The Impact of Regional Safety and Educational Attainment on Mental Health: A Multi-Factor Analysis

Rongrong Wang, Jinying Xing, Shengji Jin, Shujun Jiang

Abstract

This study examines how state-level educational attainment and crime rates interact to influence individual mental health outcomes in the U.S. Using synthetic EHR data linked with public datasets, we analyzed 3,000 individuals across eight states. While state-level crime and education were not independently associated with depression severity, a significant interaction was found: college-educated individuals showed better mental health in highly educated states, but this pattern did not hold for those without college degrees. Additionally, the protective effect of education diminished in high-crime environments. These findings highlight the importance of considering contextual interactions when evaluating structural determinants of mental health.

Keywords: mental health, education, crime, multilevel analysis, social determinants of health

1. Introduction

In recent years, mental health problems in the United States have continued to increase, emerging as a significant public health concern. According to the Centers for Disease Control and Prevention (CDC), approximately one in five adults experiences a mental disorder each year. Data from the 2023 National Health Interview Survey (NHIS) further indicate that nearly 30% of adults report symptoms of depression or anxiety—substantially higher than pre-pandemic levels.Mental health outcomes are shaped by both individual-level factors and broader social determinants of health (SDOH), such as educational attainment, income inequality, crime rates, and access to healthcare services. Among these, state-level educational attainment and crime rates are particularly influential yet unevenly distributed across regions. States with higher education levels often exhibit greater economic stability and social cohesion, which may promote psychological well-being. In contrast, high crime rates can serve as chronic stressors, undermining individuals' sense of safety and trust in their communities. However, prior research offers mixed evidence on the independent and joint effects of these structural factors.

Recent developments in multilevel theory and social-ecological models highlight the importance of interactions between individuals and their environments. Specifically, the mental health impact of an individual's educational attainment may depend on the broader educational context in which they live. For instance, individuals with lower education levels may experience higher stress in more educated communities due to perceptions of relative disadvantage.Despite these theoretical advancements, few empirical studies in the U.S. have applied multilevel modeling to examine how both individual- and state-level educational attainment and crime rates influence mental health. Research that accounts for individual characteristics—such as age, gender, and education—and incorporates clinically validated mental health measures (e.g., PHQ-9 scores, suicidal ideation, depression diagnosis) remains limited.

This study seeks to address this gap by using individual-level data from multiple U.S. states to examine how state-level education and crime rates jointly affect mental health outcomes. It further investigates whether the relationship between state-level education and mental health is moderated by individual educational attainment—specifically, whether higher education buffers the negative effects of crime, or under certain conditions, may amplify psychological distress and contribute to elevated suicide risk.

2. Related Work

Although some studies have focused on the impact of state-level structural factors on mental health, the evidence on their mechanisms of action and interaction effects is still not completely consistent, and further research is still necessary.Research shows that the strength of the association between education and health varies significantly across U.S. states. A large study of approximately 1.7 million adults found that the health gap related to educational attainment differs widely by state—for example, in West Virginia, only 69% of adults without a college degree reported good health, compared to 90% of those with a college degree (a 21-point gap); whereas in Utah, the gap

was only 9 percentage points (Montez, Zajacova, & Hayward, 2022). Notably, states with the largest educational disparities also tended to have the poorest overall health outcomes (Montez et al., 2022). These differences suggest that state-level factors—such as labor market conditions, social support systems, and health policies—can either enhance or weaken the mental health benefits of educational attainment. In states with fewer resources or weaker social supports, low educational attainment may pose greater mental health risks. Conversely, states that invest in employment, income support, and health services may see smaller mental health disparities between educational groups. These findings underscore the importance of incorporating multilevel models into mental health research—models that account for both individual educational attainment and the broader state context.

Related research has also shown that even the perception of an unsafe environment can cause psychological harm. Several non-U.S. studies have found that fear of crime is associated with poorer self-rated health and higher levels of anxiety (Stafford, Chandola, & Marmot, 2007; Baranyi et al., 2021). These findings likely apply to the U.S. context as well, suggesting that both objective crime rates and subjective feelings of safety can influence mental health. Overall, this literature supports a social-ecological perspective: community safety—or the lack thereof—is a distal factor that, along with individual-level risks, shapes mental health outcomes.

3.Methods

3.1 Research Questions and Hypotheses

Research Questions

Is regional safety, measured by crime rates, negatively associated with mental health outcomes?
Is higher educational attainment positively associated with mental health outcomes?
Does educational attainment moderate the relationship between crime rates and mental health, such that higher education mitigates the negative effects of high crime rates?

Research Hypotheses

Regions with higher crime rates will have residents with poorer mental health outcomes.
Regions with higher levels of educational attainment will have residents with better mental health outcomes.
The negative impact of high crime rates on mental health will be buffered by higher levels of educational attainment.

3.2 Datasets Preparation

In order to construct an EHR database (fake data) that conforms to the OMOP model, we used code to generate 3,000 randomized data for basic EHR tables to simulate the healthcare environment of a location to begin our study. These data are stored in the Data Core and can be queried with the following commands.

```
select * from [prj0138].[CUMC\shj4017].[final_person];
select * from [prj0138].[CUMC\shj4017].[final_concept];
select * from [prj0138].[CUMC\shj4017].[final_drug_exposure];
select * from [prj0138].[CUMC\shj4017].[final_observation];
select * from [prj0138].[CUMC\shj4017].[final_measurement];
select * from [prj0138].[CUMC\shj4017].[final_condition_occurrence];
```

In addition to this, our external structural data comes primarily from four official public databases, including education data provided by the U.S. Census Bureau (U.S. Census Bureau), crime statistical trends from the FBI (Crime Data Explorer), county-level health assessments (County Health Rankings & Roadmaps), and macroeconomic data from the U.S. Bureau of Economic Analysis (BEA). These data provide important support for our analysis of the impact of state-level social structural factors on mental health. These data are also stored in Data Core after we have structured them to be queried with the following commands.

```
select * from [prj0138].[CUMC\row4007].[mental_health_data];
select * from [prj0138].[CUMC\shj4017].[Crimes_Against_Offenses];
select * from [prj0138].[CUMC\shj4017].[Annual_Summary_Statistics];
select * from [prj0138].[CUMC\shj4017].[Edu_Level_Survey];
select * from [prj0138].[CUMC\shj4017].[Edu_Level_Survey_5_years];
select * from [prj0138].[CUMC\shj4017].[Edu_Survey_des];
```

select * from [prj0138].[CUMC\shj4017].[Edu_Survey_5_years_des];

3.3 ER Diagrams of Tables

The database schema integrates both EHR data and publicly available datasets through a shared relational structure. The central entity is the final_person table, which stores individual-level demographic and administrative information, including person ID, gender, race, ethnicity, and education level. The final_person table serves as the core entity representing individual patients. Each record is uniquely identified by person_id, which is used to link all clinical event tables in a one-to-many (1:m) relationship. This design allows multiple clinical records (e.g., diagnoses, medications, measurements) to be associated with a single patient.

final_drug_exposure captures medication records, including the drug concept, dose, and time period. It is linked to final_person via the person_id foreign key, meaning each patient can have multiple drug exposures recorded over time.

final_condition_occurrence stores diagnostic data such as condition concepts, start/end dates, and condition types. This table also links to final_person through person_id, allowing a patient to have multiple diagnoses tracked across visits or years.

final_observation includes non-diagnostic clinical events like lifestyle factors, symptom reporting, or test findings not classified as formal conditions. These observations are also related to final_person via person_id.

final_measurement contains quantitative results from laboratory tests or physical exams (e.g., blood pressure, BMI, lab values). Each measurement is tied to a person using person_id, and can further be associated with a visit through visit_occurrence_id.

Each of these clinical tables is connected to final_person via a 1:m (one-to-many) relationship, allowing multiple events per individual. Additionally, medical concept IDs (e.g., drug, condition, measurement) across these tables are mapped to standardized definitions in the final_concept table, in accordance with the OMOP Common Data Model. This relational structure supports a comprehensive, person-centric view of clinical history, while maintaining interoperability and scalability in alignment with the OMOP Common Data Model.

And to support broader social and environmental health analysis, the schema also incorporates several public datasets:
mental_health_data and Crimes_Against_Offenses store county-level statistics on mental distress and crime rates, linked to final_person via the state or FIPS identifier in a many-to-many (m:m) relationship.
Annual_Summary_Statistics contains time-series data on socioeconomic indicators by GeoFips and is joined with the person-level state field.
Edu_Level_Survey and Edu_Level_Survey_5_years provide regional educational attainment data using GEO_ID and Year. These are mapped to states, allowing indirect linkage to EHR individuals.
The corresponding description tables (Edu_Survey_des, Edu_Survey_5_years_des) provide human-readable labels for each column in the education tables, facilitating interpretation.

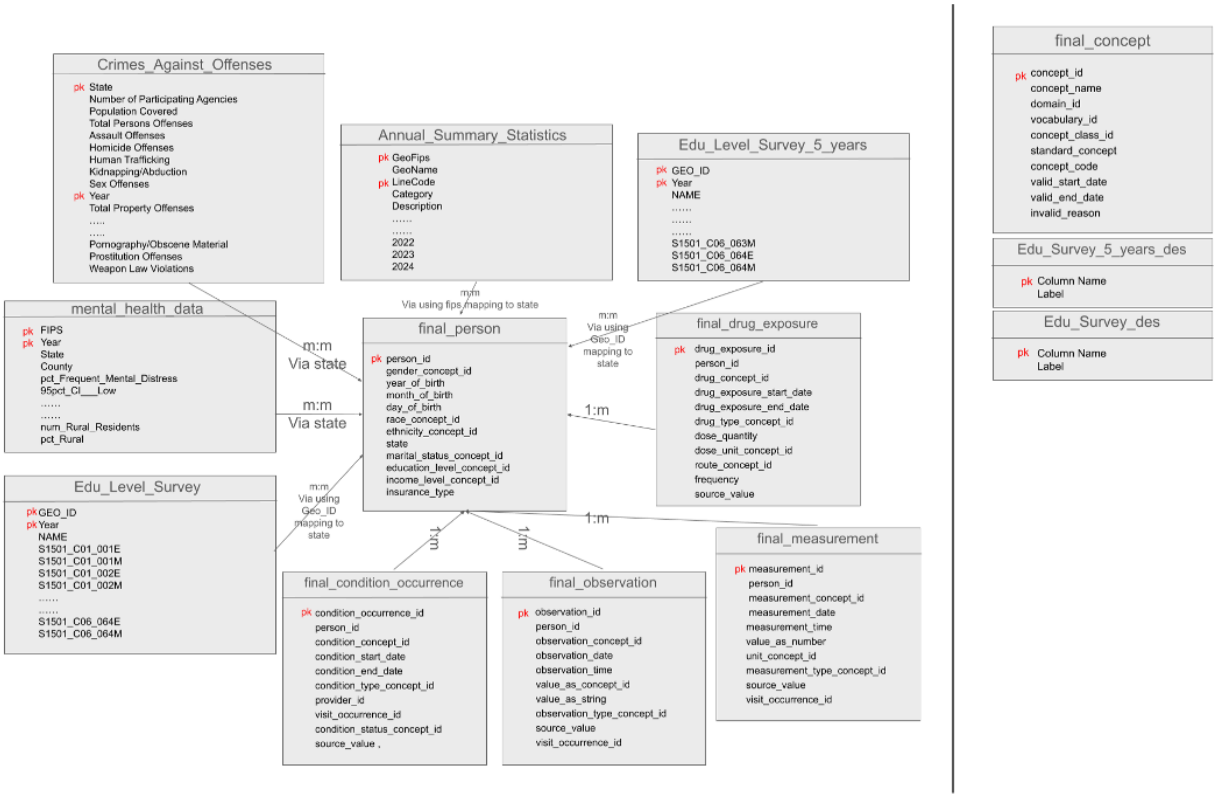


Figure 1. ER Diagrams Connecting EHR and Public Datasets

3.4 Table Schema

To provide a clearer and more intuitive understanding of the information contained in each table and their primary keys, we created a table schema. The diagram below illustrates the structure of the OMOP EHR tables.

final_person(person_id, gender_concept_id, year_of_birth, month_of_birth, day_of_birth, race_concept_id, ethnicity_concept_id, state, marital_status_concept_id, education_level_concept_id, income_level_concept_id, insurance_type)
final_drug_exposure(drug_exposure_id, person_id, drug_concept_id, drug_exposure_start_date, drug_exposure_end_date, drug_type_concept_id, dose_quantity, dose_unit_concept_id, route_concept_id, frequency, source_value)
final_condition_occurrence(condition_occurrence_id, person_id, condition_concept_id, condition_start_date, condition_end_date, condition_type_concept_id, provider_id, visit_occurrence_id, condition_status_concept_id, source_value)
final_observation(observation_id, person_id, observation_concept_id, observation_date, observation_time, value_as_concept_id, value_as_string, observation_type_concept_id, source_value, visit_occurrence_id)
final_measurement(measurement_id, person_id, measurement_concept_id, measurement_date, measurement_time, value_as_number, unit_concept_id, measurement_type_concept_id, source_value, visit_occurrence_id)
final_concept(concept_id, concept_name, domain_id, vocabulary_id, concept_class_id, standard_concept, concept_code, valid_start_date, valid_end_date, invalid_reason)

In addition, the table schemas below show the information contained in the public data tables used by our institute and the primary keys.

Annual_Summary_Statistics(GeoFips, GeoName, LineCode, Category, Description, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024)

Crimes_Against_Offenses(State, Number_of_Participating_Agencies, Population_Covered, Total_Person_Offenses, Assault_Offenses, Homicide_Offenses, Human_Trafficking, Kidnapping_Abduction, Sex_Offenses, Year, Total_Property_Offenses, Arson, Bribery, Burglary_Breaking_and_Entering, Counterfeiting_Forgery, Destruction_Damage_Vandalism, Embezzlement, Extortion_Blackmail, Fraud_Offenses, Larceny_Theft_Offenses, Motor_Vehicle_Theft, Robbery, Stolen_Property_Offenses, Total_Society_Offenses, Animal_Cruelty, Drug_Narcotic_Offenses, Gambling_Offenses, Pornography_Obscene_Material, Prostitution_Offenses, Weapon_Law_Violations)
Edu_Level_Survey(GEO_ID, NAME, S1501_C01_001E, S1501_C01_001M, S1501_C01_002E, S1501_C01_002M, ..., S1501_C06_064E, S1501_C06_064M, Year)
Edu_Level_Survey_5_years(GEO_ID, NAME, S1501_C01_001E, S1501_C01_001M, S1501_C01_002E, S1501_C01_002M, ..., S1501_C06_064E, S1501_C06_064M, Year)
Edu_Survey_des(Column_Name, Label)
Edu_Survey_5_years_des(Column_Name, Label)
mental_health_data(FIPS, Year, State, County, pct_Frequent_Mental_Distress, 95pct_CI___Low, 95pct_CI___High, num_Deaths, Suicide_Rate_Age_Adjusted, 95pct_CI___Low.1, 95pct_CI___High.1, Crude_Rate, Suicide_Rate_AIAN, Suicide_Rate_AIAN_95pct_CI___Low, Suicide_Rate_AIAN_95pct_CI___High, Suicide_Rate_Asian, Suicide_Rate_Asian_95pct_CI___Low, Suicide_Rate_Asian_95pct_CI___High, Suicide_Rate_Black, Suicide_Rate_Black_95pct_CI___Low, Suicide_Rate_Black_95pct_CI___High, Suicide_Rate_Hispanic, Suicide_Rate_Hispanic_95pct_CI___Low, Suicide_Rate_Hispanic_95pct_CI___High, Suicide_Rate_White, Suicide_Rate_White_95pct_CI___Low, Suicide_Rate_White_95pct_CI___High, Population, pct_Less_than_18_Years_of_Age, pct_65_and_Over, num_Black, pct_Black, num_American_Indian_or_Alaska_Native, pct_American_Indian_or_Alaska_Native, num_Asian, pct_Asian, num_Native_Hawaiian_or_Other_Pacific_Islander, pct_Native_Hawaiian_or_Other_Pacific_Islander, num_Hispanic, pct_Hispanic, num_Non_Hispanic_White, pct_Non_Hispanic_White, num_Not_Proficient_in_English, pct_Not_Proficient_in_English, 95pct_CI___Low.2, 95pct_CI___High.2, pct_Female, num_Rural_Residents, pct_Rural)

It is important to note that these external databases often have very complex data structures. When integrating them into our research database, we created forms in the Data Core using composite primary keys. For example, the table Annual_Summary_Statistics uses a composite key made up of GeoFips and LineCode, representing a specific type of survey conducted in a particular region. Other tables, such as mental_health_data, Crimes_Against_Offenses, Edu_Level_Survey, and Edu_Level_Survey_5_years, use composite keys consisting of FIPS, State, or GEO_ID combined with Year, representing the mental health status, crime rates, or educational attainment of a given region in a specific year.

Edu_Level_Survey and Edu_Level_Survey_5_years are two highly complex tables. In the above section, we presented a simplified version of their schema. The complete schemas are provided in the appendix. Below, we introduce the data dictionary, which we believe will give you a deeper understanding of these datasets.

3.5 Data Dictionary

In this section, we will show the data dictionary for a total of 13 tables from the EHR table and external databases. The figure below presents the data dictionary for the final_person table. It defines the structure and constraints of each field used to capture demographic and socioeconomic information, such as gender, birth date, race, ethnicity, marital status, education, income, and insurance type. These standardized fields facilitate consistent data integration and analysis across the dataset.

Field Name	Data Type	Description	Constraints	Example
person_id	INT	Unique identifier for each person.	Must be unique	1
gender_concept_id	INT	Concept ID representing the person's gender.	Not null	8507
year_of_birth	INT	Year the person was born.	Not null	2004
month_of_birth	INT	Month the person was born.		1
day_of_birth	INT	Day the person was born.		9
race_concept_id	INT	Concept ID for the person's race.		8527
ethnicity_concept_id	INT	Concept ID for the person's ethnicity.		38003564
state	VARCHAR	State abbreviation where the person resides.		OH
marital_status_concept_id	INT	Concept ID representing marital status.		9000003
education_level_concept_id	INT	Concept ID representing education level.		9000013
income_level_concept_id	INT	Concept ID representing income level.		9000020
insurance_type	VARCHAR	Type of insurance the person has.		Medicaid

Figure 2. final_person Dictionary

This figure below shows the data dictionary for the final_concept table. It defines each concept used in the database, including identifiers, names, domains, vocabularies, and classification. The table also tracks whether a concept is standard, along with its validity period and invalidation reasons, ensuring precise and consistent use of medical and demographic terminology across datasets.

Field Name	Data Type	Description	Constraints	Example
concept_id	INT	Unique identifier for a concept.	Must be unique	8507
concept_name	VARCHAR	Name of the concept.		Female
domain_id	VARCHAR	Domain to which the concept belongs.		Gender
vocabulary_id	VARCHAR	Vocabulary that the concept belongs to.		Gender
concept_class_id	VARCHAR	Class of the concept.		Gender
standard_concept	CHAR(1)	Whether it is a standard concept (S or C).		S
concept_code	VARCHAR	Code of the concept in the vocabulary.		F
valid_start_date	DATE	Start date when the concept becomes valid.		2000-01-01
valid_end_date	DATE	End date when the concept becomes invalid.		2099-12-31
invalid_reason	CHAR(1)	Reason for the concept being invalid (if any).		Has new one

Figure 3. final_concept Dictionary

This table outlines the structure of the final_condition_occurrence dataset, which captures information about diagnosed conditions. Each entry is uniquely identified and linked to a person through a person_id. It includes the condition's concept ID, its start and end dates, as well as the type and status of the condition. Additional fields record the associated healthcare provider, visit details, and the original source code. This design ensures traceability and accurate representation of clinical events within the data.

Field Name	Data Type	Description	Constraints	Example
condition_occurrence_id	INT	Unique identifier for the condition occurrence record.	Must be unique	1
person_id	INT	Person ID to whom the condition applies.	Not null	1
condition_concept_id	INT	Concept ID for the condition.	Not null	201826
condition_start_date	DATE	Start date of the condition.	Not null	10/13/2023
condition_end_date	DATE	End date of the condition.		
condition_type_concept_id	INT	Concept ID representing the type of condition.	Not null	32020
provider_id	INT	ID of the provider associated with the diagnosis.		140
visit_occurrence_id	INT	ID of the visit during which the condition was recorded.		1757
condition_status_concept_id	INT	Concept ID representing the condition's status.		2
source_value	VARCHAR	Source code of the condition from the source system.		E11

Figure 3. final_condition_occurrence Dictionary

The final_drug_exposure table is designed to record detailed information on patients' medication usage. Each row corresponds to a distinct instance of drug administration, linked to a specific individual through person_id. The record includes the standardized drug identifier (drug_concept_id), start and end dates of the exposure period, and the concept describing the type of exposure. Additional fields specify the dosage amount and its unit, the method of administration, how frequently the drug is taken, and the original drug name or code from the source system.

Field Name	Data Type	Description	Constraints	Example
drug_exposure_id	INT	Unique identifier for the drug exposure record.	Must be unique	2
person_id	INT	ID of the person receiving the drug.	Not null	2
drug_concept_id	INT	Standard concept ID of the drug.	Not null	1125315
drug_exposure_start_date	DATE	Start date of drug exposure.	Not null	2023-07-01
drug_exposure_end_date	DATE	End date of drug exposure.		2023-09-01
drug_type_concept_id	INT	Concept ID representing the type of drug exposure.	Not null	38000177
dose_quantity	FLOAT	Quantity of dose administered.		20.0
dose_unit_concept_id	INT	Concept ID representing the unit of dose.		8576
route_concept_id	INT	Concept ID for the route of administration.		4132161
frequency	VARCHAR	Frequency of drug administration.		Daily
source_value	VARCHAR	Source code or name of the drug from the source system.		Fluoxetine 20mg

Figure 4. final_drug_exosure Dictionary

The final_measurement table captures structured clinical measurement data for individuals. Each entry includes the person_id to identify the patient, and records the type of measurement via a measurement_concept_id. It logs both the date and time the measurement was taken, alongside its numeric result (value_as_number) and unit (unit_concept_id). The table also includes the measurement classification (measurement_type_concept_id), the original value from the source system, and the corresponding visit in which the measurement occurred.

Field Name	Data Type	Description	Constraints	Example
measurement_id	INT	Unique identifier for the measurement record.	Must be unique	2
person_id	INT	ID of the person receiving the measurement.	Not null	2
measurement_concept_id	INT	Concept ID for the measurement type.	Not null	44818822
measurement_date	DATE	Date the measurement was taken.	Not null	2023-07-24
measurement_time	TIME	Time the measurement was taken.		09:00:00
value_as_number	FLOAT	Numeric result of the measurement.		6
unit_concept_id	INT	Concept ID for the unit of measurement.		
measurement_type_concept_id	INT	Concept ID for the type of measurement.	Not null	44818702
source_value	VARCHAR	Original source value of the measurement.		PHQ-9
visit_occurrence_id	INT	ID of the visit during which the measurement was taken.		1363

Figure 5. final_measurement Dictionary

The final_observation table is used to store non-numeric clinical or behavioral observations linked to patients. Each record contains a unique observation_id and is associated with a specific individual via person_id. The type and value of the observation are encoded through concept IDs, with support for both coded (value_as_concept_id) and text-based (value_as_string) values. The table also records when the observation was made—both date and time—along with the type of observation, source terminology, and the related clinical visit.

Field Name	Data Type	Description	Constraints	Example
observation_id	INT	Unique identifier for the observation record.	Must be unique	2
person_id	INT	ID of the person receiving the observation.	Not null	2
observation_concept_id	INT	Concept ID representing the type of observation.	Not null	43020454
observation_date	DATE	Date the observation was recorded.	Not null	2023-07-30
observation_time	TIME	Time the observation was recorded.		10:00:00
value_as_concept_id	INT	Concept ID representing the value of the observation.		0
value_as_string	VARCHAR	String representation of the observation value.	Not null	None
observation_type_concept_id	INT	Concept ID representing the type of observation.		44814721
source_value	VARCHAR	Original source value of the observation.		Suicidal Thoughts
visit_occurrence_id	INT	ID of the visit during which the observation was made.		1564

Figure 6. final_observation Dictionary

In this section, we will show the data dictionary for a total of 13 tables from the EHR table and external databases. The mental_health_data table consolidates state-level mental health and suicide-related statistics, along with key demographic indicators. Each record is identified by a FIPS code and includes metrics such as the percentage of adults experiencing frequent mental distress, age-adjusted suicide rates across different racial and ethnic groups, and associated confidence intervals. Additionally, the table captures population counts segmented by race, ethnicity, language proficiency, and age, providing valuable context for understanding mental health disparities across regions. This rich combination of mental health outcomes and sociodemographic data supports in-depth, equity-focused public health analyses.

Field Name	Data Type	Description	Constraints	Example
FIPS	NVARCHAR(MAX)	Federal Information Processing Standards code used to identify c	Not null	1009
Year	INT	Year of data survey	Not null	2023
State	NVARCHAR(MAX)	Name of the state.	Nullable	Alabama
County	NVARCHAR(MAX)	Name of the county.	Nullable	Blount
pct_Frequent_Mental_Distress	FLOAT	Percentage of adults reporting frequent mental distress.	Nullable (numeric)	14.85927
95pct_CI__Low	FLOAT	Lower bound of 95% confidence interval for mental distress.	Nullable (numeric)	15
95pct_CI__High	FLOAT	Upper bound of 95% confidence interval for mental distress.	Nullable (numeric)	17
num_Deaths	FLOAT	Number of suicide deaths.	Nullable (numeric)	3910
Suicide_Rate_Age_Adjusted	FLOAT	Age-adjusted suicide rate per 100,000 population.	Nullable (numeric)	16
95pct_CI__Low.1	FLOAT	Lower bound of 95% CI for age-adjusted suicide rate.	Nullable (numeric)	15
95pct_CI__High.1	FLOAT	Upper bound of 95% CI for age-adjusted suicide rate.	Nullable (numeric)	16
Crude_Rate	FLOAT	Crude suicide rate.	Nullable (numeric)	16
Suicide_Rate_AIAN	FLOAT	Suicide rate among American Indian/Alaska Native population.	Nullable (numeric)	40
Suicide_Rate_AIAN_95pct_CI__Low	FLOAT	Lower CI for AIAN suicide rate.	Nullable (numeric)	35
Suicide_Rate_AIAN_95pct_CI__High	FLOAT	Upper CI for AIAN suicide rate.	Nullable (numeric)	25
Suicide_Rate_Asian	FLOAT	Suicide rate among Asian population.	Nullable (numeric)	14
Suicide_Rate_Asian_95pct_CI__Low	FLOAT	Lower CI for Asian suicide rate.	Nullable (numeric)	16
Suicide_Rate_Asian_95pct_CI__High	FLOAT	Upper CI for Asian suicide rate.	Nullable (numeric)	29
Suicide_Rate_Black	FLOAT	Suicide rate among Black/African American population.	Nullable (numeric)	48
Suicide_Rate_Black_95pct_CI__Low	FLOAT	Lower CI for Black suicide rate.	Nullable (numeric)	87
Suicide_Rate_Black_95pct_CI__High	FLOAT	Upper CI for Black suicide rate.	Nullable (numeric)	97
Suicide_Rate_Hispanic	FLOAT	Suicide rate among Hispanic population.	Nullable (numeric)	13
Suicide_Rate_Hispanic_95pct_CI__Low	FLOAT	Lower CI for Hispanic suicide rate.	Nullable (numeric)	39
Suicide_Rate_Hispanic_95pct_CI__High	FLOAT	Upper CI for Hispanic suicide rate.	Nullable (numeric)	35
Suicide_Rate_White	FLOAT	Suicide rate among White population.	Nullable (numeric)	29
Suicide_Rate_White_95pct_CI__Low	FLOAT	Lower CI for White suicide rate.	Nullable (numeric)	12
Suicide_Rate_White_95pct_CI__High	FLOAT	Upper CI for White suicide rate.	Nullable (numeric)	29
Population	NVARCHAR(MAX)	Total county population.	Nullable	55601
pct_Less_than_18_Years_of_Age	FLOAT	Percentage of population under 18.	Nullable (numeric)	23.7
pct_65_and_Over	FLOAT	Percentage of population aged 65 and older.	Nullable (numeric)	15.6
num_Black	NVARCHAR(MAX)	Number of Black/African American residents.	Nullable	10755
pct_Black	FLOAT	Percentage of Black/African American residents.	Nullable (numeric)	19.3
num_American_Indian_or_Alaska_Native	NVARCHAR(MAX)	Number of AIAN residents.	Nullable	267
pct_American_Indian_or_Alaska_Native	FLOAT	Percentage of AIAN residents.	Nullable (numeric)	0.5
num_Asian	NVARCHAR(MAX)	Number of Asian residents.	Nullable	681
pct_Asian	FLOAT	Percentage of Asian residents.	Nullable (numeric)	1.2
num_Native_Hawaiian_or_Other_Pacific_Is	NVARCHAR(MAX)	Number of NHOPI residents.	Nullable	146
pct_Native_Hawaiian_or_Other_Pacific_Isl	FLOAT	Percentage of NHOPI residents.	Nullable (numeric)	0.1
num_Hispanic	NVARCHAR(MAX)	Number of Hispanic residents.	Nullable	10131
pct_Hispanic	FLOAT	Percentage of Hispanic residents.	Nullable (numeric)	4.6
num_Non_Hispanic_White	NVARCHAR(MAX)	Number of non-Hispanic White residents.	Nullable	181201
pct_Non_Hispanic_White	FLOAT	Percentage of non-Hispanic White residents.	Nullable (numeric)	83.1
num_Not_Proficient_in_English	NVARCHAR(MAX)	Number of residents not proficient in English.	Nullable	1068
pct_Not_Proficient_in_English	FLOAT	Percentage of residents not proficient in English.	Nullable (numeric)	1
95pct_CI__Low.2	FLOAT	Lower CI for English proficiency.	Nullable (numeric)	0
95pct_CI__High.2	FLOAT	Upper CI for English proficiency.	Nullable (numeric)	1
pct_Female	FLOAT	Percentage of female population.	Nullable (numeric)	50.1
num_Rural_Residents	FLOAT	Number of rural residents.	Nullable (numeric)	150027
pct_Rural	FLOAT	Percentage of rural population.	Nullable (numeric)	41

Figure 7. mental_health_data Dictionary

The Crimes_Against_Offenses table captures detailed crime statistics reported by state-level agencies. Anchored by the state name (used in combination with the year as the primary key), the table lists the population covered and the counts of various reported offenses. These include violent crimes such as assault, sex offenses, and human trafficking, as well as property-related crimes like burglary, fraud, and embezzlement. Additional categories cover offenses such as gambling, prostitution, and animal cruelty. This dataset provides a comprehensive overview of criminal activity, supporting analyses of public safety trends across different regions.

state	VARCHAR	Name of the U.S. state.	Primary key (with year)	Alabama
number_of_participating_agencies	INT	Number of law enforcement agencies reporting in the state.		131
population_covered	FLOAT	Population covered by the reporting agencies.		715130.0
total_persons_offenses	INT	Total number of person-related offenses reported.		4384
assault_offenses	INT	Number of assault offenses.		4214
homicide_offenses	INT	Number of homicide offenses.		24
human_trafficking	INT	Number of human trafficking offenses.		0
kidnapping_abduction	INT	Number of kidnapping or abduction offenses.		52
sex_offenses	INT	Number of sex offenses.		94
year	INT	Year of the reported data.	Primary key (with state)	2020
total_property_offenses	INT	Number of bribery offenses.		7981
arson	INT	Number of burglary or breaking and entering offenses.		34
bribery	INT	Number of counterfeiting or forgery offenses.		0
burglary_breaking__entering	INT	Number of destruction, damage, or vandalism offenses.		910
counterfeiting_forgery	INT	Number of embezzlement offenses.		186
destruction_damage_vandalism	INT	Number of extortion or blackmail offenses.		1375
embezzlement	INT	Number of fraud offenses.		40
extortion_blackmail	INT	Number of larceny or theft offenses.		13
fraud_offenses	INT	Number of motor vehicle theft offenses.		820
larceny_theft_offenses	INT	Number of robbery offenses.		3840
motor_vehicle_theft	INT	Number of stolen property offenses.		541
robbery	INT	Total number of society-related offenses reported.		66
stolen_property_offenses	INT	Number of animal cruelty offenses.		156
total_society_offenses	INT	Number of drug or narcotic offenses.		3235
animal_cruelty	INT	Number of gambling offenses.		43
drug_narcotic_offenses	INT	Number of pornography or obscene material offenses.		2903
gambling_offenses	INT	Number of prostitution offenses.		1
pornography_obscene_material	INT	Number of weapon law violations.		11
prostitution_offenses	INT	Number of reported offenses of this category.		0
weapon_law_violations	INT	Number of reported offenses of this category.		277

Figure 7. Crimes_Against_Offenses Dictionary

The following is a data dictionary for both Edu_Level_Survey and Edu_Level_Survey_5_years, which share the same structure—thus, only one image is shown below. Due to the complexity of the dataset, a simplified version of the dictionary is presented here, while the full data dictionary is included in the appendix. This simplified dictionary outlines the variable names, data types, and brief descriptions, covering key information such as survey year, geographic identifiers, and education levels. It is intended to assist users in understanding the meaning and use of each column in the dataset.

Field Name	Data Type	Description	Constraints	Example
GEO_ID	VARCHAR	Geography	Primary key	0400000US
NAME	VARCHAR	Geographic Area Name		Alabama
S1501_C01_001E	INT	Estimated Total AGE BY EDUCATIONAL ATTAINMENT Population 18 to 24 years		488349
S1501_C01_001M	INT	Margin of error for Total AGE BY EDUCATIONAL ATTAINMENT Population 18 to 24 years		5182
S1501_C01_002E	FLOAT	Estimated Total AGE BY EDUCATIONAL ATTAINMENT Population 18 to 24 years Less than high school graduate		19.9
S1501_C06_019M	VARCHAR	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 35 to 44 years		0.5
S1501_C06_020E	FLOAT	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 35 to 44 years High school graduate or higher		0.5
S1501_C06_020M	FLOAT	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 35 to 44 years High school graduate (		0.5
S1501_C06_021E	FLOAT	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 35 to 44 years Bachelor's degree or higher		0.5
S1501_C06_021M	FLOAT	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 35 to 44 years Bachelor's degree or h		0.5
S1501_C06_022E	VARCHAR	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years		0.5
S1501_C06_022M	VARCHAR	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years		0.5
S1501_C06_023E	FLOAT	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years High school graduate or higher		0.5
S1501_C06_023M	FLOAT	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years High school graduate (		0.5
S1501_C06_024E	FLOAT	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years Bachelor's degree or higher		0.5
S1501_C06_024M	FLOAT	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 45 to 64 years Bachelor's degree or h		0.5
S1501_C06_025E	VARCHAR	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 65 years and over		0.5
S1501_C06_025M	VARCHAR	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 65 years and over		0.5
S1501_C06_026E	FLOAT	Estimated Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 65 years and over High school graduate or hig		0.5
S1501_C06_026M	FLOAT	Margin of error for Percent Female AGE BY EDUCATIONAL ATTAINMENT Population 65 years and over High school gradu		0.5
S1501_C06_035M	VARCHAR	Margin of error for Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT Black alone H		0.5
S1501_C06_036E	VARCHAR	Estimated Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT Black alone Bachelor's		0.5
S1501_C06_036M	VARCHAR	Margin of error for Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT Black alone B		0.5
S1501_C06_037E	VARCHAR	Estimated Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Indian or Al		0.5
S1501_C06_037M	VARCHAR	Margin of error for Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Ind		0.5
S1501_C06_038E	VARCHAR	Estimated Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Indian or Al		0.5
S1501_C06_038M	VARCHAR	Margin of error for Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Ind		0.5
S1501_C06_039E	VARCHAR	Estimated Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Indian or Al		0.5
S1501_C06_039M	VARCHAR	Margin of error for Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT American Ind		0.5
S1501_C06_040E	VARCHAR	Estimated Percent Female RACE AND HISPANIC OR LATINO ORIGIN BY EDUCATIONAL ATTAINMENT Asian alone		0.5
S1501_C06_055E	FLOAT	Estimated Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STATUS IS DE		0.5
S1501_C06_055M	FLOAT	Margin of error for Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STAT		0.5
S1501_C06_056E	FLOAT	Estimated Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STATUS IS DE		0.5
S1501_C06_056M	FLOAT	Margin of error for Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STAT		0.5
S1501_C06_057E	FLOAT	Estimated Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STATUS IS DE		0.5
S1501_C06_057M	FLOAT	Margin of error for Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STAT		0.5
S1501_C06_058E	FLOAT	Estimated Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STATUS IS DE		0.5
S1501_C06_058M	FLOAT	Margin of error for Percent Female POVERTY RATE FOR THE POPULATION 25 YEARS AND OVER FOR WHOM POVERTY STAT		0.5
S1501_C06_059E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_059M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
S1501_C06_060E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_060M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
S1501_C06_061E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_061M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
S1501_C06_062E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_062M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
S1501_C06_063E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_063M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
S1501_C06_064E	VARCHAR	Estimated Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLARS) Popul		0.5
S1501_C06_064M	VARCHAR	Margin of error for Percent Female MEDIAN EARNINGS IN THE PAST 12 MONTHS (IN 2023 INFLATION-ADJUSTED DOLLAR		0.5
Year	INT	Year of the data collected	Primary key	2010

Figure 8. Edu_Level_Survey(_*5*_years) Dictionary

Below is the data dictionary for Edu_Survey and Edu_Survey_5_years. Since both datasets share the same structure, only one table is shown here. It is structured to provide an explanation of the variables of Edu_Level_Survey.

Field Name	Data Type	Description	Constraints	Example
Column Name	VARCHAR	Name of the column or variable	Must be unique	S1501_C01_001E
Label	VARCHAR	Human-readable label or title	Not null	Estimate: Total population

Figure 9. Edu_Survey(_*5*_years)_des Dictionary

This dictionary provides metadata for each variable, including field names, data types, descriptions, and example values. Key variables include geographic identifiers (geofips), economic indicator codes (linecode), and yearly

values of economic indicators such as Real GDP, spanning from 1999 to 2023. This structure supports longitudinal economic analysis across different regions.

Field Name	Data Type	Description	Constraints	Example
geofips	INT	Geographic FIPS code (0 for the U.S. total).		0
geoname	VARCHAR	Name of the U.S. state or region.	Primary key (with linecode)	United States
linecode	INT	Code identifying the line/item of economic statistics.	Primary key (with geoname)	1
category	VARCHAR	Category of economic indicator (e.g., real dollars, current dollars).		Real dollar statistics
description	VARCHAR	Description of the economic indicator.		Real GDP (millions of chained 2017 dollars) 1
1998	VARCHAR	Value of the indicator in the year 1998.		12924876
1999	VARCHAR	Value of the indicator in the year 1999.		13543774
2000	VARCHAR	Value of the indicator in the year 2000.		14096033
2001	VARCHAR	Value of the indicator in the year 2001.		14230726
2002	VARCHAR	Value of the indicator in the year 2002.		14472712
2003	VARCHAR	Value of the indicator in the year 2003.		14877312
2004	VARCHAR	Value of the indicator in the year 2004.		15449757
2005	VARCHAR	Value of the indicator in the year 2005.		15987957
2006	VARCHAR	Value of the indicator in the year 2006.		16433148
2007	VARCHAR	Value of the indicator in the year 2007.		16762445
2008	VARCHAR	Value of the indicator in the year 2008.		16781485
2009	VARCHAR	Value of the indicator in the year 2009.		16349110
2010	VARCHAR	Value of the indicator in the year 2010.		16789750
2011	VARCHAR	Value of the indicator in the year 2011.		17052410
2012	VARCHAR	Value of the indicator in the year 2012.		17442759
2013	VARCHAR	Value of the indicator in the year 2013.		17812167
2014	VARCHAR	Value of the indicator in the year 2014.		18261714
2015	VARCHAR	Value of the indicator in the year 2015.		18799622
2016	VARCHAR	Value of the indicator in the year 2016.		19141672
2017	VARCHAR	Value of the indicator in the year 2017.		19612102
2018	VARCHAR	Value of the indicator in the year 2018.		20193896
2019	VARCHAR	Value of the indicator in the year 2019.		20715671
2020	VARCHAR	Value of the indicator in the year 2020.		20267585
2021	VARCHAR	Value of the indicator in the year 2021.		21494798
2022	VARCHAR	Value of the indicator in the year 2022.		22034828
2023	VARCHAR	Value of the indicator in the year 2023.		22671096
2024	VARCHAR	Value of the indicator in the year 2024.		23305023

Figure 10. Annual_Summary_Statistics Dictionary

3.6 Data Analysis and Visualization

In this section, we will use some maps to show the educational level of each state. The map shows the percentage of people with a bachelor's degree or above in each state. The educational level in the Northeast and the West Coast is higher. For example, the proportion of states such as Massachusetts, Colorado, and Washington is significantly higher than the national average. In comparison, the proportion of higher education in some southern states is relatively low.

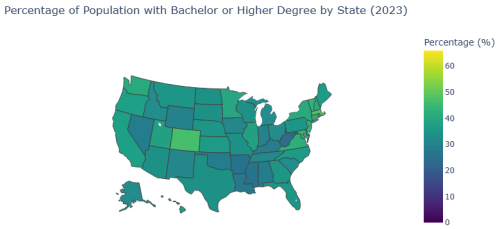


Figure 11. Percentage of Population with Bachelor or Higher Degree by State

And for the proportion of people with high school education or above, it is generally high, with most states exceeding 90%, indicating that the basic education penetration rate is good, but compared with the distribution of undergraduate and above education, the regional differences are small.

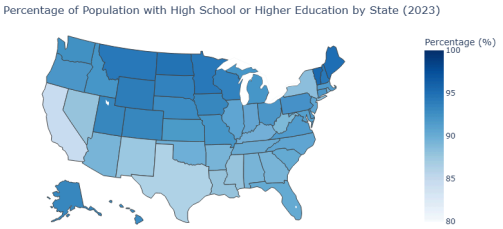


Figure 12. Percentage of Population with High School or Higher Education by State

These four maps below show the proportion of bachelor's degrees or above by age group. As can be seen from the figure, the proportion of young people (25-34 years old) with higher education is generally higher than that of the elderly (especially 65 years old and above), indicating that the penetration rate of higher education has increased in recent years.

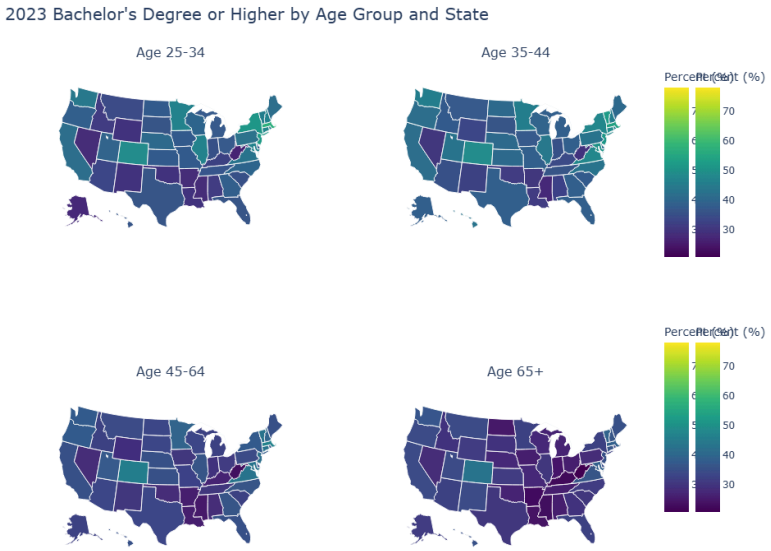


Figure 13. Bachelor's Degree or Higher by Age Group and State

This bar chart displays the top 10 states with the highest percentage of individuals holding a bachelor's degree, grouped by age categories. It clearly illustrates that younger adults aged 25–34 and 35–44 tend to have a higher percentage of bachelor's degrees compared to older age groups, particularly those aged 65 and over. The District of Columbia leads across all age groups, indicating strong overall educational attainment in the area.

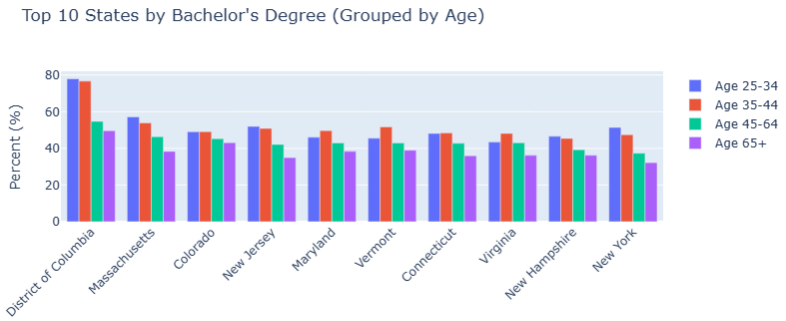


Figure 14. Top 10 States by Bachelor's Degree

Across states and age groups, bachelor's degree attainment varies notably, with the highest levels observed in the District of Columbia, Massachusetts, and Colorado. A generational pattern emerges, as younger cohorts are more likely to hold a bachelor's degree than older ones—reflecting a nationwide trend of expanding access to higher education in recent decades.

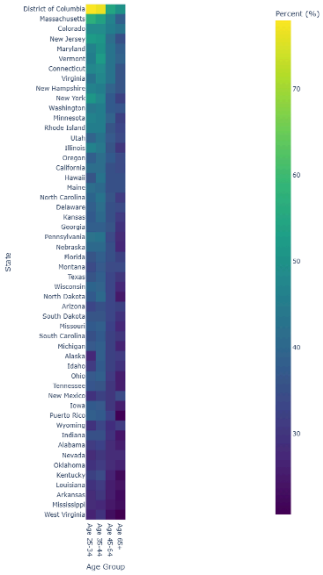


Figure 15. Bachelor's Degree or Higher by Age Group and State

Overall, there are obvious differences in the educational level of the United States between regions and generations. These differences may have an impact on variables such as mental health and crime rates in subsequent analysis, so they have important research value.

For the analysis of crime data, we still mapped these images including the overall crime rate, the distribution of different crime types, and the top ten and bottom ten states with the highest/lowest crime rates to help us with our analysis.

The map below shows the total crime rate (the number of crimes per 100,000 people) of each state. The darker the color, the higher the crime rate. It can be seen that the crime rates in the southwest (such as New Mexico and Colorado) and some states in the southeast are higher.

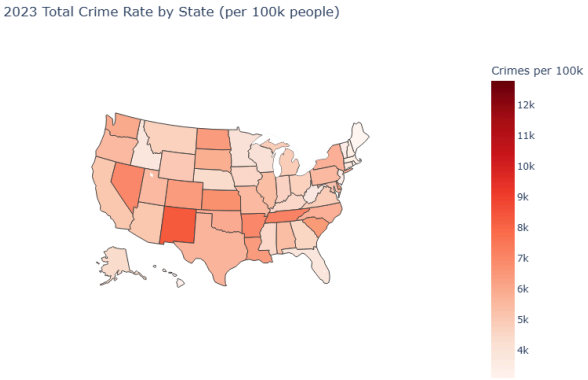


Figure 16. Total Crime Rate by State

And the three small maps below further break down the types of crimes:
Persons Crime: Mainly distributed in the west and south, such as Arizona, Nevada, etc.
Property Crime: Higher in the west and southeast.
Society Crime (social order crimes, such as drugs, gambling, etc.): Most serious in the central and southern regions.

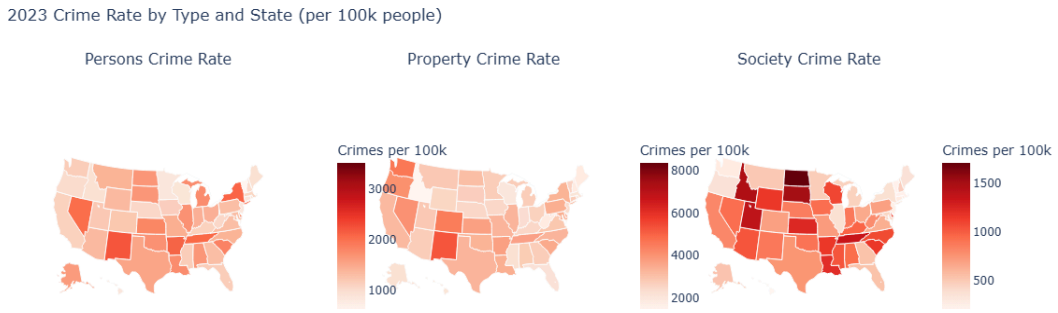


Figure 17. Crime Rate by Type and State

The bar chart in the left corner lists the top ten states with the highest crime rates. The capital Washington, DC is far ahead, with a crime rate of more than 12,000/100,000 people, followed by New Mexico, Tennessee and Nevada.The right corner shows the ten states with the lowest crime rates, which are mainly concentrated in the northeast, such as Maine, New Jersey, Connecticut, etc., with a crime rate of less than half.

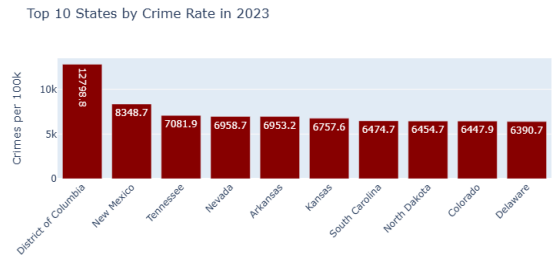


Figure 18. Top 10 States by Crime Rate

Figure 19. Bottom 10 States by Crime Rate

Overall, there are significant regional differences. The east, especially New England, is relatively safe, while the southwest and some central states have relatively poor security. This provides a geographical basis for subsequent discussions on the impact of crime rates on mental health.

Suicide makes up a large component of this data and is an indirect indication of mental health. We will draw some figures below to show the suicide rate per 100,000 people in each US state in 2023, which intuitively reflects the spatial distribution characteristics of negative mental health outcomes.

The map below uses red shades to indicate the level of suicide rates, and the darker the color, the higher the suicide rate.High-risk areas are concentrated in the western and north-central mountainous states (such as Wyoming, Montana, New Mexico, and Alaska), where the suicide rate generally exceeds 110 per 100,000 people, and the color is dark red.Low-risk areas are mostly distributed in the northeast and some coastal states (such as New Jersey, Massachusetts, and New York), and the color is lighter.This spatial heterogeneity suggests that there may be complex social structures, resource availability, and life stress factors behind the suicide rate.

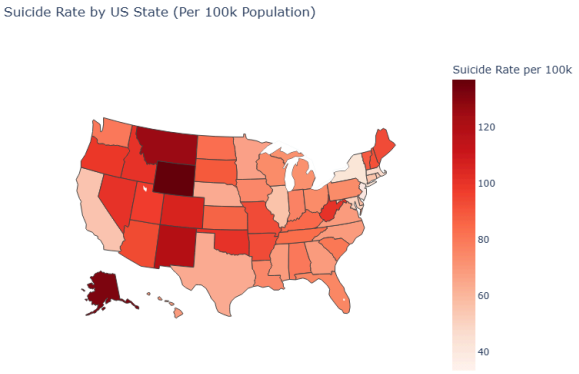


Figure 20. Suicide Rate by State (per 100k population)

The ten states with the highest suicide rates figure shows that Wyoming (136.9), Alaska (131.7) and Montana (125.2) rank in the top three, far above the national average. Such states usually have low population density, uneven resource distribution, and weak social support systems, which may increase loneliness and psychological burden. And the right figure shows the ten states with the lowest suicide rates like Washington, D.C. (33.3), New Jersey (39.9) and New York (42.5), are the lowest in the United States.These areas usually have high education levels, complete social services, and more accessible mental health resources, which may have a protective effect on individuals.

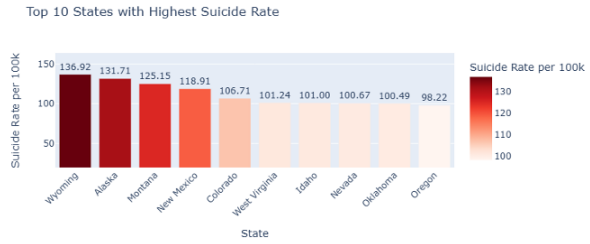


Figure 21. Top 10 States by Suicide Rate

Figure 22. Bottom 10 States by Suicide Rate

3.7 Measures

Outcome Variables:

Primary Outcome: Depressive symptom severity, measured by the individual's PHQ-9 score (treated as a continuous variable).

Secondary Outcomes: Binary indicators for (1) the presence of suicidal thoughts (derived from final_observation table, concept ID 43020454) and (2) a recorded diagnosis of depression (derived from final_condition_occurrence table, concept ID 31967).

Predictor Variables:

State-Level Safety: Violent crime rate per 100,000 population (continuous) and Property crime rate per 100,000 population (continuous) for the year 2023.

State-Level Educational Attainment: Percentage of the state population aged 25+ with a Bachelor's degree or higher in 2023 (continuous).

Individual-Level Educational Attainment: A binary variable indicating whether the individual had achieved college-level education (College/Associate degree [concept ID 9000013] or Graduate degree [concept ID 9000014]) derived from final_person.

Covariates (Individual-Level): Age (continuous, calculated as 2023 - year of birth) and Gender (categorical, based on gender_concept_id). Other available demographics (race, ethnicity, marital status, income, insurance) were noted but not included as covariates in the final reported models.

3.7 Statistical Analysis

All statistical analyses were conducted using Python (version 3.12.3) with the statsmodels, pandas, scipy, and sqlalchemy libraries. A p-value threshold of < 0.05 was used to determine statistical significance.

To address Research Question 1 (RQ1), the association between state-level crime rates (violent and property, separately) and individual PHQ-9 scores was assessed using Ordinary Least Squares (OLS) regression.

To address Research Question 2 (RQ2), the association between educational attainment and mental health was examined at both the state and individual levels. First, Pearson correlations were calculated between state-level educational attainment and state-level average PHQ-9 scores, percentage reporting suicidal thoughts, and percentage with depression diagnosis. Second, OLS regression models were used to assess the association between state-level educational attainment and individual PHQ-9 scores, controlling for individual education. An additional OLS model tested for an interaction effect between state-level and individual-level educational attainment on PHQ-9 scores. Finally, logistic regression models were used to examine the association between state-level educational attainment and the binary outcomes of suicidal thoughts and depression diagnosis.

To address Research Question 3 (RQ3) concerning the potential moderating effect of educational attainment on the relationship between crime rates and mental health outcomes (specifically suicide rates, as an indicator of severe mental health outcomes), we employed a multi-stage modeling approach using state-level aggregate data. Initial analyses utilized Ordinary Least Squares (OLS) regression to model state-level suicide rates (per 100,000 population) as a function of state-level crime rates (violent or total, centered for interpretability), state-level educational attainment (percentage of adults 25+ with Bachelor's degree or higher, centered), and their linear interaction term (crime_centered * education_centered).

Recognizing the potential for nonlinear relationships and interactions, we subsequently employed Generalized Additive Models (GAMs) using the pygam library in Python. GAMs allow for flexible, data-driven modeling of complex associations by using smoothing splines. We tested GAM specifications that included smooth terms for both the main effects of crime rate and educational attainment. To explicitly test Hypothesis 3 within this nonlinear framework, a final GAM specification incorporated a tensor product smooth interaction term (te(crime_rate, education_level) or using basis functions bs(crime_rate):bs(education_level) as indicated in the appendix) to capture potential nonlinear moderation effects. Model performance across linear and nonlinear specifications was compared using goodness-of-fit metrics such as R-squared and Root Mean Squared Error (RMSE). Diagnostic plots (residuals vs. fitted values and Q-Q plots) were generated to assess the assumptions and fit of the final selected model.

4. Results（See Appendix for more figures and analysis process）

4.1 Research Question 1: Regional Safety and Mental Health

Hypothesis 1 proposed that higher state-level crime rates would be associated with higher individual PHQ-9 scores (poorer mental health). OLS regression models with cluster-robust standard errors, controlling for individual age, gender, and college education status, were used to test this.

The association between state-level violent crime rate and individual PHQ-9 score was not statistically significant (Coef. = 0.0003, SE = 0.000, p = 0.237).

The association between state-level property crime rate and individual PHQ-9 score was also not statistically significant (Coef. = 0.0001, SE = 0.000, p = 0.197).

State-level correlation analyses (N=8 states) showed weak, non-significant positive correlations between average PHQ-9 scores and both violent (r=0.34) and property (r=0.16) crime rates. Consistent with the regression analyses, these preliminary ecological correlations provide no support for Hypothesis 1.

Overall, these findings do not support Hypothesis 1; no significant association was detected between state-level crime rates and individual depressive symptom severity in this sample.

4.2 Research Question 2: Educational Attainment and Mental Health

Hypothesis 2 proposed that higher educational attainment (state or individual) would be associated with lower PHQ-9 scores (better mental health).

State-Level Correlations: Pearson correlations based on state averages (N=8) revealed no significant associations between the state percentage of adults with a Bachelor's degree or higher and state average PHQ-9 score ($r = 0.27$, $p = 0.51$), state percentage reporting suicidal thoughts ($r = 0.04$, $p = 0.92$), or state percentage with depression diagnosis ($r = -0.11$, $p = 0.79$).

Individual-Level Regression (Main Effects): An OLS regression model predicting individual PHQ-9 scores, controlling for age and gender and including main effects for both state-level educational attainment and individual college education status (binary), found neither predictor to be statistically significant.

State Educational Attainment (% Bachelor's+): Coef. = 0.057, SE = 0.072, p = 0.429.

Individual College Education: Coef. = 0.435, SE = 0.368, p = 0.237.

Individual-Level Regression (Interaction Effect): An OLS model including an interaction term between state-level and individual-level education revealed a statistically significant negative interaction (Coef. = -0.326, SE = 0.066, $p < 0.001$). This indicates that the association between state-level educational attainment and individual PHQ-9 scores significantly differs based on the individual's own education. The negative coefficient suggests that higher state-level education is associated with lower PHQ-9 scores (better mental health) for individuals with college education, but potentially associated with higher PHQ-9 scores for those without college education.

Secondary Outcomes (Logistic Regression):

State educational attainment showed a marginally significant positive association with the odds of reporting suicidal thoughts (Coef. = 0.336, SE = 0.196, p = 0.085). No significant association was found between state educational attainment and having a depression diagnosis (Coef. = -0.006, SE = 0.013, p = 0.647).

4.3 Research Question 3: Education level buffers the impact of high crime rates on mental health

Hypothesis 3 proposed that higher educational attainment would buffer the negative impact of crime on mental health outcomes, thereby mitigating suicide risk in high-crime environments. To test this hypothesis, both linear and nonlinear models were employed using state-level data on crime rates, education levels (percent with a bachelor's degree or higher), and suicide rates.

Initial linear regression models incorporating interaction terms between crime and education did not yield a statistically significant moderation effect, failing to support Hypothesis 3 (F-test p = 0.00288; interaction term ns). While higher educational attainment alone was associated with lower suicide rates, crime rate by itself was not a significant predictor, and the linear interaction between crime and education was not statistically meaningful.

However, when generalized additive models (GAMs) were used to model nonlinear interactions, the results provided partial support for Hypothesis 3. Specifically, the interaction between crime and education became significant, and model performance improved notably (R^2 increased from 0.30 to 0.48; RMSE decreased from 18.66 to 16.05). These findings indicate that the relationship between crime, education, and suicide risk is more complex than a simple linear association and may vary depending on the levels of both predictors.

In particular, the nonlinear model showed that while higher education levels generally lower suicide rates, this protective effect is not uniform. In areas with extremely high crime, the mental health benefits of education diminish and may even reverse. Thus, while linear analysis did not support Hypothesis 3, the results from nonlinear modeling offer nuanced evidence that education moderates the crime-suicide relationship in a context-dependent manner.

5. Discussion and Conclusion

This study aimed to investigate the complex interplay between regional social determinants—specifically safety (crime rates) and educational attainment—and mental health outcomes using both individual-level (PHQ-9 scores, suicidal ideation, depression diagnosis) and state-level (suicide rates) data. Our findings highlight nuanced relationships that challenge simple, direct associations and underscore the importance of considering context and interactions.

5.1 Regional Safety and Mental Health (RQ1)

Our first hypothesis (H1), positing that higher regional crime rates are associated with poorer individual mental health (higher PHQ-9 scores), was not supported by our individual-level regression analysis. Neither state-level violent crime rates ($p = 0.237$) nor property crime rates ($p = 0.197$) showed a statistically significant association with individual PHQ-9 scores after controlling for age, gender, and individual education, using cluster-robust standard errors across 8 states (N=1833). While some literature suggests fear of crime impacts mental well-being (Stafford et al., 2007; Baranyi et al., 2021), our analysis using objective state-level crime rates did not detect a significant link to depressive symptom severity in this sample. This could be due to several factors, including the potential for PHQ-9 scores to be less sensitive to safety-related anxiety compared to broader distress, the aggregation level of crime data (state vs. neighborhood), the limited number of states in the individual-level analysis, or the possibility that other individual or community factors more strongly influence depressive symptoms than aggregate crime rates. Preliminary state-level correlations were also weak and non-significant, though limited by the small sample size (N=8) and potential ecological fallacy.

5.2 Educational Attainment and Mental Health (RQ2)

Hypothesis 2 proposed that higher educational attainment, both at the state and individual level, would be associated with better mental health outcomes (lower PHQ-9 scores). Our findings suggest a more complex relationship than hypothesized.

At the state level, no significant correlation was found between the percentage of adults with a Bachelor's degree or higher and average PHQ-9 scores ($p=0.51$), suicidal thought prevalence ($p=0.92$), or depression diagnosis prevalence ($p=0.79$). At the individual level, the main effects regression model showed no significant direct association between either state-level educational attainment ($p=0.43$) or an individual having college education ($p=0.24$) and their PHQ-9 score. However, a statistically significant negative interaction ($p < 0.001$) emerged between state-level and individual-level education. This suggests that the mental health benefits associated with living in a state with higher overall education levels are primarily accrued by individuals who themselves possess a college education (experiencing lower PHQ-9 scores in such contexts). Conversely, for individuals without college education, higher state-level education was associated with slightly higher PHQ-9 scores in this sample, potentially hinting at mechanisms like relative deprivation, though further research is needed. Furthermore, secondary analyses using logistic regression indicated a marginally significant positive association between higher state educational attainment and increased odds of reporting suicidal thoughts ($p=0.085$), a counterintuitive finding requiring cautious interpretation and validation. No significant link was found with depression diagnosis prevalence ($p=0.65$).

Therefore, H2 is not supported in its simple form. The relationship between education and mental health appears context-dependent, significantly moderated by an individual's own educational background. Higher education is not universally protective at the state level in this dataset, and its benefits may be unequally distributed.

5.3 Moderating Role of Education on Crime's Impact (RQ3)

Hypothesis 3 explored whether higher state-level educational attainment buffers the negative impact of high crime rates on suicide rates.

Linear regression models failed to find a significant interaction effect, thus not supporting H3 in a linear framework. While higher state education was independently associated with lower suicide rates, crime rate was not a significant predictor on its own, and their linear interaction was non-significant. However, Generalized Additive Models (GAMs) capable of capturing nonlinear relationships revealed a significant interaction between crime and education in predicting suicide rates. The GAM significantly outperformed the linear model (R^2 improved from 0.30 to 0.48).

The nonlinear analysis suggests that while education generally exhibits a protective effect against suicide, this effect diminishes and potentially reverses in environments with extremely high crime rates. This provides partial, nuanced support for H3, indicating that education's buffering capacity is context-dependent and may be overwhelmed by severe environmental stressors. This aligns with a social-ecological perspective where individual resources (like education) interact complexly with environmental challenges (like crime).

5.4 Synthesis

Overall, this study underscores the complexity of relationships between broad social determinants like regional crime and education, and mental health outcomes. Simple, direct associations were largely absent or weak in our analyses. State-level crime rates did not show a significant link to individual depressive symptoms (RQ1). State-level education did not demonstrate a straightforward protective effect on mental health; instead, its association with individual PHQ-9 scores was significantly moderated by personal education level, benefiting college-educated individuals more (RQ2). While linear models did not support education as a buffer against the crime-suicide link, nonlinear GAMs suggested a context-dependent buffering effect that weakens under high-crime conditions (RQ3).

These findings emphasize the need for multilevel approaches that account for both individual characteristics and environmental context, as well as potential interactions and nonlinearities, when studying the social determinants of mental health. Relying solely on ecological correlations or simple linear models may obscure important nuances.

6. Limitations

Several limitations should be acknowledged. The study's cross-sectional design prevents causal inferences. The individual-level analysis was limited to 8 states, potentially restricting generalizability and statistical power for detecting state-level effects. The range of indicators was limited; other factors like income inequality, healthcare access, social support, and perceived safety were not included but could confound or mediate the observed relationships. The use of aggregate state-level data for crime and education may mask significant within-state heterogeneity. Finally, potential unmeasured confounders at both individual and state levels could influence the results.

7. Future directions

Future research should aim to address these limitations. Incorporating longitudinal data would allow for stronger causal assessment. Expanding the geographic scope to include more states and potentially finer geographic units (e.g., counties) would enhance generalizability and allow for investigation of regional variations. Including a broader set of individual and contextual indicators, particularly measures of perceived safety, social cohesion, and economic factors, would provide a more comprehensive model. Subgroup analyses by demographics (age, gender,

race/ethnicity) could reveal differential impacts. Further exploration of the nonlinear interaction between crime and education using advanced statistical techniques is warranted, as is investigation into the mechanisms underlying the observed interaction between state and individual education levels on depressive symptoms and the counterintuitive finding regarding suicidal thoughts.

Appendix

1. Full Table Schemas for Education Level Survey tables

Edu_Level_Survey(_5_years)(GEO_ID, NAME, S1501_C01_001E, S1501_C01_001M, S1501_C01_002E, S1501_C01_002M, S1501_C01_003E, S1501_C01_003M, S1501_C01_004E, S1501_C01_004M, S1501_C01_005E, S1501_C01_005M, S1501_C01_006E, S1501_C01_006M, S1501_C01_007E, S1501_C01_007M, S1501_C01_008E, S1501_C01_008M, S1501_C01_009E, S1501_C01_009M, S1501_C01_010E, S1501_C01_010M, S1501_C01_011E, S1501_C01_011M, S1501_C01_012E, S1501_C01_012M, S1501_C01_013E, S1501_C01_013M, S1501_C01_014E, S1501_C01_014M, S1501_C01_015E, S1501_C01_015M, S1501_C01_016E, S1501_C01_016M, S1501_C01_017E, S1501_C01_017M, S1501_C01_018E, S1501_C01_018M, S1501_C01_019E, S1501_C01_019M, S1501_C01_020E, S1501_C01_020M, S1501_C01_021E, S1501_C01_021M, S1501_C01_022E, S1501_C01_022M, S1501_C01_023E, S1501_C01_023M, S1501_C01_024E, S1501_C01_024M, S1501_C01_025E, S1501_C01_025M, S1501_C01_026E, S1501_C01_026M, S1501_C01_027E, S1501_C01_027M, S1501_C01_028E, S1501_C01_028M, S1501_C01_029E, S1501_C01_029M, S1501_C01_030E, S1501_C01_030M, S1501_C01_031E, S1501_C01_031M, S1501_C01_032E, S1501_C01_032M, S1501_C01_033E, S1501_C01_033M, S1501_C01_034E, S1501_C01_034M, S1501_C01_035E, S1501_C01_035M, S1501_C01_036E, S1501_C01_036M, S1501_C01_037E, S1501_C01_037M, S1501_C01_038E, S1501_C01_038M, S1501_C01_039E, S1501_C01_039M, S1501_C01_040E, S1501_C01_040M, S1501_C01_041E, S1501_C01_041M, S1501_C01_042E, S1501_C01_042M, S1501_C01_043E, S1501_C01_043M, S1501_C01_044E, S1501_C01_044M, S1501_C01_045E, S1501_C01_045M, S1501_C01_046E, S1501_C01_046M, S1501_C01_047E, S1501_C01_047M, S1501_C01_048E, S1501_C01_048M, S1501_C01_049E, S1501_C01_049M, S1501_C01_050E, S1501_C01_050M, S1501_C01_051E, S1501_C01_051M, S1501_C01_052E, S1501_C01_052M, S1501_C01_053E, S1501_C01_053M, S1501_C01_054E, S1501_C01_054M, S1501_C01_055E, S1501_C01_055M, S1501_C01_056E, S1501_C01_056M, S1501_C01_057E, S1501_C01_057M, S1501_C01_058E, S1501_C01_058M, S1501_C01_059E, S1501_C01_059M, S1501_C01_060E, S1501_C01_060M, S1501_C01_061E, S1501_C01_061M, S1501_C01_062E, S1501_C01_062M, S1501_C01_063E, S1501_C01_063M, S1501_C01_064E, S1501_C01_064M, S1501_C02_001E, S1501_C02_001M, S1501_C02_002E, S1501_C02_002M, S1501_C02_003E, S1501_C02_003M, S1501_C02_004E, S1501_C02_004M, S1501_C02_005E, S1501_C02_005M, S1501_C02_006E, S1501_C02_006M, S1501_C02_007E, S1501_C02_007M, S1501_C02_008E, S1501_C02_008M, S1501_C02_009E, S1501_C02_009M, S1501_C02_010E, S1501_C02_010M, S1501_C02_011E, S1501_C02_011M, S1501_C02_012E, S1501_C02_012M, S1501_C02_013E, S1501_C02_013M, S1501_C02_014E, S1501_C02_014M, S1501_C02_015E, S1501_C02_015M, S1501_C02_016E, S1501_C02_016M, S1501_C02_017E, S1501_C02_017M, S1501_C02_018E, S1501_C02_018M, S1501_C02_019E, S1501_C02_019M, S1501_C02_020E, S1501_C02_020M, S1501_C02_021E, S1501_C02_021M, S1501_C02_022E, S1501_C02_022M, S1501_C02_023E, S1501_C02_023M, S1501_C02_024E, S1501_C02_024M, S1501_C02_025E, S1501_C02_025M, S1501_C02_026E, S1501_C02_026M, S1501_C02_027E, S1501_C02_027M, S1501_C02_028E, S1501_C02_028M, S1501_C02_029E, S1501_C02_029M, S1501_C02_030E, S1501_C02_030M, S1501_C02_031E, S1501_C02_031M, S1501_C02_032E, S1501_C02_032M, S1501_C02_033E, S1501_C02_033M, S1501_C02_034E, S1501_C02_034M, S1501_C02_035E, S1501_C02_035M, S1501_C02_036E, S1501_C02_036M, S1501_C02_037E, S1501_C02_037M, S1501_C02_038E, S1501_C02_038M, S1501_C02_039E, S1501_C02_039M, S1501_C02_040E, S1501_C02_040M, S1501_C02_041E, S1501_C02_041M, S1501_C02_042E, S1501_C02_042M, S1501_C02_043E, S1501_C02_043M, S1501_C02_044E, S1501_C02_044M, S1501_C02_045E, S1501_C02_045M, S1501_C02_046E, S1501_C02_046M, S1501_C02_047E, S1501_C02_047M, S1501_C02_048E, S1501_C02_048M, S1501_C02_049E, S1501_C02_049M, S1501_C02_050E, S1501_C02_050M, S1501_C02_051E, S1501_C02_051M, S1501_C02_052E, S1501_C02_052M, S1501_C02_053E, S1501_C02_053M, S1501_C02_054E, S1501_C02_054M, S1501_C02_055E, S1501_C02_055M, S1501_C02_056E, S1501_C02_056M, S1501_C02_057E, S1501_C02_057M, S1501_C02_058E, S1501_C02_058M, S1501_C02_059E, S1501_C02_059M, S1501_C02_060E, S1501_C02_060M, S1501_C02_061E, S1501_C02_061M, S1501_C02_062E, S1501_C02_062M, S1501_C02_063E, S1501_C02_063M, S1501_C02_064E, S1501_C02_064M, S1501_C03_001E, S1501_C03_001M, S1501_C03_002E, S1501_C03_002M, S1501_C03_003E, S1501_C03_003M, S1501_C03_004E, S1501_C03_004M, S1501_C03_005E, S1501_C03_005M, S1501_C03_006E, S1501_C03_006M, S1501_C03_007E, S1501_C03_007M, S1501_C03_008E, S1501_C03_008M, S1501_C03_009E, S1501_C03_009M, S1501_C03_010E, S1501_C03_010M, S1501_C03_011E, S1501_C03_011M, S1501_C03_012E, S1501_C03_012M, S1501_C03_013E, S1501_C03_013M, S1501_C03_014E, S1501_C03_014M, S1501_C03_015E, S1501_C03_015M, S1501_C03_016E, S1501_C03_016M, S1501_C03_017E, S1501_C03_017M, S1501_C03_018E, S1501_C03_018M, S1501_C03_019E, S1501_C03_019M, S1501_C03_020E, S1501_C03_020M, S1501_C03_021E, S1501_C03_021M, S1501_C03_022E, S1501_C03_022M, S1501_C03_023E, S1501_C03_023M, S1501_C03_024E, S1501_C03_024M, S1501_C03_025E, S1501_C03_025M, S1501_C03_026E, S1501_C03_026M, S1501_C03_027E, S1501_C03_027M, S1501_C03_028E, S1501_C03_028M, S1501_C03_029E, S1501_C03_029M, S1501_C03_030E, S1501_C03_030M, S1501_C03_031E, S1501_C03_031M, S1501_C03_032E, S1501_C03_032M, S1501_C03_033E, S1501_C03_033M, S1501_C03_034E, S1501_C03_034M, S1501_C03_035E, S1501_C03_035M, S1501_C03_036E, S1501_C03_036M, S1501_C03_037E, S1501_C03_037M, S1501_C03_038E, S1501_C03_038M, S1501_C03_039E, S1501_C03_039M, S1501_C03_040E, S1501_C03_040M, S1501_C03_041E, S1501_C03_041M, S1501_C03_042E, S1501_C03_042M, S1501_C03_043E, S1501_C03_043M, S1501_C03_044E, S1501_C03_044M, S1501_C03_045E, S1501_C03_045M, S1501_C03_046E, S1501_C03_046M, S1501_C03_047E, S1501_C03_047M, S1501_C03_048E, S1501_C03_048M, S1501_C03_049E, S1501_C03_049M, S1501_C03_050E, S1501_C03_050M, S1501_C03_051E, S1501_C03_051M, S1501_C03_052E, S1501_C03_052M, S1501_C03_053E, S1501_C03_053M, S1501_C03_054E, S1501_C03_054M, S1501_C03_055E, S1501_C03_055M, S1501_C03_056E, S1501_C03_056M, S1501_C03_057E, S1501_C03_057M, S1501_C03_058E, S1501_C03_058M, S1501_C03_059E, S1501_C03_059M, S1501_C03_060E, S1501_C03_060M, S1501_C03_061E, S1501_C03_061M, S1501_C03_062E, S1501_C03_062M, S1501_C03_063E, S1501_C03_063M, S1501_C03_064E, S1501_C03_064M, S1501_C04_001E, S1501_C04_001M, S1501_C04_002E, S1501_C04_002M, S1501_C04_003E, S1501_C04_003M, S1501_C04_004E, S1501_C04_004M, S1501_C04_005E, S1501_C04_005M, S1501_C04_006E, S1501_C04_006M, S1501_C04_007E, S1501_C04_007M, S1501_C04_008E, S1501_C04_008M, S1501_C04_009E, S1501_C04_009M, S1501_C04_010E, S1501_C04_010M, S1501_C04_011E, S1501_C04_011M, S1501_C04_012E, S1501_C04_012M, S1501_C04_013E, S1501_C04_013M, S1501_C04_014E, S1501_C04_014M, S1501_C04_015E, S1501_C04_015M, S1501_C04_016E, S1501_C04_016M, S1501_C04_017E, S1501_C04_017M, S1501_C04_018E, S1501_C04_018M, S1501_C04_019E, S1501_C04_019M, S1501_C04_020E, S1501_C04_020M, S1501_C04_021E, S1501_C04_021M, S1501_C04_022E, S1501_C04_022M, S1501_C04_023E, S1501_C04_023M, S1501_C04_024E, S1501_C04_024M, S1501_C04_025E, S1501_C04_025M, S1501_C04_026E, S1501_C04_026M, S1501_C04_027E, S1501_C04_027M, S1501_C04_028E, S1501_C04_028M, S1501_C04_029E, S1501_C04_029M, S1501_C04_030E, S1501_C04_030M, S1501_C04_031E, S1501_C04_031M, S1501_C04_032E, S1501_C04_032M, S1501_C04_033E, S1501_C04_033M, S1501_C04_034E, S1501_C04_034M, S1501_C04_035E, S1501_C04_035M, S1501_C04_036E, S1501_C04_036M, S1501_C04_037E, S1501_C04_037M, S1501_C04_038E, S1501_C04_038M, S1501_C04_039E, S1501_C04_039M, S1501_C04_040E, S1501_C04_040M, S1501_C04_041E, S1501_C04_041M, S1501_C04_042E, S1501_C04_042M, S1501_C04_043E, S1501_C04_043M, S1501_C04_044E, S1501_C04_044M, S1501_C04_045E, S1501_C04_045M, S1501_C04_046E, S1501_C04_046M, S1501_C04_047E, S1501_C04_047M, S1501_C04_048E, S1501_C04_048M, S1501_C04_049E, S1501_C04_049M, S1501_C04_050E, S1501_C04_050M, S1501_C04_051E, S1501_C04_051M, S1501_C04_052E, S1501_C04_052M, S1501_C04_053E, S1501_C04_053M, S1501_C04_054E, S1501_C04_054M, S1501_C04_055E, S1501_C04_055M, S1501_C04_056E, S1501_C04_056M, S1501_C04_057E, S1501_C04_057M, S1501_C04_058E, S1501_C04_058M, S1501_C04_059E, S1501_C04_059M, S1501_C04_060E, S1501_C04_060M, S1501_C04_061E, S1501_C04_061M, S1501_C04_062E, S1501_C04_062M, S1501_C04_063E, S1501_C04_063M, S1501_C04_064E, S1501_C04_064M, S1501_C05_001E, S1501_C05_001M, S1501_C05_002E, S1501_C05_002M, S1501_C05_003E, S1501_C05_003M, S1501_C05_004E, S1501_C05_004M, S1501_C05_005E, S1501_C05_005M, S1501_C05_006E, S1501_C05_006M, S1501_C05_007E, S1501_C05_007M, S1501_C05_008E, S1501_C05_008M, S1501_C05_009E, S1501_C05_009M, S1501_C05_010E, S1501_C05_010M, S1501_C05_011E, S1501_C05_011M, S1501_C05_012E, S1501_C05_012M, S1501_C05_013E, S1501_C05_013M, S1501_C05_014E, S1501_C05_014M, S1501_C05_015E, S1501_C05_015M, S1501_C05_016E, S1501_C05_016M, S1501_C05_017E, S1501_C05_017M, S1501_C05_018E, S1501_C05_018M, S1501_C05_019E, S1501_C05_019M, S1501_C05_020E, S1501_C05_020M, S1501_C05_021E, S1501_C05_021M, S1501_C05_022E, S1501_C05_022M, S1501_C05_023E, S1501_C05_023M, S1501_C05_024E, S1501_C05_024M, S1501_C05_025E, S1501_C05_025M, S1501_C05_026E, S1501_C05_026M, S1501_C05_027E, S1501_C05_027M, S1501_C05_028E, S1501_C05_028M, S1501_C05_029E, S1501_C05_029M, S1501_C05_030E, S1501_C05_030M, S1501_C05_031E, S1501_C05_031M, S1501_C05_032E, S1501_C05_032M, S1501_C05_033E, S1501_C05_033M, S1501_C05_034E, S1501_C05_034M, S1501_C05_035E, S1501_C05_035M, S1501_C05_036E, S1501_C05_036M, S1501_C05_037E, S1501_C05_037M, S1501_C05_038E, S1501_C05_038M, S1501_C05_039E, S1501_C05_039M, S1501_C05_040E, S1501_C05_040M, S1501_C05_041E, S1501_C05_041M, S1501_C05_042E, S1501_C05_042M, S1501_C05_043E, S1501_C05_043M, S1501_C05_044E, S1501_C05_044M, S1501_C05_045E, S1501_C05_045M, S1501_C05_046E, S1501_C05_046M, S1501_C05_047E, S1501_C05_047M, S1501_C05_048E, S1501_C05_048M, S1501_C05_049E, S1501_C05_049M, S1501_C05_050E, S1501_C05_050M, S1501_C05_051E, S1501_C05_051M, S1501_C05_052E, S1501_C05_052M, S1501_C05_053E, S1501_C05_053M, S1501_C05_054E, S1501_C05_054M, S1501_C05_055E, S1501_C05_055M, S1501_C05_056E, S1501_C05_056M, S1501_C05_057E, S1501_C05_057M, S1501_C05_058E, S1501_C05_058M, S1501_C05_059E, S1501_C05_059M, S1501_C05_060E, S1501_C05_060M, S1501_C05_061E, S1501_C05_061M, S1501_C05_062E, S1501_C05_062M, S1501_C05_063E, S1501_C05_063M, S1501_C05_064E, S1501_C05_064M, S1501_C06_001E, S1501_C06_001M, S1501_C06_002E, S1501_C06_002M, S1501_C06_003E, S1501_C06_003M, S1501_C06_004E, S1501_C06_004M, S1501_C06_005E, S1501_C06_005M, S1501_C06_006E, S1501_C06_006M, S1501_C06_007E, S1501_C06_007M, S1501_C06_008E, S1501_C06_008M, S1501_C06_009E, S1501_C06_009M, S1501_C06_010E, S1501_C06_010M, S1501_C06_011E, S1501_C06_011M, S1501_C06_012E, S1501_C06_012M, S1501_C06_013E, S1501_C06_013M, S1501_C06_014E, S1501_C06_014M, S1501_C06_015E, S1501_C06_015M, S1501_C06_016E, S1501_C06_016M, S1501_C06_017E, S1501_C06_017M, S1501_C06_018E, S1501_C06_018M, S1501_C06_019E, S1501_C06_019M, S1501_C06_020E, S1501_C06_020M, S1501_C06_021E, S1501_C06_021M, S1501_C06_022E, S1501_C06_022M, S1501_C06_023E, S1501_C06_023M, S1501_C06_024E, S1501_C06_024M, S1501_C06_025E, S1501_C06_025M, S1501_C06_026E, S1501_C06_026M, S1501_C06_027E, S1501_C06_027M, S1501_C06_028E, S1501_C06_028M, S1501_C06_029E, S1501_C06_029M, S1501_C06_030E, S1501_C06_030M, S1501_C06_031E, S1501_C06_031M, S1501_C06_032E, S1501_C06_032M, S1501_C06_033E, S1501_C06_033M, S1501_C06_034E, S1501_C06_034M, S1501_C06_035E, S1501_C06_035M, S1501_C06_036E, S1501_C06_036M, S1501_C06_037E, S1501_C06_037M, S1501_C06_038E, S1501_C06_038M, S1501_C06_039E, S1501_C06_039M, S1501_C06_040E, S1501_C06_040M, S1501_C06_041E, S1501_C06_041M, S1501_C06_042E, S1501_C06_042M, S1501_C06_043E, S1501_C06_043M, S1501_C06_044E, S1501_C06_044M, S1501_C06_045E, S1501_C06_045M, S1501_C06_046E, S1501_C06_046M, S1501_C06_047E, S1501_C06_047M, S1501_C06_048E, S1501_C06_048M, S1501_C06_049E, S1501_C06_049M, S1501_C06_050E, S1501_C06_050M, S1501_C06_051E, S1501_C06_051M, S1501_C06_052E, S1501_C06_052M, S1501_C06_053E, S1501_C06_053M, S1501_C06_054E, S1501_C06_054M, S1501_C06_055E, S1501_C06_055M, S1501_C06_056E, S1501_C06_056M, S1501_C06_057E, S1501_C06_057M, S1501_C06_058E, S1501_C06_058M, S1501_C06_059E, S1501_C06_059M, S1501_C06_060E, S1501_C06_060M, S1501_C06_061E, S1501_C06_061M, S1501_C06_062E, S1501_C06_062M, S1501_C06_063E, S1501_C06_063M, S1501_C06_064E, S1501_C06_064M, Year)

2. Full Data Dictionary for Education Level Survey tables

You can find the Education Level Survey data dictionary through this link:

https://docs.google.com/spreadsheets/d/1xfE6IOZzx7dS2OwOwcM9sztu5wJ56982zFUPOIGgIEI/edit?usp=sharing

3. Hypothese 1

```
OLS Results with Cluster-Robust SE (Violent Crime -> PHQ-9):
                            OLS Regression Results
==============================================================================
Dep. Variable:             phq9_score   R-squared:                       0.002
Model:                            OLS   Adj. R-squared:                 -0.001
Method:                 Least Squares   F-statistic:                    0.5278
Date:                Tue, 08 Apr 2025   Prob (F-statistic):              0.720
Time:                        11:24:14   Log-Likelihood:                -6421.4
No. Observations:                1833   AIC:                          1.285e+04
Df Residuals:                    1828   BIC:                          1.288e+04
Df Model:                           4
Covariance Type:              cluster
================================================================================================
                                   coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------
Intercept                       12.9995      0.579     22.441      0.000      11.864      14.135
C(gender_concept_id)[T.8532]     0.4217      0.563      0.749      0.454      -0.682       1.525
state_violent_crime              0.0003      0.000      1.183      0.237      -0.000       0.001
age                             -0.0062      0.012     -0.503      0.615      -0.030       0.018
has_college_education            0.4228      0.362      1.168      0.243      -0.287       1.132
==============================================================================
Omnibus:                     1362.587   Durbin-Watson:                   1.946
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              108.483
Skew:                          -0.003   Prob(JB):                     2.77e-24
Kurtosis:                       1.808   Cond. No.                     8.66e+03
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 8.66e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

OLS Results with Cluster-Robust SE (Property Crime -> PHQ-9):

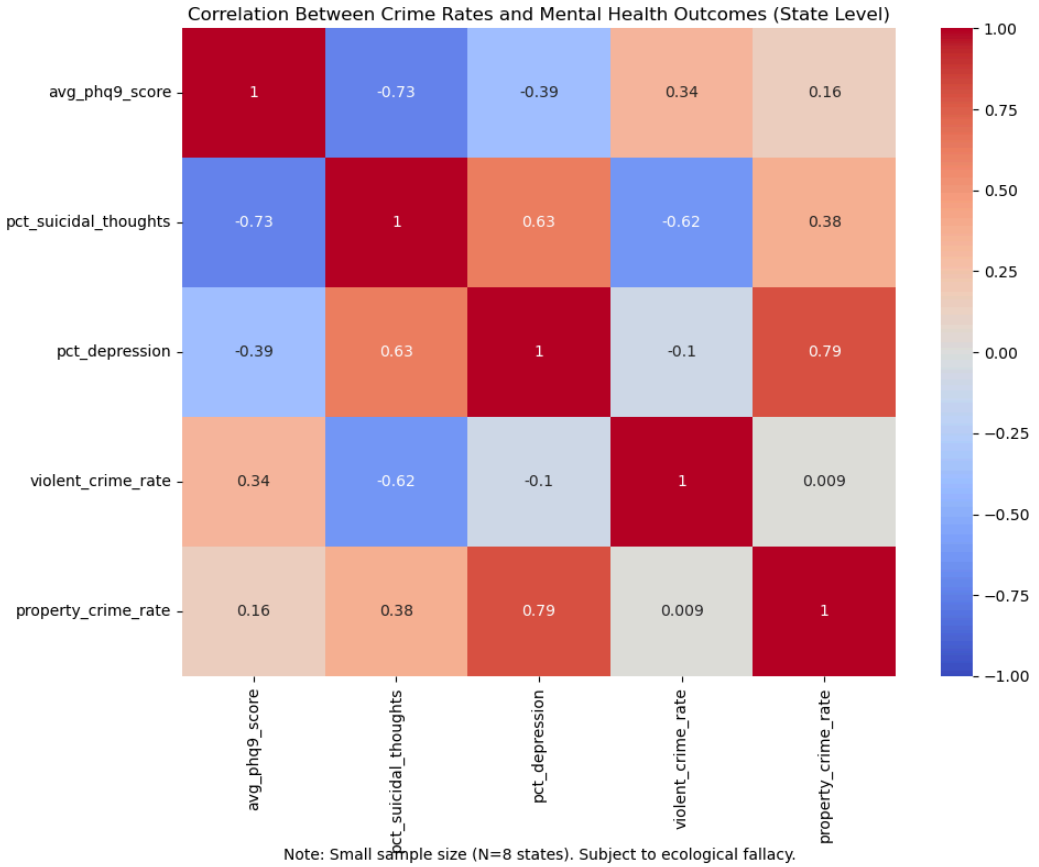
OLS Regression Results

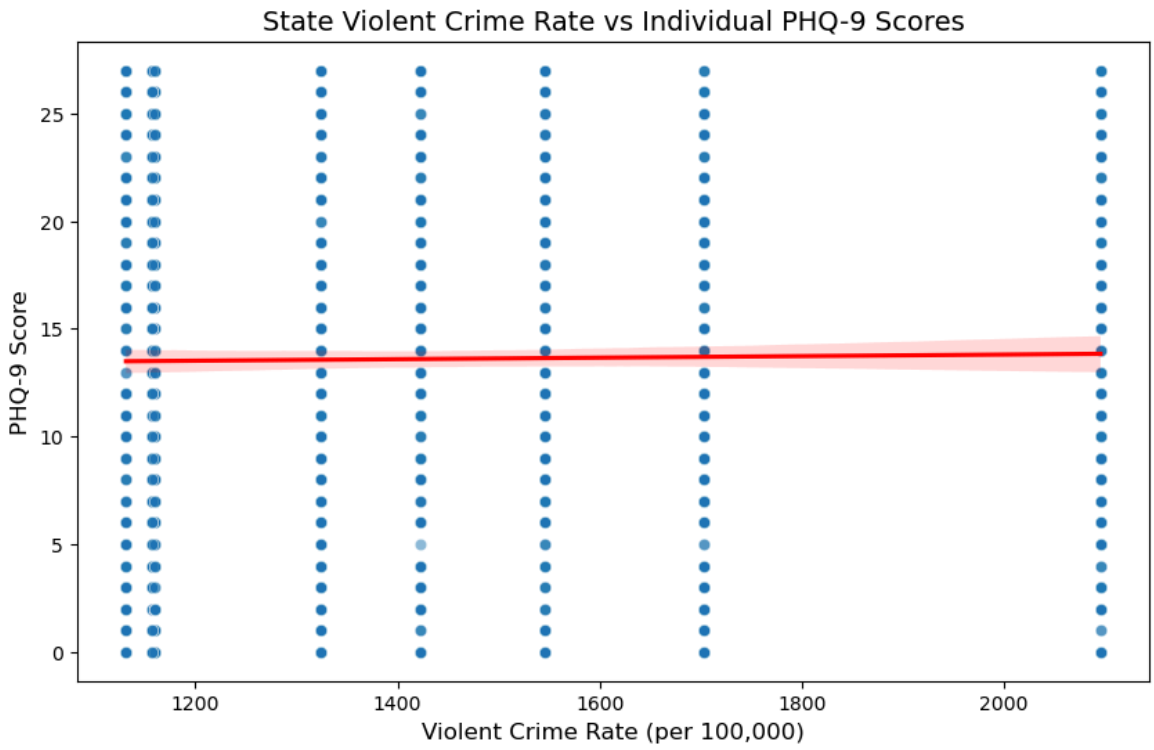
Dep. Variable:	phq9_score	R-squared:	0.002
Model:	OLS	Adj. R-squared:	-0.001
Method:	Least Squares	F-statistic:	1.988
Date:	Tue, 08 Apr 2025	Prob (F-statistic):	0.201
Time:	11:24:14	Log-Likelihood:	-6421.5
No. Observations:	1833	AIC:	1.285e+04
Df Residuals:	1828	BIC:	1.288e+04
Df Model:	4		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	13.1632	0.538	24.464	0.000	12.109	14.218
C(gender_concept_id)[T.8532]	0.4249	0.560	0.759	0.448	-0.672	1.522
state_property_crime	9.865e-05	7.65e-05	1.290	0.197	-5.12e-05	0.000
age	-0.0065	0.012	-0.537	0.592	-0.030	0.017
has_college_education	0.4319	0.364	1.185	0.236	-0.282	1.146

Omnibus:	1363.889	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	108.508
Skew:	-0.005	Prob(JB):	2.74e-24
Kurtosis:	1.808	Cond. No.	2.00e+04

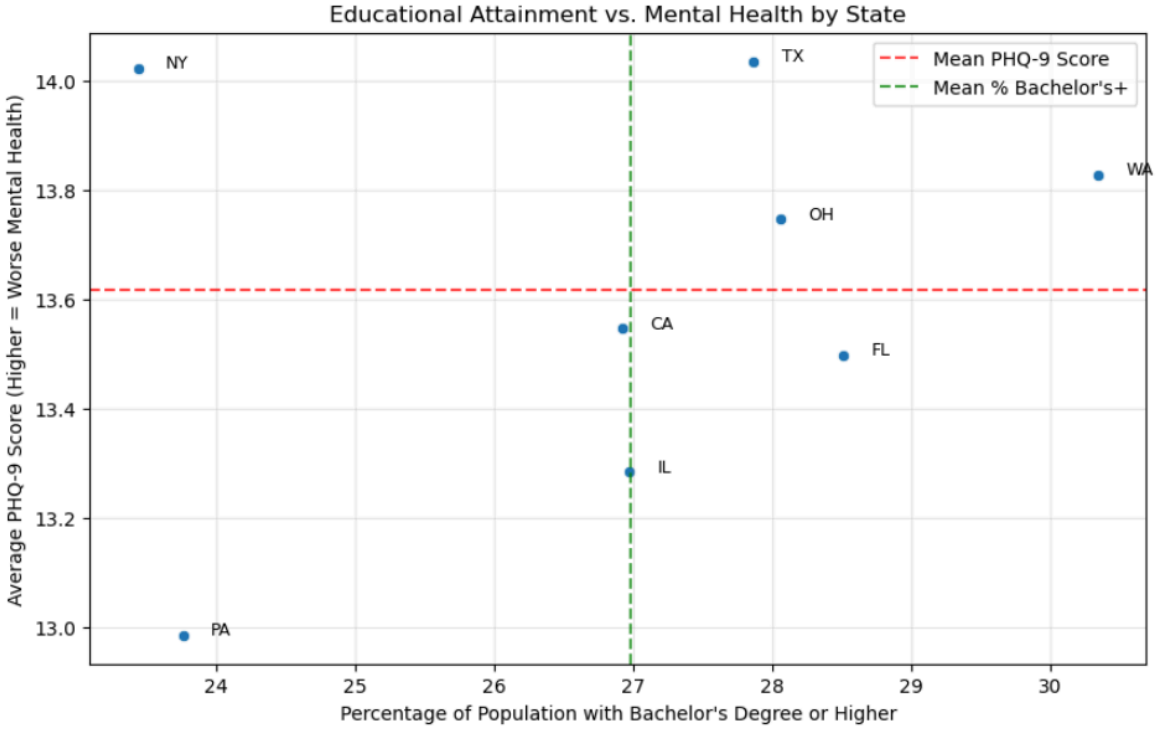
Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
[2] The condition number is large, 2e+04. This might indicate that there are strong multicollinearity or other numerical problems.





4. Hypothese 2

```
State-Level Correlations (Educational Attainment vs. Mental Health):
PHQ-9 Score vs. % Bachelor's or higher: r = 0.274, p = 0.511
% Suicidal Thoughts vs. % Bachelor's or higher: r = 0.043, p = 0.920
% Depression vs. % Bachelor's or higher: r = -0.111, p = 0.793
```



Regression Results (State Education Level -> PHQ-9):

OLS Regression Results

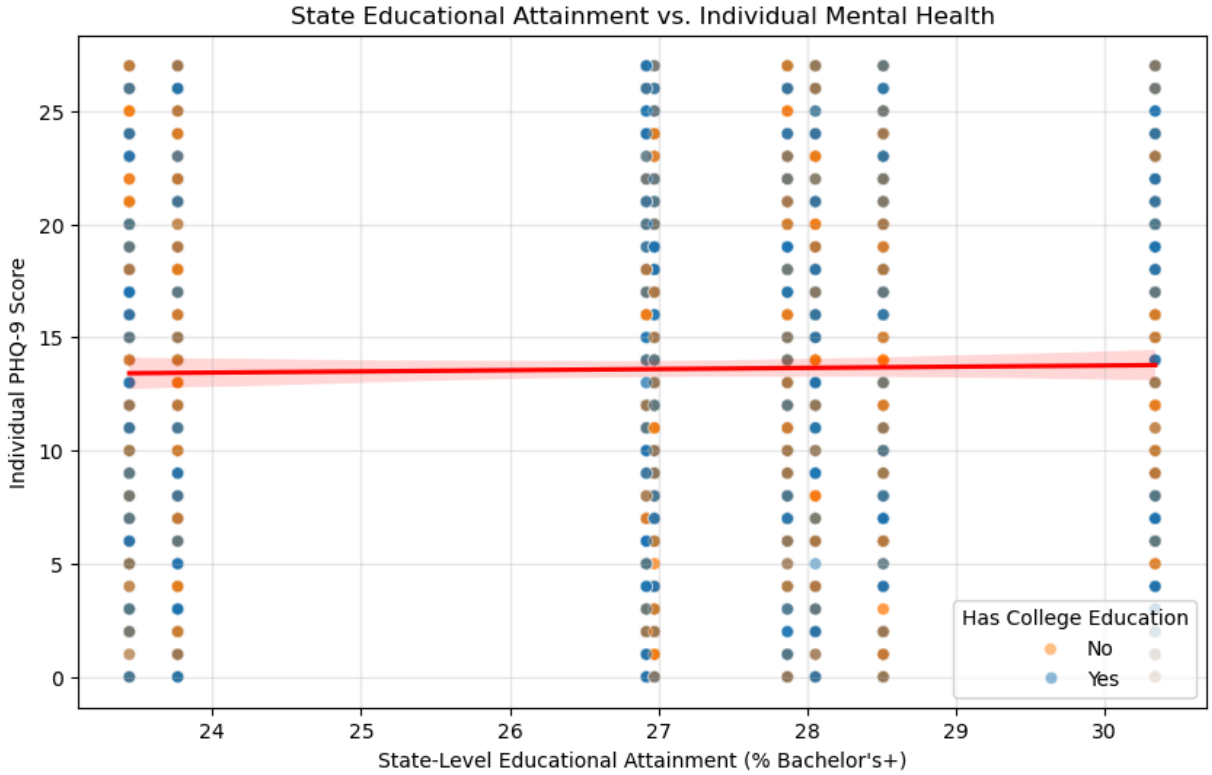
Dep. Variable:	phq9_score	R-squared:	0.002
Model:	OLS	Adj. R-squared:	-0.000
Method:	Least Squares	F-statistic:	0.5386
Date:	Tue, 08 Apr 2025	Prob (F-statistic):	0.713
Time:	11:24:16	Log-Likelihood:	-6421.3
No. Observations:	1833	AIC:	1.285e+04
Df Residuals:	1828	BIC:	1.288e+04
Df Model:	4		
Covariance Type:	cluster		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	11.9612	2.221	5.386	0.000	7.608	16.314
C(gender_concept_id)[T.8532]	0.4241	0.562	0.754	0.451	-0.678	1.526
state_edu_attainment	0.0569	0.072	0.790	0.429	-0.084	0.198
age	-0.0068	0.012	-0.563	0.573	-0.030	0.017
has_college_education	0.4352	0.368	1.182	0.237	-0.286	1.157

Omnibus:	1346.241	Durbin-Watson:	1.944
Prob(Omnibus):	0.000	Jarque-Bera (JB):	108.222
Skew:	-0.006	Prob(JB):	3.16e-24
Kurtosis:	1.810	Cond. No.	629.

Notes:

[1] Standard Errors are robust to cluster correlation (cluster)



Logistic Regression Results (State Education Level -> Suicidal Thoughts):

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	15.8135	5.249	3.013	0.003	5.525	26.102
C(gender_concept_id)[T.8532]	0.0835	0.037	2.245	0.025	0.011	0.156
state_edu_attainment	0.3364	0.196	1.720	0.085	-0.047	0.720
age	0.0717	0.001	113.188	0.000	0.070	0.073
has_college_education	0.0907	0.068	1.337	0.181	-0.042	0.224

Logistic Regression Results (State Education Level -> Depression Diagnosis):

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.1660	0.372	0.447	0.655	-0.562	0.894
C(gender_concept_id)[T.8532]	0.0048	0.073	0.066	0.948	-0.138	0.147
state_edu_attainment	-0.0061	0.013	-0.457	0.647	-0.032	0.020
age	-0.0011	0.002	-0.477	0.633	-0.006	0.003
has_college_education	-0.0131	0.137	-0.095	0.924	-0.281	0.255

Interaction Model Results:

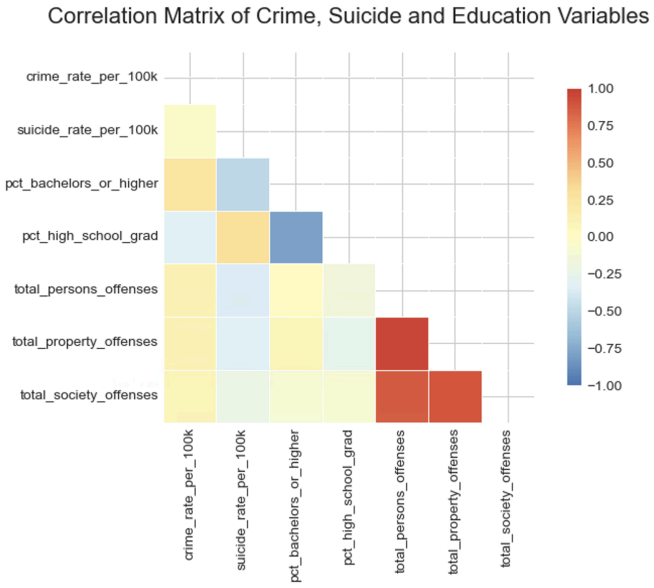
	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	8.3989	2.230	3.766	0.000	4.027	12.770
C(gender_concept_id)[T.8532]	0.4194	0.567	0.740	0.460	-0.692	1.531
state_edu_attainment	0.1901	0.077	2.472	0.013	0.039	0.341
has_college_education	9.2254	1.744	5.291	0.000	5.808	12.643
state_edu_attainment:has_college_education	-0.3261	0.066	-4.911	0.000	-0.456	-0.196
age	-0.0076	0.012	-0.634	0.526	-0.031	0.016

5. Hypothese 3

This figure shows the correlation matrix between crime rate, suicide rate and education level. The darker the color, the stronger the correlation. Red represents positive correlation and blue represents negative correlation.

The correlation between crime rate and suicide rate is weak (lighter color), indicating that there is no strong linear relationship between them, and there may be mediating or moderating factors. Education level (especially the proportion of undergraduate and above) is negatively correlated with suicide rate (blue block): that is, the higher the education level, the lower the suicide rate, suggesting that education may be a protective factor. High school graduation rate is also negatively correlated with suicide rate, but not as significant as undergraduate.

Overall, this figure provides a theoretical basis for our subsequent research: education may buffer mental health risks in a high-crime environment, but the direct relationship between crime itself and suicide is not obvious, which is worth further exploration with more complex models (such as interaction terms or nonlinear models).



This figure is a summary table of the linear regression results of the interactive model (Model 3), which is used to test the impact of crime rate, education level and their interaction terms on suicide rate.
F-test p-value = 0.00288: The overall model is significant, indicating that at least one variable has a statistical impact on suicide rate.
Education level has a significant effect on reducing suicide rate, which is consistent with the second hypothesis (H2). Crime rate has no significant effect on suicide rate, which does not support the first hypothesis (H1). The buffering effect of education on the relationship between crime and suicide has not been statistically significantly supported, which does not support the third hypothesis (H3).

Model 3 (Interaction):

OLS Regression Results

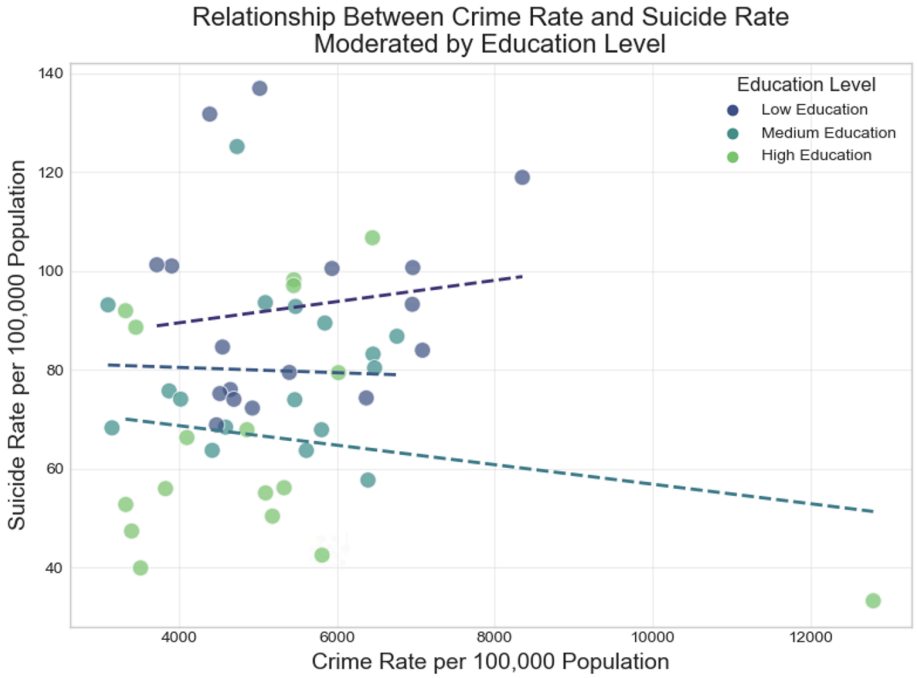
Dep. Variable:	suicide_rate_per_100k	R-squared:	0.256
Model:	OLS	Adj. R-squared:	0.208
Method:	Least Squares	F-statistic:	5.379
Date:	Fri, 04 Apr 2025	Prob (F-statistic):	0.00288
Time:	12:01:08	Log-Likelihood:	-223.25
No. Observations:	51	AIC:	454.5
Df Residuals:	47	BIC:	462.2
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	79.7229	2.835	28.118	0.000	74.019	85.427
crime_centered	0.0019	0.002	0.802	0.426	-0.003	0.007
education_centered	-1.6155	0.505	-3.196	0.002	-2.632	-0.599
crime_x_education	-3.893e-05	0.000	-0.289	0.774	-0.000	0.000

Omnibus:	7.951	Durbin-Watson:	2.215
Prob(Omnibus):	0.019	Jarque-Bera (JB):	7.501
Skew:	0.929	Prob(JB):	0.0235
Kurtosis:	3.282	Cond. No.	3.31e+04

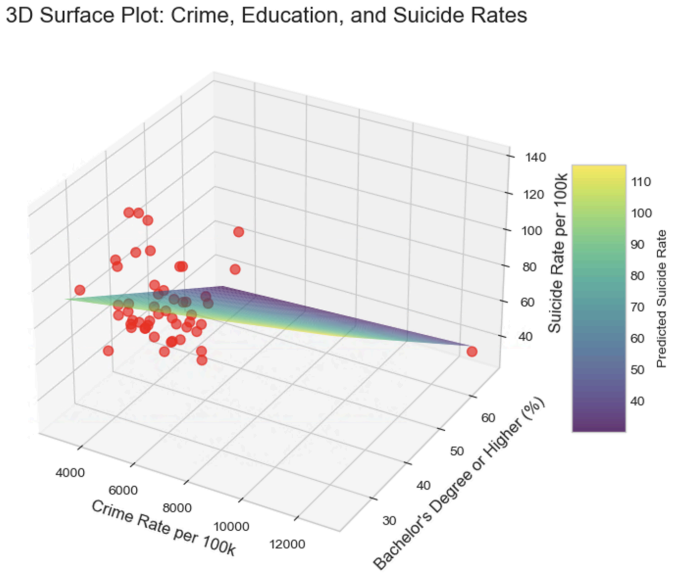
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.31e+04. This might indicate that there are strong multicollinearity or other numerical problems.

At high education levels, the negative impact of crime rate on suicide rate is significantly weakened or even reversed, which supports the hypothesis that "high education has a protective effect."



However, The 3D surface

- The surface tilts a bit in the middle, which means there might be a small interaction between crime and education.
- The red dots (real data) don't all sit on the surface — some are higher, some lower. So the model doesn't fit perfectly, which matches what we saw in the residual plot.



Given the limitations of linear models in capturing the complex relationships between crime rates, educational attainment, and suicide rates, we expanded our analysis by testing several nonlinear modeling approaches. Among

them, the Generalized Additive Model (GAM) emerged as the most effective based on both goodness-of-fit and prediction accuracy.

	R^2	RMSE
GAM	0.301704	18.661878
Polynomial Regression	0.269830	20.315443
SVR (RBF Kernel)	0.261420	19.921303
Neural Network	0.251708	20.051855
Random Forest	0.250580	20.066970
Gradient Boosting	−0.140685	24.757224

	R^2	RMSE
GAM with Interaction	0.483700	16.046738
GAM	0.301704	18.661878
Polynomial Regression	0.269830	20.315443
SVR (RBF Kernel)	0.261420	19.921303
Neural Network	0.251708	20.051855
Random Forest	0.250580	20.066970
Gradient Boosting	−0.140685	24.757224

```
interaction_formula = """suicide_rate_per_100k ~
                bs(crime_rate_per_100k, df=4) +
                bs(pct_bachelors_or_higher, df=4) +
                bs(crime_rate_per_100k,
df=3):bs(pct_bachelors_or_higher, df=3)"""
```

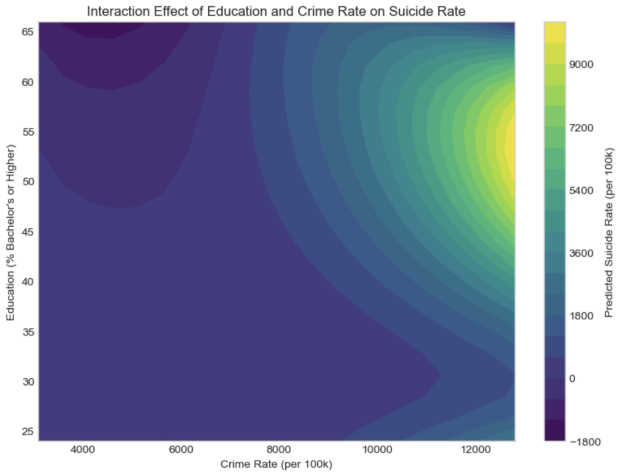
To further enhance the model's explanatory power, we introduced a nonlinear interaction term between crime rate and education level. The final GAM specification incorporated spline transformations for both predictors as well as their interaction:

```
suicide_rate_per_100k ~
      bs(crime_rate_per_100k, df=4) +
      bs(pct_bachelors_or_higher, df=4) +
      bs(crime_rate_per_100k, df=3):bs(pct_bachelors_or_higher, df=3)
```

This model significantly improved performance metrics compared to both the baseline GAM and other nonlinear models. Specifically, the R-squared increased from 0.3017 to 0.4837, an improvement of 18.2 percentage points, indicating that the model with the interaction term accounts for nearly half of the variance in suicide rates across states. In addition, the Root Mean Squared Error (RMSE) decreased from 18.66 to 16.05, reflecting a substantial reduction in prediction error.

Importantly, the inclusion of the interaction term revealed that the relationship between crime rate, education level, and suicide rate is not uniform across all values. Instead of a simple additive or linear relationship, the interaction GAM uncovers more nuanced patterns — for example, in some regions, high education levels appear to buffer the effects of high crime on suicide, while in others, this effect may be weaker or absent. These findings underscore the importance of allowing for flexible, data-driven interactions in public health modeling and suggest that linear models may oversimplify real-world phenomena.

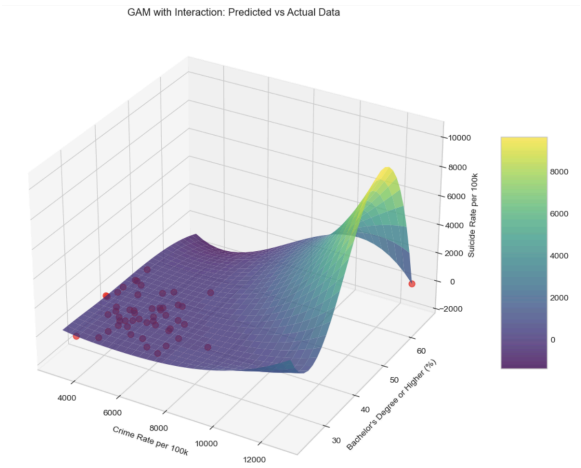
The figure below shows a contour plot showing predicted suicide rates for different crime rates (x-axis) and education levels (y-axis, measured as the percentage of the population with a bachelor's degree or higher). The visualization is based on a generalized additive model (GAM) that includes a nonlinear interaction term between crime and education.



So we could get two conclusion from this plot:

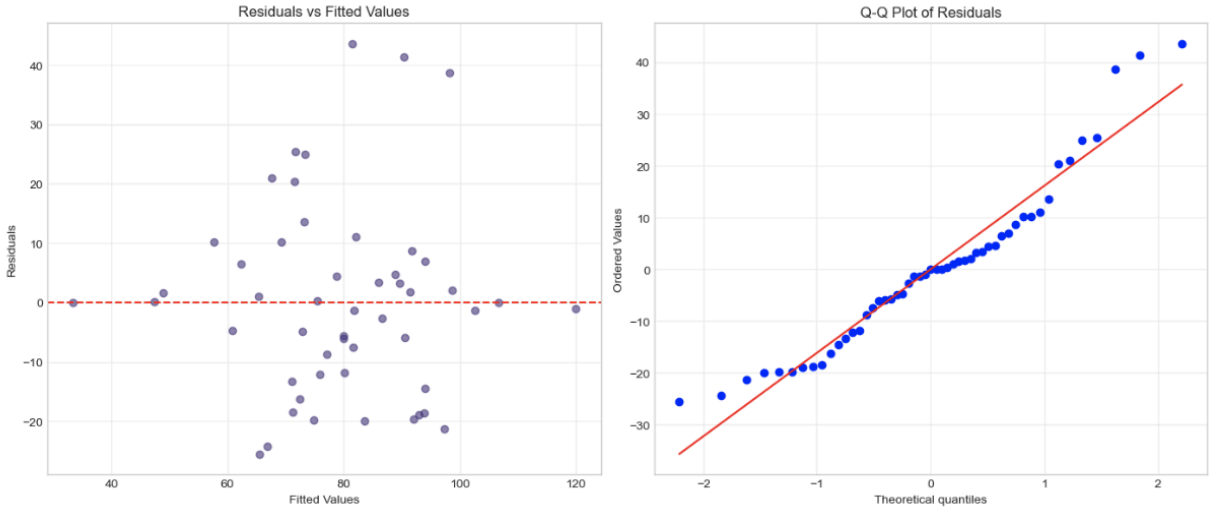
- Crime alone does not linearly predict suicide rates
    - Along the horizontal axis (crime), the color change is not strictly linear—in some areas, increased crime is associated with increased suicide rates (yellow cluster on the right), but in other areas, the effect appears to be flat or even decreasing. This suggests that the impact of crime depends on education level.
- The Protective Effect of Higher Education
    - For states with lower education levels (lower portion of the figure, 25-40%), increased crime is associated with modest increases in suicide rates, although the colors remain mostly dark to light blue, indicating limited variation. In contrast, for areas with moderately high education levels (50-60%), the interaction is more pronounced: predicted suicide rates rise sharply as crime rates rise, eventually reaching a bright yellow peak in the upper right corner. This suggests a nonlinear and potentially counterintuitive interaction—in states with very high education levels, extreme crime may overwhelm the protective effect of education.

This 3D surface plot visualizes the nonlinear surface fitted by a generalized additive model (GAM) that includes an interaction term between crime rate and education level (percent of population with a bachelor's degree or higher) to predict suicide rate per 100,000 population.



And for this figure, we also found two interesting points:

- Evidence of Interaction
    - In some areas, increasing education level lowers the predicted suicide rate, especially when crime rates are moderate. However, in areas where both crime and education levels are high (upper right), the predicted suicide rate spikes. This suggests that education does not buffer the effects of crime evenly, especially at the extremes.
- Model Fit vs. Actual Data
    - The red dots represent actual data points in the dataset. The dots are slightly scattered above and below the prediction surface, indicating a reasonable but imperfect fit. While the model captures the overall pattern well, some local deviations suggest that other unobserved factors may still be at play.



The residuals vs. fitted values plot (left) shows that residuals are fairly symmetrically distributed around zero across the range of fitted values, suggesting that the model does not exhibit strong systematic bias. There is no clear funnel shape or curvature in the pattern, which indicates that the assumption of homoscedasticity (constant variance) is reasonably met.

The Q-Q plot (right) further supports this interpretation, as most residuals lie close to the red diagonal line, implying that the residuals are approximately normally distributed. While there are minor deviations at both tails, particularly for larger residuals, these are not severe enough to indicate a major violation of the normality assumption.

These diagnostic plots suggest that the model fits the data adequately, with no major violations of key linear regression assumptions such as linearity, normality, or homoscedasticity. However, slight deviations at the extremes may point to the presence of a few influential observations or nonlinear patterns not fully captured by the model, which may warrant further exploration in future modeling steps.

Reference

1. Baranyi, G., Sieber, S., Pearce, J., Keppner, J., Mooney, S. J., & Galea, S. (2021). *Neighbourhood crime and mental health: A systematic review and meta-analysis of multilevel studies.*
2. Montez, J. K., & Cheng, K. (2022). Educational disparities in adult health across U.S. states: Larger disparities reflect economic factors. *Frontiers in Public Health, 10*, 966434. https://doi.org/10.3389/fpubh.2022.966434
3. Stafford, M., Chandola, T., & Marmot, M. (2007). *Association between fear of crime and mental health and physical functioning. American Journal of Public Health, 97*(11), 2076–2081. https://doi.org/10.2105/AJPH.2006.097154