ANALYZING LEADING CAUSES OF DEATHS BASED ON DRUG USAGE AND VIOLATION OF CRIME

Harsh Manoj Chudasama(x18187340)
School of Computing
National College of Ireland
Dublin, Ireland
x18187340@student.ncirl.ie

Vijit Laxman Chekkala(x18199429)
School of Computing
National College of Ireland
Dublin, Ireland
x18199429@student.ncirl.ie

Muhammad Imran Shaikh (17119308)
School of Computing
National College of Ireland
Dublin, Ireland
x17119308@student.ncirl.ie

Venkata Devaraju Nandimandalam(x18181422)
School of Computing
National College of Ireland
Dublin, Ireland
x18181422@student.ncirl.ie

***Abstract*—The main aim of the project is to find four parallel data sets on a specific domain and performing exploratory data analysis. Insights produced from this process are visualized using different plots. After thorough research, we had summarized the root cause of deaths due to factors like shooting incidents, drug use, criminal courts and other leading causes of deaths in the United States of America. Also, we estimated the death rates as per factors like gender, race, targeted locations, and boroughs. The highest number of deaths based on the factors above belong to the race of White-Hispanic whereas the deaths of Black-Hispanic were low. The leading cause of deaths due to drug consumption with overdose of cocaine were majority Males. Location-wise Staten Island has the least number of deaths based on shooting incidents, drug use, and criminal courts. Thus, the data sets complement each other, and key insights were drawn from them to analyze death rates.**

I. INTRODUCTION

The objective of the project is to investigate multiple aspects of Deaths, Accidental Drug-related deaths, Criminal Courts Summons, Shooting incidents. Deaths data set[1] is mainly based on death rates in New York, Accidental Drug[2] describes the death rate based on the drug's consumption, Criminal courts data set[3] represents the summons issued to every criminal by first and second quarter of 2019, shooting incidents data[4] list every shooting incident that happened from the year 2006 to the previous year.

II. MOTIVATION

Shooting incidents has the number of incidents that took place from the year 2006 till the past year 2018. The data given here has been extracted every quarter and every record is updated for proper analysis. Every row here presents the activity of the shooting incident with the shooting hours, targeted location and borough. Accidental drug deaths data set highlights that the US has increased in drug-related deaths and in 2017 the greater number of deaths are from the state of Connecticut. Most of the people are using drugs without the prescription or taking them in overdosage which is leading to most of the deaths. This Miss usage of the drugs and taking them along with some other substances are leading to the growth of deaths. This analysis helps in safeguarding the people from taking the drugs in hazards amounts and to downsize the deaths. Criminal courts summons data set[3] is focused on the violation of law by the criminals and the issuance of summons to every criminal in the 1st and 2nd quarter of 2019. The key elements for considerations are criminals Age groups, sex, race, Boroughs and its precinct of occurring and law description according to the summons issued to criminals. Death's data set is pointing to Chronic disease that has increased over the period of time in New York City. Most leading causes of deaths are related to Diseases, over usage of the drug and drug poisoning. Hence, want to explore data-set further and get key insights based on death rate and gender-wise comparison.

The reason for choosing the criminal court data set is based on its relevance with the Shooting Incidents data set, as both data sets are extracted from Same City i.e. New York and contain common key elements like sex, race, and boroughs, etc. which was further helpful while analyzing complex queries. The justification of Accidental drug deaths data set relies on its objective and relevance with Deaths data set, the Accidental drug data set explains the finding of which drug is leading the growth of

the death rate in various states of USA to analyze the deaths of drug consumption, as in deaths data set the deaths causing category includes Drugs related death which is related to Accidental drug deaths data sets in which data rate is based on drug consumption for later final combine analysis.

Research Questions:

- What is the most targeted location in New York City where the shooting incident has taken place?
- Which year had the highest number of shooting incidents based on the borough in New York City?
- Analyzing the overdose of the drugs which is leading to the greatest number of deaths
- Evaluating which gender is undergoing the greatest number of deaths for consuming drugs
- Analyzing the precinct of occurring and in boroughs of New York by age- group, gender, and race.
- Analyzing the counts of summons issued to criminals by age group and boroughs of New York.
- Examine no of the count of Law description as per summons issued to criminals.
- To evaluate the leading causes of deaths in the USA.
- To find the death rate based on factors like gender and race-ethnicity.

III. Related Work

Based on the 2012 edition of the New York City Police Department (NYPD), where the data of the shooting incident has been examined from 1963 found out that there was no pattern found similar or any method which could be analyzed for the shooting incidents. There was a big variance in the type of shooting incidents taking place based on the targeted location, age, and sex of the attacker. They searched over 230 cases and found out that most of the attackers were Male and comparing the female count, the results were 3 percent of the Male attackers. The age group of the attacker/perpetrator responsible for the shooting incident were 15-19 in school facilities areas and 30-34 age group in the places which were not close to school are the number of the shooting incidents were increasing every with the certain amount[5]. Another study for the active shooters from the year 2000-2013 showed that the shooting incident that took place was mostly carried by Male. The targeted locations were found to be of commerce areas followed by education areas and open spaces[6].
78 shooting incidents were analyzed from the year 2013 to 2016 where the trend was noticed and the majority of the shooting incident took place in open spaces, school areas and military areas[7].

From a similar analysis of the overdosage deaths were high in Connecticut, the paper describes the analysis is taken from 2012-2018, the overdosage is largely due to fentanyl and polysubstance usage. The findings may not speak more about other states since the deaths are more related to fentanyl spread in Connecticut. This is a similar case analysis from this project, but the number of records varies. These stratified analyses of deaths are taken from various cities of the USA and the most widespread overdosage drug is fentanyl and from Connecticut from the research findings and related works. While this slightly varies with the analysis of what we did due to the number of records we have taken [2], [8].

As per book[9], death is a phenomenon that can't be predicted and it has a drastic impact on personal mental health. As per the research conducted by Peter A Briss and his team, chronic disease is the leading cause of death in the 21st century[9]. Factors like usage of tobacco, physical idleness, obesity and malnutrition lead to chronic disease. In a recent research paper, 60 percent of deaths occurred between 2 A.M. till 8 A.M.[10] This research had one limitation that it couldn't predict death cause and time for a person aging over 65. So that's why we decided to dig deeper and find the root cause based on data collected from the NYC Open Data website[1].

As the extract API data set is based on criminals' court's summons which is a breakdown of every criminal summons issued in NYC by current data supported by field names and descriptions. The current study demonstrates the offense type like public urination, noise, parks offenses littering, etc. in which adults from age 36-65 years were less appeared in their summons while 16-17 years old child and females responded to summons[11] and show their appearance in civil court almost
62.52 percent, the most offended crime is public consumption but 47.15 percent of them appeared in criminal courts, the borough where these offenses have occurred. In the borough, The Brooklyn is the borough where most criminals appeared in criminal courts and large no of summon issued. The appearance of the male was lower than the appearance of the female which is 54.99 percent. The low-level offenses are less attended by criminals in the courts. The rate of appearance in criminal court is slightly higher than in civil court. The overall appearance in court was (50.98 percent criminal vs 47.11 percent civil see[reference].

IV. Methodology

The shooting data set is a collection of incidents that had occurred in New York City from the year 2006 to the previous year. This data set has 18 columns where each row justifies the shooting incident. The accidental drug-related data set is a collection of data with listings of accidental death associated with a drug overdose. The data have been derived from an investigation by a medical examiner with toxicity reports and scene investigations which lead to death. This data set has 41 columns where each row justifies death. The leading cause of death contains data which is derived from the death certificate that is issued for every death in New York City. It has
7 columns where each of the columns explains the cause, race ethnicity and death rate. The data set NYPD Criminal Court Summons describes the list of Summons issued to any criminal during the current calendar years in New York. The key information data set includes are types of Summon Category Type, Age-groups, Law description, Sex, New York City Boroughs, etc. The reason for choosing this data set is that it has key variables that are useful for its analysis insights and its relationship with the rest of the data sets.

The data sets information provided above has been fetched through API using python language from the open-source web application Jupyter Notebook. The fetched data is in a raw format and it is converted to JSON format to store it in

MongoDB cloud platform. This platform is a fully automated cloud service that has good security and operational function built-in. This unstructured data is pre-processed and then the structured data is stored into the Postgres Database which is running in the virtual machine. The data is further explored and visualized according to the project objectives.

Pandas and NumPy along with other libraries like Matplotlib, Seaborn, and Plotly are used for data processing, analysis, and visualization. Dealing with missing values, feature selection and handling different data types to ensure data is in the correct format with the project objective.

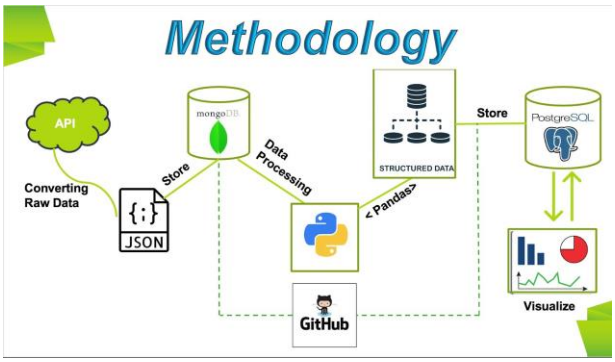

Fig. 1. KDD Methodology

The resource technologies/tools which are used in this project are Python, MongoDB[12], PostgreSQL[13], Jupyter Notebook[14]. Python which is a general-purpose programming language, open-source, variety of libraries for data analytics like Panda and NumPy[15]. MongoDB, a NoSQL database used to store unstructured data and provide a connection driver for multiple programming languages[12]. PostgreSQL, an object-relational database management system, used to store structured database and handy for defining own data types and functional languages, it also has a database driver for python. Jupyter Notebook, a browser-based IDE, which supports Python, Visualization within IDE, good for documentation, creates a cleaner and explainable script. GitHub[16] is a SaaS platform and it is used for version control.

V. Results

Here we look at the year when the shooting incident took place. The states have been displayed as a hue in the count plot and it is easy to analyze the visualization. From the graph, we can note that the highest number of shooting incidence taking place in which borough and in what year. Brooklyn has the highest number of shooting incidents that have taken place in the past years which is followed by the Bronx. Queens is the safest borough as it has uncommon shooting incidence taking place compared to all other boroughs [Fig. 2]

The graph tells us about the location which was targeted in a time zone. It portrays that shooting is observed throughout the day in a public house and apartment building. Based on this analysis, these places should develop a good security and defense system for the safety of the civilians.

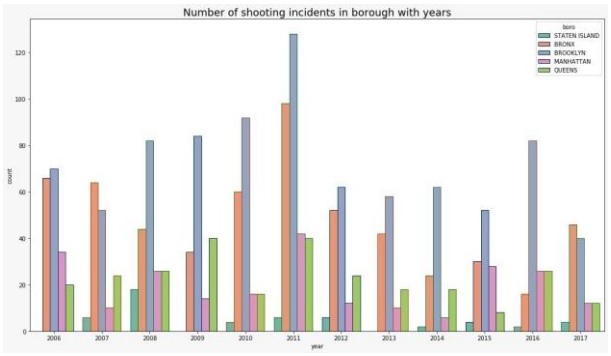

Fig. 2.

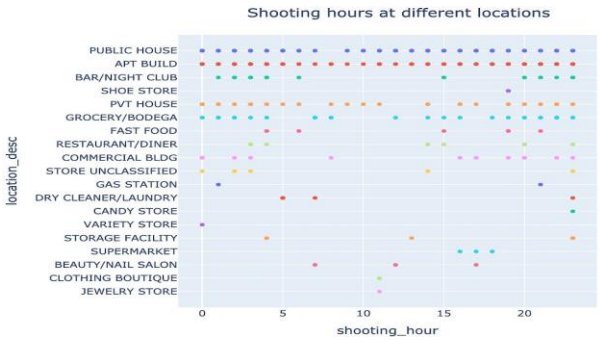

Fig. 3.

Whereas bar/night club and fast food places have observed the shooting incident during the midnight period These shooting hours is very essential to note the shooting incident taking place at targeted locations [Fig. 3].

The below horizontal bar plot shows the total number of deaths, and the total number of deaths count by one drug intake, two drug intake, three-drug intake, four drug intake, five drug intake, and six drug intake. Based on the figure below the highest death rate is by consuming 2 drugs and the total number of deaths by taking 2 drugs is 370, while the least number of deaths are by taking 5 drugs and there are no deaths by consuming 6 drugs [Fig. 4].

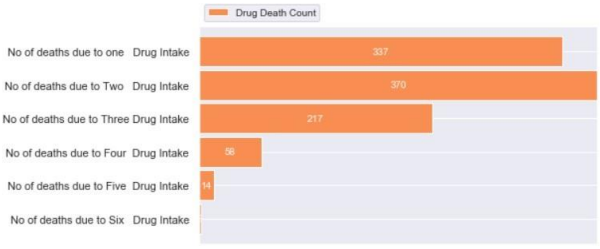

Fig. 4.

The graph is a bar plot where the number of deaths based on gender happening in various locations. Male is dominating females in terms of death in these locations [Fig. 5].

The chart below is in a donut style, every section of them do not figure below describes the various types of drugs and

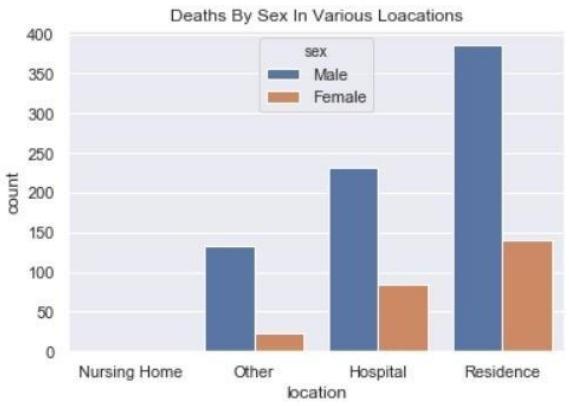

Fig. 5.

their percentages of death rate because of them. Every area of the donut is deaths by consuming that one drug alone [Fig. 6].

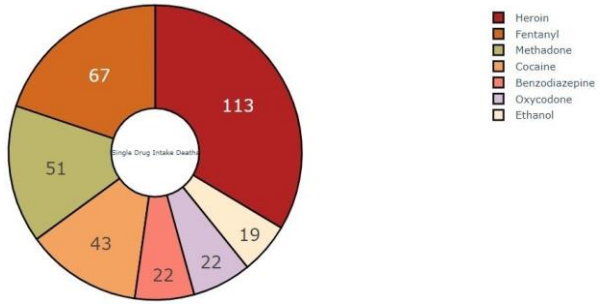

Fig. 6.

The red color is representing the death rate by consuming heroin individually and it is in major amounts compared to all other drugs, whereas ethanol is the least death affected drug which is just 19 in number [fig. 6].

This is a stacked bar plot where we are analyzing Deaths by Gender over the period between 2007 to 2013. Male death count increases gradually every year except for the years 2010 and 2013. Whereas for Females, it was shocking to see no deaths in the year 2007 and post 2007, it kept on rising every year. Following the trend, we can say that more Females will be targeted next year as compared to men [Fig. 7].

This result is cross-validated by plotting a Pie chart which depicts the same results as shown below in figure [Fig. 8].

The below count plot is plotted against no of counts of summons issues to criminals on the y-axis, age-group on the x-axis with a hue of the borough to the right side of the plot. This plot is analyzed into two categories. In terms of age group ages between 25-44 have been called for criminal courts summons. The lower graph is shown for the age of 65+ and the ages less than 18 are less responsible for any criminal violation. The age between 18-24 is placed as the second most category for summons call. On the other hand, if we see our graph with the prospective of the borough. It can be clearly seen that the Bronx

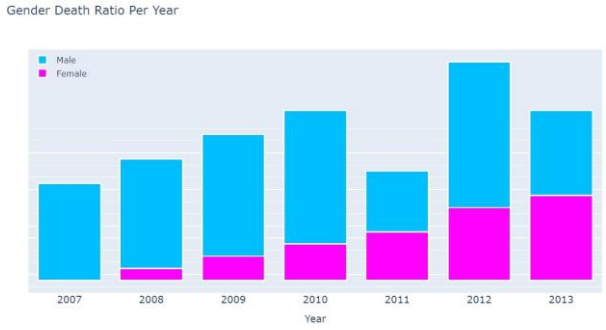

Fig. 7.

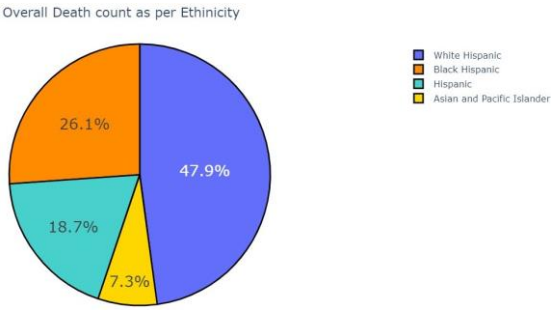

Fig. 8.

share the highest contribution in the count of summons issued and Staten Island participated lowest among all the borough. The rest of the borough fluctuated at different scale as per the age group [Fig. 9].

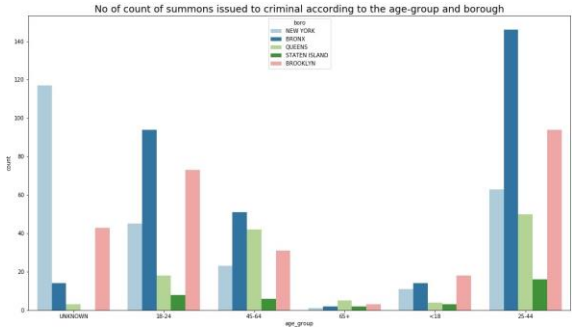

Fig. 9.

The above plot explains the categories of law description and it's no of counts. It is shown in the figure that penal law has a maximum no of counts which is offended by criminals who attended court summons. NYS Transportation and Administrative code share almost identical counts whereas lowest counts law descriptions are Sanitary Code, Tax Law and general Business Law [Fig. 10].

VI. Conclusion

Our main aim in the project was to establish a relationship between different data set that are somehow related to each

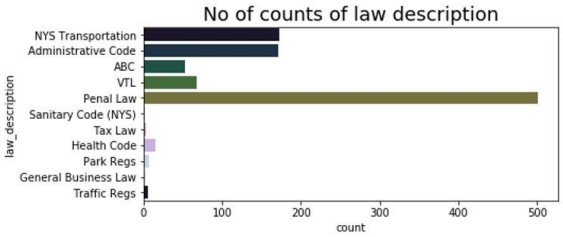

Fig. 10.

other and perform Exploratory data analysis. Looking at the recent trend of deaths happening in the United States of America, it is mainly caused by diseases and overdose of drug intake like cocaine. Also, shooting incidents occur on a frequently basis in locations like Public Houses, Apartment and Private houses at any time of the day. In case of crime events sited at jewelry store usually happen at 1 P.M. day time. Explicit discussion of the research question(s) in the context of your findings We combined the data sets using PostgreSQL and formed complex queries to find the common relationship and the following are the results of it.

First, we plotted category plot, deaths count vs borough (locations) with category as race ("Black Hispanic", "white Hispanic"), here ratio of deaths based on category race-White Hispanic was more than race-Black Hispanic, except for Staten Island where the only White Hispanic race was targeted [Fig. 11].

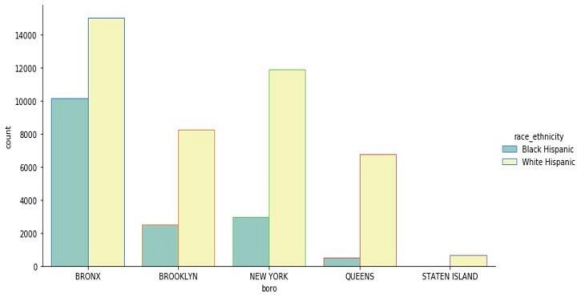

Fig. 11.

Secondly, using count plot we estimated Males who died by an overdose of cocaine intake based on location (Hospital, Residence and other). Hospital has most records and others contain the least record [Fig.12].

Thirdly, we plotted a strip plot to observe deaths happening at different locations, here we observed Staten Island is the safest area as compared to other locations as fewer crime incidents were reported in a day [Fig.13].

Our only limitation was that we're only able to fetch 1000 rows through the public link as no key was required.

If we had more time, we would like to automate the script in order to save time with the execution of code running in the background. We would like to do geospatial analysis to pinpoint the exact location of deaths. Also, we would like to gather all data at once with the Private key.

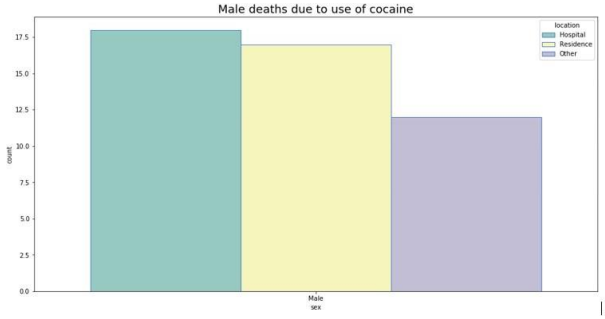

Fig. 12.

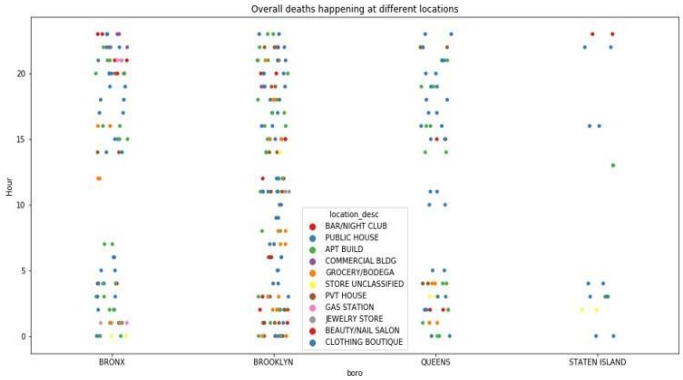

Fig. 13.

REFERENCES

[1] "New York City Leading Causes of Death | NYC Open Data." [Online]. Available: https://data.cityofnewyork.us/Health/New-York-City-Leading-Causes-of-Death/jb7j-dtam. [Accessed: 13-Dec-2019].

[2] "Drug Overdose Deaths | Drug Overdose | CDC Injury Center." [Online]. Available: https://www.cdc.gov/drugoverdose/data/statedeaths.html. [Accessed: 14-Dec-2019].

[3] "NYPD Criminal Court Summons Incident Level Data (Year To Date) | NYC Open Data." [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Criminal-Court-Summons-Incident-Level-Data-Ye/mv4k-y93f. [Accessed: 14-Dec-2019].

[4] "NYPD Shooting Incident Data (Historic) | NYC Open Data." [Online]. Available: https://data.cityofnewyork.us/Public-Safety/NYPD-Shooting-Incident-Data-Historic-/833y-fsy8. [Accessed: 14-Dec-2019].

[5] "Active Shooter 2012 Edition | Emergency Evacuation | Violence." [Online]. Available: https://www.scribd.com/document/117685801/Active-Shooter-2012-Edition. [Accessed: 14-Dec-2019].

[6] J. A. Capellan and J. R. Silva, "An Investigation of Mass Public Shooting Attacks Against Government Targets in the United States," *Stud. Confl. Terror.*, 2018.

[7] J. Blair, "A study of active shooter incidents in the United States between 2000 and 2013.," p. 47, 2014.

[8] T. G. Rhee, J. S. Ross, R. A. Rosenheck, L. E. Grau, D. A. Fiellin, and W. C. Becker, "Accidental drug overdose

deaths in Connecticut, 2012–2018: The rise of polysubstance detection?," *Drug Alcohol Depend.*, vol. 205, Dec. 2019.
[9] H. Paul, Ed., *Critical Terms in Futures Studies*. Cham: Springer International Publishing, 2019.
[10] M. M. Mitler, R. M. Hajdukovic, R. Shafor, P. M. Hahn, and D. F. Kripke, "When people die. Cause of death versus time of death," *Am. J. Med.*, vol. 82, no. 2, pp. 266–274, 1987.
[11] J. K. Doerner and S. Demuth, "Gender and Sentencing in the Federal Courts: Are Women Treated More Leniently?" *Crim. Justice Policy Rev.*, vol. 25, no. 2, pp. 242–269, Mar. 2014.
[12] "MongoDB Cloud Database Solutions | MongoDB." [Online]. Available: https://www.mongodb.com/cloud. [Accessed: 14-Dec-2019].
[13] "PostgreSQL: The world's most advanced open source database." [Online]. Available: https://www.postgresql.org/. [Accessed: 14-Dec-2019].
[14] "Project Jupyter | Home." [Online]. Available: https://jupyter.org/. [Accessed: 14-Dec-2019].
[15] "Welcome to Python.org." [Online]. Available: https://www.python.org/. [Accessed: 14-Dec-2019].
[16] "GitHub." [Online]. Available: https://github.com/. [Accessed: 14-Dec-2019].

APPENDIX

Team Member Workload Distribution

Team Members	Workload Percentage
Imran(17119308)	25%
DevaRaju(18181422)	25%
Harsh(18187340)	25%
Vijit(18199429)	25%

Team Work in achieving goals:

- All members supported each other very well and divided the workload equally.
- We were continuously accessing progress every week and discussing on group chat in Microsoft Teams and Video call conference on Zoom.
- Below were the main task performed by individual team members:
- ***Harsh(18187430):***
    - Performing EDA on the leading cause of death dataset.
    - Documenting the research done.
    - Preparing PPTs for respective topics assigned.
    - Connecting PostgreSQL and defining complex queries.
- ***Vijit(18199429):***
    - Performing EDA on the Shooting dataset.
    - Documenting the research done.
    - Preparing PPTs for respective topics assigned.
    - Connecting PostgreSQL and defining complex queries.
- ***Imran(17119308):***
    - Performing EDA on the Criminal court dataset.
    - Documenting the research done.
    - Preparing PPTs for respective topics assigned.
    - Connecting MongoDB to Python.
- ***Devaraju(18181422):***
    - Performing EDA on Drug dataset.
    - Documenting the research done.
    - Preparing PPTs for respective topics assigned.
    - Connecting MongoDB to Python.