Crime Data Mining and Victimization Analysis across United States

*Naga Tejaswi Veluri Tharun Reddy Somannagari Dileep Reddy Suddamalla   
Volgenau School of Engineering, Volgenau School of Engineering, Volgenau School of Engineering,  
George Mason University George Mason University George Mason University*

*Email:* [*nveluri@gmu.edu*](mailto:nveluri@gmu.edu) *Email: tsomanna@gmu.edu Email: dsuddama@gmu.edu*

1. **Abstract**

**These days, technology has grown to an extent where it is transforming the traditional way of tackling the problems. One such field, where advancement in technology have its prominent role is security. With the help new methods and technologies, we can have complete picture of the incidents that took place in past and present and analyse the incident there by understanding the possible patterns.**

**During this project, We analysed three data sets to predict the state safety Index, they are State Crime (Data1), Human trafficking (Data2) and Police Shootings (Data3), which are extracted from corgis.data and data.world.**

*Keywords:* ***Security, incidents ,state safety***

1. **Introduction**

According to the [Global Peace Index](http://visionofhumanity.org/app/uploads/2019/06/GPI-2019-web003.pdf),  [United States](https://worldpopulationreview.com/countries/united-states-population) stands in 128 place out of 163 countries in the world. Safety plays a vital role while choosing a place to live for an one. Terrorist attacks, mass shootings, crimes, and natural disasters, are always on the headlines, creating a fearful environment for both citizens and visitors of the location. These factors are intern acting as a reasons for families or citizens looking to relocation.

Safety varies from place to place based on time and depending on various factors. Some states are safe in few factors but may fall prey to others. To determine which states are the safest, we used indicators such as Human trafficking history, State- crime records and Police shooting data to analyse and determine the safest states

State Crimes include Violent crimes and Property crime analysis, with the help of human trafficking data set, we tried to uncover the rates of trafficking based on different aspects and police shooting helped us to understand and estimate the factors and frequency of shootings that took place in a state.

Also, using these datasets, we try to find if there are any patterns followed in any of the locations to commit a crime.

Which would act as a precautionary measure and caution both the security department and public which would in turn enhance safety and enable police/ security determine the resource allocation according to the crime rate of the location.

1. **Significance**

This analysis is used to determine the pattern of happening of any incident, which is used as guide for both public and officials, to foresee the happening of incident and take measures accordingly.

In turn, this analysis is used to can be used as a precautionary guide to enforce the security department to ensure required action is taken before hand before happening of the event.

1. **About the Data**

**State Crime:**

This data set describes the crime rates on multiple incidents across United states from 1960 -2019. Which mainly categorised into two factors, namely property and violent crime. Property crime refers to incidents related burglary, larceny and motor. And Violent crimes are related to assault, murder, rape and robbery.

**Police Shooting:**

This data set can be widely classified into three categories, which describes incident details, victim details and factors that led to shooting.

Incident details explains us about the place, State and data of the incident, whereas victim’s details brief us about the gender, age and race details. Incident details are totally about the weapons used during the attack and victim’s actions post incident and other factors are described.

This data set also contains the mental health condition of victim and overall threat level of incident.

**Human Trafficking:**

This data set gives us state wise count of trafficking activity, and it mainly classified into two categories. Namely commercial sex acts and involuntary servitude and there are other attributes which are used to describe the actions.

1. **Literature Survey**

Being a sensitive topic, there are only few works related to this topic,

[1] is used to extract a useful inference from time series data collection, a combination of big data tools for data administration and Generalized Linear Regression for statistical analysis is used. Similar crime tendencies among diverse crime locations can be detected with this augmentation, which can help criminal site selection. As a result, the Auto Regressive Integrated Moving Average model enables the prediction model to minimize error. The goal of this study is to more precisely determine the offender's whereabouts in advance. By establishing the association among the crime attributes, Auto Regression Techniques were used to accurately forecast the crime with the least amount of error for such time series data. Only the linear data relationship fit the auto regressive model proposed in this study.

Here in [2], The algorithm prioritizes attributes with fewer values but greater importance, which solves the classification issue of choosing attributions with more values by introducing attribute-importance. The experimental results reveal that Advanced ID3 algorithm has better classification accuracy than regular ID3 algorithm and can generate more acceptable, effective, and rational classification rules. The new technology will assist detectives in recognizing suspicious e-mail and in obtaining information in a timely manner so that required actions can be taken to reduce criminal activity.

1. Paper [3], depicted the percentage variation in crime accusing in each state, the frequency of crime by type, and crime against women and girls for police departments to use this data to make strategic decisions and for the government to evaluate existing plans to reduce crime and crime rates, as shown in this paper under Apache Pig. This report also identifies numerous crime incident drift points for future goals. This large-scale crime data analysis can assist our police department in examining crime scenes and coming up with decisions and solutions to reduce incident-based crimes.

Using [5], the authors tried to find if there is an relation or patterns found while working on the data, and this data is fed into multiple machine learning models such as k-means clustering , Naïve bayes, correlation and clustering models to predict optimal values for the future years using tools such as R and Weka.

Segmented Multiple Metric Similarity Measure (SMMSM) clustering method ,was used in [6] to find the patterns or any activities that would lead to crime. As traditional clustering method is predicting the patterns with relatively low accuracy. The attributes in the data set are segmented based on their effect on the crimes that take place. Which would help us to accurately predict the occurrence of events in future.

[7] project gives the analysis of majority data attribute and tries to find the rates based on the categories present in data. And also explores which category of data is mostly involved in the incidents. Here, we can get the count of total incidents in the state irrespective of years. And the author used two different graphs which represents race and age of victim.

In the other project[8], the developer tried to show the vary in rate of shootings by years based on race. He used a bar graph that explain the number of persons who were affected in the attack by grouping them based on race, different colours were used to differentiate the years in the plot. But this analysis does not give any clear picture of data attributes present in the data set. This analysis gives a very brief view of data.

The project[9], which explained about the most contributing factor that led to shooting. As per their analysis, they did not get adequate data to analyse if there is any relation between race and occurrence of incident, but in this analysis, I would like to find the relation between race and age of individuals. During this project, they tried to find if there is any pattern followed in states, but in the proposed analysis, I would like to see if there is any change in pattern based on years.

Project[10], gives us a box plot was used to see the age range of individual for every race, but did not explain the its change during years. Attacking style and weapons used for attacking were not considered as impacting factors in this project.

Here[16], authors Andrii Shalaginov, Jan William Johnsen, Katrin Franke, states that When it comes to cyber-enabled and cyber-dependent crimes, big data is becoming a difficulty for criminal forensic investigators. Traditional investigative methods and digital forensics technologies grow less effective as their ability to deliver required results in a timely and resource-constrained manner deteriorates. The use of computational forensics based on modern data analytics to prevent and combat cybercrime is one potential alternative for Cyber Crime Investigations. As a result, machine learning and computer modelling should be included in the investigations. One of the options is computational forensics, which provides quick and efficient data analysis in order to discover minute errors.

In[17], author Xingchen Yu, states that the cross-border cybercrime detection system uses TFTP server, Uboot network development board, and OpenStack crime information detection component to build a complete hardware operation environment and detect information characteristics and cross-border network in the context of PSE and big data analysis. A more stable software execution environment is established with the help of three areas of detection domain and cybercrime information orientation planning data. In terms of practicality, this new type of technology has a fast cross-border cybercrime detection speed, which can considerably increase criminal location accuracy and totally replace traditional detection methods.

The author in [18], ChhayaChauhan, SmritiSehgal, states that during the examination of experimental data, the enhanced ID3 algorithm provides more reasonable and effective categorization criteria. The purpose of the hidden link method was to find hidden linkages in co-offender networks, which revealed the potential future crime partner and a distinct network outside the genuine network. Using Bayes theorem, classification approaches yielded above 90% accuracy. The forensic kit tool creates a file and examines the data, as well as analyzing the victim system where the attack is taking place. The accuracy of the criminal investigation analysis (CIA) technology, which is used to help investigate violent crimes, was limited. (A. Shalaginov, 2017)

1. **Proposed Method and Approach**

Generating Safety Score to each State

Our goal is to find the patterns of an incident based on previous incidents to caution the security department and public to ensure needed measures are taken to prevent it. And aiding to allocate more security resources to ensure there no disturbance in activities.

Initially, the data is pre-processed by removing all the null values, unwanted data is eliminated and the parameters are analysed .

Finally, upon pre-processing, we have created dashboards to find the hidden patterns

To attain this, the whole process is evenly broken down to multiple steps:

1. Data Collection and Cleaning:

The data is collected from corgis.data and data.world. which is then cleaned by removing the null values, duplicate entries and other unwanted entries. Also if any of the column has more than 25% of null values, they are also eliminated as it cannot be used for prediction.

The if the co-relation is less than 25% then such variables are also eliminated.

1. Creating the Final Data:

As the data is collected from multiple sources, the pre-processed data is broken into required form a data repository to eliminate redundancies and eliminating unwanted data.

Where, the main division is on state crime data set, which is divided into two data sets namely Violent crimes and property crimes datasets.

1. Exploratory Data Analysis:

During this the attribute in the data is studied by plotting various graphs, there by giving us the initial picture of the rates of each category based on year and the state.

1. Prediction of Results:

Once the exploratory data analysis is done, the impacting factors are recognised using correlation and dashboards accordingly based on state and year.

Diagram

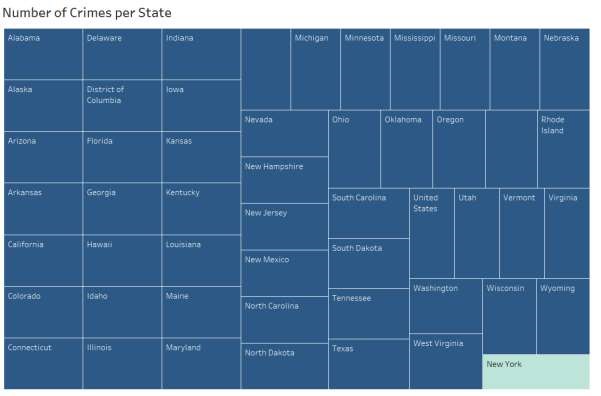
Description automatically generated

**Figure :1**

**[Approach to**

**Solve the Problem]**

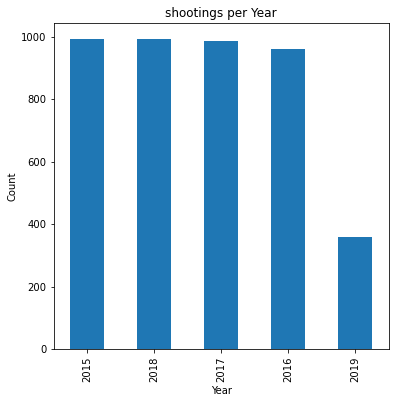
1. **Exploratory Data Analysis**



**Figure: 2**

**[Total Number of Incidents based on Year ]**

By the above analysis, Alabama has the most number of crimes and least being the states from east coast.



**Figure:3**

**[ year Wise Analysis]**

Here, we can see approximately same count of incidents took place during 2015,2016,2017 and 2018 but we can huge down fall in number of incidents that took place in 2019 throughout USA .

Graphical user interface

Description automatically generated with low confidence

**Figure:4[Weapons Used]**

we observe guns are widely used during the attack and followed by pipe and hatchet and other tools, irrespective of time of incident and place where it took place. If We get into in-depth analysis, that trying to figure out if there is any change in usage of weapons during year.

Chart, pie chart

Description automatically generated

**Figure:5**

**[Fleeing Analysis]**

In general, when attacked by police, majority of the victims did not try to flee but few tried to escape by car or through feet irrespective of years and location incident.

Chart

Description automatically generated with medium confidence

**Figure:6**

**[Shooting Manner]**

we can determine that there is very negligible rate of shot and tasered shootings to understand this better, we perform a year wise analysis.

A picture containing crossword puzzle

Description automatically generated

**Figure:7**

**[Threat Level Analysis]**

From this, we can say that most of the incidents took place as they were attacking. To understand it bit clear, let’s have yearly view of data.

Chart

Description automatically generated

**Figure:8**

**[Gender Wise Analysis]**

We notice males are majorly involved in the incidents. To understand the change in rates, year wise analysis is done.

Chart, histogram

Description automatically generated

**Figure:9**

**[Age Wise Analysis]**

In generic analysis, we can notice young adults aging between 25-35 are mostly involved in shootings to understand the change in pattern, data is analysed based on year.

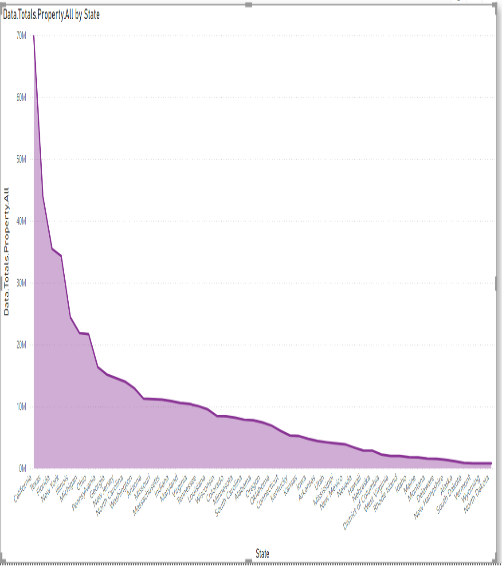
A picture containing chart

Description automatically generated

**Figure:10**

**[Race Wise Analysis]**

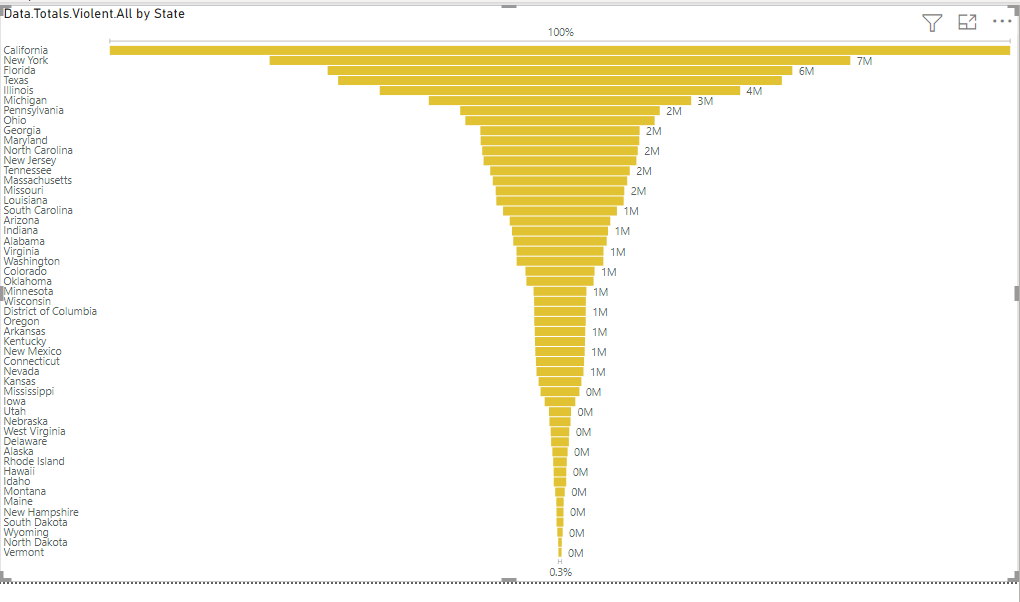
Here, we see most of the victims are white skinned, to have a clear picture to check and analyse if there are any patterns involved, year wise analysis is performed.



**Figure:11**

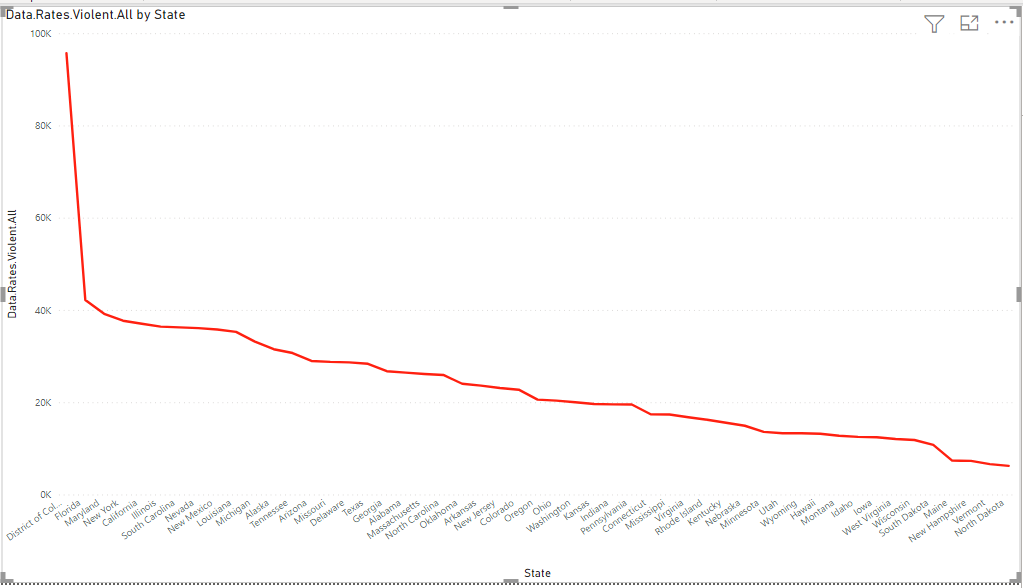
**[Property Crime Analysis]**

Most number of property crimes are observed in California and least in east coast region

****

**Figure:12[State Wise Crime rate]**

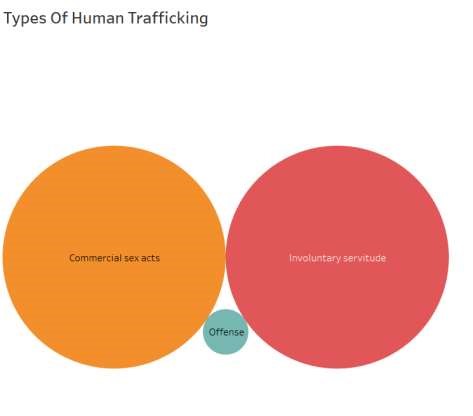
Generic state wise analysis is done to find the state with most number of crimes.

****

**Figure :13**

**[Violent Crime Analysis]**

As per the analysis, we find similar pattern in case of least number of criminal activities recorded similar to property crimes but on the flip side most number of cases are recorded in capital city of the country.



**Figure:14**

**[Human Trafficking Analysis]**

The above picture clearly depicts that majority of the human trafficking are due to commercial and involuntary acts. And minute number of cases in offense category.

Exploratory analysis is made on all four data sets, initial three data sets were divided into four data sets. That is the data sets which are obtained after data pre-processing.

Using which we were able to get higher level view of the data set and the impacting factors.

We also found that males are mostly involved in activities, but analysing the data on month wise and year wise would help us to determine the change in the ratio or proportion based on gender over years.

We also can observe guns are widely used for attacking , again year wise analysis is used to find the hidden patterns occurring in the data set.

1. **Results**

**Graphical user interface, application

Description automatically generated**

**Figure: 15**

**[Human Trafficking Dashboard]**

From the above dashboard, we can get all the patterns and possibilities of happening of an incident. During which the factors such as actual count of trafficking, cleared count and uncleared count are analysed to check what is the actual count rate of missing and to which extent , the victims are found with out trafficking.

Graphical user interface, application

Description automatically generated

**Figure:16**

**[Property crime Analysis]**

This dashboard is used to uncover the all the patterns present in property crime data set to estimate the happening of events in future.

Graphical user interface, application

Description automatically generated

**Figure:17**

**[police Shooting Dashboard]**

Using this, dash board, we are analysing the events that took place between 2015 to 2019. Where we can observe a pattern of usage of guns with attacking natured victims lead to most of the shootings. One good note is that most of the victims did not try flee which lead to very few shot and tasered cases.

Graphical user interface, application

Description automatically generated

**Figure: 18**

**[ Violent Crime Analysis]**

Incidents such as rape, murder, assault and robbery are analysed during this analysis.

State wise and year wise analysis is done to uncover the all the approximating patterns of events and then verifying it with the rates present in the data set to check the accuracy of the analysis.

1. **Conclusion**

With the help of above dashboards and Exploratory data analysis,

We notice CA, TX and FL are the states with highest shoot outs and during 2015 and 2016, we notice higher activities during mid of year but there is quite opposite scenario when it comes to 2017,2018 and 2019. We see more incidents in the beginning of the year.

Also, we can conclude that we can see a pattern of usage of guns with attacking natured victims lead to most of the shootings. One good note is that most of the victims did not try flee which lead to very few shot and tasered cases.

A pattern with males and age group of 25-35 are mostly involved in the shooting activity.

We also notice slight raise in proportion of female involvement as years passes.

We also notice less crime rate both in terms of property and violent crimes in east coast region, which can be considered as the safe region when compared to any other regions in USA.

Where as the most number of property crimes are recorded in California and western regions of united states. To ensure safety, a firm policy must be made to reduce the risk.

In terms of Human Trafficking Texas is considered the most risky region with highest number of cases recorded over the past years.

And East coastline have least number of cases recorded when compared to other regions in USA.

East coast region have least number of cases recorded from 2015 to mid 2019.

From the above analysis, East coast can be considered as the safest region in USA.

1. **Road Blocks**

One of data set contains data ranging only from 2015 to Mid-2019, which restricted us to perform analysis and find the patterns in the same time frame. And most of the data being categorical data disabled us to perform any statistical models on data.

A better ranged data set helps the analyser to perform and unfold the patterns and the relation among the data.

1. **Future Scope**

Many other factors can be considered into account while proposing a Safety Index for a State like Road Safety Analysis, Environmental hazards etc.

The datasets that are available in the market are not feasible for prediction modelling since they lack the dependent variable.

This has to be taken into consideration for further study.

the datasets are also having different variables since they used different variables to create datasets for each type of crime.

Proper Datasets for each crime need to be generated with the crime rate for an area so that can be processed into a single dataset with locations and crime rate for different types of crimes.

1. **References**
2. R. Yadav and S. Kumari Sheoran, "Crime Prediction Using Auto Regression Techniques for Time Series Data," 2018 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering (ICRAIE), 2018, pp. 1-5, doi: 10.1109/ICRAIE.2018.8710407.
3. M. Sharma, "Z - CRIME: A data mining tool for the detection of suspicious criminal activities based on decision tree," 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014, pp. 1-6, doi: 10.1109/ICDMIC.2014.6954268.
4. Monika and A. Bhat, "An analysis of Crime data under Apache Pig on Big Data," 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2019, pp. 330-335, doi: 10.1109/I-SMAC47947.2019.9032565.

[4]Yadav, S., Timbadia, M., Yadav, A., Vishwakarma, R., & Yadav, N. (2017, April 22). Crime pattern detection, analysis & prediction. Retrieved from IEEE Xplore

[5] Guangzhu Yu, ,. S. (2017). Mining Crime Data by Using New Similarity Measure. *IEEE Xplore*.

[6]Data Source:corgis-edu. (n.d.). *corgis-edu*. Retrieved from corgis-edu: https://corgisedu.github.io/corgis/csv/police\_shootings/

[7]Referred to understand the previously developed projects:statista.com. (n.d.). *statista.com*. Retrieved from statista.com:

https://www.statista.com/statistics/585152/people-shot-to-death-by-us-police-by-race/

[8]Referred to estimate the process of analysis:towardsdatascience.com. (n.d.). *towardsdatascience.com*. Retrieved from towardsdatascience.com: https://towardsdatascience.com/an-examination-of-fatal-forceby-police-in-the-us-db897d97085c

[9]Refereed to understasnd previously developed models:washingtonpost.com. (n.d.). *washingtonpost.com*. Retrieved from washingtonpost.com:

https://www.washingtonpost.com/graphics/investigations/police-shootings-database/

1. *Retrieved From: Used While data preprocesinghttps://pandas.pydata.org/docs/user\_guide/categorical.html*. (2021). Pandas.

https://pandas.pydata.org/docs/user\_guide/categorical.html

1. micromapST. (2020). micromapST. Is referred to complete the project.

https://rdocumentation.org/packages/micromapST/versions/1.1.1/topics/micromapSTpackage

1. Retrieved From: Used for Data preprocessing, Working with missing data — pandas 1.4.1 documentation. (2021). Pandas. https://pandas.pydata.org/docs/user\_guide/missing\_data.html
2. Retrieved From: Used for Data preprocessing,ggplot2 documentation. (2021). Ggplot2. <https://ggplot2.tidyverse.org/reference/>
3. Time series / date functionality — pandas 1.4.1 documentation. (2021). Timeseries Data.

https://pandas.pydata.org/docs/user\_guide/timeseries.html

[16]A. Shalaginov, J. W. (2017). Cybercrime investigations in the era of big data. *IEEE.* 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 3672-3676, doi: 10.1109/BigData.2017.8258362.

[17]Sehgal, C. C. (2017). A review: Crime analysis using data mining techniques and algorithms. *IEEE* (pp. 21-25). 2017 International Conference on Computing, Communication and Automation (ICCCA).

[18]Yu, X. (2020). Design of Cross-border Network Crime Detection System Based on PSE and Big Data Analysis. *EEE International Conference on Power* (pp. 480-483). 2020 IEEE International Conference on Power, Intelligent Computing and Systems (ICPICS), .

**Data Sets:**

[19](n.d.). Retrieved from https://corgis-edu.github.io/corgis/csv/state\_crime/

[20](n.d.). Retrieved from https://corgis-edu.github.io/corgis/csv/police\_shootings/ [21](n.d.). Retrieved from https://data.world/datasets/human-trafficking