ETL Implementation and Analysis for Rate of Crime Offences with Respect to Awaiting Court Room Trials in US

Dharesh Vadalia
MSc. Data Analytics
National College of Ireland
x18192076@student.ncirl.ie

Ashish Patel
MSc. Data Analytics
National College of Ireland
x18182445@student.ncirl.ie

Amit Sahoo
MSc. Data Analytics
National College of Ireland
x18188851@student.ncirl.ie

Rashmikant Shukla
MSc. Data Analytics
National College of Ireland
x18181236@student.ncirl.ie

----------***----------

***Abstract*— Criminal offences is the bane on healthily developing society. Which makes it very important for the law to monitor these criminal activities to maintain the balance in the society. Project focuses on implementing ETL (Extraction, Transformation, Loading) process over collection of unstructured data fetched via API call from various data sources based on criminal offences in US and awaiting court room trials. Analysis over cleaned structured data is performed to determine the trend in various type of criminal acts taking place in US and most dominantly used ammo for committing crime, influence of drug and also find the correlation with the rate to awaiting trials of convicted inmates. To determine this relation, area of analysis is narrowed down to one of the most crime prone state of US, Connecticut, in year 2016 with crime rate of 227 per 100,000. **[1]** Performed analysis aims at finding the trend of various criminal offences in US.

Keywords— ETL, MongoDB, PostgreSQL, Crime Analysis, United States, Connecticut Trials, Gun Violence, Drug Influence

I. INTRODUCTION

Crime rate in United States have shown steep fall of 51% on a long-term run from 1990's to 2018, though FBI reported gradual increase in crime rate from 2014 to 2018 **[2]** **[3]**.Using published criminal offence dataset of US, project aims at visualizing the trends in various category of violent criminal offences such as gun violence, terrorism **[4]** and drug influence. And based on these trends, most popularly used ammo by the convicts. Going ahead we narrow down our analysis to one of the most crime prone state reported in US, Connecticut. Based on dataset published by government of Connecticut regarding inmates waiting for court room trials **[1]**, project aims at determining the correlation between the crime rate in specific category in proportion to criminals waiting for trials in that category at Connecticut Court of Justice. To understand the factors affecting the increase in crime rate for 2016.

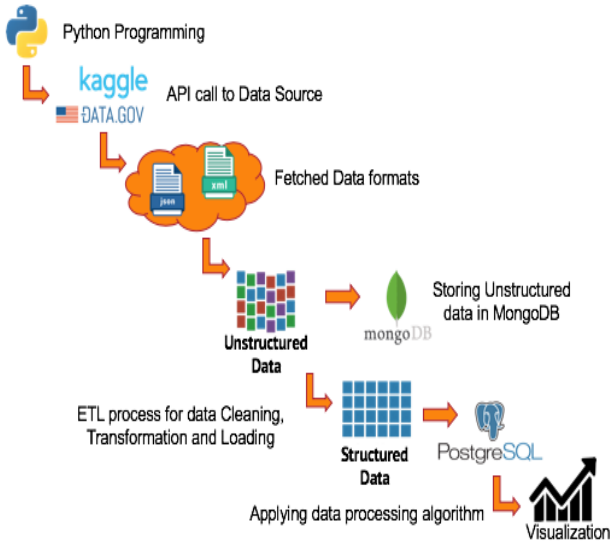

Fig. 1: Followed Procedure

Automation in imposed process flow is achieved by fetching the raw unstructured data via API call, structure fetched data into dictionary format to be able to store it on NoSQL database MongoDB, later perform ETL process on this unstructured data and then load clean data on PostgreSQL and later generate visualizations using plotting libraries in Python and infer the desired relations over cleaned data. Visualization will give an understanding on crime trend, most popular choice of ammunition, mean age group of convicted criminals and what is the impact of judiciary lawsuit trails on future crime rate.

II. Related Work

Crime rate of United States have been predominantly high in early 1900's, which have significantly declined on a long-term run. According to an article published by Pew Research Center and crime report release by FBI in 2018, [3] [2] rate of violent criminal offence shows increasing slope from 2014. As stated in one of the papers published in 2019, That aims at analyzing the trend in committed gun violence's and try to understand laws imposed to manage rate of gun violence and which time of the year is criminally active. Also gives inferences on this shift in trend towards the rate of committed crime in America which are highly correlated to the rate of alcohol consumption, rate of depression and laws enforced against gun violence in State [4].

No one can forget the devastation caused by the terror attacks against the United states on Tuesday, September 11, 2011. Similar Attacks are also observed in other countries like India and Philippines. Therefore, detecting such attacks can be very helpful to the society. We have taken inspiration from the paper [5] for analyzing and visualizing the growing trends of terrorism and its target group over the last 50 years. By 2017, the number of deaths because of terrorism has reduced to 25,673, which is 22%less than its peak value observed in 2014 [6]. We have also referred this document [7] to get better inputs about the arms used by terrorist to carry such activities.

Drug abuse is a constant headache for the authorities and study on these contribute immensely in development of an ideal social structure to live in. Drug related death is prominent field in crime related studies, as many offenders are often are under substance influence like many gun violence offenders are the one with substance abuse [8]. This motivate our study of Connecticut deaths due to drugs.

There are many comparative study papers on performance of MongoDB and Relational Database [9] [10] these shows that individually both lack few features. So, we have used MongoDB and Postgres for smooth management of data, as we are fetching data in json. Before using it in python, changes are made in Postgres to make it consistent for analysis. Graphical analysis of the data set is done as in the paper [11] author explore different content found in accidental drug death cases.

III. Methedology

A. Dataset Description

(1) Weapon Violence Archive (GVA) is a non-revenue driven organization formed in 2013 to give online community access to the firearm related savagery in the United States. GVA will gather and check for exactness of the data, which was put in Kaggle as 'Gun Violence Data'. Thus, in order to study the gun related violence in USA this data set was chosen [12].

(2) GTD (Global Terrorism Database) is an open source database which contains information of more than 180K terrorist activities (both domestic and International) occurred across the globe for the last 50 year. This database is maintained by researchers for START (Study of Terrorism and Response to Terrorism) and kept in Kaggle [13].

(3) Government of Connecticut, a state of US in 2016 released data of all the inmate jailed for criminal offence and waiting for court room trials. Dataset is published on public data portal Kaggle as 'Connecticut inmates awaiting trial'. In order to study and perform analyze over convicted inmates [14].

(4) This dataset is taken from data.gov it includes data of accidental death by drugs reported in Connecticut from 2012 to 2018. These records are originally collated from toxicity report, death certificate, by the office of chief medical examiner [15].

B. Data Gathering and Handling

After Identifying the datasets that will serve the objective of proposed analysis. Following data gathering and handling steps are adopted:

- The dataset (1), (2) & (3) was already in structured form i.e. is in .CSV format, the file was fetched in the local with the help of official API for Kaggle data store, which could be accessed by the use of command line tool for python. The Kaggle API has been used because normal GET-Request are blocked by Kaggle admin. Later, dataset (3) is transformed into XML format to meet the requirement of project. Dataset (4) is fetched from US government datastore Data.gov via API call in JSON format.
- Later dataset (3), is converted into dictionary format to load the unstructured data on NoSQL database MongoDB. Similarly dataset (4) was loaded on MongoDB.
- Data of dataset (3) & (4) is then extracted from MongoDB to perform data cleaning and transformation to structured data format.
- Later all data from datasets (1), (2), (3) & (4) is then moved to the PostgreSQL relational database where the blank cell was converted to the NULL values so that when they are moved to data frame it is recorded as None.

- From Postgres the data is moved to the Pandas Dataframe where all the visualization is carried out with the help matplotlib, plotly etc.

C. *Data Processing*

End to end automation is achieved for the flow designed to process ETL over huge dataset. A controller class handle the data gathering process and submits the data to its respective process job to clean and transform unstructured data into cleaned structured data. Jobs are designed according to individual data set processing requirement to handle different format of unstructured data. To improve the performance of the analytical process for generating visualization and infer huge dataset, chunk processing of dataset is adopted. Also, text processing NLP tool wordcloud is used, which takes the text based raw data and find the most used word among them which could be displayed in cloud style plot, this tool was specifically chosen because there were more than few columns in database where text based raw data was given.

Following technologies and database were used:

- **Python:** Python was chosen as programming language because there were various third-party libraries such as pandas, wordcloud, plotly, matplotlib etc.
- **PostgreSQL:** Relational database was chosen because it's an open source database which could be used without licencing.
- **MongoDB:** Non-Relational database was chosen for storing unstructured data and it is high efficient open source No-SQL database.

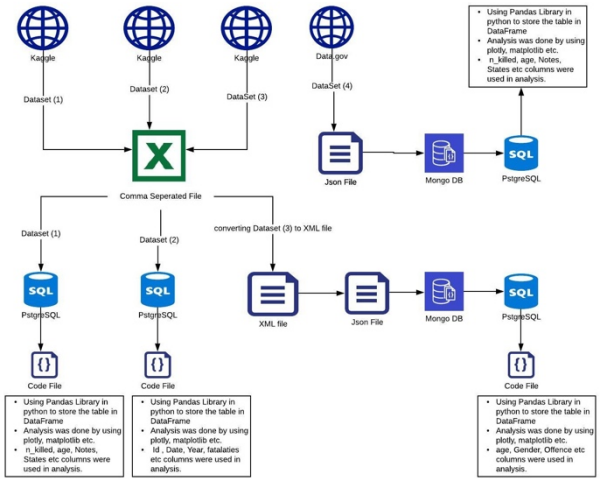

Fig. 2: Process Flow Chart

D. *Data Analysis*

From dataset (1), following visualizations are inferred.

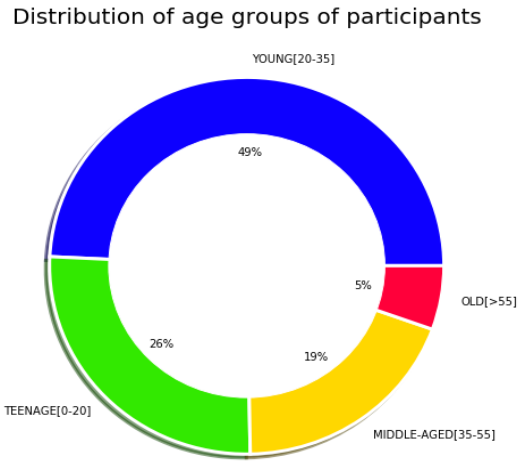

Fig. 3: Age wise distribution of criminals

Above, doughnut chart, we could make out that mostly young age people were involved in the gun violence crime.

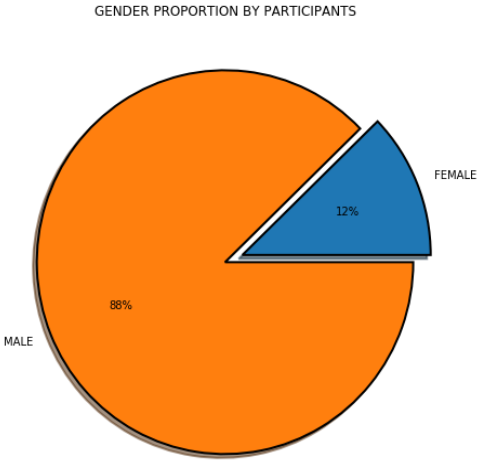

Fig. 4: Gender ratio of criminals

Above, pie chart we could make out that people responsible for the gun violence mostly were male.

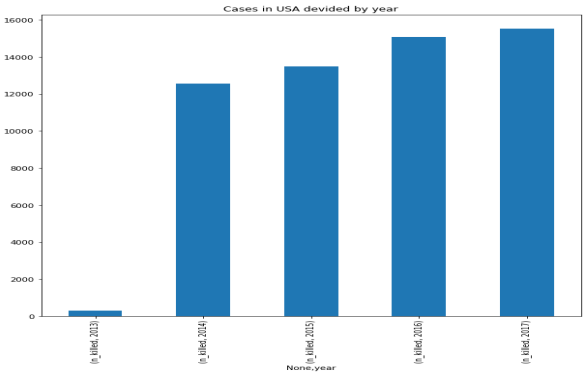

Fig. 5: Increasing rate of gun violence 2013-2017

Above, histogram we could infer that gun violence is increasing per year in united states.

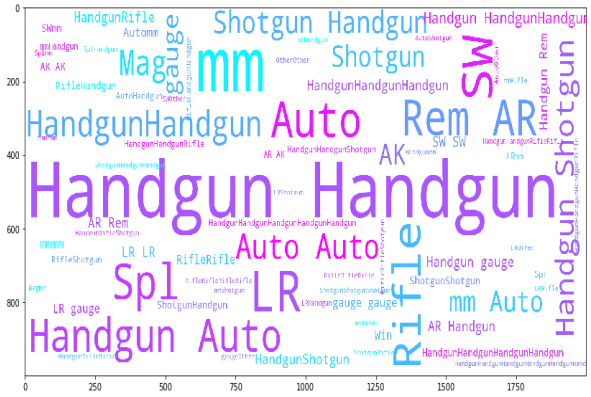

Fig. 6: Most popular choice of Ammo

The inference from the wordcloud chart could be made that mostly handguns were used in the gun violence activities.

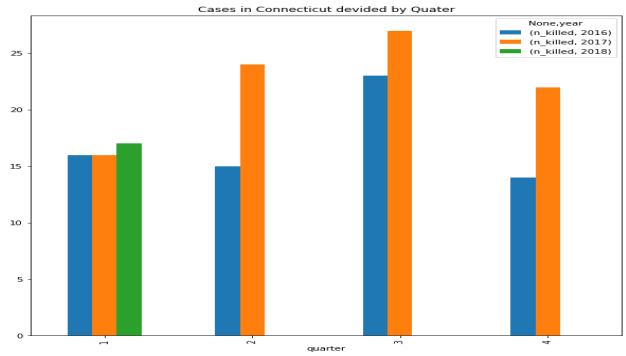

Fig. 7: Rate of gun violence in Connecticut 2016-2018

Drilling down further on the state we look in the Connecticut state there is a constant increase in the number of people killed in each quarter from 2016, 2017 and 2018 by gun violence.

From dataset (2), following visualizations are inferred.

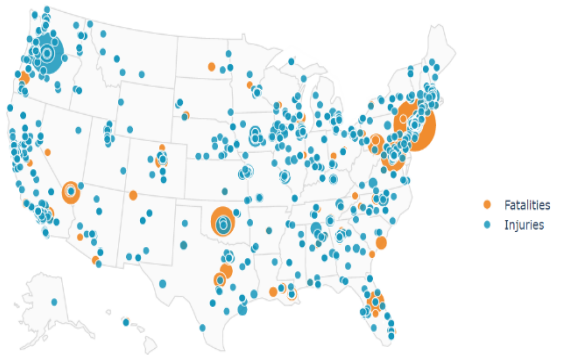

Fig. 8: Impact of terrorist attack in US

Above, graph shows all the terrorist occurred in USA with the magnitude of fatalities and Injuries occurred.

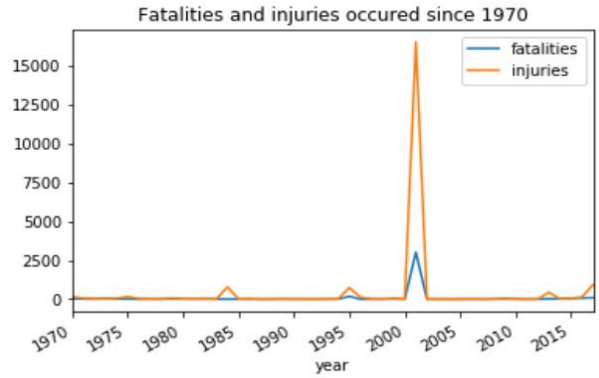

Fig. 8: Impact of terrorist attack in US

Number of deaths and Injuries due to terrorist activity in USA has always been in control, except the one case in 9th Sept 2001. This can be shown be shown in the above graph

Fig. 9: Trend of terrorist attack in US since 1970

Number of terrorist activity happened per Year since 1970. The number is its peak in 1970 and remains the same till 2015.

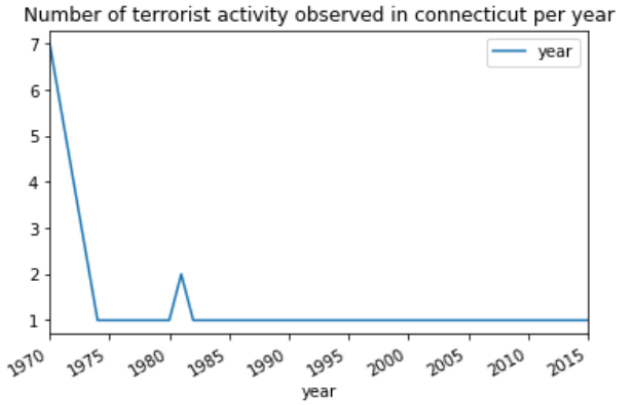

Fig. 10: Trend of terrorist attack at Connecticut

We have drill down to the terrorist attacks in one of the states of USA – Connecticut, which is our area of focus, we could see that, terrorist activity was maximum in 1970 and almost negligible post 1974.

From dataset (3), following visualizations are inferred.

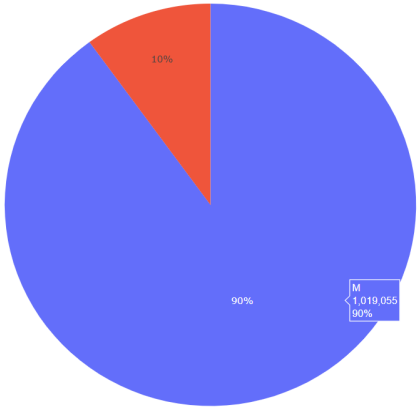

Fig. 11: gender ratio of inmate awaiting trials, 2016

Above pie chart, demonstrates the gender ratio of convicts waiting for trial. Also infers that almost 90% percent of total inmates are male.

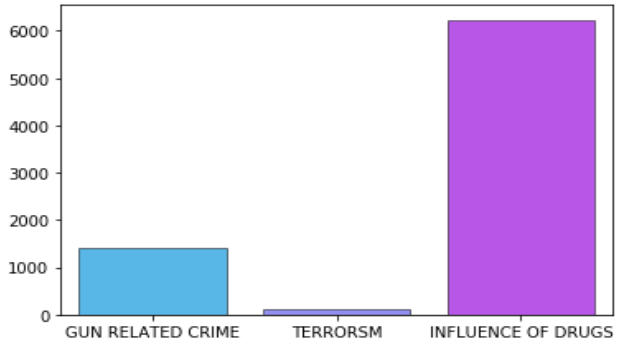

Fig. 12: count of inmate for various category 2016

Above bar chart, demonstrate the count of inmate held for different categories of offences. Chart clearly pictures that most of the crime are committed in influence of drugs.

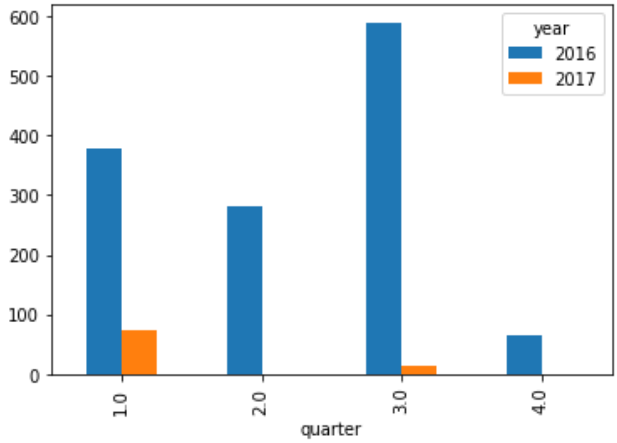

Fig. 13: Count of trails for gun violence

Above bar chart, pictures the quarterly count of inmates jailed for gun violence for year 2016 and 2017. Count of inmates held for gun violence has dropped significantly from 2016 to 2017.

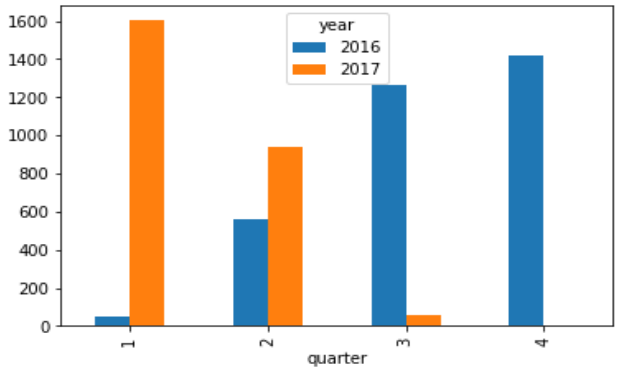

Fig. 14: Count of trials for drug related crime

Above bar chart, depicts the count of criminals jailed for drug related crime. Clearly, influence of drug on crime have increased from 2016 to 1st Quarter of 2017 and then have significantly reduced.

From dataset (4), following visualizations are inferred.

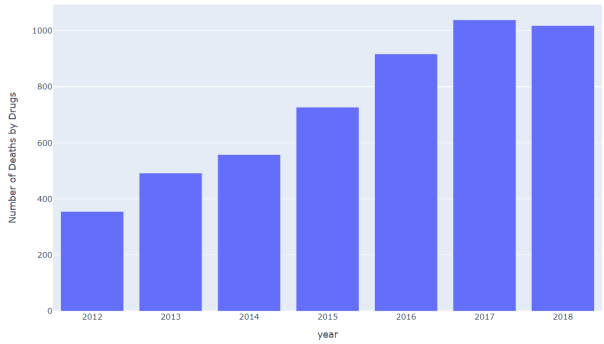

Fig. 15: Connecticut accidental death by drugs

From Above graph, Connecticut accidental death by drugs are continuously on surge from 2012 to 2017. In 2017 it is maximum, which approximately tripled since 2012.

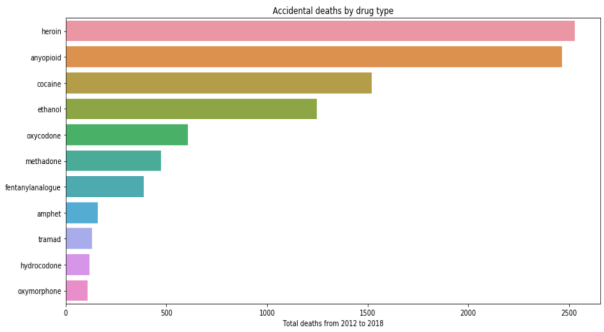

Fig. 16: Total death due to drugs from 2012-2018

Above graph infers, most of the cases are caused by Heroin, Cocaine and Ethanol. Other than this there is interesting trend for “Any opioid” it is almost same as

of heroin. Any Opioid cases are those in which multiple drugs were found.

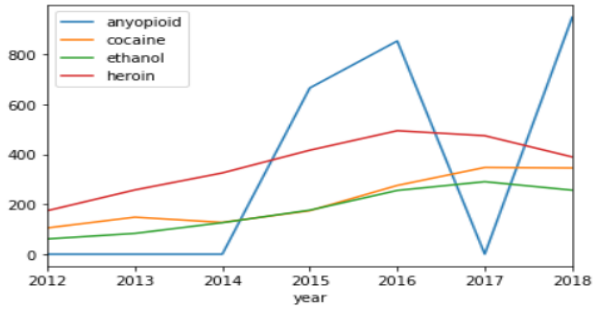

Fig. 17: Number of deaths by various Drug 2012-2018

Drilling further if we see the trend of most found drugs. Here Opioid trend is very unpredictable, while others are on continuous rise.

IV. Result

From Fig 3,4 we can observe that most of the criminal cases are against the age group of 20-35 and are predominantly Male with Handgun as the major weapon of attack (Fig 6) .

From Fig16 ,7 we can see a steady increase in the total deaths in drugs and rate of gun violence respectively. Which suggest that there is strong correlation between them, as referred in a journal.

Since most the crimes are committed by Males (Fig 4), so it's quite obvious that percentage of Male in trial cases is also high as compared to the opposite gender (Fig 11)

The crime rates in Connecticut is in an alarming state as there is a steady increase in the number of deaths due to gun violence and Drug cases. This can be evident from Fig 7 ,Fig 15 . This increase is in sync with the ongoing courtroom trials (evident form Fig13, 14). However, the percentage of terrorism violence (Fig 10) in the ongoing trials is significantly low.

V. Conclusion and Future Work

From the above inferences, we conclude that the number of court room trials awaiting justice in Connecticut state is increasing with each passing year, this correlates with the increasing trend of Gun violence and Drug cases reported in the same state. However, the contribution of Terrorism towards the number of trials is negligible.

References

[1] "Crime in Connecticut," 2016. [Online]. Available: https://www.dpsdata.ct.gov/dps/ucr/data/2016/Crime%20in%20Connecticut%202016.pdf.

[2] "2018 Crime in United States," FBI, 2018. [Online]. Available: https://ucr.fbi.gov/crime-in-the-u.s/2018/crime-in-the-u.s.-2018/topic-pages/violent-crime.pdf.

[3] J. GRAMLICH, "5 facts about crime in the U.S.," [Online]. Available: https://www.pewresearch.org/fact-tank/2019/10/17/facts-about-crime-in-the-u-s/.

[4] C. Moore, "Analysis of American Gun Crime," 06 05 2019. [Online]. Available: http://trap.ncirl.ie/3873/1/charlenemoore.pdf.

[5] A. Magpantay, "Data Analysis and Visualization of Terrorist Attacks in the Philippines".

[6] "Global Terrorism Index," 2017.

[7] J. Bonomo, G. Bergamo, D. . R. Frelinger, J. Gordon IV and B. A. Jackson, "Stealing the Sword," 2007.

[8] G. Banks, K. Hadenfeldt, M. Janoch, C. Manning, K. Ramos, D. A. and P. S. Wolf, "Gun Violence and Substance Abuse," 2017.

[9] T. Jia, X. Zhao, Z. Wang, D. Gong and G. Ding, "Model Transformation and Data Migration from Relational Database to MongoDB," *IEEE,* 2016.

[10] M.-G. Jung, S.-A. Youn, J. Bae and Y.-L. Choi, "A Study on Data Input and Output Performance Comparison of MongoDB and PostgreSQL in the Big Data Environment," *IEEE,* 2015.

[11] T. G. Rheea, J. S. Ross, R. A. Rosenheck, L. E. Grau, D. A. Fiellin and W. C. Becker, "Accidental drug overdose deaths in Connecticut," 2019.

[12] "Gun Violence Data," kaggle.com, 15 04 2018. [Online]. Available: https://www.kaggle.com/jameslko/gun-violence-data/metadata.

[13] "Global Terrorism Database," kaggle.com, 10 09 2018. [Online]. Available: https://www.kaggle.com/START-UMD/gtd/metadata.

[14] "Connecticut inmates awaiting trial," kaggle.com, 26 07 2017. [Online]. Available: https://www.kaggle.com/Connecticut-open-data/connecticut-inmates-awaiting-trial/metadata.

[15] "Accidental Drug Related Deaths 2012-2018," data.gov, 08 05 2019. [Online]. Available: https://catalog.data.gov/dataset/accidental-drug-related-deaths-january-2012-sept-2015.