DSA 210 Project Report – Crime Rate vs. Socio-Economic Factors
Efe Sözer - 31012

---

1. Introduction

This study investigates how five socio-economic indicators relate to **Crime Rate**:

Feature	Brief description
Education Level %	Share of residents with at least a high-school diploma
Employment Rate %	Share of the labour force that is employed
Median Income (USD)	Median household income
Poverty Rate %	Share of the population below the poverty line
Population Density	Residents per km²

---

2. Data Collection and Processing

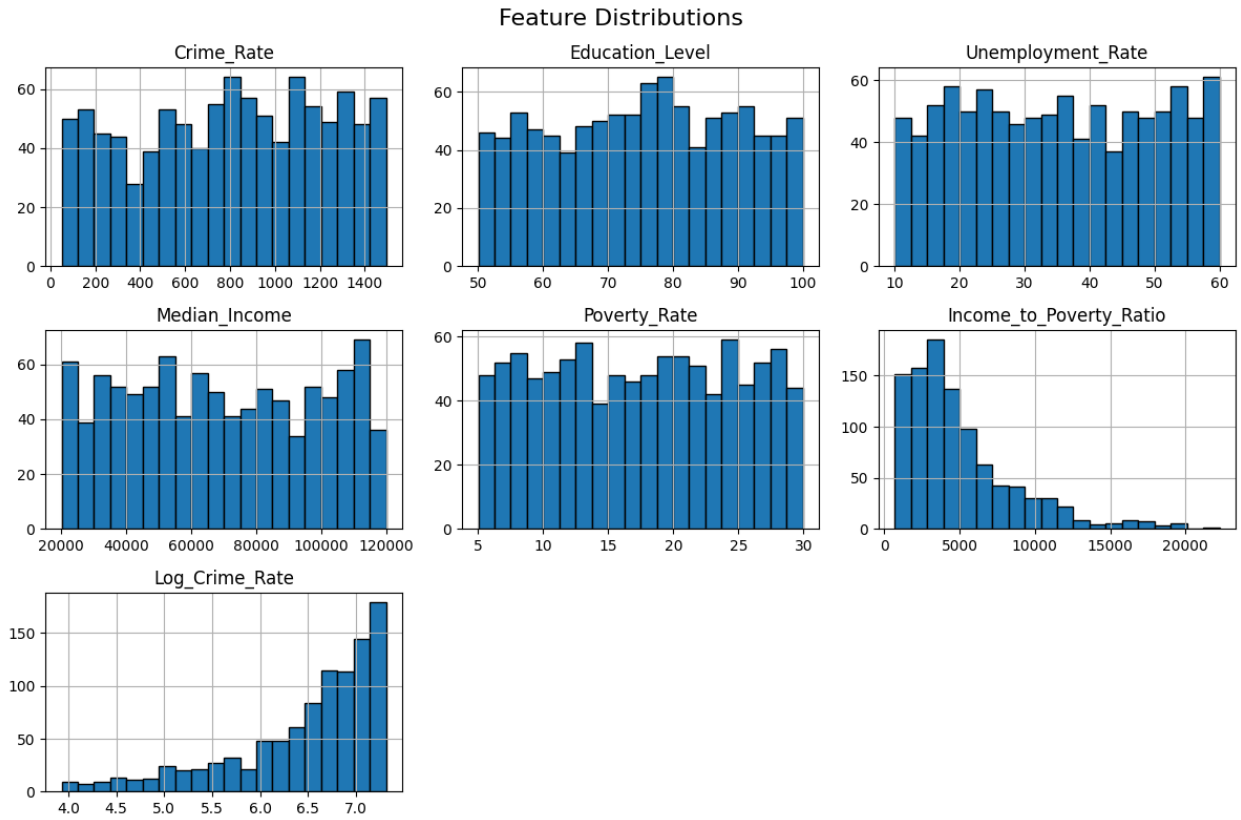
- **Source:** Kaggle dataset *"Crime Rate vs Socio-Economic Factors"*
- **Feature engineering**
  - Unemployment_Rate = 100 – Employment_Rate
  - Income_to_Poverty_Ratio = Median_Income / Poverty_Rate
  - Log_Crime_Rate (natural log) applied to reduce skew
  - Poverty_Density = Poverty_Rate × Population_Density
- No rows were removed; the dataset was free of missing values and extreme outliers.

---

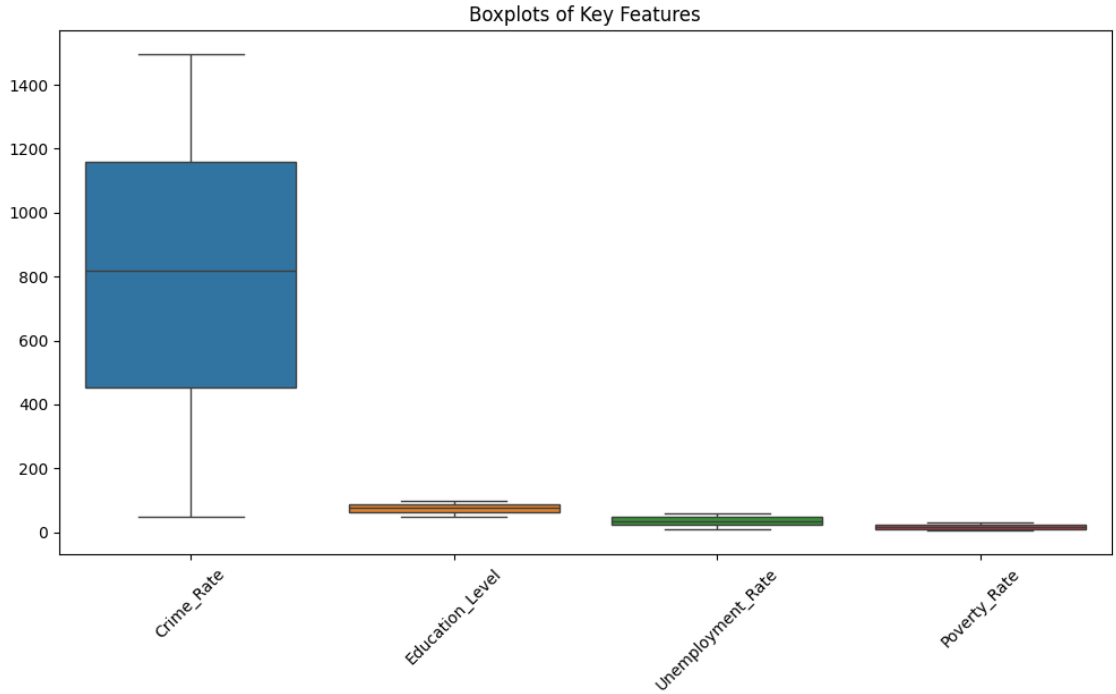
3. Exploratory Data Analysis

Statistic	Crime Rate	Education	Unemployment	Median Income	Poverty
Mean	802.43	75.39	35.18	69 427.93	17.47
Std Dev	418.19	14.12	14.65	29 219.03	7.23
Median	818.50	76.14	34.97	67 484.00	17.55

3.1 Histograms



3.2 Box-plots



4. Hypothesis Tests

Hypothesis 1 – Unemployment Rate vs Crime Rate
• Null hypothesis: There is no significant association between Unemployment Rate and the Crime Rate.
• Alternative hypothesis: There is a significant association between Unemployment Rate and the Crime Rate.
• Statistical test applied: Pearson product-moment correlation.
• Correlation coefficient: 0.02 P-value: 0.8923
→ Decision ($\alpha = 0.05$): fail to reject H_0; no evidence of a linear relationship.

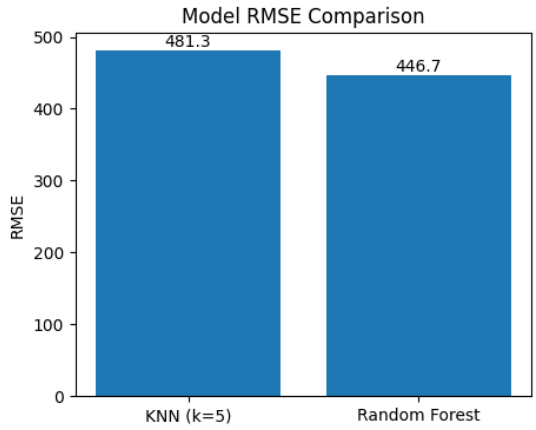
Hypothesis 2 – Median Household Income vs Crime Rate
• Null hypothesis: There is no significant association between Median Household Income and the Crime Rate.
• Alternative hypothesis: There is a significant association between Median Household Income and the Crime Rate.
• Statistical test applied: Pearson product-moment correlation.
• Correlation coefficient: 0.01 P-value: 0.8067
→ Decision ($\alpha = 0.05$): fail to reject H_0; median income does not appear to relate linearly to crime.

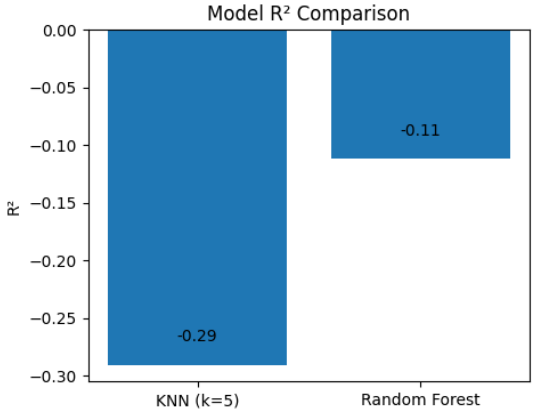
Hypothesis 3 – Education Level vs Crime Rate
• Null hypothesis: There is no significant association between the percentage of residents who completed high school (Education Level) and the Crime Rate.
• Alternative hypothesis: There is a significant association between Education Level and the Crime Rate.
• Statistical test applied: Pearson product-moment correlation.
• Correlation coefficient: -0.06 P-value: 0.0565
→ Decision ($\alpha = 0.05$): fail to reject H_0; the relationship is not statistically significant, although it is close to the threshold.

Hypothesis 4 – Poverty Rate as a Predictor when Density Is High
• Null hypothesis: There is no significant difference in Crime Rate between high-poverty and low-poverty regions (regardless of population density).
• Alternative hypothesis: Crime Rate differs significantly between high-poverty and low-poverty regions.
• Statistical test applied: Independent-samples t-test (regions split at median Poverty Rate).
• t-statistic: -1.41 P-value: 0.1597
→ Decision ($\alpha = 0.05$): fail to reject H_0; poverty level alone is not a significant divider of crime in this dataset.

5. Machine-Learning Models

Model	RMSE	R^2
K-Nearest Neighbors (k = 5)	**481.26**	**–0.29**
Random Forest	**446.69**	**–0.11**





---

6. Discussion

- The classical socio-economic variables studied show **no strong standalone relationship** with crime.
- Education exhibits the clearest negative trend.
- Tree-based ensemble (Random Forest) improves RMSE versus KNN but still leaves most variance.

- Crime drivers are likely multifactorial, with factors not captured in this dataset.

---

7. Limitations and Future Work

1. Single-year, aggregated data – a panel covering multiple years could capture temporal trends.
2. Possible multicollinearity (e.g., income vs poverty) – future work might apply regularisation or PCA.
3. Important omitted variables – police presence, social-service expenditure, and urban-design metrics.
4. Spatial autocorrelation not addressed – geographically weighted regression or spatial lag models are recommended.

---

8. Conclusion

Within this dataset, socio-economic indicators alone explain little of the observed variation in crime rates. More granular data and advanced spatial or causal modelling are needed to yield actionable insights.