Business Analytics Midterm Project

MGT 7104

Yacine Ndiaye

March 17th,2025

Instructor: Jaime Velasco Sanchez

Richmond The American University in London

Word Count: 1953

Data set: USArrest

TABLE OF CONTENTS

ABSTRACT

The “USArrest” dataset contains statistics regarding arrests for murder, rape, assault and Urban population in 50 states within the U.S. The data set contains 50 observations (states) and 4 variables (murder, rape, assault, urban population).

This paper will be Analysing and visualizing the information provided by the dataset . A Linear regression model will be used to explain variables and their relations as well as make predictions.

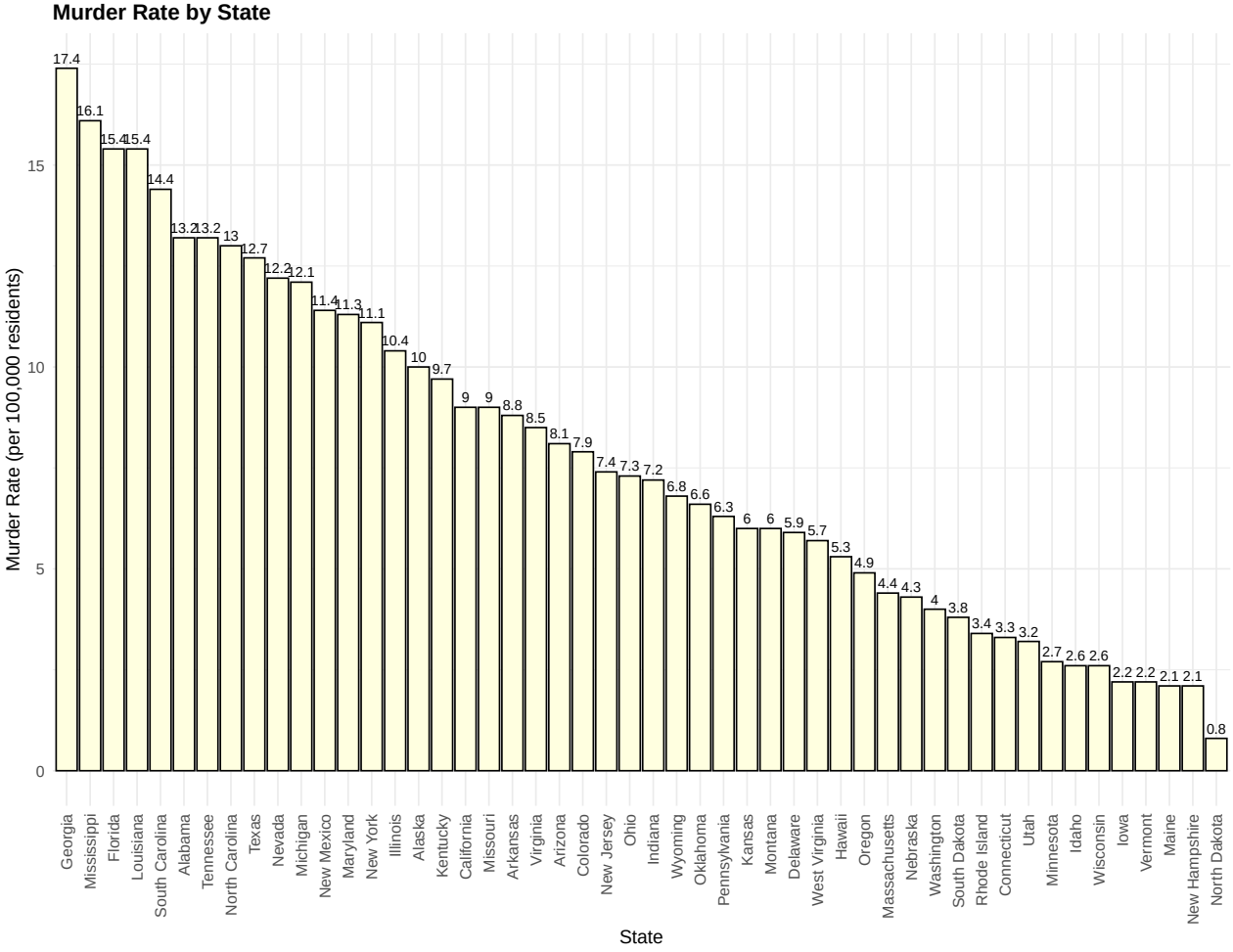
I. DATA ANALYSIS

A. Data analysis by state

This section will analyse murder, rape, assault and urban population rate based on state. Summary of the data set shows that the average murder, rape and assault rate respectively are 7.79 ; 21.23 ; 170.76 per 100,000 residents. As for the Urbanization percentage, it ranges from 32% to 91%. The purpose is to analyse variations for each variable in all 50 states.

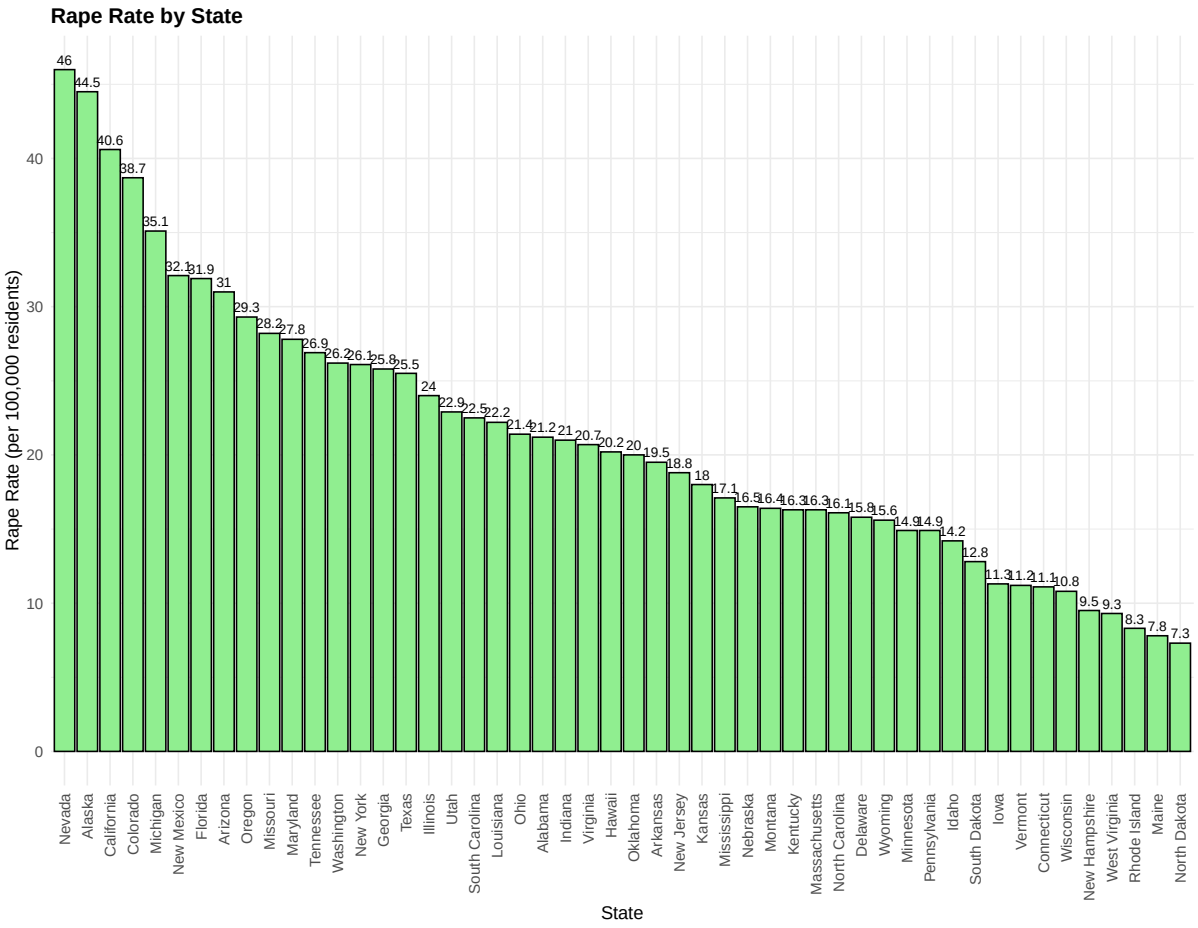
1) Murder by state

Southern states such as Georgia(17.4), Mississippi(16.1) and Florida(15.4) display high murder rates. The states with the lowest rate are Maine(0.8), New Hampshire an North-Dakota(both 2.1). The murder arrest rate is calculated per 100,000 residents within the 50 states and ranges from 0.8 to 17.4 with a mean of 7.7 and a 7.2 median.



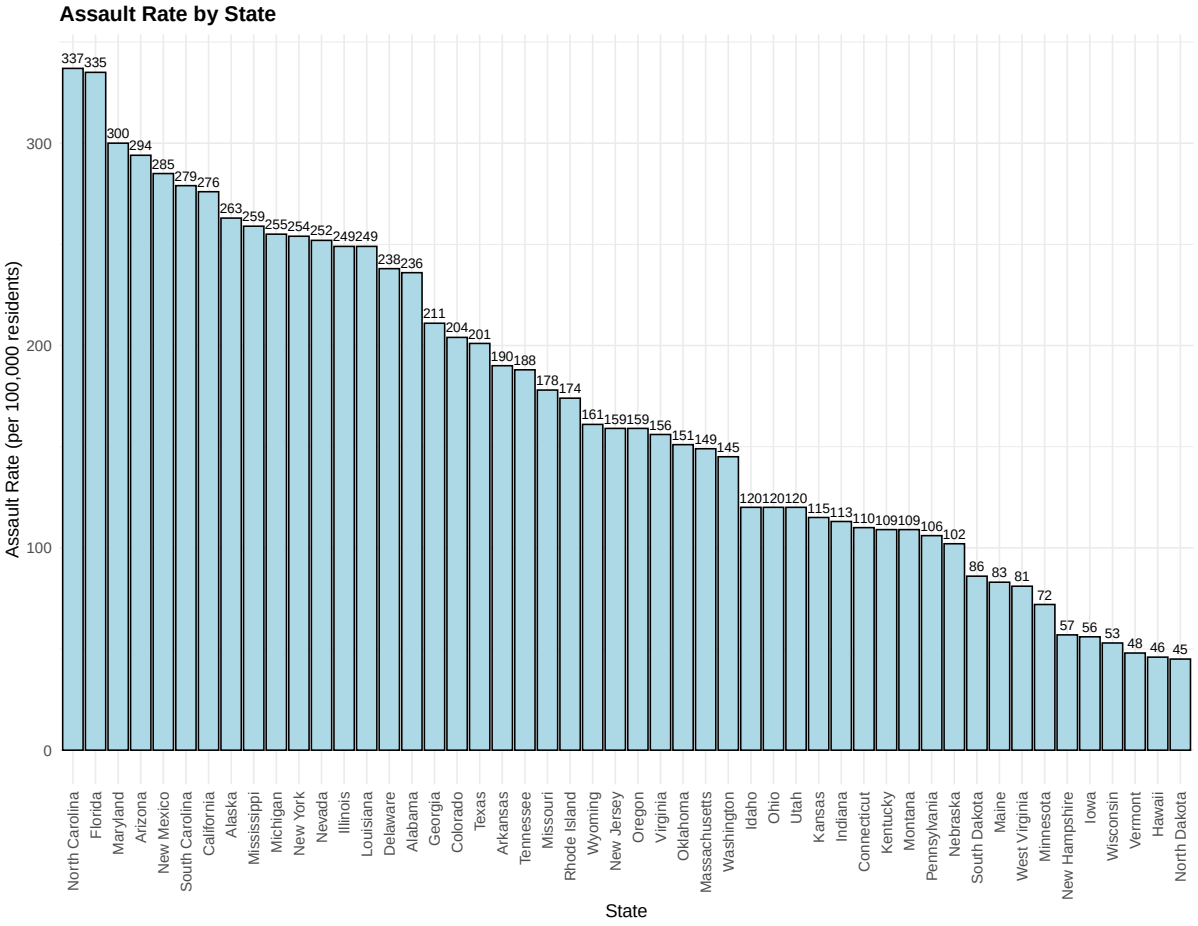
2) Rape by state

Western states such as Nevada(46), Alaska(44.5), California(40.6) and Colorado(38.7) showcase the highest rape rates. The rates gradually decrease with the lowest averages found in Rhode Island(8.3), Maine(7.8) and North Dakota(7.3). With a mean of 21.23 and a median of 20.10, the rape arrest rate ranges from 7.3 to 46 arrests per 100,000 inhabitant which is pretty low overall.



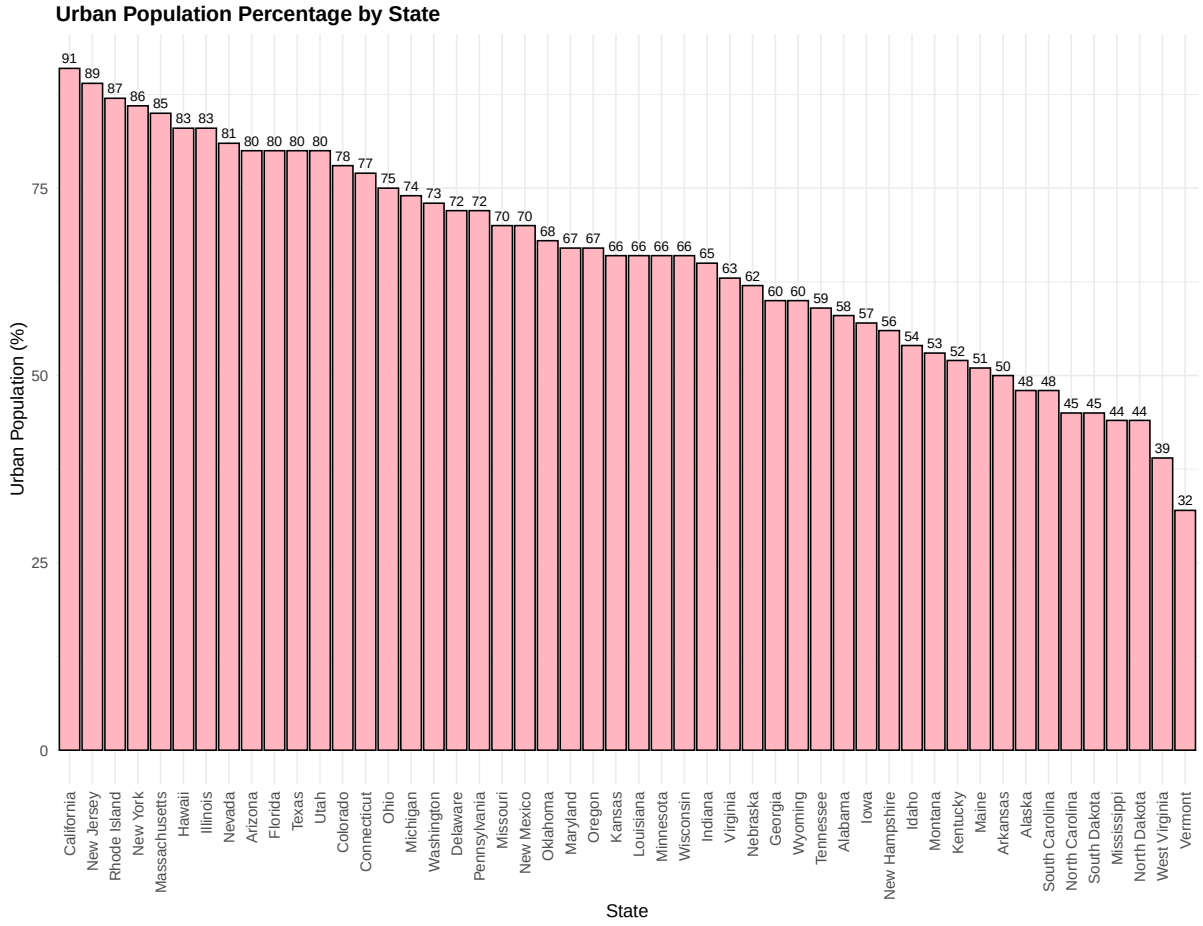
3) Assault by state

North Carolina(337) and Florida(335) hold the highest assault arrest rates. Evolution of arrest rates is gradual except for a few considerable drops here and there (Alabama, 236 and Georgia 211). Vermont(48), Hawaii(46) and North-Dakota(45) showcase fewer rates. Assault arrest rates vary from 45 to 337 with a mean of 170.8 per 100,000 residents and a median of 159.0.



4) Urban population by state

The overall Urban population Rate is high ranging from 32 to 91 with a mean of 65.54 and a median of 66.0. The highest percentages are recorded in California(91), New Jersey(89) and Rhode island (87). The graph displays low Percentage variations from state to state with lower percentiles in West Virginia(39) and Vermont(32).



B. Data analysis by urban population

This section will delve into murder rape and assault arrest rates per each category of urban population . the latter has been divided into three categories. 26 to 50; 51 to 75 and 76 to 100 urbanization percentage. the purpose is to analyse the effect of urbanization on crime rates.

1) Murder by Urban population

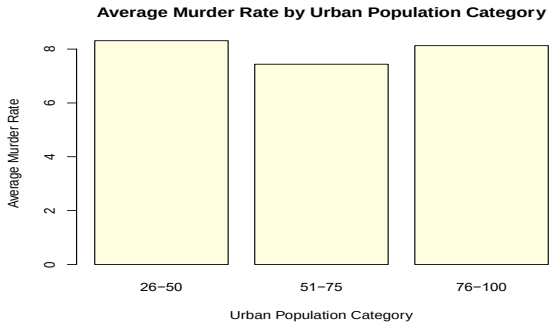
areas with 26-50% and 76-100% urban population hold the highest average murder rate (above 8). Areas with 51-75% have overall lowest average murder rate

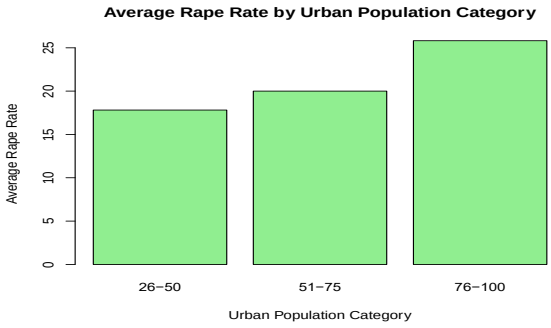
2) Rape by Urban population

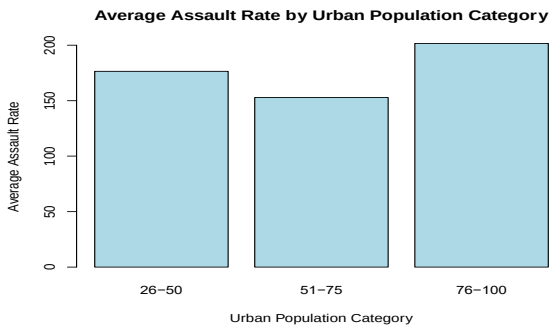
The average rape rate has a positive correlation with the percentage of urban population. The category with the lowest percentage (26-50%) presents the lowest average (below 20) and higher categories display higher rates (20 for 51-75) and (25 for 76-100)

3) Assault by Urban population

Assault rates don't positively correlate with urbanization percentiles. 26-50% hold higher average assault rate (above 150 and below 200) while areas with 51-75% urbanization percentage hold the lowest average for assault rates (150). The categories with the highest urban population however holds a positive correlation with the average assault rate (above 200)







II. LINEAR REGRESSION MODEL BY VARIABLES

The linear regression model is used to evaluate this dataset due to the continuous numerical nature of its variables (murder, assault, rape and UrbanPop). It allows for transparency and eases the modelling of relationships between variables and the prediction of outcome through simple visualization. (Kavita,2025).

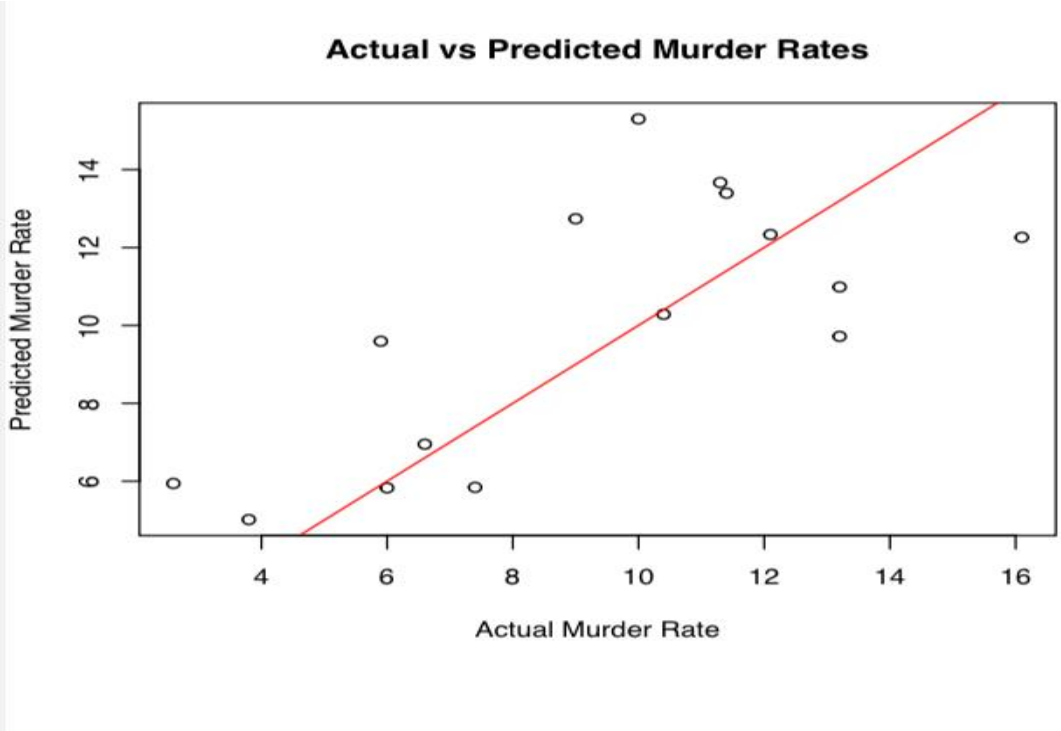
To further assess the Model performance before putting it to use, the data was split into 70% training sets and 30% test sets comparing murder (intercept) to rape, assault and UrbanPop (independent variables). Results are as Follow.

The estimate coefficient for murder is 2.9; Assault (0.03) and Rape (0.14) show a positive relationship while UrbanPop (-0.06) has minimal influence on murder rates.

P-value for murder is 0.1; That of Assault (3.32e-06) and Rape (0.11) indicates their significance as predictors on murder rate while UrbanPop (0.09) does not.

Multiple R-squared is 0.70, therefore 70% of the variance in murder rate is explained by the predictors. Adjusted R-squared is 0.67 indicating a good model fit as 67% of the model is explained after adjusting for the number of predictors.

The RSE is 2.56 on 31 degrees freedom. The typical prediction error is 2.56. the predictions provided by this model are moderately accurate.



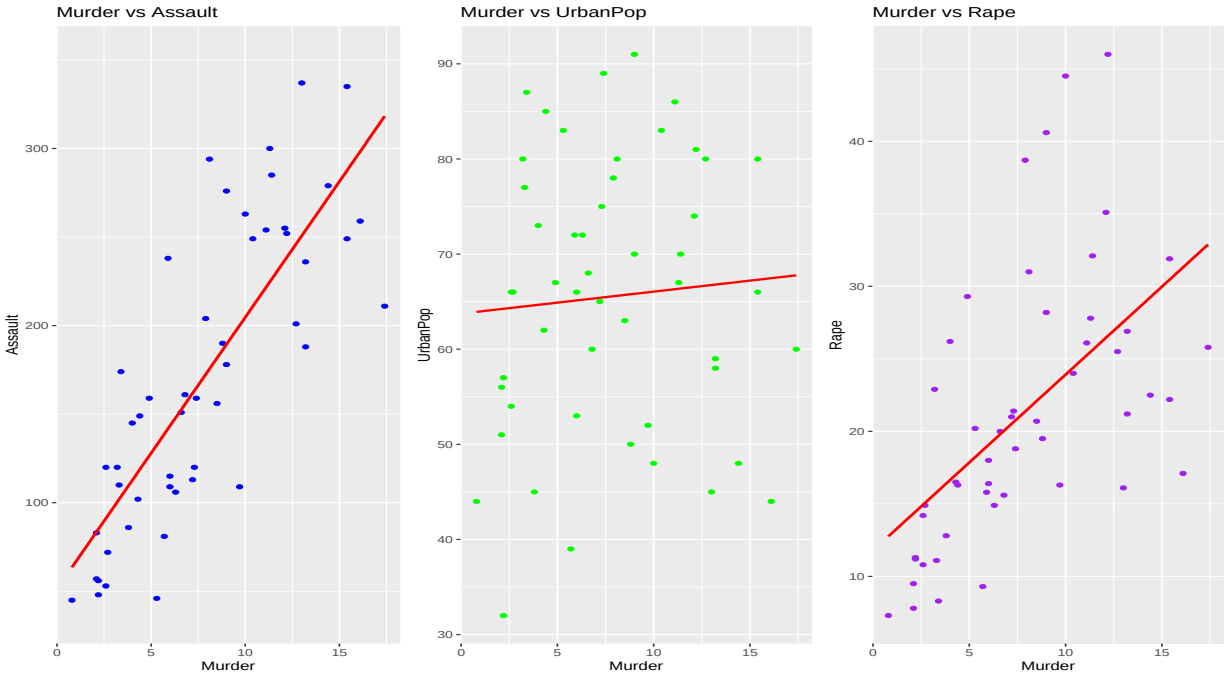
A. Linear regression by murder

Murder is a dependant variable with assault, rape and urban pop as independent predictors.

The coefficient estimate for murder is 3.2; Assault (0.03) and rape (0.06) show a positive relationship, with rape having a stronger influence. UrbanPop has a negative coefficient (-0.5), suggesting minimal influence on murder rates.

P-value for murder is 0.06. That of Urban population (0.05) and rape (0.02) shows their insignificance in predicting murder rates. As for Assault (2.33e-0.8) it is a significant predictor.

Multiple R-squared is 0.67, therefore 67% of the variance in murder rate is explained by the predictors. The adjusted R-squared is 0.65 indicating a good model fit as 65% of the model is explained after adjusting for the number of predictors.



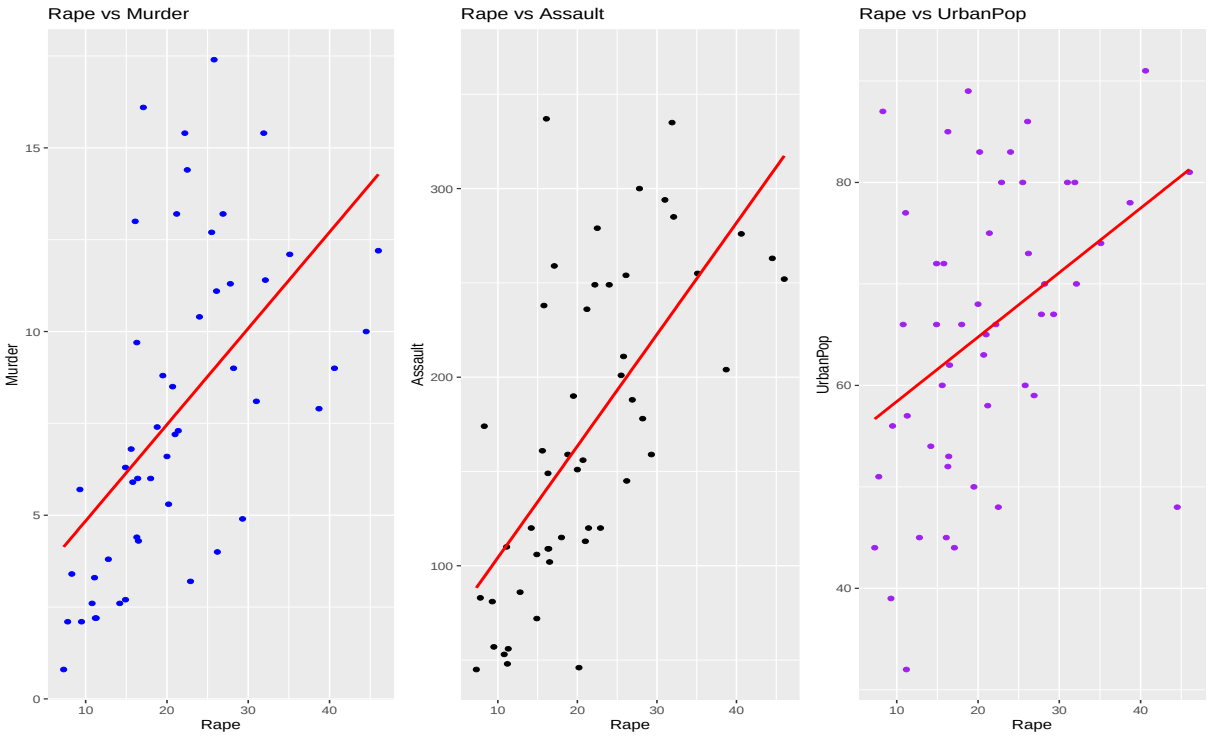
- Assault: regression line slopes upward indicating a positive corelation. Increase in assault rates will cause an increase in murder rates. Cluster points are close to the regression line showing strong relationship between two variables.
- UrbanPop: regression line is relatively flat and the points are widely scattered suggesting little to no relationship. Urban population percentage doesn't have a significant influence on the murder rate.
- Rape: the regression slope is upward and the cluster points are dispersed, suggesting a positive yet weak relationship between the two variables.

B. Linear regression by rape

Rape is a dependant variable with Murder, Assault and UrbanPop as independent predictors. its coefficient estimate is -2.4. a unit increase in murder (0.41), assault (0.04) and UrbanPop (0.18) means an increase in rape by the value of their respective coefficient.

The P-value for Rape is 0.6. Assault (0.02) and UrbanPop (0.01) are significant predictors for Rape while Murder (0.2) is not.

The Model's R-squared is 0.51, with an adjusted R-squared of 0.48 indicating that 51% of the variance is explained by the predictors and 48% after adjusting for them, suggesting moderate model significance.



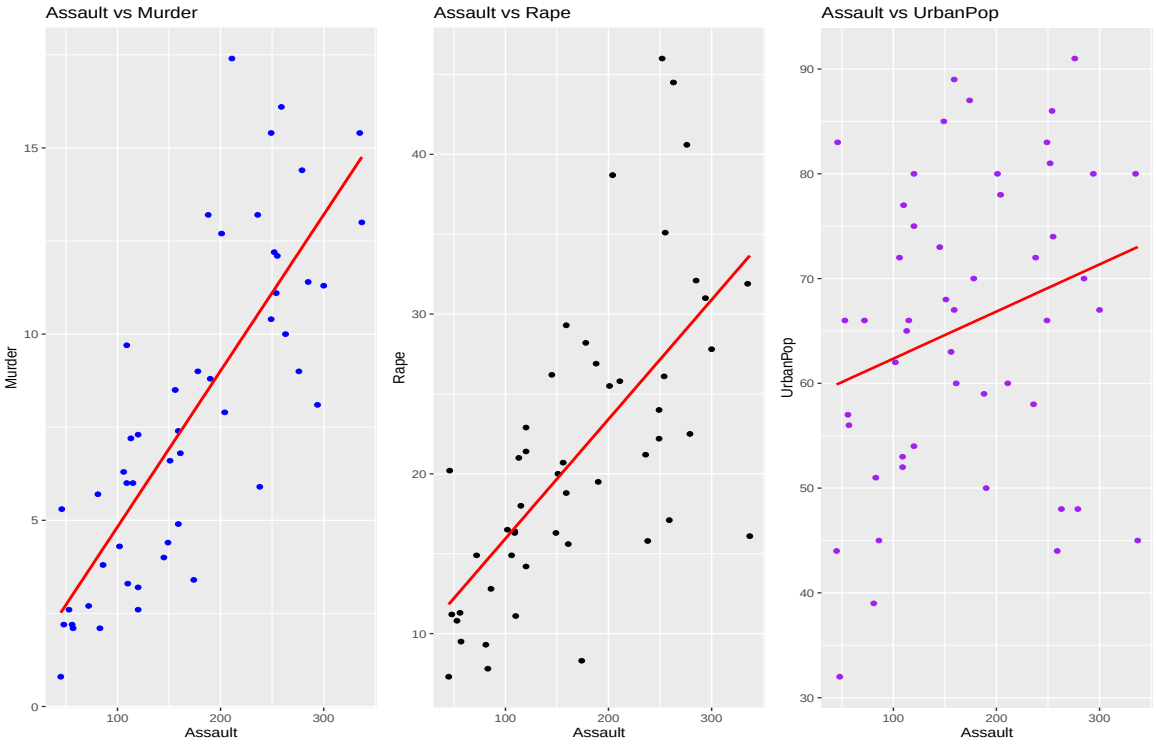
- The linear regression line for Murder and Assault present an upward slope with cluster points close to the line implying a strong positive correlation between Murder/assault rates with rape rates.
- The linear regression line for UrbanPop has a slightly flat slope with cluster points around the line although more dispersed than for murder and assault. The relationship is still positive however it's weaker than that of the previous two.

C. Linear regression by assault

Assault is a dependant variable with Murder, Rape and urban pop as independent predictors. its coefficient estimate is -15.45. a unit increase in murder(12.47), rape(2.25) and urban population(0.6) means an increase in Assault by the value of their respective coefficient.

P-value for Assault is 0.63. Murder (2.33e-08) and Rape (0.02), are significant predictors for Assault while UrbanPop (0.2) is not.

The model's R-squared is 0.71, and adjusted R-squared is 0.70, therefore 71% of the variance in Assault rate is explained by the predictor and 70% after adjustment suggesting a good model fit.



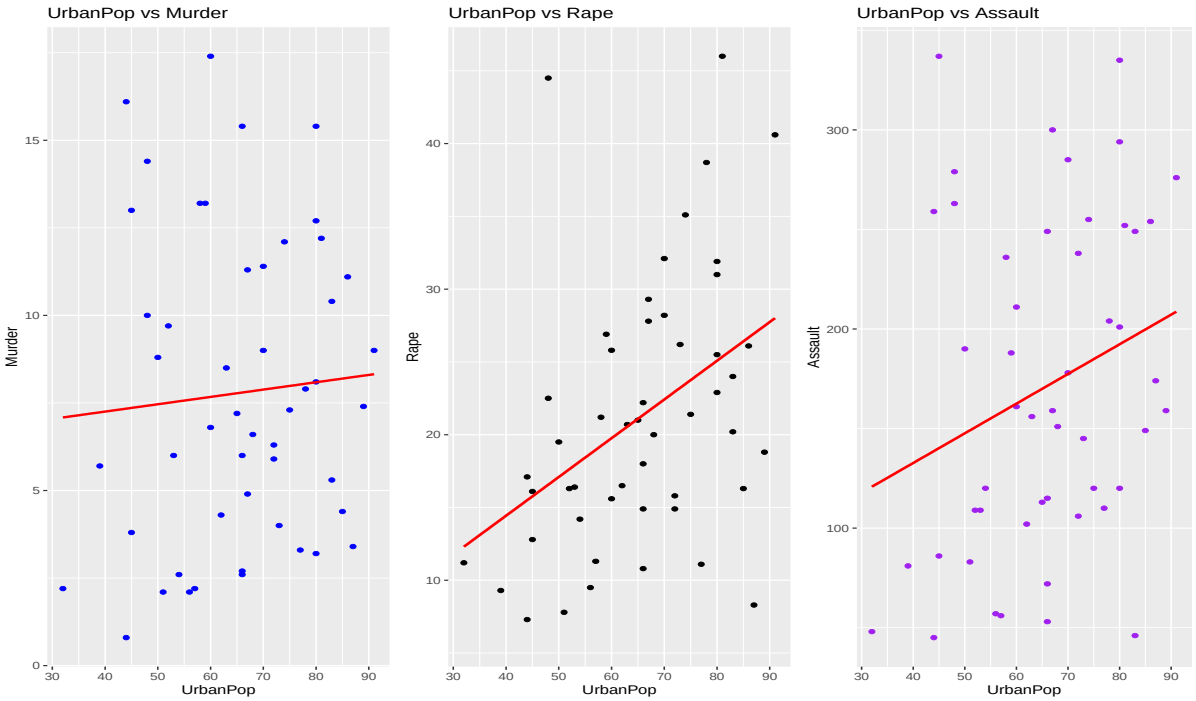
- The linear regression line for Murder and Rape present an upward slope with cluster points close to the line implying a strong positive correlation between Murder/rape rates with Assault rates.
- The linear regression line for urban pop has a relatively flat slope with cluster points widely scattered suggesting little to no relationship. Urban population percentage doesn't have a significant influence on the Assault rate.

D. Linear regression by urban population

UrbanPop is a dependant variable with Murder, Rape and Assault as independent predictors. its coefficient estimate is 52.84. Rape (0.69) and Assault (0.05)show a positive relationship, Rape even more so than Assault. Murder (-1.4) has a negative, negligible influence.

P-values for UrbanPop is 2.09e-14 (-11.90). That of Rape (0.012) shows it's a significant predictor while P-value of murder(0.05) and Assault(0.21) shows their insignificance in predicting UrbanPop percentages.

Multiple R-squared is 0.23 and adjusted R-squared is 0.18, meaning 23% of the variance in urban pop percentage is explained by the predictors and 18% after adjustment, indicating a modest model fit.

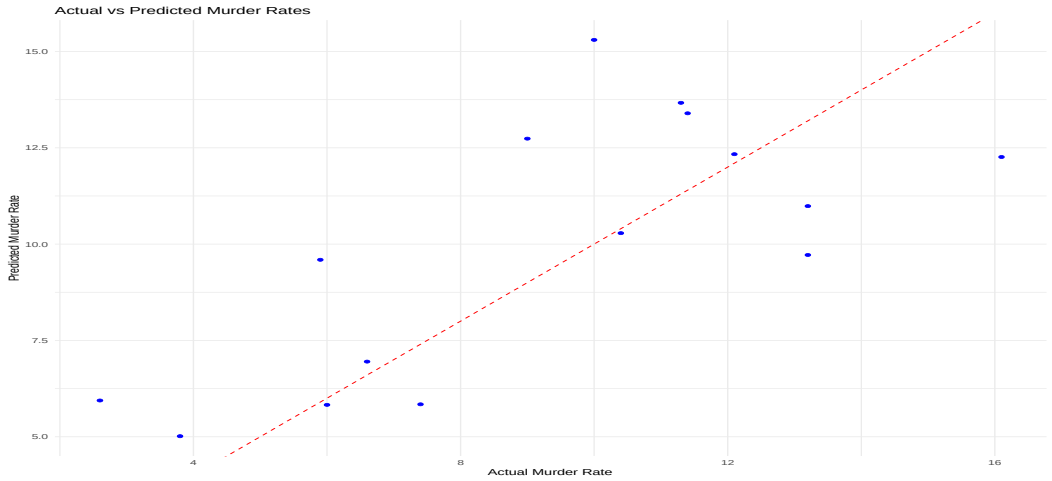


- The linear regression line for murder has a relatively flat slope with cluster points widely scattered suggesting little to no relationship. Murder rates doesn't have a significant influence on the urban population percentage.
- The linear regression line for Rape has a slightly upward slope with few cluster points around the line although dispersed enough to show a positive yet weak relationship between the two.
- The linear regression line for assault has a relatively flat slope with cluster points widely scattered suggesting little to no relationship. Assault rates don't have a significant influence on the urban population percentages.

III. PREDICTIONS BASED ON MODEL

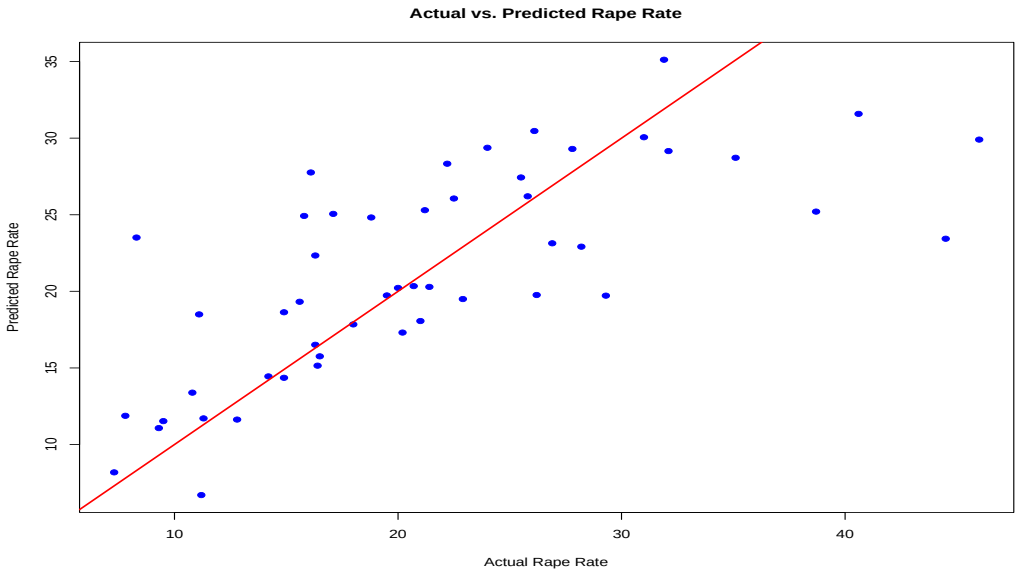
1) Predicted Murder

The RSE is 2.574 on 46 degrees freedom. The typical prediction error is 2.574 murder per 100,000 people. With an MSE of 6.0 and an RMSE of 2.4, a wide spread of cluster points suggests the model is struggling with making accurate predictions.



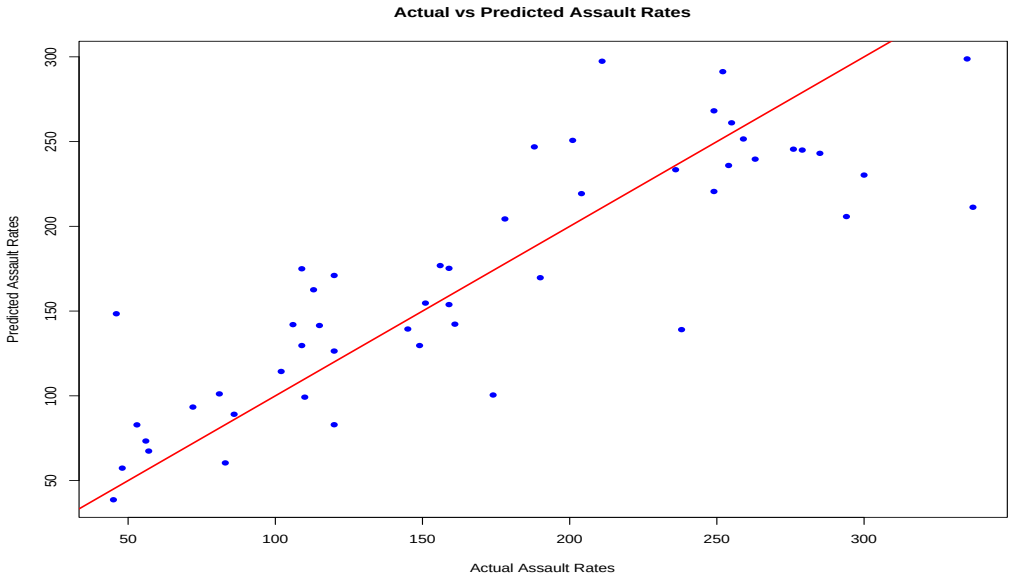
2) Predicted Rape

The RSE is 6.721 on 46 degrees freedom. The typical prediction error is 6.721 rape per 100,000 people with an MSE of 45.2 and an RMSE of 6.7. The cluster points spread indicates a moderately accurate prediction from the model



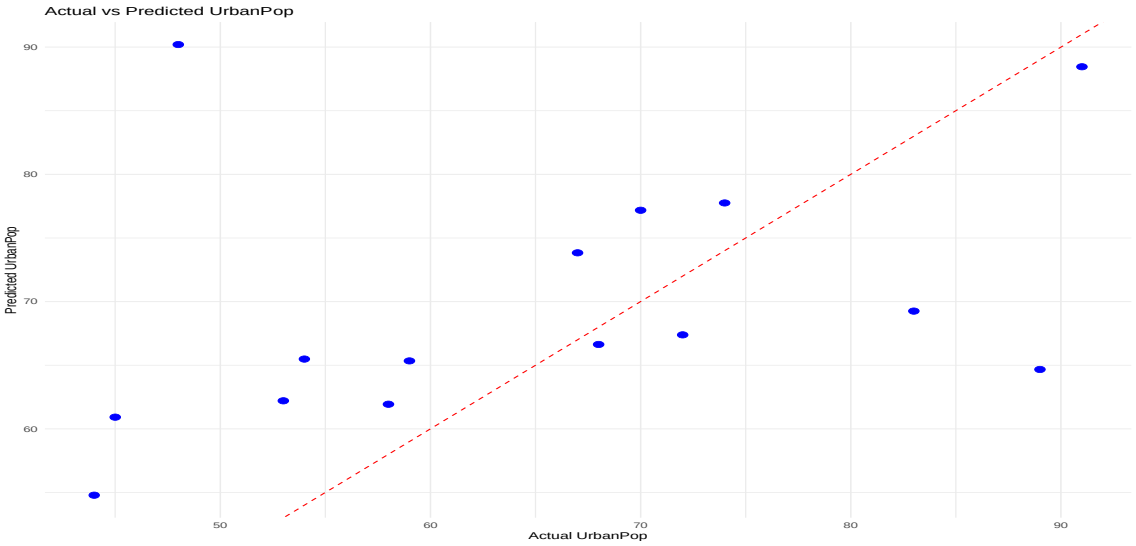
3) Predicted Assault

The RSE is 45.58 on 46 degrees freedom. The typical prediction error is 45.58 assault per 100,000 people with a 1996.20 MSE and 44.6 RMSE. The cluster point spread suggests a moderate accuracy in assault rate predictions



4) Predicted UrbanPop

The RSE is 13.08 on 46 degrees freedom. On average, the model's prediction of urban pop percentages deviate by 13.08 from the actual values with a 163.39 MSE and 12.78 RMSE. Visualization suggests a non-accurate prediction of UrbanPop percentage.



CONCLUSION

A thorough analysis on this dataset and the application of the linear regression model on each variable has brought forth a couple insights. The main one being the positive yet limited influence of urbanization on crime rates in recent years. Although studies conducted by the department of justice confirm that urban crimes victimization rate (24.5 out of 1,000 people aged 12 or older) exceed those in rural areas (11.1). (USAFactsteam, 2023), the linear regression model on urban population compared to other variable shows that its influence is limited. In this context it is safe to consider that other socio-economic factors such as poverty, inequality, unemployment etc might be more predictive of overall crime rates. Education, family structure and community resource also play an important role in shaping crime rates as all the aforementioned factors are recognized in criminology as fuel for engaging in illegal activities all under the assumptions of the economic deprivation theory. (Vargas,2023). The factors considered by the theory can also explain the reason why predictions of crime rates and urban population are moderately accurate at best. The model only analyses the influence of crime and urban population upon crime without accounting for any other outside influence on presented variables. All these extraneous variables make for non-accurate predictions.

analysis of the influence of variables upon the incidence/predictability of other variables shows that while the influence of urban population on other variables might be limited; Assault, rape and murder positively influence one another. Sexual homicide is an example of a “hybrid” crime category born form rape and murder although it’s infrequent and vastly overlooked in criminology. (Beauregard, Chopin, 2020). The influence of Assault on murder is high as shown by provided data .

This analysis has provided a deeper understanding of crime rates and urbanization in all 50 states. The linear regression model granted a clear, interpretable framework for understanding the dynamics across variables. Insights gained from this analysis can better/more accurately inform criminologists and help policy makers design targeted interventions to reduce overall crime rates.

REFERENCES

Kavita, 2025. “what is linear regression”. Analytics Vidhya. Available at: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/. Accessed 15/03/2025

USAFactsteam, 2023. “ where are crime victimization rates higher: urban or rural areas?”. Available at: https://usafacts.org/articles/where-are-crime-victimization-rates-higher-urban-rural-areas/ . Accessed: 16/03/2025

Vargas,J., 2023. “ the impact of socioeconomic factors on crime rates” Allied Academies. Available at: https://www.alliedacademies.org/articles/the-impact-of-socioeconomic-factors-oncrimerates26135.html#:~:text=Poverty%20increases%20the%20likelihood%20of,socioeconomic%20factor%20linked%20to%20crime. Accessed: 16/03/2025

Bearegard,E., Chopin,J., 2020). “ The lesser of two evils? Sexual homicide as and “hybrid” offense”. Journal of criminal Justice. Available at: https://www.sciencedirect.com/science/article/abs/pii/S0047235220302166 accessed: 16/03/2025

INDEX

UrbanPop: Urban population

RSE: residual standard error

MSE: `mean squared error

RMSE: root mean squared error

R-CODE

```
#INTRODUCTION
# STEP 1: LOAD AND VIEW DATA SET
data("USArrests")
View(USArrests)
#STEP 2: DISPLAY AND STRUCTURE
head(USArrests)
str(USArrests)
#STEP 3: SUMMARY
summary(USArrests)



# DATA VISUALIZATION 1 BY STATE
library(ggplot2)
library(gridExtra)

# ORDER STATE BY MURDER. RATE
USArrests$State <- rownames(USArrests)
USArrests <- USArrests[order(USArrests$Murder, decreasing = TRUE), ]
#BAR PLOT MURDER RATE
barplot(USArrests$Murder,
        names.arg = USArrests$State,
        las = 2,                    # Rotate state names for readability
        col = "lightyellow",
        main = "Murder Rate by State",
        ylab = "Murder Rate",
        cex.names = 0.7)            # Adjust label size
#ADVANCED BAR PLOT
ggplot(USArrests, aes(x = reorder(State, -Murder), y = Murder)) +
  geom_bar(stat = "identity", fill = "lightyellow", color = "black") +
  labs(title = "Murder Rate by State",
       x = "State",
       y = "Murder Rate (per 100,000 residents)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(face = "bold", size = 14),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)) +
  geom_text(aes(label = round(Murder, 1)), vjust = -0.5, size = 3, color = "black")



# ORDER STATE BY RAPE RATE
USArrests$State <- rownames(USArrests)
USArrests <- USArrests[order(USArrests$Rape, decreasing = TRUE), ]
```

```
# BAR PLOT RAPE RATE
barplot(USArrests$Rape,
        names.arg = USArrests$State,
        las = 2,                  # Rotate state names for readability
        col = "lightgreen",
        main = "Rape Rate by State",
        ylab = "Rape Rate",
        cex.names = 0.7)          # Adjust label size
#ADVANCED BAR PLOT
USArrests$State <- rownames(USArrests)
ggplot(USArrests, aes(x = reorder(State, -Rape), y = Rape)) +
  geom_bar(stat = "identity", fill = "lightgreen", color = "black") +
  labs(title = "Rape Rate by State",
       x = "State",
       y = "Rape Rate (per 100,000 residents)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1),
        plot.title = element_text(face = "bold", size = 14),
        axis.title = element_text(size = 12),
        axis.text = element_text(size = 10)) +
  geom_text(aes(label = round(Rape, 1)), vjust = -0.5, size = 3, color = "black")
library(ggplot2)



# ORDER STATE BY ASSAULT RATE
USArrests$State <- rownames(USArrests)
USArrests <- USArrests[order(USArrests$Assault, decreasing = TRUE), ]
#BAR PLOT ASSAULT RATE
barplot(USArrests$Assault,
        names.arg = USArrests$State,
        las = 2,                  # Rotate state names for readability
        col = "lightblue",
        main = "Assault Rate by State",
        ylab = "Assault Rate",
        cex.names = 0.7)          # Adjust label size
#ADVANCED BAR PLOT
ggplot(USArrests, aes(x = reorder(State, -Assault), y = Assault)) +
  geom_bar(stat = "identity", fill = "lightblue", color = "black") +
  labs(title = "Assault Rate by State",
       x = "State",
       y = "Assault Rate (per 100,000 residents)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
        plot.title = element_text(face = "bold", size = 14),
        axis.title = element_text(size = 12),
```

```
      axis.text = element_text(size = 10)) +
  geom_text(aes(label = round(Assault, 1)), vjust = -0.5, size = 3, color = "black")
library(ggplot2)

# ORDER STATE BY ASSAULT RATE
USArrests$State <- rownames(USArrests)
USArrests <- USArrests[order(USArrests$UrbanPop, decreasing = TRUE), ]
#BAR PLOT ASSAULT RATE
barplot(USArrests$Assault,
      names.arg = USArrests$State,
      las = 2,                  # Rotate state names for readability
      col = "lightpink",
      main = "Urbanpop Rate by State",
      ylab = "Urbanpop Rate",
      cex.names = 0.7)          # Adjust label size
#ADVANCED BAR PLOT
ggplot(USArrests, aes(x = reorder(State, -UrbanPop), y = UrbanPop)) +
  geom_bar(stat = "identity", fill = "lightpink", color = "black") +
  labs(title = "Urban Population Percentage by State",
      x = "State",
      y = "Urban Population (%)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1),
      plot.title = element_text(face = "bold", size = 14),
      axis.title = element_text(size = 12),
      axis.text = element_text(size = 10)) +
  geom_text(aes(label = round(UrbanPop, 1)), vjust = -0.5, size = 3, color = "black")
library(ggplot2)

#DATA VISUALIZATION 2 BY URBANPOP
#MURDER BY URBANPOP
# Create a new column for UrbanPop categories (binned)
USArrests$UrbanPopCategory <- cut(USArrests$UrbanPop,
                  breaks = c(0, 25, 50, 75, 100),
                  labels = c("0-25", "26-50", "51-75", "76-100"))
# Calculate the average Murder rate for each UrbanPop category
avg_murder_by_urbanpop <- aggregate(Murder ~ UrbanPopCategory, data = USArrests, FUN = mean)
# Create the barplot
barplot(avg_murder_by_urbanpop$Murder,
      names.arg = avg_murder_by_urbanpop$UrbanPopCategory,
      main = "Average Murder Rate by Urban Population Category",
      xlab = "Urban Population Category",
      ylab = "Average Murder Rate",
      col = "lightyellow")
```

```
#RAPE BY URBANPOP
# Calculate the average Rape rate for each UrbanPop category
avg_rape_by_urbanpop <- aggregate(Rape ~ UrbanPopCategory, data = USArrests, FUN = mean)
# Create the barplot
barplot(avg_rape_by_urbanpop$Rape,
        names.arg = avg_rape_by_urbanpop$UrbanPopCategory,
        main = "Average Rape Rate by Urban Population Category",
        xlab = "Urban Population Category",
        ylab = "Average Rape Rate",
        col = "lightgreen")

#ASSAULT BY URBANPOP
# Calculate the average Assault rate for each UrbanPop category
avg_assault_by_urbanpop <- aggregate(Assault ~ UrbanPopCategory, data = USArrests, FUN = mean)
# Create the bar plot
barplot(avg_assault_by_urbanpop$Assault,
        names.arg = avg_assault_by_urbanpop$UrbanPopCategory,
        main = "Average Assault Rate by Urban Population Category",
        xlab = "Urban Population Category",
        ylab = "Average Assault Rate",
        col = "lightblue")


################################################################################
# TEST AND TRAIN LINEAR REGRESSION MODEL
# Set a seed for reproducibility
set.seed(123)
# Split data into training (70%) and test (30%) sets
sample_index <- sample(1:nrow(USArrests), size = 0.7 * nrow(USArrests))
train_data <- USArrests[sample_index, ]
test_data <- USArrests[-sample_index, ]
# Fit a linear regression model on the training data
model <- lm(Murder ~ Assault + UrbanPop + Rape, data = train_data)
# Display the summary of the model
summary(model)
# Predict on the test set
predictions <- predict(model, newdata = test_data)
# View the first few predictions
head(predictions)
# Calculate Mean Squared Error (MSE)
mse <- mean((predictions - test_data$Murder)^2)
# Print MSE
print(paste("Mean Squared Error: ", mse))
# Plot Actual vs Predicted values
plot(test_data$Murder, predictions,
```

```
    main = "Actual vs Predicted Murder Rates",
    xlab = "Actual Murder Rate", ylab = "Predicted Murder Rate")
abline(0, 1, col = "red")  # Add a line for perfect predictions

################################################################################


# LINEAR REGRESSION MODEL ON THE MURDER VARIABLE
# Fit the linear regression model
model_lr <- lm(Murder ~ Assault + UrbanPop + Rape, data = USArrests)
# Show the summary of the model
summary(model_lr)
#MURDER VS OTHER VARIABLES+ VISUALIZATION
# Create scatterplots using ggplot2
p1 <- ggplot(USArrests, aes(x = Murder, y = Assault)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Murder vs Assault")

p2 <- ggplot(USArrests, aes(x = Murder, y = UrbanPop)) +
  geom_point(color = "green") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Murder vs UrbanPop")

p3 <- ggplot(USArrests, aes(x = Murder, y = Rape)) +
  geom_point(color = "red") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Murder vs Rape")
# Arrange the plots in a 2x2 grid
grid.arrange(p1, p2, p3,ncol = 3)


#PREDICTION
# Fit the linear regression model (if not already done)
model <- lm(Murder ~ Assault + UrbanPop + Rape, data = train_data)
# Make predictions on the test set
predictions <- predict(model, newdata = test_data)
# Create a data frame for actual and predicted values
comparison <- data.frame(Actual = test_data$Murder, Predicted = predictions)
# Plot actual vs predicted murder rates
library(ggplot2)
ggplot(comparison, aes(x = Actual, y = Predicted)) +
  geom_point(color = 'blue') +
  geom_abline(intercept = 0, slope = 1, color = 'red', linetype = 'dashed') +
  labs(title = "Actual vs Predicted Murder Rates",
```

```
      x = "Actual Murder Rate",
      y = "Predicted Murder Rate") +
  theme_minimal()

#PREDICTION USING FITTED MODEL
# Predict on the existing dataset
predictions <- predict(model_lr, newdata = USArrests)
# View the first few predictions
head(predictions)
# Calculate MSE and RMSE
mse <- mean((USArrests$Murder - predictions)^2)
rmse <- sqrt(mse)
# Print MSE and RMSE
cat("MSE:", mse, "\n")
cat("RMSE:", rmse, "\n")
###############################################################################

# LINEAR REGRESSION MODEL ON THE RAPE VARIABLE
# Fit the linear regression model for Rape
model_lr_rape <- lm(Rape ~ Assault + UrbanPop + Murder, data = USArrests)
# Show the summary of the model
summary(model_lr_rape)
# Create scatterplots using ggplot2
p1 <- ggplot(USArrests, aes(x = Rape, y = Murder)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Rape vs Murder")

p2 <- ggplot(USArrests, aes(x = Rape, y = Assault)) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Rape vs Assault")

p3 <- ggplot(USArrests, aes(x = Rape, y = UrbanPop)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Rape vs UrbanPop")
# Arrange the plots in a 2x2 grid
grid.arrange(p1, p2, p3,ncol = 3)

#PREDICTIONS
# Fit the model using Rape as the dependent variable
model_rape <- lm(Rape ~ Murder + Assault + UrbanPop, data = USArrests)
# Make predictions
predicted_rape <- predict(model_rape, newdata = USArrests)
```

```
# Create the plot
plot(USArrests$Rape, predicted_rape,
     xlab = "Actual Rape Rate",
     ylab = "Predicted Rape Rate",
     main = "Actual vs. Predicted Rape Rate",
     pch = 19, col = "blue")
# Add a diagonal reference line
abline(0, 1, col = "red", lwd = 2)

#PREDICTION USING FITTED MODEL
# Fit the model (renaming it as model_lr_Rape)
model_lr_rape <- lm(Rape ~ Murder + Assault + UrbanPopCategory, data = train_data)
# Predict the Rape rate using the model
predictions_rape <- predict(model_lr_rape, newdata = USArrests)
# View the first few predictions
head(predictions_rape)
# Calculate MSE and RMSE for the Rape predictions
mse_rape <- mean((USArrests$Rape - predictions_rape)^2)
rmse_rape <- sqrt(mse_rape)
# Print MSE and RMSE
cat("MSE for Rape:", mse_rape, "\n")
cat("RMSE for Rape:", rmse_rape, "\n")
# Show the summary of the model
summary(model_lr_rape)
################################################################################

# LINEAR REGRESSION MODEL ON THE ASSAULT VARIABLE
# Fit the linear regression model for Rape
model_lr_Assault <- lm(Assault ~ Rape + UrbanPop + Murder, data = USArrests)
# Show the summary of the model
summary(model_lr_Assault)
# Create scatterplots using ggplot2
p1 <- ggplot(USArrests, aes(x = Assault, y = Murder)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Assault vs Murder")
p2 <- ggplot(USArrests, aes(x = Assault, y = Rape)) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Assault vs Rape")
p3 <- ggplot(USArrests, aes(x = Assault, y = UrbanPop)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("Assault vs UrbanPop")
# Arrange the plots in a 2x2 grid
```

```
grid.arrange(p1, p2, p3,ncol = 3)


#PREDICTIONS
# Fit a linear model for Assault
model_assault <- lm(Assault ~ Murder + UrbanPop + Rape, data = USArrests)
# Make predictions for Assault
predicted_assault <- predict(model_assault)
# Plot actual vs predicted Assault rates
plot(USArrests$Assault, predicted_assault,
    xlab = "Actual Assault Rates", ylab = "Predicted Assault Rates",
    main = "Actual vs Predicted Assault Rates",
    pch = 16, col = "blue")
# Add a 45-degree line (perfect prediction line)
abline(0, 1, col = "red", lwd = 2)

#PREDICTION USING FITTED MODEL
# Fit the model (renaming it as model_lr_Assault)
model_lr_Assault <- lm(Assault ~ Murder + Rape + UrbanPopCategory, data = train_data)
# Predict the Assault rate using the model
predictions_assault <- predict(model_lr_Assault, newdata = USArrests)
# View the first few predictions
head(predictions_assault)
# Calculate MSE and RMSE for the Assault predictions
mse_assault <- mean((USArrests$Assault - predictions_assault)^2)
rmse_assault <- sqrt(mse_assault)
# Print MSE and RMSE
cat("MSE for Assault:", mse_assault, "\n")
cat("RMSE for Assault:", rmse_assault, "\n")
# Show the summary of the model
summary(model_lr_Assault)

################################################################################

# LINEAR REGRESSION MODEL ON THE URBANPOP VARIABLE
# Fit the linear regression model for UrbanPop
model_lr_urbanpop <- lm(UrbanPop ~ Murder + Assault + Rape, data = USArrests)
# Show the summary of the model
summary(model_lr_urbanpop)
# Create scatterplots using ggplot2
p1 <- ggplot(USArrests, aes(x = UrbanPop, y = Murder)) +
  geom_point(color = "blue") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("UrbanPop vs Murder")
```

```
p2 <- ggplot(USArrests, aes(x = UrbanPop, y = Rape)) +
  geom_point(color = "black") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("UrbanPop vs Rape")

p3 <- ggplot(USArrests, aes(x = UrbanPop, y = Assault)) +
  geom_point(color = "purple") +
  geom_smooth(method = "lm", se = TRUE, color = "red") +
  ggtitle("UrbanPop vs Assault")
# Arrange the plots in a 2x2 grid
grid.arrange(p1, p2, p3,ncol = 3)

#PREDICTIONS
# Fit linear model to predict UrbanPop
model_urban <- lm(UrbanPop ~ Murder + Assault + Rape, data = train_data)
# Make predictions
predicted_urban <- predict(model_urban, newdata = test_data)
# Create the data frame for plotting
plot_data <- data.frame(
  Actual = test_data$UrbanPop,
  Predicted = predicted_urban)
# Plotting actual vs predicted
ggplot(plot_data, aes(x = Actual, y = Predicted)) +
  geom_point(color = "blue", size = 3) +                # Actual vs. Predicted points
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +  # Ideal fit line
  labs(title = "Actual vs Predicted UrbanPop",
       x = "Actual UrbanPop",
       y = "Predicted UrbanPop") +
  theme_minimal()

#PREDICTION USING FITTED MODEL
# Fit the model (renaming it as model_lr_urbanpop)
model_lr_urbanpop <- lm(UrbanPop ~ Murder + Assault + Rape, data = train_data)
# Predict UrbanPop using the fitted model
predictions_urbanpop <- predict(model_lr_urbanpop, newdata = USArrests)
# View the first few predictions
head(predictions_urbanpop)
# Calculate MSE and RMSE
mse_urbanpop <- mean((USArrests$UrbanPop - predictions_urbanpop)^2)
rmse_urbanpop <- sqrt(mse_urbanpop)
# Print MSE and RMSE
cat("MSE for UrbanPop:", mse_urbanpop, "\n")
cat("RMSE for UrbanPop:", rmse_urbanpop, "\n")

##############################################################################
```