CL-205 PROJECT REPORT

Team Members:

Rahul (23B0305)

A Naveen (23B0327)

Rishit Kesharwani (23B0439)

PROBLEM STATEMENT

Analyzing how crime rates are influenced by a wide array of socioeconomic factors like income inequality, education level, unemployment rate, per capita expenditure on police protection, etc. across 47 states in the United States using a dataset from 1960. Our main goal is to see which of these factors have a significant impact on the crime rates.

RELEVANT LINKS

Code: **Analysis using Python**

Dataset: **US Crime Dataset**

OBJECTIVES

- Calculating statistics of features in the dataset such as crime rate, wealth, income inequality, etc. across states and evaluating the confidence interval of crime rate
- Comparing mean crime rates among the northern and southern rates and calculating confidence interval for difference in means

- Using t-tests and Pearson's correlation to perform hypothesis testing, determining statistical significance of all features to describe the relation of socioeconomic factors and crime rates
- Performing linear regression on using specific predictors of choice, and highlighting the 95% prediction interval, specifying the mean square errors (MSE) and the R-value
- Plotting these regression line fits with the prediction intervals and the residual errors for visual analysis

OBJECTIVE 1: *Calculate statistics of every feature in the dataset such as crime rate, wealth inequality and determine the* ***95% confidence interval*** *for crime rate in the states of US for 1960.*

Feature	Mean	Median	Std Dev	Variance
Percentage of Males Aged 14–24	13.86	13.60	1.26	1.58
Southern State Indicator	0.34	0.00	0.48	0.23
Mean Years of Schooling (25+)	10.56	10.80	1.12	1.25
Police Expenditure 1960 (Po1)	8.50	7.80	2.97	8.83
Police Expenditure 1959 (Po2)	8.02	7.30	2.80	7.82
Labour Force Participation (14-24)	0.56	0.56	0.04	0.00
Male-to-Female Ratio	98.30	97.70	2.95	8.68
State Population (100k)	36.62	25.00	38.07	1449.42
Percentage of Nonwhites	10.11	7.60	10.28	105.74
Unemployment Rate (Urban 14-24)	0.10	0.09	0.02	0.00
Unemployment Rate (Urban 35-39)	3.40	3.40	0.84	0.71
Median Family Income	5253.83	5370.00	964.91	931050.23
Income Inequality	19.40	17.60	3.99	15.92
Probability of Imprisonment	0.05	0.04	0.02	0.00
Average Time Served (Months)	26.60	25.80	7.09	50.22
Crime Rate (per 100k population)	905.09	831.00	386.76	149585.38

Table 1: Statistical Summary of the features

95% Confidence interval for crime rate can be evaluated using:

$$\text{Confidence Interval (CI)} = \bar{x} \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

After calculation, it is evaluated as (791.527, 1018.642)

OBJECTIVE 2: *Comparing mean crime rates among the northern and southern rates and calculating confidence interval for difference in means*

$$\text{Mean} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{Sample Standard Deviation} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Category	Mean Crime Rate (per 100,000 population)	Standard Deviation
South	856.8125	270.1760
Northern	930.0	437.0154

Table 2: Mean and Standard Deviation of States

Difference in mean crime rates: -73.1875

Confidence Interval (95%) for difference in means: (-293.902, 147.527)

It can be seen that the mean crime rate in the northern states is slightly higher, however the standard deviation is also higher which indicates that the difference is not statistically significant. Moreover, the confidence interval for the difference in

mean crime rates varies over a wide range center, supporting our conclusion in the previous statement.

OBJECTIVE 3: *Using t-tests and Pearson's correlation to perform hypothesis testing, determining statistical significance of all features to describe the relation of socioeconomic factors and crime rates*

We have used two-sample t-test and Pearson correlation to perform the hypothesis testing on all the features of the dataset with respect to crime rate to see if there is a statistically significant correlation. Summarizing the formulae here:

Hypothesis Test	Null Hypothesis H_0	Alternative Hypothesis H_1	Test Statistic t	Degrees of Freedom df
Two-Sample T-Test	$H_0 : \mu_1 = \mu_2$	$H_1 : \mu_1 \neq \mu_2$	$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$
Pearson Correlation Test	$H_0 : \rho = 0$	$H_1 : \rho \neq 0$	$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	$df = n - 2$

Table 3: Specifications of the tests used

Here ρ represents the correlation, where a non-zero correlation indicates that there is indeed a statistical significance of the factor on the crime rate.

After running the hypothesis testing on selected features, we have the following observations:

Factor	T-Test Result ($p < 0.05$)	T-Test Result ($p < 0.01$)	Correlation Result ($p < 0.05$)	Correlation Result ($p < 0.01$)
Ed	Not Significant	Not Significant	Significant	Not Significant
Ineq	Not Significant	Not Significant	Not Significant	Not Significant
U1	Not Significant	Not Significant	Not Significant	Not Significant
U2	Not Significant	Not Significant	Not Significant	Not Significant
LF	Significant	Not Significant	Not Significant	Not Significant
Po1	Significant	Significant	Significant	Significant
Po2	Significant	Significant	Significant	Significant
M	Not Significant	Not Significant	Not Significant	Not Significant
NW	Significant	Significant	Not Significant	Not Significant
Wealth	Significant	Significant	Significant	Significant
Time	Not Significant	Not Significant	Not Significant	Not Significant

Table 4: Hypothesis Testing Results

Some Key Observations:

1. **Significant Factors at Both Levels**:
    - *Per capita expenditure on police protection (1959 & 1960)* and *wealth* are statistically significant at both the 0.05 and 0.01 levels for both T-Test and Pearson correlation analysis.
    - *Percentage of nonwhites in the population* is statistically significant for T-Test at both levels but not for correlation.
2. **Factors Significant Only at 0.05 Level**:
    - *Labor force participation rate of civilian urban males* is significant at the 0.05 level in T-Test but not in correlation.
    - *Education level* shows significance in correlation at the 0.05 level but not in T-Test.

3. **Non-Significant Factors**:
    - *Income inequality, unemployment rates, number of males in the 14-24 age group and average time in months served by offenders* are not statistically significant in both T-Tests and Correlation analysis at either significance level.

According to societal philosophy, we would expect to see some correlation between income inequality and unemployment rates with the crime rate, i.e. more income inequality should correspond to more crime rate, and similarly with unemployment rate. The reason we don't see the result from these tests might be because of the absence of sufficient data points and the change over time.

OBJECTIVE 4 : *Performing **linear regression** on using specific predictors of choice, and highlighting the **95% prediction interval**, specifying the **mean square errors (MSE)** and the **R-value***

The R squared value of comparison is calculated using the formula,

$$R^2 = \frac{SS_R}{SS_T}$$

MSE is the measure of averaged squared difference between actual values and predicted values. It is useful for quantifying the model's prediction error, with smaller values indicating better model performance.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 : \text{ total variability in response variable}$$

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 : \text{ variability captured by regression line}$$

$$SS_{Res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 : \text{ variability left unexplained}$$

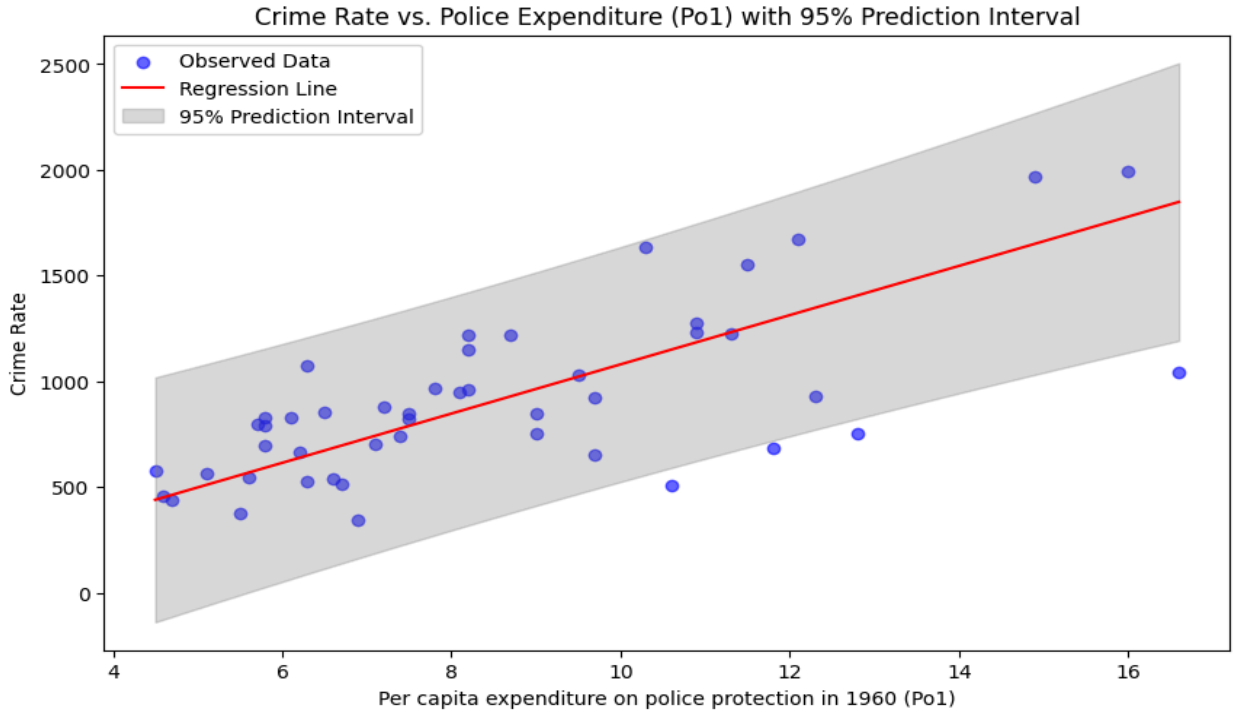
Metric	Value
Intercept	469.1779
Slope for Po1	116.4202
Slope for Wealth	-0.1054
Mean Squared Error (MSE)	73335.1911
R-squared (R^2)	0.4991

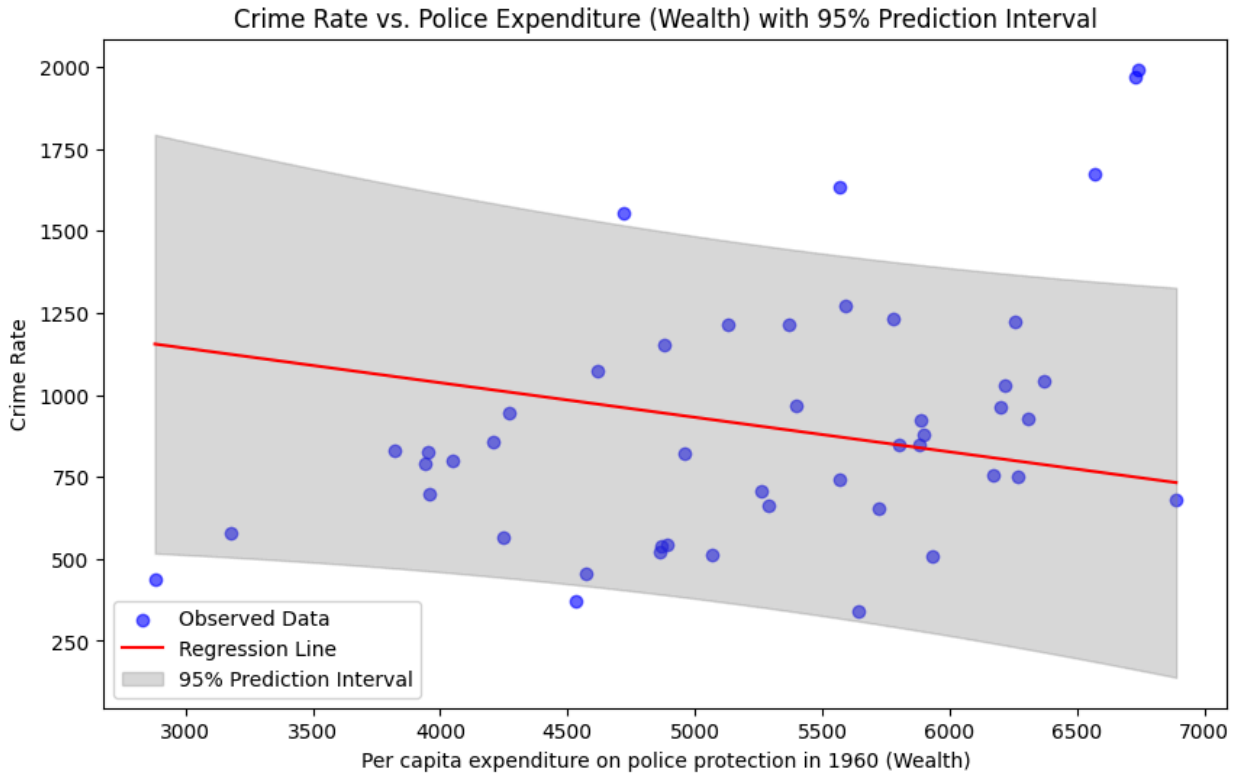
Table 5: Estimates of Slope and intercept and metrics of Linear Regression

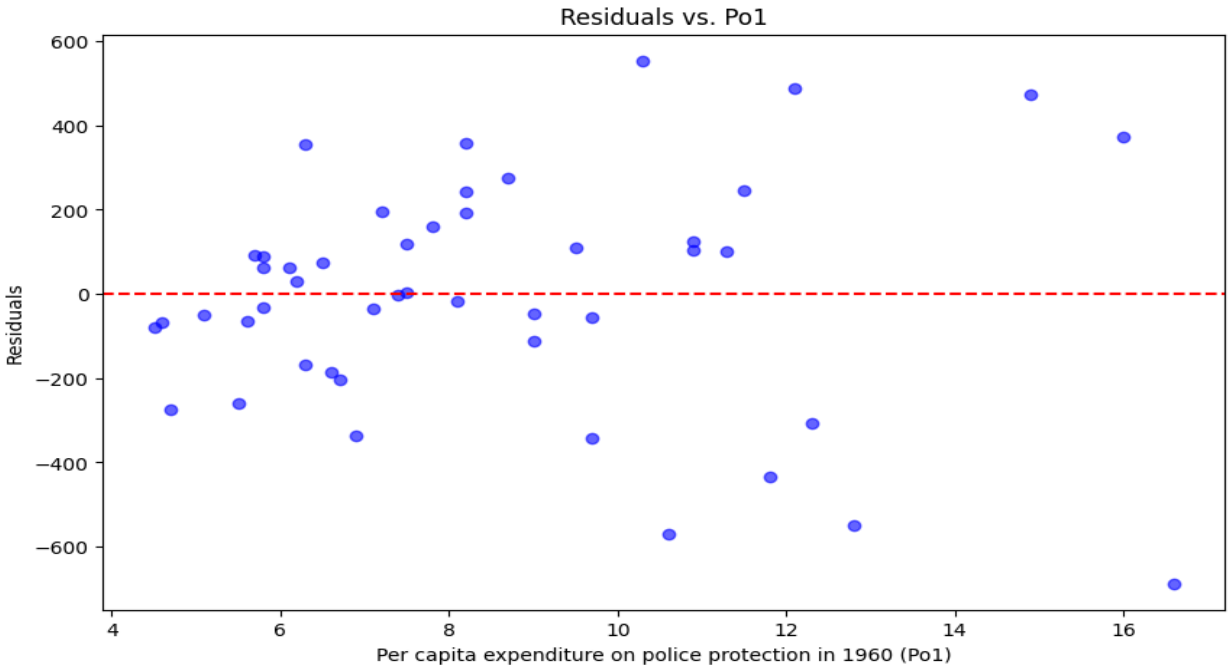
OBJECTIVE 5: *Calculating residual errors and plotting these regression line fits with the prediction intervals and the residual errors for visual analysis*

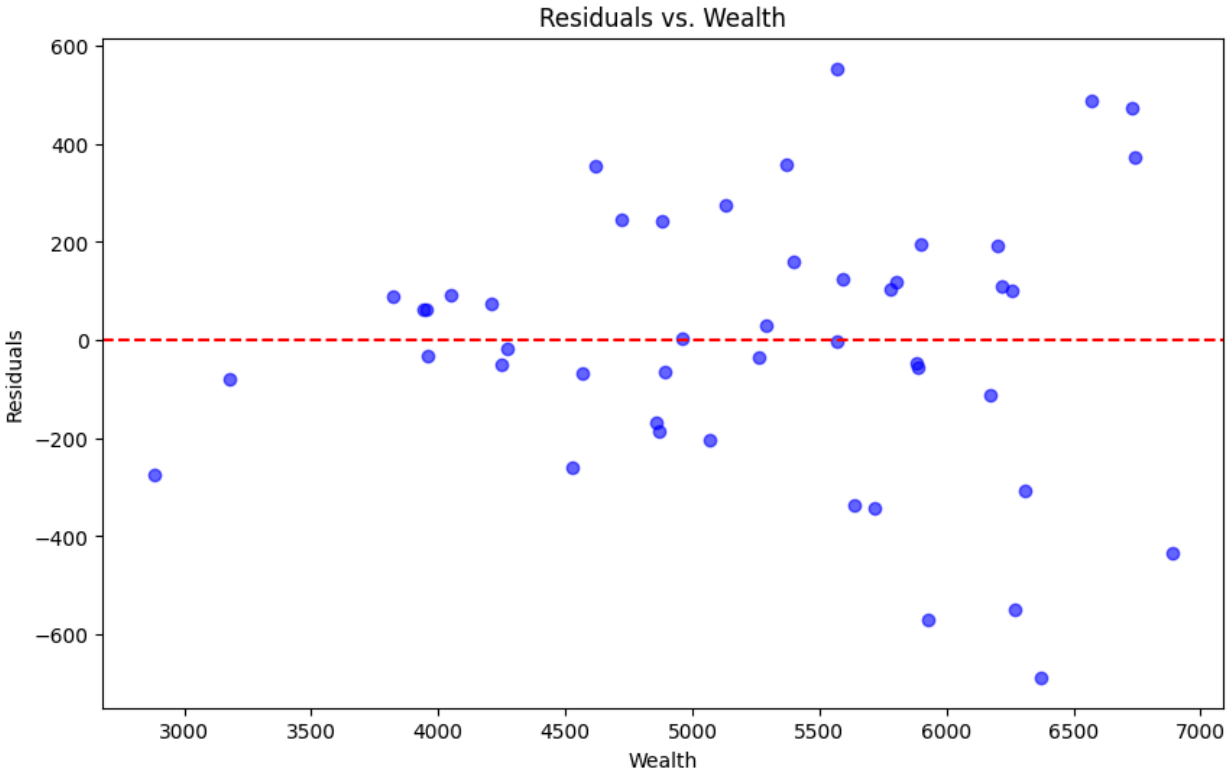
$$e_i = y_i - \hat{y}_i$$

This is how the residual error is expressed, i.e. the difference between the actual observed value and the predicted value. We have done this for police expenditure (1960) and the wealth factors. We can see qualitatively that the points are distributed almost equally above and below the zero line, which is what we would expect from normally distributed noise.


Crime Rate vs. Police Expenditure (Po1) with 95% Prediction Interval
Observed Data
Regression Line
95% Prediction Interval
Crime Rate
Per capita expenditure on police protection in 1960 (Po1)


Crime Rate vs. Police Expenditure (Wealth) with 95% Prediction Interval
Observed Data
Regression Line
95% Prediction Interval
Crime Rate
Per capita expenditure on police protection in 1960 (Wealth)


Residuals vs. Po1
Residuals
Per capita expenditure on police protection in 1960 (Po1)
600
400
200
0
−200
−400
−600
4
6
8
10
12
14
16


Residuals vs. Wealth
Residuals
Wealth
600
400
200
0
−200
−400
−600
3000
3500
4000
4500
5000
5500
6000
6500
7000