Jordan Gil-Baptiste

Prof. Singh

Spring 2025

Assignment #3

Report

a) Problem Selected

The problem I have selected and am looking to analyze focuses on crime prevention through state-level risk analysis. The primary objective is to identify socioeconomic, community, and psychological factors that correlate with higher violent crime rates across U.S. states. This includes variables such as the percentage of single-parent households, substance use disorder prevalence, unemployment rates, and mental health conditions like depression. By examining how these factors relate to violent crime per 100,000 residents, this analysis can inform more targeted, data-driven prevention strategies that benefit both communities and public safety agencies.

b) Model Explained

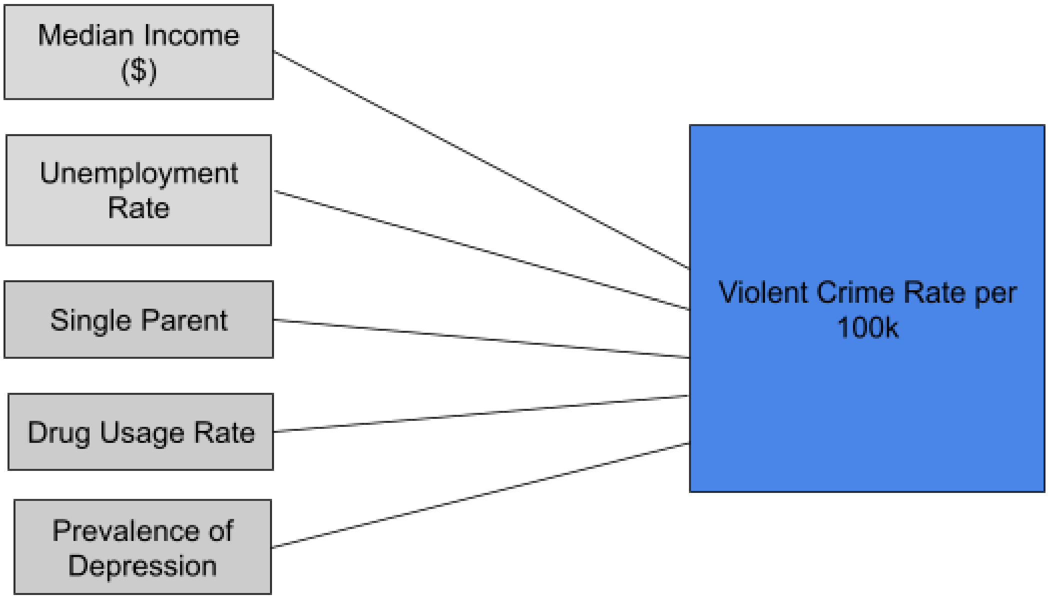
This study will utilize a multi-linear regression model to analyze how various socioeconomic, environmental, and psychological factors contribute to state-level violent crime rates. With multiple independent variables and a single dependent variable (violent crime per 100,000 residents) this model is appropriate for identifying significant predictors and assessing the strength of their relationship to crime rates across states.

Regression Equation:

$(\text{Violent Crime Rate per 100k})_i = B_0 + (\text{Median Income})_i B_1 + (\text{UnemploymentRate})_i B_2 +$

$(\text{Single-Parent Households})_i B_3 + (\text{Drug Usage})_i B_4 + (\text{Depression Rate})_i B_{65}$

Diagram:



c) **Variables**

Independent Variables (x):

i) **Socioeconomic Factors:**

1) Median Household Income ($- US Dollars)
2) Unemployment Rate (% of population)

ii) **Community/Environment Factors**

1) Percntages of Single-Parent Households (% of State Population)
2) Drug Usage Rate (% of population)

iii) **Psychological Factors:**

1) Prevalence of Depression (% of Community)

Dependent Variable (y)

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

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

d) <u>Data Source</u>

Independent Variables (x):

- **Median Household Income by State:**

  historical-income-households.html

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

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

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

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

- **Unemployment Rates**

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

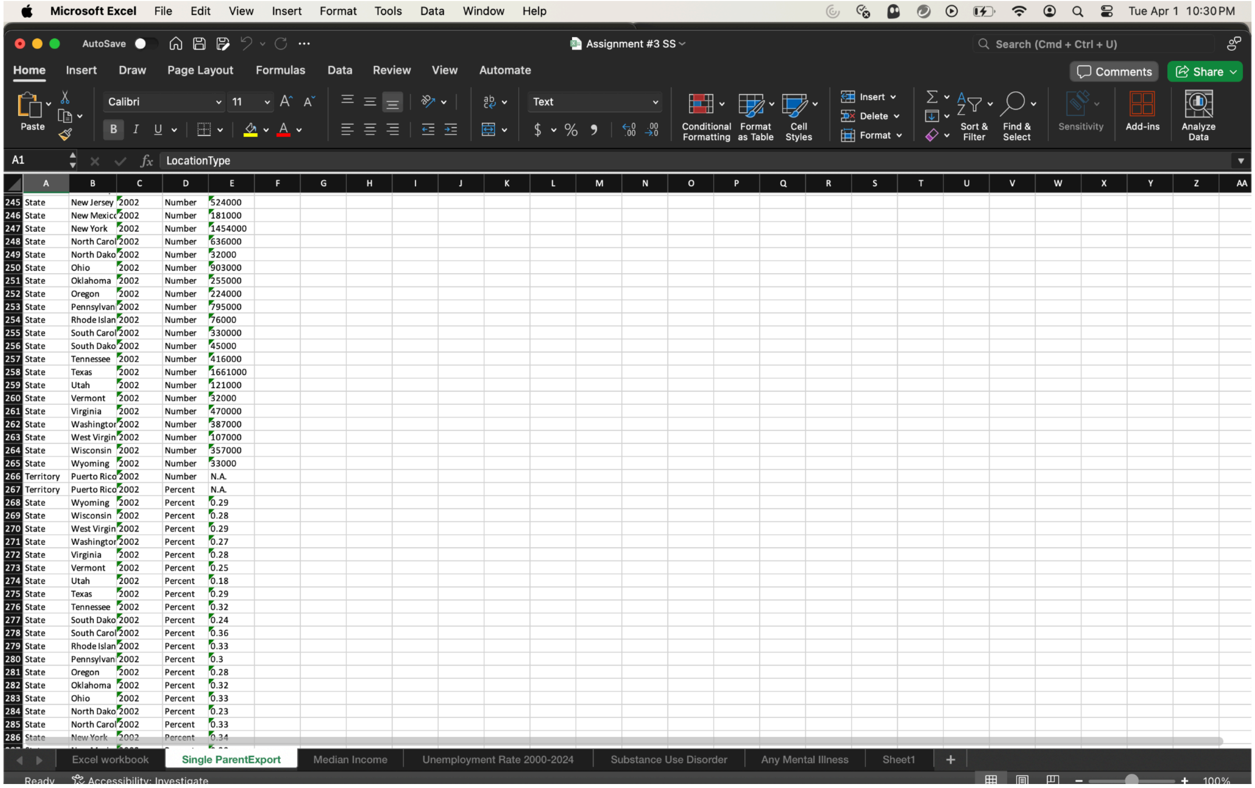
Dependent Variable (Y)

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

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

e) Curation Process

To curate the data analysis process, the data must be cleaned and structured in a clear, organized format for easy interpretation and tabulation. When selecting a file format for data extraction, several options were available but I ultimately went with Excel as it best suited my needs for the analysis. As most of my sources came from government sources or other government agencies missing data or inputs wasn't necessarily a concern. However, changing or deleting variables was an issue as some datasets included percentages and numbers (photo listed below). As I am looking to focus on percentages as most of my data is in percentages making sure to clean the data is crucial. Additionally, as these sets provided data from a long range of years I wanted to focus on more recent years. I'm looking to focus on a range from 2022-2023 this was the latest data for most sets at the moment. If needed taking into account a multi-year approach can also be an alternative as it can help pick up on trends.

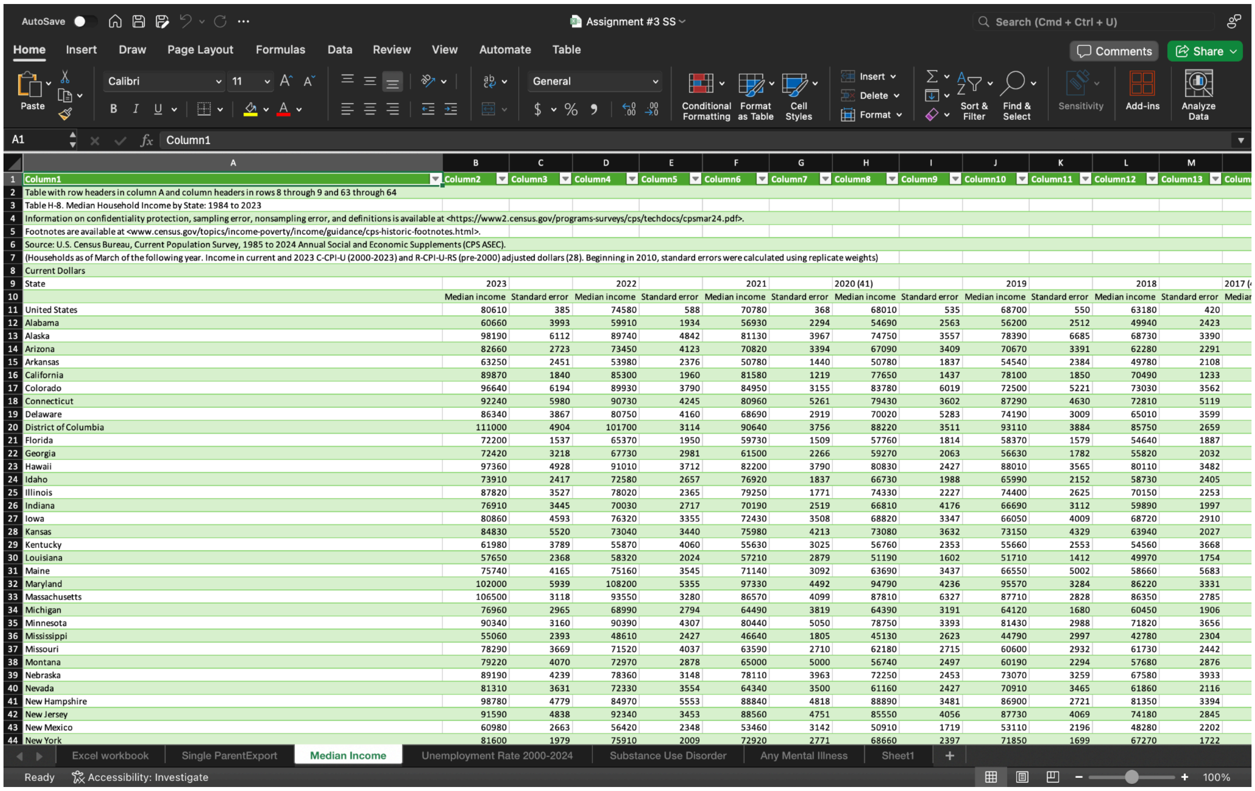


A1 LocationType

	A	B	C	D	E
245	State	New Jersey	2002	Number	524000
246	State	New Mexicc	2002	Number	181000
247	State	New York	2002	Number	1454000
248	State	North Carol	2002	Number	636000
249	State	North Dako	2002	Number	32000
250	State	Ohio	2002	Number	903000
251	State	Oklahoma	2002	Number	255000
252	State	Oregon	2002	Number	224000
253	State	Pennsylvani	2002	Number	795000
254	State	Rhode Islan	2002	Number	76000
255	State	South Carol	2002	Number	330000
256	State	South Dako	2002	Number	45000
257	State	Tennessee	2002	Number	416000
258	State	Texas	2002	Number	1661000
259	State	Utah	2002	Number	121000
260	State	Vermont	2002	Number	32000
261	State	Virginia	2002	Number	470000
262	State	Washingtor	2002	Number	387000
263	State	West Virgin	2002	Number	107000
264	State	Wisconsin	2002	Number	357000
265	State	Wyoming	2002	Number	33000
266	Territory	Puerto Rico	2002	Number	N.A.
267	Territory	Puerto Rico	2002	Percent	N.A.
268	State	Wyoming	2002	Percent	0.29
269	State	Wisconsin	2002	Percent	0.28
270	State	West Virgin	2002	Percent	0.29
271	State	Washingtor	2002	Percent	0.27
272	State	Virginia	2002	Percent	0.28
273	State	Vermont	2002	Percent	0.25
274	State	Utah	2002	Percent	0.18
275	State	Texas	2002	Percent	0.29
276	State	Tennessee	2002	Percent	0.32
277	State	South Dako	2002	Percent	0.24
278	State	South Carol	2002	Percent	0.36
279	State	Rhode Islan	2002	Percent	0.33
280	State	Pennsylvani	2002	Percent	0.3
281	State	Oregon	2002	Percent	0.28
282	State	Oklahoma	2002	Percent	0.32
283	State	Ohio	2002	Percent	0.33
284	State	North Dako	2002	Percent	0.23
285	State	North Carol	2002	Percent	0.33
286	State	New York	2002	Percent	0.34

(Raw Data Single Parent Households)

As seen above although these years aren't the ones I'm using it provides a perspective as to not only how much data is in this set but also as stated how it includes both percentages and numbers. This is displaying single-parent households and using numbers wouldn't do justice as the number of single-parent households will naturally be higher in more populous areas. Using percentages allows for a fair comparison across different states. The data is also not listed year by year for example 2001, 2002, 2003, etc. and there are times when it jumps from 2009 to 2018 and then back to 2011 so ensuring the years are in chronological order is highly beneficial for analysis. Also, deleting the years I will not be using is crucial as that is not what I'm looking to observe.



	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	Column1	Column2	Column3	Column4	Column5	Column6	Column7	Column8	Column9	Column10	Column11	Column12	Column13	Colum
2	Table with row headers in column A and column headers in rows 8 through 9 and 63 through 64													
3	Table H-8. Median Household Income by State: 1984 to 2023													
4	Information on confidentiality protection, sampling error, nonsampling error, and definitions is available at <https://www2.census.gov/programs-surveys/cps/techdocs/cpsmar24.pdf>.													
5	Footnotes are available at <www.census.gov/topics/income-poverty/income/guidance/cps-historic-footnotes.html>.													
6	Source: U.S. Census Bureau, Current Population Survey, 1985 to 2024 Annual Social and Economic Supplements (CPS ASEC).													
7	(Households as of March of the following year. Income in current and 2023 C-CPI-U (2000-2023) and R-CPI-U-RS (pre-2000) adjusted dollars (28). Beginning in 2010, standard errors were calculated using replicate weights)													
8	Current Dollars													
9	State	2023		2022		2021		2020 (41)		2019		2018		2017 (
10		Median income	Standard error	Median income	Standard error	Median income	Standard error	Median income	Standard error	Median income	Standard error	Median income	Standard error	Mediar
11	United States	80610	385	74580	588	70780	368	68010	535	68700	550	63180	420	
12	Alabama	60660	3993	59910	1934	56930	2294	54690	2563	56200	2512	49940	2423	
13	Alaska	98190	6112	89740	4842	81130	3967	74750	3557	78390	6685	68730	3390	
14	Arizona	82660	2723	73450	4123	70820	3394	67090	3409	70670	3391	62280	2291	
15	Arkansas	63250	2451	53980	2376	50780	1440	50780	1837	54540	2384	49780	2108	
16	California	89870	1840	85300	1960	81580	1219	77650	1437	78100	1850	70490	1233	
17	Colorado	96640	6194	89930	3790	84950	3155	83780	6019	72500	5221	73030	3562	
18	Connecticut	92240	5980	90730	4245	80960	5261	79430	3602	87290	4630	72810	5119	
19	Delaware	86340	3867	80750	4160	68690	2919	70020	5283	74190	3009	65010	3599	
20	District of Columbia	111000	4904	101700	3114	90640	3756	88220	3511	93110	3884	85750	2659	
21	Florida	72200	1537	65370	1950	59730	1509	57760	1814	58370	1579	54640	1887	
22	Georgia	72420	3218	67730	2981	61500	2266	59270	2063	56630	1782	55820	2032	
23	Hawaii	97360	4928	91010	3712	82200	3790	80830	2427	88010	3565	80110	3482	
24	Idaho	73910	2417	72580	2657	76920	1837	66730	1988	65990	2152	58730	2405	
25	Illinois	87820	3527	78020	2365	79250	1771	74330	2227	74400	2625	70150	2253	
26	Indiana	76910	3445	70030	2717	70190	2519	66810	4176	66690	3112	59890	1997	
27	Iowa	80860	4593	76320	3355	72430	3508	68820	3347	66050	4009	68720	2910	
28	Kansas	84830	5520	73040	3440	75980	4213	73080	3632	73150	4329	63940	2027	
29	Kentucky	61980	3789	55870	4060	55630	3025	56760	2353	55660	2553	54560	3668	
30	Louisiana	57650	2368	58320	2024	57210	2879	51190	1602	51710	1412	49970	1754	
31	Maine	75740	4165	75160	3545	71140	3092	63690	3437	66550	5002	58660	5683	
32	Maryland	102000	5939	108200	5355	97330	4492	94790	4236	95570	3284	86220	3331	
33	Massachusetts	106500	3118	93550	3280	86570	4099	87810	6327	87710	2828	86350	2785	
34	Michigan	76960	2965	68990	2794	64490	3819	64390	3191	64120	1680	60450	1906	
35	Minnesota	90340	3160	90390	4307	80440	5050	78750	3393	81430	2988	71820	3656	
36	Mississippi	55060	2393	48610	2427	46640	1805	45130	2623	44790	2997	42780	2304	
37	Missouri	78290	3669	71520	4037	63590	2710	62180	2715	60600	2932	61730	2442	
38	Montana	79220	4070	72970	2878	65000	5000	56740	2497	60190	2294	57680	2876	
39	Nebraska	89190	4239	78360	3148	78110	3963	72250	2453	73070	3259	67580	3933	
40	Nevada	81310	3631	72330	3554	64340	3500	61160	2427	70910	3465	61860	2116	
41	New Hampshire	98780	4779	84970	5553	88840	4818	88890	3481	86900	2721	81350	3394	
42	New Jersey	91590	4838	92340	3453	88560	4751	85550	4056	87730	4069	74180	2845	
43	New Mexico	60980	2663	56420	2348	53460	3142	50910	1719	53110	2196	48280	2202	
44	New York	81600	1979	75910	2009	72920	2771	68660	2397	71850	1699	67270	1722	

(Raw Data of Median Household Income)

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

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

Excel

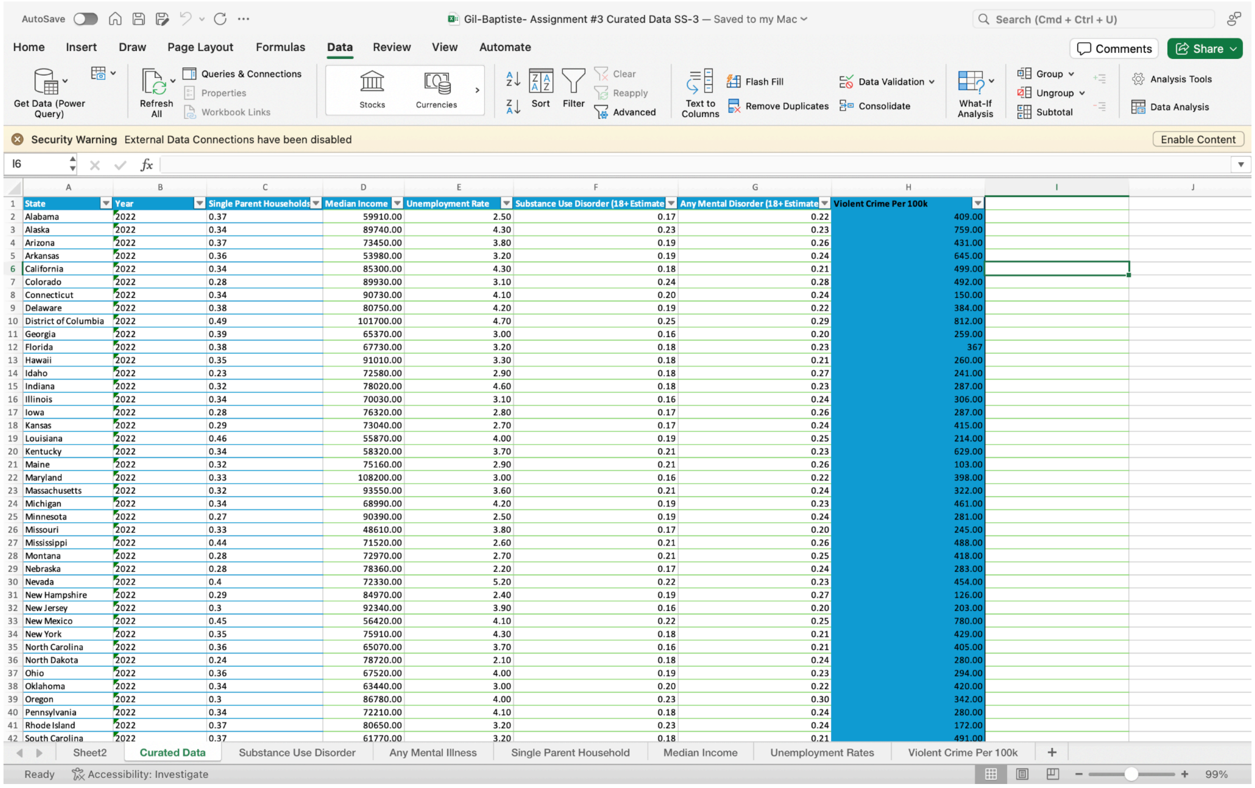
- 1st Excel Sheet: Curated data with proper years, proper data needed, number in percentages etc.
- 2nd Sheet Substance Abuse Disorders 2022-2023, 3rd Sheet: Any Mentall illness 2022-2023, 4th Sheet: Single Parent Households 2000-2023, 5th Sheet: Median Income 1984-2023, 6th Sheet: Unemployment Rates 2012-2024

Summary of Data Curation:

1. Identify the problem
2. Find necessary data
3. Import data
4. Make changes where needed and organize
5. Ensure Consistency

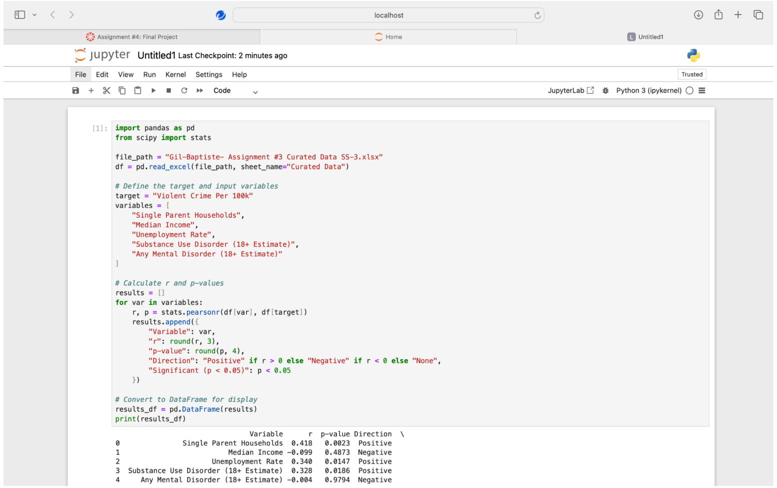
Assignment 4 Section

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



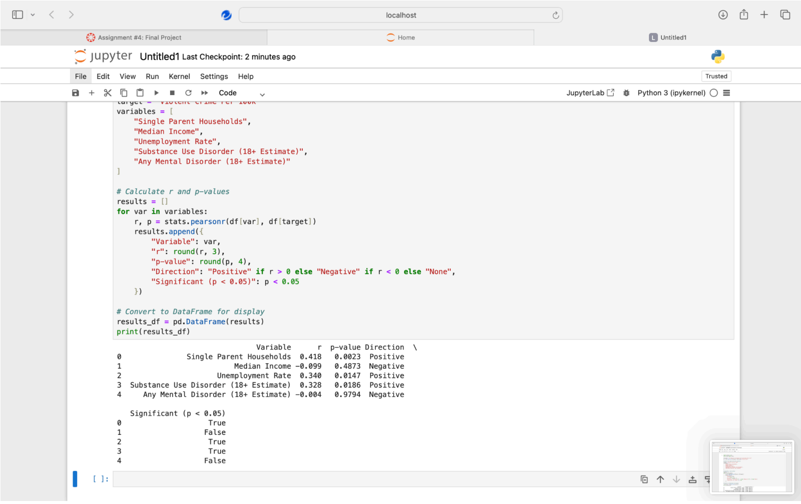
	A	B	C	D	E	F	G	H
1	State	Year	Single Parent Households	Median Income	Unemployment Rate	Substance Use Disorder (18+ Estimate	Any Mental Disorder (18+ Estimate	Violent Crime Per 100k
2	Alabama	2022	0.37	59910.00	2.50	0.17	0.22	409.00
3	Alaska	2022	0.34	89740.00	4.30	0.23	0.23	759.00
4	Arizona	2022	0.37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

f) Correlations found between different input variables



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

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

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

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

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

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


localhost
Assignment #4: Final Project
Home
Untitled1
jupyter Untitled1 Last Checkpoint: 2 minutes ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
[ ]:

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

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

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

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

   Significant (p < 0.05)
0                    True
1                   False
2                    True
3                    True
4                   False
```

a. Correlation score for the following variables:

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

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

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

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

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

b. Significance (p-value):

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

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

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

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

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

c. Direction of correlation

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

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

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

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

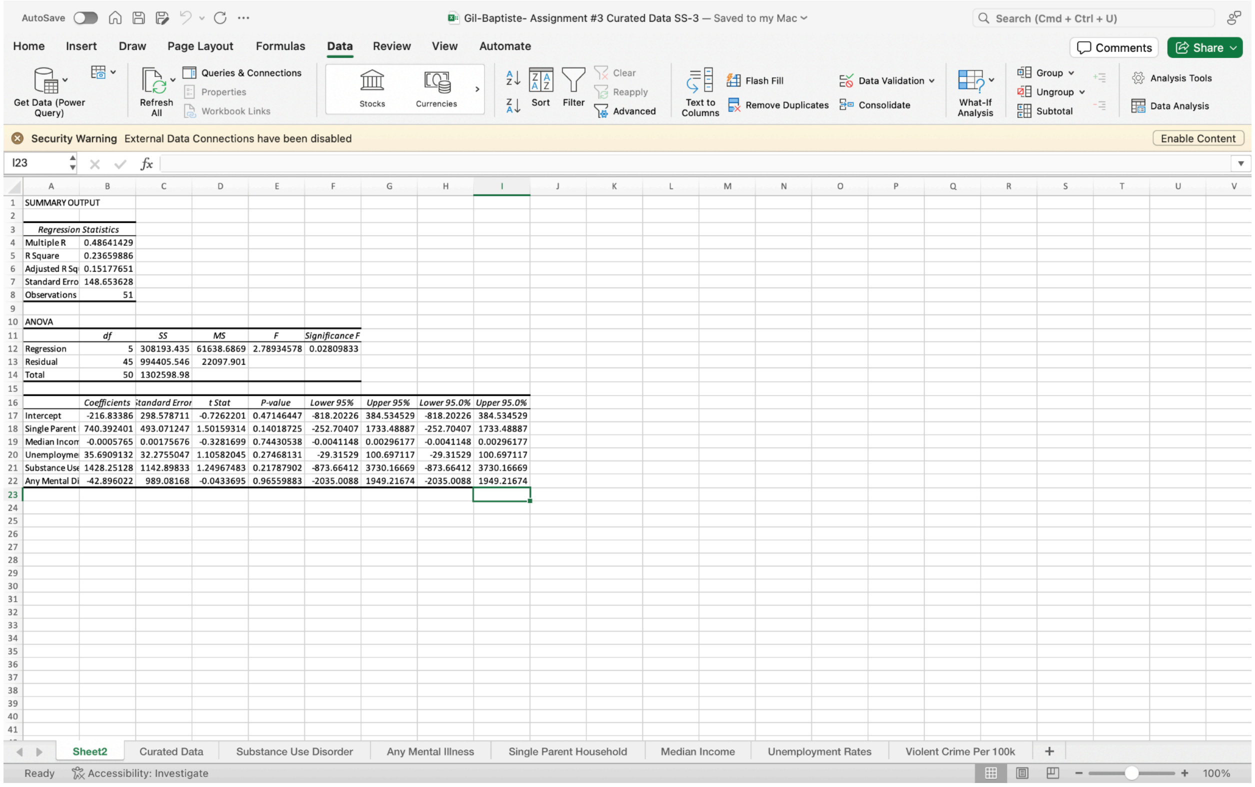
v. Violet crime per 100k & Any mental disorder -> negative

d. Interpretation of Findings

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

g) **<u>Multi Liner Regression</u>**

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



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.48641429
R Square	0.23659886
Adjusted R Sq	0.15177651
Standard Erro	148.653628
Observations	51

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	5	308193.435	61638.6869	2.78934578	0.02809833
Residual	45	994405.546	22097.901		
Total	50	1302598.98			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-216.83386	298.578711	-0.7262201	0.47146447	-818.20226	384.534529	-818.20226	384.534529
Single Parent	740.392401	493.071247	1.50159314	0.14018725	-252.70407	1733.48887	-252.70407	1733.48887
Median Incom	-0.0005765	0.00175676	-0.3281699	0.74430538	-0.0041148	0.00296177	-0.0041148	0.00296177
Unemployme	35.6909132	32.2755047	1.10582045	0.27468131	-29.31529	100.697117	-29.31529	100.697117
Substance Use	1428.25128	1142.89833	1.24967483	0.21787902	-873.66412	3730.16669	-873.66412	3730.16669
Any Mental Di	-42.896022	989.08168	-0.0433695	0.96559883	-2035.0088	1949.21674	-2035.0088	1949.21674

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

$$\text{Violent Crime Rate} = -216.83 + 740.39(\text{Single Parent}) - 0.00058(\text{Median Income}) + 35.69(\text{Unemployment}) + 1428.25(\text{Substance Use}) - 42.90(\text{Any Mental Disorder})$$

a. Percentage of variance explaine model

   i. R = .2366 (23.66%)

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

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

b. Signifcance of Variables in the Model

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

ii.