**Exploring Factors Affecting Homicide Rates in Countries**

Country data came from Kaggle. Based on 38 traits and 204 entries, a detailed study of key social and economic indicators in several countries was offered. Due to its wide range of measurements, worldwide trends in education, the environment, and GDP may be studied in detail. The structured file comprises environmental, education, and GDP datasets from countries. The most important factors in the data set are homicide rates per 100,000 people, urban population growth, internet users, GDP per capita, refugees, and population growth.

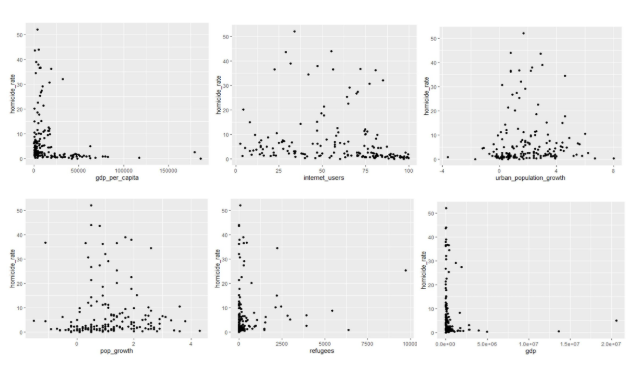
In the project, we will study urban population growth, internet use, GDP per capita, refugees, and population increase. The data set has issues with missing and repeated data points. Another difficulty is that its creation date is unknown. We propose that shrinking urban population, fewer refugees, higher GDP per capita, larger GDP, smaller population, and more internet user's lower homicide rates.

## Data Cleaning Process

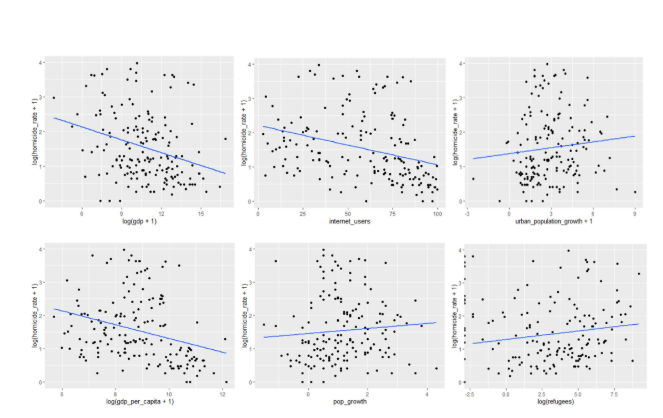
We cleaned and refined web data using R's dplyr tool to make it statistically viable. The study started by selecting only the relevant columns: nation name, homicide rate, GDP, urban population growth, population growth, internet users, GDP per capita, and refugees. The dataset was reduced to focus on analytical variables by removing extraneous columns. Next, missing homicide rate values were deleted and the duplicate country names. This step was necessary to maintain data integrity because observations with missing homicide rate values could skew the analysis.

## Association Analysis

In our association analysis, we compared homicide rates and socioeconomic factors across nations. The homicide rate was plotted against GDP, urban population growth, population growth, internet user percentage, GDP per capita, and refugee population. Homicide rate scatterplots against urban population growth, population growth, and internet users showed considerable non-linearity and non-monotonicity. The scatterplots of homicide rate against GDP, GDP per capita, and refugees showed nonlinearity, high skewness, and considerable non-monotonicity. Non-monotonic correlations prevented Spearman's rank correlation assessment of these variables.



The initial scatterplot showed non-linear and non-monotonic correlations, therefore we methodically performed and plotted log transformations on the predictors and/or response variable and picked the best plots for testing. The scatterplot's right-skewness forced us to transform GDP, GDP per capita, and refugee population predictor and response variables. While preserving the predictors in their original scale, we log-transformed the homicide rate for urban population growth, population growth, and internet user percentage. Using log(x+1) prevented undefined log(0) values without materially modifying the data.



Pearson correlations were possible since the adjustments moderately linearized the relationships. We found a moderate negative linear association between log-transformed homicide rate and GDP. Similar patterns were seen when plotting log(homicide rate) versus internet user percentage and GDP per capita. Urban, refugee, and population growth had weaker connections.

Three predictors showed significant Pearson correlation coefficients. The association between log-transformed homicide rate and GDP (r = -0.2951072; p < 0.001) suggests that nations with higher GDP have lower homicide rates. A moderate negative connection (r = -0.3205548, p < 0.001) suggests that increasing internet usage is associated with lower homicide rates. Log(GDP per capita) had a significant negative connection (r = -0.3025433, p < 0.001). Urban population growth (r = 0.1014501, p = 0.186), population growth (r = 0.0818263, p = 0.268), and refugee population (r = 0.05412809, p = 0.522) showed no statistically significant associations with homicide rates.

|  |  |  |  |
| --- | --- | --- | --- |
| **Transformed Variables** | **p-value** | **Pearson's r value** | **Statistically Significant** |
| log(Homicide Rate) vs log(GDP) | 0 | -0.2951072 | Yes |
| log(Homicide Rate) vs Urban Population Growth | .186 | 0.1014501 | No |
| log(Homicide Rate) vs Population Growth | .268 | 0.08182263 | No |
| log(Homicide Rate) vs Internet Users | 0 | -0.3205548 | Yes |
| log(Homicide Rate) vs log(GDP Per Capita) | 0 | -0.3025433 | Yes |
| log(Homicide Rate vs log(Refugee) | .522 | 0.05412809 | No |

These findings show that economic and technological development indicators predict homicide rates more than demographic factors in our dataset. The considerable negative correlations between log transformed homicide rates, GDP, internet usage, and GDP per capita suggest these variables may be valuable predictors but may be better suited for non-linear modeling. The relationship between log(homicide rates) and internet user percentage may be attributable to economic variables like GDP.

## Simple Linear Regression

To evaluate the relationship between each predictor variable (GDP, urban population growth, population growth, internet users, GDP per capita, and refugees) and homicide rates, simple linear regression was conducted independently for each predictor. The primary objective was to assess the strength and significance of the relationship between each predictor and homicide rates.

For each regression, a linear model was fit to the data using the lm() function in R, with homicide rate as the dependent variable and the respective predictor as the independent variable (e.g., lm(homicide\_rate ~ predictor\_variable)). This approach yielded key metrics such as the slope, intercept, R², p-values, and confidence intervals. The summary() function was used to extract detailed model outputs.

The findings revealed that GDP, urban population growth, and refugees had insignificant relationships (p > 0.05) with homicide rates. These models were characterized by low R² values and wide confidence intervals, further confirming their weak explanatory power. Conversely, internet users and GDP per capita showed statistically significant relationships (p < 0.05), yet their R² values (0.029 and 0.046, respectively) highlighted limited predictive power.

Overall, these results suggest that none of the individual predictors provides a robust explanation of homicide rate variability. This underscores the need to explore additional variables or develop more complex multivariate models to better understand the factors influencing homicide rates.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Formula** | **Intercept** | **Slope** | **R2** | **P-value** |
| Homicide Rate~GDP | 7.329 | -4.053\*107 | 0.006004 | 0.318 |
| Homicide Rate~Urban Population Growth | 6.7657 | 0.1666 | 0.0008061 | 0.714 |
| Homicide Rate~Internet Users | 10.89886 | -0.06505 | 0.0291 | 0.0275\*\* |
| Homicide Rate~GDP per capita | 8.562 | -8.282\*105 | 0.04619 | 0.00515\*\* |
| Homicide Rate~Refugees | 6.55575 | 0.000873 | 0.01022 | 0.205 |

## Tests performed

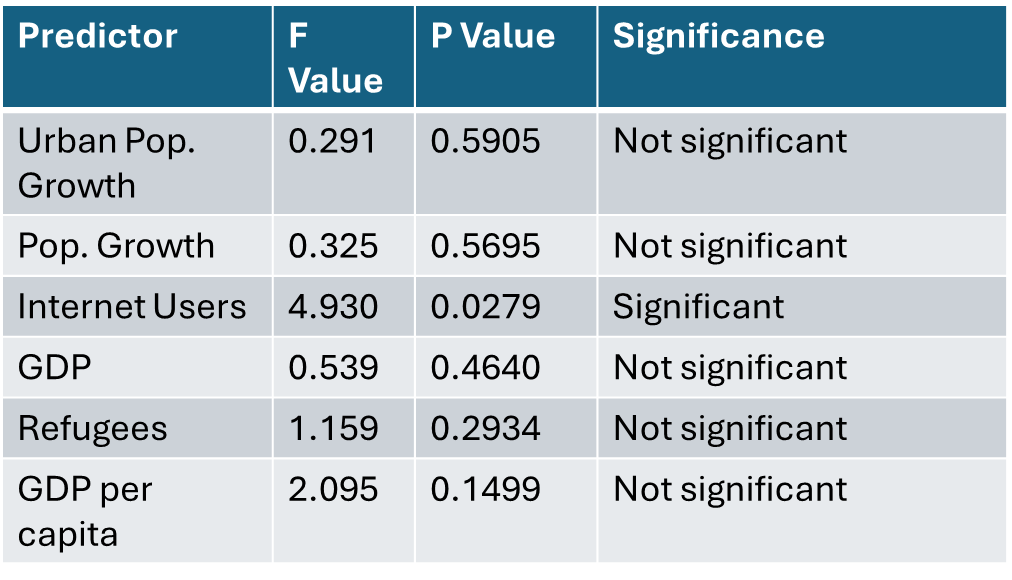
**Analysis of Variance (ANOVA)**

The Analysis of Variance (ANOVA) test was conducted to evaluate the relationship between the dependent variable, homicide rate, and the independent variables: GDP, urban population growth, pop growth, internet users, GDP per capita, and refugees. The purpose of the ANOVA test was to determine which predictors, if any, had a statistically significant effect on the variance in homicide rate.

The null hypothesis for the ANOVA was that each independent variable does not explain a significant portion of the variance in homicide rate, meaning that changes in those variables would have no measurable impact on the dependent variable. A significance level of α = 0.05 was used to evaluate the results.

The ANOVA results revealed that internet users were the only variable with a statistically significant impact on homicide rate, with a p-value of 0.0279 (p < 0.05). This finding indicates that there is evidence to reject the null hypothesis for this variable, suggesting a meaningful relationship between internet access and homicide rates. In contrast, other variables, such as GDP, urban population growth, pop growth, GDP per capita, and refugees, had p-values greater than 0.05, indicating that they did not have a significant effect on homicide rate within this model.

The results of the ANOVA highlight the potential importance of internet accessibility in influencing homicide rates. Increased internet usage may contribute to crime reduction through mechanisms such as improved education, access to resources, and communication. However, the non-significant predictors may still be relevant in broader contexts or in interaction with other actors, warranting further investigation.



### Multicollinearity Diagnostics

Multicollinearity diagnostics were conducted to evaluate the relationships among the independent variables in the dataset and to ensure the reliability of the statistical analysis. Multicollinearity occurs when two or more predictors are highly correlated, which can distort the interpretation of their individual effects on the dependent variable (homicide rate). High multicollinearity inflates the standard errors of coefficient estimates, making it difficult to assess the true impact of individual predictors in a model. To measure multicollinearity, two approaches were employed: Variance Inflation Factor (VIF) and the Condition Number.

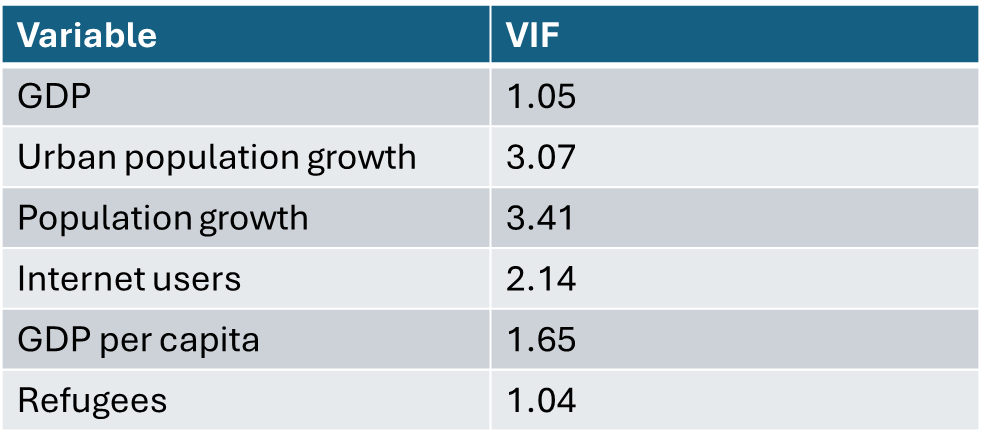
#### 1. Variance Inflation Factor (VIF)

VIF quantifies how much the variance of a regression coefficient is inflated due to multicollinearity. A VIF value greater than 5 typically indicates moderate multicollinearity, while a value greater than 10 suggests severe multicollinearity. In this analysis, all predictors (GDP, urban population growth, pop growth, internet users, GDP per capita, and refugees) had VIF values below 3.5, with the highest VIF being 3.41 for pop growth. This result confirmed that none of the predictors were excessively correlated with one another, indicating an acceptable level of multicollinearity.

#### 2. Condition Number

The Condition Number assesses the overall multicollinearity in the dataset by examining the ratio of the largest to the smallest eigenvalue of the predictor correlation matrix. A Condition Number below 30 generally indicates low multicollinearity, while a value above 30 suggests serious multicollinearity issues. In this analysis, the Condition Number was calculated to be 14.65, well below the threshold of concern. This further proved that multicollinearity was not a significant issue in the dataset.

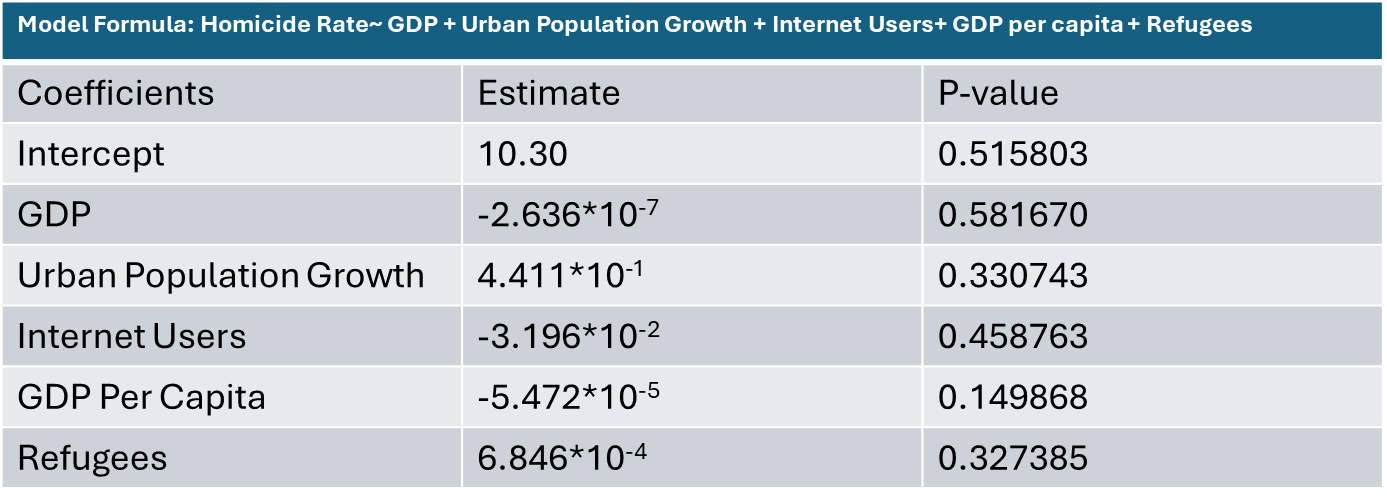
#### Implications



The multicollinearity diagnostics demonstrated that the predictors in the dataset were independent and did not provide redundant information. This ensured that the statistical model was reliable, and the results of subsequent tests, such as ANOVA, were not biased by overlapping contributions of the predictors by confirming the absence of high multicollinearity, the diagnostics supported the inclusion of all variables in the model without the need for additional transformations or removal of predictors.

### Multivariate Regression Analysis

A multiple linear regression model was fitted using all predictor variables (GDP, urban population growth, population growth, internet users, GDP per capita, and refugees) to evaluate their combined impact on homicide rates and to identify potential confounding effects. The model, specified as lm(homicide\_rate ~ gdp + urban\_population\_growth + pop\_growth + internet\_users + gdp\_per\_capita + refugees), was analyzed using the summary() function, which provided p-values for individual predictors and their coefficients. Each coefficient quantified the expected change in homicide rate for a one-unit increase in the predictor, holding other variables constant.

The results indicated that all predictors had p-values greater than 0.05, suggesting no statistically significant relationships with homicide rates in the presence of other variables. These finding highlights that none of the predictors demonstrate a strong, independent association with homicide rates when considered together, underscoring the potential need for alternative predictors or modeling approaches to better understand the factors driving homicide rates.

The multiple regression analysis demonstrates that while some predictors may show significance in isolation (as observed in the simple linear models), their collective effect is insignificant. These findings show that the variables in the model fail to explain the variability in homicide rates, suggesting that key influencing factors may be missing from the analysis. Future analyses should consider additional predictors, interaction terms, or non-linear models to capture more complex relationships and enhance the model's ability to predict homicide rates more effectively.

In conclusion, homicide rates can be calculated using GDP, internet users, GDP per capita, urban population growth, refugee population increase, and general population development. This study shows that GDP, GDP per capita, and internet use are the best predictors of homicide rates after data transformation. Insignificant data sets included refugee population, population growth, and urban population growth. The study implies that increasing GDP, internet users, and GDP per capita can reduce homicides in a nation. The study's biggest problem was data production date and lack of data points. More study can be done utilizing the country dataset by using a variety of criteria to predict other variables.

## Appendix 1

**Saadullah:** Introduction, conclusion, and Proofreader.

**Taylor:** VIF model, Multicollinearity, Condition Number, and Data Cleaning.

**Robert:** Scatter Plots and association analysis.

**Saisrijith:** Regression Analysis and Multi regression analysis.