In [227…
```
import numpy as np
import seaborn as sns
import pandas as pd
import matplotlib.pyplot as plt
import duckdb
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from statsmodels.api import OLS
from scipy.stats import bootstrap
from math import sqrt
from sklearn.metrics import mean_squared_error
from sklearn.metrics import root_mean_squared_error
from sklearn.metrics import mean_absolute_error
from sklearn.model_selection import train_test_split
```

Team members: Jasmine Ren jr2293, Carina Lau cl2623, Savitta Sivapalan ss2849

Phase 5 Outline

Introduction

- Research Question
- Key Terminologies
- Background on the Research Question

Data description and cleaning

- Cleaning
- Data Description
    - Income Dataset
    - Poverty Dataset

- Population Dataset
- Unemployment Dataset
- Crime Dataset

Preregistration statement

- Hypothesis 1
- Hypothesis 2
- Hypothesis 3

Data Analysis

- Hypothesis 1
- Hypothesis 2
- Hypothesis 3

Evaluation of Significance

- Hypothesis 1
- Hypothesis 2
- Hypothesis 3

Conclusion

Limitations

Citations

Introduction

Key Terminologies

- Violent crime according to the FBI.gov website is defined by the composition of 4 types of crimes:
    - Aggravated assault: an unlawful attack by one person upon another for the purpose of inflicting severe or aggravated bodily injury
    - Rape: penetration, no matter how slight, of the vagina or anus with any body part or object, or oral penetration by a sex organ of another person, without the consent of the victim
    - Homicide: the willful (nonnegligent) killing of one human being by another
    - Robbery: the taking or attempting to take anything of value from the care, custody, or control of a person or persons by force or threat of force or violence and/or by putting the victim in fear
- Unemployment rate: the percentage of people in the labor force who are unemployed

*The unemployment rate was determined by the average across all the unemployment dataset's years (1980-2018) instead of the current day unemployment rate of 4.2% to account for the change in years and economic conditions

- Income: the total earnings of an individual from various sources such as wages, investment ventures, and other sources of income
- Poverty: the state of not having enough money or resources to meet basic needs, such as food, clothing, and shelter
- Population: the number of people or inhabitants in a country or region

Research Question:

What set of factors (income, poverty, population, unemployment) is most influential in determining adult violent crime rates across U.S. states (and the District of Columbia) from 1979-2018? How could we use these factors to accurately predict future violent crime rates?

Background on the Research Question

Violent crime is a recurring and deeply ingrained issue in the United States. It dominates news channels and political debates. This election year, crime was Ranked #10 as an

"Extremely Important" topic to registered voters in the U.S. and has continuously been of concern in the country. 63% of Americans describe the crime problem in U.S. as either extremely or very serious, which was up from the 54% last measured in 2021. Headlines about rising crime rates in urban centers, coupled with polarizing opinions on how to address the root causes of such violence, have made this topic relevant. Conversations around violent crime are frequently grounded in stereotypes where impoverished areas are often labelled as inherently dangerous, unemployment is seen as a precursor to crime and low-income populations are disproportionately stigmatized. While these beliefs may oversimplify the issue, they underscore a broader societal need to truly understand the underlying factors driving violent crime in America and how they can be used to interpret ongoing violent crime trends and patterns.

With unemployment spiking to 7.5% during the recession of the early 1980 and poverty widening throughout the 1990s and into the Great Recession of 2008–2009, concerns about the potential link between economic hardship and violent crime repeatedly surfaced. There has always been speculation that these economic stressors are fueling violence in states. However, despite these claims, the relationship between socioeconomic factors, such as income, poverty, population size, unemployment, and crime remains a topic of debate. Are these stereotypes rooted in statistical truth or do they obscure more nuanced explanations? Are there geographic-crime patterns that hold for when we analyze densely populated urban versus rural states?

Our interest in this research was sparked by our shared interest in watching crime documentaries and between public perceptions and the complexities revealed in crime data. While the media and public discourse often assume a direct connection between economic hardship and violence, the actual predictive strength of variables like income or unemployment in determining violent crime rates, we believe, may not be as straightforward as believed.

Through a data-driven lens, using crime data from 1979 to 2018 across all U.S. states (including the District of Columbia), we aim to identify which socioeconomic factors are most influential in driving violent crime and to challenge or validate prevailing assumptions. By focusing on statistical patterns rather than stereotypes, we hope to shed light on the broader dynamics of crime, while providing practical insights that could inform policymaking and public safety strategies. As violent crime continues to shape American society, understanding its root causes has never been more important.

Cleaning

link: https://github.com/w0ahnder/INFO2950_project/blob/main/phase4_datacleaning/phase4_cleaning.ipynb

Data Description

Income Dataset

The rows of this dataset are the states and their respective values for the columns of this dataset which are the median income and standard error by year (1984 to 2018). This dataset was created by the U.S. Census for the United States governmnet to gather information about the overall income of different states, perhaps to investigate income disparities among states. Some processes that might have influenced what data was observed and recorded is the feasability of the data. This data comes from people who actually fill out the U.S. Census, which we know from history isn't typically everyone. So, the data might not be completly accurate, which explains why the U.S. Census believed it was necessary to include the "standard error" column. The preprocessing and cleaning that we performed on this dataset mainly revolved around looking specifically from the years 2013 to 2018, while also looking specifically at the median income column, since that is what we're interested in, in determining its effect on the research question (violent crime rate). People were involved and they were aware of the data collection. This is because this is the U.S. Census, which people fill out, knowing that their data will be used, likely assuming that it would be used to learn more about the general population of the United States. The raw source can be found here: https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xls

Poverty Dataset

The rows of this dataset are the states and their respective values for columns that consist of the total population, number in poverty, margin of error, percent in poverty, and margin of error. This dataset was created by the U.S. Census for the United States government to gather more information and learn more about the overall number in poverty and percent of poverty for different states. This could show some economic disparities among states, perhaps as a way for the government to better understand which states need more funding in preventing poverty and homelessness. Some processes that might have influenced what data was observed and recorded is that not everyone in the United States fills out the U.S. Census, which can make the data not necessarily representative of the population, which is likely why the U.S. Census added in the margin of error columns for the number in poverty and percent in poverty values. The preprocessing and cleaning that we performed was breaking up the data by year, specifically 2013-2018, and concatenating all the data from each year together. People were involved and they were aware of the data collection. This is because this is the U.S. Census, which people fill out, knowing that their data will be used, likely assuming that it would be used to learn more about the general population of the United States. The raw source can be found here: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html (table 19)

Population Dataset

- What are the observations (rows) and the attributes (columns)?

The observations (rows) in this dataset represent the United States, including all fifty states, Washington D.C., and Puerto Rico. The attributes (columns) include Census data,

Estimates Base, and annual population estimates from 2010 to 2019. This dataset was created to provide updated and accurate estimates of the population for U.S. states and other regions during the years 2010 to 2019, helping to track population trends and emigration patterns. Such information is crucial for resource allocation, cultural representation, and policymaking. The dataset was funded by the U.S. Census Bureau, a government organization under the U.S. Department of Commerce.

The observed and recorded data may have been influenced by migration trends, survey methodologies, and reporting mechanisms such as birth and death records. Data accuracy could vary based on resource availability and the capacity to collect data in less focused or rural regions. Undocumented residents in such areas might also impact the comprehensiveness of the dataset. Preprocessing of the data likely involved cleaning, validation, and formatting to ensure clarity and usability for analysis. As the dataset was made available in Excel format, it was structured to facilitate straightforward analysis by users.

While the public may not have been directly aware of the specific collection of this population data, it is common practice in this era to use collected data for larger analysis and decision-making purposes. The raw source data can be found at the U.S. Census Bureau website, specifically at this link: https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html. Users should refer to the section labeled “Annual Estimates of the Resident Population for the United States...” for the original dataset.

Unemployment Dataset

The rows of the original dataset include the annual unemployemnt rate for each of the states and the columns specifies which year the rate is for. Our cleaned dataset only includes columns for years 2013-2018

The dataset was created by Iowa State University using data collectd from the Bureau of Labor Statistics (BLS) using Local Area Unemployment Statistics (LAUS) and the Current Population Survey (CPS) and its purpose was to compile all of the unemployment rates over the years into a single table for the states.

The creation of the dataset does not seem to have been funded.

The survey is conducted as an interview where participants are asked a series of questions relating to employment status, age, occupation, etc. If participants did not feel comfortable answering a certain question, they did not need to anwer it. This could affect the data collection since there would be missing information from some participants and not for others.

In terms of prepocessing, the BLS removes all conidential information from their survey results (like names and addresses), and then proceed to analayze and publish their statistics.

The people interviewed for the survey are aware that their responses will be used to create statistical information about the labor market, and that personal information would be

confidential. link to data https://www.icip.iastate.edu/tables/employment/unemployment-states

Crime Dataset

The observations (rows) in this dataset represent years from 1979 to 2023, while the attributes (columns) include state abbreviations, state names, population, violent crime, homicide, legacy rape, revised rape, robbery, aggravated assault, property crime, burglary, larceny, and motor vehicle theft. This dataset was created to provide transparent access to crime data across the United States and is part of the broader Uniform Crime Reporting (UCR) Program, which was established in 1930 to standardize and report crime statistics consistently across states, agencies, and local jurisdictions. By creating a uniform way to collect and analyze crime data, the dataset ensures comparability. Additionally, as seen on the front page of the website, the backend data helps track crime trends and showcases the FBI's progress in reducing crimes and improving their track record over time. It serves as a tool for holding the FBI accountable for enhancing the safety and security of the nation.

While the website does not explicitly state the funding source, research indicates that the FBI, which oversees the dataset, is predominantly funded by the U.S. government, with Congress annually providing 80% of its budget. Likely, Congress allocates part of this budget to the Bureau of Justice Statistics (BJS), which funds the FBI's development of automated data capture specifications adhering to the UCR program's standards (https://ucr.fbi.gov/crime-in-the-u.s/2014/crime-in-the-u.s.-2014/resource-pages/about-ucr#:~:text=Upon%20selecting%20the%20South%20Carolina,of%20contributing%20law%20enforcement%20agencies).

The data observed and recorded are influenced by the FBI's crime category definitions (e.g., violent crimes and rape), which local law enforcement agencies are required to follow. The dataset follows the "hierarchy rule," meaning only the most serious offense in a case is counted, with the order of violent crimes descending from homicide, rape, robbery, and aggravated assault. Property crimes, such as burglary, larceny-theft, and motor vehicle theft, are included as well, although the hierarchy rule does not apply to arson. Certain crimes may be excluded due to these definitions and hierarchical requirements. Additionally, not all law enforcement agencies report data consistently, often due to resource limitations, leading to potential gaps in the dataset. Legal definitions of crimes can also change over time; for example, the definition of rape was revised in 2013, meaning earlier data might not align with more recent records.

Crime data is submitted to the FBI through the Uniform Crime Reporting (UCR) system (https://www.fbi.gov/how-we-can-help-you/more-fbi-services-and-information/ucr). It is entered either manually or automatically via digital reporting platforms. The FBI performs quality checks on the submitted data to ensure consistency with the UCR format, identifying and addressing missing, incomplete, or erroneous data. Those involved in crime incidents (both offenders and victims) are generally aware that their actions or experiences are recorded, especially if reported to law enforcement. However, they may not always know how this data is later aggregated or used in a broader dataset tracking crime trends nationwide.

The raw source data for this dataset can be accessed at https://cde.ucr.cjis.gov/LATEST/webapp/#/pages/downloads. To locate the dataset, navigate to "Additional Dataset" and

select “Summary Reporting System"

Pregistration Statement

Hypotheses:

Hypothesis #1: Income is a better predictor for robbery than burglary across the U.S. states using the data for years 1984-2018

The motivation behind Hypothesis #1 stems from our Phase 2 where we got a negative correlation between violent crime and income and how there is a nuanced distinction between robbery and burglary as defined by the FBI. We expected a positive relationship, so wanted to look into the relationship further but with robbery and burglary because we were initially conufsed between their differences. Robbery is classified as a violent crime because it involves direct confrontation with a victim, often with threats, while burglary is categorized as a property crime, typically involving unlawful entry without personal confrontation. Despite this difference, both crimes share a common thread of stealing... Are the root causes driving these crimes fundamentally the same? Or are there other factors influencing why someone might commit one over the other?

One common stereotype is that people rob because they are poor, driven by the urgent need for money. If this stereotype holds any merit, we would expect income to be a strong predictor for robbery rates. However, burglary also involves theft, and it may be similarly influenced by financial desperation.

Analysis:

Our approach will involve encoding each state as a binary variable (e.g., XNew York, XCalifornia) alongside predictors for income and year to capture both geographic and temporal variations. The dependent variable (Y) will alternate between robbery and burglary rates, allowing us to compare their respective relationships with income. We will implement a train-test split, using data from 1984–2015 for training and 2016–2018 for testing, ensuring that our model can be evaluated on unseen data. Ordinary Least Squares (OLS) regressions will be conducted separately for robbery and burglary, and model performance will be assessed using residual plots to check for randomness. Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) will be used to measure predictive accuracy. Additionally, we will visually compare actual versus predicted values for 20 randomly selected states in years 2016, 2017, and 2018 to ensure there are no geographical biases and to observe patterns in prediction accuracy.

Hypothesis #2: States that are below the U.S. unemployment average over the years 1980 to 2018 correspond to higher violent crime rates

In our Phase 2, we found that violent crime and unemployment had a 0.53 correlation, which we expected to be a lot higher. So we wanted a hypothesis that challenges the stereotype that higher unemployment directly causes more crime. This question is critical to our overarching goal of understanding the socioeconomic factors influencing violent

crime. By examining unemployment from both perspectives, above and below the national average, we aim to uncover whether employment stability paradoxically contributes to higher crime rates in certain states.

Analysis:

To test this, we followed a similar approach as in Hypothesis #1. States were encoded as binary variables (e.g., XNew York, XCalifornia, XAlabama), and we included year and unemployment rate inputs to predict the Y variable, which was the average violent crime rate for a state. We will be running a train-test split to evaluate the model's ability to predict violent crime based on test X inputs. Metrics such as variance, betas, residual plots will provide insights into the model’s ability to measure for states below versus above the national unemployment average.

We will visualize the relationship using a scatterplot of average unemployment versus average violent crime rates, marking the national average unemployment rate (calculated as the average from 1980 to 2018) as a reference. Then, we will divide states into two groups: those with unemployment rates above and below the national average. For each group, we will run OLS regressions to test the significance of unemployment rates in predicting violent crime. We will then build predictive models using a train-test split (1980–2015 for training and 2016–2018 for testing) and include dummy variables for states to account for geographic variation. RMSE, MAE, and residual plots will be used to evaluate model accuracy for both groups.

Hypothesis #3: At least one of income, population, unemployment, and poverty (+ year) have a relationship with the 4 categories of violent crime for each state regardless of the year

Hypothesis #3 was designed to encompass the full range of socioeconomic factors and violent crime categories we considered in our overarching question. By including income, population, unemployment, and poverty as predictors, and homicide, aggravated assault, rape, and robbery as dependent variables, this hypothesis allowed us to explore whether these socioeconomic variables have a relationship with violent crime across all states, irrespective of the year. We found that (these factors correlate most strongly with homicide) [https://pmc.ncbi.nlm.nih.gov/articles/PMC6660557/], so we were curious to see how these relationships would manifest when applied to all categories of violent crime using OLS regression analysis. This approach broadened the scope of our investigation, enabling us to identify patterns across different types of violent crime rather than limiting our focus to a single category.

Analysis:

We will run Multivariable Ordinary Least Squares (OLS) regressions for each crime type, using socioeconomic factors, year, and state dummies as predictors. The dependent variable for each model will be the rate of the specific crime, and a train-test split will be applied, with data from 1980–2015 used for training and 2016–2018 for testing. We will evaluate the models using metrics such as Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) to measure predictive accuracy and assess the significance of each predictor’s coefficient to identify meaningful relationships. For each crime type, we will examine whether the null hypothesis, that none of the factors have a significant relationship

with the crime rate, can be rejected.

Poverty

In [228…
```
# importing cleaned csv from previous phase
poverty_concat = pd.read_csv('poverty_concat.csv')
poverty_concat.head()
```

Out[228…

	State	TotalPop	PovertyTotal	PovertyPercent	Year
0	Alabama	4751	796	16.7	2013
1	Alaska	696	76	10.9	2013
2	Arizona	6645	1345	20.2	2013
3	Arkansas	2940	504	17.1	2013
4	California	38050	5675	14.9	2013

Income 1984-2018

In [229…
```
# importing cleaned csv from previous phase
income_df = pd.read_csv('income_df.csv')
income_df.head()
```

Out[229…

	State	2018	2017	2016	2015	2014	2013	2013.1	2012	2011	...	1993	1992	1991	1990	1989	1988	1987	19
0	Alabama	49936.0	51113.0	47221.0	44509.0	42278.0	47320.0	41381.0	43464.0	42590.0	...	25082.0	25808.0	24346.0	23357.0	21284.0	19948.0	19734.0	1913
1	Alaska	68734.0	72231.0	75723.0	75112.0	67629.0	72472.0	61137.0	63648.0	57431.0	...	42931.0	41802.0	40612.0	39298.0	36006.0	33103.0	33233.0	3135
2	Arizona	62283.0	61125.0	57100.0	52248.0	49254.0	52611.0	50602.0	47044.0	48621.0	...	30510.0	29358.0	30737.0	29224.0	28552.0	26435.0	26749.0	2550
3	Arkansas	49781.0	48829.0	45907.0	42798.0	44922.0	39376.0	39919.0	39018.0	41302.0	...	23039.0	23882.0	23435.0	22786.0	21433.0	20172.0	18827.0	1873
4	California	70489.0	69759.0	66637.0	63636.0	60487.0	60794.0	57528.0	57020.0	53367.0	...	34073.0	34903.0	33664.0	33290.0	33009.0	30287.0	30146.0	2901

5 rows × 37 columns

In [230…
```
# importing cleaned csv from previous phase
income_melt = pd.read_csv('income_melt_df.csv')
income_melt.head()
```

Out[230…

	State	Year	Median_Income
0	Alabama	2018	49936.0
1	Alaska	2018	68734.0
2	Arizona	2018	62283.0
3	Arkansas	2018	49781.0
4	California	2018	70489.0

Unemployment

In [231…
```
# importing cleaned csv from previous phase
job_df = pd.read_csv('job_df.csv')
```

```
job_melt = pd.read_csv('job_melt.csv')
```

Crime 1979-2022

In [232…

```
# importing cleaned csv from previous phase
crime_df = pd.read_csv('crime_df.csv')
crime_df
```

Out[232…

	Year	State_Abbreviation	State	Population	Violent_Crime	Homicide	Rape	Robbery	Aggravated_Assault	Property_Crime	Burglary
0	1979	AK	Alaska	406000	1994	54	292	445	1203	23193	5616
1	1979	AL	Alabama	3769000	15578	496	1037	4127	9918	144372	48517
2	1979	AR	Arkansas	2180000	7984	198	595	1626	5565	70949	21457
3	1979	AZ	Arizona	2450000	14528	219	1120	4305	8884	177977	48916
4	1979	CA	California	22696000	184087	2952	12239	75767	93129	1511021	496310
...	...	...	...	...	...	...	...	...	...	...	...
2239	2022	VA	Virginia	8683619	20624	641	2791	3360	13832	148845	10944
2240	2022	WA	Washington	7785786	29504	400	3208	6766	19130	262437	43987
2241	2022	WV	West Virginia	1775156	5213	95	909	210	3999	23663	3561
2242	2022	WI	Wisconsin	5892539	17889	322	2452	2350	12765	80703	9137
2243	2022	WY	Wyoming	581381	1188	15	373	48	752	9590	1257

2244 rows × 11 columns

DATA ANALYSIS

Hypothesis 1:

Income is a better predictor for robbery than burglary across the states.

Training the model for predicting robbery using income

As explained in our introduction, our inputs (X) are income, year and for the states we decided to use binaries, each time the equation is run only one state can = 1

In [233…
```
income_robbery = duckdb.sql('''SELECT  Robbery,C.State, C.Year, Median_Income
                            FROM crime_df AS C JOIN income_melt AS I
                            ON C.State = I.State AND C.Year = I.Year
                            ORDER BY C.Year ASC''').df()
income_robbery = pd.get_dummies(income_robbery,prefix='', \
                                prefix_sep='', columns=['State'], dtype=int, drop_first=True)
income_robbery.head()
```

Out[233…

	Robbery	Year	Median_Income	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	...	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Wash
0	547	1984	32356.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	3833	1984	17310.0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	1587	1984	15674.0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	
3	4003	1984	21425.0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	
4	83924	1984	25287.0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	

5 rows × 52 columns

In [234…
```
X_inc_rob_train, X_inc_rob_test, y_inc_rob_train, y_inc_rob_test = \
```

```
                  train_test_split(income_robbery.iloc[:,1:],
                                   income_robbery['Robbery'], test_size=0.3, shuffle=False)
X_inc_rob_train = sm.add_constant(X_inc_rob_train)
inc_rob_model =sm.OLS(y_inc_rob_train, X_inc_rob_train).fit()
```

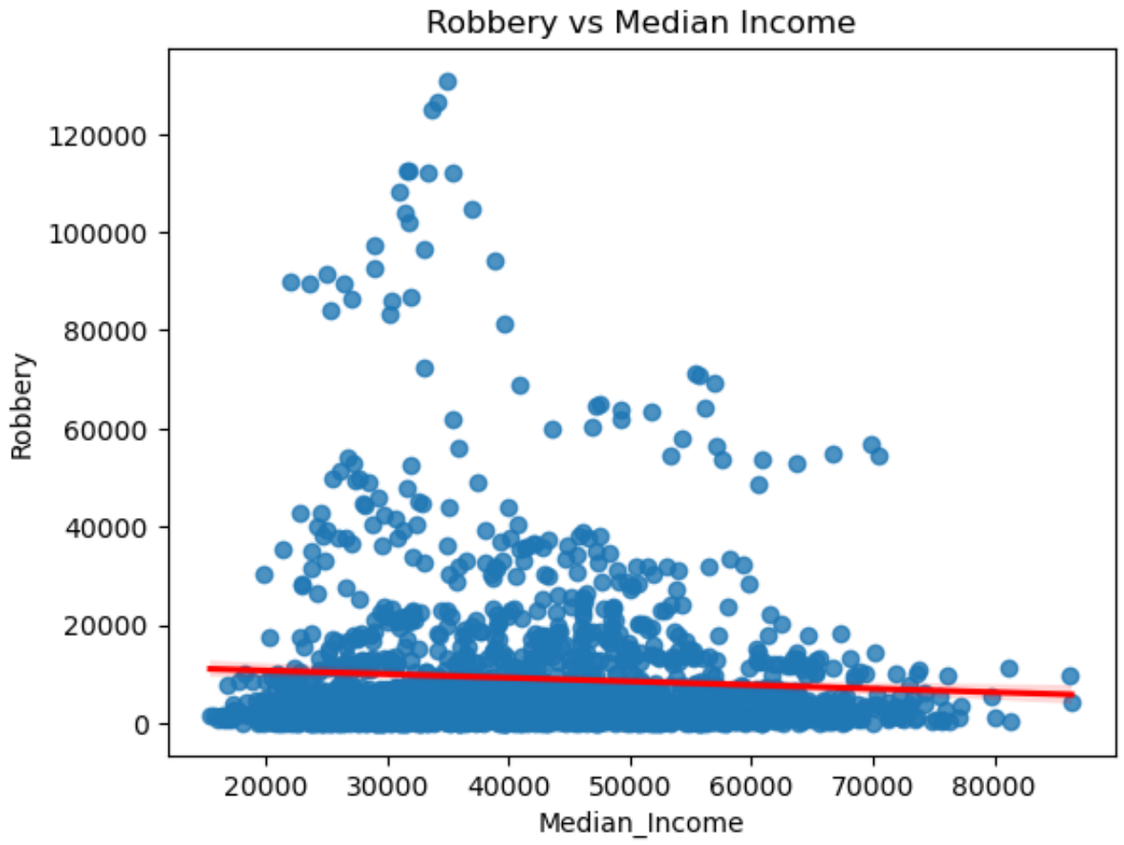
In [235…

```
#make predictions for robberies using income
X_inc_rob_test = sm.add_constant(X_inc_rob_test)
inc_rob_preds = inc_rob_model.predict(X_inc_rob_test)
print( f"MAE for predicting robbery using income: \
      {mean_absolute_error(y_inc_rob_test, inc_rob_preds)}")
print( f"RSME for predicting robbery using income: \
      {root_mean_squared_error(y_inc_rob_test, inc_rob_preds)}")
```

```
MAE for predicting robbery using income:       3903.309439874085
RSME for predicting robbery using income:       7759.480589023579
```

In [236…

```
sns.regplot(x = income_robbery['Median_Income'], y = income_robbery['Robbery'], line_kws = {"color":"red"});
plt.title("Robbery vs Median Income");
```



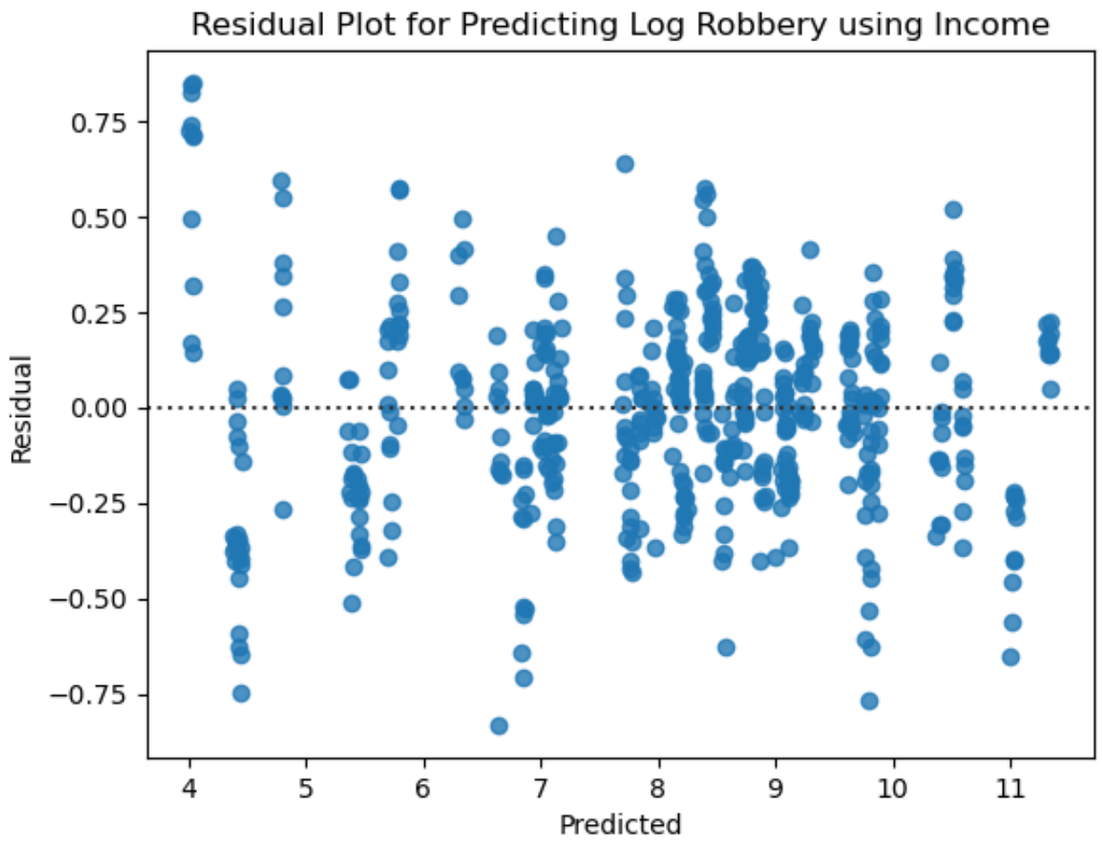
Since the y value rises exponentionally comparatively to the x value. We want to take a log transform of the robbery (y value). Also, based on where the linear regression line is positioned, we can see that the y values have a lot of outliers that can be dealt with using a log transform on the y axis.

In [237…

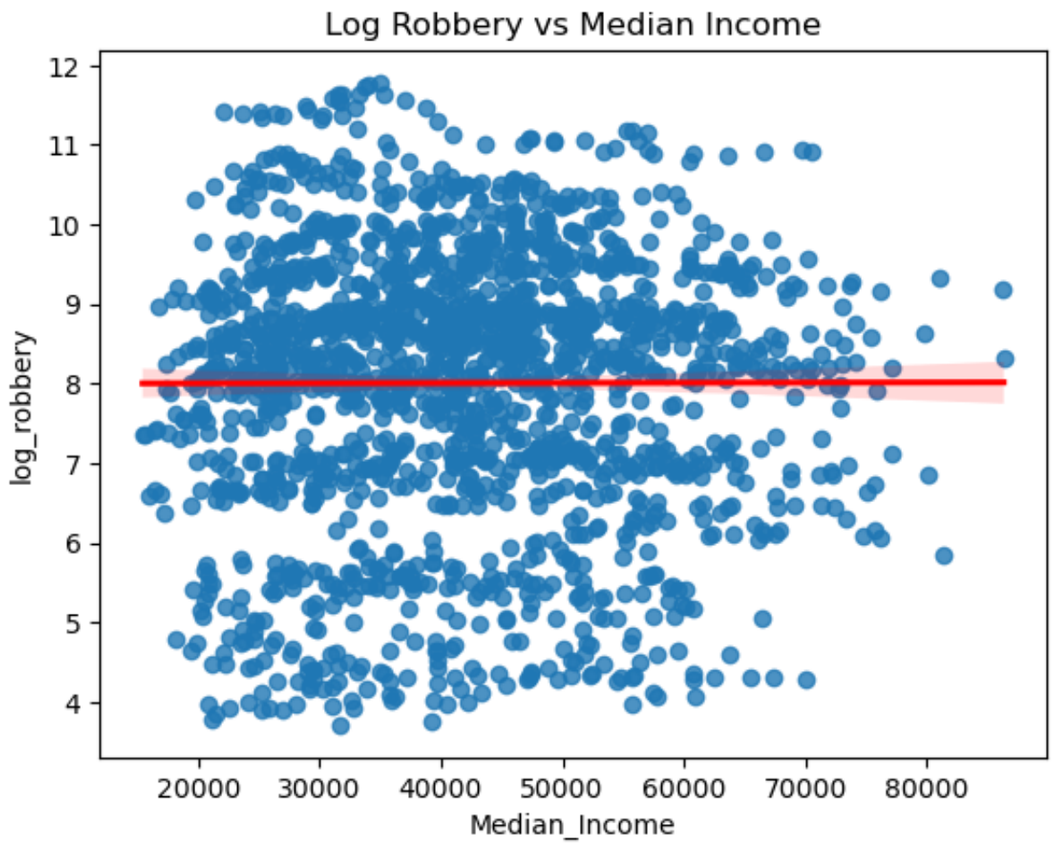
```
income_robbery['log_robbery'] = np.log(income_robbery['Robbery'])

#print(income_robbery.iloc[:, 47:])
#print((income_robbery.iloc[:, 1:-1]).columns)
```

```
X_inc_rob_train, X_inc_rob_test, y_inc_rob_train, y_inc_rob_test = \
                  train_test_split(income_robbery.iloc[:,1:-1],
                                   income_robbery['log_robbery'], test_size=0.3, shuffle=False)
X_inc_rob_train = sm.add_constant(X_inc_rob_train)
inc_rob_model =sm.OLS(y_inc_rob_train, X_inc_rob_train).fit()

X_inc_rob_test = sm.add_constant(X_inc_rob_test)
inc_rob_preds = inc_rob_model.predict(X_inc_rob_test)
print( f"MAE for predicting robbery using income: {mean_absolute_error(y_inc_rob_test, inc_rob_preds)}")
print( f"RSME for predicting robbery using income: {root_mean_squared_error(y_inc_rob_test, inc_rob_preds)}")


inc_rob_df = pd.DataFrame({"Preds": inc_rob_preds, "Actual": y_inc_rob_test})
ax = sns.residplot(x = inc_rob_preds, y = y_inc_rob_test)
ax.set(xlabel = "Predicted", ylabel= "Residual", \
       title = "Residual Plot for Predicting Log Robbery using Income");
```

```
MAE for predicting robbery using income: 0.29952513809241593
RSME for predicting robbery using income: 0.3845879612089225
```



In [238…
```
sns.regplot(x = income_robbery['Median_Income'], y = income_robbery['log_robbery'], \
            line_kws = {"color":"red"});
plt.title("Log Robbery vs Median Income");
```



In [239…
```
print(inc_rob_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            log_robbery   R-squared:                       0.986
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                     1655.
Date:                Mon, 09 Dec 2024   Prob (F-statistic):               0.00
Time:                        22:15:42   Log-Likelihood:                180.57
```

```
No. Observations:                1260   AIC:                            -257.1
Df Residuals:                    1208   BIC:                             10.09
Df Model:                          51
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const             -3.4351      6.542     -0.525      0.600     -16.270       9.400
Year               0.0061      0.003      1.851      0.064      -0.000       0.013
Median_Income  -4.422e-06   2.73e-06     -1.619      0.106   -9.78e-06    9.35e-07
Alaska            -2.3042      0.076    -30.455      0.000      -2.453      -2.156
Arizona            0.1590      0.061      2.613      0.009       0.040       0.278
Arkansas          -0.8965      0.060    -14.995      0.000      -1.014      -0.779
California         2.6702      0.066     40.407      0.000       2.541       2.800
Colorado          -0.4553      0.067     -6.811      0.000      -0.586      -0.324
Connecticut       -0.0632      0.075     -0.848      0.397      -0.209       0.083
Delaware          -1.5334      0.066    -23.392      0.000      -1.662      -1.405
Florida            1.8850      0.060     31.469      0.000       1.767       2.003
Georgia            0.9150      0.061     15.000      0.000       0.795       1.035
Hawaii            -1.5972      0.071    -22.440      0.000      -1.737      -1.458
Idaho             -3.2400      0.061    -53.267      0.000      -3.359      -3.121
Illinois           1.7249      0.065     26.641      0.000       1.598       1.852
Indiana            0.0998      0.061      1.626      0.104      -0.021       0.220
Iowa              -1.5631      0.061    -25.428      0.000      -1.684      -1.443
Kansas            -0.9365      0.061    -15.231      0.000      -1.057      -0.816
Kentucky          -0.5892      0.060     -9.823      0.000      -0.707      -0.472
Louisiana          0.3759      0.060      6.265      0.000       0.258       0.494
Maine             -2.9921      0.061    -49.169      0.000      -3.112      -2.873
Maryland           1.0248      0.075     13.687      0.000       0.878       1.172
Massachusetts      0.4287      0.069      6.174      0.000       0.292       0.565
Michigan           1.1107      0.064     17.405      0.000       0.985       1.236
Minnesota         -0.2574      0.068     -3.793      0.000      -0.391      -0.124
Mississippi       -0.7952      0.061    -13.115      0.000      -0.914      -0.676
Missouri           0.3920      0.061      6.383      0.000       0.271       0.512
Montana           -3.3308      0.060    -55.529      0.000      -3.449      -3.213
Nebraska          -1.7409      0.062    -28.161      0.000      -1.862      -1.620
Nevada            -0.2954      0.065     -4.573      0.000      -0.422      -0.169
```

```
New Hampshire     -2.8215      0.072    -39.046      0.000      -2.963      -2.680
New Jersey         1.1304      0.076     14.920      0.000       0.982       1.279
New Mexico        -1.0176      0.060    -16.966      0.000      -1.135      -0.900
New York           2.3480      0.063     37.526      0.000       2.225       2.471
North Carolina     0.5829      0.060      9.636      0.000       0.464       0.702
North Dakota      -4.6497      0.060    -77.180      0.000      -4.768      -4.532
Ohio               1.1300      0.062     18.149      0.000       1.008       1.252
Oklahoma          -0.5420      0.060     -9.031      0.000      -0.660      -0.424
Oregon            -0.4606      0.062     -7.392      0.000      -0.583      -0.338
Pennsylvania       1.2060      0.062     19.373      0.000       1.084       1.328
Rhode Island      -1.8280      0.064    -28.590      0.000      -1.953      -1.703
South Carolina    -0.0856      0.060     -1.418      0.156      -0.204       0.033
South Dakota      -3.8982      0.060    -64.662      0.000      -4.016      -3.780
Tennessee          0.5097      0.060      8.497      0.000       0.392       0.627
Texas              1.8234      0.061     29.899      0.000       1.704       1.943
Utah              -1.6111      0.066    -24.346      0.000      -1.741      -1.481
Vermont           -4.2612      0.063    -67.706      0.000      -4.385      -4.138
Virginia           0.2581      0.069      3.734      0.000       0.123       0.394
Washington         0.1090      0.066      1.651      0.099      -0.021       0.239
West Virginia     -2.0901      0.061    -34.490      0.000      -2.209      -1.971
Wisconsin         -0.2269      0.064     -3.531      0.000      -0.353      -0.101
Wyoming           -4.2375      0.062    -68.679      0.000      -4.359      -4.116
==============================================================================
Omnibus:                       12.452   Durbin-Watson:                   1.784
Prob(Omnibus):                  0.002   Jarque-Bera (JB):               12.679
Skew:                          -0.220   Prob(JB):                     0.00177
Kurtosis:                       3.217   Cond. No.                    4.08e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.08e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Looking at this OLS regression, we observe that the p values for income and year are insignificant, 0.106 and 0.064 respectively (based on a 0.05 significance level). This shows that the insignificant coefficient for income is -4.422e-06. This shows that for every one dollar increase in income, e^(-4.422e-06) unit decrease of robberies. Although our

hypothesis is not exactly testing income having a relationship with a states' number of robberies, the p-value still now tells us that it is not a strong determinor of robberies

For each of the states, using the test set, we get the predicted amount of robbery, along with the true value per year.

In [240…
```
# for each of the states, get the predicted value and the actual value
X_inc_rob_test2 = X_inc_rob_test
X_inc_rob_test2['actual_robbery'] = y_inc_rob_test
X_inc_rob_test2['preds_robbery'] = inc_rob_preds
inc_rob_melt = pd.melt(X_inc_rob_test2, id_vars=['const','Year', 'Median_Income', 'actual_robbery','preds_robbery'], \
                       var_name = 'State', value_name='Binary')
inc_rob_melt = inc_rob_melt[ inc_rob_melt['Binary']==1]
inc_rob_melt.head()
```

Out[240…

	const	Year	Median_Income	actual_robbery	preds_robbery	State	Binary
40	1.0	2010	57848.0	6.386879	6.347382	Alaska	1
90	1.0	2011	57431.0	6.356108	6.355366	Alaska	1
140	1.0	2012	63648.0	6.445720	6.334015	Alaska	1
190	1.0	2013	61137.0	6.434547	6.351259	Alaska	1
229	1.0	2013	72472.0	6.434547	6.301135	Alaska	1

We select 20 random states to see how well the model does on predicting the annual robbery totals.

In [241…
```
twenty_states = pd.DataFrame(np.random.choice(inc_rob_melt['State'], size =20, replace = False))
#print(twenty_states)
chosen_stats_robbery = duckdb.sql(''' SELECT * FROM inc_rob_melt
                                  WHERE Year>=2014 AND Year<=2018
                                  AND State IN (SELECT * FROM twenty_states)''').df()
chosen_stats_robbery.head()
```

Out[241…

	const	Year	Median_Income	actual_robbery	preds_robbery	State	Binary
0	1.0	2014	71223.0	6.858565	7.019839	Hawaii	1
1	1.0	2015	64514.0	6.989335	7.055647	Hawaii	1
2	1.0	2016	72133.0	6.892642	7.028096	Hawaii	1
3	1.0	2017	73575.0	6.981935	7.027859	Hawaii	1
4	1.0	2018	80108.0	6.857514	7.005111	Hawaii	1

Now we want to look at how the model did, visually, at predicting robbery for 20 random states. We chose to do 20 random ones so ensure there is no geographical bias or underlying pattern that the model is overfitting.

Income and Robbery (Expected vs Predicted) for 2016

Here we create two dataframes: robbery_2016_actual which has the actual values for the annual amounts of robbery for our 20 chosen states, and robbery_2016_preds which has the predicted values from our model

In [242…

```
robbery_2016= chosen_stats_robbery[chosen_stats_robbery['Year']==2016]
robbery_2016_actual = robbery_2016.drop(columns=['const','Binary','preds_robbery'])
robbery_2016_actual['Type'] = 'Actual'
robbery_2016_actual = robbery_2016_actual.rename(columns = {'actual_robbery':'robbery'})

robbery_2016_preds = robbery_2016.loc[:,['Year','Median_Income', 'preds_robbery','State']]
robbery_2016_preds = robbery_2016_preds.rename(columns = {'preds_robbery':'robbery'})
robbery_2016_preds['Type'] = 'Predicted'
print(robbery_2016_actual.head())
robbery_2016_preds.head()
```

```
     Year  Median_Income   robbery     State    Type
2    2016        72133.0  6.892642    Hawaii  Actual
7    2016        56094.0  8.894396   Indiana  Actual
12   2016        59094.0  7.047517      Iowa  Actual
17   2016        56810.0  7.471363    Kansas  Actual
22   2016        73760.0  9.288689  Maryland  Actual
```

Out[242…

	Year	Median_Income	robbery	State	Type
2	2016	72133.0	7.028096	Hawaii	Predicted
7	2016	56094.0	8.795943	Indiana	Predicted
12	2016	59094.0	7.119791	Iowa	Predicted
17	2016	56810.0	7.756547	Kansas	Predicted
22	2016	73760.0	9.642867	Maryland	Predicted

Merge the two data frames to get a single dataframe containing annual actual and predicted robbery rates.

In [243…
```
robbery_2016_preds_actual = pd.concat([robbery_2016_actual , robbery_2016_preds], ignore_index=True, axis=0)
robbery_2016_preds_actual.head()
```

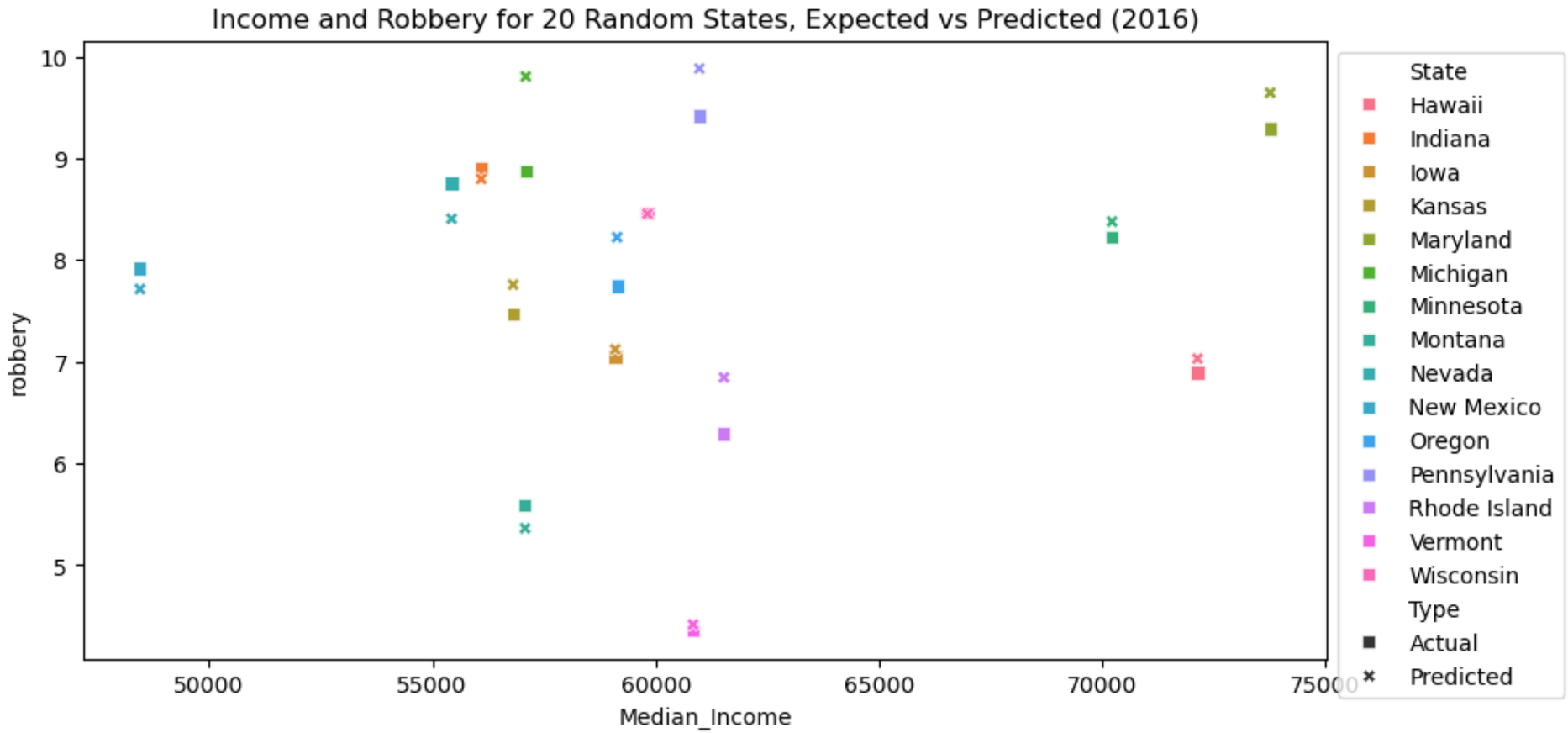
Out[243…

	Year	Median_Income	robbery	State	Type
0	2016	72133.0	6.892642	Hawaii	Actual
1	2016	56094.0	8.894396	Indiana	Actual
2	2016	59094.0	7.047517	Iowa	Actual
3	2016	56810.0	7.471363	Kansas	Actual
4	2016	73760.0	9.288689	Maryland	Actual

In [244…
```
markers = {"Predicted":'X', 'Actual':'s'}
g = sns.scatterplot(data = robbery_2016_preds_actual, x='Median_Income',y='robbery',\
```

```
                 hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Robbery for 20 Random States, Expected vs Predicted (2016)")
```

Out[244… [Text(0.5, 1.0, 'Income and Robbery for 20 Random States, Expected vs Predicted (2016)')]



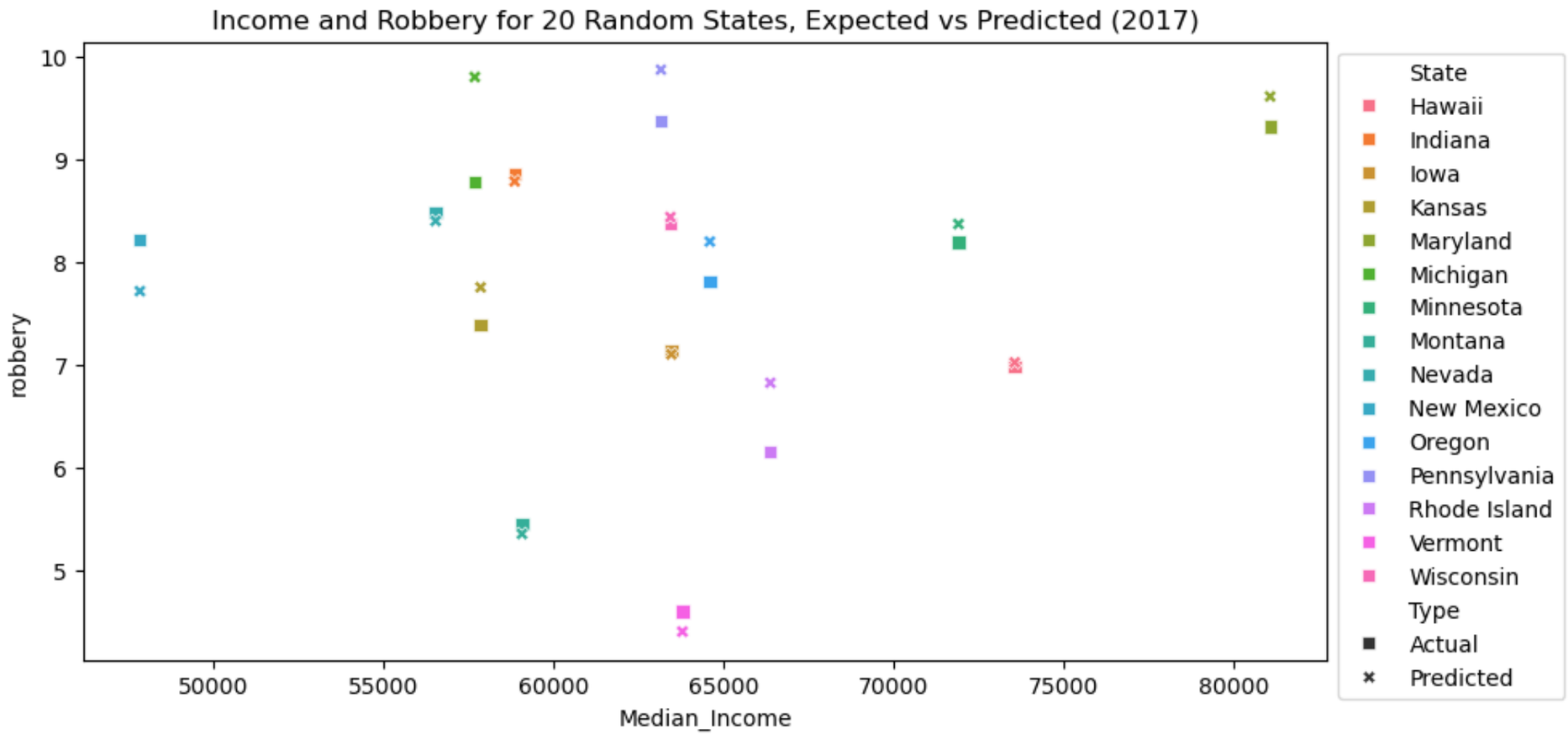
Income and Robbery (Expected vs Predicted) for 2017

Here we create the dataframe robbery_2017_preds_actual which has the actual and predicted robbery rates for our 20 chosen states for the year 2017

In [245…
```
robbery_2017= chosen_stats_robbery[chosen_stats_robbery['Year']==2017]
robbery_2017_actual = robbery_2017.drop(columns=['const','Binary','preds_robbery'])
robbery_2017_actual['Type'] = 'Actual'
robbery_2017_actual = robbery_2017_actual.rename(columns = {'actual_robbery':'robbery'})

robbery_2017_preds = robbery_2017.loc[:,['Year','Median_Income', 'preds_robbery','State']]
robbery_2017_preds = robbery_2017_preds.rename(columns = {'preds_robbery':'robbery'})
robbery_2017_preds['Type'] = 'Predicted'
robbery_2017_preds_actual = pd.concat([robbery_2017_actual , robbery_2017_preds], ignore_index=True, axis=0)
```

In [246…
```
markers = {"Predicted":'X', 'Actual':'s'}
g = sns.scatterplot(data = robbery_2017_preds_actual, x='Median_Income',y='robbery',\
                hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Robbery for 20 Random States, Expected vs Predicted (2017)")
```

Out[246… `[Text(0.5, 1.0, 'Income and Robbery for 20 Random States, Expected vs Predicted (2017)')]`



Income and Robbery (Expected vs Predicted) for 2018

The dataframe robbery_2018_preds_actual which has the actual and predicted robbery rates for the chosen states in the year 2018

In [247…

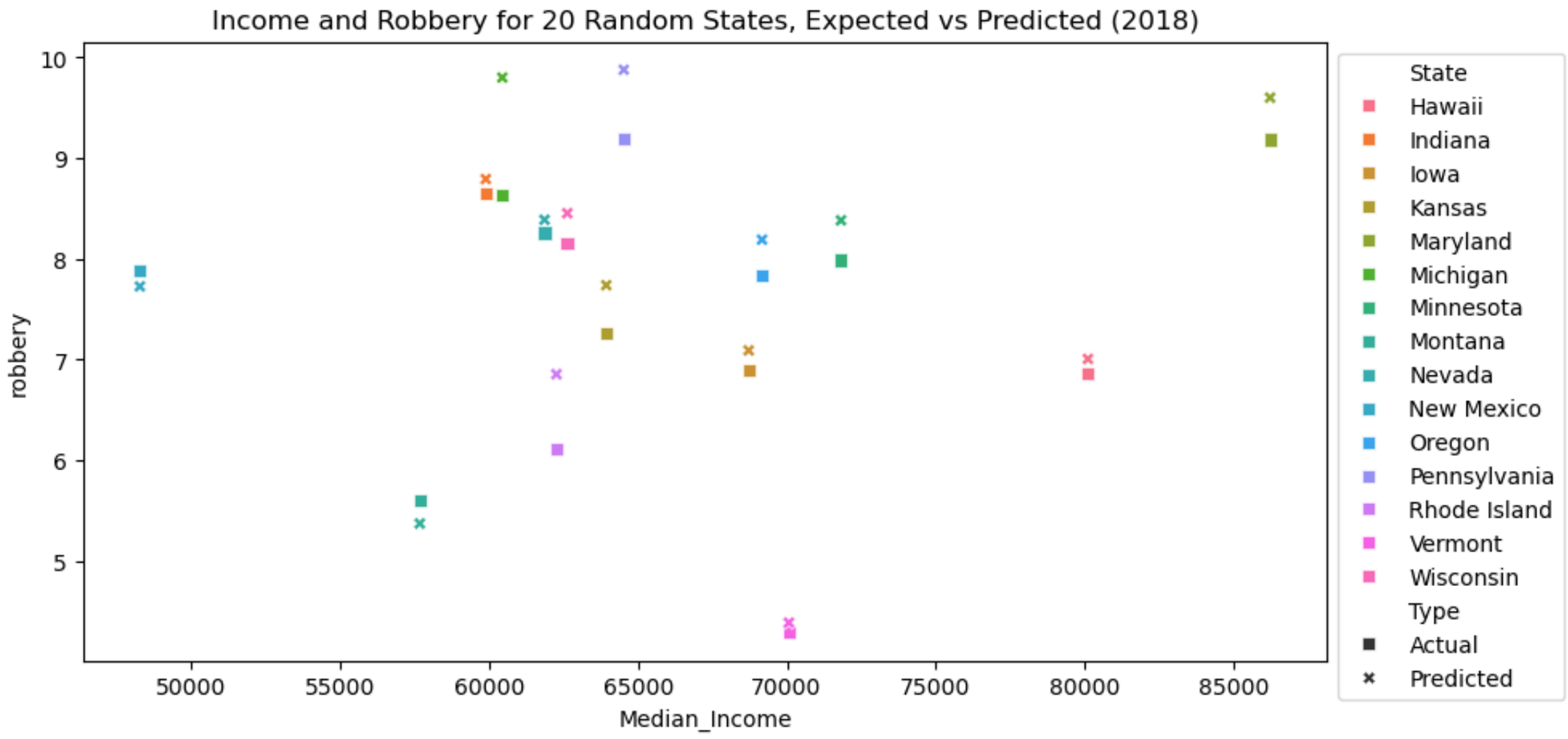
```
robbery_2018= chosen_stats_robbery[chosen_stats_robbery['Year']==2018]
robbery_2018_actual = robbery_2018.drop(columns=['const','Binary','preds_robbery'])
```

```
robbery_2018_actual['Type'] = 'Actual'
robbery_2018_actual = robbery_2018_actual.rename(columns = {'actual_robbery':'robbery'})

robbery_2018_preds = robbery_2018.loc[:,['Year','Median_Income', 'preds_robbery','State']]
robbery_2018_preds = robbery_2018_preds.rename(columns = {'preds_robbery':'robbery'})
robbery_2018_preds['Type'] = 'Predicted'
robbery_2018_preds_actual = pd.concat([robbery_2018_actual , robbery_2018_preds], ignore_index=True, axis=0)
```

In [248…

```
markers = {"Predicted":'X', 'Actual':'s'}
g = sns.scatterplot(data = robbery_2018_preds_actual, x='Median_Income',y='robbery',\
                hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Robbery for 20 Random States, Expected vs Predicted (2018)")
```

Out[248… `[Text(0.5, 1.0, 'Income and Robbery for 20 Random States, Expected vs Predicted (2018)')]`



Now, we train a different model for predicting burglary given income

In [249…

```
income_burglary = duckdb.sql('''SELECT  Burglary,C.State, C.Year, Median_Income
                            FROM crime_df AS C JOIN income_melt AS I
                            ON C.State = I.State AND C.Year = I.Year
                            ORDER BY C.Year ASC''').df()
income_burglary = pd.get_dummies(income_burglary,prefix='', \
```

```
                                prefix_sep='', columns=['State'], dtype=int, drop_first=True)
income_burglary.head()
```

Out[249…

	Burglary	Year	Median_Income	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	...	South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Wash
0	6184	1984	32356.0	1	0	0	0	0	0	0	...	0	0	0	0	0	0	
1	39970	1984	17310.0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	
2	20810	1984	15674.0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	
3	52327	1984	21425.0	0	1	0	0	0	0	0	...	0	0	0	0	0	0	
4	443094	1984	25287.0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	

5 rows × 52 columns

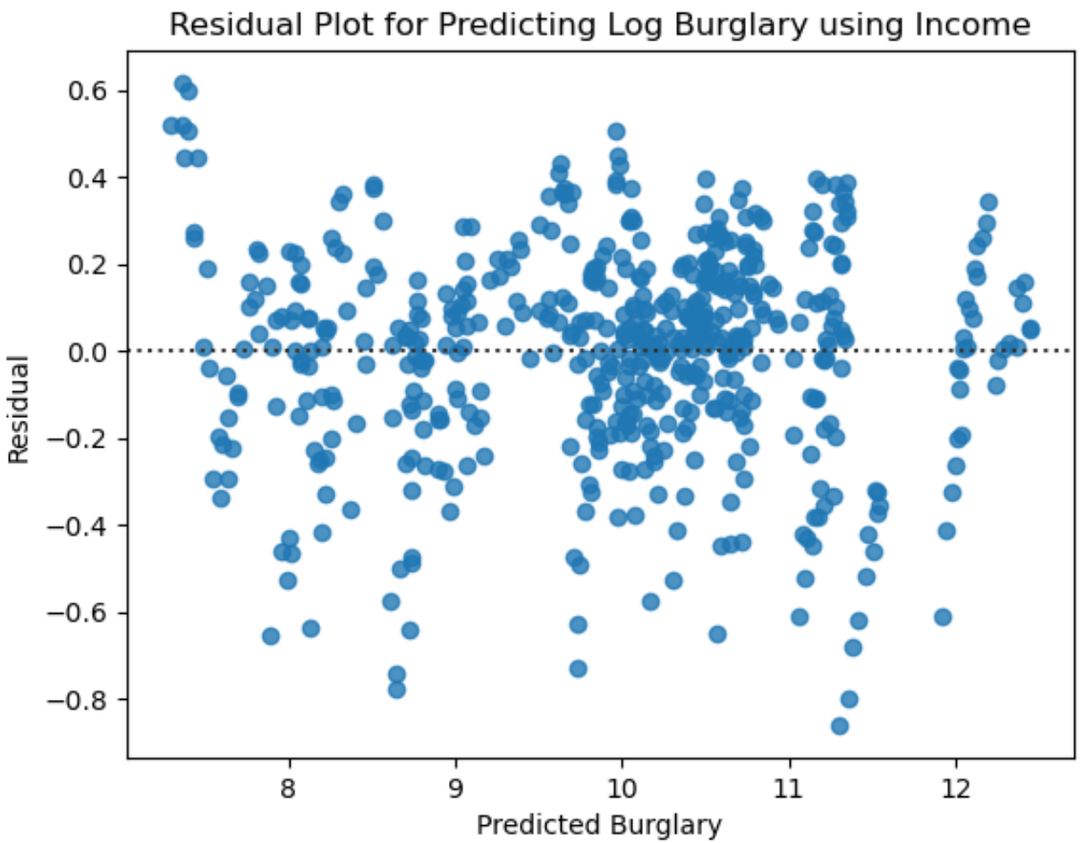
For this regression, we to utilize a log transform on the y (burglary) variable

In [250…

```
income_burglary['log_burglary'] = np.log(income_burglary['Burglary'])

X_inc_burg_train, X_inc_burg_test, y_inc_burg_train, y_inc_burg_test = \
                 train_test_split(income_burglary.iloc[:,1:-1],
                                  income_burglary['log_burglary'], test_size=0.3, shuffle=False)
X_inc_burg_train = sm.add_constant(X_inc_burg_train)
inc_burg_model =sm.OLS(y_inc_burg_train, X_inc_burg_train).fit()

X_inc_burg_test = sm.add_constant(X_inc_burg_test)
inc_burg_preds = inc_burg_model.predict(X_inc_burg_test)
print( f"MAE for predicting burglary using income: {mean_absolute_error(y_inc_burg_test, inc_burg_preds)}")
print( f"RSME for predicting burglary using income: {root_mean_squared_error(y_inc_burg_test, inc_burg_preds)}")

inc_burg_df = pd.DataFrame({"Preds": inc_burg_preds, "Actual": y_inc_burg_test})
ax = sns.residplot(x = inc_burg_preds, y = y_inc_burg_test)
ax.set(xlabel = "Predicted Burglary", ylabel= "Residual",\
```

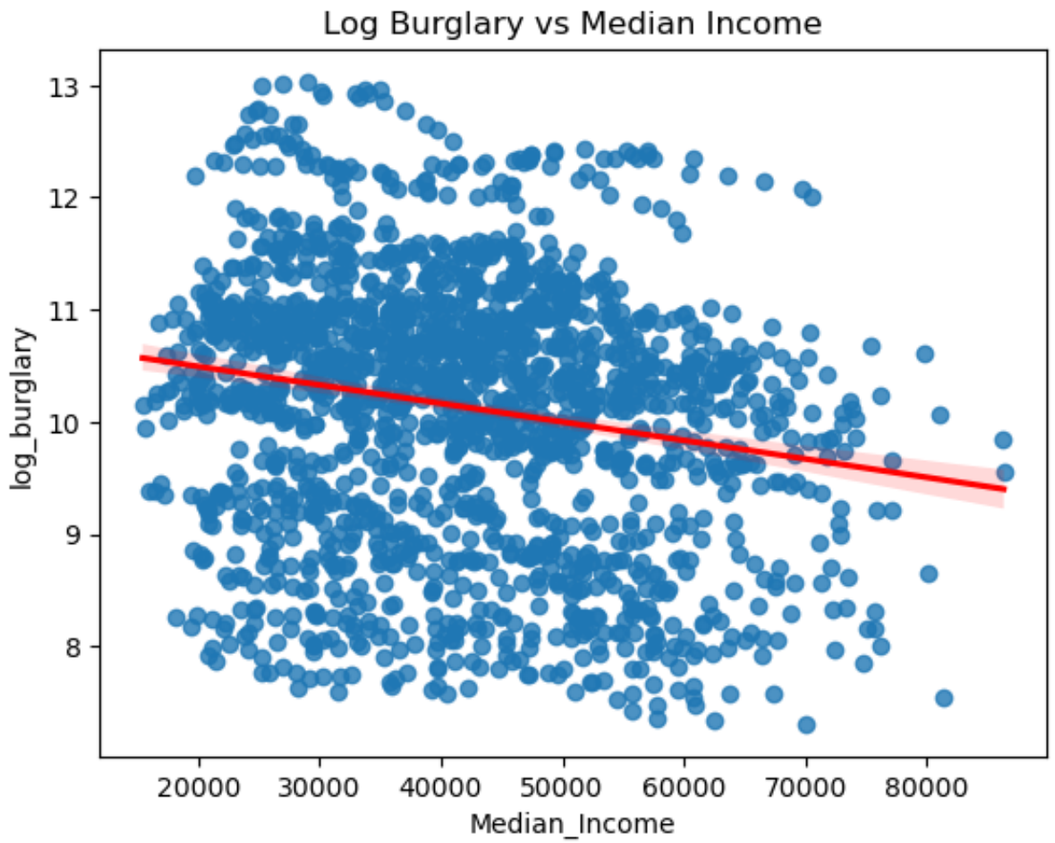
```
        title = "Residual Plot for Predicting Log Burglary using Income");
```

```
MAE for predicting burglary using income: 0.18949502456210937
RSME for predicting burglary using income: 0.25217229608913777
```



In [251…

```
sns.regplot(x = income_burglary['Median_Income'],
            y = income_burglary['log_burglary'],
            line_kws = {"color":"red"})
plt.title("Log Burglary vs Median Income");
```



In [252…
```
print(inc_burg_model.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:           log_burglary   R-squared:                       0.986
Model:                            OLS   Adj. R-squared:                  0.985
Method:                 Least Squares   F-statistic:                     1661.
Date:                Mon, 09 Dec 2024   Prob (F-statistic):               0.00
Time:                        22:15:45   Log-Likelihood:                684.01
```

No. Observations:	1260	AIC:	-1264.			
Df Residuals:	1208	BIC:	-996.8			
Df Model:	51					
Covariance Type:	nonrobust					

	coef	std err	t	P>|t|	[0.025	0.975]
const	21.8900	4.387	4.989	0.000	13.283	30.497
Year	-0.0054	0.002	-2.448	0.015	-0.010	-0.001
Median_Income	-1.028e-05	1.83e-06	-5.616	0.000	-1.39e-05	-6.69e-06
Alaska	-2.1246	0.051	-41.874	0.000	-2.224	-2.025
Arizona	0.2958	0.041	7.249	0.000	0.216	0.376
Arkansas	-0.5578	0.040	-13.912	0.000	-0.636	-0.479
California	2.0601	0.044	46.486	0.000	1.973	2.147
Colorado	-0.1112	0.045	-2.481	0.013	-0.199	-0.023
Connecticut	-0.4197	0.050	-8.398	0.000	-0.518	-0.322
Delaware	-1.8651	0.044	-42.427	0.000	-1.951	-1.779
Florida	1.5848	0.040	39.453	0.000	1.506	1.664
Georgia	0.7022	0.041	17.165	0.000	0.622	0.782
Hawaii	-1.1825	0.048	-24.775	0.000	-1.276	-1.089
Idaho	-1.6700	0.041	-40.940	0.000	-1.750	-1.590
Illinois	0.9077	0.043	20.906	0.000	0.823	0.993
Indiana	0.1012	0.041	2.459	0.014	0.020	0.182
Iowa	-0.7443	0.041	-18.054	0.000	-0.825	-0.663
Kansas	-0.5518	0.041	-13.383	0.000	-0.633	-0.471
Kentucky	-0.4739	0.040	-11.781	0.000	-0.553	-0.395
Louisiana	0.1539	0.040	3.826	0.000	0.075	0.233
Maine	-1.6549	0.041	-40.551	0.000	-1.735	-1.575
Maryland	0.2032	0.050	4.046	0.000	0.105	0.302
Massachusetts	0.1554	0.047	3.337	0.001	0.064	0.247
Michigan	0.7841	0.043	18.322	0.000	0.700	0.868
Minnesota	-0.1655	0.046	-3.638	0.000	-0.255	-0.076
Mississippi	-0.4469	0.041	-10.991	0.000	-0.527	-0.367
Missouri	0.1439	0.041	3.493	0.000	0.063	0.225
Montana	-2.2653	0.040	-56.315	0.000	-2.344	-2.186
Nebraska	-1.3940	0.041	-33.625	0.000	-1.475	-1.313
Nevada	-0.7668	0.043	-17.703	0.000	-0.852	-0.682

```
New Hampshire     -1.9385      0.048    -40.003      0.000      -2.034      -1.843
New Jersey         0.4347      0.051      8.555      0.000       0.335       0.534
New Mexico        -0.6701      0.040    -16.658      0.000      -0.749      -0.591
New York           1.0954      0.042     26.106      0.000       1.013       1.178
North Carolina     0.8046      0.041     19.835      0.000       0.725       0.884
North Dakota      -2.9270      0.040    -72.449      0.000      -3.006      -2.848
Ohio               0.8659      0.042     20.739      0.000       0.784       0.948
Oklahoma          -0.0843      0.040     -2.094      0.036      -0.163      -0.005
Oregon            -0.2597      0.042     -6.215      0.000      -0.342      -0.178
Pennsylvania       0.4744      0.042     11.363      0.000       0.392       0.556
Rhode Island      -1.6123      0.043    -37.603      0.000      -1.696      -1.528
South Carolina     0.0263      0.040      0.651      0.515      -0.053       0.106
South Dakota      -2.5563      0.040    -63.230      0.000      -2.636      -2.477
Tennessee          0.2977      0.040      7.401      0.000       0.219       0.377
Texas              1.7287      0.041     42.268      0.000       1.648       1.809
Utah              -0.9711      0.044    -21.884      0.000      -1.058      -0.884
Vermont           -2.2919      0.042    -54.302      0.000      -2.375      -2.209
Virginia          -0.0358      0.046     -0.772      0.440      -0.127       0.055
Washington         0.4264      0.044      9.630      0.000       0.340       0.513
West Virginia     -1.4444      0.041    -35.542      0.000      -1.524      -1.365
Wisconsin         -0.2873      0.043     -6.666      0.000      -0.372      -0.203
Wyoming           -2.7145      0.041    -65.604      0.000      -2.796      -2.633
==============================================================================
Omnibus:                       29.839   Durbin-Watson:                   1.882
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               63.849
Skew:                           0.027   Prob(JB):                     1.37e-14
Kurtosis:                       4.102   Cond. No.                     4.08e+07
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.08e+07. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Looking at this OLS regression, we observe that the p values for income and year are significant, 0.000 and 0.015, respectively (based on a 0.05 significance level). This shows that the significant coefficient for income is -1.028e-05. This shows that for every one dollar increase in income, e^(-1.028e-05) unit decrease of burglaries. Although our hypothesis is

not exactly testing income having a relationship with a states' number of burglaries, the p-value still now tells us that it is a strong determinor of burglaries.

Now, plot the predicted vs actual values of burglary when using income as a predictor

In [253…
```
X_inc_burg_test2 = X_inc_burg_test
X_inc_burg_test2['actual_burglary'] = y_inc_burg_test
X_inc_burg_test2['preds_burglary'] = inc_burg_preds
inc_burg_melt = pd.melt(X_inc_burg_test2, \
                        id_vars=['const','Year', 'Median_Income', 'actual_burglary','preds_burglary'], \
                       var_name = 'State', value_name='Binary')
inc_burg_melt = inc_burg_melt[ inc_burg_melt['Binary']==1]
inc_burg_melt.head()
```

Out[253…

	const	Year	Median_Income	actual_burglary	preds_burglary	State	Binary
40	1.0	2010	57848.0	8.040769	8.225878	Alaska	1
90	1.0	2011	57431.0	7.955776	8.224721	Alaska	1
140	1.0	2012	63648.0	7.989560	8.155338	Alaska	1
190	1.0	2013	61137.0	7.978311	8.175717	Alaska	1
229	1.0	2013	72472.0	7.978311	8.059143	Alaska	1

We again choose 20 states at random to focus on in looking at the predicted and actual values of burglary rates

In [254…
```
chosen_stats_burglary = duckdb.sql(''' SELECT * FROM inc_burg_melt
                                WHERE Year>=2014 AND Year<=2018
                                 AND State IN (SELECT * FROM twenty_states)''').df()
chosen_stats_burglary.head()
```

Out[254…

	const	Year	Median_Income	actual_burglary	preds_burglary	State	Binary
0	1.0	2014	71223.0	8.918650	9.008606	Hawaii	1
1	1.0	2015	64514.0	8.826147	9.072159	Hawaii	1
2	1.0	2016	72133.0	8.696677	8.988357	Hawaii	1
3	1.0	2017	73575.0	8.621373	8.968082	Hawaii	1
4	1.0	2018	80108.0	8.649799	8.895449	Hawaii	1

We follow a similar methodology from the previous model where we used income to predict burglary. Here, we will look at separate plots for years 2016, 2017, 2018 containing the predicted and actual values of burglary when using income in our regression

Income and Burglary (Expected vs Predicted) for 2016

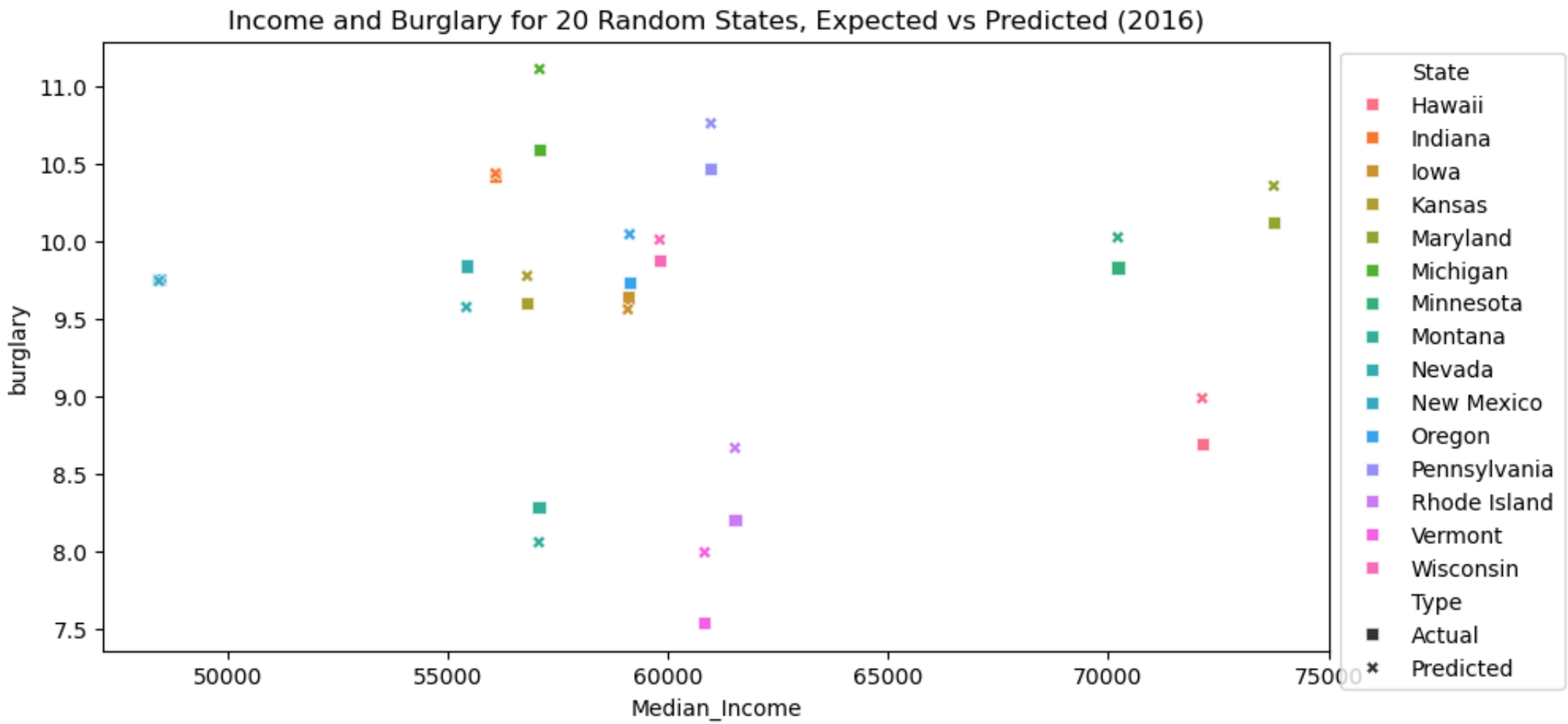
In [255…

```
burglary_2016= chosen_stats_burglary[chosen_stats_burglary['Year']==2016]
burglary_2016_actual = burglary_2016.drop(columns=['const','Binary','preds_burglary'])
burglary_2016_actual['Type'] = 'Actual'
burglary_2016_actual = burglary_2016_actual.rename(columns = {'actual_burglary':'burglary'})

burglary_2016_preds = burglary_2016.loc[:,['Year','Median_Income', 'preds_burglary','State']]
burglary_2016_preds = burglary_2016_preds.rename(columns = {'preds_burglary':'burglary'})
burglary_2016_preds['Type'] = 'Predicted'
burglary_2016_preds_actual = pd.concat([burglary_2016_actual , burglary_2016_preds], ignore_index=True, axis=0)
#burglary_2016_preds_actual
```

In [256…

```
g = sns.scatterplot(data = burglary_2016_preds_actual, x='Median_Income',y='burglary',\
                hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Burglary for 20 Random States, Expected vs Predicted (2016)")
```

Out[256… [Text(0.5, 1.0, 'Income and Burglary for 20 Random States, Expected vs Predicted (2016)')]



Income and Burglary (Expected vs Predicted) for 2017

In [257…

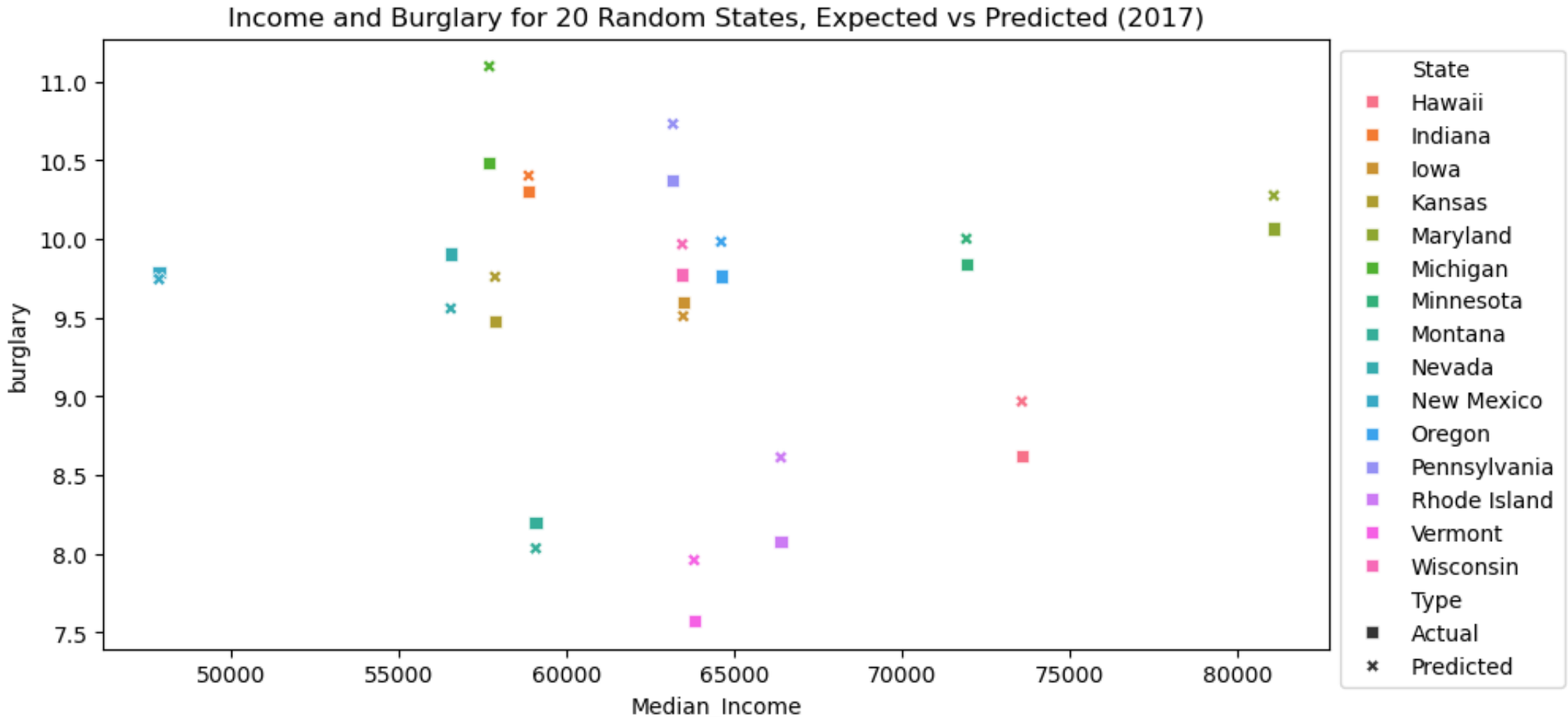
```
burglary_2017= chosen_stats_burglary[chosen_stats_burglary['Year']==2017]
burglary_2017_actual = burglary_2017.drop(columns=['const','Binary','preds_burglary'])
```

```
burglary_2017_actual['Type'] = 'Actual'
burglary_2017_actual = burglary_2017_actual.rename(columns = {'actual_burglary':'burglary'})

burglary_2017_preds = burglary_2017.loc[:,['Year','Median_Income', 'preds_burglary','State']]
burglary_2017_preds = burglary_2017_preds.rename(columns = {'preds_burglary':'burglary'})
burglary_2017_preds['Type'] = 'Predicted'
burglary_2017_preds_actual = pd.concat([burglary_2017_actual , burglary_2017_preds], ignore_index=True, axis=0)
#burglary_2017_preds_actual
```

In [258…

```
g = sns.scatterplot(data = burglary_2017_preds_actual, x='Median_Income',y='burglary',\
                hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Burglary for 20 Random States, Expected vs Predicted (2017)")
```

Out[258… `[Text(0.5, 1.0, 'Income and Burglary for 20 Random States, Expected vs Predicted (2017)')]`



Income and Burglary (Expected vs Predicted) for 2018

In [259…

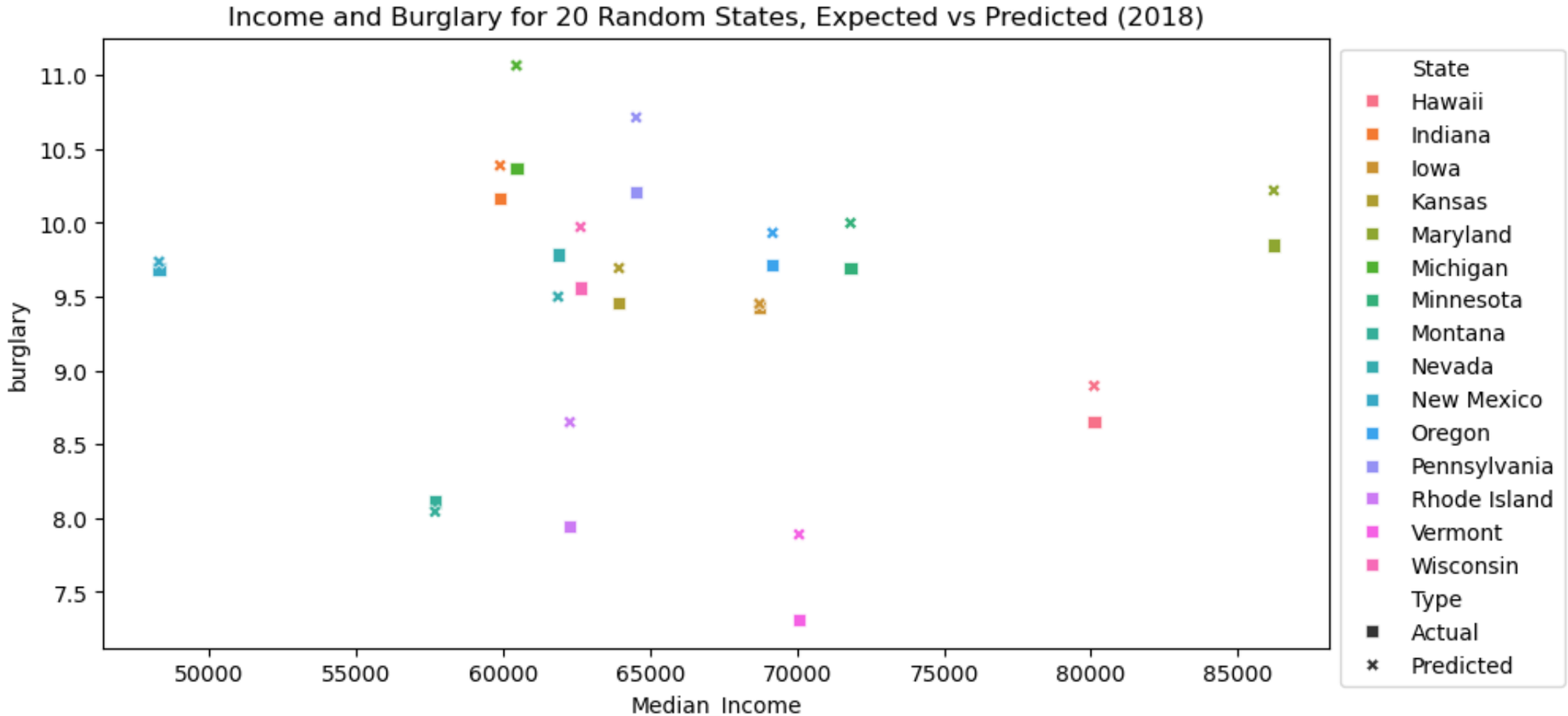
```
burglary_2018= chosen_stats_burglary[chosen_stats_burglary['Year']==2018]
burglary_2018_actual = burglary_2018.drop(columns=['const','Binary','preds_burglary'])
burglary_2018_actual['Type'] = 'Actual'
burglary_2018_actual = burglary_2018_actual.rename(columns = {'actual_burglary':'burglary'})
```

```
burglary_2018_preds = burglary_2018.loc[:,['Year','Median_Income', 'preds_burglary','State']]
burglary_2018_preds = burglary_2018_preds.rename(columns = {'preds_burglary':'burglary'})
burglary_2018_preds['Type'] = 'Predicted'
burglary_2018_preds_actual = pd.concat([burglary_2018_actual , burglary_2018_preds], ignore_index=True, axis=0)
#burglary_2018_preds_actual
```

In [260…

```
g = sns.scatterplot(data = burglary_2018_preds_actual, x='Median_Income',y='burglary',\
                hue='State', style = 'Type', markers = markers)
g.figure.set_size_inches(10,5)
sns.move_legend(g, "upper left", bbox_to_anchor=(1, 1))
g.set(title = "Income and Burglary for 20 Random States, Expected vs Predicted (2018)")
```

Out[260… [Text(0.5, 1.0, 'Income and Burglary for 20 Random States, Expected vs Predicted (2018)')]



Hypothesis 2

There is a relationship among states that are above the U.S. unemployment average over the years 1980 to 2018 in predicting violent crime rates, and there is a relationship among states that are below the U.S. unemploymnent average over the years 1980 to 2018 in predicting violent crime rates.

In [261…
```
#avg unemployment in the US
us_avg = job_df.iloc[0, 1:]
#print(us_avg)
us_average = us_avg.sum()/us_avg.size
print(us_average)
```

```
6.264102564102564
```

Looking across all the years from 1980 to 2018, the U.S. unemployment rate average is about 6.264.

In [262…
```
#created a dictionary with all states and their respective unemployment averages across the years 1980 to 2018
dict_avg = {}
states = job_df.iloc[1:, 0].reset_index(drop=True) #rows 0,1,...50
#print(states)

#for x in range(1, 38):
for x in range(0,states.size):
    the_avg = job_df.iloc[x+1, 1:] #get the row for this state
    dict_avg[states[x]] = the_avg.sum()/(2018-1980) + 1

print(dict_avg)
```

```
{'Alabama': 8.339473684210525, 'Alaska': 8.976315789473684, 'Arizona': 7.371052631578947, 'Arkansas': 7.628947368421051, 'California': 8.3973684210526
3, 'Colorado': 6.46578947368421, 'Connecticut': 6.518421052631578, 'Delaware': 6.278947368421052, 'District of Columbia': 8.736842105263161, 'Florida':
7.197368421052631, 'Georgia': 7.1236842105263145, 'Hawaii': 5.7026315789473685, 'Idaho': 7.094736842105262, 'Illinois': 8.157894736842106, 'Indiana':
7.342105263157896, 'Iowa': 5.755263157894736, 'Kansas': 5.88157894736842, 'Kentucky': 8.06578947368421, 'Louisiana': 8.507894736842106, 'Maine': 6.7763
15789473684, 'Maryland': 6.360526315789474, 'Massachusetts': 6.457894736842107, 'Michigan': 9.092105263157896, 'Minnesota': 5.978947368421053, 'Mississ
ippi': 8.807894736842107, 'Missouri': 7.123684210526315, 'Montana': 6.8921052631578945, 'Nebraska': 4.6421052631578945, 'Nevada': 7.75, 'New Hampshir
e': 5.3578947368421055, 'New Jersey': 7.221052631578949, 'New Mexico': 7.849999999999999, 'New York': 7.4710526315789485, 'North Carolina': 6.989473684
210526, 'North Dakota': 4.9184210526315795, 'Ohio': 7.9026315789473704, 'Oklahoma': 6.3500000000000005, 'Oregon': 8.228947368421053, 'Pennsylvania': 7.
513157894736843, 'Rhode Island': 7.563157894736842, 'South Carolina': 7.702631578947368, 'South Dakota': 4.83421052631579, 'Tennessee': 7.5868421052631
56, 'Texas': 7.24210526315789, 'Utah': 5.984210526315789, 'Vermont': 5.5815789473684205, 'Virginia': 5.781578947368419, 'Washington': 8.07631578947368
4, 'West Virginia': 9.45526315789474, 'Wisconsin': 6.6921052631578934, 'Wyoming': 6.218421052631579}
```

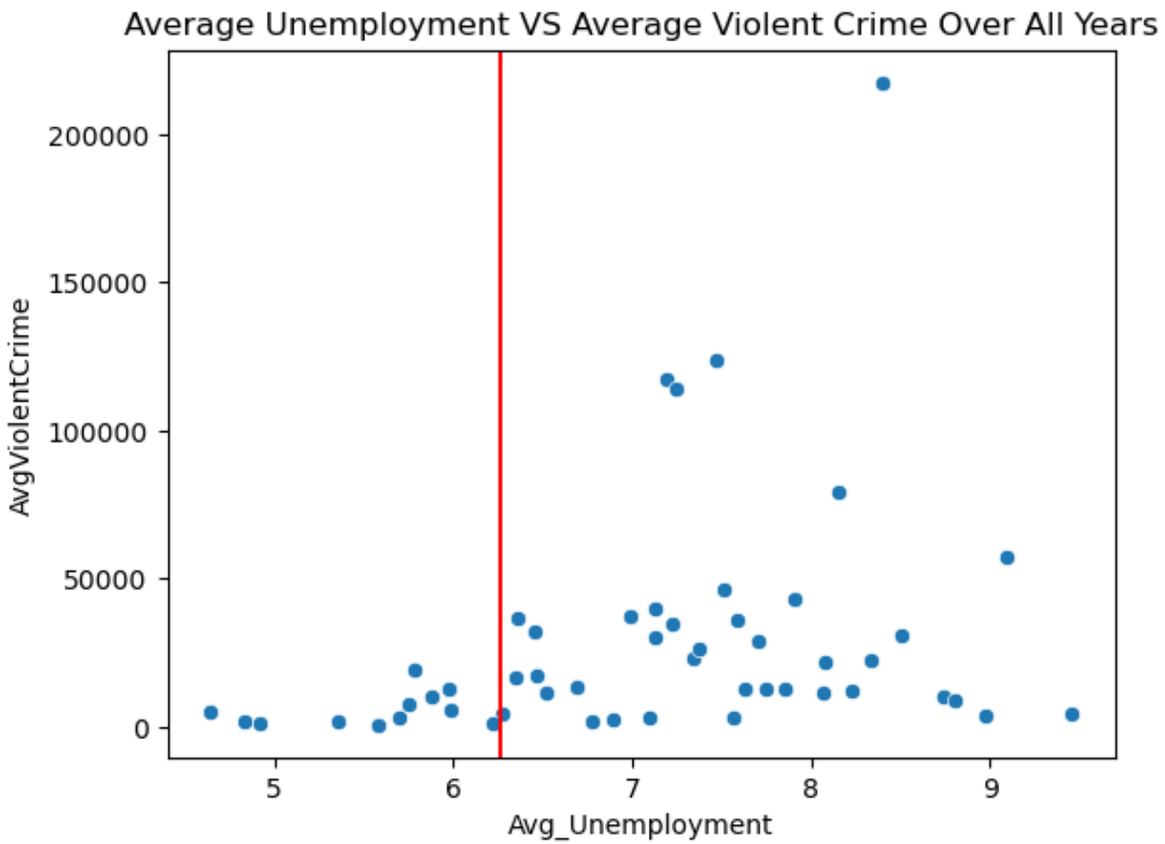
Here we create a dataframe containing the average amount of violent crime, and the average unemployment rate for each state taken over all the years

In [263…
```
states_unemp_avg_df = pd.DataFrame({'State':np.asarray(states), 'Avg_Unemployment':dict_avg.values()})
avg_violent_crime_states = duckdb.sql('''SELECT
                                    SUM(Violent_Crime)/Count(*) AS AvgViolentCrime,
                                    State
                                    FROM crime_df AS C
                                    GROUP BY State''').df()
states_unemp_avg_df = duckdb.sql(''' SELECT A.*, Avg_Unemployment
                                FROM avg_violent_crime_states
                                AS A JOIN states_unemp_avg_df AS S
                                 ON A.State  = S.State''').df()
states_unemp_avg_df.head()
```

Out[263…

	AvgViolentCrime	State	Avg_Unemployment
0	13012.511628	Arkansas	7.628947
1	3258.279070	Idaho	7.094737
2	12886.255814	New Mexico	7.850000
3	28915.255814	South Carolina	7.702632
4	21883.976744	Washington	8.076316

In [264…
```
g= sns.scatterplot(data = states_unemp_avg_df, x ='Avg_Unemployment', y='AvgViolentCrime');
plt.axvline(x = us_average, color = "red");
plt.title('Average Unemployment VS Average Violent Crime Over All Years');
```



This scatterplot is a data visualization that demonstrates the national average unemployment rate as a vertical line. Then, it also represents the 50 states that are scattered across, showing that there are significantly more states that have unemployment rates greater than the national average, comparatively to the number of states below the national average.

Particularly, we plotted this average unemployment rate against the average violent crime (y axis). The general trend that this plot exemplifies is that a greater average unemployment rate corresponds to a greater average violent crime rate. We seek to investigate this through a comparison of two different models, one for states that are below the national unemployment rate average and one for states that are above the national unemployment rate average.

First, we create two lists: one with states having unemployment rates above the antional average, and another for those below.

In [265…
```
# put all states that are lower than the national average in one list
# put all states that are above the national average in another list

less_dict = []
more_dict = []
count = 2
dict_avg[states[count]]

for x in dict_avg:
    rate = dict_avg[x]
    if rate < us_average:
        less_dict.append(x)
    else:
        more_dict.append(x)

print(less_dict)
print(more_dict)
```

```
['Hawaii', 'Iowa', 'Kansas', 'Minnesota', 'Nebraska', 'New Hampshire', 'North Dakota', 'South Dakota', 'Utah', 'Vermont', 'Virginia', 'Wyoming']
['Alabama', 'Alaska', 'Arizona', 'Arkansas', 'California', 'Colorado', 'Connecticut', 'Delaware', 'District of Columbia', 'Florida', 'Georgia', 'Idah
o', 'Illinois', 'Indiana', 'Kentucky', 'Louisiana', 'Maine', 'Maryland', 'Massachusetts', 'Michigan', 'Mississippi', 'Missouri', 'Montana', 'Nevada', '
New Jersey', 'New Mexico', 'New York', 'North Carolina', 'Ohio', 'Oklahoma', 'Oregon', 'Pennsylvania', 'Rhode Island', 'South Carolina', 'Tennessee', '
Texas', 'Washington', 'West Virginia', 'Wisconsin']
```

In [266…
```
#now get the info for states that have above average
states_above_df = job_melt[ job_melt['Area'].isin(more_dict)]
states_above_df.head()
```

Out[266…

	Area	Year	UnemploymentRate
1	Alabama	1980	8.9
2	Alaska	1980	9.6
3	Arizona	1980	6.6
4	Arkansas	1980	7.6
5	California	1980	6.8

In [267…

```
states_above_df = duckdb.sql('''Select  C.Violent_Crime, S.*
                            FROM states_above_df AS S JOIN crime_df AS C
                            ON  S.Year = C.Year AND S.Area = C.State''').df()
states_above_df.head()
```

Out[267…

	Violent_Crime	Area	Year	UnemploymentRate
0	1919	Alaska	1980	9.6
1	17320	Alabama	1980	8.9
2	7656	Arkansas	1980	7.6
3	210290	California	1980	6.8
4	15215	Colorado	1980	5.8

In [268…

```
states_above_df = pd.get_dummies(states_above_df, prefix='', prefix_sep='',\
                                 drop_first=True, dtype=int, columns=['Area'])
states_above_df['Year'] = states_above_df['Year'].astype(int)
states_above_df.head()
```

Out[268…

	Violent_Crime	Year	UnemploymentRate	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware	...	Oklahoma	Oregon	Pennsylvania	Rhode Island	Sout Carolin
0	1919	1980	9.6	1	0	0	0	0	0	0	...	0	0	0	0	
1	17320	1980	8.9	0	0	0	0	0	0	0	...	0	0	0	0	
2	7656	1980	7.6	0	0	1	0	0	0	0	...	0	0	0	0	
3	210290	1980	6.8	0	0	0	1	0	0	0	...	0	0	0	0	
4	15215	1980	5.8	0	0	0	0	1	0	0	...	0	0	0	0	

5 rows × 41 columns

Train the model to predict crime rate given unemployment rate using data for years 1980-2018 corresponding to states with an average unemployment rate above the national average

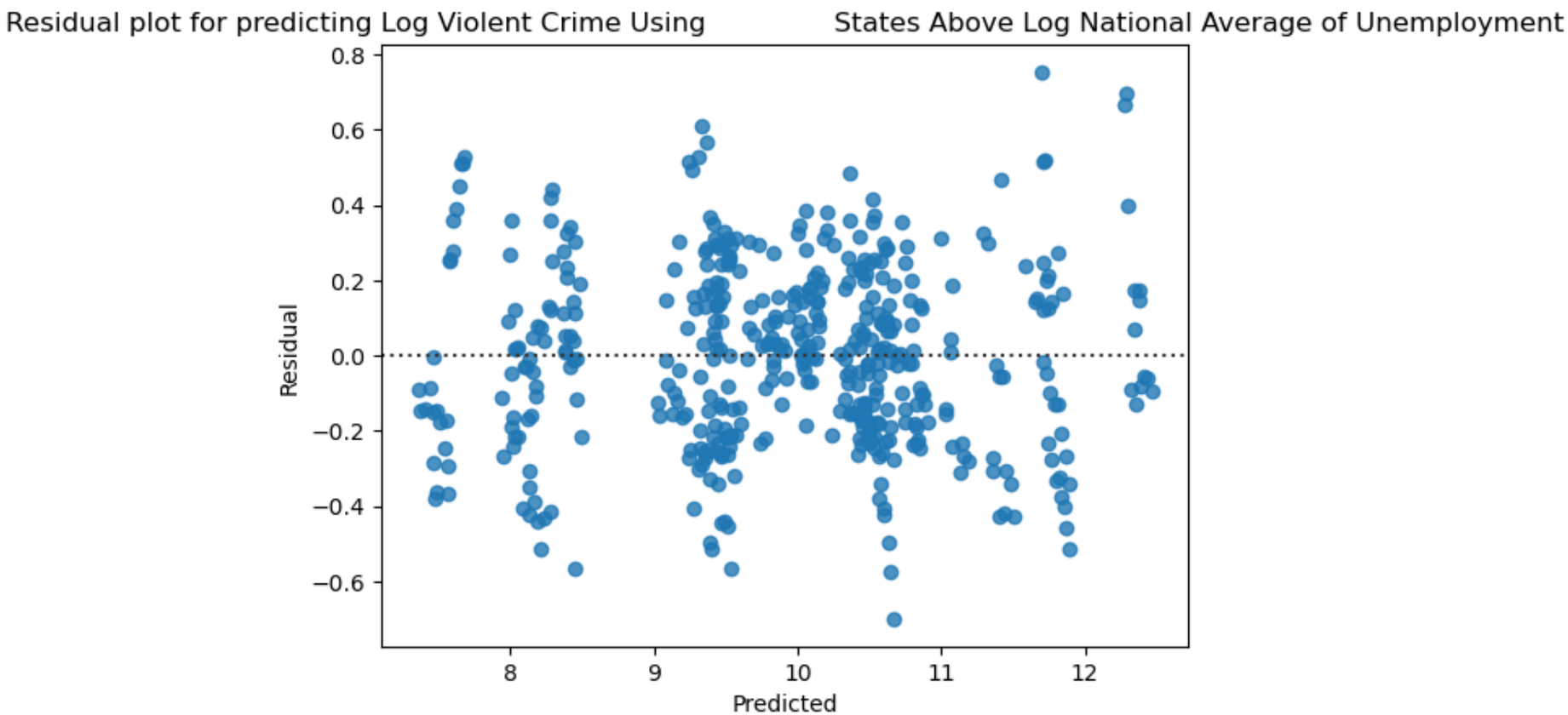
We apply a log transform on the y axis (violent crime rate), as well as on the x axis (unemployment rate).

In [269…

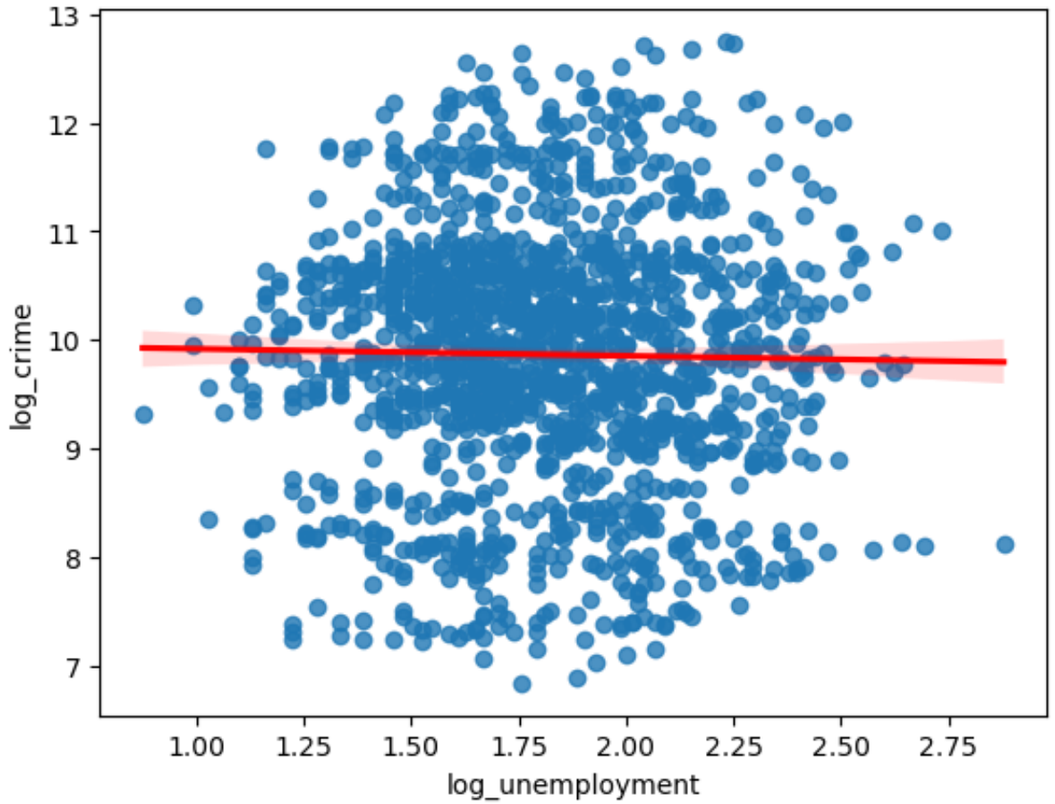
```
states_above_df['log_unemployment'] = np.log(states_above_df['UnemploymentRate'])
states_above_df['log_crime'] = np.log(states_above_df['Violent_Crime'])
modified_df = states_above_df.drop(columns = "UnemploymentRate")

#print(modifed_df.iloc[:, 1:-1].columns)
X_train_above, X_test_above, Y_train_above, Y_test_above = \
  train_test_split(modified_df.iloc[:, 1:-1],modified_df['log_crime'],\
                   test_size=.30, shuffle=False)
X_train_above = sm.add_constant(X_train_above)
model_above = sm.OLS( Y_train_above,X_train_above).fit()

X_test_above = sm.add_constant(X_test_above)
preds_above = model_above.predict(X_test_above)
y_above_df = pd.DataFrame(data = {"Preds":preds_above, "Actual":Y_test_above})
```

```
y_above_df.head()

ax = sns.residplot(data= y_above_df, x = 'Preds', y = 'Actual')
ax.set(xlabel = "Predicted", ylabel= "Residual",
       title="Residual plot for predicting Log Violent Crime Using \
              States Above Log National Average of Unemployment");

print( f"MAE for predicting violent crime rate using states above national average: \
      {mean_absolute_error(preds_above,Y_test_above)}")
print( f"RSME for predicting violent crime rate using states above national average: \
      {root_mean_squared_error(preds_above,Y_test_above)}")
```

```
MAE for predicting violent crime rate using states above national average:       0.21877660270961435
RSME for predicting violent crime rate using states above national average:       0.272094301110766
```



In [270…
```
sns.regplot(x = states_above_df['log_unemployment'], y = states_above_df['log_crime'], line_kws = {"color":"red"});
```



In [271…
```
print(model_above.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_crime   R-squared:                       0.970
Model:                            OLS   Adj. R-squared:                  0.969
Method:                 Least Squares   F-statistic:                     820.0
Date:                Mon, 09 Dec 2024   Prob (F-statistic):               0.00
Time:                        22:15:49   Log-Likelihood:                180.36
No. Observations:                1064   AIC:                            -278.7
```

```
Df Residuals:                     1023   BIC:                          -74.95
Df Model:                           40
Covariance Type:            nonrobust
================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const                    5.3260      1.420      3.752      0.000       2.540       8.112
Year                     0.0025      0.001      3.561      0.000       0.001       0.004
Alaska                  -1.7643      0.055    -31.951      0.000      -1.873      -1.656
Arizona                  0.0934      0.055      1.703      0.089      -0.014       0.201
Arkansas                -0.6329      0.054    -11.662      0.000      -0.739      -0.526
California               2.3249      0.056     41.320      0.000       2.215       2.435
Colorado                -0.3461      0.055     -6.276      0.000      -0.454      -0.238
Connecticut             -0.6069      0.055    -10.986      0.000      -0.715      -0.499
Delaware                -1.6620      0.056    -29.510      0.000      -1.772      -1.551
District of Columbia    -0.6842      0.055    -12.501      0.000      -0.792      -0.577
Florida                  1.7291      0.055     31.527      0.000       1.621       1.837
Georgia                  0.5738      0.056     10.258      0.000       0.464       0.684
Idaho                   -2.0017      0.058    -34.739      0.000      -2.115      -1.889
Illinois                 1.3705      0.056     24.356      0.000       1.260       1.481
Indiana                 -0.0293      0.056     -0.524      0.600      -0.139       0.080
Kentucky                -0.5814      0.055    -10.628      0.000      -0.689      -0.474
Louisiana                0.3910      0.057      6.874      0.000       0.279       0.503
Maine                   -2.6097      0.056    -46.613      0.000      -2.720      -2.500
Maryland                 0.5177      0.057      9.105      0.000       0.406       0.629
Massachusetts            0.4204      0.056      7.553      0.000       0.311       0.530
Michigan                 1.0405      0.055     18.834      0.000       0.932       1.149
Mississippi             -0.8595      0.058    -14.947      0.000      -0.972      -0.747
Missouri                 0.2970      0.054      5.505      0.000       0.191       0.403
Montana                 -2.4862      0.056    -44.509      0.000      -2.596      -2.377
Nevada                  -0.7122      0.055    -12.995      0.000      -0.820      -0.605
New Jersey               0.5258      0.056      9.412      0.000       0.416       0.635
New Mexico              -0.6179      0.057    -10.864      0.000      -0.730      -0.506
New York                 1.7491      0.055     31.641      0.000       1.641       1.858
North Carolina           0.4937      0.055      8.959      0.000       0.386       0.602
Ohio                     0.7271      0.057     12.647      0.000       0.614       0.840
Oklahoma                -0.2948      0.057     -5.152      0.000      -0.407      -0.183
```

```
Oregon                -0.5417      0.057     -9.424      0.000      -0.655      -0.429
Pennsylvania           0.7768      0.056     13.788      0.000       0.666       0.887
Rhode Island          -1.9353      0.056    -34.297      0.000      -2.046      -1.825
South Carolina         0.3071      0.057      5.400      0.000       0.195       0.419
Tennessee              0.4196      0.055      7.665      0.000       0.312       0.527
Texas                  1.6153      0.054     29.728      0.000       1.509       1.722
Washington             0.0109      0.056      0.193      0.847      -0.100       0.121
West Virginia         -1.6539      0.056    -29.627      0.000      -1.763      -1.544
Wisconsin             -0.6074      0.055    -11.127      0.000      -0.714      -0.500
log_unemployment      -0.1784      0.023     -7.817      0.000      -0.223      -0.134
==============================================================================
Omnibus:                        0.948   Durbin-Watson:                   1.577
Prob(Omnibus):                  0.623   Jarque-Bera (JB):                0.858
Skew:                          -0.065   Prob(JB):                        0.651
Kurtosis:                       3.049   Cond. No.                     4.44e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.44e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

In [272…
```
print(model_above.params)
```

```
const                  5.325987
Year                   0.002511
Alaska                -1.764312
Arizona                0.093420
Arkansas              -0.632854
California             2.324917
Colorado              -0.346144
Connecticut           -0.606923
Delaware              -1.661968
District of Columbia  -0.684183
Florida                1.729079
Georgia                0.573794
Idaho                 -2.001730
```

```
Illinois              1.370516
Indiana              -0.029317
Kentucky             -0.581431
Louisiana             0.390955
Maine                -2.609738
Maryland              0.517662
Massachusetts         0.420433
Michigan              1.040505
Mississippi          -0.859466
Missouri              0.296987
Montana              -2.486209
Nevada               -0.712211
New Jersey            0.525756
New Mexico           -0.617922
New York              1.749060
North Carolina        0.493699
Ohio                  0.727094
Oklahoma             -0.294843
Oregon               -0.541732
Pennsylvania          0.776770
Rhode Island         -1.935292
South Carolina        0.307100
Tennessee             0.419632
Texas                 1.615296
Washington            0.010853
West Virginia        -1.653911
Wisconsin            -0.607364
log_unemployment     -0.178411
dtype: float64
```

Looking at this OLS regression, we observe that the p value for unemployment rate (what we are interested in) is significant (based on a 0.05 significance level). This shows that we can reject the null hypothesis (that there is no relationship between unemployment rate among states that are above the national average and violent crime rates). There exists a relationship, as exemplified by our p value that is less than the significance level.

Our model takes the form of

$$\ln(\text{crime}) = \beta_0 \cdot \ln(\text{unemploymentRate}) + \beta_1 x_1 + \beta_2 x_2 + \dots$$

$$\exp^{\ln(\text{crime})} = \exp^{\beta_0 \cdot \ln(\text{unemploymentRate})} \cdot \exp^{\beta_1 x_1} \cdot \exp^{\beta_2 x_2} \cdot$$

$$\text{crime} = (\text{unemploymentRate})^{\beta_0} \cdot \exp^{\beta_1 x_1} \cdot \exp^{\beta_2 x_2} \dots$$

This means that when the unemployemnt rate changes by a factor of c, the amount of violent crime changes by a factor of c_0^{β} where $\beta_0 = -0.178411$ is the coefficient from the model for ln(unemployment)

Now we train a model to predict violent crime rate using the unemployment rates for states below the national average

In [273…
```
states_below_df = job_melt[ job_melt['Area'].isin(less_dict)]
states_below_df.head()
```

Out[273…

	Area	Year	UnemploymentRate
12	Hawaii	1980	5.0
16	Iowa	1980	6.0
17	Kansas	1980	4.4
24	Minnesota	1980	5.8
28	Nebraska	1980	3.9

In [274…
```
states_below_df = duckdb.sql('''Select  C.Violent_Crime, S.*
                            FROM states_below_df AS S JOIN crime_df AS C
                            ON  S.Year = C.Year AND S.Area = C.State''').df()
states_below_df.head()
```

Out[274…

	Violent_Crime	Area	Year	UnemploymentRate
0	9168	Kansas	1980	4.4
1	352	North Dakota	1980	4.9
2	3512	Nebraska	1980	3.9
3	4425	Utah	1980	6.2
4	16355	Virginia	1980	5.2

In [275…

```
states_below_df = pd.get_dummies(states_below_df, prefix='', prefix_sep='',\
                                 drop_first=True, dtype=int, columns=['Area'])
states_below_df['Year'] = states_below_df['Year'].astype(int)
states_below_df.head()
```

Out[275…

	Violent_Crime	Year	UnemploymentRate	Iowa	Kansas	Minnesota	Nebraska	New Hampshire	North Dakota	South Dakota	Utah	Vermont	Virginia	Wyoming
0	9168	1980	4.4	0	1	0	0	0	0	0	0	0	0	0
1	352	1980	4.9	0	0	0	0	0	1	0	0	0	0	0
2	3512	1980	3.9	0	0	0	1	0	0	0	0	0	0	0
3	4425	1980	6.2	0	0	0	0	0	0	0	1	0	0	0
4	16355	1980	5.2	0	0	0	0	0	0	0	0	0	1	0

We apply a log transformation on the y axis (violent crime rates)

In [276…

```
states_below_df['log_crime'] = np.log(states_below_df['Violent_Crime'])

X_train_below, X_test_below, Y_train_below, Y_test_below = \
  train_test_split(states_below_df.iloc[:,1:-1],
                   states_below_df['log_crime'], test_size=.30, shuffle=False)
#print(Y_test)
```

```
X_train_below = sm.add_constant(X_train_below)
model_below = sm.OLS( Y_train_below,X_train_below).fit()

X_test_below = sm.add_constant(X_test_below)
preds_below = model_below.predict(X_test_below)
y_below_df = pd.DataFrame(data = {"Preds":preds_below, "Actual":Y_test_below})
y_below_df.head()

print( f"MAE for predicting violent crime rate using states below national average: \
      {mean_absolute_error(preds_below,Y_test_below)}")
print( f"RSME for predicting violent crime rate using states below national average: \
      {root_mean_squared_error(preds_below,Y_test_below)}")
```

```
MAE for predicting violent crime rate using states below national average:       0.22514088308828198
RSME for predicting violent crime rate using states below national average:       0.2960296142006233
```

In [277…

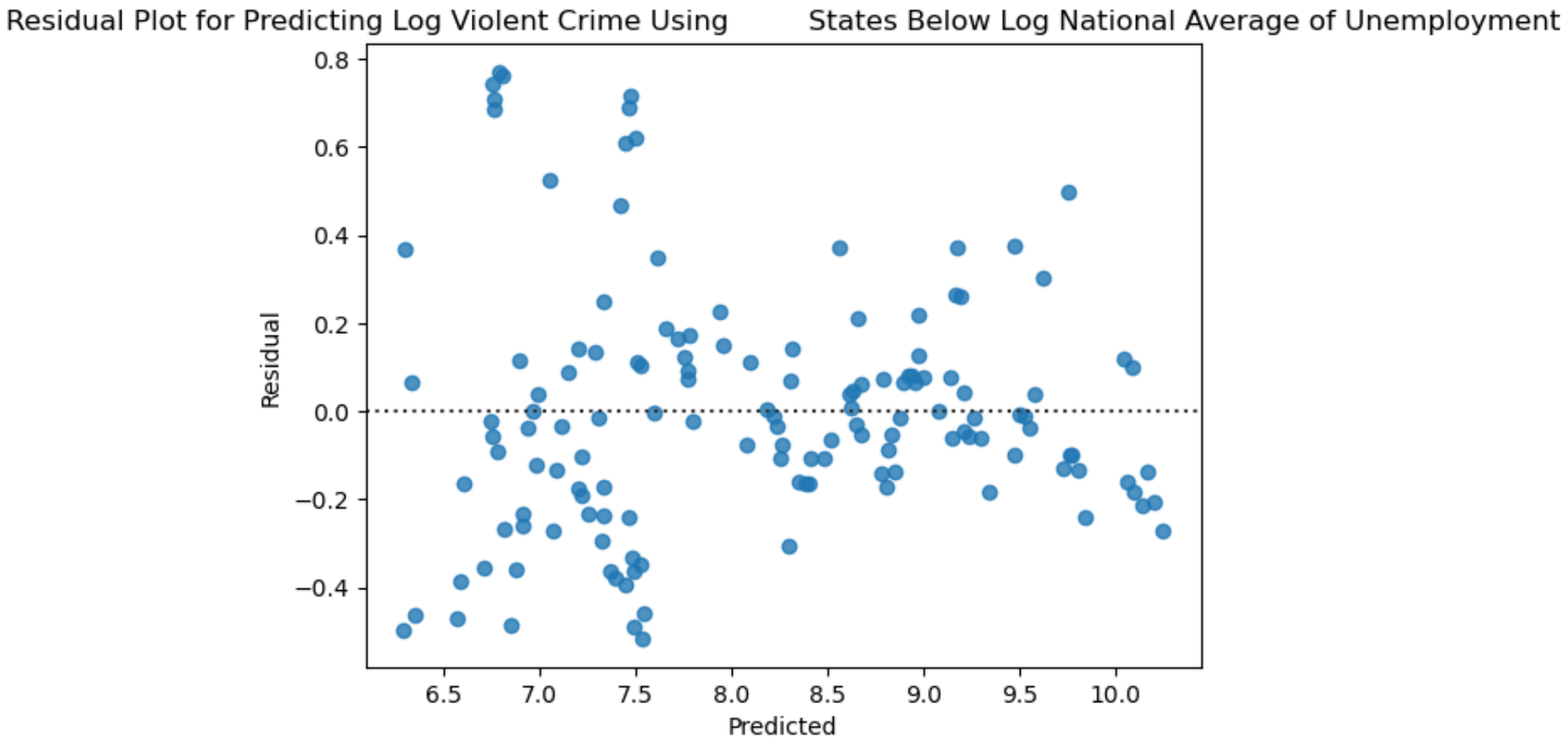
```
print(model_below.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              log_crime   R-squared:                       0.970
Model:                            OLS   Adj. R-squared:                  0.969
Method:                 Least Squares   F-statistic:                     781.0
Date:                Mon, 09 Dec 2024   Prob (F-statistic):          1.10e-229
Time:                        22:15:49   Log-Likelihood:                 63.834
No. Observations:                 327   AIC:                            -99.67
Df Residuals:                     313   BIC:                            -46.61
Df Model:                          13
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const            -18.2185      2.408     -7.566      0.000     -22.956     -13.481
Year               0.0132      0.001     11.016      0.000       0.011       0.016
UnemploymentRate  -0.0450      0.009     -4.889      0.000      -0.063      -0.027
Iowa               0.9423      0.056     16.812      0.000       0.832       1.053
Kansas             1.2214      0.056     21.951      0.000       1.112       1.331
```

```
Minnesota            1.4660      0.055     26.738      0.000       1.358       1.574
Nebraska             0.4690      0.057      8.243      0.000       0.357       0.581
New Hampshire       -0.6022      0.058    -10.336      0.000      -0.717      -0.488
North Dakota        -1.5855      0.056    -28.390      0.000      -1.695      -1.476
South Dakota        -0.8801      0.058    -15.289      0.000      -0.993      -0.767
Utah                 0.5912      0.056     10.522      0.000       0.481       0.702
Vermont             -1.4290      0.056    -25.301      0.000      -1.540      -1.318
Virginia             1.8752      0.056     33.768      0.000       1.766       1.985
Wyoming             -0.7824      0.059    -13.241      0.000      -0.899      -0.666
==============================================================================
Omnibus:                       77.367   Durbin-Watson:                   2.087
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              197.140
Skew:                           1.119   Prob(JB):                     1.55e-43
Kurtosis:                       6.076   Cond. No.                     4.27e+05
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.27e+05. This might indicate that there are
strong multicollinearity or other numerical problems.
```

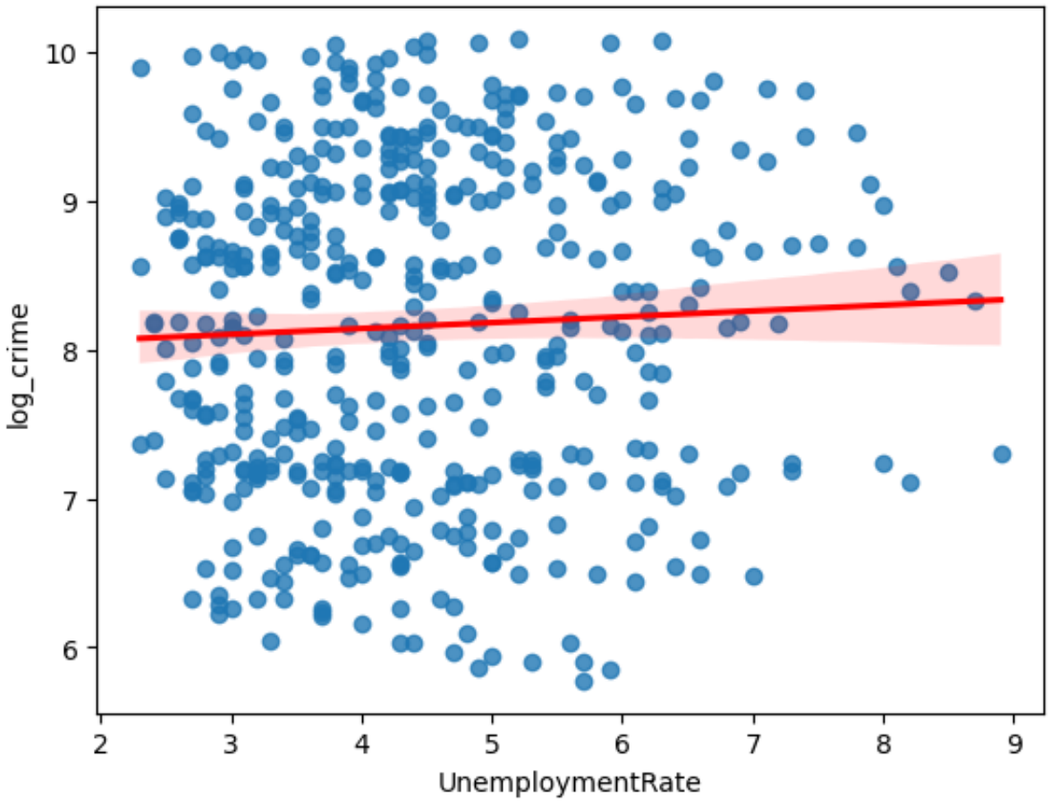
In [278…

```
ax = sns.residplot(data= y_below_df, x = 'Preds', y = 'Actual')
ax.set(xlabel = "Predicted", ylabel= "Residual",
       title="Residual Plot for Predicting Log Violent Crime Using \
        States Below Log National Average of Unemployment");
```



In [279…

```
sns.regplot(x = states_below_df['UnemploymentRate'], \
            y = states_below_df['log_crime'], line_kws = {"color":"red"});
```



Looking at this OLS regression, we observe that the p value for unemployment rate is significant (based on a 0.05 significance level). Moreover, we observe that the p value for the year is insignificant since it is greater than the 0.05 signficance level. This shows that the significant coefficient for unemployment rate, which is -0.0450. This shows that for every one unit increase in the unemployment rate, that corresponds to a e^(-0.045) increase in the amount of violent crime rate. Due to the p value being lower than the signfiicance level, we can reject the null hypothesis (that there is no relationship between unemployment rates among states lower than the national average and violent crime rates). There is a relationship that exists.

Hypothesis 3

At least one of income, population, unemployment, and poverty (+ year) have a relationship with the 4 categories of violent crime for each state regardless of the year.

In [280…

```
crime_job = duckdb.sql('''SELECT C.State, C.Year, UnemploymentRate,
                              C.Population,
                              FROM crime_df AS C
                              JOIN job_melt AS J ON
                              C.Year = J.Year AND J.Area = C.State
                              ''').df()
cj_poverty = duckdb.sql('''SELECT C.State, C.Year, C.Population,
                         UnemploymentRate, PovertyPercent
                              FROM crime_job AS C
                              JOIN poverty_concat AS P ON
                              C.Year = P.Year AND C.State = P.State
                              ''').df()
four_factors= duckdb.sql('''SELECT C.State, C.Year, C.Population,
                         UnemploymentRate, PovertyPercent,
                        Median_Income
                              FROM cj_poverty AS C
                              JOIN income_melt AS I ON
                              C.Year = I.Year AND C.State = I.State
                              ORDER BY C.Year ASC''').df()
four_factors.head() #2013 - 2018
```

Out[280…

	State	Year	Population	UnemploymentRate	PovertyPercent	Median_Income
0	Alabama	2013	4833996	7.2	16.7	47320.0
1	Alaska	2013	737259	7.0	10.9	72472.0
2	Arizona	2013	6634997	7.7	20.2	52611.0
3	Arkansas	2013	2958765	7.2	17.1	39376.0
4	California	2013	38431393	8.9	14.9	60794.0

3.1 Hypothesis

At least one of income, population, unemployment, and poverty (+ year) have a relationship with rape of each state regardless of the year.

MODEL FOR RAPE

In [281…

```
rape_fourfactors= duckdb.sql('''SELECT C.Rape, F.*
                                FROM crime_df AS C
                                JOIN four_factors AS F ON
                                C.Year = F.Year AND C.State = F.State
                                ORDER BY C.Year ASC''').df()
rape_fourfactors = pd.get_dummies(rape_fourfactors, prefix='', prefix_sep='',\
                                  drop_first=True, dtype=int, columns=['State'])
#print(rape_fourfactors)
X4_train_rape, X4_test_rape, Y4_train_rape, Y4_test_rape = \
  train_test_split(rape_fourfactors.iloc[:,1:], rape_fourfactors['Rape'],\
                   test_size=.30, shuffle=False)
X4_train_rape = sm.add_constant(X4_train_rape)
model_rape = sm.OLS(Y4_train_rape,X4_train_rape).fit()
print(model_rape.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   Rape   R-squared:                       0.991
Model:                            OLS   Adj. R-squared:                  0.988
Method:                 Least Squares   F-statistic:                     372.1
Date:                Mon, 09 Dec 2024   Prob (F-statistic):          3.06e-167
Time:                        22:15:49   Log-Likelihood:                -1693.3
No. Observations:                 245   AIC:                             3497.
Df Residuals:                     190   BIC:                             3689.
Df Model:                          54
Covariance Type:            nonrobust
==============================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
```

const	4.335e+04	6.7e+04	0.647	0.519	−8.89e+04	1.76e+05
Year	−25.2902	33.294	−0.760	0.448	−90.964	40.384
Population	0.0019	0.000	11.033	0.000	0.002	0.002
UnemploymentRate	−66.5489	37.403	−1.779	0.077	−140.328	7.231
PovertyPercent	13.4160	17.912	0.749	0.455	−21.916	48.748
Median_Income	0.0143	0.008	1.785	0.076	−0.002	0.030
Alaska	6405.1410	767.548	8.345	0.000	4891.131	7919.151
Arizona	−2599.2507	387.206	−6.713	0.000	−3363.026	−1835.475
Arkansas	3296.8095	378.494	8.710	0.000	2550.220	4043.399
California	−5.512e+04	5861.979	−9.403	0.000	−6.67e+04	−4.36e+04
Colorado	−183.7796	263.086	−0.699	0.486	−702.724	335.165
Connecticut	871.7753	359.199	2.427	0.016	163.245	1580.306
Delaware	5597.2542	718.121	7.794	0.000	4180.740	7013.768
Florida	−2.382e+04	2616.636	−9.103	0.000	−2.9e+04	−1.87e+04
Georgia	−9033.2682	929.201	−9.722	0.000	−1.09e+04	−7200.393
Hawaii	4612.3950	665.988	6.926	0.000	3298.715	5926.075
Idaho	4575.6542	604.977	7.563	0.000	3382.321	5768.988
Illinois	−1.243e+04	1397.072	−8.894	0.000	−1.52e+04	−9670.028
Indiana	−3161.3509	353.543	−8.942	0.000	−3858.725	−2463.977
Iowa	2155.4384	397.188	5.427	0.000	1371.975	2938.902
Kansas	2926.0533	408.584	7.161	0.000	2120.111	3731.996
Kentucky	389.3660	213.415	1.824	0.070	−31.600	810.332
Louisiana	−34.0387	196.832	−0.173	0.863	−422.296	354.219
Maine	4980.6499	651.382	7.646	0.000	3695.780	6265.520
Maryland	−2846.1972	332.964	−8.548	0.000	−3502.977	−2189.417
Massachusetts	−3792.8940	401.228	−9.453	0.000	−4584.327	−3001.461
Michigan	−4983.1334	892.946	−5.581	0.000	−6744.495	−3221.772
Minnesota	−1455.4546	274.739	−5.298	0.000	−1997.384	−913.525
Mississippi	2670.6673	373.461	7.151	0.000	1934.005	3407.329
Missouri	−2065.9710	285.154	−7.245	0.000	−2628.445	−1503.497
Montana	5612.9533	704.762	7.964	0.000	4222.790	7003.116
Nebraska	4177.3261	586.833	7.118	0.000	3019.782	5334.870
Nevada	3314.9961	397.850	8.332	0.000	2530.226	4099.766
New Hampshire	4874.4189	699.906	6.964	0.000	3493.834	6255.004
New Jersey	−8754.3212	739.720	−11.835	0.000	−1.02e+04	−7295.202
New Mexico	4723.9865	512.144	9.224	0.000	3713.768	5734.205
New York	−2.546e+04	2572.619	−9.895	0.000	−3.05e+04	−2.04e+04

```
North Carolina   -9267.8935    901.078    -10.285      0.000    -1.1e+04   -7490.492
North Dakota      5777.1106    777.520      7.430      0.000    4243.430    7310.791
Ohio             -9951.1700   1172.842     -8.485      0.000   -1.23e+04   -7637.706
Oklahoma          1760.4547    264.864      6.647      0.000    1238.004    2282.905
Oregon            1040.3729    259.245      4.013      0.000     529.005    1551.741
Pennsylvania     -1.328e+04   1377.351     -9.644      0.000    -1.6e+04   -1.06e+04
Rhode Island      5525.8703    694.921      7.952      0.000    4155.119    6896.621
South Carolina     294.2983    196.304      1.499      0.135     -92.916     681.513
South Dakota      5765.5423    750.301      7.684      0.000    4285.553    7245.531
Tennessee        -2697.1895    346.288     -7.789      0.000   -3380.253   -2014.126
Texas            -3.292e+04   3831.795     -8.592      0.000   -4.05e+04   -2.54e+04
Utah              2762.4683    436.118      6.334      0.000    1902.212    3622.724
Vermont           5816.2570    791.668      7.347      0.000    4254.670    7377.844
Virginia         -6525.1385    635.767    -10.263      0.000   -7779.207   -5271.070
Washington       -3763.9658    444.359     -8.471      0.000   -4640.476   -2887.456
West Virginia     4361.4114    551.649      7.906      0.000    3273.268    5449.555
Wisconsin        -2122.4036    259.031     -8.194      0.000   -2633.350   -1611.457
Wyoming           5985.3770    789.009      7.586      0.000    4429.034    7541.720
==============================================================================
Omnibus:                       85.238   Durbin-Watson:                   1.921
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1143.027
Skew:                          -0.945   Prob(JB):                    6.24e-249
Kurtosis:                      13.411   Cond. No.                     3.65e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.65e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

If we look exclusively at the four factors (+ year): Year, Population, Unemployment Rate, Poverty Percent, and Median Income, we can see that there is significant coefficient (based on a 0.05 signficance level) for population. Thus, we can reject the null hypothesis that none of the four factors (+ year) have a significant relationship with rape.

In [282…

```
#make predictions using rape model
X4_test_rape = sm.add_constant(X4_test_rape)
rape4_preds = model_rape.predict(X4_test_rape)
```

```
print( f"MAE for predicting rape using the four factors: \
      {mean_absolute_error(Y4_test_rape, rape4_preds)}")
print( f"RSME for predicting rape using the four factors: \
      {root_mean_squared_error(Y4_test_rape, rape4_preds)}")
```

```
MAE for predicting rape using the four factors:       311.5111383524392
RSME for predicting rape using the four factors:       513.7375186153721
```

3.2 Hypothesis

At least one of income, population, unemployment, and poverty (+ year) have a relationship with homicide of each state regardless of the year.

MODEL FOR HOMICIDE

In [283…

```
homicide_fourfactors= duckdb.sql('''SELECT C.Homicide, F.*
                                FROM crime_df AS C
                                JOIN four_factors AS F ON
                                C.Year = F.Year AND C.State = F.State
                                ORDER BY C.Year ASC''').df()
homicide_fourfactors = pd.get_dummies(homicide_fourfactors, prefix='', prefix_sep='',\
                                drop_first=True, dtype=int, columns=['State'])
#print(homicide_fourfactors)

X4_train_homi, X4_test_homi, Y4_train_homi, Y4_test_homi = \
  train_test_split(homicide_fourfactors.iloc[:,1:], homicide_fourfactors['Homicide'],\
                   test_size=.30, shuffle=False)
X4_train_homi = sm.add_constant(X4_train_homi)
model_homicide = sm.OLS(Y4_train_homi,X4_train_homi).fit()
print(model_homicide.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               Homicide   R-squared:                       0.991
Model:                           OLS   Adj. R-squared:                  0.989
Method:                Least Squares   F-statistic:                     402.7
```

```
Date:                Mon, 09 Dec 2024   Prob (F-statistic):          1.84e-170
Time:                        22:15:50   Log-Likelihood:                -1201.4
No. Observations:                 245   AIC:                             2513.
Df Residuals:                     190   BIC:                             2705.
Df Model:                          54
Covariance Type:            nonrobust
================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const          -2.392e+04   8998.286     -2.658      0.009   -4.17e+04   -6166.430
Year              11.6447      4.470      2.605      0.010       2.827      20.462
Population         0.0002   2.32e-05      7.072      0.000       0.000       0.000
UnemploymentRate   2.8045      5.022      0.558      0.577      -7.101      12.710
PovertyPercent    -2.9717      2.405     -1.236      0.218      -7.715       1.772
Median_Income      0.0007      0.001      0.694      0.489      -0.001       0.003
Alaska           345.2953    103.049      3.351      0.001     142.027     548.563
Arizona         -302.7253     51.986     -5.823      0.000    -405.268    -200.182
Arkansas         155.7467     50.816      3.065      0.002      55.511     255.982
California     -4122.4606    787.018     -5.238      0.000   -5674.875   -2570.046
Colorado        -280.7717     35.321     -7.949      0.000    -350.444    -211.099
Connecticut      -76.4630     48.225     -1.586      0.115    -171.589      18.663
Delaware         336.8930     96.414      3.494      0.001     146.715     527.071
Florida        -1809.8695    351.304     -5.152      0.000   -2502.827   -1116.912
Georgia         -597.4079    124.753     -4.789      0.000    -843.486    -351.330
Hawaii           226.1713     89.414      2.529      0.012      49.799     402.543
Idaho            210.1941     81.223      2.588      0.010      49.979     370.409
Illinois        -883.9677    187.568     -4.713      0.000   -1253.951    -513.984
Indiana         -268.4234     47.466     -5.655      0.000    -362.051    -174.795
Iowa             -15.4397     53.326     -0.290      0.772    -120.626      89.747
Kansas            87.7421     54.856      1.600      0.111     -20.462     195.946
Kentucky         -67.1190     28.653     -2.343      0.020    -123.637     -10.601
Louisiana        209.9811     26.426      7.946      0.000     157.854     262.108
Maine            250.0053     87.453      2.859      0.005      77.501     422.509
Maryland        -118.7489     44.703     -2.656      0.009    -206.927     -30.571
Massachusetts   -541.0515     53.868    -10.044      0.000    -647.308    -434.795
Michigan        -584.8956    119.885     -4.879      0.000    -821.373    -348.419
Minnesota       -356.7134     36.886     -9.671      0.000    -429.472    -283.955
```

Mississippi	222.3707	50.140	4.435	0.000	123.468	321.273
Missouri	-116.4016	38.284	-3.040	0.003	-191.918	-40.885
Montana	313.0892	94.620	3.309	0.001	126.449	499.730
Nebraska	186.7957	78.787	2.371	0.019	31.386	342.205
Nevada	158.8729	53.415	2.974	0.003	53.511	264.235
New Hampshire	218.7536	93.968	2.328	0.021	33.399	404.108
New Jersey	-658.8656	99.313	-6.634	0.000	-854.764	-462.967
New Mexico	247.8324	68.759	3.604	0.000	112.202	383.462
New York	-2163.8243	345.395	-6.265	0.000	-2845.125	-1482.523
North Carolina	-653.1976	120.977	-5.399	0.000	-891.828	-414.567
North Dakota	337.2439	104.388	3.231	0.001	131.335	543.153
Ohio	-927.3333	157.463	-5.889	0.000	-1237.934	-616.732
Oklahoma	32.6034	35.560	0.917	0.360	-37.540	102.747
Oregon	-117.9779	34.806	-3.390	0.001	-186.633	-49.323
Pennsylvania	-1024.5008	184.920	-5.540	0.000	-1389.262	-659.740
Rhode Island	289.9932	93.299	3.108	0.002	105.959	474.028
South Carolina	5.3643	26.355	0.204	0.839	-46.622	57.351
South Dakota	334.8478	100.734	3.324	0.001	136.147	533.548
Tennessee	-226.1157	46.492	-4.864	0.000	-317.822	-134.409
Texas	-2730.0887	514.449	-5.307	0.000	-3744.854	-1715.323
Utah	3.6746	58.552	0.063	0.950	-111.822	119.171
Vermont	341.2756	106.288	3.211	0.002	131.620	550.931
Virginia	-563.0608	85.357	-6.597	0.000	-731.430	-394.692
Washington	-534.8952	59.659	-8.966	0.000	-652.574	-417.217
West Virginia	236.5015	74.063	3.193	0.002	90.410	382.593
Wisconsin	-315.5816	34.777	-9.074	0.000	-384.180	-246.983
Wyoming	357.1918	105.931	3.372	0.001	148.240	566.143

Omnibus:	158.883	Durbin-Watson:	1.954
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2692.105
Skew:	2.220	Prob(JB):	0.00
Kurtosis:	18.620	Cond. No.	3.65e+10

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.65e+10. This might indicate that there are

```
strong multicollinearity or other numerical problems.
```

If we look exclusively at the four factors (+ year): Year, Population, Unemployment Rate, Poverty Percent, and Median Income, we can see that there are significant coefficients (based on a 0.05 signficance level) for year and population. Thus, we can reject the null hypothesis that none of the four factors (+ year) have a significant relationship with homicide.

In [284…
```
#make predictions using homicide model
X4_test_homi = sm.add_constant(X4_test_homi)
homicide4_preds = model_homicide.predict(X4_test_homi)
print( f"MAE for predicting homicide using the four factors: \
      {mean_absolute_error( Y4_test_homi,homicide4_preds)}")
print( f"RSME for predicting homicide using the four factors: \
      {root_mean_squared_error(Y4_test_homi, homicide4_preds)}")
```

```
MAE for predicting homicide using the four factors:       56.87366280713931
RSME for predicting homicide using the four factors:       71.35952922187123
```

3.3 Hypothesis

At least one of income, population, unemployment, and poverty (+ year) have a relationship with aggravated assault of each state regardless of the year.

MODEL FOR AGGRAVATED ASSAULT

In [285…
```
assault_fourfactors= duckdb.sql('''SELECT C.Aggravated_Assault, F.*
                                FROM crime_df AS C
                                JOIN four_factors AS F ON
                                C.Year = F.Year AND C.State = F.State
                                ORDER BY C.Year ASC''').df()
assault_fourfactors = pd.get_dummies(assault_fourfactors, prefix='', prefix_sep='',\
                                     drop_first=True, dtype=int, columns=['State'])
#print(assault_fourfactors)

X4_train_assault, X4_test_assault, Y4_train_assault, Y4_test_assault = \
  train_test_split(assault_fourfactors.iloc[:,1:], assault_fourfactors['Aggravated_Assault'],\
```

```
                    test_size=.30, shuffle=False)
X4_train_assault = sm.add_constant(X4_train_assault)
model_assault = sm.OLS(Y4_train_assault,X4_train_assault).fit()
print(model_assault.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Aggravated_Assault   R-squared:                       0.997
Model:                            OLS   Adj. R-squared:                  0.997
Method:                 Least Squares   F-statistic:                     1345.
Date:                Mon, 09 Dec 2024   Prob (F-statistic):          6.70e-220
Time:                        22:15:50   Log-Likelihood:                -2022.1
No. Observations:                 245   AIC:                             4154.
Df Residuals:                     190   BIC:                             4347.
Df Model:                          54
Covariance Type:            nonrobust
================================================================================================
                       coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------------------------
const            -1.722e+05   2.56e+05     -0.671      0.503   -6.78e+05    3.34e+05
Year                80.2801    127.388      0.630      0.529    -170.997     331.557
Population           0.0051      0.001      7.773      0.000       0.004       0.006
UnemploymentRate  -139.3850    143.110     -0.974      0.331    -421.674     142.904
PovertyPercent     -16.1359     68.534     -0.235      0.814    -151.321     119.050
Median_Income        0.0273      0.031      0.891      0.374      -0.033       0.088
Alaska             9334.4523   2936.728      3.179      0.002    3541.673    1.51e+04
Arizona           -6055.1126   1481.497     -4.087      0.000   -8977.408   -3132.817
Arkansas           5812.5060   1448.161      4.014      0.000    2955.967    8669.045
California        -9.434e+04   2.24e+04     -4.206      0.000   -1.39e+05   -5.01e+04
Colorado          -7614.2644   1006.598     -7.564      0.000   -9599.808   -5628.721
Connecticut       -4228.4634   1374.339     -3.077      0.002   -6939.385   -1517.542
Delaware           8133.9796   2747.615      2.960      0.003    2714.231    1.36e+04
Florida           -3.093e+04      1e+04     -3.089      0.002   -5.07e+04   -1.12e+04
Georgia           -1.925e+04   3555.231     -5.415      0.000   -2.63e+04   -1.22e+04
Hawaii             4072.1046   2548.148      1.598      0.112    -954.188    9098.398
Idaho              4171.6853   2314.712      1.802      0.073    -394.149    8737.519
Illinois          -2.692e+04   5345.362     -5.036      0.000   -3.75e+04   -1.64e+04
```

Indiana	-8854.8466	1352.698	-6.546	0.000	-1.15e+04	-6186.612
Iowa	38.1079	1519.687	0.025	0.980	-2959.517	3035.732
Kansas	2522.9792	1563.289	1.614	0.108	-560.652	5606.611
Kentucky	-7616.2686	816.549	-9.327	0.000	-9226.935	-6005.602
Louisiana	3630.9420	753.104	4.821	0.000	2145.424	5116.460
Maine	4114.9044	2492.265	1.651	0.100	-801.159	9030.968
Maryland	-5376.8858	1273.958	-4.221	0.000	-7889.803	-2863.969
Massachusetts	-6673.7866	1535.143	-4.347	0.000	-9701.900	-3645.673
Michigan	-1.278e+04	3416.517	-3.740	0.000	-1.95e+04	-6039.740
Minnesota	-1.164e+04	1051.182	-11.075	0.000	-1.37e+04	-9568.097
Mississippi	16.8066	1428.905	0.012	0.991	-2801.749	2835.362
Missouri	-1613.3075	1091.031	-1.479	0.141	-3765.398	538.783
Montana	7298.9035	2696.502	2.707	0.007	1979.978	1.26e+04
Nebraska	3081.3503	2245.291	1.372	0.172	-1347.549	7510.250
Nevada	6661.1389	1522.219	4.376	0.000	3658.518	9663.760
New Hampshire	3978.4353	2677.923	1.486	0.139	-1303.844	9260.715
New Jersey	-2.449e+04	2830.256	-8.654	0.000	-3.01e+04	-1.89e+04
New Mexico	9283.0885	1959.522	4.737	0.000	5417.876	1.31e+04
New York	-4.586e+04	9843.141	-4.659	0.000	-6.53e+04	-2.64e+04
North Carolina	-1.845e+04	3447.629	-5.352	0.000	-2.53e+04	-1.16e+04
North Dakota	7066.1227	2974.883	2.375	0.019	1198.081	1.29e+04
Ohio	-3.416e+04	4487.432	-7.612	0.000	-4.3e+04	-2.53e+04
Oklahoma	1796.4282	1013.399	1.773	0.078	-202.529	3795.386
Oregon	-4197.2791	991.901	-4.232	0.000	-6153.832	-2240.727
Pennsylvania	-3.265e+04	5269.906	-6.195	0.000	-4.3e+04	-2.23e+04
Rhode Island	6224.5998	2658.849	2.341	0.020	979.946	1.15e+04
South Carolina	3453.0884	751.080	4.597	0.000	1971.561	4934.616
South Dakota	7541.9023	2870.738	2.627	0.009	1879.291	1.32e+04
Tennessee	6136.3445	1324.940	4.631	0.000	3522.863	8749.826
Texas	-6.154e+04	1.47e+04	-4.198	0.000	-9.05e+04	-3.26e+04
Utah	-1860.5653	1668.640	-1.115	0.266	-5152.005	1430.875
Vermont	6835.3362	3029.013	2.257	0.025	860.523	1.28e+04
Virginia	-2.372e+04	2432.520	-9.752	0.000	-2.85e+04	-1.89e+04
Washington	-1.454e+04	1700.169	-8.553	0.000	-1.79e+04	-1.12e+04
West Virginia	5651.0149	2110.674	2.677	0.008	1487.652	9814.378
Wisconsin	-9780.8091	991.084	-9.869	0.000	-1.17e+04	-7825.868
Wyoming	7654.2885	3018.843	2.536	0.012	1699.536	1.36e+04

```
==============================================================================
Omnibus:                       78.448   Durbin-Watson:                   1.975
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2481.618
Skew:                           0.478   Prob(JB):                         0.00
Kurtosis:                      18.562   Cond. No.                     3.65e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.65e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

If we look exclusively at the four factors (+ year): Year, Population, Unemployment Rate, Poverty Percent, and Median Income, we can see that there is significant coefficient (based on a 0.05 signficance level) for population. Thus, we can reject the null hypothesis that none of the four factors (+ year) have a significant relationship with aggravated assault.

In [286…
```
#make predictions using assault model
X4_test_assault = sm.add_constant(X4_test_assault)
assault4_preds = model_assault.predict(X4_test_assault)
print( f"MAE for predicting aggravated assault using the four factors: \
      {mean_absolute_error( Y4_test_assault,assault4_preds)}")
print( f"RSME for predicting aggravated assault using the four factors: \
      {root_mean_squared_error(Y4_test_assault,assault4_preds)}")
```

```
MAE for predicting aggravated assault using the four factors:       1381.4995644397807
RSME for predicting aggravated assault using the four factors:       2261.3542980047837
```

3.4 Hypothesis

At least one of income, population, unemployment, and poverty (+ year) have a relationship with robbery of each state regardless of the year.

MODEL FOR ROBBERY

In [287…
```
robbery_fourfactors= duckdb.sql('''SELECT C.Robbery, F.*
                                FROM crime_df AS C
```

```
                              JOIN four_factors AS F ON
                              C.Year = F.Year AND C.State = F.State
                              ORDER BY C.Year ASC''').df()
robbery_fourfactors = pd.get_dummies(robbery_fourfactors , prefix='', prefix_sep='',\
                                     drop_first=True, dtype=int, columns=['State'])
#print(robbery_fourfactors)

X4_train_robbery, X4_test_robbery, Y4_train_robbery, Y4_test_robbery = \
  train_test_split(robbery_fourfactors .iloc[:,1:], robbery_fourfactors ['Robbery'],\
                   test_size=.30, shuffle=False)
X4_train_robbery = sm.add_constant(X4_train_robbery)
model_robbery = sm.OLS(Y4_train_robbery,X4_train_robbery).fit()
print(model_robbery.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Robbery   R-squared:                       0.996
Model:                            OLS   Adj. R-squared:                  0.995
Method:                 Least Squares   F-statistic:                     946.8
Date:                Mon, 09 Dec 2024   Prob (F-statistic):          1.79e-205
Time:                        22:15:50   Log-Likelihood:                -1904.5
No. Observations:                 245   AIC:                             3919.
Df Residuals:                     190   BIC:                             4112.
Df Model:                          54
Covariance Type:            nonrobust
=====================================================================================
                        coef    std err          t      P>|t|      [0.025      0.975]
-------------------------------------------------------------------------------------
const             -2.958e+05   1.59e+05     -1.864      0.064   -6.09e+05    1.72e+04
Year                148.2717     78.846      1.881      0.062      -7.255     303.798
Population            0.0002      0.000      0.569      0.570      -0.001       0.001
UnemploymentRate    388.3461     88.577      4.384      0.000     213.625     563.067
PovertyPercent      -88.8426     42.419     -2.094      0.038    -172.515      -5.170
Median_Income        -0.0087      0.019     -0.459      0.647      -0.046       0.029
Alaska            -3361.2988   1817.672     -1.849      0.066   -6946.708     224.110
Arizona            1721.6003    916.965      1.877      0.062     -87.138    3530.339
Arkansas          -1760.2523    896.331     -1.964      0.051   -3528.291       7.787
```

California	3.984e+04	1.39e+04	2.870	0.005	1.25e+04	6.72e+04
Colorado	-1359.0718	623.028	-2.181	0.030	-2588.013	-130.131
Connecticut	-1529.0730	850.639	-1.798	0.074	-3206.983	148.837
Delaware	-2402.7752	1700.621	-1.413	0.159	-5757.299	951.748
Florida	1.375e+04	6196.596	2.219	0.028	1526.481	2.6e+04
Georgia	6588.3377	2200.491	2.994	0.003	2247.808	1.09e+04
Hawaii	-2263.7243	1577.162	-1.435	0.153	-5374.720	847.272
Idaho	-3384.8025	1432.678	-2.363	0.019	-6210.800	-558.805
Illinois	9667.8009	3308.482	2.922	0.004	3141.727	1.62e+04
Indiana	1932.2752	837.245	2.308	0.022	280.786	3583.764
Iowa	-2716.7049	940.602	-2.888	0.004	-4572.068	-861.342
Kansas	-2187.0110	967.589	-2.260	0.025	-4095.608	-278.415
Kentucky	-1051.6153	505.399	-2.081	0.039	-2048.529	-54.702
Louisiana	1341.8201	466.129	2.879	0.004	422.367	2261.273
Maine	-3326.7918	1542.574	-2.157	0.032	-6369.562	-284.022
Maryland	5144.1452	788.509	6.524	0.000	3588.789	6699.501
Massachusetts	1082.0673	950.169	1.139	0.256	-792.167	2956.301
Michigan	2440.8107	2114.634	1.154	0.250	-1730.365	6611.986
Minnesota	-634.7861	650.623	-0.976	0.330	-1918.159	648.587
Mississippi	-1735.5432	884.413	-1.962	0.051	-3480.073	8.986
Missouri	953.8770	675.288	1.413	0.159	-378.147	2285.901
Montana	-3108.0007	1668.985	-1.862	0.064	-6400.120	184.119
Nebraska	-2113.9860	1389.710	-1.521	0.130	-4855.229	627.257
Nevada	1002.7354	942.169	1.064	0.289	-855.720	2861.191
New Hampshire	-2870.3487	1657.486	-1.732	0.085	-6139.786	399.089
New Jersey	4656.7584	1751.771	2.658	0.009	1201.340	8112.177
New Mexico	-1541.6802	1212.835	-1.271	0.205	-3934.032	850.672
New York	1.676e+04	6092.357	2.751	0.007	4740.337	2.88e+04
North Carolina	3205.9711	2133.891	1.502	0.135	-1003.189	7415.131
North Dakota	-2527.8083	1841.288	-1.373	0.171	-6159.800	1104.183
Ohio	7174.5154	2777.471	2.583	0.011	1695.876	1.27e+04
Oklahoma	-824.2846	627.238	-1.314	0.190	-2061.528	412.959
Oregon	-2298.8665	613.932	-3.744	0.000	-3509.864	-1087.869
Pennsylvania	6999.7072	3261.779	2.146	0.033	565.756	1.34e+04
Rhode Island	-3785.2743	1645.680	-2.300	0.023	-7031.424	-539.125
South Carolina	-811.5589	464.877	-1.746	0.082	-1728.542	105.424
South Dakota	-2619.6005	1776.827	-1.474	0.142	-6124.442	885.241

```
Tennessee         2454.4598    820.064      2.993      0.003     836.860    4072.060
Texas             2.265e+04   9074.281      2.496      0.013    4748.686    4.05e+04
Utah             -2377.1412   1032.796     -2.302      0.022   -4414.360    -339.923
Vermont          -3041.0169   1874.791     -1.622      0.106   -6739.095     657.061
Virginia          -654.9319   1505.595     -0.435      0.664   -3624.760    2314.896
Washington         395.6018   1052.310      0.376      0.707   -1680.109    2471.313
West Virginia    -3197.1326   1306.390     -2.447      0.015   -5774.023    -620.242
Wisconsin          117.5978    613.426      0.192      0.848   -1092.402    1327.598
Wyoming          -3210.3994   1868.496     -1.718      0.087   -6896.060     475.261
==============================================================================
Omnibus:                       96.064   Durbin-Watson:                   2.090
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1458.502
Skew:                          -1.090   Prob(JB):                         0.00
Kurtosis:                      14.752   Cond. No.                     3.65e+10
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.65e+10. This might indicate that there are
strong multicollinearity or other numerical problems.
```

If we look exclusively at the four factors (+ year): Year, Population, Unemployment Rate, Poverty Percent, and Median Income, we can see that there are significant coefficients (based on a 0.05 signficance level) for unemployment rate and population. Thus, we can reject the null hypothesis that none of the four factors (+ year) have a significant relationship with robbery.

In [288…

```
#make predictions using robbery model
X4_test_robbery = sm.add_constant(X4_test_robbery)
robbery4_preds = model_robbery.predict(X4_test_robbery)
print( f"MAE for predicting robbery using the four factors: \
      {mean_absolute_error( Y4_test_robbery,robbery4_preds)}")
print( f"RSME for predicting robbery using the four factors: \
      {root_mean_squared_error(Y4_test_robbery,robbery4_preds)}")
```

```
MAE for predicting robbery using the four factors:       922.7968416893488
RSME for predicting robbery using the four factors:       1557.5517439986706
```

Evaluation of Significance

Hypothesis 1: Income is a better predictor for robbery than burglary across the states.

Although the main method we used to comapare income being a predictor of robbery vs burglary is train test split and comparing relative MAE and RMSE values, we also ran OLS regression tests for both income ~ robbery and income ~ burglary to test some sub-hypotheses. We also included the different states and year in our OLS regression because we believe that its input into the model were important, though due to the numerous number of states, we only evaluated the significance of the variables we were interested in (income).

Essentially, we investigated sub-hypotheses within this hypothesis, based on the OLS regression. If the p value is less than the signficance level of 0.05, we can reject the null hyptohesis that there is no relationship between income and (either burlary or robbery). However, if the p value is greater than the significance level of 0.05, we fail to reject the null hypothesis.

Again, this does not directly answer our main hypothesis 1 — we aimed to answer this broader question not only by using OLS regressions/t tests and evaluation of significance but also by using train test split.

In [289…
```
# Income vs Robbery OLS regression
print(f"Coefficient: {inc_rob_model.params['Median_Income']}")
print(f"p-value: {inc_rob_model.pvalues['Median_Income']}")
```

```
Coefficient: -4.4220649094349295e-06
p-value: 0.1056196023555338
```

Since median income's p value of 0.719 is not less than the signficance level of 0.05, we fail to reject the null hypothesis that there is no relationship between the median income and robbery.

In [290…
```
# Income vs Burglary
print(f"Coefficient: {inc_burg_model.params['Median_Income']}")
print(f"p-value: {inc_burg_model.pvalues['Median_Income']}")
```

```
Coefficient: -1.0284426843920932e-05
p-value: 2.4211526999809017e-08
```

Since median income's p value of 0.015 is less than the signficance level of 0.05, we can reject the null hypothesis that there is no relationship between the median income and burglary.

Therefore, all in all, as it relates to the main hypothesis 1, we can note that the OLS regression for median income vs. robbery yielded an insignificant coefficient for median income, while the OLS regression for median income vs. burglary yielded a significant coefficient for median income. Again, signficance tells us that there is a relationship between the two variables.

Hypothesis 2: There is a relationship among states that are above the U.S. unemployment average over the years 1980 to 2018 in predicting violent crime rates, and there is a relationship among states that are below the U.S. unemploymnent average over the years 1980 to 2018 in predicting violent crime rates.

Similar to the previous hypothesis, we not only used t-tests and OLS regressions to test our hypothesis, but also used train test split. Moreover,

In [291…
```
# OLS regression for states above the average unemployment rate
print(f"Coefficient: {model_above.params['log_unemployment']}")
print(f"p-value: {model_above.pvalues['log_unemployment']}")
```

```
Coefficient: -0.17841104425957266
p-value: 1.337397300273809e-14
```

Instead of looking at the entire summary, I simply looked at the coefficient and p value for unemployment, since it appeared towards the end of the summary, so it didn't appear in the truncated summary.

Based on the small p value of 1.33e-14, which is less than the significance level of 0.05, we can reject the null hypothesis that there is no relationship between unemployment rates and violent crime rates (for states that are above the average unemployment rate).

In [292…
```
print(f"Coefficient: {model_below.params['UnemploymentRate']}")
print(f"p-value: {model_below.pvalues['UnemploymentRate']}")
```

```
Coefficient: -0.04495800997152745
p-value: 1.62373260511353e-06
```

Based on the p value of 1.62e-06, which is less than the significance level of 0.05, we can reject the null hypothesis that there is no relationship between unemployment rates and violent crime rates (for states that are below the average unemployment rate).

Hypothesis 3: At least one of income, population, unemployment, and poverty (+ year) have a relationship with the 4 categories of violent crime for each state regardless of the year.

This broader hypothesis had the most sub-hypotheses that we tested using t-tests to observe the significance of whether or not a relationship existed between a category and our 4 factors (+ year).

In [293…
```
# Rape

print(f"p-value for Year: {model_rape.pvalues['Year']}")
print(f"p-value for Population: {model_rape.pvalues['Population']}")
print(f"p-value for Unemployment Rate: {model_rape.pvalues['UnemploymentRate']}")
print(f"p-value for Poverty: {model_rape.pvalues['PovertyPercent']}")
print(f"p-value for Income: {model_rape.pvalues['Median_Income']}")
```

```
p-value for Year: 0.44844052893252007
p-value for Population: 3.4218610490530756e-22
p-value for Unemployment Rate: 0.07680291203642022
p-value for Poverty: 0.4547912781839283
p-value for Income: 0.07590774900443466
```

Since the p-value for population is less than 0.05, we can reject the null hypothesis that there is no relationship between rape and one of income, population, unemployment, poverty, and year.

In [294…
```
# Homicide

print(f"p-value for Year: {model_homicide.pvalues['Year']}")
print(f"p-value for Population: {model_homicide.pvalues['Population']}")
print(f"p-value for Unemployment Rate: {model_homicide.pvalues['UnemploymentRate']}")
print(f"p-value for Poverty: {model_homicide.pvalues['PovertyPercent']}")
print(f"p-value for Income: {model_homicide.pvalues['Median_Income']}")
```

```
p-value for Year: 0.009912848847460944
p-value for Population: 2.847509756470276e-11
p-value for Unemployment Rate: 0.5771784664955727
p-value for Poverty: 0.21809624514341858
p-value for Income: 0.48880910691965695
```

Since the p-values for population and year are less than 0.05, we can reject the null hypothesis that there is no relationship between homicide and one of income, population, unemployment, poverty, and year.

In [295…
```
# Aggravated Assault

print(f"p-value for Year: {model_assault.pvalues['Year']}")
print(f"p-value for Population: {model_assault.pvalues['Population']}")
print(f"p-value for Unemployment Rate: {model_assault.pvalues['UnemploymentRate']}")
print(f"p-value for Poverty: {model_assault.pvalues['PovertyPercent']}")
print(f"p-value for Income: {model_assault.pvalues['Median_Income']}")
```

```
p-value for Year: 0.5293211718213582
p-value for Population: 4.710400576733012e-13
p-value for Unemployment Rate: 0.33130953358010895
p-value for Poverty: 0.814118317823932
p-value for Income: 0.3741392021884128
```

Since the p-value for population is less than 0.05, we can reject the null hypothesis that there is no relationship between aggravated assault and one of income, population, unemployment, poverty, and year.

In [296…
```
# Robbery

print(f"p-value for Year: {model_robbery.pvalues['Year']}")
print(f"p-value for Population: {model_robbery.pvalues['Population']}")
print(f"p-value for Unemployment Rate: {model_robbery.pvalues['UnemploymentRate']}")
print(f"p-value for Poverty: {model_robbery.pvalues['PovertyPercent']}")
print(f"p-value for Income: {model_robbery.pvalues['Median_Income']}")
```

```
p-value for Year: 0.061568006176528314
p-value for Population: 0.5697681336755915
p-value for Unemployment Rate: 1.9243531015311026e-05
p-value for Poverty: 0.03754921210965942
p-value for Income: 0.6467755312700574
```

Since the p-values for unemployment rate and poverty are less than 0.05, we can reject the null hypothesis that there is no relationship between robbery and one of income, population, unemployment, poverty, and year.

Conclusion

Hypothesis 1: Income is a better predictor for robbery than burglary across the states.

After creating a regression model comparing income vs robbery and income vs burglary, we can see the differences in the RMSE and MAE metrics between the two models.

Income vs Robbery

- RSME for predicting robbery using income: 0.3845879612087546
- MAE for predicting robbery using income: 0.2995251380922235

Income vs Burglary

- RSME for predicting burglary using income: 0.25217229608909597
- MAE for predicting burglary using income: 0.1894950245621107

We observe that income vs burglary produces slightly better results across the board since it has a lower MAE and RMSE value, though the difference is quite small. Moreover, besides the train test split that created this model, we also ran a hypothesis test that overall aimed to demonstrate whether or not income vs robbery had a statistically signficant relationship (vs. not having a relationship) and whether or not income vs burglary had a statistically significant relationship (vs. not having a relationship). We found that there is only a statistically significant relationship between income and burglary, rejecting the null hypothesis that there is no relationship between those two variables.

As we noticed along the way, our hypothesis 1 is quite broad, which is why we also broke it down to hypotheis tests regarding whether or not there is a statistically significant

relationship between income vs robbery or burglary. Moreover, we also went a step further by creating two models that utilized income as a predictor for robbery and burglary. In both the statistical significance test and the predictive regression model, we have found that income vs burglary has a more accurate model and income vs burglary has a statistically significant relationship (while income vs robbery doesn't).

Hypothesis 2: There is a relationship among states that are above the U.S. unemployment average over the years 1980 to 2018 in predicting violent crime rates, and there is a relationship among states that are below the U.S. unemploymnent average over the years 1980 to 2018 in predicting violent crime rates.

In conclusion, when looking at the average unemployment rate over the years vs average violent crime rates over the years, we got a significant coefficient, showing a positive relationship between the two. This allowed us to reject the null hypothesis (that there was no relationship between average unemployment rate over the years vs average violent crime rates over the years), as we observed a significant p value. From there, to actually test our specific hypothesis, we looked at two groups: states that have an unemployment rate that is lower than the national average and states that have an unemployment rate that is greater than the national average. With this information, we created two models that can predict the violent crime rates based on the category of unemployment rate the states fall into (either below or above the national average).

After log transformations, we observed the following about our predictive models:

States Below National Average:

- RMSE: 0.29602961420061813
- MAE: 0.22514088308826197

States Above National Average:

- RSME: 0.27352759131431054
- MAE: 0.21924054627991002

With the relatively low and similar RMSE and MAE values across the states that are above and below the national average, we can conclude that both of our models are quite accurate at predicting violent crime rates.

Moreover, with OLS regression models for both of the two groups, we also performed signficance testing showing that for states that are below the national average, the coefficient for unemployment rate vs violent crime rates was significant, and for states that are above the national average, the coefficient for unemployment rate vs violent crime rates was

also signficant. So, for both of these groups, we should reject the null hypothesis (that there is no relationship between unemployment rate and violent crime rates). There is a relationship for both models (above and below the national average) between unemployment rate and violent crime rates.

The broader implication is that unemployment rate over the years is a good predictor of average violent crime rates, regardless of whether or not a state is above or below the average unemployment rate across the United States.

Hypothesis 3: At least one of income, population, unemployment, and poverty (+ year) have a relationship with the 4 categories of violent crime for each state regardless of the year.

This was by far our most broad hypothesis, which is why we broke it down into many sub-hypotheses accoridng to the 4 categories of violent crime for each state: rape, homicide, aggravated assault, and robbery. Each of these sub hypotheses were essentially testing whether or not there is a statistically significant relationship between at least one of the following factors we have been exploring throughout our project (income, population, unemployment, poverty, year) and the category of violent crime.

In conclusion, we found that when creating an OLS regression including all of these factors and the category of interest, each sub-hypothesis showed at least one of the factors had a statistically significant coefficient. This shows that, when all these factors are included in the regression, at least one of them will have a statistically signficiant coefficient with each of the different categories of violent crime.

Here is a breakdown of what was factor was statistically significant for each:

Rape:

- Population

Homicide:

- Population
- Year

Aggravated Assault:

- Population

Robbery:

- Unemployment Rate
- Poverty

The broader implication of this is that these findings highlight that there is a complex relationship between socioeconomic and demographic factors and different categories of violent crime. With this data, the federal or state agencies can enact change and create stratgies about crime prevention. More sepcifically, the significance of population across multiple categories suggests that densely populated areas may require more interventions, patrol, and community resources. Moreover, economic factors like unemployment rate and poverty suggest the importance of addressing economic disparities through support programs and emphasis on job creation and growth. Lastly, the significance of year suggests the importance of time in creating change, especially with societal trends and policies that can influence violent crime rates.

Limitations

Data set related

- Our data is not encompassing of all the kinds of counties/regions within a state. For our factors (income, poverty, unemployment and population), they are measured state by state, there is no consideration for the urban, suburban or rural areas, so a crime rate that is measured of 5% may not be applicable to all regions of the state. This contributes to the residual we have that are +/- a couple thousand (refer to hypothesis #1).

- Our analysis has different time periods. We tried to find datasets as close to the 1979 to 2018 time range as possible, but data sets like income only starts at 1984-2018 while unemployment starts at 1980-2018. We would wanted to have looked at the most recent year, 2023. The crime data frame had that range, but the factors we were analyzing did not. We would have had even more data to train, especially given that our data is generalized for across each state.

- The original income data set had a margin of error for each state's income each year. We removed it to create consistency across our data by just taking the average income. However, the +/- could have been useful in seeing how it would predict vioelnt crime rates if we were more conservative in our research

- There is collinarity amongst the coefficients, but is quite common social economic regressions as many social and economic variables tend to be interconnected and correlated with each other, making it likely for multiple independent variables in a model to show significant relationships with each other (Professor Koenecke)

Analysis related

- Did not use Regression Discontinuity Design (RDD), which could provide a stronger framework for causal inference. If we knew how to properly implement RDD, we could use it to look at factors like national average unemployment rate as a threshold to compare violent crime rates in states just above and below this cutoff. By focusing on observations near the threshold, we could isolate the causal impact of unemployment on violent crime, offering more insights into the relationships between unemployment (ando ther factors of similar analysis) versus violent crime rates

- Did not create heatmaps, where we would have plotted the different predictions of violent crime for a given factor like income vs burglary and income vs robbery (hypothesis #1). We would have shown the actual versus predicted versions of the maps as a visual in observing any geographic patterns, like if better predictions were made on the west versus east coast. We did not know how to plot data respective to a state and for only the U.S. and not the entire world

- There are definitely more than 4 factors that affect crimes, so including more related beyond personal socio-economic status would be beneficial. For example, political changes over the years as we mentioned in the overarching question the dynamics that politics play into crime and the enforcement of it.

- We could have looked at other crimes outside of the violent crime scope to better understand the relationship crimes have with the four factors we chose, such as exploring property crimes like larceny or motor vehicle theft. Including a broader range of crime types would provide a more comprehensive view of how socioeconomic factors impact crime as a whole, rather than limiting our understanding to violent crime alone

- The most accurate dataset available for crime was from the FBI government website, but it did not include the pandemic years, so we did not have the most up to date data that we could compare with recent current events.

- As suggested in our feedback for Phase 3 by our TAs, if we knew how to run random forests or gradient boosting, we could have improved predictive accuracy by capturing non-linear relationships and interactions between factors

Citations

Forming our hypotheses were based on some online research about the topic of violent crimes (and its subcategories) and the factors that do and do not affect them

Source regarding the minimal effect of disparity of income on crime

- Jawadi, Fredj, et al. “Does higher unemployment lead to greater criminality? revisiting the debate over the business cycle.” Journal of Economic Behavior & Organization, vol. 182, Feb. 2021, pp. 448–471, https://doi.org/10.1016/j.jebo.2019.03.025.

Source discussing possible effect of unemployment on violent and non-violent crimes

- Pazzona, Matteo. “Revisiting the income inequality-crime puzzle.” World Development, vol. 176, Apr. 2024, p. 106520, https://doi.org/10.1016/j.worlddev.2023.106520.

Source discussing the effects of factors such as income, unemployment, etc. on the amount of violent crime

- Wilkins, Natalie J., et al. “Societal determinants of violent death: The extent to which social, economic, and structural characteristics explain differences in violence across Australia, Canada, and the United States.” SSM - Population Health, vol. 8, Aug. 2019, p. 100431, https://doi.org/10.1016/j.ssmph.2019.100431.

General research:

- Brenan, Megan. “Economy Most Important Issue to 2024 Presidential Vote.” Gallup.Com, Gallup, 15 Nov. 2024, news.gallup.com/poll/651719/economy-important-issue-2024-presidential-vote.aspx.
- Jones, Jeffrey M. “More Americans See U.S. Crime Problem as Serious.” Gallup.Com, Gallup, 16 Oct. 2024, news.gallup.com/poll/544442/americans-crime-problem-serious.aspx.
- Westcott, Diane N, and Robert W Bednarzik. Employment and Unemployment: A Report on 1980, Monthly Labor Review , Feb. 1981, www.bls.gov/opub/mlr/1981/02/art1full.pdf.

Unemployment dataset

- “Annual Unemployment Rates by State.” Annual Unemployment Rates by State | Iowa Community Indicators Program, Iowa State University, Apr. 2019, www.icip.iastate.edu/tables/employment/unemployment-states.
- Raw data link: https://github.com/w0ahnder/INFO2950_project/blob/97d2b8a7d6583dd30dd254fc0b3b3449b99d1338/phase4/data/emp_table.csv

Crime dataset

- FBI Crime Data Explorer, cde.ucr.cjis.gov/LATEST/webapp/#/pages/downloads. Accessed 9 Dec. 2024.
- Raw data link: https://github.com/w0ahnder/INFO2950_project/blob/97d2b8a7d6583dd30dd254fc0b3b3449b99d1338/data/crime.csv

Income dataset

- United States Census Bureau, https://www2.census.gov/programs-surveys/cps/tables/time-series/historical-income-households/h08.xls
- Raw data link:

https://github.com/w0ahnder/INFO2950_project/blob/97d2b8a7d6583dd30dd254fc0b3b3449b99d1338/phase4/data/income.csv

Poverty dataset

- U.S. Census Bureau, "Historical Poverty Tables: People and Families - 1959 to 2023" Table 18 https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html
- Raw data link:

https://github.com/w0ahnder/INFO2950_project/blob/main/phase4_datacleaning/data/poverty.csv