Due to this the objective and analysis is changed from a classification problem to a regressor problem where the prisoner count in each state is predicted based on the various crimes affecting the imprisonment.

The various regressor models build for the prediction of the prisoner count in each state are **Linear Regression Model, Decision Tree Regressor, and Random Forest Regressor models**. The features extracted for the training of these models were based on the feature selection and correlation plot. Due to the high collinearity between the independent variables, certain crime features were dropped and only the important features are considered for the prediction. The state population parameter, includes jail or not, and state parameter are the other features considered along with the types of crime variables.

The recommendation here to the researchers is that the prisoner count for each state is monitored based on the different crime types and thus **aggravated assault and murder manslaughter** are the two types of crimes which highly affect the crime rate and the prisoner count increases for the increase in these crimes. The researchers need to consider the reporting change factor based on the prisoner count prediction and for an **increase in the prisoner count** for the state, actions need to be taken which could help in the safety of public in that particular state.

The evaluation of the regressor models are based on the **MAE, MSE, RMSE, and R-squared** values. The accuracy obtained for the models are **91% for Linear Regressor model**, and **99% for Decision Tree and Random Forest Regressor model.** Hence, DT or RF model can be considered as the best fit model. The table below compares the evaluation of the various regressor

models implemented in order to determine the best fit model for the prediction of the prisoner count in each state.

Table 1. Regressor Model Comparison

	Machine Learning Regressor Models		
Evaluation Metric	Linear Regressor	Decision Tree Regressor	Random Forest Regressor
Accuracy	91.01%	98.6%	**98.8%**
Mean Absolute Error	5857.4	2371.9	2310.6
Mean Squared Error	94183008.6	14738704.8	12862887.4
Root Mean Squared Error	9704.8	3839.1	**3586.5**
R-Squared value	0.91	0.986	0.988

Thus, based on the model evaluation and comparison, it is observed that the accuracy for Random Forest Regressor is more as compared to other two models. The MAE and MSE values represent the difference between the actual and predicted values extracted by the mean error over the dataset. The RMSE is the error rate where the R-Squared represents how well the value fits compared to the original values, and the higher the value of R-Squared the better the model is. Thus, based on the evaluation metrices, it is observed that the **root mean squared error** is less in Random Forest model which implies that the prediction rate error is less when compared it to the other models. Hence, **Random Forest Regressor model would be recommended to use for the prediction of the prisoner count in each state**.

Conclusion

The project deals with analyzing the data to understand the relationship between crime incarceration rates and crime rates in order to analyze and answer the question of whether or not it improved public safety. The dataset thus consists of data on crimes committed and prisoner count in every 50 state and the interested crime types to analyze are violent crime total, murder manslaughter, and aggravated assaults. The important and significant features would help in determining the crime estimated within the state and help predict the whether or not the state changed reporting system had an affect with comparison to the previous years in United States.

In order to achieve the required business objective of predicting the prisoner count for each state, the data was analyzed based on the crime and incarceration dataset for which the descriptive and statistical analysis was performed. This gave an overview of what exactly does the dataset look like and what features does the data contain that can be used for further analysis. Here, the prisoner count analysis within the state and analysis of the crime type based on the average count within the state for the year gives an understanding to the researchers who can consider this insights in order to compare it to previous reports and understand the crimes estimated with respect to the different crime types within the state for the year of 2001 to 2016.

The prisoner count prediction will help the researchers in determining the **total count of prisoner** within each state for the various crimes estimated. The researchers need to monitor for the **murder manslaughter crime and aggravated assault crime** as these are the types of crimes having an increase in the prisoner count in the state. The reports generated by the researcher need to concentrate and update on these crime type to avoid the prisoner count which will then lead to public safety.

The model recommended in order to achieve this business objective is the **Random Forest Regressor** model which gave **a 99% accuracy**, and the **error rate of prediction is also less** as compared to the Linear Regressor model and Decision Tree Regressor model. The future scope would be to have more of data with respect to prisoner count in each state and some additional features for which a better model could be trained to improve the efficiency of the model as this data deals with the issue of multicollinearity.

References

Goyal, C. (2021b, July 1). How to Detect and Remove Outliers | Outlier Detection And Removal. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

Kumar, A. (2022a, January 23). Python - Replace Missing Values with Mean, Median & Mode. Data Analytics. https://vitalflux.com/pandas-impute-missing-values-mean-median-mode/

Visualizing distributions of data — seaborn 0.11.2 documentation. (2021). Seaborn. https://seaborn.pydata.org/tutorial/distributions.html

sklearn.linear_model.LinearRegression. (2022). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Decision Tree Regression. (2022). Scikit-Learn. https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

sklearn.ensemble.RandomForestRegressor. (2022). Scikit-Learn. https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

D. (2019b, October 10). Regression Accuracy Check in Python (MAE, MSE, RMSE, R-Squared). Data Tech Notes. https://www.datatechnotes.com/2019/10/accuracy-check-in-python-mae-mse-rmse-r.html