Crime Rate Prediction

Name	Roll Number	Class & Section
Battula Sai Charan	AP18110010214	CSE-D
Paladugu Sirivanth	AP18110010245	CSE-D
Yakkala Tarun Sai	AP18110010469	CSE-H
Neeraj Boggarapu	AP18110010474	CSE-H
Modekurty Sri Krishna Kumar	AP18110010506	CSE-H

Abstract :

The main objective of our project is to predict the Crime Rate in different regions using specific parameters like density, country, crime rate, central etc. In this project we considered a dataset of a particular Country. The region we selected is the United States of America. We used Linear Regression model to predict the crime rate. Finally a graph is plotted after implementing Linear Regression. A graph is plotted between density and crime rate to enhance distribution of crime rate in a particular region.

Introduction :

Analysis of crime is a methodological approach to the identification and assessment of criminal patterns and trends. Before starting this project we have gone through many datasets all over the world, but in the USA a state named North Carolina has given us sufficient data to start off with. The variables with highest correlation in regards with crime rate are urban and density. The reason for highest correlation for urban and density variables is that the urban areas of North Carolina are densely populated. As a result, there is a high probability of multicollinearity between the density and urban features. The combination of density and location features will help us in predicting the crime rate in North Carolina. The columns of wage, wfed and wtrd are positively correlated with the density feature. From this we can instinctively understand that the weekly wages would be higher in urban localities. The features wtrd and wfir are positively correlated with wfed and wloc respectively. Both the wfir and wtrd are moderately correlated. The features crime rate, density, urban, wfed, wtrd and taxpc are very highly correlated with the Crime Rate. The prediction of future crime location can be determined by geographical data mining approaches.

Problem survey:

In our problem we will be evaluating and examining the large pre existing databases in order to generate new information which would help us to find the solution. The prediction is based on the extraction of the new information using the existing datasets. The main aim of this problem is to perform the survey on certain algorithms which helps us to analyse the crime rate.

Dataset description:

In our dataset we have considered several parameters like country, year, crime rate, density etc. After once we get the raw dataset we then clean the dataset. After we clean the dataset we then preprocess it. We then use linear regression to obtain our results, the results are obtained in the form of graphs. We then use a statistical approach to find values like count, mean, standard deviation, minimum and maximum values. For each and every parameter we have given a model summary. A combination of density and location (west/ central/ urban) can help aid crime rate prediction.The variables with highest correlation in regards with crime rate are urban and density.

The drive link for the dataset collected is as follows:
https://drive.google.com/file/d/1vUtFfG8JVoV3TExI5s9iM6UQ3OBZQaqF/view?usp=sharing

Data Preprocessing:

The important step in the entire process is data preprocessing. Data preprocessing allows us to remove the unwanted data with the help of data cleaning, this allows the user to have a dataset which contains more valuable information after the preprocessing stage in the mining process.

Data Cleaning :

```
crimeData = crimeData[crimeData.county != 185]
crimeData = crimeData[crimeData.county != 115]
crimeData = crimeData[crimeData['prbarr'] < 1]
crimeData = crimeData[crimeData['prbconv'] < 1]
crimeData = crimeData[crimeData['west']+crimeData['central'] <= 1]
crimeData = crimeData.drop('year', axis=1)
print (crimeData.shape)
```

```
(80, 24)
```

Preprocessing the cleaned data :

```
import statsmodels.api as sm
y = crimeData['crmrte']
X = crimeData['density']
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
density_pvalue = model.pvalues['density']
model.summary()
```

Implementation:

We have implemented the problem using an algorithm called Linear Regression. We have developed a python code based on the algorithm. For this problem we also calculated R Square, P - Value Difference, Number of Observations, Covariance type, etc.

Parameters : Crime Rate :

```
y = crimeData['crmrte']
X = crimeData.drop('crmrte', axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Parameters : Density, Urban :

```
y = crimeData['crmrte']
X = crimeData[['density', 'urban']]
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
density_pvalue_upd = model.pvalues['density']
print('Difference in P-Value = ' + str(density_pvalue_upd - density_pvalue))
model.summary()
```

```
Difference in P-Value = 2.2791391426354033e-05
```

Parameters : Crime Rate, Urban, Country :

```
y = crimeData['crmrte']
X = crimeData.drop(['crmrte', 'urban', 'county'], axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Parameters : Crime Rate, Urban, Country :

```
y = crimeData['crmrte']
X = crimeData.drop(['crmrte', 'urban', 'county','wmfg', 'prbpris', 'wloc', 'west', 'wtuc'], axis=1)
X = sm.add_constant(X)
model = sm.OLS(y, X).fit()
model.summary()
```

Resultant graphs:

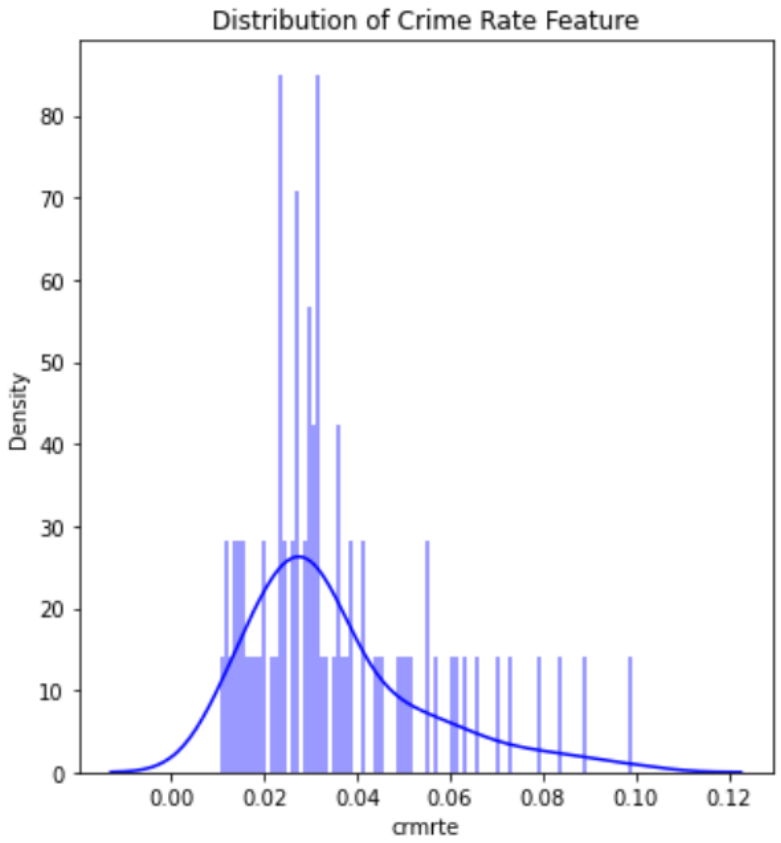
Statistical approach towards Crime Rate Prediction :

```
print('Statistics of Crime Rate: \n')
print(crimeData['crmrte'].describe())
plt.figure(figsize=(6,6.5))
plt.title('Distribution of Crime Rate Feature')
sns.distplot(crimeData['crmrte'], color='b', bins=100, hist_kws={'alpha': 0.4})
```

```
Statistics of Crime Rate:

count    80.000000
mean      0.035126
std       0.018846
min       0.010623
25%       0.023359
50%       0.030342
75%       0.041639
max       0.098966
Name: crmrte, dtype: float64
```

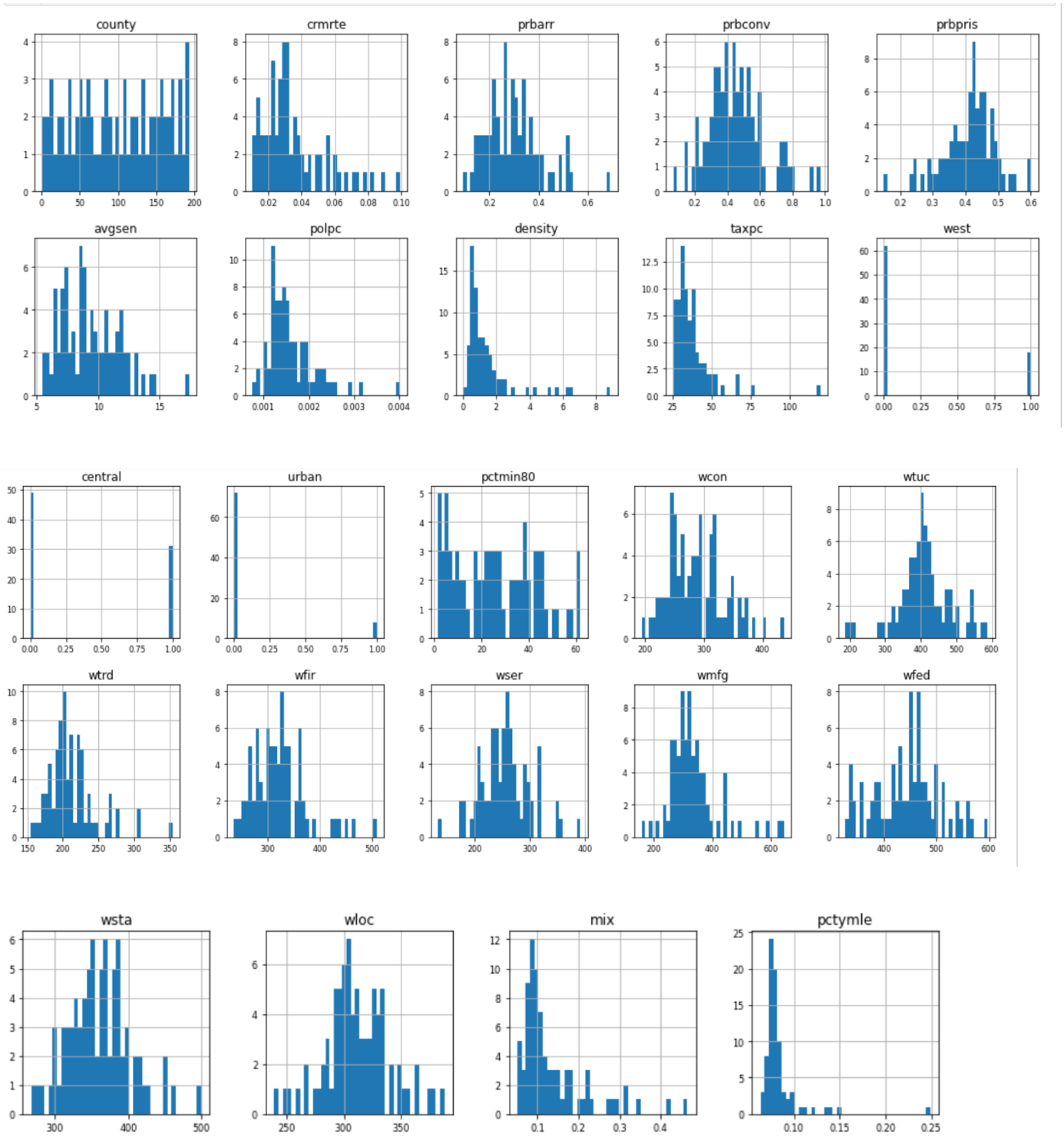
Graph Output:



Crime Data Mined in different perspectives :

```
1 crimeData.hist(figsize=(18,18), bins=40, xlabelsize=8, ylabelsize=8);
```

Graph Outputs :



Conclusion:

The map of Actual vs. Predicted is linear.
This indicates that the forecast is right. The amount of data available for input is small. The plot could be more linear if there was more evidence.

We may also incorporate the boolean features west, centre, and urban into a single feature with categorical values 1,2, and 3 to create a single feature. A single function like this may be more useful for prediction. On features, functional transformations (such as log, function) can be useful. If there is a possibility to incorporate functionality, using 'unemployment rate' as a proxy for crime rate might be useful.