Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Comprehensive Analysis of the Determinants Influencing Sexual Assault Crime Rates Across the United States Using Machine Learning and Explainable AI

Samhita Konduri
University of California, Berkeley
Vishnu Pendyala
vishnu.pendyala@sjsu.edu
San Jose State University

Research Article

Keywords: Sexual Assault, Explainable AI, Random Forest, SHAP values

Posted Date: October 28th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11484590/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

A Comprehensive Analysis of the Determinants Influencing Sexual Assault Crime Rates Across the United States Using Machine Learning and Explainable AI

Samhita Konduri[1] and Vishnu S. Pendyala[2*]

[1]University of California, Berkeley, CA, USA.
[2*]Department of Applied Data Science, College of Information, Data, and Society, San Jose State University, 1 Washington Square, San Jose, 95192, CA, USA.

*Corresponding author(s). E-mail(s): vishnu.pendyala@sjsu.edu;
Contributing authors: samhita.konduri@gmail.com;

Abstract

Sexual assault (SA) crimes are horrific crimes against humanity. It is critical for our society to find ways to reduce and ultimately eradicate this issue. However, it is not yet clear which of the possible factors impact the occurrence of SA crimes more than others. There is also a lack of quantitative studies on SA crime rates. This leads to less focused and targeted efforts when trying to reduce the occurrence of SA. This paper fills these gaps by leveraging machine learning techniques. The work explores the impact of various legal, societal, economic, and demographic factors across all 50 US states on SA crime. A Random Forest regression model is used to determine the relative influence of each factor on SA crime rates. SHAP values are used to identify the features that correlate the most and least with SA crime rates. The work suggests that factors such as drug usage and rural population density are more strongly correlated with SA crime rates than legal factors such as the strictness of statute penalties. The findings indicate that rather than focusing solely on increasing statutory strictness, directing efforts to other solutions, such as educational programs and community initiatives, will be more impactful. The study also contradicts conclusions from previous studies by finding that states with more registered sex offenders have lower SA crime rates.

Keywords: Sexual Assault, Explainable AI, Random Forest, SHAP values

1 Introduction

In the United States of America, there are 13 forcible rapes of adult women every 10 minutes (for Victims of Crime, for Victims of Crime). Addressing and reducing the occurrence of such sexual assault (SA) crimes is a high priority for society. One important aspect of combating SA is primary prevention, which focuses on preventing SA crimes before they occur (Porat et al., 2024). In order to improve the efficacy of these primary prevention strategies, it is necessary to know which aspects of our society and legal system influence the occurrence of SA crime. This can allow for more targeted and informed approaches. There have been very few studies, especially quantitative studies, to investigate which factors contribute to higher SA crime rates. This work addresses this question using a unique, comprehensive, and quantitative study and leveraging artificial intelligence (AI) techniques. The study analyzes differences in SA crime rates in US states across a comprehensive set of factors.

In the field of quantitative social sciences, AI is well-suited for analyzing human behavior (Grossmann et al., 2023). There are even calls to use AI for better democracy (Pendyala, 2023b) and policy-making (Pendyala, 2023a). Previous work (Konduri and Pendyala, 2024) explored the use of LLMs for categorizing the relative strictness of SA punishments in U.S. statutes. The study concluded that there is little to weak correlation between stricter SA offender punishments in laws and SA crime rates. The research question for this work ensues from that conclusion as follows.

- RQ: If the strictness of SA laws has a weak correlation with crime rates, then are there other factors that have a stronger correlation?

The project sets out to answer how different societal, economic, and demographic characteristics of the 50 U.S. states (drug usage, rural population density, literacy rate, etc.) correlate with SA crime rates, especially in comparison to legal factors such as the strictness of SA statutes.

1.1 Contribution

To the best of our knowledge, this is the most comprehensive, quantitative study examining the factors that influence SA crime in the United States. Specifically, previous work has explored SA crime concerning specific, individual factors, such as sex offender recidivism rates and location. This study takes a novel approach to analyzing multiple different factors across legal, societal, economic, and demographic domains together.

In the field of social science, there are few purely quantitative studies. This study provides a framework for 1) how machine learning (ML) can assist in deriving quantitative conclusions about society and 2) how decision-makers can potentially utilize these quantitative conclusions to make policies and actions more efficient.

Additionally, ML techniques have shown potential for use in crime rate prediction and other social science applications. This study affirms this and showcases the value of ML for much smaller datasets than were previously used with ML models. This is also the first use of ML for SA crime analysis, which is one of the most sensitive and violent crimes in human society.

2 Related Work

The existing body of research on SA crime is extensive, encompassing various analytical approaches and focusing on diverse aspects of the issue. Our work builds upon and differentiates itself from prior studies by offering a purely quantitative analysis of overall SA crime rates, utilizing non-linear regression methods to predict the impact of societal and demographic factors, and employing machine learning strategies for predictive modeling and associational analysis. This section reviews relevant literature, categorizing it into two main areas: existing SA crime analysis and relevant machine learning strategies.

2.1 SA Crime Analysis

There have been many studies on factors that influence various parts of SA crime. In particular, one study explored factors that influence the seriousness of SA charges (Bradmiller and Walters, 1985). Amount of force and type of sexual activity were found to be the most powerful predictors of charge. Other work studied sex offender recidivism in connection with criminal history, substance use, and other factors (Lievore, 2004). The study found that a substantial number of recidivists misuse substances, and the likelihood of SA recidivism decreases with increasing age. However, this study focuses on a specific aspect of SA crime and the specific subset of recidivists. This study analyzes overall SA crime, instead of specific aspects.

Other studies have analyzed various sociocultural factors and their influence on all SA crime, beyond specific aspects. Specifically, one study examined the impact of societal norms and expectations regarding sexual aggression (Stermac et al., 1990). However, these studies lack quantitative correlation values and other results. A quantitative approach offers an alternative viewpoint that can potentially be more precise.

Among the few quantitative studies of SA, General Estimating Equations (GEEs) were used to predict how environmental factors affect serial SA crimes (Hewitt and Beauregard, 2014). GEEs use generalized linear models to model correlations and account for repeated observations. A study about a slightly different topic used hierarchical multiple regression analysis (Bradmiller and Walters, 1985). The technique also assumes that the relationship between factors and outcomes is linear. This study, however, utilizes a non-linear regression method, which is better suited to the real-life data.

Additionally, many of these studies focus on survivors and victims after the crime has taken place (Cook et al., 2011). Though important, this does not address the primary prevention strategies that this work focuses on, to address factors that contribute to the occurrence of crime. Furthermore, multiple studies combine quantitative and qualitative methods for a new "mixed methods" approach (Campbell et al., 2011). This is helpful for an understanding of SA crime; however, a purely quantitative study can provide an additional point of view.

2.2 Machine Learning Strategies

Machine learning tools have been found to have considerable potential for predictive modeling/associational analysis tasks in social science (Grimmer et al., 2021). Predictive modeling tries to forecast future events or variable values from current trends, while associational analysis is discovering patterns and co-occurrences between variables, both of which are perfectly suited to this work. In particular, there have been many applications of machine learning models for crime prediction. Previously, K-Means clustering and neural networks were used for crime prediction in Buenos Aires (Reier Forradellas et al., 2020). However, neural networks rely on large amounts of data that were not available for this work. Therefore, Random Forest models were used instead. Overall, there is a lack of ML applications for SA crimes specifically, which this work addresses.

Our work, however, is specifically focused on crime rate prediction. This differs from crime prediction because, instead of where crimes may take place, crime rate prediction uses a more general, macro lens. There have been studies that used historical crime rates (He and Zheng, 2021) and records (Mahmud et al., 2021). Another study assessed the impact of location on crime rates in India (Bandekar and Vijayalakshmi, 2020). To the best of our knowledge, there are no studies that take into account various societal and demographic characteristics in order to predict crime rates.

3 Methodology

The influence of 18 factors on SA crime rates is compared using regression models and SHAP machine learning explainability. The following subsections further detail the methodology.

3.1 Data

The wording of different types of sexual crimes is critical, especially in a legal context. The terms "forcible rape" and "sexual assault" can be used interchangeably[1], and this is the main crime addressed in this study. Although other articles and papers may distinguish these two categories, this paper follows the delineation offered by Duncan, which is justified by the fact that many modern state penal codes have been replacing wording of "rape" to "sexual assault". The forcible rape crime rates per US state were accessed from FBI data (FBI, 2024) published on Statista, a data collection platform. This statistic will henceforth be referred to as "SA crime rate."

A list of the factors likely to influence SA crime was compiled through a literature survey of social science papers, and amounted to 18 total factors. Quantitative data for these 18 factors were collected from a variety of sources. In particular, among the legal factors, a quantitative measure of the strictness of SA statutes in each state was sourced from a previous study that used LLMs to analyze the statute (Konduri and Pendyala, 2024). A list of the sources and more details about each feature can be found in Table 2 in the Appendix.

[1] A legal article from the University of Houston's Law Center (Duncan, 2007) writes: "Throughout this Article, I use the terms "forcible rape" and "sexual assault" interchangeably. The term "nonconsensual sex" is not interchangeable with these terms."

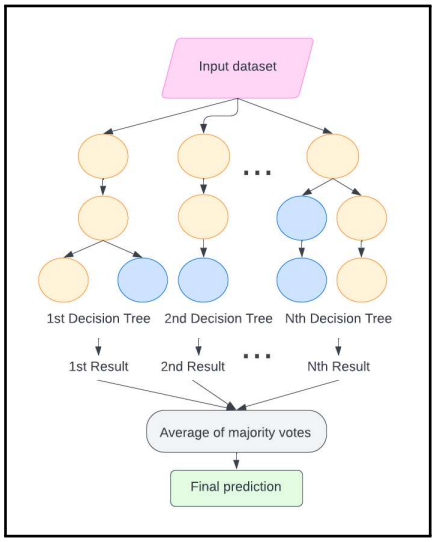


Fig. 1 A schematic diagram of a random forest regression model.

3.2 Regression Models and Machine Learning Explainability

In a previous study, 18 features were identified as likely to influence SA crime rates, based on previous studies and analyses (detailed in the Appendix). The comparative importance of these features in predicting SA crime rates is analyzed.

These features span a range of domains: demographical (rural population density, median population age, etc.), societal (drug usage score, loneliness score, etc.), legal (age of consent, strictness ratings etc.), and economic (percent under the poverty level, amount of federal grant for SA services). The goal of analyzing these factors is not to shift blame away from SA offenders or to attribute responsibility to these 18 factors. Rather, the intention is to identify the best areas for governments and organizations to focus on to address SA crime effectively.

The quantitative data for these features were sourced from various organizations, including census data, as explained in Table 2 in the Appendix. The table also includes the rationale for why each feature could correlate with SA crime rates and was included in this study.

The final dataset for this part of the study consists of 50 samples, each with 18 input features and one numerical dependent variable. This comes from each of the 50 states, with the 18 features that are likely to influence SA crime rates, and the SA crime rates.

Regression models can determine how changes in input features influence the output variable. The most basic regression model, linear regression, achieves this by determining the best-fit line. In contrast, Random Forest regression models are more complex. They operate by constructing multiple decision trees, each trained on randomly selected subsets of the features and data samples. They then make final predictions by averaging the predictions from all of the decision trees, as shown in Figure 1.

Random Forest models are considered to be better suited for datasets with smaller data sizes (between 10 and 80 (Luan et al., 2020)) and for predicting numerical values (AnalytixLabs, 2023). Because the data has a relatively small sample size of 50 and the model will be predicting the numerical values of crime rates, a Random Forest regression model was chosen. The dataset was split into training and testing sets with a ratio of 7:3, meaning 70% of the samples were randomly selected for the training set, and the rest were in the testing set. In practice, this meant the model was trained on 35 states' data, and model performance was evaluated using the remaining 15 states.

The regression model's performance was evaluated using three metrics: MAE, RMSE, and R^2. MAE, or mean absolute error, represents the average absolute error between the actual and predicted crime rates (average residuals). RMSE is the square root of the average squared difference between the actual and predicted crime rates (standard deviation of residuals). Residuals are the differences between the actual crime rate and the predicted crime rate for every sample, representing the error terms for each sample. MAE and RMSE values both range from 0 to infinity. R^2 values range from 0 to 1, and represent the proportion of the variance in the crime rates which is explained by the regression model. Higher performing models have MAE and RMSE values closer to 0 and an R^2 close to 1.

The SHapley Additive exPlanations (SHAP) framework is regarded as the "gold standard" for model explainability (Mosca et al., 2022). SHAP values, derived from Shapley values in game theory, are used in machine learning to show the relative importance of various features in a model (Lipovetsky and Conklin, 2001). SHAP values quantify each feature's contribution to the model's final prediction. This can be used to show which features have more influence on the output variable.

Once the Random Forest model was trained and evaluated, SHAP values were used to determine which features had higher importance in predicting SA crime rates. SHAP values are displayed and analyzed through the beeswarm plot in Section 5.

4 Experiments

All of the code was written in Python and executed on Google Colaboratory (Colab) Jupyter Notebooks. The SA crime data and data for all of the influencing factors were stored in Google Drive and accessed by mounting the drive on Colab.

The Python library scikit-learn (sklearn) was used for the regression model. The method sklearn.model_selection.train_test_split() was used to split the dataset into training and testing sets, with test_size = 0.3. The method sklearn.ensemble.RandomForestRegressor() was used to fit the Random Forest regression model, with all of the default hyperparameters. The SHAP values and graph were created using the Python library shap and the method shap.plots.beeswarm(). A random state of 109 was used for both the train_test_split and Random Forest functions.

5 Results

As Table 1 shows, the Random Forest regression model achieved acceptable performance, with an R^2 of 0.8342 that is close to the ideal value of 1. This indicates that the

	Metrics for Random Forest Regression Model
MAE	3.4243
RMSE	4.3566
R^2	0.8342

Table 1 The MAE, RMSE, and R^2 metrics for the Random Forest regression model fitted to the dataset of states and the 18 indicator features being studied. The training set was 35 states, and these metrics are based on the differences between actual and predicted crime rates for the testing set of 15 states. All values are rounded to the nearest ten-thousandths place.

model is able to explain 83.42% of the variance in crime rates, which is relatively high, especially for predictive models in social science. The MAE and RMSE values are also smaller and closer to 0, suggesting that crime rates are being predicted accurately.

It is expected that regression models trained with these 18 features are not likely to reach extremely high performance levels (such as R^2 values greater than 0.9). This is because crime rates are extremely complex and difficult to model, due to the numerous factors that can affect them. Thus, an R^2 of 0.8342 is high enough, showing that the Random Forest regression model acceptably models crime rates in U.S. states. It serves the goal of this study to compare the relative importance of the 18 features in predicting SA crime rates.

SHAP values are calculated for each of the 18 features that were used to train the Random Forest regression model. These values are displayed in a beeswarm plot in Figure 2. A beeswarm plot visually summarizes how the features in the dataset impact a machine learning model's predictions. In the plot, each instance is a dot on each feature row. The dot's x-position indicates the feature's effect on the model's output. Densities are shown by the dots, and colors represent the original feature values. Features with data points spread out farther have higher SHAP values. Therefore, they are more important for the model's crime rate predictions.

The color distribution for each feature shows the general relationship between the feature and SA crime rates. For example, for drug_usage_score, higher drug usage scores have higher SHAP values. This indicates that states with higher drug usage scores have higher SA crime rates.

In contrast, for registered_sex_offenders_per_100K (number of registered sex offenders per 100,000 population), higher numbers of registered sex offenders have lower SHAP values. This indicates that states that have more registered sex offenders actually tend to have lower SA crime rates. Because it is among the top 4 most important features in this study, sex offender registration is an important predictor of SA crime rates, and there seems to be a negative association between the two variables. This challenges previous research findings that there is no consistent relationship between the two (Stucky and Ottensmann, 2016). Future work can further investigate and detangle this relationship.

Interestingly, for the frac_female feature (fraction of the state population that is female), lower feature values (lower female populations) have higher SHAP values. This indicates that states that have smaller female populations tend to have higher

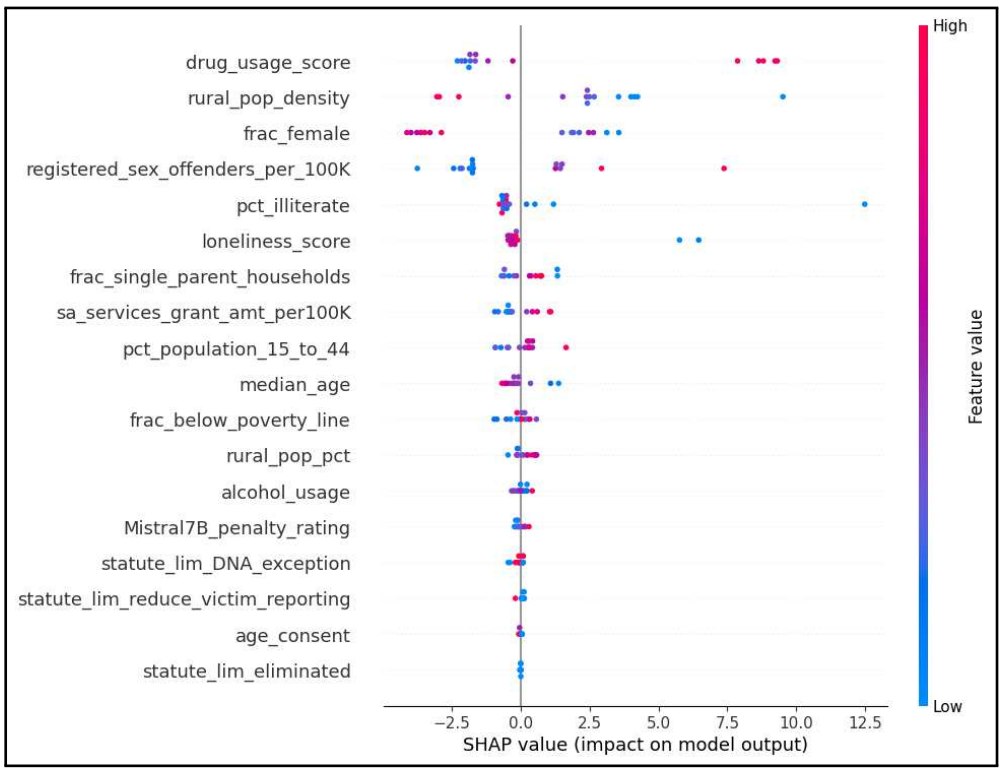


Fig. 2 A beeswarm plot displaying the SHAP values for each of the 18 features studied. The features are sorted (on the y-axis) in order of importance (mean absolute SHAP value), from highest importance at the top to the least importance at the bottom. For every row, each point is one of the 15 states in the testing dataset. The color of the points represents the variable's raw value, with red representing high values of that variable and blue representing lower values of that variable. For example, higher rural_pop_density (red dots) have lower SHAP values (towards the left of the x-axis).

SA crime rates, which has many implications given that SA victims are overwhelmingly women (for Victims of Crime, for Victims of Crime). There are likely many confounding variables, which may include population composition and reporting patterns, that affect female population and SA crime rates, which should be studied in future work via targeted studies.

The strictness of statutes (indicated by the Mistral7B_penalty_rating feature) has low importance to predicting SA crime rates. In fact, the five legal features (age of consent, statute of limitations, statute penalty strictness) have lower importance in predicting SA crimes than all of the other features. This suggests that only making federal and state statutes stricter is not the most effective strategy to address SA crime, compared to strategies that can impact the other factors.

The most important features for the model to predict SA crime rates are drug usage, rural population density, and the fraction of the population that is female. Drug usage is the most impactful feature. Higher drug usage scores had higher SHAP values,

which means that states with higher drug usage also have higher SA crime rates. This can be addressed through policy and outreach programs. Rural population density is the next most impactful feature. States with lower rural population densities have higher SHAP values, meaning they are associated with higher SA crime rates. This could mean that focusing on the most remote areas in each state is a good strategy for addressing the overall SA crime rates. Although the female population of each state was the next most impactful feature, this is unfortunately a feature that cannot be significantly influenced by policy or programs.

5.1 Limitations

One major limitation of this study is the low sample size. There are only 35 states in the training set, and only 15 test states. Given there are only 50 U.S. states, and including other countries' data would introduce much more variability in legal systems and other confounding variables, this was the maximum possible sample size for this study. Future studies can try to analyze more countries/provinces/states with more standardized variables. However, for this study, the authors note that there is a risk of overfitting and the R^2 value should be cautiously interpreted since it is only based on 15 states.

Another limitation of the study is our reliance on correlational data. Due to the nature of the available data, our analysis only permits the identification of associations and correlations. Therefore, the findings do not establish causal relationships, nor do they ensure that the suggested policy changes will be effective. These can be worked on in future studies. This study uses correlational findings to only suggest potential priority areas for policy changes.

Additionally, since all of the data was collected between 2022 and 2024, it is possible that the COVID-19 pandemic impacted the collected features, which means the relationship between these factors may have been different prior to the pandemic. The FBI's SA crime reporting methods and standards may change in the future, which can have an impact on the results of this study.

6 Conclusion

This study's research question can be answered as follows: Legal factors (strictness of SA statutes in terms of penalties, statutes of limitations, and ages of consent) correlate far less with SA crime rates than other features. Drug usage and rural population density are the features most correlated with SA crime rates.

The study explores various legal, societal, demographic, and economic factors across all 50 states, applying a Random Forest regression model alongside SHAP values to determine each factor's influence on SA crime rates. The findings demonstrate that drug usage, rural population density, and female population fraction tend to be more impactful for SA crime rate prediction than legal factors.

The findings suggest that solely making federal or state statutes stricter may not be the most effective strategy for reducing SA crime rates, and grassroots campaigns or changes to the enforcement and outreach processes may also be effective solutions to pursue. Correlation findings for some of these features are also highlighted. Drug

usage has the highest importance for predicting SA crime rates, with states with higher drug usage scores having higher SA crime rates. States that have more registered sex offenders tend to have lower SA crime rates.

These trends and their suggestions for policy decisions can be studied in future work. More regression models can be compared as well. This work can also be extended with a larger set of features, such as military base concentrations. The methodology of this project can be applied to analyze SA crimes around the world, helping to identify global trends in these crimes.

This study showcases the potential for using Machine Learning techniques for legal analysis and modeling crime. The methodology can be applied to a number of other types of crimes beyond SA, such as driving under the influence (DUI) or thefts, making our society and lawmakers/policymakers more informed about the key factors that influence these crime rates.

Declarations

- No Funding has been received for this project.
- The authors have no conflicts of interest or competing interests.
- Ethics approval and consent to participate does not apply to this study.
- Both authors consent to the publication of this article.
- Data availability: all data used in this study is available in the public domain, and cited in the Appendix and/or Methodology sections.
- Materials availability: Not applicable.
- Code availability: All code used in this study will be provided by the authors upon request.
- Author contribution: Conceptualization: [Samhita Konduri, Vishnu Pendyala]; Methodology: [Samhita Konduri, Vishnu Pendyala]; Formal analysis and investigation: [Samhita Konduri]; Writing - original draft preparation: [Samhita Konduri]; Writing - review and editing: [Samhita Konduri, Vishnu Pendyala]; Supervision: [Vishnu Pendyala]; Data Curation: [Samhita Konduri]; Software and Debugging: [Samhita Konduri]

Appendix

The following table contains details for the 18 features used in the regression model and SHAP explainability analysis. Statutes of limitations are laws that set a time limit for when legal proceedings can begin after an event takes place. The common justification is that smaller statutes of limitations make it harder for victims to report their cases, which increases the chances of perpetrators escaping punishment and continuing to commit these crimes. The results from this study contradict this justification.

Variable Name	Description	Source	Reason
LEGAL FACTORS			

age_consent	Age at which a person is considered mature enough to consent to sexual activities as of 2024	(Review, 2024a) [2]	(Snyder, 2000)
Mistral7B_penalty_rating	Strictness rating for state SA penalties from Mistral 7B LLM	(Konduri and Pendyala, 2024)	(DeMatteo et al., 2015)
statute_lim_DNA_exception	Whether this state eliminated the statute of limitations for all felony sex crimes: 0 for no, 1 for yes	(Rainn, 2024) [2]	(Purdue and Vars, 2023)*
statute_lim_eliminated	Whether a victim's decision to report or not to report reduces the state's statute of limitations for sex crimes: 0 for no, 1 for yes	(Rainn, 2024) [2]	(Purdue and Vars, 2023)*
statute_lim_reduce_victim_reporting	Whether this state has exceptions to statutes of limitations for DNA evidence: 0 for no, 1 for yes	(Rainn, 2024) [2]	(Purdue and Vars, 2023)*
SOCIETAL FACTORS			
alcohol_usage	Rate of ethanol consumption in gallons of spirits per 100,000 population, in 2022	(Zhu, 2024) [3]	(Carpenter and Dobkin, 2010)
drug_usage_score	Comprehensive drug usage score combining 20 metrics ranging from overdose rates to drug testing laws, as of 2024	(Kiernan, 2024) [4]	(Bianchi et al., 2024)
loneliness_score	Comprehensive loneliness score based on factors ranging from searches on friendship apps and number of single-person households, 2022	(Bloom, 2022)	(Wielinga et al., 2021)
ECONOMIC FACTORS			
frac_below_poverty_line	Fraction of state population below poverty level, Census-based estimates for 2023	(Bureau, 2024b)[5]	(Pennsylvania Coalition Against Rape, 2007)

[2]Data collected from state codes.
[3]Data from the National Institute on Alcohol Abuse and Alcoholism (NIAAA).
[4]Data collected from U.S. Census Bureau, Bureau of Labor Statistics, Centers for Disease Control and Prevention, and other government groups.
[5]Data based on U.S. Census.

sa_services_grant_amt_per100K	Rate of federal Sexual Assault Services Formula Grant Program (SASP) amount awarded to state per 100,000 residents, in 2023	(of Justics Office on Violence Against Women (OVW), 2024)	(Morrison et al., 2004)

DEMOGRAPHICAL FACTORS

frac_female	Fraction of state population identifying as female, Census-based estimates for 2023	(KFF, 2024)[6]	(Morgan and Kena, 2022)
frac_single_parent_households	Percentage of families in the state where the householder parent is single, from 2021 Census survey	(Davis, 2024)[6]	(Boccella et al., 2022)
median_age	Median age of state population, Census-based estimates for 2023	(Bureau, 2024a)[6]	(Felson and Cundiff, 2014)
pct_illiterate	Percent of state population with low literacy levels, from Barbara Bush Foundation (2024)	(Review, 2024c)[7]	(Pennsylvania Coalition Against Rape, 2023)
pct_population_15_to_44	Percent of state population between ages 15 and 44, Census-based estimates for 2023	(Bureau, 2024a)[6]	(Felson and Cundiff, 2014)
registered_sex_offenders_per_100K	Rate of registered sex offenders per 100,000 residents, as of 2024	(Gabriele, 2024) [8]	(Quinsey et al., 1995)

[6]Data based on U.S. Census.

[7]Data from PIAAC International Adult Literacy Survey, NAEP reading assessment. 4th grade ELA test performance data comes from edopportunity.org, American Community Survey, and others.

[8]Data from state online sex offender registries, law enforcement public information officers, U.S. Department of Health and Human Services 2022 Child Maltreatment Report, and U.S. Census.

rural_pop_density	Population in rural areas divided by total rural land area from 2020 Census	(Review, 2024b)[9]	(Salt et al., 2024)
rural_pop_pct	Percent of state population residing in rural areas from 2020 Census	(Review, 2024b)[9]	(Salt et al., 2024)

Table 2: Explanations of the 18 factors analyzed for RQ3 with the regression model. The "Source" column includes the source of the data. The "Reason" column includes previous work that has justified why the factor may play a role in SA crime rates. A * next to the citation in the "Reason" column for the statutes of limitations indicates that the rationale is explained further above this table.

References

AnalytixLabs. 2023. Random forest regression – how it helps in predictive analytics?

Bandekar, S.R. and C. Vijayalakshmi. 2020. Design and analysis of machine learning algorithms for the reduction of crime rates in india. *Procedia Computer Science* 172: 122–127 .

Bianchi, F., C. Giraudo, D. Favretto, A. Amagliani, F. Boccardo, P. Mortara, G. Borretta, M. Onorati, R. Beux, M. Luparia, et al. 2024. Victims of drug facilitated sexual assault aged 13-24. *International Journal of Legal Medicine*: 1–10. https://doi.org/10.1007/s00414-024-03197-0 .

Bloom, L.B. 2022. The loneliest places in america will surprise you.

Boccella, L., J.J. Mandemakers, and R. van Gaalen. 2022. Single-parent families and adolescent crime: Unpacking the role of family structure. *Journal of Developmental and Life-Course Criminology* 8: 231–255. https://doi.org/10.1007/s40865-021-00183-7 .

Bradmiller, L.L. and W.S. Walters. 1985. Seriousness of sexual assault charges: Influencing factors. *Criminal Justice and Behavior 12*(4): 463–484 .

Bureau, U.C. 2024a. Age and sex.

Bureau, U.C. 2024b. Poverty status in the past 12 months.

[9]Data based on U.S. Census.

Campbell, R., D. Patterson, and D. Bybee. 2011. Using mixed methods to evaluate a community intervention for sexual assault survivors: A methodological tale. *Violence against women 17*(3): 376–388 .

Carpenter, C. and C. Dobkin. 2010. Alcohol use and crime: Findings from a longitudinal sample of u.s. adolescents and young adults. *Journal of Health Economics 29*(6): 1001–1014 .

Cook, J.M., S. Dinnen, and C. O'Donnell. 2011. Older women survivors of physical and sexual violence: A systematic review of the quantitative literature. *Journal of Women's Health 20*(7): 1075–1081 .

Davis, M. 2024. Best and worst places for raising a family as a single parent.

DeMatteo, D., M. Galloway, S. Arnold, and U. Patel. 2015. Sexual assault on college campuses: A 50-state survey of criminal sexual assault statutes and their relevance to campus sexual assault. *Psychology, Public Policy, and Law 21*(3): 227 .

Duncan, M.J. 2007. Sex crimes and sexual miscues: The need for a clearer line between forcible rape and nonconsensual sex. *Wake Forest L. Rev.* 42: 1087 .

FBI. 2024. Forcible rape rate per 100,000 inhabitants in the united states in 2022, by state.

Felson, R.B. and P.R. Cundiff. 2014. Sexual assault as a crime against young people. *Archives of sexual behavior* 43: 273–284 .

for Victims of Crime, N.C. Sexual assault victimization.

Gabriele, R. 2024. Sex offender registry statistics: 2024 data for all 50 states.

Grimmer, J., M.E. Roberts, and B.M. Stewart. 2021. Machine learning for social science: An agnostic approach. *Annual Review of Political Science 24*(1): 395–419 .

Grossmann, I., M. Feinberg, D.C. Parker, N.A. Christakis, P.E. Tetlock, and W.A. Cunningham. 2023. Ai and the transformation of social science research. *Science 380*(6650): 1108–1109 .

He, J. and H. Zheng. 2021. Prediction of crime rate in urban neighborhoods based on machine learning. *Engineering Applications of Artificial Intelligence* 106: 104460 .

Hewitt, A. and E. Beauregard. 2014. Sexual crime and place: The impact of the environmental context on sexual assault outcomes. *Journal of Criminal Justice 42*(5): 375–383 .

KFF. 2024. Population distribution by sex.

Kiernan, J.S. 2024. Drug use by state (2024).

Konduri, S. and V.S. Pendyala 2024. Large language models for quantitative analysis of statutory stringency and its correlation with sexual assault crime rates. In *International Advanced Computing Conference*, pp. 81–94. Springer.

Lievore, D. 2004. Recidivism of sexual assault offenders: Rates, risk factors and treatment efficacy .

Lipovetsky, S. and M. Conklin. 2001. Analysis of regression in game theory approach. *Applied stochastic models in business and industry 17*(4): 319–330 .

Luan, J., C. Zhang, B. Xu, Y. Xue, and Y. Ren. 2020. The predictive performances of random forest models with limited sample size and different species traits. *Fisheries Research* 227: 105534 .

Mahmud, S., M. Nuha, and A. Sattar 2021. Crime rate prediction using machine learning and data mining. In *Soft Computing Techniques and Applications: Proceeding of the International Conference on Computing and Communication (IC3 2020)*, pp. 59–69. Springer.

Morgan, R.E. and G. Kena. 2022. The global prevalence of sexual assault: A systematic review of international research since 2010. *Trauma, Violence, & Abuse 23*(5): 1451–1464. https://doi.org/10.1177/1524838020965837 .

Morrison, S., J. Hardison, A. Mathew, and J. O'Neil. 2004. An evidence-based review of sexual assault preventive intervention programs. *Washington, DC: Department of Justice* .

Mosca, E., F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh 2022. Shap-based explanation methods: a review for nlp interpretability. In *Proceedings of the 29th international conference on computational linguistics*, pp. 4593–4603.

of Justics Office on Violence Against Women (OVW), U.D. 2024. Sexual assault services formula grant program (sasp).

Pendyala, V.S. 2023a, December. Artificial intelligence can take the politics out of policymaking. *Fulcrum* .

Pendyala, V.S. 2023b, November. Toward an ai-driven democracy. *Newsweek* .

Pennsylvania Coalition Against Rape. 2007. Poverty and sexual violence: Building prevention and intervention responses .

Pennsylvania Coalition Against Rape. 2023. How does literacy relate to sexual violence? *PCAR* .

Porat, R., A. Gantman, S.A. Green, J.H. Pezzuto, and E.L. Paluck. 2024. Preventing sexual violence: A behavioral problem without a behaviorally informed solution. *Psychological Science in the Public Interest 25*(1): 4–29 .

Purdue, J.M. and F.E. Vars. 2023. Time to heal: Trauma's impact on rape & sexual assault statutes of limitations. *Tex. A&M L. Rev.* 11: 125 .

Quinsey, V.L., M.L. Lalumière, M.E. Rice, and G.T. Harris. 1995. Predicting sexual offenses. *Assessing dangerousness: Violence by sexual offenders, batterers, and child abusers*: 114–137 .

Rainn. 2024. State by state guide on statutes of limitations.

Reier Forradellas, R.F., S.L. Náñez Alonso, J. Jorge-Vazquez, and M.L. Rodriguez. 2020. Applied machine learning in social sciences: neural networks and crime prediction. *Social Sciences 10*(1): 4 .

Review, W.P. 2024a. Age of consent by state 2024.

Review, W.P. 2024b. Most rural states 2024.

Review, W.P. 2024c. U.s. literacy rates by state 2024.

Salt, E., I. Erickson, A.T. Wiggins, T. Borders, C. Curtsinger, A. Wallace, and M.K. Rayens. 2024. Comparing the demographic characteristics of victims of sexual assault in rural versus urban areas. *Journal of forensic nursing 20*(3): E34–E42 .

Snyder, H.N. 2000. Sexual assault of young children as reported to law enforcement: Victim, incident, and offender characteristics. *Bureau of Justice Statistics* .

Stermac, L.E., Z.V. Segal, and R. Gillis. 1990. Social and cultural factors in sexual assault, *Handbook of sexual assault: Issues, theories, and treatment of the offender*, 143–159. New York: Springer.

Stucky, T.D. and J.R. Ottensmann. 2016. Registered sex offenders and reported sex offenses. *Crime & Delinquency 62*(8): 1026–1045 .

Wielinga, F., K. Margeotes, and M.E. Olver. 2021. Clinical and risk relevance of intimacy and loneliness in a treated sample of men who have offended sexually. *Journal of Sexual Aggression 27*(1): 35–51 .

Zhu, K. 2024. Mapped: America's spirits consumption by state.