Crime Data Analysis Report

Report

a) Problem Selected

The problem I have selected and am looking to analyze focuses on crime prevention

through state-level risk analysis. The primary objective is to identify socioeconomic, community,

and psychological factors that correlate with higher violent crime rates across U.S. states. This

includes variables such as the percentage of single-parent households, substance use disorder

prevalence, unemployment rates, and mental health conditions like depression. By examining

how these factors relate to violent crime per 100,000 residents, this analysis can inform more

targeted, data-driven prevention strategies that benefit both communities and public safety
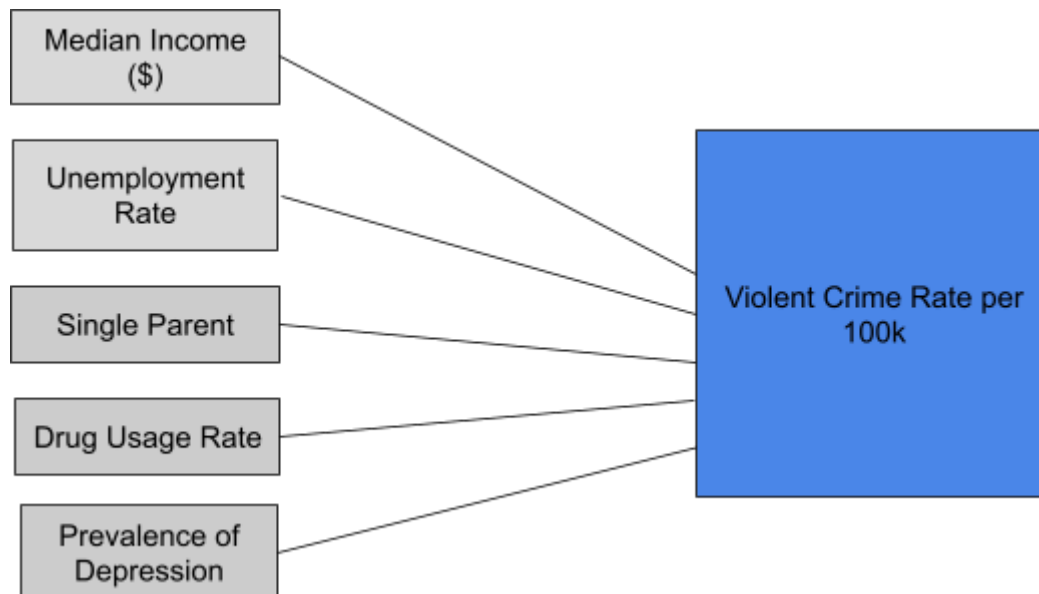
agencies.

b) Model Explained

This study will utilize a multi-linear regression model to analyze how various

socioeconomic, environmental, and psychological factors contribute to state-level violent crime

rates. With multiple independent variables and a single dependent variable (violent crime per

100,000 residents) this model is appropriate for identifying significant predictors and assessing

the strength of their relationship to crime rates across states.

Regression Equation:

$(\text{Violent Crime Rate per 100k})_i = B_0 + (\text{Median Income})_i B_1 + (\text{UnemploymentRate})_i B_2 + (\text{Single-Parent Households})_i B_3 + (\text{Drug Usage})_i B_4 + (\text{Depression Rate})_i B_{65}$

Diagram:

c) **Variables**

Independent Variables (x):

    i)   **Socioeconomic Factors:**

        1) Median Household Income ($- US Dollars)

        2) Unemployment Rate (% of population)

    ii)   **Community/Environment Factors**

        1) Percntages of Single-Parent Households (% of State Population)

2) Drug Usage Rate (% of population)

    iii)   Psychological Factors:

        1) Prevalence of Depression (% of Community)

Dependent Variable (y)

    iv)   Violent Crime Rate per 100,000 residents (By state)

All variables are in the numeric category. The independent variables include Socioeconomic Factors: Median Household Income (USD), Unemployment Rate (% of the population); Community/Environmental Factors: Percentage of Single-Parent Households (% of the state population), Crime Rate, Drug Usage Rate (% of the population); and Psychological Factors: Prevalence of Depression (% of the community). The dependent variable is theviolent crime rate per 100,000 residents, measured by state level.

    d) **Data Source**

Independent Variables (x):

- **Median Household Income by State:**

  historical-income-households.html

- **Children in Signle-Parent Households by State**

  https://datacenter.aecf.org/data/tables/106-children-in-single-parent-families?loc=1&loct=1#detailed/2/2-53/true/2545,1095,2048,1729,37,871,870,573,869,36/any/429

- **Drug Usage and Mental Illnesses by State:**

  https://www.samhsa.gov/data/data-we-collect/nsduh-national-survey-drug-use-and-health/state-releases/2022-2023

- **Unemployment Rates**

  https://www.bls.gov/web/laus/laumstrk.htm

Dependent Variable (Y)

- **Violent Crime Rate per 100,000 residents**

  https://www.visualcapitalist.com/mapped-violent-crime-rates-by-u-s-state/

  e) **Curation Process**

  To curate the data analysis process, the data must be cleaned and structured in a clear, organized format for easy interpretation and tabulation. When selecting a file format for data extraction, several options were available but I ultimately went with Excel as it best suited my needs for the analysis. As most of my sources came from government sources or other government agencies missing data or inputs wasn't necessarily a concern. However, changing or
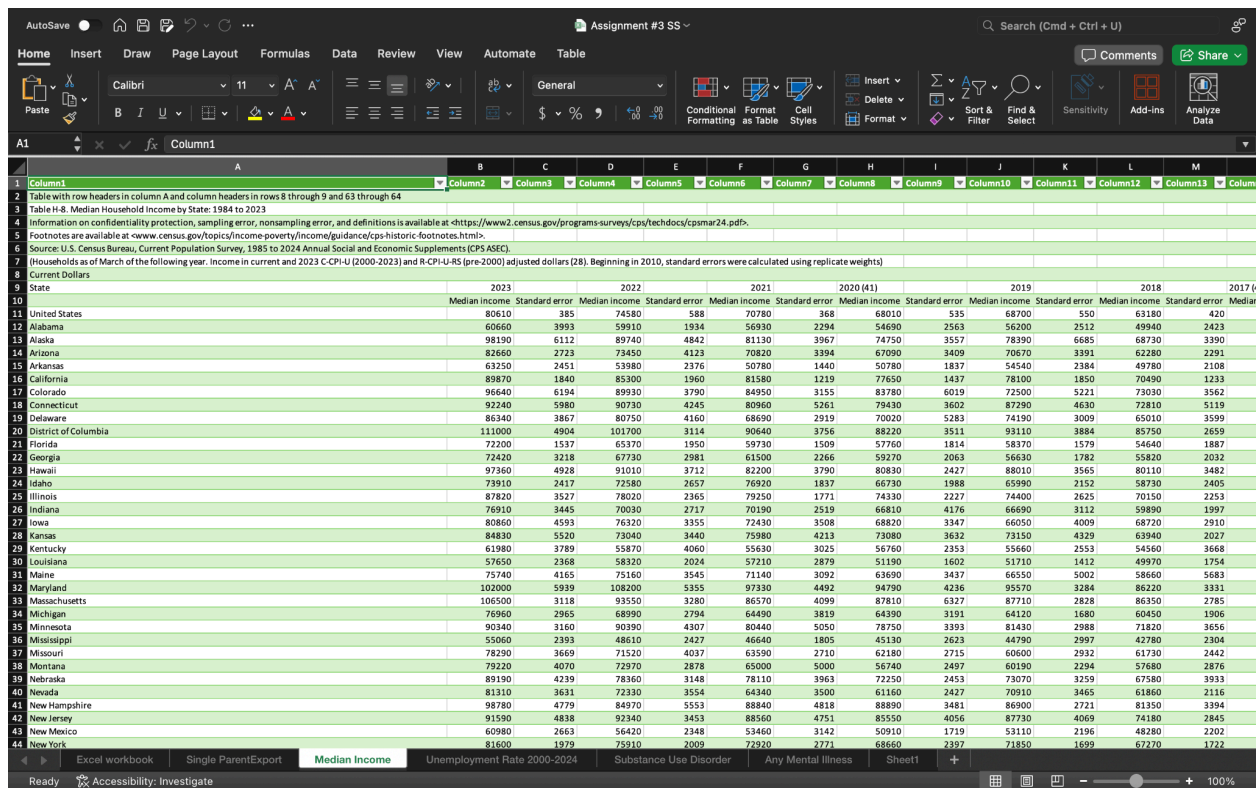
deleting variables was an issue as some datasets included percentages and numbers (photo listed

below). As I am looking to focus on percentages as most of my data is in percentages making

sure to clean the data is crucial. Additionally, as these sets provided data from a long range of

years I wanted to focus on more recent years. I'm looking to focus on a range from 2022-2023

this was the latest data for most sets at the moment. If needed taking into account a multi-year

approach can also be an alternative as it can help pick up on trends.

(Raw Data Single Parent Households)

As seen above although these years aren't the ones I'm using it provides a perspective as

to not only how much data is in this set but also as stated how it includes both percentages and

numbers. This is displaying single-parent households and using numbers wouldn't do justice as

the number of single-parent households will naturally be higher in more populous areas. Using

percentages allows for a fair comparison across different states. The data is also not listed year

by year for example 2001, 2002, 2003, etc. and there are times when it jumps from 2009 to 2018 and then back to 2011 so ensuring the years are in chronological order is highly beneficial for analysis. Also, deleting the years I will not be using is crucial as that is not what I'm looking to observe.

(Raw Data of Median Household Income)

As stated this examples provides context into the range of years that are listed this dataset is showing median income all the way back from 1985. Utilizing data from that long ago wouldn't be beneficial as not only pay rates raised but many different societal factors have also taken place in this timeframe.

Furthermore, these are only two of the six datasets that need to be curated but they paint a picture as to how the other sets looks. As mentioned I will have to delta rows in columns to not only display percentages but also make sure the data is from the right range of years.

Excel

- 1st Excel Sheet: Curated data with proper years, proper data needed, number in percentages etc.

- 2nd Sheet Substance Abuse Disorders 2022-2023, 3rd Sheet: Any Mentall illness 2022-2023, 4th Sheet: Single Parent Households 2000-2023, 5th Sheet: Median Income 1984-2023, 6th Sheet: Unemployment Rates 2012-2024

Summary of Data Curation:

1. Identify the problem

2. Find necessary data

3. Import data

4. Make changes where needed and organize

5. Ensure Consistency

Assignment 4 Section

For this assignment, the data was curated and analysed in Excel as seen above I made sure to remove missing data or change values in order for them to be analysed properly and statistical analysis to be conducted.

f) Correlations found between different input variables

State	Year	Single Parent Households	Median Income	Unemployment Rate	Substance Use Disorder (18+ Estimate)	Any Mental Disorder (18+ Estimate)	Violent Crime Per 100k
Alabama	2022	0.37	59910.00	2.50	0.17	0.22	409.00
Alaska	2022	0.34	89740.00	4.30	0.23	0.23	759.00
Arizona	2022	0.37	73450.00	3.80	0.19	0.26	431.00
Arkansas	2022	0.36	53980.00	3.20	0.19	0.24	645.00
California	2022	0.34	85300.00	4.30	0.18	0.21	499.00
Colorado	2022	0.28	89930.00	3.10	0.24	0.28	492.00
Connecticut	2022	0.34	90730.00	4.10	0.20	0.24	150.00
Delaware	2022	0.38	80750.00	4.20	0.19	0.22	384.00
District of Columbia	2022	0.49	101700.00	4.70	0.25	0.29	812.00
Georgia	2022	0.39	65370.00	3.00	0.16	0.20	259.00
Florida	2022	0.38	67730.00	3.20	0.18	0.23	367
Hawaii	2022	0.35	91010.00	3.30	0.18	0.21	260.00
Idaho	2022	0.23	72580.00	2.90	0.18	0.27	241.00
Indiana	2022	0.32	78020.00	4.60	0.18	0.23	287.00
Illinois	2022	0.34	70030.00	3.10	0.16	0.24	306.00
Iowa	2022	0.28	76320.00	2.80	0.17	0.26	287.00
Kansas	2022	0.29	73040.00	2.70	0.17	0.24	415.00
Louisiana	2022	0.46	55870.00	4.00	0.19	0.25	214.00
Kentucky	2022	0.34	58320.00	3.70	0.21	0.23	629.00
Maine	2022	0.32	75160.00	2.90	0.21	0.26	103.00
Maryland	2022	0.33	108200.00	3.00	0.16	0.22	398.00
Massachusetts	2022	0.32	93550.00	3.60	0.21	0.24	322.00
Michigan	2022	0.34	68990.00	4.20	0.19	0.23	461.00
Minnesota	2022	0.27	90390.00	2.50	0.19	0.24	281.00
Missouri	2022	0.33	48610.00	3.80	0.17	0.20	245.00
Mississippi	2022	0.44	71520.00	2.60	0.21	0.26	488.00
Montana	2022	0.28	72970.00	2.70	0.21	0.25	418.00
Nebraska	2022	0.28	78360.00	2.20	0.17	0.24	283.00
Nevada	2022	0.4	72330.00	5.20	0.22	0.23	454.00
New Hampshire	2022	0.29	84970.00	2.40	0.19	0.27	126.00
New Jersey	2022	0.3	92340.00	3.90	0.16	0.20	203.00
New Mexico	2022	0.45	56420.00	4.10	0.22	0.25	780.00
New York	2022	0.35	75910.00	4.30	0.18	0.21	429.00
North Carolina	2022	0.36	65070.00	3.70	0.16	0.21	405.00
North Dakota	2022	0.24	78720.00	2.10	0.18	0.24	280.00
Ohio	2022	0.36	67520.00	4.00	0.19	0.23	294.00
Oklahoma	2022	0.34	63440.00	3.00	0.20	0.22	420.00
Oregon	2022	0.3	86780.00	4.00	0.23	0.30	342.00
Pennsylvania	2022	0.34	72210.00	4.10	0.18	0.24	280.00
Rhode Island	2022	0.37	80650.00	3.20	0.23	0.24	172.00
South Carolina	2022	0.37	61770.00	3.20	0.18	0.21	491.00


```python
import pandas as pd
from scipy import stats

file_path = "Gil-Baptiste- Assignment #3 Curated Data SS-3.xlsx"
df = pd.read_excel(file_path, sheet_name="Curated Data")

# Define the target and input variables
target = "Violent Crime Per 100k"
variables = [
    "Single Parent Households",
    "Median Income",
    "Unemployment Rate",
    "Substance Use Disorder (18+ Estimate)",
    "Any Mental Disorder (18+ Estimate)"
]

# Calculate r and p-values
results = []
for var in variables:
    r, p = stats.pearsonr(df[var], df[target])
    results.append({
        "Variable": var,
        "r": round(r, 3),
        "p-value": round(p, 4),
        "Direction": "Positive" if r > 0 else "Negative" if r < 0 else "None",
        "Significant (p < 0.05)": p < 0.05
    })

# Convert to DataFrame for display
results_df = pd.DataFrame(results)
print(results_df)

                                Variable      r  p-value Direction  \
0               Single Parent Households  0.418   0.0023  Positive
1                          Median Income -0.099   0.4873  Negative
2                      Unemployment Rate  0.340   0.0147  Positive
3  Substance Use Disorder (18+ Estimate)  0.328   0.0186  Positive
4     Any Mental Disorder (18+ Estimate) -0.004   0.9794  Negative
```

```
        target =
        variables = [
            "Single Parent Households",
            "Median Income",
            "Unemployment Rate",
            "Substance Use Disorder (18+ Estimate)",
            "Any Mental Disorder (18+ Estimate)"
        ]

# Calculate r and p-values
results = []
for var in variables:
    r, p = stats.pearsonr(df[var], df[target])
    results.append({
        "Variable": var,
        "r": round(r, 3),
        "p-value": round(p, 4),
        "Direction": "Positive" if r > 0 else "Negative" if r < 0 else "None",
        "Significant (p < 0.05)": p < 0.05
    })

# Convert to DataFrame for display
results_df = pd.DataFrame(results)
print(results_df)

                                    Variable      r   p-value Direction  \
0                    Single Parent Households  0.418   0.0023  Positive
1                               Median Income -0.099   0.4873  Negative
2                           Unemployment Rate  0.340   0.0147  Positive
3       Substance Use Disorder (18+ Estimate)  0.328   0.0186  Positive
4          Any Mental Disorder (18+ Estimate) -0.004   0.9794  Negative

   Significant (p < 0.05)
0                    True
1                   False
2                    True
3                    True
4                   False
```
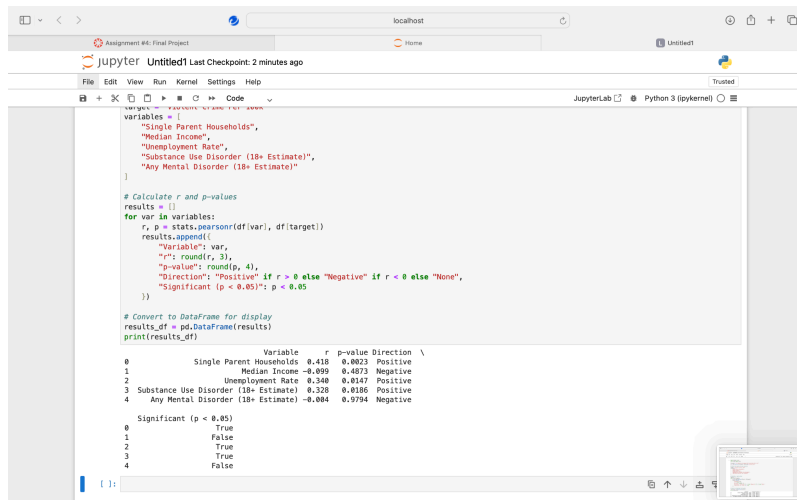
a. Correlation score for the following variables:

    i.   Violent crime per 100k & Single Parent Households -> 0.418

    ii.   Violent crime per 100k & Median income -> -0.099

    iii.   Violent crime per 100k& Unemployment Rate -> 0.340

    iv.   Violent crime per 100k & Susbtance Use Disorder -> 0.328

    v.   Violet crime per 100k & Any mental disorder -> -0.004

b. Significance (p-value):

    i.   Violent crime per 100k & Single Parent Households -> 0.0023

    ii.   Violent crime per 100k & Median income -> 0.4873

    iii.   Violent crime per 100k& Unemployment Rate -> 0.0147

    iv.   Violent crime per 100k & Susbtance Use Disorder -> 0.0186

    v.   Violet crime per 100k & Any mental disorder -> 0.9794

c. Direction of correlation

    i.   Violent crime per 100k & Single Parent Households -> positive

      ii.   Violent crime per 100k & Median income -> negative

     iii.   Violent crime per 100k& Unemployment Rate -> positive

     iv.   Violent crime per 100k & Susbtance Use Disorder -> positive

     v.   Violet crime per 100k & Any mental disorder -> negative

    d. Interpretation of Findings

The correlation results were partly expected but included some surprises. At first I assumed all the variables would be directly related to violent crime, but the data showed a more nuanced picture. Single-parent households, unemployment, and substance use disorder all had moderate positive and statistically significant correlations with violent crime. This supports the idea that social and economic instability are linked to higher crime rates. Median income showed a very weak negative and non-significant correlation with crime, which was surprising. I expected a stronger relationship, but the data suggests income alone isn't a strong predictor of violent crime in this case. Most notably, mental disorder prevalence had no meaningful correlation with violent crime. This goes against common assumptions and suggests that mental health rates alone don't explain variations in violent crime across states. Overall, the findings highlight that while some variables have clear relationships with crime, others may be less influential or only matter when combined with broader factors.

g) Multi Liner Regression

(Violent Crime Rate per 100k)$_i$ = B_0+(Median Income)$_i B_1$ + (UnemploymentRate)$_i B_2$ + (Single-Parent Households)$_i B_3$ + (Drug Usage)$_i B_4$ + (Depression Rate)$_i B_{65}$

	A	B	C	D	E	F	G	H
1	SUMMARY OUTPUT							
3	*Regression Statistics*							
4	Multiple R	0.48641429						
5	R Square	0.23659886						
6	Adjusted R Sq	0.15177651						
7	Standard Erro	148.653628						
8	Observations	51						
10	ANOVA							
11		df	SS	MS	F	Significance F		
12	Regression	5	308193.435	61638.6869	2.78934578	0.02809833		
13	Residual	45	994405.546	22097.901				
14	Total	50	1302598.98					
16		Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0% Upper 95.0%
17	Intercept	-216.83386	298.578711	-0.7262201	0.47146447	-818.20226	384.534529	-818.20226 384.534529
18	Single Parent	740.392401	493.071247	1.50159314	0.14018725	-252.70407	1733.48887	-252.70407 1733.48887
19	Median Incom	-0.0005765	0.00175676	-0.3281699	0.74430538	-0.0041148	0.00296177	-0.0041148 0.00296177
20	Unemployme	35.6909132	32.2755047	1.10582045	0.27468131	-29.31529	100.697117	-29.31529 100.697117
21	Substance Use	1428.25128	1142.89833	1.24967483	0.21787902	-873.66412	3730.16669	-873.66412 3730.16669
22	Any Mental Di	-42.896022	989.08168	-0.0433695	0.96559883	-2035.0088	1949.21674	-2035.0088 1949.21674

After performing statistical analysis using Excel's Data Analysis, the model coefficients were identified and used to construct the following regression equation:

Violent Crime Rate = –216.83 + 740.39(Single Parent) – 0.00058(Median Income) + 35.69(Unemployment) + 1428.25(Substance Use) – 42.90(Any Mental Disorder)

a. Percentage of variance explaine model

   i.   R = .2366 (23.66%)

   ii. This indicates that 23.66% of the variability in violent crime rates across the dataset is explained by the combination of the five input variables.

   iii. While not very high, this value suggests that the model captures a portion of the trend, but other factors outside the model may also play substantial roles in influencing violent crime rates.

b. Significance of Variables in the Model

      i.   Based on the p-values obtained through the regression output, none of the variables were statistically significant (all had p-values > 0.05).Since none of the predictors had a p-value below 0.05, we cannot confidently say that any of them individually explain changes in the violent crime rate in a statistically significant way. However, we can still consider the direction of the coefficients: Positive relationships: Single Parent Households, Unemployment Rate, Substance Use. Negative relationships: Median Income, Any Mental Disorder.

<u>h) How will you interpret the results?</u>

- Interpreatation of results

    When interpreting the results, we must first consider which variables are statistically significant, as this guides how much weight we can place on the observed relationships. In this analysis, only a subset of the independent variables showed statistically significant correlations with violent crime rates. Among them, the percentage of single-parent households, unemployment rate, and substance use disorder were found to be positively correlated and statistically significant. This indicates that as the proportion of single-parent households, unemployment, or substance use disorder increases, the violent crime rate also tends to increase. These findings are consistent with social theories suggesting that economic instability, family structure, and drug abuse contribute to criminal behavior. nterestingly, some variables like median household income and mental disorder prevalence did not show statistically significant

relationships. This was somewhat surprising particularly in the case of income since it's often assumed that poverty or low income directly leads to higher crime.

Furthemore, This result may suggest that income alone doesn't fully explain crime rates and may interact with other socioeconomic or systemic factors. Additionally, the regression model's R^2 value was around 0.2366 (23.66%), which means the included variables only account for a small portion of the variance in violent crime rates. While helpful, this indicates that there are other unobserved factors influencing crime that were not captured in this model. It also serves as a reminder that correlation does not imply causation, especially in complex social systems.

- Implications

These findings highlight key areas for targeted intervention. For instance, policies aimed at reducing substance use disorder such as funding for rehabilitation programs or preventive education could potentially help lower crime rates. Similarly, supporting single-parent households through childcare support, job training, and family counseling may reduce vulnerability and strain that can contribute to criminal activity. One unexpected implication is that increasing median income may not directly reduce violent crime unless it is paired with broader support systems. This means policymakers should focus not just on income-based economic growth, but also on structural issues like employment stability and family welfare.

The results also show how adding more variables in a multivariable regression can change the story. Variables that seemed significant in isolation may become insignificant when considered alongside others. This reflects how intertwined social factors are in real-world policy issues.

- Aspects Important for Policymakers

For policymakers, the most actionable insights come from the significant variables particularly unemployment, substance use, and family structure. These are areas where government can intervene through economic policy, public health services, and community support. Programs that create stable job opportunities or reduce drug dependency may indirectly reduce crime. It's also important for policymakers to understand that addressing violent crime requires a multi-faceted approach. The low R coefficent suggests that no single solution will suffice; rather, broad, integrated strategies across sectors (ex. education, housing, healthcare, etc.) are needed. Furthermore, more data collection and analysis may be needed to uncover other contributing factors not captured in the current model.

i) What were the limitations and ethical implications of this study? What will you change if you had more data or more time?

One of the main limitations of this study is the relatively small set of variables included. While the analysis focused on five key factors such as unemployment, substance use disorder, and family structure the model explained only about 23.66% of the variation in violent crime between states. This suggests that many other important influences on crime were not captured in the model. Another limitation is the use of state-level data, which can hide important differences within states. Crime patterns often vary widely between urban and rural areas, or even between neighborhoods in the same city. Looking only at state averages may miss these local dynamics. Additionally, it is important to recognize that crime is a very broad and complex issue, influenced by a wide range of social, economic, psychological, and environmental factors. No

single model can fully explain all the reasons why crime happens. This study only scratches the surface. Additionally, some of the variables in the model may influence each other. For example, higher unemployment might be linked with higher substance use in certain states. When factors are connected like this, it can be harder to clearly separate their individual effects on violent crime, which makes the interpretation of the results more challenging. Lastly, differences in how data was collected or reported across states may have affected the consistency or accuracy of the analysis.

Furthemore, if more time or data were available, the study could be improved in several ways. First, more variables could be added to capture a broader picture of what drives violent crime such as poverty rates, education levels, housing instability, access to social services, or policing practices. These factors could help build a more complete and accurate model. Second, using more detailed data like city-level or neighborhood-level information instead of state-level averages would allow for a more focused analysis that reflects local conditions. Also, looking at changes over time instead of a single year could help identify trends and reveal whether certain variables are leading indicators of crime. Finally, with more time, the analysis could explore how different factors might combine or interact to influence crime, which could lead to deeper and more useful insights.