Northeastern University
College *of* Professional Studies

ALY 6040 : DATA MINING APPLICATIONS

CRIME & INCARCERATION IN THE UNITED STATES

Presented by-

VAMSHI KRISHNAMURTHY

RICHA UMESH RAMBHIA

JYOTHSNA AKULA

SAI MADHAV AVAKU

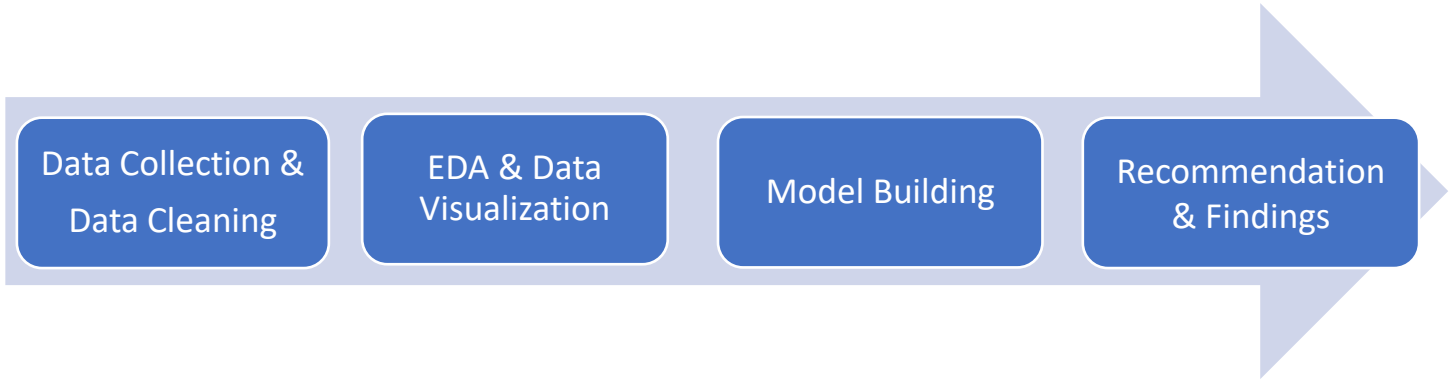
AISHWARYA JANGAM

Crime & Incarceration in the U.S.

Crime and Incarceration in the United States contains data on crimes that are committed, and the prisoner counts in every 50 state, covering from 2001-2017 where the most interest crime types in the data are violent_crime_total, murder_manslaughter, and age_assault.

Problem Statement:

- To analyze the dataset using EDA to gain further insights about the crimes data.
- To determine the crimes estimated and prisoner count in each state.
- Analyzing the prisoner count based on the various crimes committed to recommend the alertness and strictness of those crimes.



Tasks Performed:

1. Data Collection & Data Preparation
2. Exploratory Data Analysis
3. Data Cleaning
4. Data Visualization
5. Pre-Modeling Steps
6. Model Building
7. Recommendation & Findings

Crime & Incarceration in the U.S.

Data Collection & Preparation

- The dataset collected is from Kaggle which contains information about each of the crimes taken place in the United States from 2001 to 2016.
- This dataset is used in analyzing the crimes estimated to help understand the prisoner count in each state.
- It consists data consisting of various crimes having **816 rows** of crimes data and about **17 field values.**
- In order to better understand the dataset, **Descriptive & Statistical Analysis** is performed on this crimes data to gain further insights.

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

```
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')
```

Crime & Incarceration in the U.S.

Descriptive & Statistical Analysis - EDA

The descriptive analysis gives a broad overview of the dataset and the statistical analysis helped in understanding the statistics of the dataset.

- As observed in the statistical analysis, the average count of the various types of crime in the dataset are computed which clarifies the total count of the 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.

Statistical Analysis 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

Crime & Incarceration in the U.S.

Data Cleaning

- The dataset is checked for the total number of null or missing values in each column which needs to be taken care of while analysis.
- Based on the output, it is observed that there are **871 missing values** which is **6.3% missing cells** in the dataset having **no duplicate rows.**
- 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.

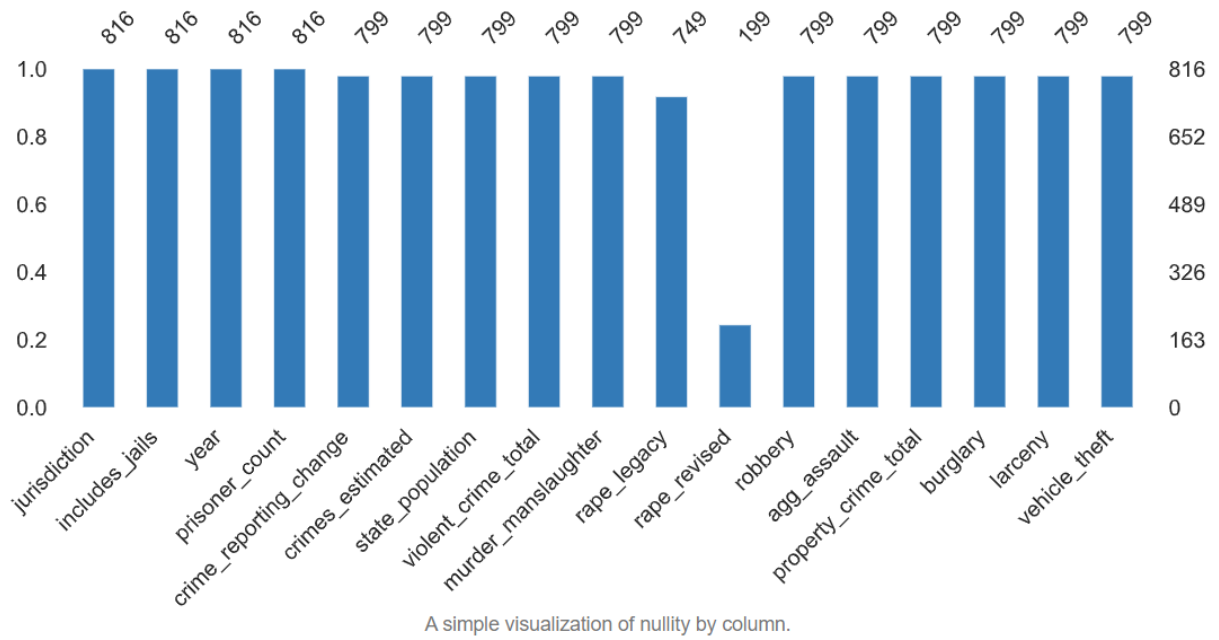
jurisdiction	0
includes_jails	0
year	0
prisoner_count	0
crime_reporting_change	17
crimes_estimated	17
state_population	17
violent_crime_total	17
murder_manslaughter	17
rape_legacy	67
rape_revised	617
robbery	17
agg_assault	17
property_crime_total	17
burglary	17
larceny	17
vehicle_theft	17

Total number of null values in each column of the dataset

Crime & Incarceration in the U.S.

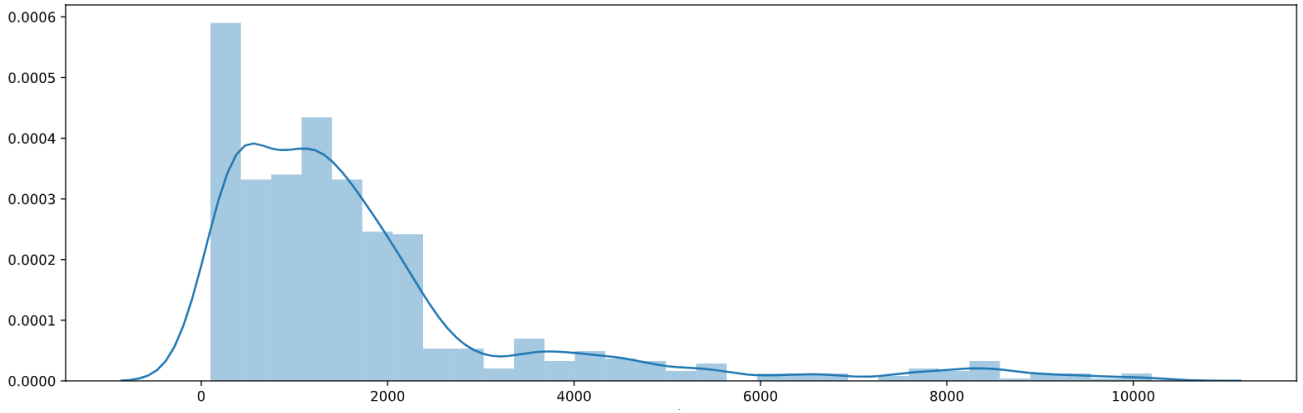
Data Cleaning

- 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, all the 17 values are dropped which results to 799 rows of data and 16 field values.
- In order to check for the rape legacy column of missing values, the distribution plot for rape legacy is taken into consideration 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.



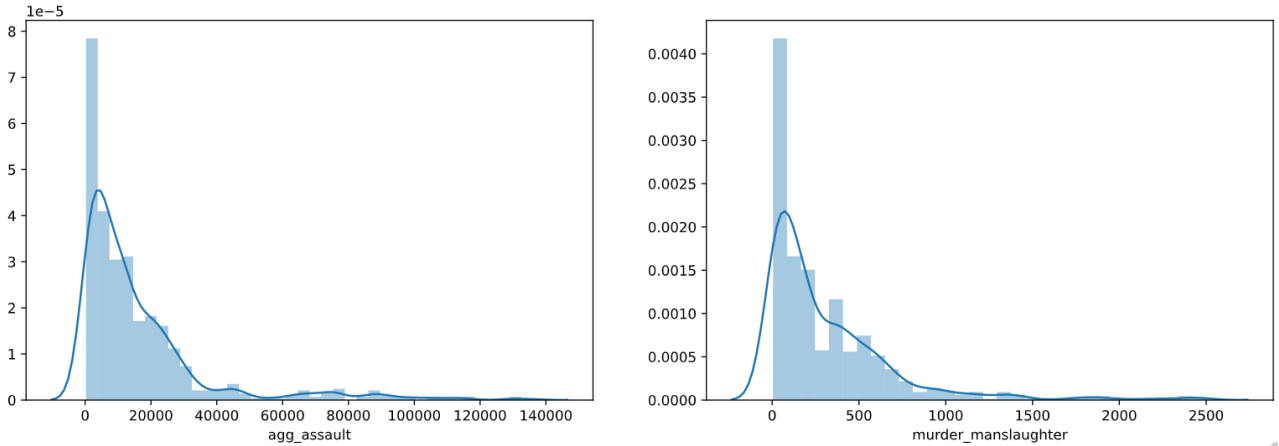
Crime & Incarceration in the U.S.

Data Visualization



Distribution Plot for Rape Legacy

The distribution plot for rape legacy, aggravated assaults, and murder manslaughter show that the data for the field values is skewed.

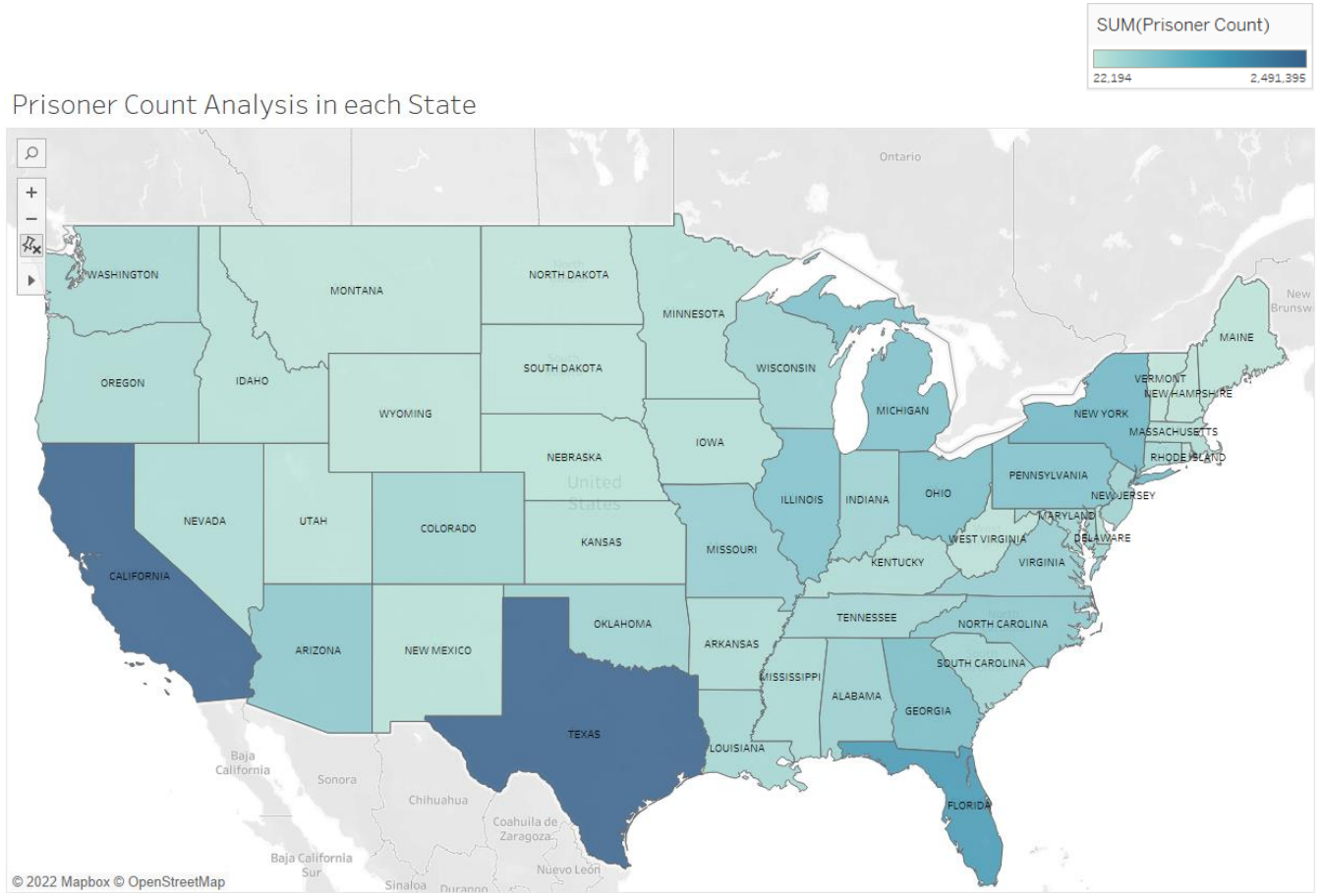


Distribution Plot

Crime & Incarceration in the U.S.

Data Visualization

- The 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.

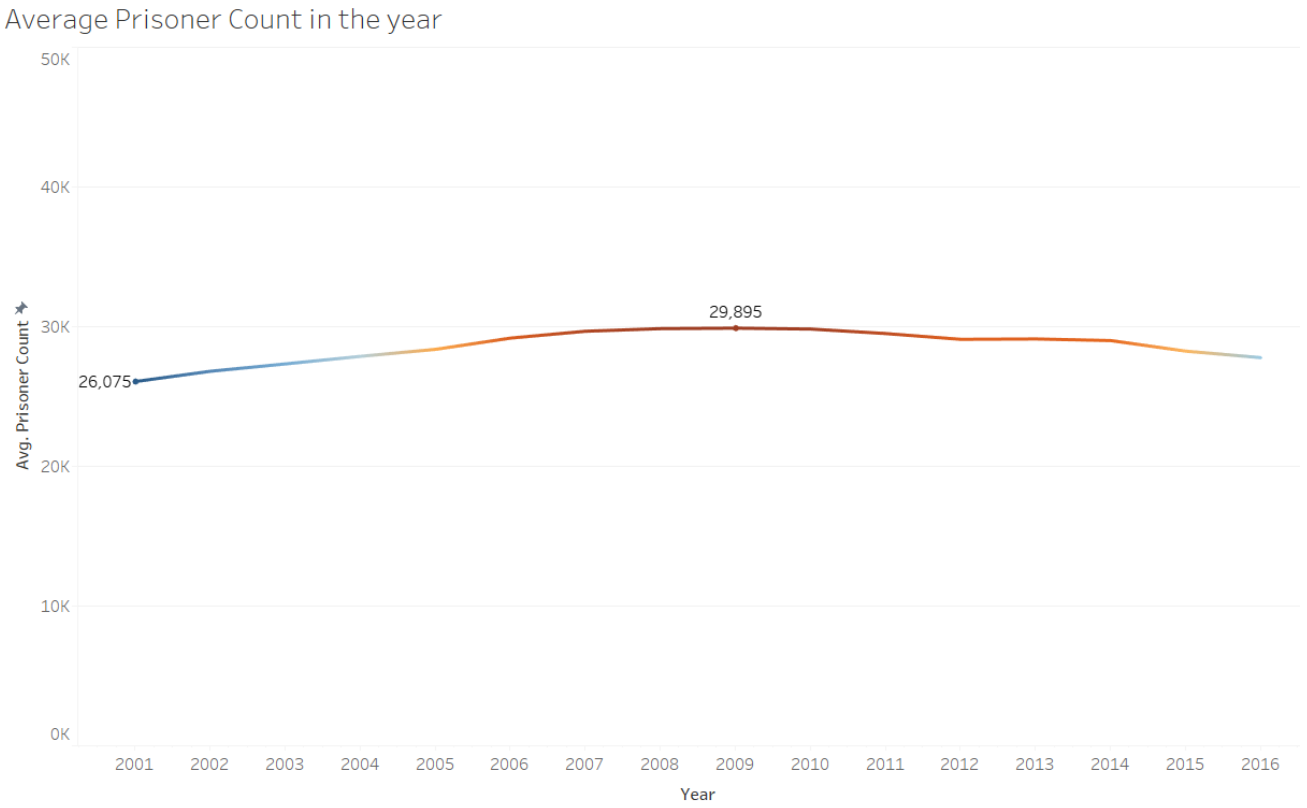


Prisoner count analysis in each State

Crime & Incarceration in the U.S.

Data Visualization

- This visual is the average prisoner count in the year which helps in understanding the prisoner count during the years from **2001 to 2016**.
- 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.

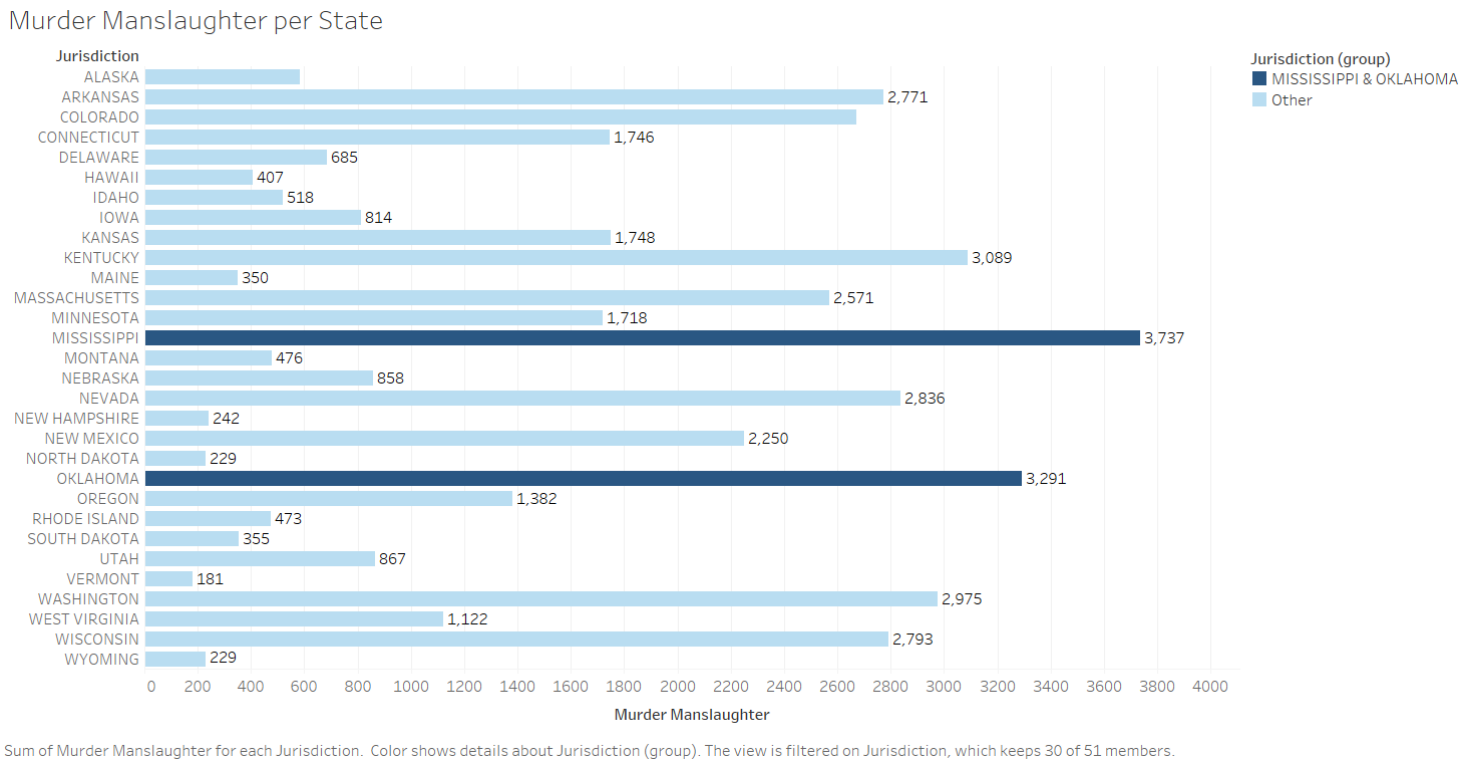


Average Prisoner Count in the year

Crime & Incarceration in the U.S.

Data Visualization

The 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.

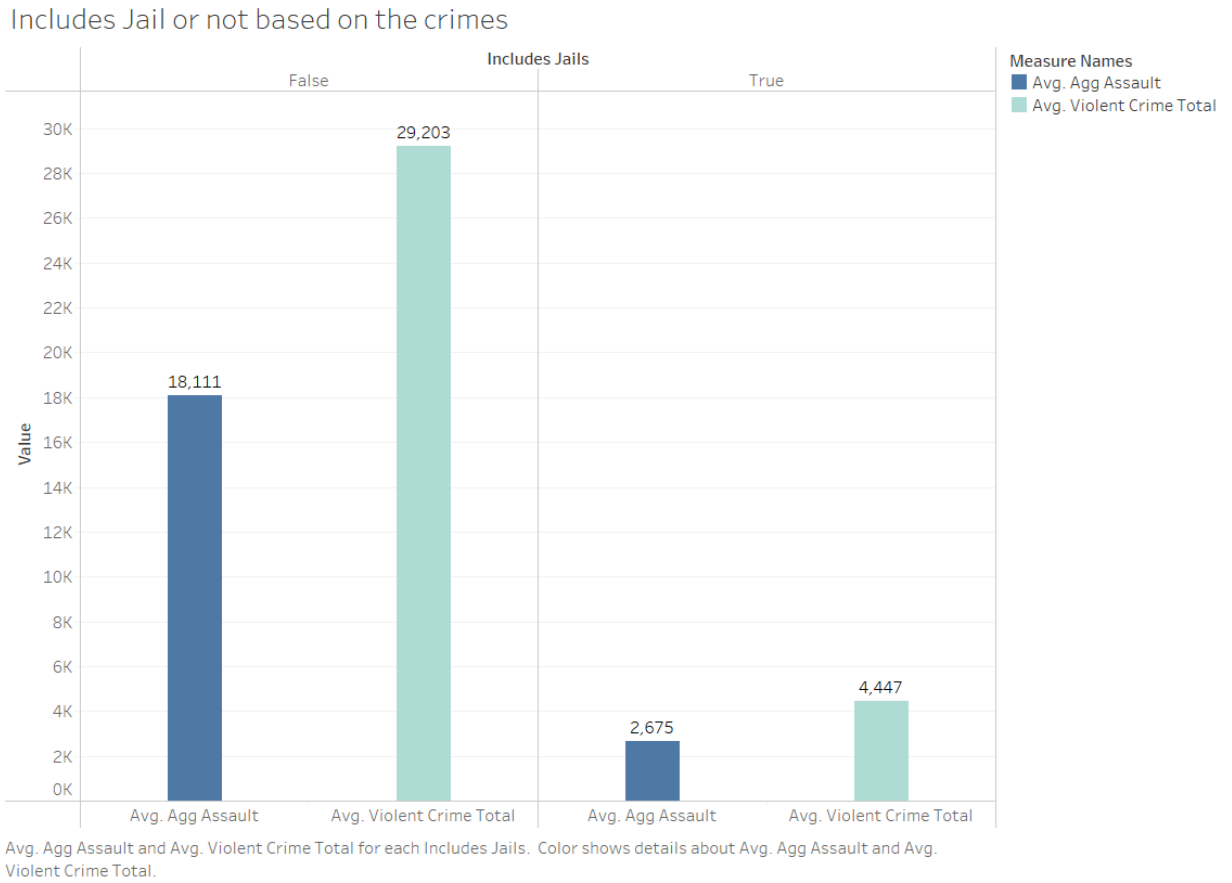


Murder Manslaughter per State

Crime & Incarceration in the U.S.

Data Visualization

- The 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.

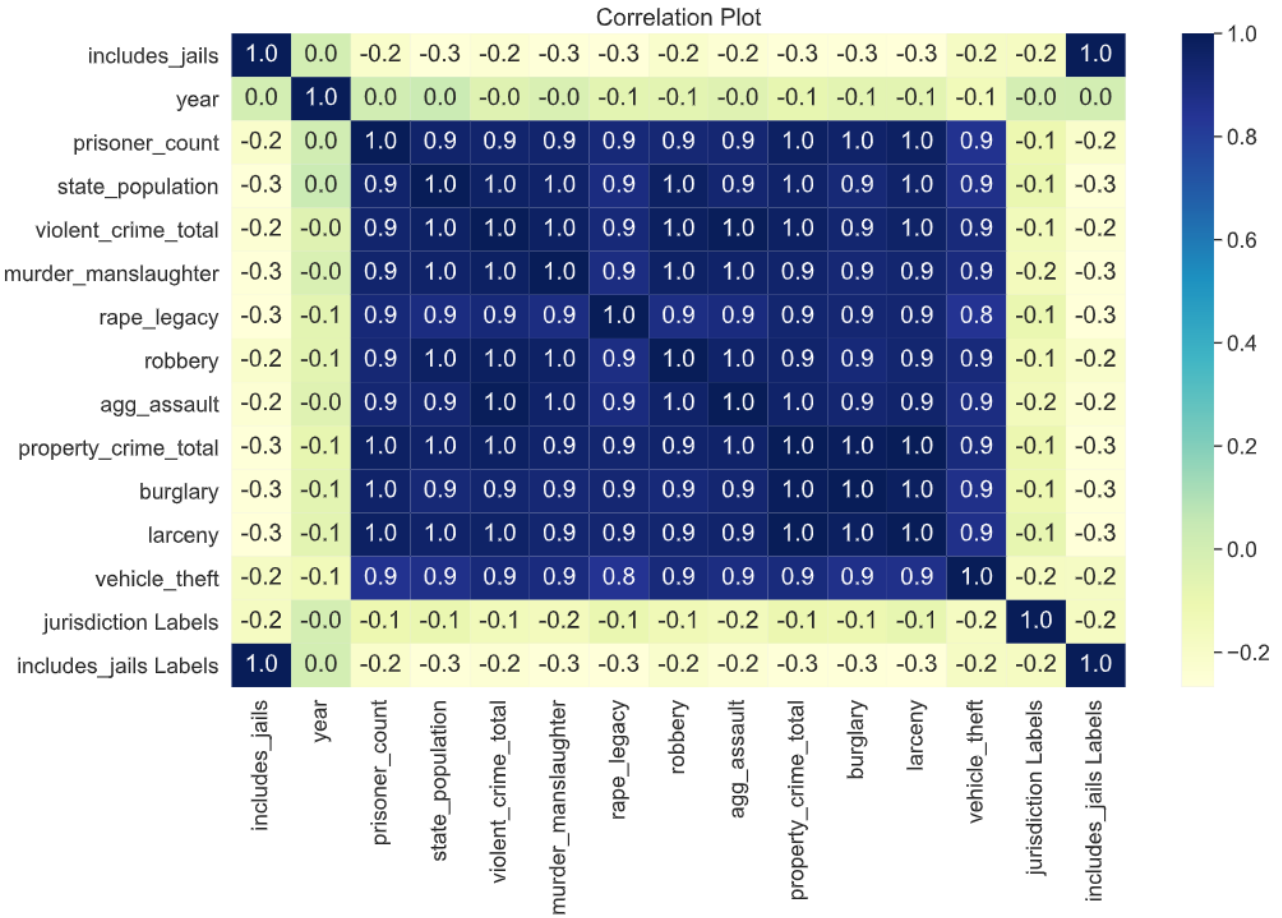


Includes Jail or not based on the Crimes

Crime & Incarceration in the U.S.

Pre-Modeling Steps

1. Feature Selection & Extraction
2. Correlation Plot
3. Label Encoding



The features that were selected are 'state_population', 'jurisdiction', 'includes_jails', 'crime_reporting_change', and 'crimes_estimated'.

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.

Crime & Incarceration in the U.S.

Model Building

Machine Learning Models Implemented.

- Linear Regressor Model
- Decision Tree Regressor Model
- Random Forest Regressor Model

Linear Regressor Model

- 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 the Linear Regressor Model

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

0.9101
```

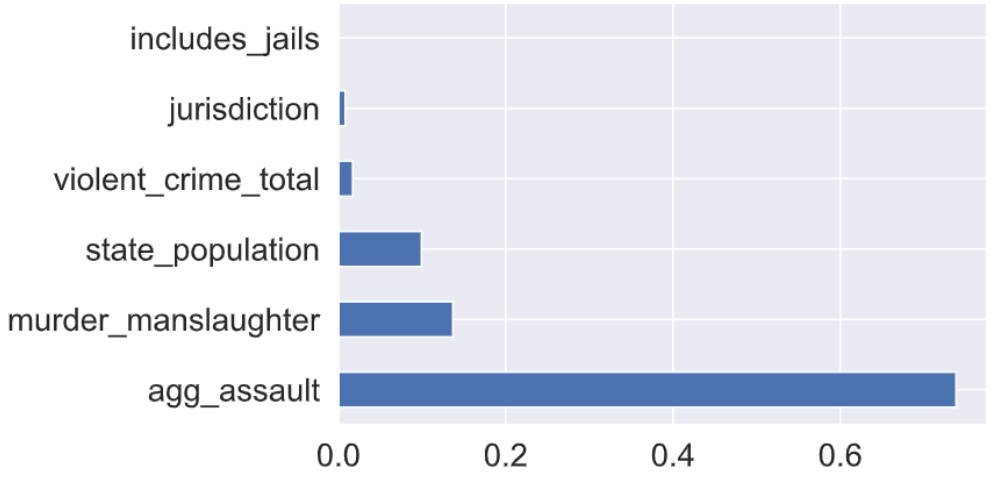
```
Model Evaluation of Linear Regression.
Mean Absolute Error: 5857.4
Mean Squared Error: 94183008.6
Root Mean Squared Error: 9704.8
R-Squared value: 0.9100756689449834
```

Model Evaluation

Crime & Incarceration in the U.S.

Decision Tree Regressor Model

- The 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.



Feature Importance of the Decision Tree Regressor Model

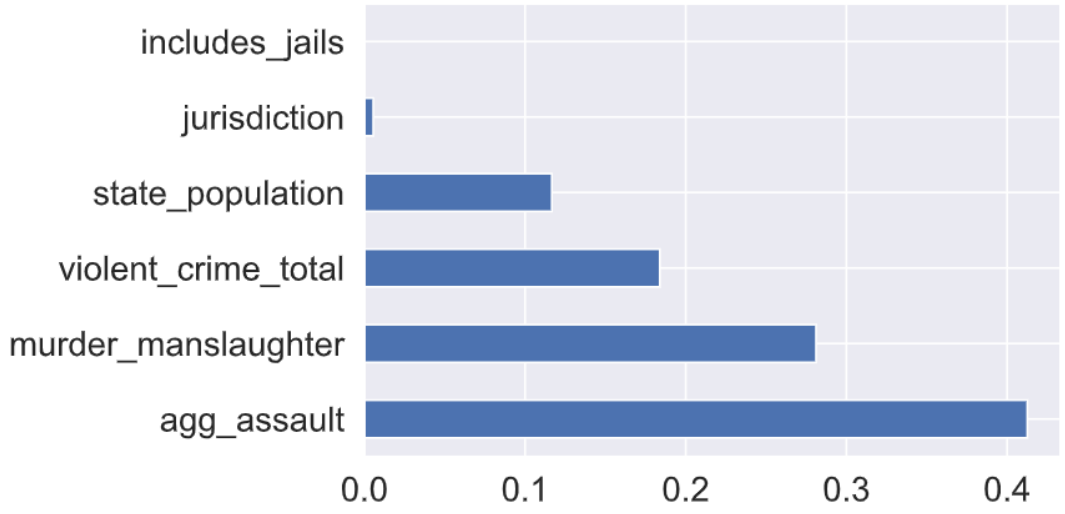
```
Model Evaluation of Decision Tree Regressor.
Mean Absolute Error: 2371.9
Mean Squared Error: 14738704.8
Root Mean Squared Error: 3839.1
R-Squared value: 0.9859277358751884
```

Model Evaluation

Crime & Incarceration in the U.S.

Random Forest Regressor 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**.



Feature Importance of the Random Forest Regressor Model

```
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
```

Model Evaluation

Crime & Incarceration in the U.S.

Model Evaluation & Comparison

- 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.
- 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.

	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

Crime & Incarceration in the U.S.

Recommendation

- 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 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.**
- But based on the model evaluation and metrices, RF model has the lowest RMSE value, implying low prediction rate error, hence RF model is recommended to use for the prediction of the prisoner count in each state.

Crime & Incarceration in the U.S.

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 it improved public safety.
- 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 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.

Crime & Incarceration in the U.S.

THANK YOU