A Statistical Analysis of Sexual Crime Trends in Illinois (2021–2024)

Client: Marcie Sheridan
Survivor Resource Center

University of Illinois Urbana-Champaign
STAT 427: Statistical Consulting

Joe Li, Weijia He, Xiaoshan Huang, Jinhong Zhu, Kexin Wang

May 7, 2025

Abstract

This project investigates county-level sexual crime patterns in Illinois between 2021 and 2024 and explores their associations with key socioeconomic and health-related variables. We integrated crime reports, mental and physical health indicators, and demographic data to construct a comprehensive dataset. Multiple modeling approaches were explored, including time series analysis, K-means clustering, Generalized Least Squares (GLS), and ultimately Generalized Estimating Equations (GEE), which proved most effective in capturing cross-county trends. Our final model identified excessive drinking and smoking behaviors, unemployment, and shortage of mental health providers as key risk factors associated with sexual crime. To support interpretation and client engagement, we developed an interactive Shiny app to visualize sexual crime patterns and predictors comparisons by county and year. The findings offer valuable insights to inform resource allocation for the Survival Resource Center (SRC).

Table of Contents

1. Introduction

Sexual violence remains a critical public health and safety concern across communities in the United States. Understanding the socioeconomic and behavioral factors that influence crime patterns is essential for designing effective interventions for SRC. SRC is a non-profit organization in Illinois serving victims of sexual violence across Vermilion, Edgar, and Clark counties. SRC provides free and confidential services including counseling, legal and medical support, case management, and a 24-hour hotline. SRC also engages in community education and prevention programs with the long-term goal of ending sexual violence. The center serves as a vital resource for individuals and communities affected by sexual assault.

This study focuses on county-level sexual crime trends in Illinois from 2021 to 2024 and aims to identify the underlying predictors associated with fluctuations in reported incidents. We integrate data from state crime reports and public health databases to build a panel dataset. This includes monthly crime counts derived from Group A offenses in the National Incident-Based Reporting System (NIBRS) database and health indicators from the County Health Rankings dataset. Variables such as average monthly mental distress days, insurance coverage rate, household income, and demographics are incorporated to form a profile of each county we chose. The subset of counties from Illinois is chosen based on their similarity in demographics and socioeconomic factors to Champaign.

After conducting exploratory data analysis, we combined time series modeling methods for a deeper understanding of sexual crimes in Illinois. This allows us to provide insights that support the work of SRC.

2. Problem Statement and Objectives

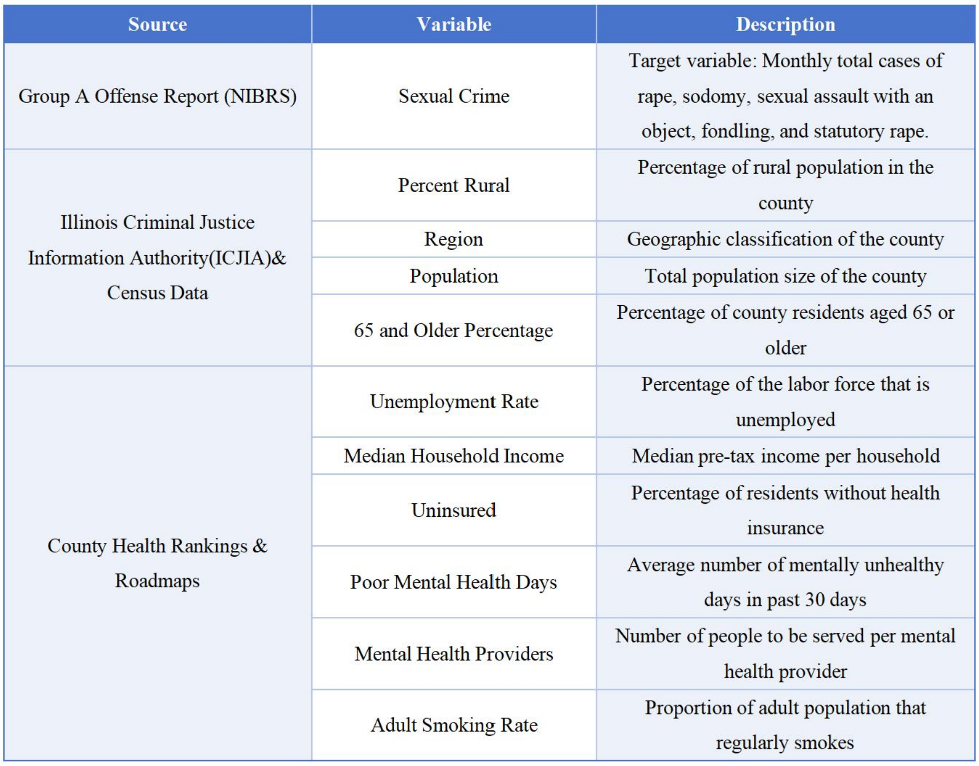
Although the government continues to promote social security, the high sexual crime rates in some areas of Illinois remain a serious problem. Sexual crimes are often influenced by a complex combination of socioeconomic and health factors. However, it is unclear which factors are most correlated with sexual crimes and to what extent they are associated. So the overall goal of this project is to assist the SRC in identifying key factors that lead to sexual crime varying with time and region.

More specifically, this project aims to quantify the strength and direction of associations between sexual crime and key socioeconomic and health factors. We seek to uncover both spatial and temporal patterns in sexual crime across counties in Illinois. Due to our client's preference for counties with similar demographics to Champaign, we identify these counties and investigate how their risk profiles evolve. For more straightforward visualizations, we also develop an interactive Shiny application that allows SRC to study data trends. Ultimately, our findings will help SRC prioritize resources that align with sexual crime distributions.

To accomplish this goal, we will analyze monthly county-level sexual crime data between 2021 and 2024 from sources including the NIBRS and County Health Rankings. This requires integrating these different data sources and manually inputting certain data. The resulting dataset serves as the foundation for our exploratory analysis and statistical modeling.

3. Methodology

3.1 Collected Data Description



Source	Variable	Description
Group A Offense Report (NIBRS)	Sexual Crime	Target variable: Monthly total cases of rape, sodomy, sexual assault with an object, fondling, and statutory rape.
Illinois Criminal Justice Information Authority(ICJIA)& Census Data	Percent Rural	Percentage of rural population in the county
	Region	Geographic classification of the county
	Population	Total population size of the county
	65 and Older Percentage	Percentage of county residents aged 65 or older
County Health Rankings & Roadmaps	Unemployment Rate	Percentage of the labor force that is unemployed
	Median Household Income	Median pre-tax income per household
	Uninsured	Percentage of residents without health insurance
	Poor Mental Health Days	Average number of mentally unhealthy days in past 30 days
	Mental Health Providers	Number of people to be served per mental health provider
	Adult Smoking Rate	Proportion of adult population that regularly smokes

Table 1. Summary of Data Sources and Variables

To investigate the factors that may impact sexual crime in Illinois, we constructed a monthly panel dataset covering the period from January 2021 to December 2024 and gathered data on sexual crime from Group A Offenses reported in NIBRS. We defined sexual crime as the sum of reported cases of rape, sodomy, sexual assault with an object, fondling, and statutory rape, following the Group A offense classification.

In addition to sexual crime, we collected county-level public health and socioeconomic indicators from the County Health Rankings & Roadmaps. Poor mental health days show the average number of poor mental health days reported in the past 30 days. Mental health providers suggest the population-to-provider ratio for mental health services. Other Health behavior variables include excessive drinking rate and adult smoking rate. Demographics are also incorporated, such as unemployment rate, median household income, uninsured rate, percentage of residents aged 65 or older, total population, percentage of rural population, and geographic region. As they are only available on an annual scale, we applied disaggregation techniques using population growth trends and interpolation methods to align them with the monthly sexual crime data.

All datasets were cleaned, standardized, and merged into a panel dataset including 12 counties in Illinois. The county selection was guided by our clustering results based on demographic and socioeconomic similarity to Champaign, as well as the specific interest of our client in Vermilion.

3.2 K-means Cluster (Code: Appendix B. Code B.1)

K-means clustering is an unsupervised machine learning technique that partitions observations into K groups based on similarity, aiming to minimize the within-cluster variance.

The clustering was based on four variables: the percentage of rural population, geographic region in Illinois, median household income, and total population. These variables were selected to capture both demographic and geographic factors that may influence socioeconomic outcomes across counties.

Although the elbow method suggests that K=3 may be the optimal number of clusters due to a noticeable flattening of the curve (Figure 1), the total within-cluster sum of squares continues to decline significantly up to K=7. Considering the potential differences among counties in terms of socioeconomic and crime-related characteristics, selecting a larger number of clusters, K=7, allows for a more nuanced segmentation.

Under this solution, we identified 10 counties grouped with Champaign, suggesting they share similar patterns in key indicators. These counties are DeKalb, LaSalle, McLean, Madison, Peoria, Rock Island, St. Clair, Sangamon, Tazewell, and Winnebago.

At the client's request, particular interest was expressed in Vermilion County. Therefore, our final dataset includes a total of 12 counties. They are DeKalb, LaSalle, McLean, Madison, Peoria, Rock Island, St. Clair, Sangamon, Tazewell, Winnebago, as well as Champaign and Vermilion.

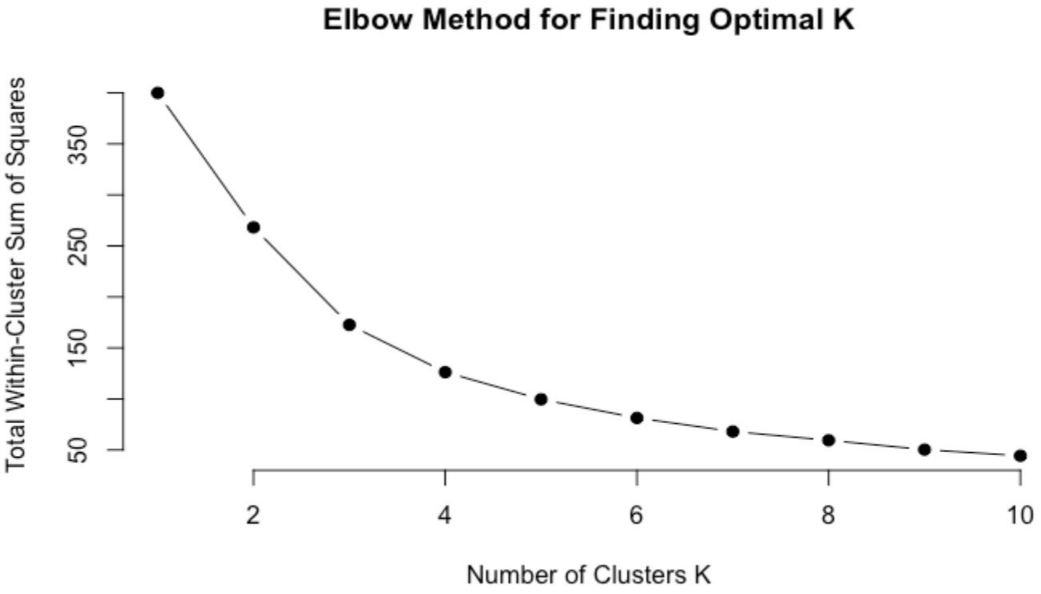
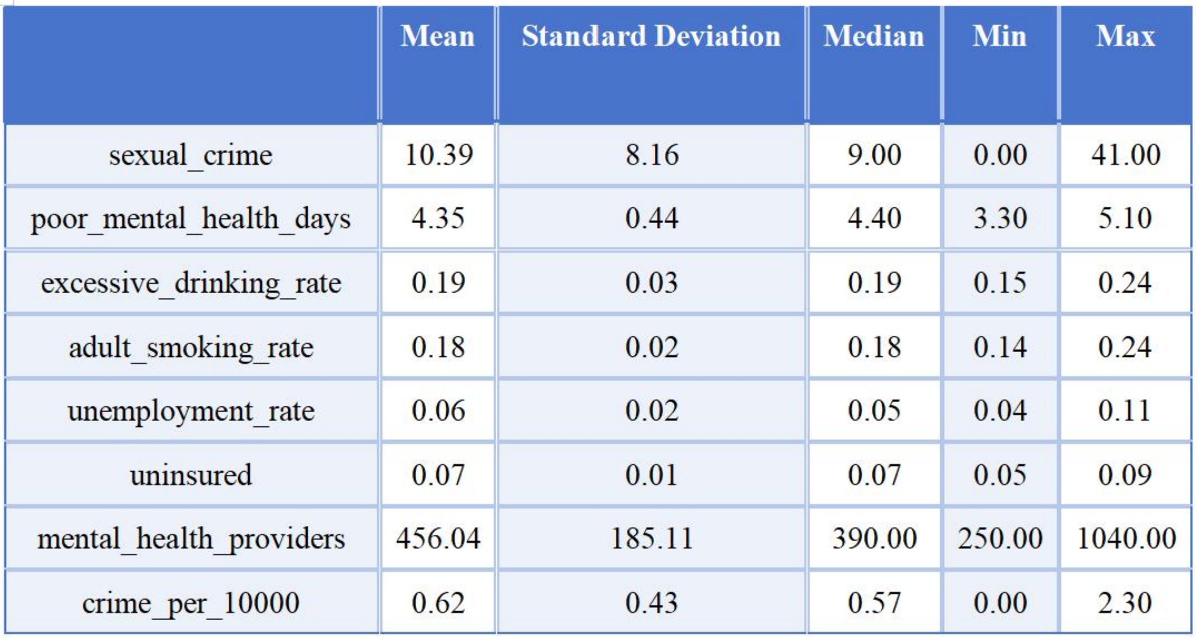


Figure 1. K-means

3.3 Basic Analysis (Code: Appendix B. Code B.2)



	Mean	Standard Deviation	Median	Min	Max
sexual_crime	10.39	8.16	9.00	0.00	41.00
poor_mental_health_days	4.35	0.44	4.40	3.30	5.10
excessive_drinking_rate	0.19	0.03	0.19	0.15	0.24
adult_smoking_rate	0.18	0.02	0.18	0.14	0.24
unemployment_rate	0.06	0.02	0.05	0.04	0.11
uninsured	0.07	0.01	0.07	0.05	0.09
mental_health_providers	456.04	185.11	390.00	250.00	1040.00
crime_per_10000	0.62	0.43	0.57	0.00	2.30

Table 2. Descriptive Statistics

This table summarizes monthly averages and distribution statistics for key variables across the 12 selected counties, over the four years from 2021 to 2024.

The average monthly number of reported sexual crimes across these counties is 10.39, with a high standard deviation of 8.16, indicating substantial variability across regions. Residents in these counties report an average of 4.35 poor mental health days per month, indicating a high mental health distress rate. Both the excessive drinking rate (mean = 0.19, SD = 0.03) and the adult smoking rate (mean = 0.18, SD = 0.02) exhibit low variability across counties, reflecting persistent and widespread public health challenges. Unemployment (mean = 0.06, SD = 0.02) and uninsured rates (mean = 0.07, SD = 0.01) show low standard deviations, implying that these economic stressors are consistently distributed across the selected areas. The average number of mental health providers is 456, but varies greatly (SD = 185.11, range: 90 to 1,040), indicating high disparities in healthcare accessibility. The standardized sexual crime rate per 10,000 residents averages 0.62, with a maximum of 2.30, also confirming that some counties face significantly higher safety risks than others. These summary statistics offer a comprehensive view of health, economic, and safety conditions across the 12 counties.

3.4 Correlation Plot (Code: Appendix B. Code B.3)

To analyze the relationship between the selected variables and to detect the presence of multicollinearity, we plotted a correlation circle plot.

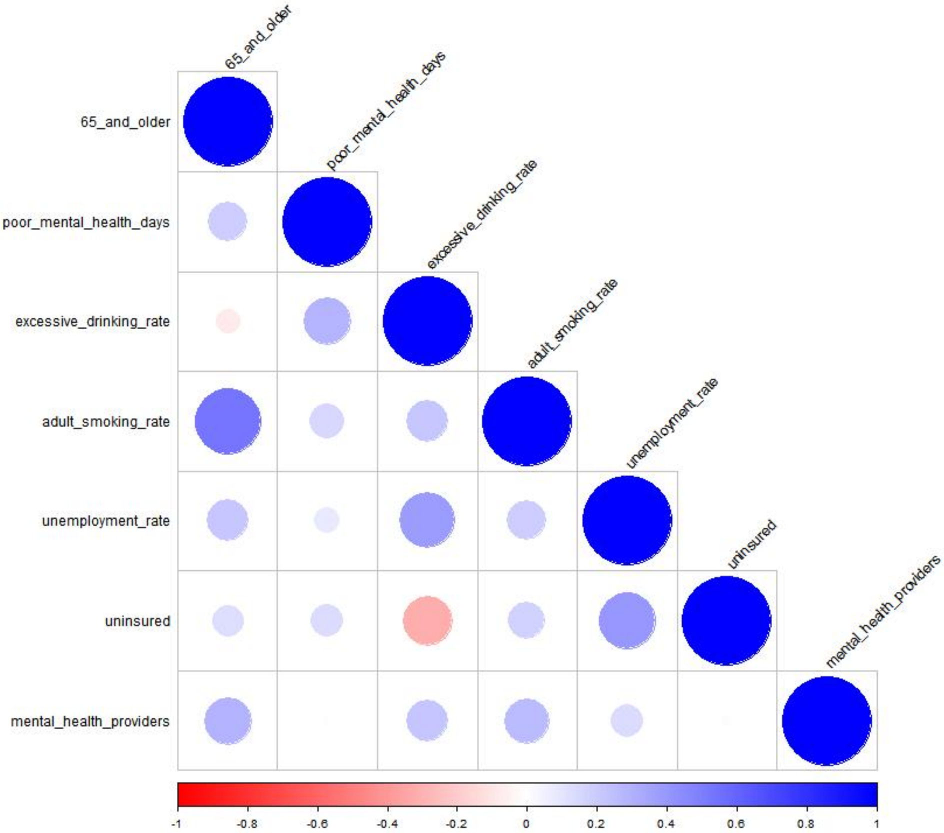


Figure 2. Correlation Plot

According to the Correlation Plot (Figure 2), all variables show correlation coefficients below 0.8, which suggests that multicollinearity is not a concern in this dataset.

3.5 Time Series Plots (Code: Appendix B. Code B.4)

To explore the temporal dynamics of sexual crime in each county, we constructed an overlapped time series plot for all counties.

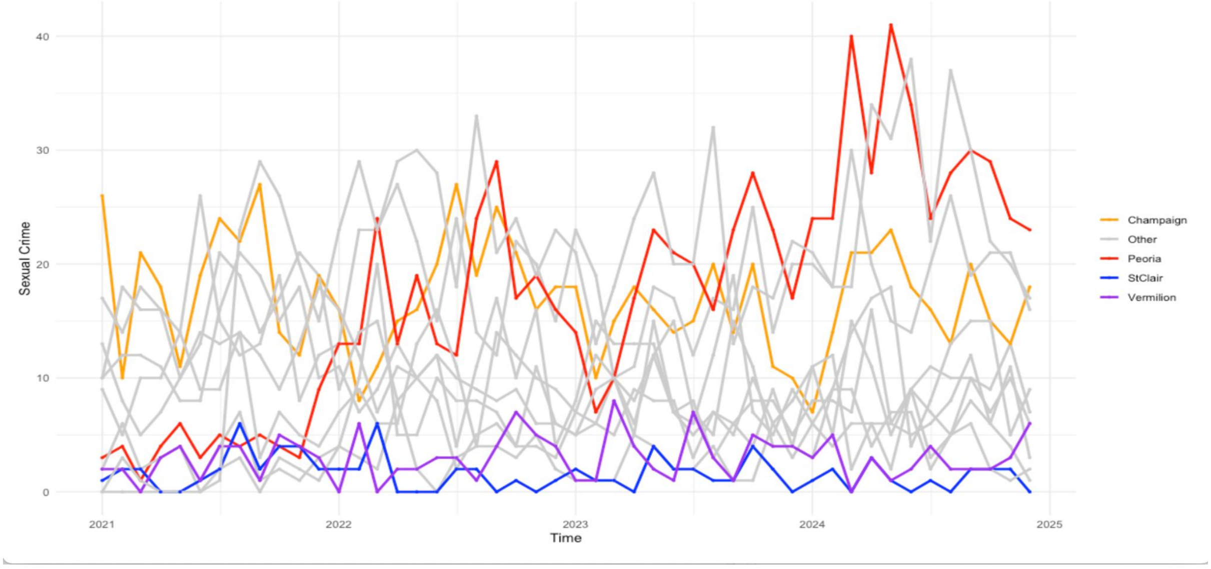


Figure 3. Sexual Crime Trends by County

This figure shows monthly trends in reported sexual crimes across counties from 2021 to 2024. Champaign was selected as the reference county in our clustering analysis. Peoria and St. Clair are examples of counties with different trends in sexual crimes. Vermilion is the area of interest for our client.

The number of sex crimes in Peoria rises rapidly from mid-2023 and peaks in early 2024, showing a clear increasing trend. Champaign also had a high and fluctuating number of overall sex crimes, with several notable peaks during the period. In contrast, St. Clair has the lowest crime rate across almost all times. The gray lines represent other counties, showing different levels of sexual crime.

Finally, we conducted ACF and PACF analyses to support further time series analyses.

3.6 Time Series Models (Code: Appendix B. Code B.5)

We applied ARIMA models to capture time-related patterns in monthly sexual crime counts across counties. These models include components for past values, past error terms, and differencing to remove trends. For each county, we used visual tools, including autocorrelation function (ACF) and partial autocorrelation function (PACF) plots, to help select model parameters. This approach follows established guidelines in time series modeling (Hyndman & Athanasopoulos, 2021). We also used the Augmented Dickey-Fuller (ADF) test to assess whether each time series was stationary. Stationarity means that its overall behavior, such as average level and variation, stays relatively stable over time, rather than drifting or trending systematically. If a series was not stationary, we applied first differencing, and the number of differences used to make it stationary is reported as “ND” in Table 3.

Table 3 summarizes the suggested ARIMA models. Most counties required one difference to become stationary, except Champaign and DeKalb. We also found evidence of an autoregressive structure in all counties. Autoregressive structure implies that the recent past helps predict the near future. For example, if crime was high last month, it's more likely to be high this month.

Because ARIMA models without predictors explained only a small portion of the variance, we further applied Generalized Least Squares (GLS) models with autocorrelated errors to include relevant predictors while still accounting for time-related structure.

County	ACF Cutoff	ACF Tail Off	PACF Cutoff	PACF Tail Off	Stationarity	ND	Suggested Model
McLean	NA	Yes	1	NA	No	1	ARIMA(1,1,0)
Madison	NA	Yes	NA	Yes	No	1	ARIMA(p,1,q)
Peoria	NA	Yes	1	NA	No	1	ARIMA(1,1,0)
Rock Island	NA	Yes	NA	Yes	No	1	ARIMA(p,1,q)
Tazewell	NA	Yes	NA	Yes	No	1	ARIMA(p,1,q)
Winnebago	NA	Yes	1	NA	No	1	ARIMA(1,1,0)
Champaign	NA	Yes	NA	Yes	Yes	0	ARMA(p,q)
DeKalb	NA	Yes	2	NA	Yes	0	ARMA(2,0)
LaSalle	NA	Yes	1	NA	No	1	ARIMA(1,1,0)
Sangamon	NA	Yes	NA	Yes	No	1	ARIMA(p,1,q)
St. Clair	NA	Yes	NA	No	No	1	ARIMA(p,1,q)
Vermilion	NA	Yes	NA	No	No	1	ARIMA(p,1,q)

Table 3. Time Series Modeling with ARIMA by County

3.7 Distribution Map (Code: Appendix B. Code B.6)

Figure 4 shows the total number of reported sexual violence cases in each county for 2024. The data varies for each year. However, counties like Champaign, Madison, and Winnebago consistently have higher total case numbers.

Figure 5 displays the rate of sexual crime per 10,000 people in each county in 2024, which adjusts for the difference in population size. Counties like Dekalb, Champaign, and Peoria had higher crime rates over the years.

Distribution maps for 2021-2023 can be found in Appendix A.

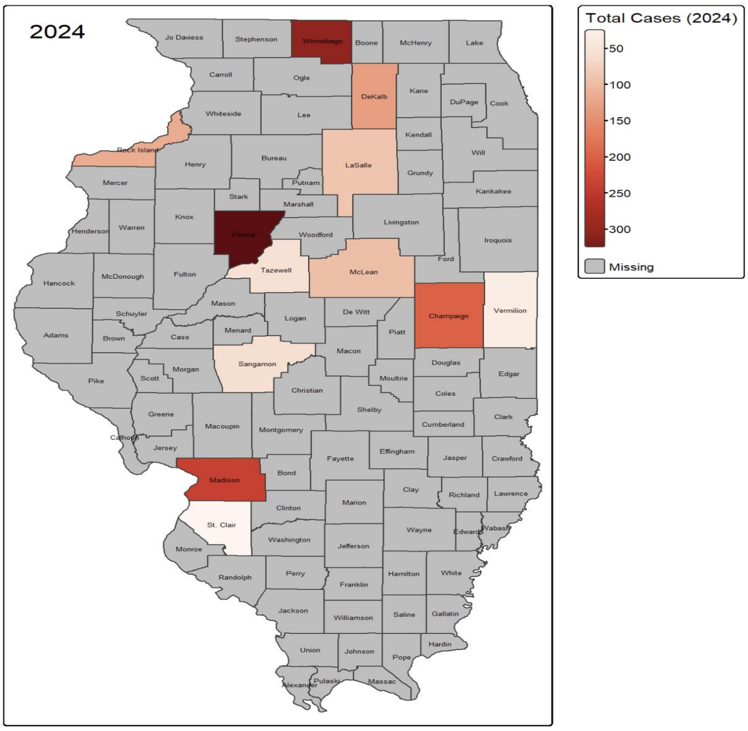


Figure 4. Distribution Map of Total Cases in 2024

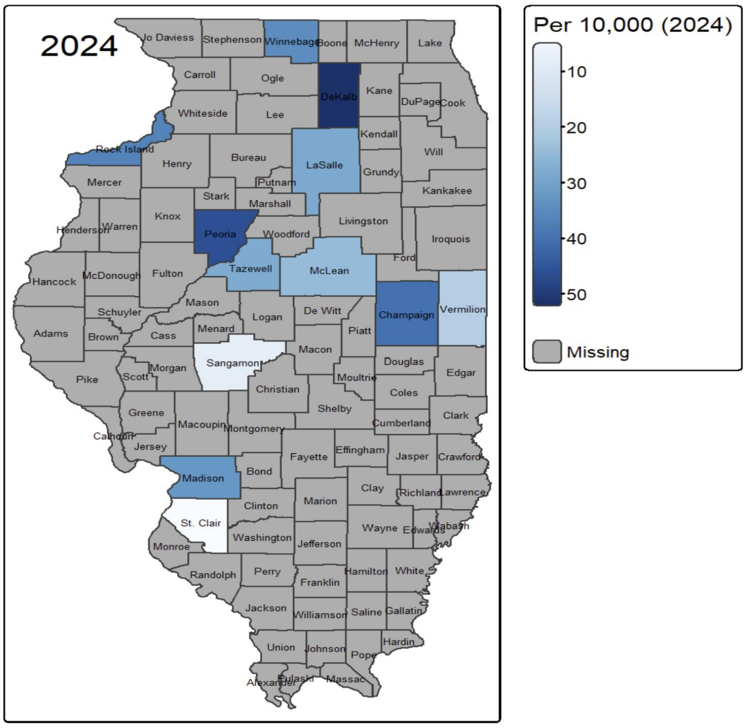


Figure 5. Distribution Map per 10,000 people in 2024

3.8 Challenges

3.8.1 Dataset Construction

One of the main challenges we faced was the lack of a clean dataset that can be used directly. As a result, we had to build the dataset manually by collecting data from multiple public sources. This process involved searching through various websites and copying and organizing the information.

Another difficulty was dealing with inconsistencies and missing values across sources. For instance, some datasets did not include recent years or have extensive zero values, others lacked data for certain counties. These limitations made the data preparation process challenging and required extensive cleaning and integrating to ensure completeness.

3.8.2 GLS Model (Failed, Code: Appendix B. Code B.7)

At the early stage of the project, we observed a clear time trend in the data. To interpret potential patterns, we applied Generalized Least Squares (GLS) modeling, a method originally proposed by Aitken (1935) for obtaining efficient estimates when the assumptions of ordinary least squares are violated.

We tested various correlation structures, including AR(1), ARMA(1,1), and ARIMA(1,1,1), and also applied log-transformations of the variable. In all the models we tried, none of the predictors were significant. Consequently, we decided to abandon GLS and adopt alternative modeling strategies.

Response	Model	Significant Predictors	Residual Pattern
Sexual Crime	GLS AR1	None	Autocorrelated
	GLS ARMA (1,1)	None	
	GLS ARIMA (1,1,1)	None	

Table 4. GLS Regression Diagnostics

4. Proposed Solution

4.1 Final Model Using GEE (Code: Appendix B. Code B.8)

In our project, we switched from GLS to Generalized Estimating Equations (GEE) because GEE is better at identifying overall trends across counties, especially when dealing with data collected over time. GLS tries to model how data points are related within each county, but this can be too strict or unreliable when the data does not follow those patterns closely. GEE, on the other hand, gives more dependable results when we are mainly interested in general patterns rather than exact details within each county. Disaggregating predictors from yearly to monthly with respect to the sexual crime can be misleading, because sexual crime is the response variable and should not be used to adjust itself. We used external population data from the U.S. Bureau of Economic Analysis (2025) to break down the yearly predictor values into monthly ones. This helped remove repetition in the data. Comparing different models, we decided to apply the GEE model, which best balances accuracy and simplicity based on a standard measure, Independence model Criterion (QIC). This helps us evaluate how county-level socioeconomic and health indicators relate to sexual crime trends over time. The approach can also account for repeated monthly measurements within each county. It can also model the deseasonalized trend component of sexual crime counts.

The response variable was derived through STL decomposition. This isolates the trend of monthly sexual crimes. Annual predictors such as mental health days, smoking rate, and unemployment were disaggregated to monthly frequencies using population growth trends and interpolation methods, using the tempdisagg package in R.

We considered all combinations of six monthly predictors and fitted GEE models with an autoregressive correlation structure AR(1). Six monthly disaggregated predictors result in 63 total combinations, excluding the null model without predictors. Each model was evaluated using the Quasi-likelihood under QIC. Figure 6 illustrates the QICs of combinations of predictors, and the significant drop in QIC is due to the inclusion of the most essential predictor, mental health providers. As shown in Table 4, the final selected model minimized QIC (25667.9) and included key predictors, including excessive_drinking_rate, adult smoking rate, unemployment rate, and the number of mental health providers per 10,000 residents. With a p-value of 0.003, access to sufficient mental health providers seems to be the most relevant predictor of reduced sexual crime. This model provides slightly lower QIC than using predictors without disaggregation (26188.4), and significantly lower QIC than a null linear model (29719.6).

Variable	Estimate	Std. Error	Wald Statistic	p-value
(Intercept)	16.4866	5.1442	10.27	0.0014**
Monthly Excessive Drinking Rate	-1.1098	9.2306	0.01	0.9043
Monthly Adult Smoking Rate	-17.3417	17.2667	1.01	0.3152
Monthly Unemployment Rate	4.2959	8.0056	0.29	0.5915
Monthly Mental Health Providers (per 10,000)	-0.1193	0.0402	8.82	0.0030**

Table 5. GEE Model Results Summary

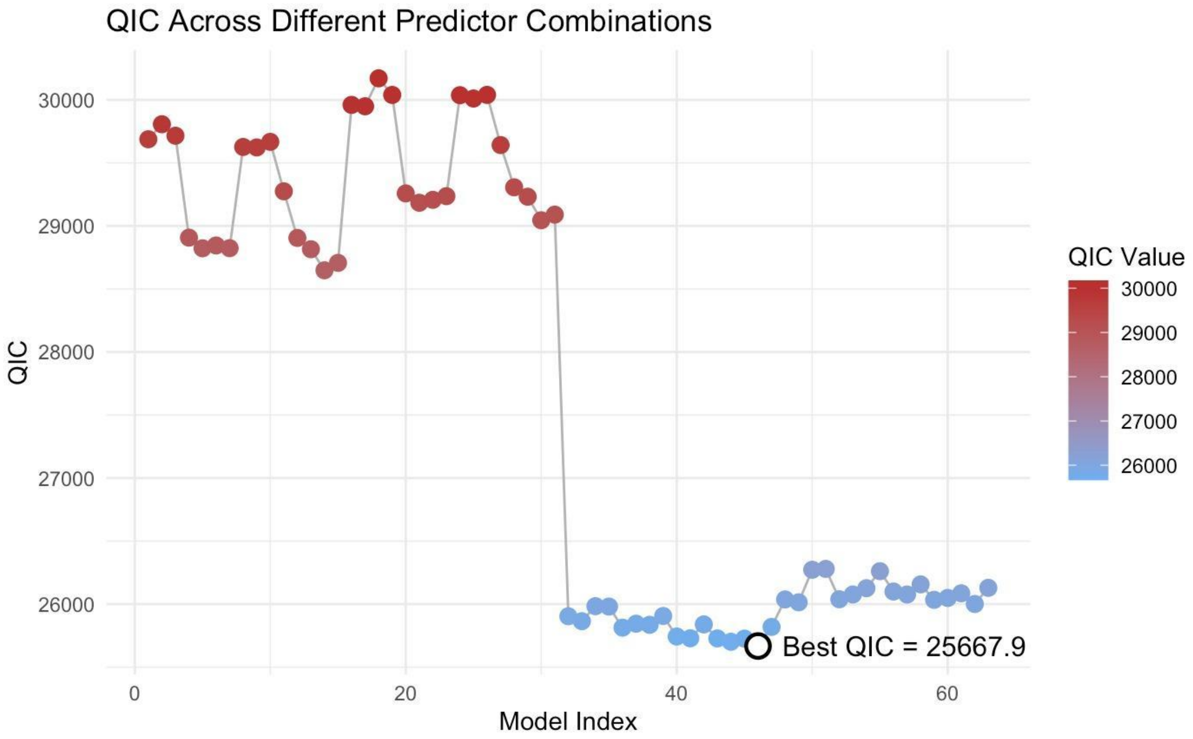


Figure 6. GEE Model QICs of Combinations of Predictors

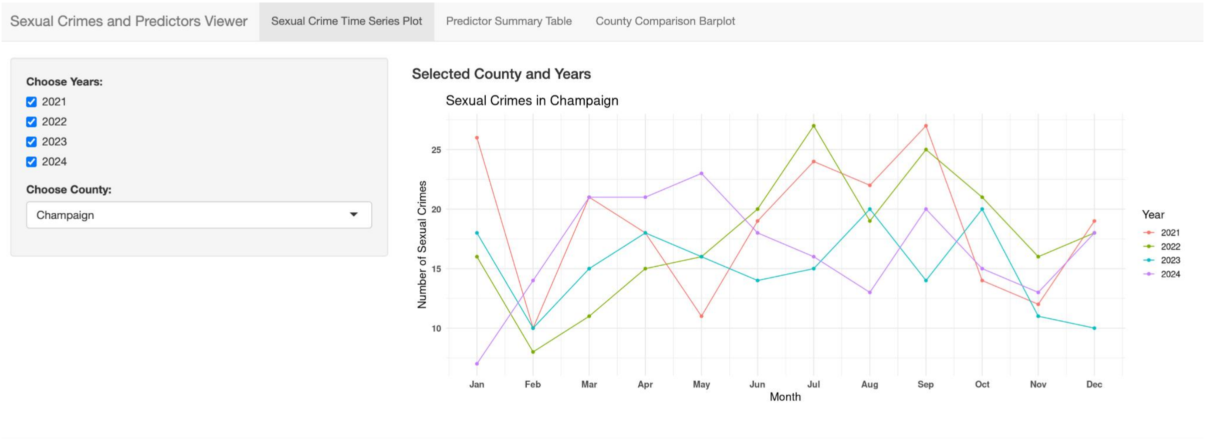
(Code: Appendix B. Code B.9)

The results suggest that higher levels of mental health burden and unemployment are associated with increased sexual crimes, while more abundant access to mental health services correlates with reductions in crime.

4.2 Shiny APP (Code: Appendix B. Code B.10)

Our Shiny application is designed to help clients explore sexual crime trends and related predictors across selected Illinois counties from 2021 to 2024.

The first tab illustrates the interactive county and year selection. Users can select one or more years and choose a specific county to view detailed monthly trends in sexual crime incidents.


Sexual Crimes and Predictors Viewer
Sexual Crime Time Series Plot
Predictor Summary Table
County Comparison Barplot
Sexual Crimes in Champaign
Number of Sexual Crimes
25
20
15
10
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month
Year
2021
2022
2023
2024

Choose Years:

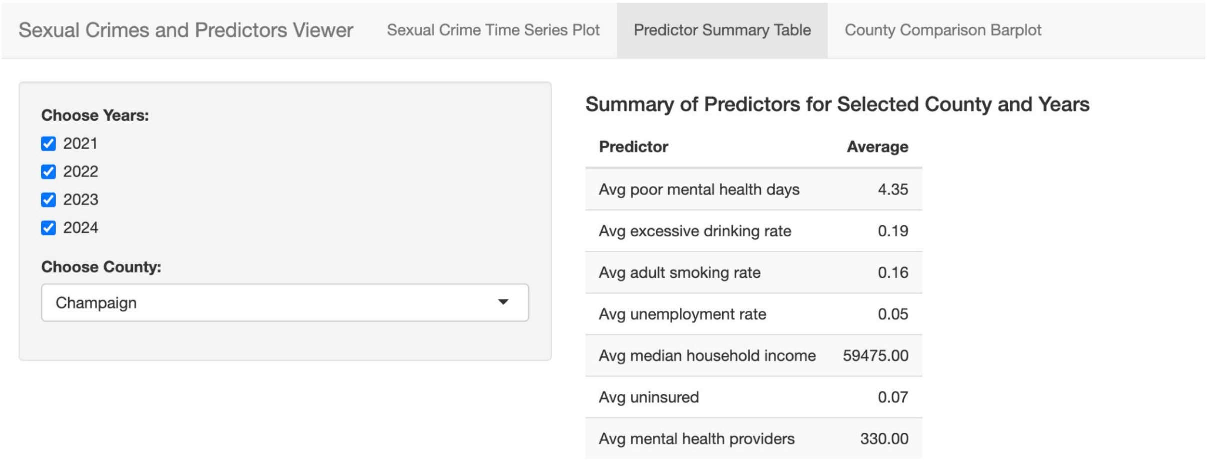
☑ 2021
☑ 2022
☑ 2023
☑ 2024

Choose County: Champaign

Selected County and Years

Figure 7. Sexual Crime Time Series Plot for Selected County and Years

The second tab is the predictor summary panel**.** Once a county and time range are selected, the app displays average values for key predictors such as poor mental health days, excessive drinking rate, smoking rate, unemployment, median household income, insurance coverage, and mental health provider availability.


Sexual Crimes and Predictors Viewer
Sexual Crime Time Series Plot
Predictor Summary Table
County Comparison Barplot

Choose Years:

☑ 2021
☑ 2022
☑ 2023
☑ 2024

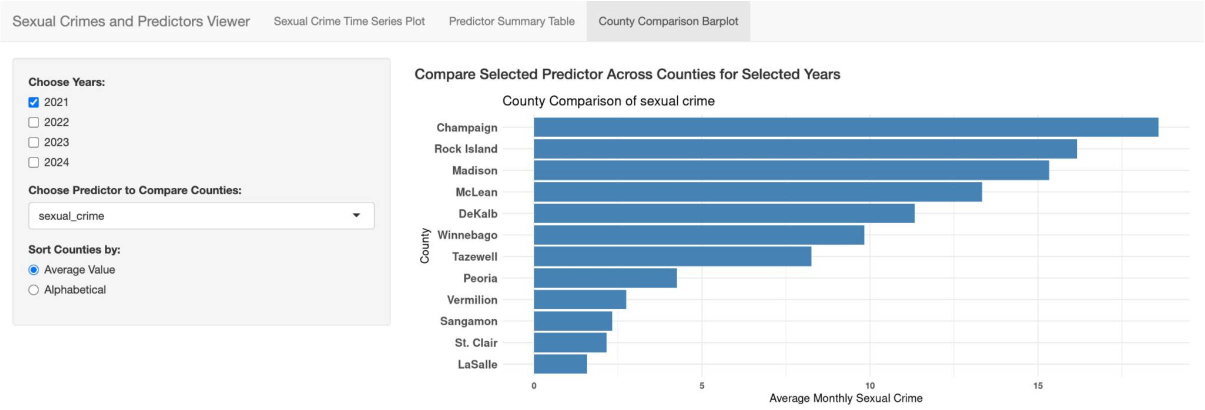
Choose County: Champaign

Summary of Predictors for Selected County and Years

Predictor	Average
Avg poor mental health days	4.35
Avg excessive drinking rate	0.19
Avg adult smoking rate	0.16
Avg unemployment rate	0.05
Avg median household income	59475.00
Avg uninsured	0.07
Avg mental health providers	330.00

Figure 8. Summary of Predictors for Selected County and Years

Finally, the third tab serves as a cross-county comparison tool**.** The app also allows the user to compare selected counties by average values of any predictor, and helps to identify counties with similar or different characteristics. Counties can be sorted by average value or alphabetically.


Sexual Crimes and Predictors Viewer
Sexual Crime Time Series Plot
Predictor Summary Table
County Comparison Barplot
Compare Selected Predictor Across Counties for Selected Years
County Comparison of sexual crime
Champaign
Rock Island
Madison
McLean
DeKalb
Winnebago
Tazewell
Peoria
Vermilion
Sangamon
St. Clair
LaSalle
County
0
5
10
15
Average Monthly Sexual Crime

Choose Years:

☑ 2021
☐ 2022
☐ 2023
☐ 2024

Choose Predictor to Compare Counties:

sexual_crime

Sort Counties by:

● Average Value
○ Alphabetical

Figure 9. County-Level Comparison of Selected Predictors

5. Final Recommendations and Conclusion

After systematically evaluating all possible combinations of six candidate variables using GEE with an AR(1) correlation structure, and assessing each model based on QIC, we concluded a subset of predictors that minimized QIC as the final model. The optimal model includes excessive drinking rate, adult smoking rate, unemployment rate, and mental health providers per 10,000 people. All these variables performed better under monthly disaggregation.

While 3 of the 4 predictors in this model are not statistically significant, mental health provider availability remains significant. This result strongly supports the interpretation that easy access to mental health services is associated with reduced sexual crime rates. The selected model achieved a lower QIC, compared with the model without monthly disaggregation. The model selection process based on combination search and QIC minimization ensures that the final results are reliable.

Several limitations remain for our project. The statistical associations from the GEE model do not imply causation. Also, the analysis focuses on variables without looking at their potential lag effects due to dataset size limitations. Delayed impacts of unemployment or mental health distress may enhance model performance. Some confounding factors, including policing practices, education levels, and housing instability, were not included. Finally, the current analysis is conducted at the county level, which limits insight from smaller-scale data. Overall, these limitations stem primarily from the lack of access to more detailed, high-frequency, and comprehensive data.

For future steps, SRC could allocate resources to high-risk counties, focusing more on mental health support according to the model's outputs. The Shiny application can also be applied as a support for decision-making. Future work should incorporate additional environmental variables and update the dataset over time. Building partnerships with state or academic institutions could further enhance the richness and scope of the analysis.

6. References

Aitken, A. C. (1935). On Least Squares and Linear Combinations of Observations. *Proceedings of the Royal Society of Edinburgh*, 55, 42–48.

County Health Rankings & Roadmaps. (2024). Frequent mental distress. University of Wisconsin Population Health Institute. https://www.countyhealthrankings.org/health-data/population-health-and-well-being/quality-of-life/mental-health/frequent-mental-distress

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts.

Illinois State Police. (n.d.). Crime in Illinois. *Illinois Uniform Crime Reporting (I-UCR) Program*. https://ilucr.nibrs.com/CrimePublication/CrimeinIllinois

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1), 13–22. https://doi.org/10.1093/biomet/73.1.13

U.S. Bureau of Economic Analysis, Population [POPTHM], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/POPTHM, April 29, 2025

7. Appendix

Appendix A: Figure Appendix

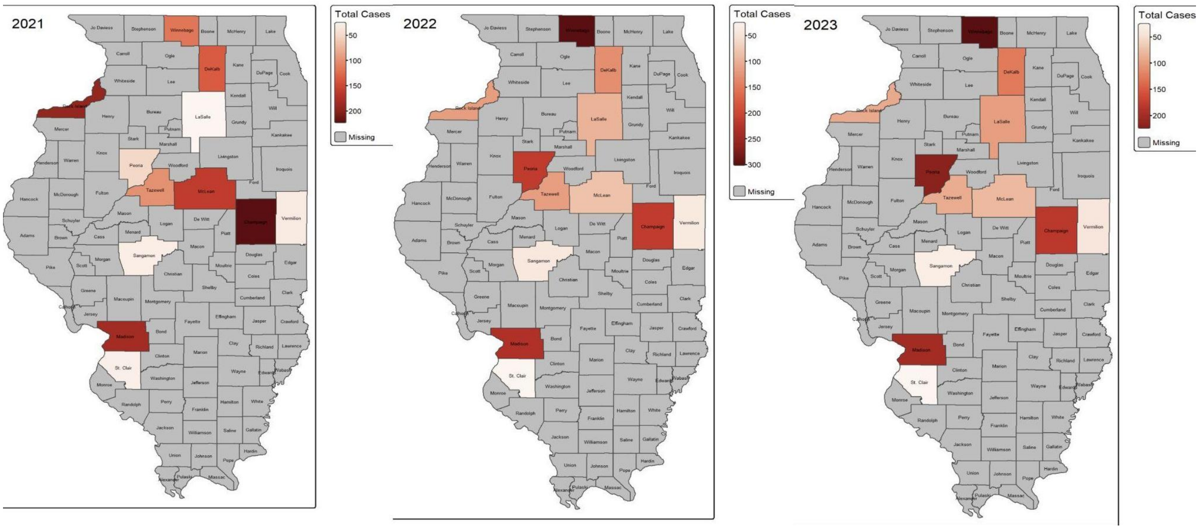


Figure A.1 Distribution Map in 2021, 2022, and 2023

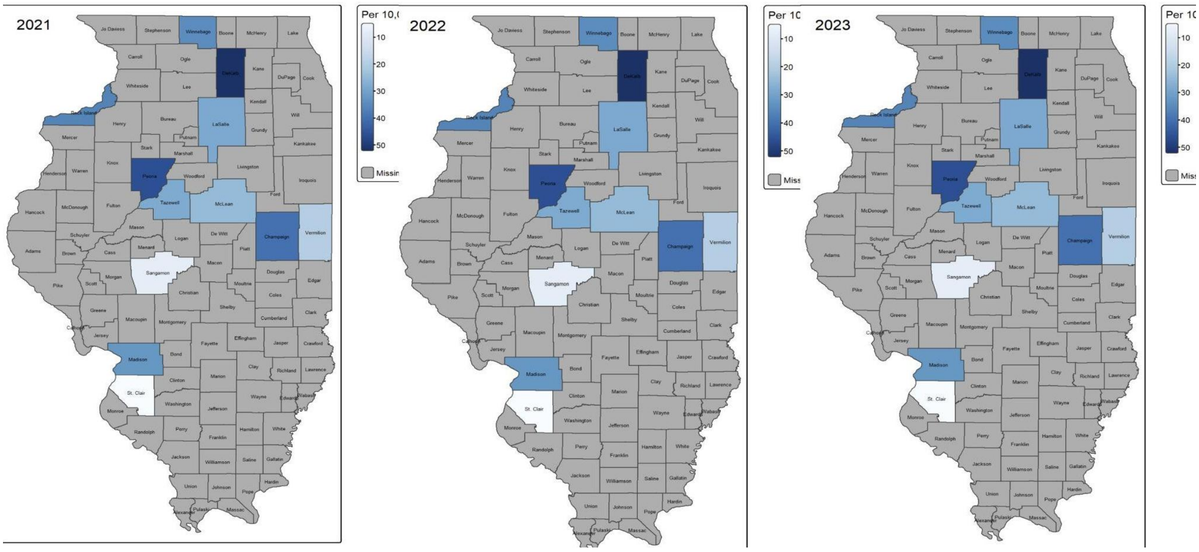


Figure A.2 Distribution Map per 10,000 people in 2021, 2022 and 2023

Appendix B: Code Appendix

In this section, the code is presented in the order it was written and run during the analysis process.

Code B.1: K-means Cluster

```
library(tidyverse)
library(factoextra)
data=read.csv('final_merged_data.csv') %>% filter(Year==2020 & County!="Cook")
data_cluster <- data %>%
  select(percent_rural, region, Median_Household_Income, population)
data_cluster$region <- as.numeric(factor(data_cluster$region))
data_cluster <- na.omit(data_cluster)
data_scaled <- scale(data_cluster)

set.seed(1)
wss <- numeric(10)
for (k in 1:10) {
    kmeans_result <- kmeans(data_scaled, centers = k, nstart = 25)
    wss[k] <- kmeans_result$tot.withinss
}

plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
     xlab = "Number of Clusters K",
     ylab = "Total Within-Cluster Sum of Squares",
     main = "Elbow Method for Finding Optimal K")

k <- 7
kmeans_result <- kmeans(data_scaled, centers = k, nstart = 25)
data_cluster$cluster <- kmeans_result$cluster
print(data_cluster$cluster)
table(data_cluster$cluster)
data_cluster$county <- data$County
print(data_cluster$county[data_cluster$cluster == 3])
```

Code B.2: Basic Analysis

```
library(psych)
library(tidyverse)
library(readxl)

data <- read_excel("dataset.xlsx")
summary(data)
numeric_vars <- data %>% select(where(is.numeric))
describe_by_county <- describeBy(numeric_vars, group = data$County, mat = TRUE)
describe(numeric_vars)
```

Code B.3: Correlation Plot

```
library(corrplot)
library(readxl)
data <- read_excel("dataset.xlsx")
selected_data <- data[, sapply(data, is.numeric)]
selected_data <- subset(selected_data,
                        select = -c(`sexual_crime`,
crime_per_10000))
corr_matrix <- c
or(selected_data, use = "complete.obs")
corrplot(corr_matrix,
         method = "circle",
         type = "lower",
         order = "original",
         tl.col = "black",
         tl.srt = 45,
         col = colorRampPalette(c("red", "white", "blue"))(200),
         diag = TRUE,
         addCoef.col = NULL,
         number.cex = 0.7
)
png("correlation_circle_plot_v2.png", width = 1000, height = 800)
corrplot(corr_matrix,
         method = "circle",
         type = "lower",
         order = "original",
         tl.col = "black",
         tl.srt = 45,
         col = colorRampPalette(c("red", "white", "blue"))(200),
         diag = TRUE)
dev.off()
```

Code B.4: Time Series Plot

```
library(readxl)
library(dplyr)
data <- read_excel("dataset.xlsx")
data$Time <- as.Date(data$Time)
data <- data %>%
  mutate(color_group = case_when(
    County == "Peoria" ~ "Peoria",
    County == "St. Clair" ~ "StClair",
    County == "Champaign" ~ "Champaign",
    County == "Vermilion" ~ "Vermilion",
    TRUE ~ "Other"
  ))
ggplot(data, aes(x = Time, y = sexual_crime , group = County)) +
  geom_line(aes(color = color_group), size = 1) +
  geom_point(aes(color = color_group), shape = 16, size = 1) +
  scale_color_manual(values = c(
    "Peoria" = "red",
    "StClair" = "blue",
    "Champaign"="orange",
    "Vermilion"="purple",
    "Other" = "gray80"

  )) +
  labs(title = "Sexual Crime by County",
       x = "Time",
       y = "Sexual Crime",
       color = NULL) +
  theme_minimal() +
  theme(legend.title = element_text(size = 10),
        legend.position = "right")
```

Code B.5: Time Series Models

The following code in this part represents a segment of the full code used for time series modeling. In this part, we applied the same procedure to each county, so we only show the code for Champaign. The suggested ARIMA model can be identified according to the ACF and PACF plots.

```
library(tseries)
library(urca)
library(tidyverse)
library(readxl)
data <- read_excel("dataset.xlsx")
data$Time <- as.Date(paste0(data$Time, "-01"))
data_Champaign <- data %>%
  filter(County == "Champaign") %>%
  arrange(Time)
y_Champaign <- ts(data_Champaign$`sexual_crime`, start = c(2021, 1), frequency =
12)
adf.test(y_Champaign) # adf test stationarity: no difference is stationary
acf(y_Champaign, lag.max = 24, main = "ACF Plot for Sexual Crime in Champaign
County", xaxt = "n") # ACF
axis(1, at = seq(0, 24, by = 1), labels = seq(0, 24, by = 1))

pacf(y_Champaign, lag.max = 24, main = "PACF Plot for Sexual Crime in
Champaign County", xaxt = "n") # PACF
axis(1, at = seq(0, 24, by = 1), labels = seq(0, 24, by = 1))
```

Code B.6: Distribution Map

The following code in this part represents a segment of the full code used for visualization, and only key components are shown here. The code corresponds to the heatmaps of Illinois in 2024.

```
library(tidyverse)
library(readxl)
library(sf)
library(tigris)
library(tmap)

data <- read_excel("dataset.xlsx")

sexual_violence_data <- data %>%
  mutate(year = format(Time, "%Y"))
```

```
il_county_shapes <- counties(state = "IL", cb = TRUE, class = "sf")

# 2024 Total Case
total_crime_2024 <- sexual_violence_data %>%
  filter(year == "2024") %>%
  group_by(County) %>%
  summarise(total = sum(sexual_crime, na.rm = TRUE), .groups = "drop")

map_total_2024 <- left_join(il_county_shapes, total_crime_2024, by = c("NAME" =
"County"))

tm_shape(map_total_2024) +
  tm_polygons(
    fill = "total",
    palette = "Reds",
    style = "cont",
    title = "Total Cases (2024)"
  ) +
  tm_text("NAME", size = 0.4) +
  tm_layout(legend.outside = TRUE, title = "2024")

# 2024 (per 10,000)
plot_rate_map_facet <- function(data, year, map_sf) {
  county_year_summary <- data %>%
    filter(year == year) %>%
    group_by(County) %>%
    summarise(total_crimes = sum(sexual_crime, na.rm = TRUE),
              population = first(population),
              .groups = "drop") %>%
    mutate(rate_per_10000 = (total_crimes / population) * 10000)

  final_map <- left_join(map_sf, county_year_summary, by = c("NAME" =
"County"))

  tm_shape(final_map) +
    tm_polygons(
      fill = "rate_per_10000",
      palette = "Blues",
      style = "cont",
      title = paste("Per 10,000 (", year, ")", sep = "")
    ) +
    tm_text("NAME", size = 0.4) +
  tm_layout(legend.outside = TRUE, title = paste(year))
```

Code B.7: GLS model

```
library(tidyverse)
library(nlme)
library(readxl)
data <- read_excel("dataset.xlsx")
data$Time <- as.numeric(as.factor(data$Time))

mod_gls_1a <- gls(sexual_crime ~ County + poor_mental_health_days +
excessive_drinking_rate +
               adult_smoking_rate + unemployment_rate +
               uninsured + mental_health_providers,
               data = data,
               correlation = corAR1(form = ~ Time | County),
               method = "ML")
summary(mod_gls_1a)
par(mfrow = c(1, 2))
acf(mod_gls_1a$residuals, main = "ACF of GLS Residuals")
pacf(mod_gls_1a$residuals, main = "PACF of GLS Residuals")
```

Code B.8: Final model

```
library(tidyverse)
library(tempdisagg)
library(lubridate)
library(geepack)
library(qif)

# 1.Preprocess-function :Disaggregate each yearly predictor to monthly

preprocess = function(county_data, pop, predictors, response) {
  monthly_ts = ts(county_data$population * pop,
                  start = c(min(lubridate::year(county_data$date)),
                            min(lubridate::month(county_data$date))),
                  frequency = 12)
  for (pred in predictors) {
    yearly_col = county_data %>%
      group_by(year) %>%
      summarise(value = first(.data[[pred]]))
    yearly_ts = ts(yearly_col$value, start = min(yearly_col$year), frequency = 1)
    td_model = td(yearly_ts ~ monthly_ts, to = 12, conversion = "average")
    monthly_pred = predict(td_model)

    new_colname = paste0("monthly_", pred)
    county_data[[new_colname]] = as.numeric(monthly_pred)
  }

# 2.Decompose the response time series using STL
  monthly_response = ts(county_data[[response]],
                        start = c(min(lubridate::year(county_data$date)),
                                  min(lubridate::month(county_data$date))),
                        frequency = 12)

  stl_response = stl(monthly_response, "periodic")


  county_data[[paste0("deseason_", response)]] = rowSums(stl_response$time.series[,
2:3])
  county_data[[paste0("trend_", response)]] = as.numeric(stl_response$time.series[,
2])

  return(county_data)
}

# 3.Reading data and initial processing data
pop = read_csv("population_il.csv")$percent_change_year_ago
raw = read_excel("dataset.xlsx")
names(raw) = c("time", "county", "sexual_crime",
               "percent_rural", "region", "population", "older_than_65",
               "poor_mental_health_days", "excessive_drinking_rate",
               "adult_smoking_rate", "unemployment_rate",
"median_household_income",
               "uninsured", "mental_health_providers", "sexual_crime_per_10000")
raw$date = as.Date(paste0(raw$time, "-01"))
raw$year = lubridate::year(raw$date)
raw$mental_health_providers_per_10000 = raw$mental_health_providers /
raw$population * 10000
raw_list = split(raw, raw$county)

# 4.Applying the preprocess() function
```

```
predictors = c("poor_mental_health_days", "excessive_drinking_rate",
               "adult_smoking_rate", "unemployment_rate",
               "uninsured", "mental_health_providers_per_10000")
response = "sexual_crime"
processed_list = lapply(raw_list, preprocess, pop=pop, predictors=predictors,
response=response)
processed = bind_rows(processed_list)
colnames(processed)
predictors <- c("monthly_poor_mental_health_days",
                "monthly_excessive_drinking_rate",
                "monthly_adult_smoking_rate",
                "monthly_unemployment_rate",
                "monthly_uninsured",
                "monthly_mental_health_providers_per_10000")

# 5. Final model selection

combination = vector("list", length(predictors))
for (i in 1:length(predictors))
  combination[[i]] = c(FALSE, TRUE)
combination = expand.grid(combination)
combination = as.matrix(combination[-1, ])
out = matrix(nrow = nrow(combination), ncol = 5)
for (i in 1:nrow(combination)) {
  f = paste0("trend_sexual_crime ~ ", paste(predictors[combination[i, ]], collapse =
"+"))
  out[i, 1] = f

  model = geeglm(as.formula(f),
                 data = processed,
                 id = factor(processed$county),
                 family = gaussian,
                 corstr = "ar1")

  out[i, 2:5] = unname(QIC(model)[c(1, 2, 4, 6)])
}
f = out[which.min(out[, 2]), 1]
model = geeglm(as.formula(f),
               data = processed,
               id = factor(processed$county),
               family = gaussian,
               corstr = "ar1")
summary(model)
```

```
QIC(model)

# 6.Null model check
model0 = geeglm(as.formula(paste0("trend_", response, "~1")),
               data=processed, id=factor(processed$county),
               family=gaussian, corstr="ar1")

QIC(model0)
```

Code B.9: Model Selection

```
library(ggplot2)
out_df$highlight <- out_df$QIC == min(out_df$QIC)
best_model_row <- subset(out_df, highlight == TRUE)
ggplot(out_df, aes(x = model, y = QIC)) +
  geom_line(aes(group = 1), color = "gray70") +
  geom_point(aes(color = QIC), size = 3) +
  geom_point(data = best_model_row,
             shape = 21, color = "black", fill = "white", size =
4, stroke = 1.2) +
  geom_text(data = best_model_row,
            aes(label = paste0("Best QIC = ", round(QIC, 1))),
            hjust = -0.1, vjust = 0.5, size = 4.2, color =
"black") +
  scale_color_gradient(low = "#56B1F7", high = "#CA0020") +
  labs(title = "QIC Across Different Predictor Combinations",
       x = "Model Index", y = "QIC", color = "QIC Value") +
  theme_minimal()
```

Code B.10: Shiny APP

```
library(shiny)
library(DT)
library(ggplot2)
library(dplyr)
library(stringr)
library(readxl)
data <- read_excel("dataset.xlsx")
data <- data %>%
  mutate(Year = as.numeric(substr(Time, 1, 4)),
         Month = as.numeric(substr(Time, 6, 7)))
county_choices <- sort(unique(data$County))
```

```
format_predictor <- function(predictor) {
 if (predictor == "sexual_crime") {
  return("Average Monthly Sexual Crime")
 } else if (predictor == "poor_mental_health_days") {
  return("Average Monthly Poor Mental Health Days")
 } else {
  return(paste("Average", str_replace_all(predictor, "[_\\.]", " ")))
 }
}
ui <- navbarPage("Sexual Crimes and Predictors Viewer",
         tabPanel("Sexual Crime Time Series Plot",
             sidebarLayout(
              sidebarPanel(
               checkboxGroupInput("years_plot", "Choose Years:", choices =
2021:2024, selected = 2021),
               selectInput("county_plot", "Choose County:", choices =
county_choices)
              ),
              mainPanel(
               h4("Selected County and Years"),
               plotOutput("time_series_plot")
              )
             )
         ),
         tabPanel("Predictor Summary Table",
             sidebarLayout(
              sidebarPanel(
               checkboxGroupInput("years_table", "Choose Years:", choices =
2021:2024, selected = 2021),
               selectInput("county_table", "Choose County:", choices =
county_choices)
              ),
              mainPanel(
               h4("Summary of Predictors for Selected County and Years"),
               tableOutput("predictor_table")
              )
             )
         ),
         tabPanel("County Comparison Barplot",
             sidebarLayout(
              sidebarPanel(
               checkboxGroupInput("years_bar", "Choose Years:", choices =
2021:2024, selected = 2021),
                              selectInput("predictor", "Choose Predictor to Compare Counties:",
                                          choices = c("sexual_crime", "poor_mental_health_days",
"excessive_drinking_rate", "adult_smoking_rate",
                                                       "unemployment_rate", "median_household_income",
"uninsured", "mental_health_providers")),
                              radioButtons("sort_order", "Sort Counties by:",
                                           choices = c("Average Value" = "value", "Alphabetical" =
"alpha"),
                                           selected = "value")
                            ),
                            mainPanel(
                              h4("Compare Selected Predictor Across Counties for Selected
Years"),
                              plotOutput("predictor_barplot")
                            )
                          )
                 )
)
server <- function(input, output, session) {
  filtered_data_plot <- reactive({
    req(input$years_plot)
    subset(data, Year %in% input$years_plot & County == input$county_plot)
  })
  output$time_series_plot <- renderPlot({
    plot_data <- filtered_data_plot()
    ggplot(plot_data, aes(x = Month, y = sexual_crime, color = factor(Year), group =
Year)) +
      geom_line() +
      geom_point() +
      scale_x_continuous(breaks = 1:12, labels = month.abb) +
      labs(x = "Month", y = "Number of Sexual Crimes", color = "Year") +
      theme_minimal(base_size = 14) +
      theme(axis.text.x = element_text(face = "bold"),
            axis.text.y = element_text(face = "bold")) +
      ggtitle(paste("Sexual Crimes in", input$county_plot))
  })

  filtered_data_table <- reactive({
    req(input$years_table)
    subset(data, Year %in% input$years_table & County == input$county_table)
  })
  output$predictor_table <- renderTable({
    filtered_data_table() %>%
      summarise(
        "Avg poor mental health days" = round(mean(poor_mental_health_days, na.rm
= TRUE), 3),
        "Avg excessive drinking rate" = round(mean(excessive_drinking_rate, na.rm =
TRUE), 3),
        "Avg adult smoking rate" = round(mean(adult_smoking_rate, na.rm = TRUE),
3),
        "Avg unemployment rate" = round(mean(unemployment_rate, na.rm = TRUE),
3),
        "Avg median household income" = round(mean(median_household_income,
na.rm = TRUE), 3),
        "Avg uninsured" = round(mean(uninsured, na.rm = TRUE), 3),
        "Avg mental health providers" = round(mean(mental_health_providers, na.rm =
TRUE), 3)
      ) %>%
      t() %>%
      as.data.frame() %>%
      tibble::rownames_to_column("Predictor") %>%
      rename("Average" = V1)
  }, striped = TRUE, spacing = "m")
  output$predictor_barplot <- renderPlot({
    plot_data <- data %>%
      filter(Year %in% input$years_bar) %>%
      group_by(County) %>%
      summarise(Average = mean(.data[[input$predictor]], na.rm = TRUE))
    if (input$sort_order == "value") {
      plot_data <- plot_data %>% arrange(Average)
    } else {
      plot_data <- plot_data %>% arrange(desc(County))
    }
    plot_data$County <- factor(plot_data$County, levels = plot_data$County)
    y_label <- format_predictor(input$predictor)
    title_label <- str_replace_all(input$predictor, "[_\\.]", " ")
    ggplot(plot_data, aes(x = County, y = Average)) +
      geom_bar(stat = "identity", fill = "steelblue") +
      coord_flip() +
      labs(x = "County", y = y_label, title = paste("County Comparison of", title_label))
+
      theme_minimal(base_size = 14) +
      theme(axis.text.y = element_text(face = "bold", size = 14),
            axis.text.x = element_text(face = "bold"))
  })
}
```

```
shinyApp(ui = ui, server = server)
```