USE CASE STUDY REPORT

A Data Driven Analysis of the Indian Prison System

IE 7275 – Data Mining in Engineering

Prof. Xuemin Jin

Student Name: Yeshwin Krishnamoorthy

20th June, 2018

Northeastern University
College *of* Engineering

I. Background and Introduction

At first glance, the Indian prison population may not seem particularly alarming; as a metric measured per million people; Indian prisoners constitute one of the lowest rates in the world. However, Indian prisons hold close to 420,000 prisoners, which is a substantial number. Since prison populations tend to indicate broader socioeconomic trends in the country, such as hyperactive policing, racial/religious discrimination, etc. it is in the interest of social scientists to analyze prison and justice systems data to validate or disprove these trends. Given that India is a vast country with multiple ethnic groups and four major religions, this data may reflect communities who are victim to discrimination by the police force or the justice system. Additionally, while the caste system has been abolished in India, deeply rooted social sentiment still lingers in most communities across the country and may even be manifest in a potential overrepresentation of traditionally lower caste groups within the prison population. Analyzing the data will thus provide a wealth of information that may be indicative of macrosocial discrimination and will enable predictions about the status of prison populations in the future.

The data used in this study is sourced from the National Crime Records Bureau of India, and licensed under the Government Open Data License – India. This dataset is very rich, containing multiple facets of information ranging from their level of education to their religion and caste. In addition to containing all the inmates' details, it also contains about the information of the prison across all states and union territories like expenditure, budget, no. of vehicles available, etc. The data contains many .csv files all of which have 2 common variables - the year and the state/union territory where all the information is present.

The goal of this study is to establish relationships between different variables i.e. connecting a link between the background information of the prisoners and their activities. By doing this, it can help in enabling a better understanding of the communities that make up these prisons. Even though they are imprisoned due to committing a crime, it is important to ensure that these prisoners are not primarily in prison because of their background and social standing. It is also the responsibility of society to ensure that prisoners get proper rehabilitation and access to education or making them better citizens once they are released from prison. As such, this study can help identify important trends that can be used by the state and national governments to create a more targeted approach in prisoner rehabilitation programs.

There are many questions about Indian prison with this dataset. Some of the interesting questions that can be raised are:

- Percentage of jails overcrowded. Is there any change in percentage over time?
- How many percentages of inmates are re-arrested?
- Which state/union territory pay more wages to the inmates?
- Which state/union territory has more capital punishment/life imprisonment inmates?
- Inmates gender ratio per state

By analyzing the data, we can come up with possible solutions that will enable:

- Better accountability of running all the prisons in all the states/union territories
- Better organization of the prison structure

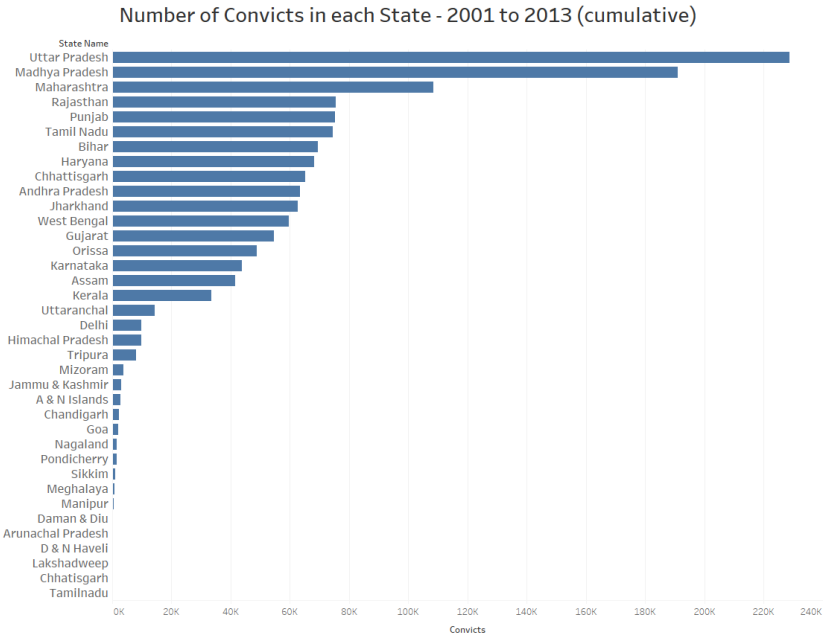
- Better assistance for inmates to make sure that they come out as better citizens and contribute to the betterment of the nation
- Prevention of people ending up in prison

For the case study report, we will first clean and preprocess the data. We will then visually display the data in the form of tables, charts, etc. Analysis will then be undertaken by using data mining techniques including linear regression, PCA, correlation and unsupervised machine learning to predict the future course of prisoners, in terms of recidivism, social rehabilitation, etc. We will also be dividing the entire dataset into a training set, a validation set and a test set. In addition to using R code, we will use Tableau and other data visualization tools to display the results since it will be visually appealing.

II. Data Exploration and Visualization

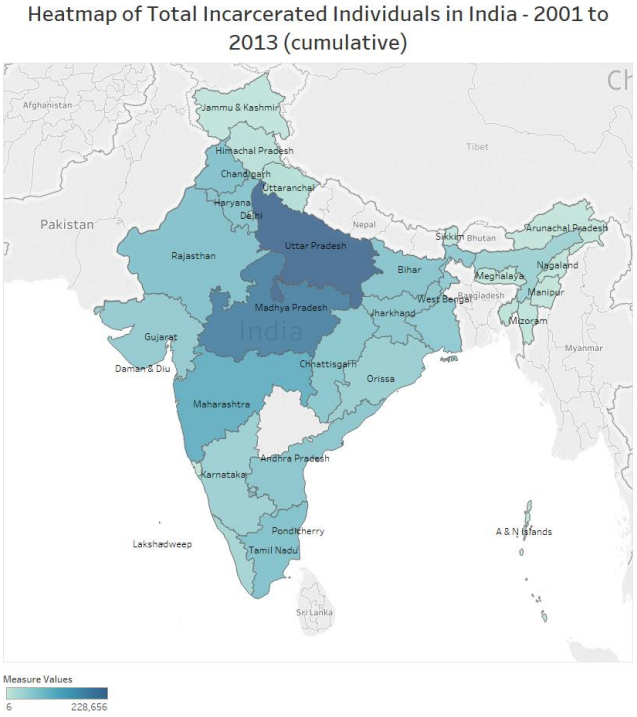
To begin processing the data set, it's necessary to visualize it in a way where noticeable patterns emerge and can be analyzed. To that end, the following bar graph showcases the cumulative number of convicts from each Indian state from 2001 to 2013.

The prison population in India up to 2013 has been overwhelmingly male. When contrasted with the ratio of the American prison system (where males make up 91.4% of the population), the male convicts in India represent almost 96% of all total prisoners. As such, for most of the following visualizations, the variables corresponding to the male populations of the state prison systems will be analyzed.



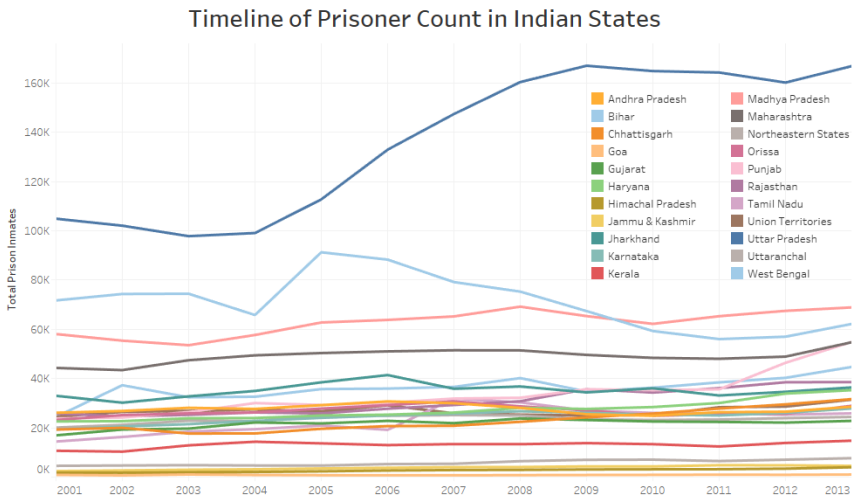
As can be seen above, there are three states that have noticeably more incarcerated people than the others: Uttar Pradesh, Madhya Pradesh, and Maharashtra. To analyze a data set that is rich enough to be useful but lean enough to easily wrangle data reasonably well. As such, the remainder of the data processing and visualization will be done on the aforementioned states, as well as three others: Rajasthan, Punjab, and Tamil Nadu, the next highest states.

To make visualizing this easier, the following heatmap can be referred to:

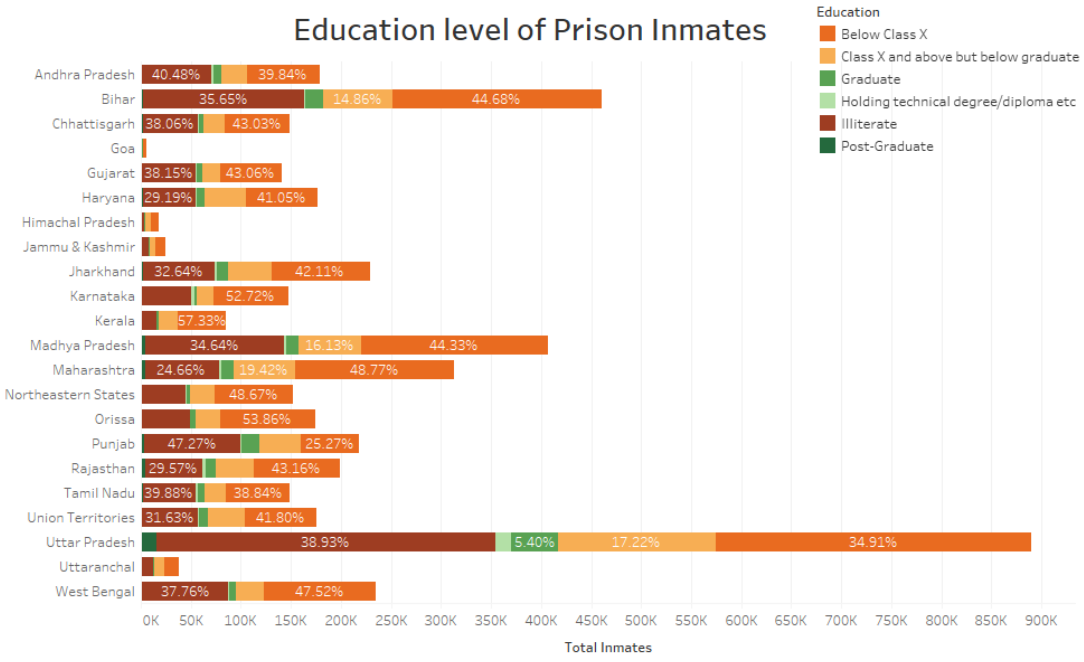


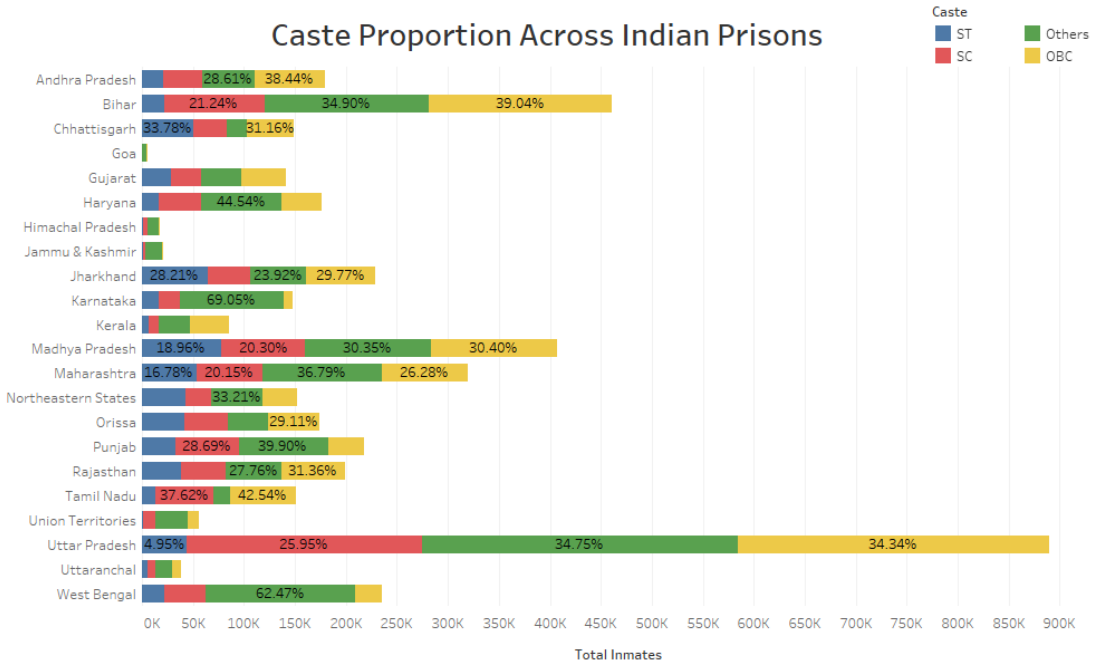
This heatmap reinforces the claims set forth previously regarding the states with a high number of imprisoned individuals. This map is useful for a cursory understanding of the distribution of criminals in India and where they're most prevalent.

Next, a sample age group (18-30 years old) is charted to try and spot noticeable trends. In this case, it can be noted that Tamil Nadu and West Bengal have a declining prisoner count, whereas Uttar Pradesh and Punjab have increasing counts. The remainder are somewhat static and don't exhibit any noticeably drastic trends.

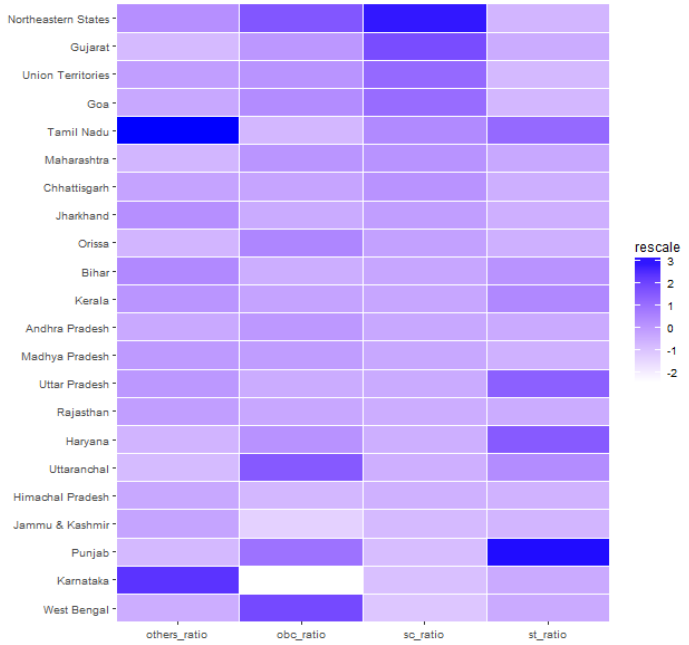


The next bar chart shows the education level of inmates. It is clear that the vast majority of Indian prisoners are poorly educated, with very few of them (proportionally speaking) holding an advanced degree. This is to be expected since uneducated populations tend to be disproportionately represented in the criminal justice system.

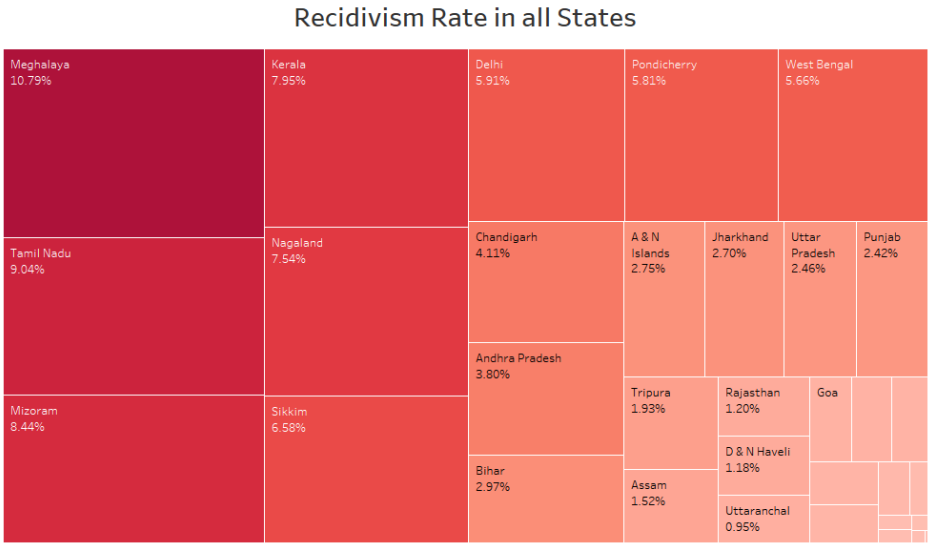




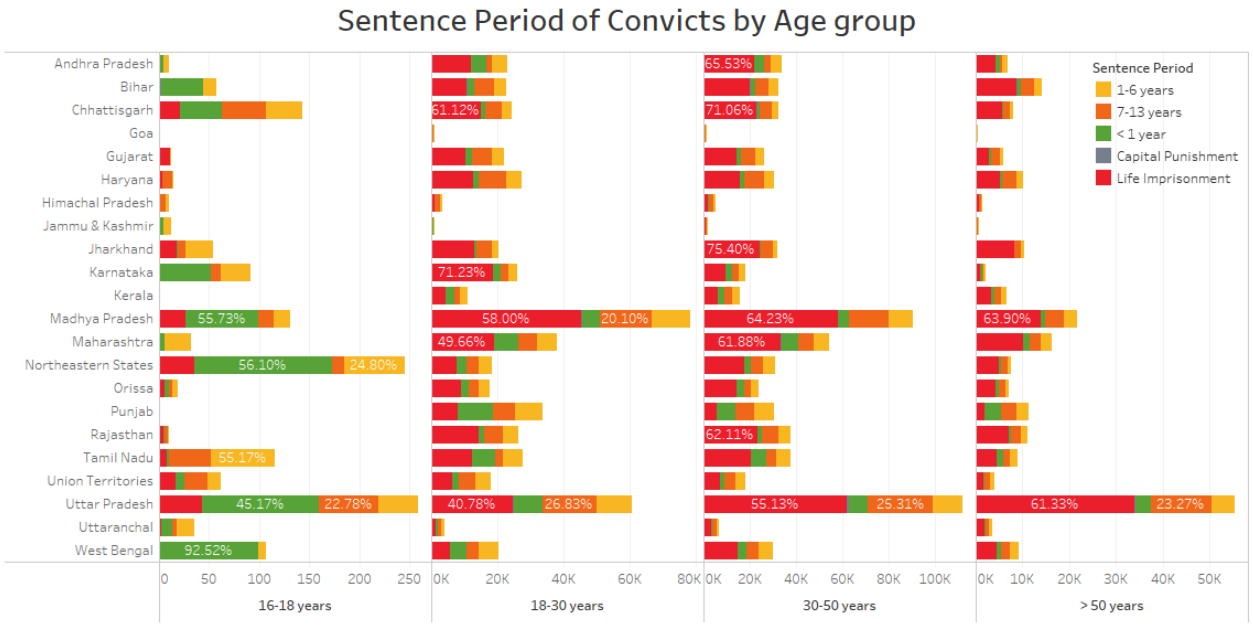
Despite casteism being banned by the government in 1947, the majority of India is still afflicted by this malaise. Due to its deep-seated cultural connection, and because lower castes across tend to be denied fair access to socioeconomic opportunities, they tend to be poorer and less educated. Due this, they are usually overrepresented in prisons, as can be seen in the figure above. In most states they form the overwhelming majority, except in Uttar Pradesh, Punjab, Haryana, and Bihar, where they still form a majority. Note: lower castes are ST, SC, and OBC. A composite representation of the upper castes is indicated by Others. This is an indication of the inequalities that persist in the modern Indian state. A more granular representation is shown below:



Here it can be seen that the Northeastern states contain the most overrepresented groups of lower caste prisoners.

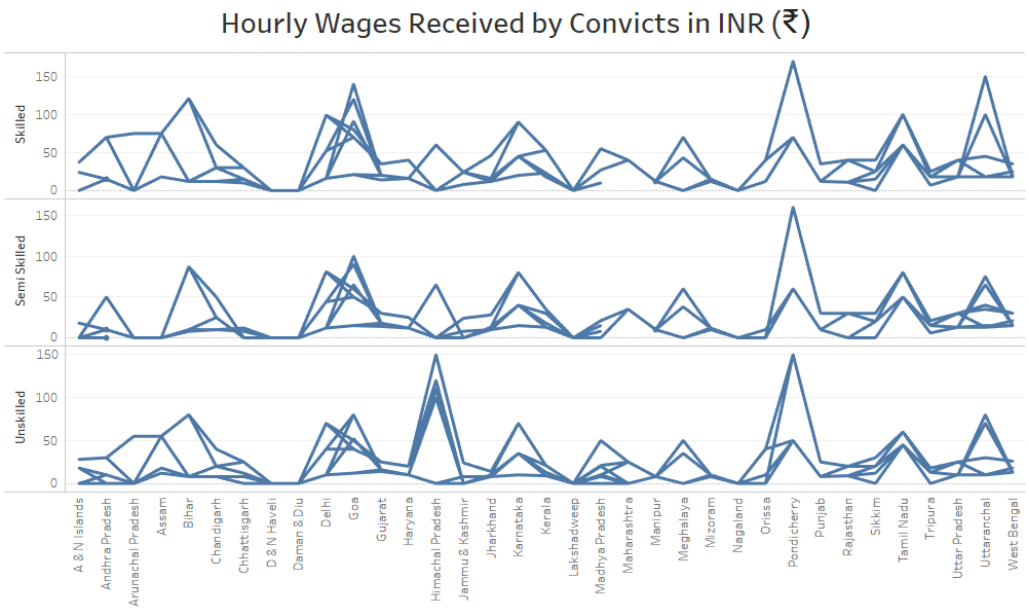


Recidivism rate is an indicator of how well the prison system functions: the lower the rate of reoffenders, the better the prison. As such, the worst offenders in this category are the states of Meghalaya, Tamil Nadu, Mizoram, Kerala, Nagaland, and Sikkim. Interestingly, with the exception of Tamil Nadu, these states have relatively small prison populations, indicating more so that these recidivists may be habitual reoffenders rather than a flawed prison system.



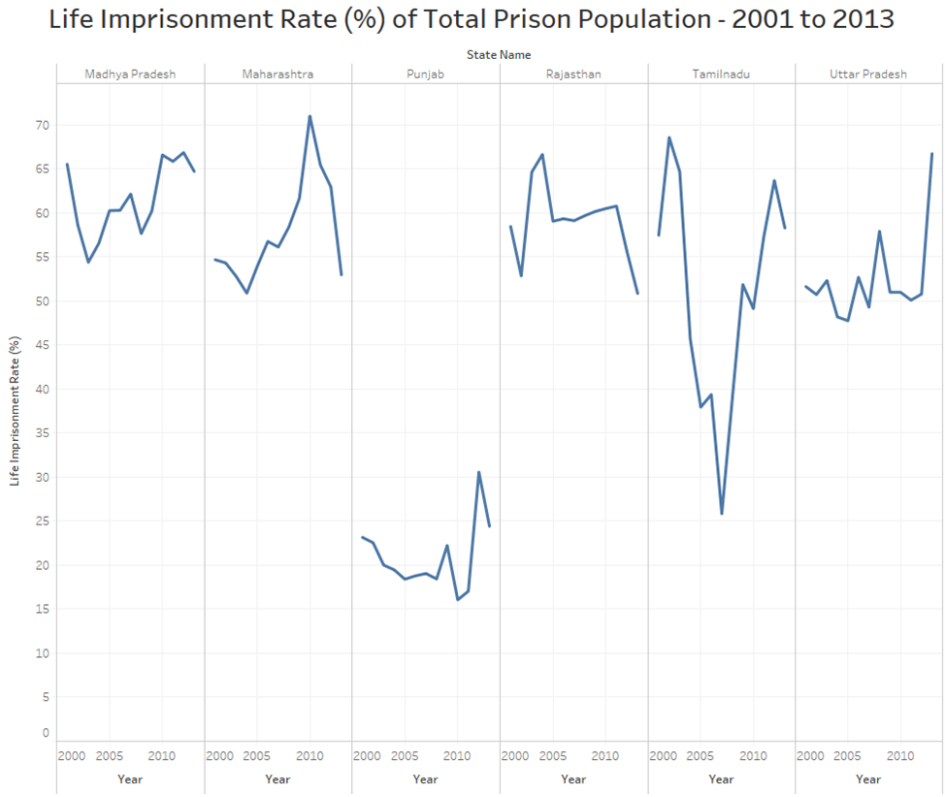
It's also interesting to analyze the age-group makeup of the prisoners in the different states. Right away, the most visible pattern is that the majority of prisoners aged 16-18 are sentenced to less than one year in prison, which is understandable because of the leniency that tends to be

awarded to that age group. Interestingly, the majority of prisoners in all other age groups have been sentenced to life imprisonment, which is quite harsh.



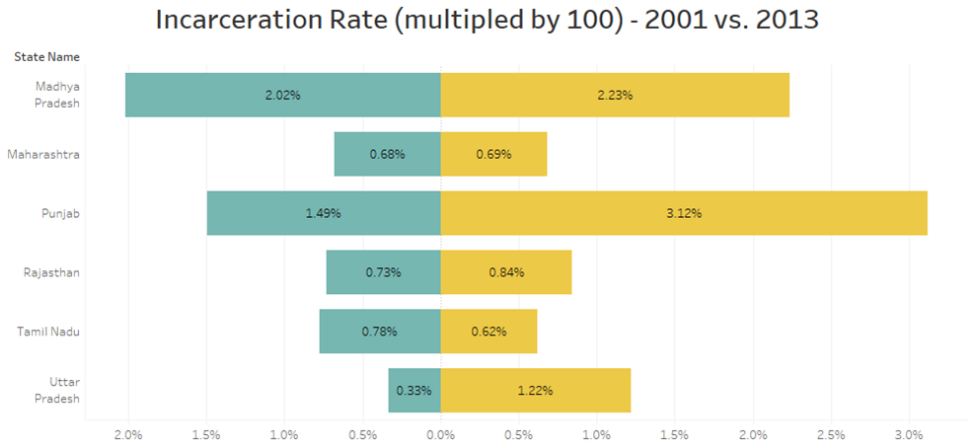
The wages a prisoner receives for working while in prison can prove to be a useful source of supplemental income. In the figure above, the most generous states are Himachal Pradesh, Pondicherry, Tamil Nadu, and Uttarakhand. This pattern applies for the three different types of labor.

For the next two visualizations, the states with the highest net prisoner count are analyzed.



The proportion of prisoners imprisoned for life in a state can be used to gauge how severe a criminal justice system is. In this case, some interesting trends can be observed: Maharashtra and Rajasthan are becoming increasingly lenient in terms of issuing life imprisonment sentences. Tamil Nadu experienced both a steep drop and uptick in prisoners incarcerated for life. Uttar Pradesh looks to be increasing its share, while Madhya Pradesh looks fairly constant. However, the most obvious outlier is Punjab, which has a very low share of convicts imprisoned for life as compared to the other states, and it's angled to keep dropping (despite experiencing a sharp uptick from 2010-2012). This could indicate a society where serious crime is relatively low and may even point to lenient/corrupt elements in the justice system.

To further understand growth/decline of prison population, it is useful to analyze the proportion of incarcerated relative to the total population of the state, and how this proportion has changed from 2001 to 2013. This phenomenon is exhibited:



The two states with the biggest increase in prisoner populations are Punjab and Uttar Pradesh, with the latter recording nearly a four-fold increase. Only Tamil Nadu has lowered proportion out of all the states, with Madhya Pradesh and Rajasthan accruing a modest increase, and Maharashtra staying almost the same.

Now that some interesting trends have been spotted within a narrow selection of the dataset, further analysis using data mining techniques are ready to be conducted, in the hope that interesting correlations may be found and exploited to uncover other patterns within the prison system in India.

Another crucial metric to analyze is the overrepresentation of religious minorities; in India's case, the Muslim population has often been subject to social discrimination. The heat map above shows that Muslim prisoners are overrepresented with respect to their general population in 14 states. Interestingly, Hindus are very overrepresented in Jammu & Kashmir, the only Muslim majority state in India. These metrics are especially important because they reflect the social prejudices that exist in different areas of India and how they seep into the judicial system.

III. Data Preparation and Preprocessing

To analyze data meaningfully, one must first clean and preprocess data better. As an exploratory foray into finding correlations between variables in the data set, two possible models are considered:

1) Correlation between Recidivism and Education Facilities in Prisons
2) Correlation between Violence in Prisons and Total Expenditure

An example of data wrangling for this study is from the correlation between recidivism: to glean useful information from the different raw data points, different variables reflecting different education rates are required. In the original data set, variables provided were the number of prisoners enrolled in different levels of education within the prison system. While important, this variable selection was deemed unsuitable for the goals of this study; it was decided that the *education rates*, i.e., the proportion of convicts in different educational stages (relative to total number of students) was more important. As such, different variables were constructed (elementary_education_rates, adult_education_rates, etc.) by applying a simple series of calculations for the different variable columns.

For the second correlation analysis, variable extrapolation was undertaken since the given dataset only contained values for total prison expenditures up to 2011. Using an average rate of growth (as reflected by past data), expenditures were extrapolated for 2012 and 2013 for the six different states. Once this was completed, further data processing was done wherein two variables were constructed (sum_incidents, sum_inmate_injuries) to paint a more composite picture of incidents and inmate injuries, which hitherto contained multiple types of incidents and injuries respectively. By obtaining a summation of the two variables for 2001-2013 in the six states, it was possible to compare aggregate values with the expenditure in each prison.
A note about data processing: although the dataset is extremely rich and contains a wealth of information, the variables for the most part weren't consistent, requiring a large deal of extrapolation and data wrangling. As such, an element of human input bias must be accounted for in the above models.

IV. Data Mining Techniques and Implementation

For this study, a classification-based approach shows the most promise, since it enables an analysis of the prison population makeup and to show if certain variables, such as expenditure, education, etc. have a proportional response in terms of inmate well-being. The data set is well structured to pursue such an analysis, given a uniform time frame with a high variable count. As such, the following describes the first problem discussed.

1) Correlation between Recidivism and Education Facilities in Prisons

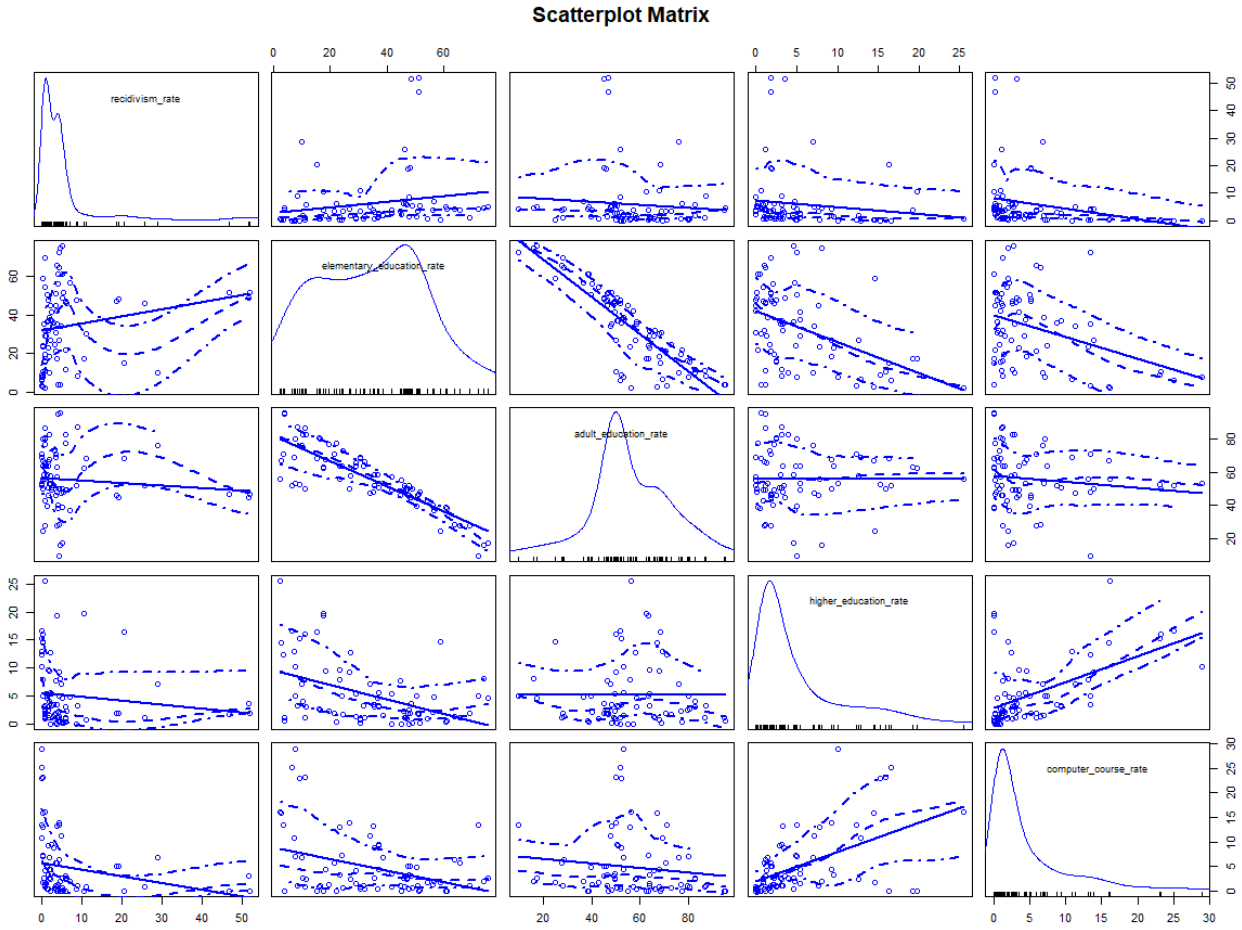


Figure 6. Scatterplot Matrix of Recidivism Rate vs Varying Education Opportunities (Cumulative – All States)

From this first level scatterplot matrix, it can be observed that recidivism rates are generally lowered with increased education levels (this isn't true for elementary education rates, interestingly). Another interesting phenomenon is the strongly negative correlation between elementary education and adult education rates; this can be explained in that as more prisoners enroll in adult education courses, it highly likely that they have completed elementary education previously.

Applying a simple multiple linear regression yields the following:

```
Call:
lm(formula = recidivism_rate ~ elementary_education_rate + adult_education_rate +
    higher_education_rate + computer_course_rate, data = corr_recidivism_education)

Residuals:
   Min     1Q Median     3Q    Max
-9.121 -4.769 -3.210  0.098 43.980

Coefficients: (1 not defined because of singularities)
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -26.5273    21.3743  -1.241   0.2185
elementary_education_rate   0.3871     0.2197   1.762   0.0822 .
adult_education_rate        0.3140     0.2202   1.426   0.1580
higher_education_rate       0.3668     0.4086   0.898   0.3723
computer_course_rate            NA         NA      NA       NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.28 on 74 degrees of freedom
Multiple R-squared:  0.06585,   Adjusted R-squared:  0.02798
F-statistic: 1.739 on 3 and 74 DF,  p-value: 0.1664
```

This early model isn’t as accurate as hoped, given the relatively low Adjusted R-squared value of 0.02798. While this doesn’t indicate a total dismissal of the model, it clearly isn’t very strong. To get a better idea of the viability of this model, the following graphs can be produced:

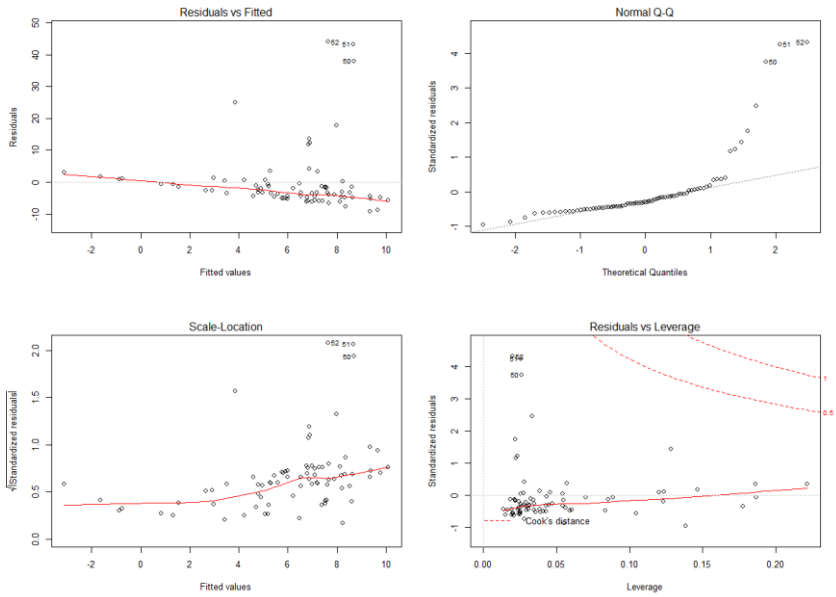


Figure 7. Plots of Linear Regression Model (Recidivism and Education)

From the Normal Q-Q Plot alone, the data diverges too much and indicates poor confidence in the model.

Applying a second order regression model, the following is obtained:

```
Call:
lm(formula = recidivism_rate ~ (elementary_education_rate + adult_education_rate +
    higher_education_rate + computer_course_rate)^2, data = corr_recidivism_education)

Residuals:
   Min     1Q Median     3Q    Max
-9.924 -5.140 -2.331  1.184 43.337

Coefficients: (1 not defined because of singularities)
                                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                                   120.580645 233.204363   0.517    0.607
elementary_education_rate                      -1.083086   2.326099  -0.466    0.643
adult_education_rate                           -1.221279   2.327291  -0.525    0.601
higher_education_rate                          -4.095136   4.471260  -0.916    0.363
computer_course_rate                                 NA         NA      NA       NA
elementary_education_rate:adult_education_rate  0.002084   0.003386   0.616    0.540
elementary_education_rate:higher_education_rate 0.018432   0.043662   0.422    0.674
elementary_education_rate:computer_course_rate -0.017000   0.026785  -0.635    0.528
adult_education_rate:higher_education_rate      0.045303   0.052566   0.862    0.392
adult_education_rate:computer_course_rate      -0.021697   0.029526  -0.735    0.465
higher_education_rate:computer_course_rate      0.017937   0.095087   0.189    0.851

Residual standard error: 10.49 on 68 degrees of freedom
Multiple R-squared:  0.1063,    Adjusted R-squared:  -0.01196
F-statistic: 0.8989 on 9 and 68 DF,  p-value: 0.5313
```

While multiple R-squared value has been improved, the adjusted R-squared value has become worse, indicating that this model falls short as well. The following plots are obtained:

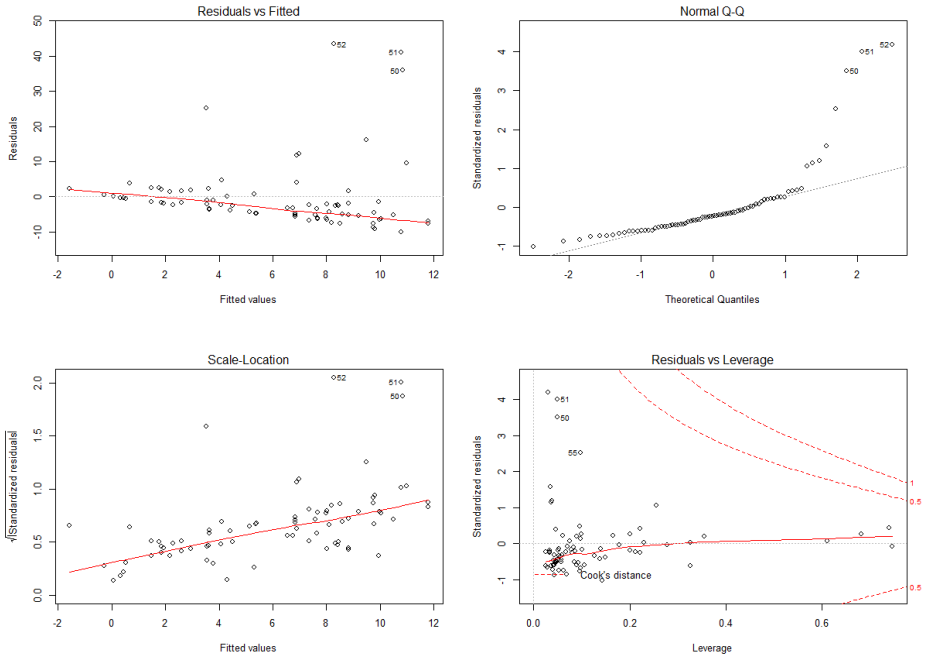


Figure 8. Plots of Second Order Regression Model (Recidivism and Education)

These plots are again indicative of a poorly calibrated model, one that will not be as useful as hoped, and whose results won't necessarily reflect accurate predictions.

3) Correlation between Violence in Prisons and Total Expenditure

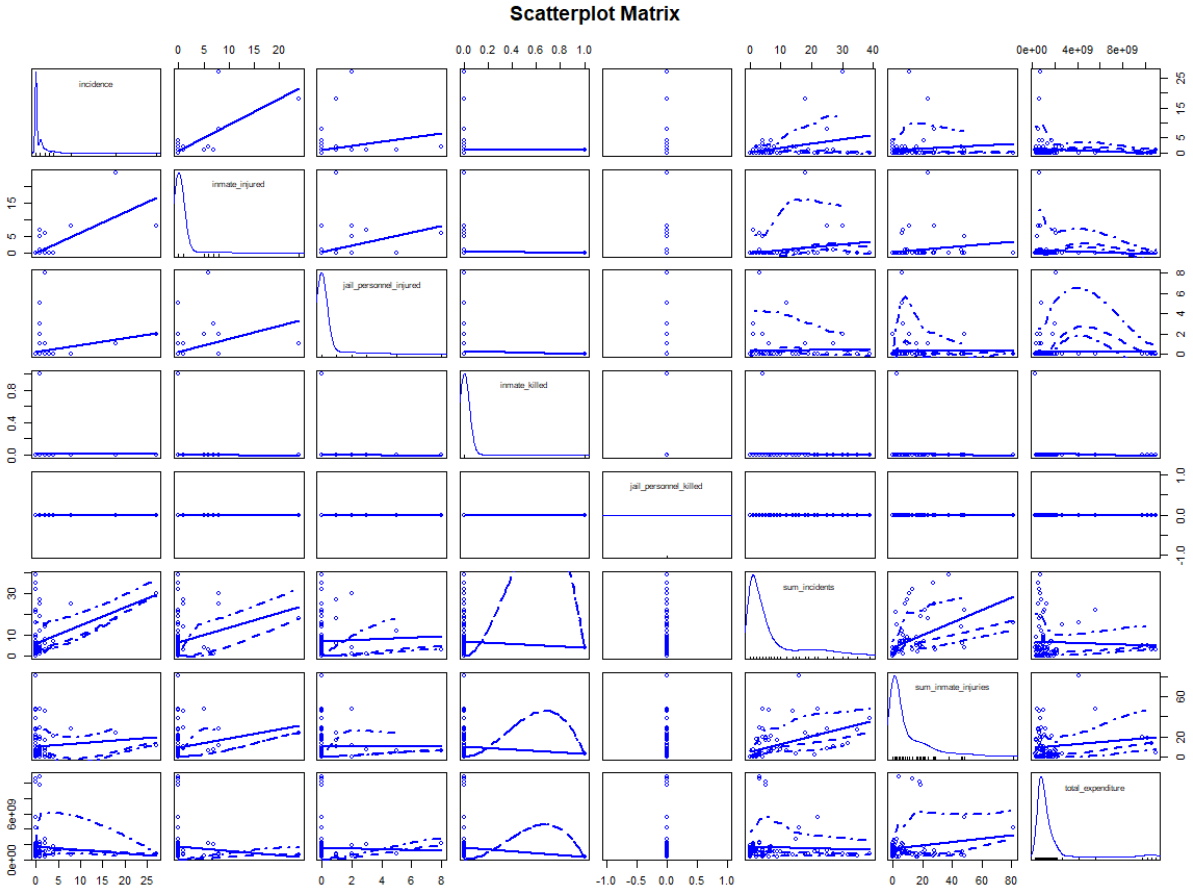


Figure 9. Scatterplot Matrix of Violent Incidents and Total Expenditure in Prison (Cumulative - All States)

Unfortunately, unlike the previous scatterplot matrix, this isn't as indicative; the only meaningful (if weak) correlation that can be observed is that of a lowered rate of total summed incidents with increased prison expenditure. Applying multiple linear regression to the model, the following results were obtained:

```
Call:
lm(formula = total_expenditure ~ incidence + inmate_injured +
    jail_personnel_injured + inmate_killed + jail_personnel_killed +
    sum_incidents + sum_inmate_injuries, data = corr_expend_tranquility)

Residuals:
       Min         1Q     Median         3Q        Max
-2.298e+09 -8.726e+08 -4.804e+08  3.084e+07  9.295e+09

Coefficients: (1 not defined because of singularities)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.565e+09  3.408e+08   4.592 1.88e-05 ***
incidence             1.876e+07  1.043e+08   0.180   0.8577
inmate_injured       -7.849e+07  1.255e+08  -0.626   0.5337
jail_personnel_injured  6.389e+06  2.346e+08   0.027   0.9784
inmate_killed        -1.247e+09  2.259e+09  -0.552   0.5825
jail_personnel_killed         NA         NA      NA       NA
sum_incidents        -3.584e+07  3.366e+07  -1.065   0.2906
sum_inmate_injuries   3.542e+07  2.006e+07   1.766   0.0818 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.239e+09 on 70 degrees of freedom
Multiple R-squared:  0.05502,   Adjusted R-squared:  -0.02598
F-statistic: 0.6793 on 6 and 70 DF,  p-value: 0.6668
```

As expected, the linear regression model is very weak, and will likely produce substandard results, which is further evidenced by the following graphs in Figure 10:

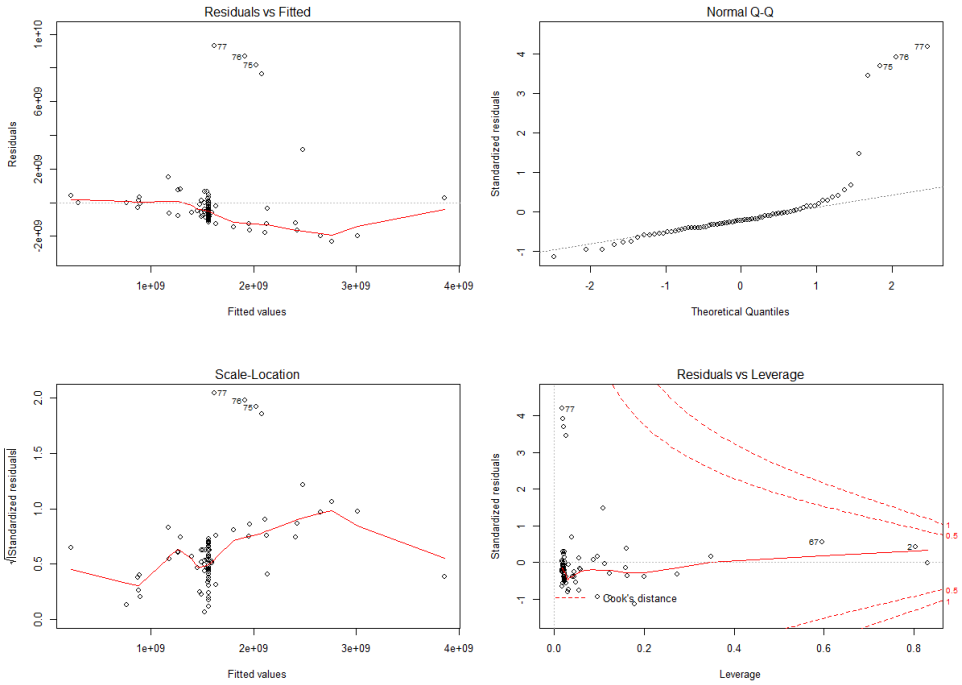


Figure 10. Plots of Multiple Linear Regression (Prison Expenditure and Tranquility)

Flowchart

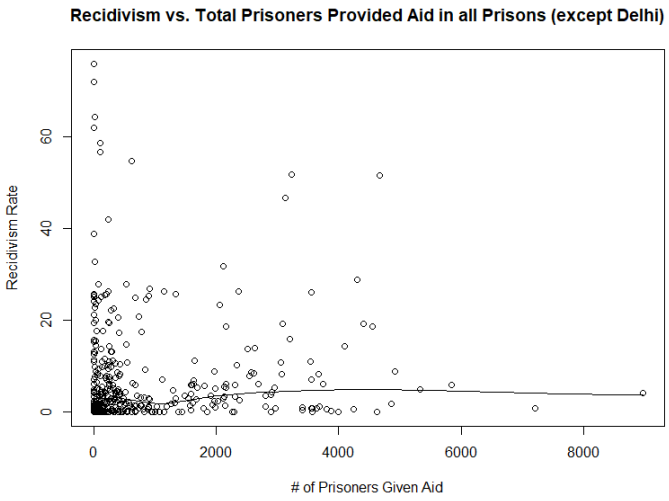
A basic mapping of the techniques used to develop models in this study is shown below:

NOTE: The data modelling that has been undertaken so far indicates that the models are not strong. As the case study progresses, more refinements will be made to ensure that thorough analysis yielding useful data can be achieved.

V. Performance Evaluation

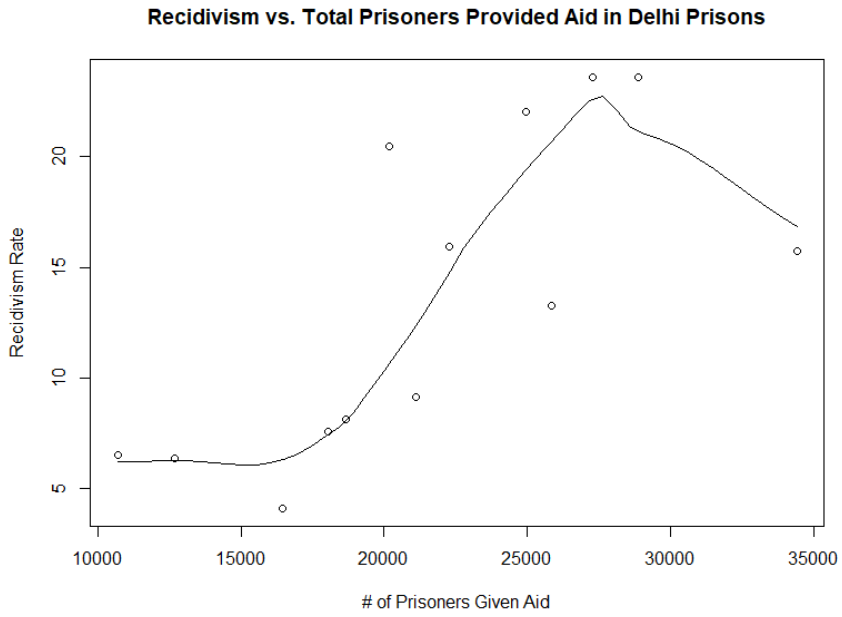
Given that the previous models didn't work so well, two new sets of data were considered. If they fail, regression diagnostics will be performed on them to determine how robust the linear regression models are. For all regression models, the datasets are split into 80% training data, and 20% validation data.

In the following regression model, the relationship between the recidivism rate of prisoners and the total number of those that received aid is considered. The hypothesis: the more aid prisoners receive, the less likely they will be to commit crimes again in the future. As a result, the two variables will be inversely proportional. To begin, a simple scatterplot matrix is drawn to establish if this relationship is true.



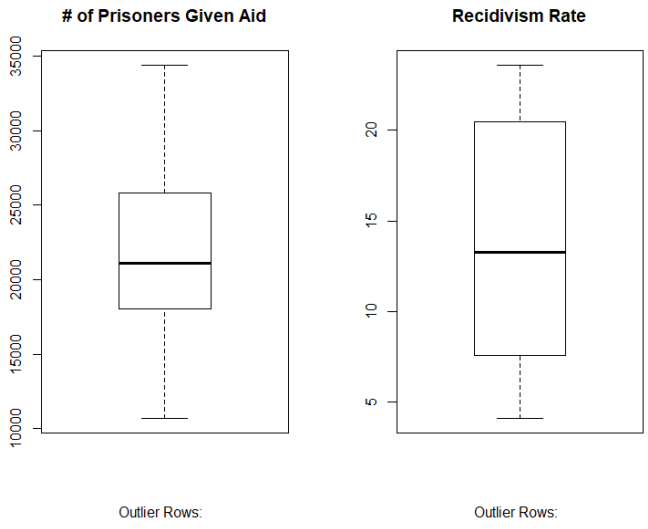
Figure

This graph shows that the hypothesis isn't satisfied, and that there's a miniscule relationship between the data (if at all). Therefore, it was decided to focus on one of the states with the highest rates of aid for the prisoners, Delhi. This yielded a far more interesting relationship, as shown below.

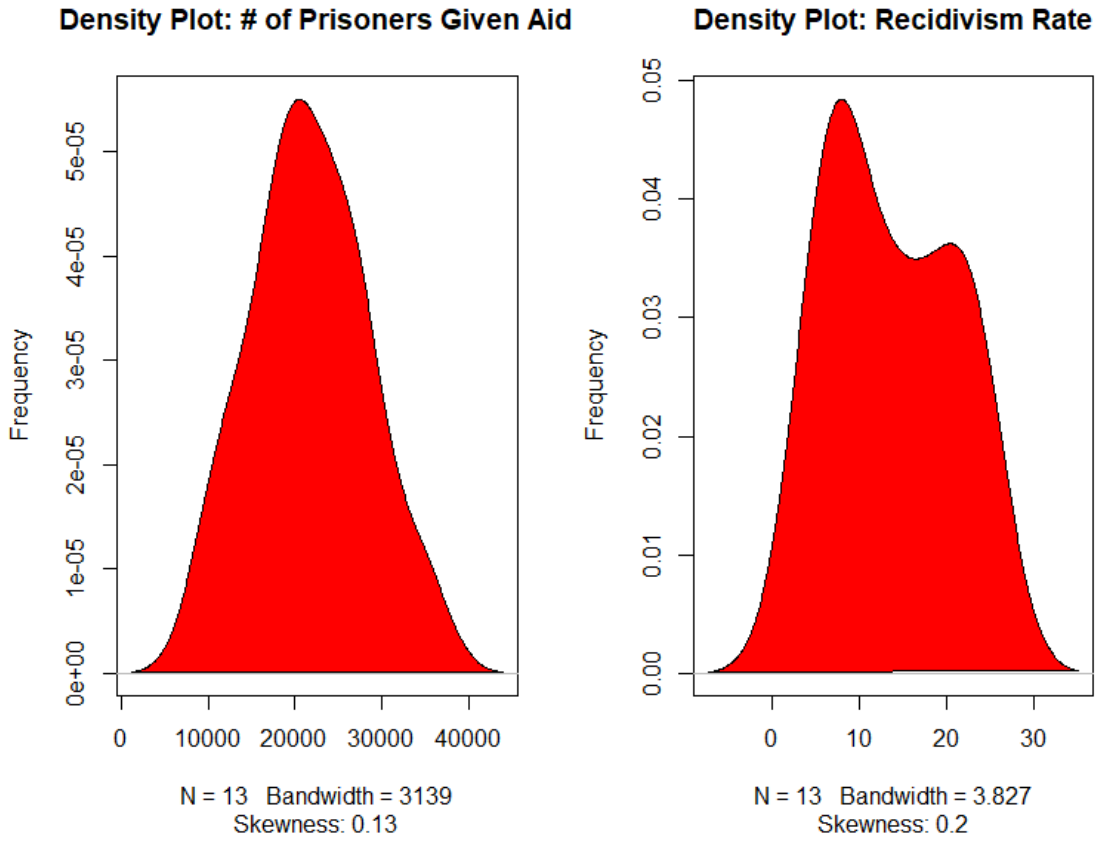


Rather than being an inversely proportional relationship, it's almost a linearly proportional one. This poses very interesting ramifications, because it indicates that despite prisoners given a lot of aid, recidivism rates actually *increase*. This could point to corruption within the prison system itself, ex. Wardens and other staff pocketing the aid for themselves and not distributing it to the prisoners. It could also mean that the hypothesis of expecting a decrease in recidivism with increased aid is wrong, which is entirely possible. Nonetheless, further data mining was conducted to try to predict the state of the relationship between recidivism and aid.

Next, a box-and-whisper plot is shown to spot any obvious outliers in the dataset.



Next, a density plot is used to establish the normality of the data distribution. The aid data is fairly normal, but the recidivism rate is bimodal.



Now, a linear model is built using the full data set.

```
Call:
lm(formula = total ~ recidivism_rate, data = rehab_reci_delhi)

Coefficients:
    (Intercept)  recidivism_rate
        12559.0            669.9
```

Following this, we must diagnose the model to ensure that it is statistically significant. The following shows the summary of the linear model.

```
Call:
lm(formula = total ~ recidivism_rate, data = rehab_reci_delhi)

Residuals:
   Min     1Q Median     3Q    Max
 -6236  -2330    397   1137  11329

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     12559.0     2983.0   4.210  0.00146 **
recidivism_rate   669.9      196.3   3.412  0.00580 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4830 on 11 degrees of freedom
Multiple R-squared:  0.5142,    Adjusted R-squared:   0.47
F-statistic: 11.64 on 1 and 11 DF,  p-value: 0.005804
```

```
> t_value
[1] 3.41209
> p_value
[1] 0.02697549
> f
   value    numdf    dendf
11.64236  1.00000 11.00000
> model_p
      value
0.005803765
> AIC(linearMod)
[1] 261.27
> BIC(linearMod)
[1] 262.9649
```

Looking at the R-square value, which has a value of 0.5142, it can be said that there is a moderately high degree of correlation between the two variables. Since the t-value is not especially high, but the p-value is lower than 0.05, it indicates that the model is somewhat robust and statistically significant. AIC and BIC were calculated for model comparison in the future.

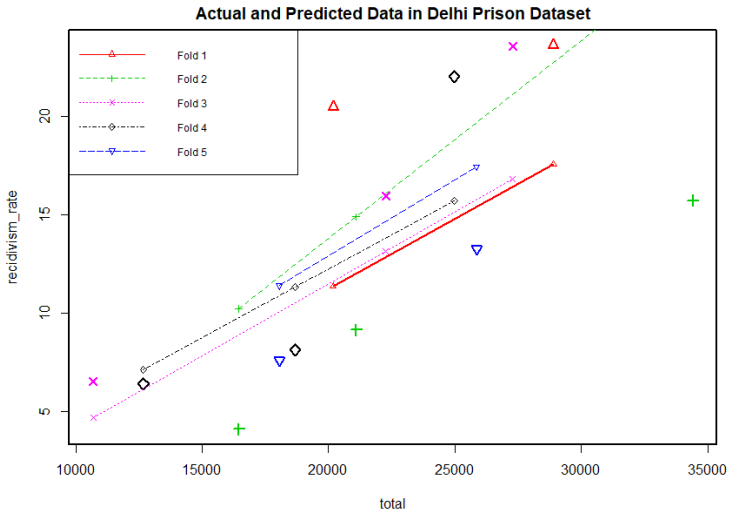
```
> summary(lmMod)

Call:
lm(formula = total ~ recidivism_rate, data = trainingData)

Residuals:
    Min      1Q  Median      3Q     Max
-7035.9 -1498.0  -109.9  1440.3 10493.7

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      13017.5     3834.5   3.395  0.00943 **
recidivism_rate    693.8      276.7   2.507  0.03652 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5324 on 8 degrees of freedom
Multiple R-squared:  0.4401,    Adjusted R-squared:  0.3701
F-statistic: 6.287 on 1 and 8 DF,  p-value: 0.03652

> AIC(lmMod)
[1] 203.7463
```

The model has an R-squared value of 0.4401, which indicates moderately strong positive correlation between the variables. The p-value is less than 0.05, indicating that it is a statistically sound model. AIC was calculated for model comparison in the future.

```
> head(actuals_preds)
  actuals predicteds
2   12674   17457.94
8   27283   29370.71
9   24973   28288.36
> min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds, 1, max))
> mape <- mean(abs((actuals_preds$predicteds - actuals_preds$actuals))/actuals_preds$actuals)
> mape
[1] 0.1955797
```

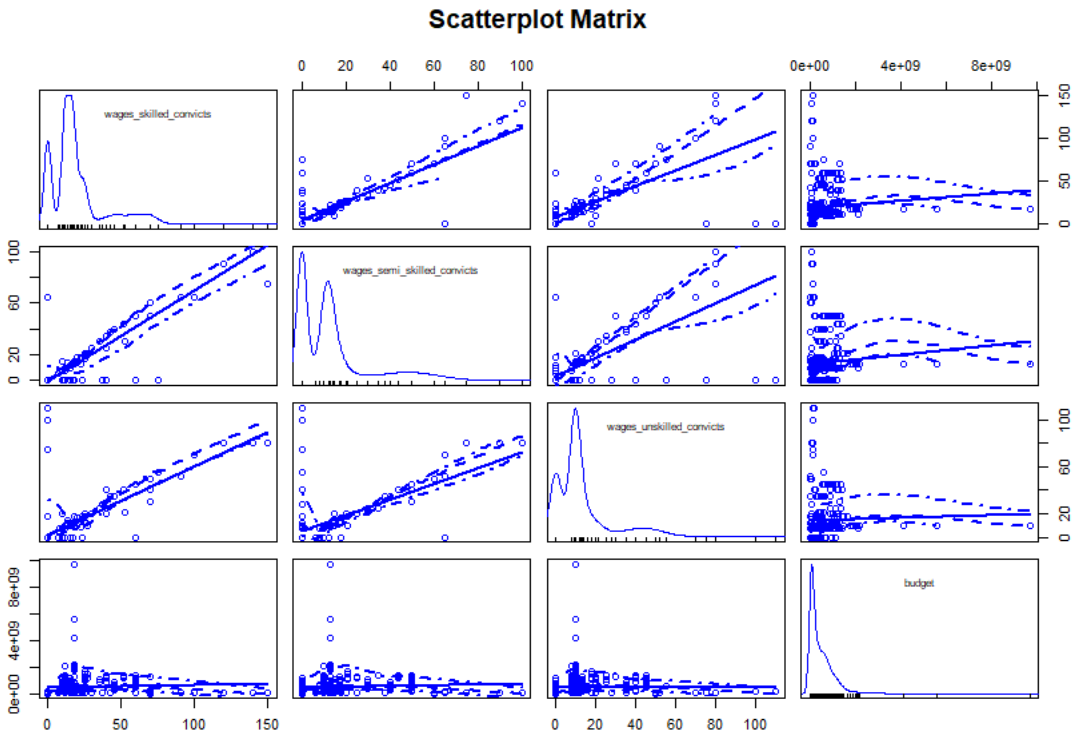
The mean absolute percentage error is 0.19558 or around 19.6%, and the Min-Max accuracy is around 68%.

Next, the k-Fold Cross validation process is undertaken to test the model's rigor.

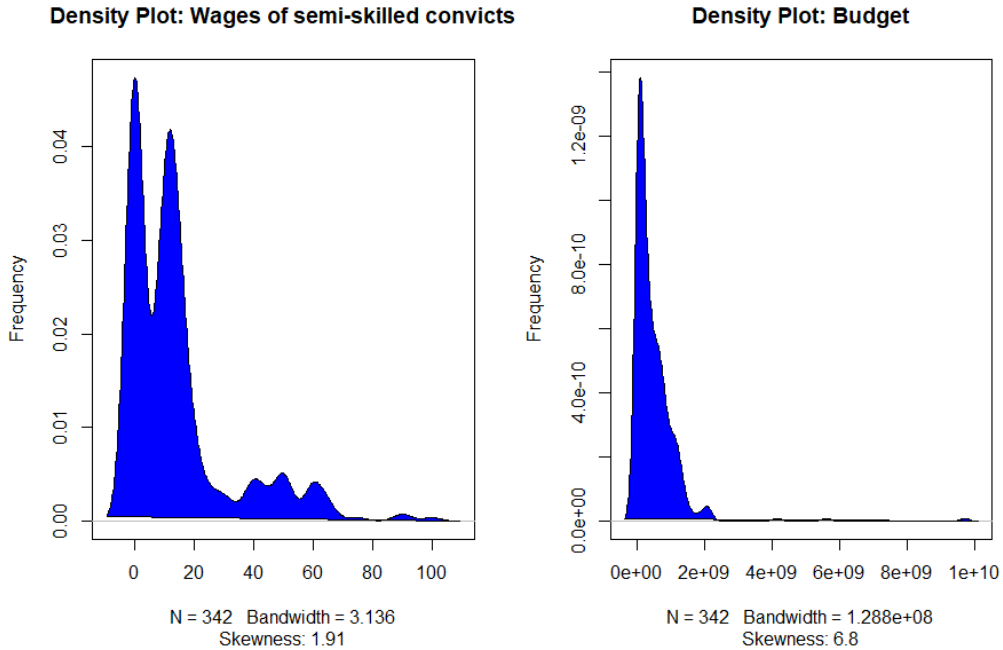


This shows that the model isn't as robust as we hoped (since the lines aren't very close to each other and aren't all parallel), but it still possesses characteristics that render it useful for a first-order analysis of the Delhi prison system. As stated before, the hypothesis that was presented was rendered false by the scatterplot as well as by the model developed in this section.

Since there was moderate success achieved with the previous linear regression model, the same approach was decided to be applied to another phenomenon: the prison budget's impact on the wages paid to prisoners. The following scatterplot matrix shows the trend of three different types of wages: unskilled, semi-skilled, and skilled vs. prison budgets for multiple states.



From this matrix, the last column's trends are analyzed: there's a slightly strong correlation for the skilled convicts, a moderately strong correlation for the semi-skilled convicts, and a very weak correlation for unskilled convicts (all positive). As such, the semi-skilled wages and budget relationship will be focused on since it appears to be the most promising trend. Next, the density plots for both are plotted.



These density plots show that both sets of data are extremely skewed, and in the case of the wages of semi-skilled convicts: not normalized at all.

```
> print(linearMod2)

Call:
lm(formula = wages_semi_skilled_convicts ~ budget, data = budget_wages)

Coefficients:
(Intercept)       budget
  1.314e+01    1.797e-09

> summary(linearMod2)

Call:
lm(formula = wages_semi_skilled_convicts ~ budget, data = budget_wages)

Residuals:
    Min      1Q  Median      3Q     Max
-17.578 -13.192  -3.664   1.657  86.810

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.314e+01  1.147e+00  11.455   <2e-16 ***
budget      1.797e-09  1.291e-09   1.393    0.165
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.99 on 340 degrees of freedom
Multiple R-squared:  0.005671,  Adjusted R-squared:  0.002747
F-statistic: 1.939 on 1 and 340 DF,  p-value: 0.1647
```

Although it appeared that the scatterplot matrix showed a relatively strong correlation between the two variables, it appears that the data is very slightly correlated, as evidenced by the R-squared values. As such, this does not bode well for the model's success.

```
> AIC(linearMod2)
[1] 2951.287
> BIC(linearMod2)
[1] 2962.792
> t_value2
[1] 1.392542
> p_value2
[1] 0.1646792
> f
NULL
```

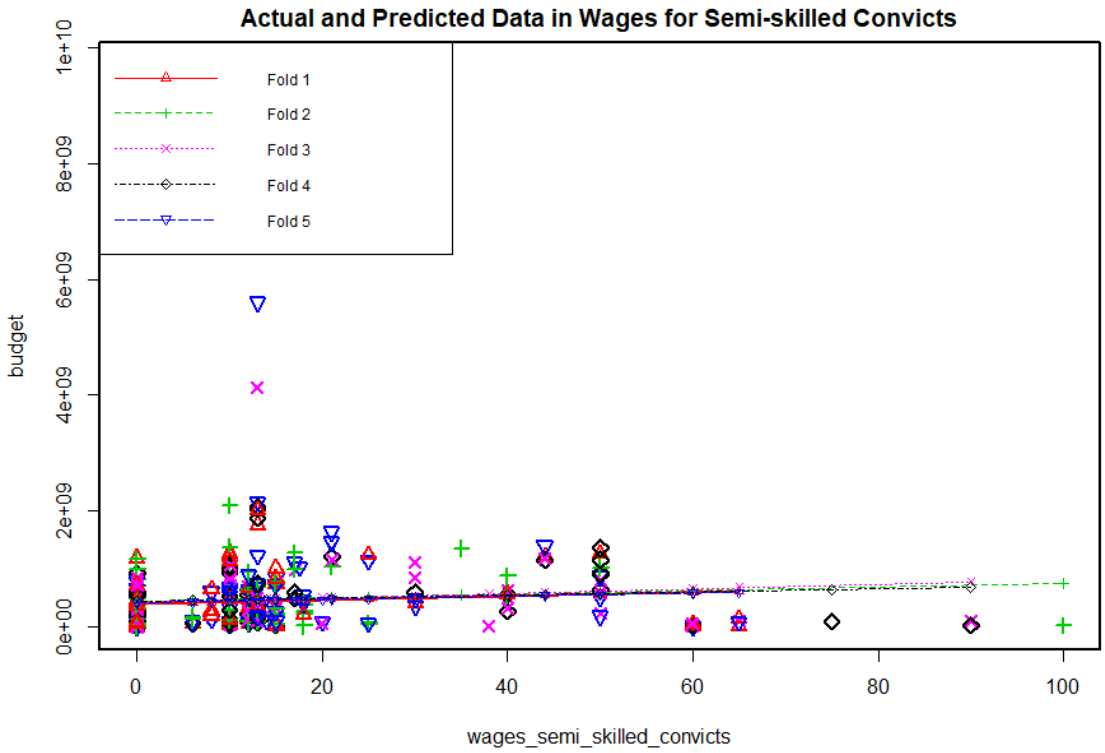
This sentiment is further proved by the t-value of 1.392542 (quite low) and p-value of 0.1646792, which is higher than the 0.05 threshold that is generally accepted. As such, it is almost certain that this model is not as robust as required.

```
> summary(lmMod2)

Call:
lm(formula = wages_semi_skilled_convicts ~ budget, data = trainingData2)

Residuals:
    Min      1Q  Median      3Q     Max
-15.047 -13.426  -3.561   1.244  86.573

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.339e+01  1.288e+00  10.392   <2e-16 ***
budget      1.413e-09  1.376e-09   1.026    0.306
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.35 on 271 degrees of freedom
Multiple R-squared:  0.003872,  Adjusted R-squared:  0.0001962
F-statistic: 1.053 on 1 and 271 DF,  p-value: 0.3056

> AIC(lmMod2)
[1] 2367.396
```

Seeing how high the AIC value is, it wouldn't be surprising if quite a few other models were better than this one, given that a lower value indicates a better model

```
> head(actuals_preds2)
   actuals predicteds
1        0   13.41265
2        0   13.42081
11      10   14.41944
13       0   15.06015
14      10   14.70902
15       0   14.70269
> min_max_accuracy2 <- mean(apply(actuals_preds2, 1, min) / apply(actuals_preds2, 1, max))
> mape2 <- mean(abs((actuals_preds2$predicteds - actuals_preds$actuals))/actuals_preds2$actuals)
> mape2
[1] Inf
> min_max_accuracy2
[1] 0.3911101
```

Defining the model's untenability is the MAPE value of infinity; this is most likely due to a set of predictions being compared with an actual value of zero. This is indicative of a high degree of volatility in the predictive model as compared to the actual values.



Finally, the k-Fold Cross validation graph showcases the extremely weak correlation between the two data sets, and also shows the weakness of the model, with many outliers and trend lines being too close to each other.

VI. Discussion and Recommendation

The approach that was used followed a general trend of identifying trends from applying visualization techniques to the different data sets and then trying to implement models based on them. These models were regression-based and attempted to chart a relationship between two different variable types. However, this approach failed to yield satisfactory results, producing models that either didn't work or were only moderately successful. The data set that was used can be thought of as a double-edged sword: although extremely rich with very useful data and variables, it was often messy and required large amounts of data wrangling in order to produce datasets that were compatible with each other. In addition, there was a high amount of extrapolation required for some variable types, since a lot of the smaller states had misleading or missing data. This was a constant challenge throughout the process. In the future, it would be worthwhile pursuing a single year, or a smaller series of years, in order to chart meaningful trends. Another improvement to make would be to try to integrate most, if not all, of the data and variables into one single CSV file. This would make it much easier to work with. Given the accelerated time frame of the course and the large dataset worked with, we feel that we have definitely extracted some meaningful and exciting information about the Indian prison population that can be built upon for future academic/social research. Given more time, we would have analyzed more varied data sets with different models and engaged in creating more sophisticated data processing models.

VII. Summary

Analyzing Indian prison system data from 2001 to 2013 yielded fascinating insights into dominant and varied trends. Visualizations show a kaleidoscope of different patterns in the prison populations, such that most prisoners are poorly educated, are overrepresented religious minorities, and are predominantly lower caste. These are indicative of broad social inequalities in Indian society, and require urgent progress to be made in social, economic, and judicial spheres. This study also attempts to showcase the relationship between the incarcerated population and the judicial system in a new light, and to establish reasonable correlations between different metrics such as prison budget, prisoner recidivism rates, prisoner education, prison tranquility, etc. Although the models applied weren't robust enough (for the most part) to indicate a strong relationship between the variables, this may be indicative of weak modelling, and not necessarily that there isn't a relationship. Data mining is the process of extracting patterns from vast data sets, and it may just be that this study dug in the wrong direction!

VIII. References:

- Dataset: https://www.kaggle.com/rajanand/prison-in-india/version/3
- India Census 2001 Data
- India Census 2011 Data
- Handbook on Social Welfare Statistics
- Demographic Status of Scheduled Tribe Population of India

Appendix: R Code for use case study

Please show the R code you generated for the use case study. Please do not show results here, only the code.

```
## Correlation between Violent Incidents and Total Expenditure in Prisons

corr_expend_tranquility <- read_excel("exp_tranquil.xlsx")
summary(corr_expend_tranquility)
scatterplotMatrix(corr_expend_tranquility[,c(4:11)], spread = FALSE,
lty.smooth = 2, main = "Scatterplot Matrix")
names(corr_expend_tranquility)

## Recidivism and Education in Prisons
install.packages("readxl")
library(readxl)
getwd()
setwd("C:/Users/skyeshwin/Documents/Course Material/Summer 2018/IE7275 37152
Data Mining in Engg - Summer 2018/Project/correlation")
corr_recidivism_education <- read_excel("recidivism_education.xlsx")
install.packages("car")
library(car)
scatterplotMatrix(corr_recidivism_education[,c(5,11:14)], spread = FALSE,
lty.smooth = 2, main = "Scatterplot Matrix")
summary(corr_recidivism_education)
names(corr_recidivism_education)
fit.linear <- lm(recidivism_rate ~ elementary_education_rate +
adult_education_rate + higher_education_rate + computer_course_rate, data =
corr_recidivism_education)
summary(fit.linear)
fit.second_order <- lm(recidivism_rate ~ (elementary_education_rate +
adult_education_rate + higher_education_rate + computer_course_rate)^2, data
= corr_recidivism_education)
summary(fit.second_order)
par(mfrow = c(2,2))
plot(fit.second_order)
crPlots(fit.linear)
qqPlot(fit.linear)
par(mfrow = c(1,1))
install.packages("gvlma")
library(gvlma)
summary(gvlma(fit.linear))
install.packages("glmnet")
library(glmnet)
vif(fit.second_order)
summary(fit.linear)
alias(lm(fit.second_order))
outlierTest(fit.linear)
outlierTest(fit.second_order)


## Indian states and their religious representation of convicts in prisons

#Heatmap
install.packages("ggplot2")
```

```
install.packages("scales")
install.packages("reshape")
install.packages("plyr")
install.packages("TDMR")
library(ggplot2)
library(reshape)
library(plyr)
library(TDMR)
library(scales)
setwd("C:/Users/skyeshwin/Documents/Course Material/Summer 2018/IE7275 37152
Data Mining in Engg - Summer 2018/Project/Indian Prison Statistics (2001 -
2013) Dataset")
religion <- read.csv("religion_ratio_2.csv")
religion$State.Name <- with(religion, reorder(State.Name, muslim_ratio))
religion.m <- melt(religion)
religion.m <- ddply(religion.m, .(variable), transform, rescale =
scale(value))
p <- ggplot(religion.m, aes(variable, State.Name)) + geom_tile(aes(fill =
rescale), colour = "white") + scale_fill_gradient(low = "green", high =
"red")
base_size <- 9
p + ggplot2::theme_grey(base_size = base_size) + labs(x = "", y = "") +
scale_x_discrete(expand = c(0, 0)) + scale_y_discrete(expand = c(0, 0)) +
labs(legend.position = "none", axis.ticks = element_blank(), axis.text.x =
element_text(size = base_size * 0.8, angle = 330, hjust = 0, colour = "red"))


## Indian states and their caste representation of convicts in prisons

#Heatmap
caste <- read.csv("caste_ratio2.csv")
caste$State.Name <- with(caste, reorder(State.Name, sc_ratio))
caste.m <- melt(caste)
caste.m <- ddply(caste.m, .(variable), transform, rescale = scale(value))
p <- ggplot(caste.m, aes(variable, State.Name)) + geom_tile(aes(fill =
rescale), colour = "white") + scale_fill_gradient(low = "white", high =
"blue")
base_size <- 9
p + ggplot2::theme_grey(base_size = base_size) + labs(x = "", y = "") +
scale_x_discrete(expand = c(0, 0)) + scale_y_discrete(expand = c(0, 0)) +
labs(legend.position = "none", axis.ticks = element_blank(), axis.text.x =
element_text(size = base_size * 0.8, angle = 330, hjust = 0, colour = "red"))


## Rehab vs Recidivism rate in other states

rehab_reci_other <- read.csv("reci_rehab_other_states.csv")
scatter.smooth(x=rehab_reci_other$total, y=rehab_reci_other$recidivism_rate,
xlab = "# of Prisoners Given Aid", ylab = "Recidivism Rate", main =
"Recidivism vs. Total Prisoners Provided Aid in all Prisons (except Delhi)")


## Rehab vs Recidivism rate in New Delhi

rehab_reci_delhi <- read.csv("reci_rehab_delhi.csv")
```

```
scatter.smooth(x=rehab_reci_delhi$total, y=rehab_reci_delhi$recidivism_rate,
xlab = "# of Prisoners Given Aid", ylab = "Recidivism Rate", main =
"Recidivism vs. Total Prisoners Provided Aid in Delhi Prisons")

#Scatterplot
par(mfrow=c(1,2))
boxplot(rehab_reci_delhi$total, main = "# of Prisoners Given Aid", sub =
paste("Outlier Rows: ", boxplot.stats(rehab_reci_delhi$total)$out))
boxplot(rehab_reci_delhi$recidivism_rate, main = "Recidivism Rate", sub =
paste("Outlier Rows: ", boxplot.stats(rehab_reci_delhi$recidivism_rate)$out))

#Density Plot
library(e1071)
par(mfrow=c(1, 2))  # divide graph area in 2 columns
plot(density(rehab_reci_delhi$total), main="Density Plot: # of Prisoners
Given Aid", ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(rehab_reci_delhi$total), 2)))  # density plot for
'total'
polygon(density(rehab_reci_delhi$total), col="red")
plot(density(rehab_reci_delhi$recidivism_rate), main="Density Plot:
Recidivism Rate", ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(rehab_reci_delhi$recidivism_rate), 2)))  # density plot
for 'recidivism_rate'
polygon(density(rehab_reci_delhi$recidivism_rate), col="red")

#Linear model
linearMod <- lm(total ~ recidivism_rate, data=rehab_reci_delhi)  # build
linear regression model on full data
print(linearMod)
summary(linearMod)

#calculate p-statistic and p-value
install.packages("QuantPsyc")
library(QuantPsyc)
modelSummary <- summary(linearMod)  # capture model summary as an object
modelCoeffs <- modelSummary$coefficients  # model coefficients
beta.estimate <- modelCoeffs["recidivism_rate", "Estimate"]  # get beta
estimate for speed
std.error <- modelCoeffs["recidivism_rate", "Std. Error"]  # get std.error
for speed
t_value <- beta.estimate/std.error  # calc t statistic
p_value <- 2*pt(-abs(t_value), df=nrow(rehab_reci_delhi)-
ncol(rehab_reci_delhi))  # calc p Value
f_statistic <- linearMod$fstatistic[1]  # fstatistic
f <- summary(linearMod)$fstatistic  # parameters for model p-value calc
model_p <- pf(f[1], f[2], f[3], lower=FALSE)
AIC(linearMod)
BIC(linearMod)

#split the dataset into training and test data
set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex <- sample(1:nrow(rehab_reci_delhi),
0.8*nrow(rehab_reci_delhi))  # row indices for training data
trainingData <- rehab_reci_delhi[trainingRowIndex, ]  # model training data
testData  <- rehab_reci_delhi[-trainingRowIndex, ]   # test data

# Build the model on training data
```

```
lmMod <- lm(total ~ recidivism_rate, data=trainingData)  # build the model
totalPred <- predict(lmMod, testData)  # predict distance

#review diagnostic measures
summary(lmMod)
AIC(lmMod)

#calculating prediction accuracy and error rate
actuals_preds <- data.frame(cbind(actuals=testData$total,
predicteds=totalPred))  # make actuals_predicteds dataframe.
correlation_accuracy <- cor(actuals_preds)
head(actuals_preds)

min_max_accuracy <- mean(apply(actuals_preds, 1, min) / apply(actuals_preds,
1, max))
mape <- mean(abs((actuals_preds$predicteds -
actuals_preds$actuals))/actuals_preds$actuals)

#K-fold cross validation
install.packages("DAAG")
library(DAAG)
cvResults <- suppressWarnings(CVlm(data=rehab_reci_delhi,
form.lm=recidivism_rate ~ total, m=5, dots=FALSE, seed=29,
legend.pos="topleft",  printit=FALSE, main = "Actual and Predicted Data in
Delhi Prison Dataset"))


## Prisoner Wages vs Budget in Prisons

#Scatterplot matrix
budget_wages <- read.csv("budget_wages.csv")
summary(budget_wages)
library(car)
scatterplotMatrix(budget_wages[,c(3:6)], spread = FALSE, lty.smooth = 2, main
= "Scatterplot Matrix")
names(budget_wages)
summary(budget_wages)

#Density Plot
library(e1071)
par(mfrow=c(1, 2))  # divide graph area in 2 columns
plot(density(budget_wages$wages_semi_skilled_convicts), main="Density Plot:
Wages of semi-skilled convicts", ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(budget_wages$wages_semi_skilled_convicts), 2)))  #
density plot for 'wages_semi_skilled_convicts'
polygon(density(budget_wages$wages_semi_skilled_convicts), col="blue")
plot(density(budget_wages$budget), main="Density Plot: Budget",
ylab="Frequency", sub=paste("Skewness:",
round(e1071::skewness(budget_wages$budget), 2)))  # density plot for 'budget'
polygon(density(budget_wages$budget), col="blue")

#Linear model
linearMod2 <- lm(wages_semi_skilled_convicts ~ budget, data=budget_wages)  #
build linear regression model on full data
print(linearMod2)
summary(linearMod2)
```

```
#calculate p-statistic and p-value
modelSummary2 <- summary(linearMod2)  # capture model summary as an object
modelCoeffs2 <- modelSummary2$coefficients  # model coefficients
beta.estimate2 <- modelCoeffs2["budget", "Estimate"]  # get beta estimate for
wages
std.error2 <- modelCoeffs2["budget", "Std. Error"]  # get std.error for wages
t_value2 <- beta.estimate2/std.error2  # calc t statistic
p_value2 <- 2*pt(-abs(t_value2), df=nrow(budget_wages)-ncol(budget_wages))  #
calc p Value
f_statistic2 <- linearMod2$fstatistic2[2]  # fstatistic
f <- summary(linearMod2)$fstatistic2  # parameters for model p-value calc
model_p <- pf(f[1], f[2], f[3], lower=FALSE)
AIC(linearMod2)
BIC(linearMod2)

#split the dataset into training and test data
set.seed(100)  # setting seed to reproduce results of random sampling
trainingRowIndex2 <- sample(1:nrow(budget_wages), 0.8*nrow(budget_wages))  #
row indices for training data
trainingData2 <- budget_wages[trainingRowIndex2, ]  # model training data
testData2  <- budget_wages[-trainingRowIndex2, ]   # test data

# Build the model on training data
lmMod2 <- lm(wages_semi_skilled_convicts ~ budget, data=trainingData2)  #
build the model
wages_semi_skilled_convictsPred <- predict(lmMod2, testData2)  # predict
distance

#review diagnostic measures
summary(lmMod2)
AIC(lmMod2)

#calculating prediction accuracy and error rate
actuals_preds2 <-
data.frame(cbind(actuals=testData2$wages_semi_skilled_convicts,
predicteds=wages_semi_skilled_convictsPred))  # make actuals_predicteds
dataframe.
correlation_accuracy2 <- cor(actuals_preds2)
head(actuals_preds2)

min_max_accuracy2 <- mean(apply(actuals_preds2, 1, min) /
apply(actuals_preds2, 1, max))
mape2 <- mean(abs((actuals_preds2$predicteds -
actuals_preds$actuals))/actuals_preds2$actuals)

#K-fold cross validation
cvResults2 <- suppressWarnings(CVlm(data=budget_wages, form.lm=budget ~
wages_semi_skilled_convicts, m=5, dots=FALSE, seed=29, legend.pos="topleft",
printit=FALSE, main = "Actual and Predicted Data in Wages for Semi-skilled
Convicts"))
par(mfrow=c(1,1))
```