An Investigative Analysis of Disparities in Healthcare Utilization, Suicide and Overdose Death Rate in the United State.

Benjamin Ayodele Kelani
School of Computing
National college of Ireland
x21226181@student.ncirl.ie

Hilal Ozcelik
School of Computing
National college of Ireland
x23218274@student.ncirl.ie

Adeola Deborah Adenijii
School of Computing
National college of Ireland
x23104201@*student.ncirl.ie*

***Abstract*—This study aims to conduct in-depth investigation of the disparity ratio of visit to healthcare facilities, on fatalities in suicide, and overdose death rates across diverse demographics over a certain period in the United States. This study relies on Python as the primary programming language, alongside MongoDB and PostgreSQL as the database system with Luigi pipeline to address the complexity of data preprocessing, automation, transformation, and storage. To provide actionable insights, the investigation analyzes three distinct structured and semi-structured datasets, to identify patterns in the disparity of health utilization and as well as to explore their correlation between death rates by suicide and overdose in the United State. The evaluation would result in appropriate visualizations to facilitate easy understanding and to uncover significant insights.**

***Keywords*— Healthcare Facilities, Disparity Ratio, demographic, MongoDB, PostgreSQL, Luigi Automation, Data Preprocessing, Transformation, Visualization.**

I. Introduction

Considering the alarming increase in drug overdose and suicide mortality and ongoing inequalities in healthcare access. The United States faces a serious public health crisis. To improve overall population health, it is critical to understand the dynamics and the disparities within these three domains. Healthcare utilization refers to how frequently individuals visit hospitals and in what way they seek medical treatments. The information provided here highlights the availability of resources for mental health, chronic illness management and preventive care. While suicide and overdose deaths have been linked to the complex social determinants of health, such as substance abuse and mental health struggle.

Across different demographics groupings, categorized by age, sex, race, and ethnicity have made significant difference in the visit to healthcare outcome and utilization. The identification and resolution of these disparities are crucial in the quest for equality in healthcare usage and ensuring widespread access to high-quality healthcare services and metal health assistance for all. The use of data mining techniques in healthcare has piqued the interest of many Researchers, Organization, and health policy makers. A literature review has been carried out to uncover recent trends and drawbacks using database analytic tools in the real world and existing techniques in healthcare management. The analysis below provides questions that the proposed projects aim to address:

Research Questions

a) What are the trends and patterns in the United States over a certain period in healthcare facilities, suicide rates and overdose death rates?

b) Are there significant disparities in demographics groups of age, gender, race, and ethnicity over time between death rates caused by suicide and drug overdose and access to healthcare facilities?

c) Is evidence indicating that healthcare visits and suicide rates, or overdose mortality and demographic factors like age, gender, race, and ethnicity are correlated or causally related?

II. Related works

[1] A review of healthcare big data management, statistical analysis, and analytical programming is done by Nazir et al. The study relies on papers published between 2015 and 2020. A methodological strategy is used in the study that looks at 127 important documents, including research articles, conference proceedings, book chapters, and reports from surveys. To assist healthcare professionals, make more informed choices, this analysis will look at previous research to find patterns and challenges in handling substantial amounts of medical data. The paper is limited since it uses secondary sources and might not cover the newest modifications or new trends related to healthcare big data management. Nevertheless, it does give an adequate summary of where the field is now. An in-depth comprehension of practitioners' needs and points of view is also rendered more difficult by the absence of first-hand data collection methods like surveys and conversations. More studies in this area that utilize primary data collection methods of large amounts of data could lead to a greater comprehension of the field and how it might impact patient health and outcomes in the future.

In a search made by [2], brought evidence before the study in the increasing recognition that a significant percentage of individuals leaving in the United States suffered deaths which was caused by opioid overdose and assumed to be not deliberate in having a suicidal component. With the use of a reliable data collection of 5386 participants and sources from the National Survey on Drug Use and Health with an implementation method of statistical analysis. Made discoveries that, over the years from 2009 to 2020, the increase in suicidal ideation has drastically increased in the US. Many overdoses and suicidal ideation have been reported in need of healthcare services but did not receive any in the last 12 months. Despite the increasing trends, the provision of mental health services used was not improved. An annual average percentage changed from 22.8% to 29.8% in young adults ranging age 18 to 25 living in non-metropolitan areas had suffered major depression episodes and with no significance of alteration in the healthcare services were observed. These outcomes underscored the prospective benefits of regular investigations into suicidal and overdosed victims. One of the limitations of this study, is that it did not cover an extensive broader trend in the healthcare utilization, suicide, and overdose death rates in the United States. Paying more attention to just young adults aged 18 to 49 with OUD, which limited the discoveries, from a generalized large population. Also, the research made no significant disparities in the demographic groups based on gender, race, and ethnicity, overlooking a wider range of correlation between the difference in mortality rates and healthcare access.

According to research done by Sally C. Curtin, in representing suicide death rates among individuals ranging from age 10 to 24 in the United States [3]. With data collection made available, a total of 100,000 population were computed for each year starting from 2000 to 2018. From a selective year in 2007 to 2018, uncovered a disturbing trend of 57.4% increase in suicide rates nationally among individuals of specified age range 10-24 years old. The analysis included all the states in the US, as the surge of majority of states indicated a spike in the death was not limited to any area. For example, New Jersey had the lowest suicide rate and Alaska the highest. One of the critical limitations to this report lies in the low number of suicides in some states within a given year, posing challenges in accuracy and evaluations in death rates over time. Diving deeper into a more complex analysis because single year assessments were not futile. Selective years average with three years intervals in 2007 to 2009 and 2016 to 2018, helped to gain more insights. Despite this approach, good quality information was yet limited. This was because of the usage of a relatively small sizes sample population, and lack of statistical power to demonstrate significance. A lot would have been accomplished if data collection, analysis methodologies and the inclusion of other related factors were to be prioritized, enhancing statistical reliability and accurate assessment to influence the suicide death rates outcome.

Holly Hadegaard, Margaret Warner, and Arialdi M. Minino conducted a study [4] on drug overdose deaths in United States. They reviewed drug fatalities, and changes in the death rates in the United States from 1999 to 2018. The join point regression program was used to analyze trends in the age-adjusted mortality rates, considering patterns, substance involvement, and variations using data from the National Vital Statistics System Mortality File. It drew attention to the significant increase in drug related facilities from 1999 to 2006 and from 2013 to 2018, despite varying the rates of variations over that period. With differences by location, the age-adjusted rate of drug overdose deaths in 2018 was 4.6% lower than 2017. Gender imbalances in the death rates are also noted in the findings, with males experiencing greater rates than females. Nevertheless, it is important to note that the limitations of this study do not specifically tackle key demographic factors, healthcare access, and other related causes of deaths such as suicides in the United States. It also ignores important aspects that are necessary for comprehensive knowledge of the problem, such as environmental elements that contribute to drug overdose deaths. Consequently, to inform more complete investigations into health-related issues and policies, feature research should strive to combine a wide range of data sources and consider the varied characteristics of drug overdose deaths.

Over the last 20 years, more than 47000 suicides were observed in the US. Increasing numbers of suicide deaths led to perform a lot of studies in these issues nowadays [5] Some of these studies can be seen below. In 2019, Conner Andrew et al. investigated suicide rates in the US from 2007 to 2014 through a nationwide with a population-based approach. The aim of the study is ensuring insights into the rates and suicides for different demographics in the USA during the period. The study provides a valuable information about suicide rates and distribution of suicide cases by analysing data and contributing prevention ways for suicide at a nation level [6]. In 2020, Choi Daejin et al. developed machine learning models to estimate the weekly suicide fatalities in the US, by using many heterogeneous data sources. The model attempts to predict weekly suicide fatalities by utilizing several kinds of data from various sources, including demographic, socioeconomic, mental health, and environmental aspects. A [7] thorough examination of the variables influencing suicide rates is made possible by the integration of several data sources, which produces forecasts that are more accurate. In the end, the model offers insightful guidance to mental health practitioners, legislators, and public health officials on how to promptly execute tactics and interventions meant to deter suicide and advance mental health on a nationwide level.

III. Methodlogy

1. *Rationale for Technology*

A thorough selection of technologies is crucial to the implementation of data analytics solutions. This section provides justifications to the technologies been used to conduct investigative analysis of disparities in healthcare utilization, suicide, and overdose death rate in the united state. Based on their efficiency, easy to use, documentation, as well as community support. As a result, the research will benefit from robust and efficient data analytics by utilizing these technologies. See below the technologies that have been used.

a) **Programming Language: Python** was chosen as the Primary programming language because it is wildly adopted in data analytics community, having extensive libraries and it easiness to use [8].

b) **Integrated Development Environment: Jupyter Notebook** [9] was selected as the IDE as it allows students and developers to explore, analyze and document data interactively.

c) Database: For a robust design in the storage and retrieval of structured and semi-structured data, the use of **MongoDB** and **PostgreSQL** databases [10] where used. This is due to their versatility and reliability.

d) **Python Packages:** [8] Multiple packages were used for data manipulation, analysis, and visualizations. They provided powerful tools used for handling data and conducting advanced analytical tasks. Packages and libraries such as Pandas, Matplotlib, Seaborn, and Scikit-learn were used.

e) **Automation Pipeline: Luigi** was utilized to create an automation pipeline to streamline efficiency workflow of the investigation during the analysis.

2. *Dataset Description and Justification*

a) **Healthcare Trips Dataset:** The datasets comprise of information on visits to physician offices, hospital outpatient departments and hospital emergency department from 1950 to 2018 in the United State. Healthcare Visit USA.. The data is categorized by age, sex, and race. The data was extracted and loaded through XML (Extensible Markup language) format. It contains 3571 numbered rows, and 16 columns. The inclusion of the healthcare trip data is justified as it provides patterns for the utilization of healthcare facilities across different demographic groups over time. Using the data to analyze trends in health visits can provide insights into disparities across healthcare services and healthcare policies to bring informative decisions and resources allocations.

b) **Death Rate for Suicide Dataset:** This dataset involves information for death rates for suicide, by drug type, sex, age, race, and Hispanic origin in the United States from 1950 to 2018. The data was downloaded through a CSV (Comma Separated Value) file format. Death Rate for Suicide USA. It contains 6390 numbered rows and 13 columns. The inclusion of the death rates for suicide is justified as it allows an in-depth analysis of its correlation between various demographics and as well as disparities with other cause of deaths. It also provides an understanding of access to healthcare and possible intervention strategies in the United States. It also analyzes the disparities between the health visits over time.

c) **Drug Overdose Death Rate Dataset:** This dataset contains information about Drug Overdose Death Rates, by Drug Type, Sex, Age, Race, And Hispanic Origin in the United State from the year 1999 to 2018. Data collection was done through a URL of a CSV file but converted to JSON (JavaScript Object Notation) immediately. Overdose Death Rate USA . It contains 6229 numbered rows and 15 columns. The inclusion of drug rate for overdose dataset is justified as it shed more lights in the ratio of deaths caused between various demographics, and its correlation with the suicides rates caused in the United States. It also analyzes the disparities between the health visits over time.

3. *Data Processing Activities*

Series of data processing activities were carried out during the investigative analysis of disparities in healthcare utilization, suicide, and overdose death rate in the United State. The data process activities were not done sequentially, as each was analyzed due to its complex characteristic but used a single method.

a) **Data Collection:** This section involves fetching each dataset from their relevant sources in Data.Gov and using a suitable structured and semi-structured format. Healthcare visits and suicide rate data were downloaded and saved into the python environment, while overdosed death rate data used URL request to fetch the data directly into the environment. The XML and CSV files that were downloaded were initially saved into python environment then immediately converted into dictionaries formart. While the URL CSV file was first converted into JSON format before dictionaries formart.

b) **Data Automation:** An automation pipeline was utilized to execute complex workflow. With Luigi, tasks such as downloading and converting data, loading data into various databases, and performing data analysis and transformations were automated to reduce manual intervention and potential errors.

c) **Data Extraction:** With the help of Luigi automation pipeline, the extraction process to retrieve the converted dictionaries, i.e. XML, CSV and JSON files were made easy. Each dataset was extracted into a Pandas data frame for easy manipulation and analysis to be made possible.

d) **Data Storage:** Initially, the data converted as dictionaries were housed in MongoDB database. It enabled flexibility and scalability, allowing easy retrieval for future analysis. MongoDB document-oriented structure suited the semi-structure nature of the data as well. Afterwards, the clean data was then transferred into another database called PostgreSQL. It was selected for its robust relational database management system, making it satisfactory for storing processed data.

e) **Data Cleaning and Transformation:** A considerable and well thought of different data cleaning and transformations were carried out due to the complexity of each data frame. For example, redundant columns that did not add much important value to the analysis were systematically removed, while missing values and NAs were checked for, and were either replaced with the target mean value or omitted entirely. Furthermore, outliers were detected to identify anomalies, ensuring data consistency and reliability. Other methods were also employed, such as Feature Selection (Example: Aggregated the estimated death rate by age group and year), Data Encoding (Example: male as 1s and female as 0s), and Feature Engineering Categorization (Example: split Race demographic from Sub_Label). All this was done to prepare the data for quality analysis.

f) **Data Visualization:** The use of visual representations played a large role during the analysis as patterns and insights were made possible and with ease. To effectively uncover trends and patterns in the distribution of disparities as well as correlations in the healthcare visits, suicide rates and drug overdose death rates. The use of various visualizations such as Pie charts, Bar plots, stacked bar plot, line plots and many others were created.

Below shows the general overview diagram of data processing activities used to answer the above research questions.

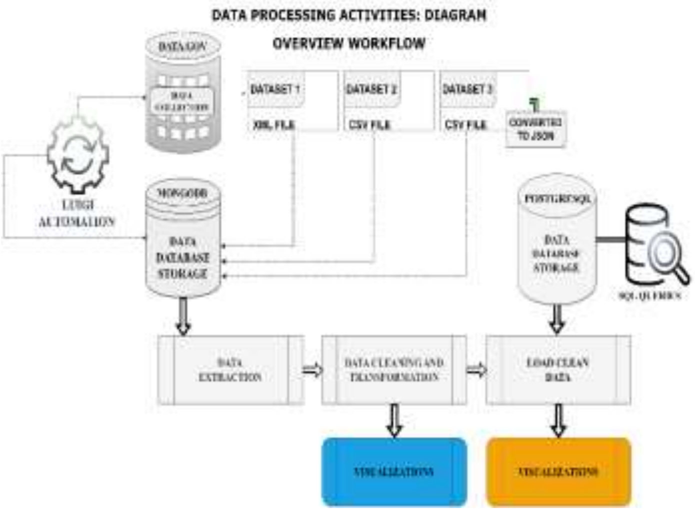

Figure 1: Data Processing Activities Overview.

IV. Results and Evaluations

a) In the search to find the trends and patterns in the United States over a certain period in healthcare visits, suicide rates and overdose death rates. The healthcare data provides information on individuals living in the United States who seek for healthcare access and their visits in utilizing those facilities. Visits are made to physical offices, hospital outpatient departments, and hospital emergency departments by a selected population in the US. The demographic characteristics of individuals who utilized the healthcare facilities are grouped based on age, sex, and races. Over the years, the total estimates of these visits made to the healthcare facilities are seen in figure 2. In 2000 up to 2011, it showed a high usage of the facilities, as it drastically reduced from 2012 to about 13314 in usage and then began to increase to 143348.10 in 2016.

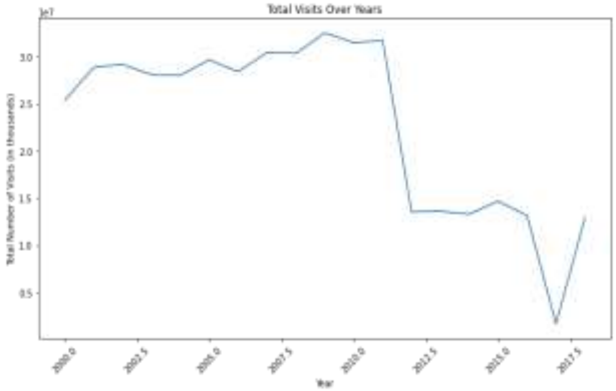

Figure 3: Healthcare visit over the years (2000 - 2017).

For suicide data, it provides information on victims in the United States who died by suicide from the year 1950 to 2018. The demographic characteristics of these victims varies from age, gender, race, and Hispanic origins. As seen from figure 4., the estimated death rates from 1950 was on a high level with about 15,000 deaths, not until 2000, which showed a down trend of about 11,000 deaths and has spiked up greatly in 2010 with about 12,8461deaths and still rising.

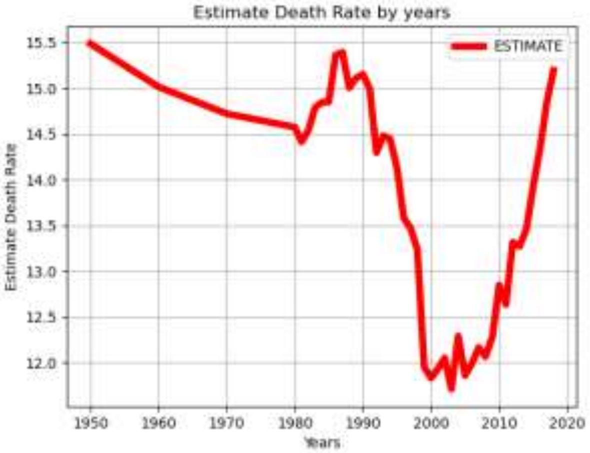

Figure 4: Suicide death rate over the years (1950 - 2020).

Drug overdose data also provides the same information and demographic characteristics as suicide rate data. The difference is that the victims died from drug overdose and not suicide. As seen in figure 5., the total estimate of individuals who died from drug overdose decreased in 2000 with about 2376.9 deaths down to 2007. Death's rate had a high peak in 2015 of about 2685.791 deaths, but did not exceed the early 2000s, and has been showing downtrends.

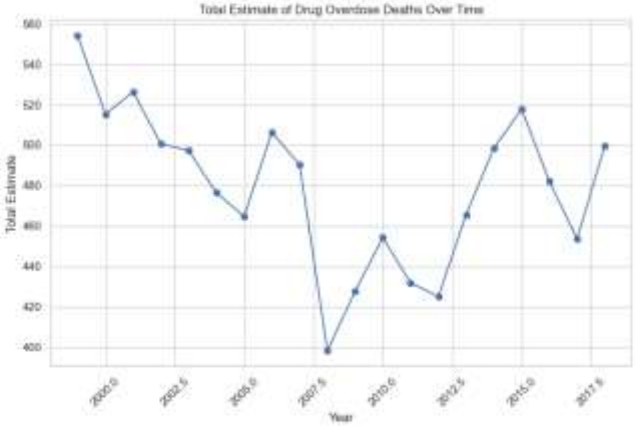

Figure 4:Death rates for drug overdose over the years (2000 - 2017).

Overall, the utilization of healthcare facilities fluctuated, with a noble increase in visits in the early years and followed by a drastic decrease. There was an initial decrease in suicide rates during the late 1990s, followed by a sharp increase in 2000s. Similarly, drug overdose death rates fluctuated overtime, showing peaks and troughs in the level of mortality brought by substance abuse. In general, these trends highlight the importance of targeted interventions to address mental health problems, substance abuse and the disparities in healthcare outcomes and access.

To investigate if there are significant disparities in demographics groups of age, gender, race, and ethnicity over time between death rates caused by suicide and drug overdose and access to healthcare facilities. Figure 6 shows variation and trends in healthcare visits with rates staying the same for certain categories and going down for others. This could suggest there are eliminations in healthcare use or modifications to the graph.

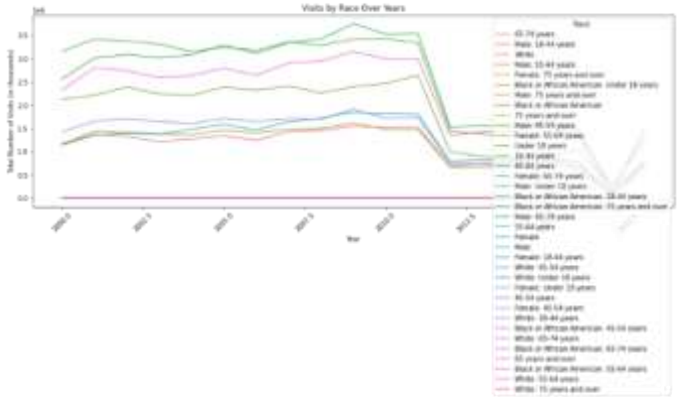

Figure 6: Healthcare visit by Race, Gender and Age-groups.

To see if there are disparities in the demographics groups of age, gender, race, and ethnicity over time. As seen in figure 7., across most age groups, female had a lower trend than male as the deaths rate increased with age.

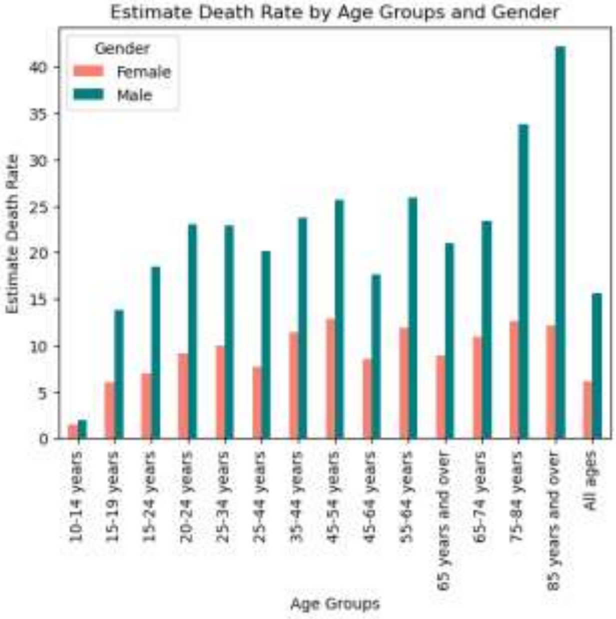

Figure 7: Suicide Rate by Age-groups.

According to the bar chart created in figure 8. Overdose deaths by gender and age group are distributed. The drug overdose deaths seen on the stack bar chart occurred more in individuals aged 85 years and over. In addition, males in the united state have higher mortality rates.

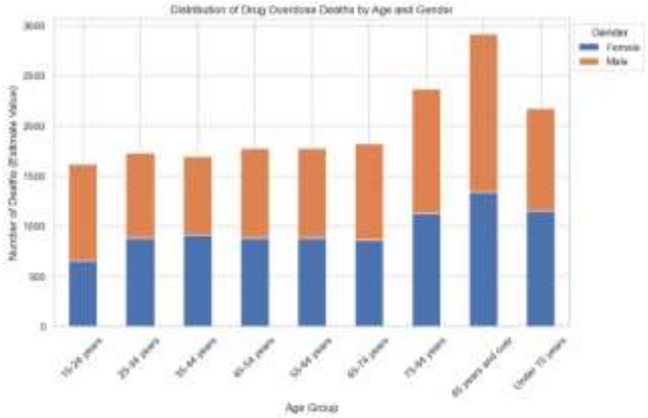

Figure 8: Overdose Death Rates by Gender and Age-groups.

Overall, the investigation uncovered significant disparities between suicide and drug overdose death rates and access to healthcare facilities across demographic groups including age, gender, and race.

b) There is evidence indicating that healthcare visits and suicide rates, or overdose mortality and demographic factors like age, gender, race, and ethnicity are correlated or causally related. Take for instance, the visit panel in figure 9., displays the healthcare data showing the percentage numbers of utilization visits of individuals who seek medical treatments. The visit panel experiences very low percentages during the analysis.

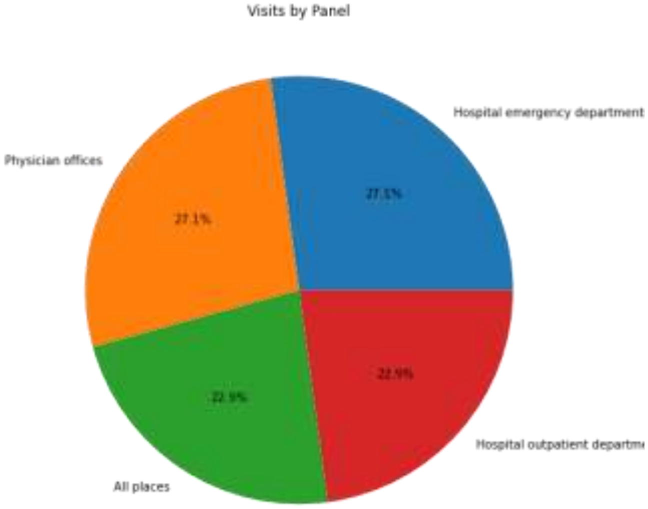

Figure 9: Healthcare Visit Panel.

Pie charts were used to show the correlation distribution of gender in male and female in the data for suicide rates and drug overdose death rate. Figure 10., displays the distribution of suicide rates as female having the highest number of deaths with over 53.3%. Unlike for drug overdose data, where the distribution of genders, male having a higher death rate of over 51.4% and female with a lower percentage of 48.6% deaths in the United States.

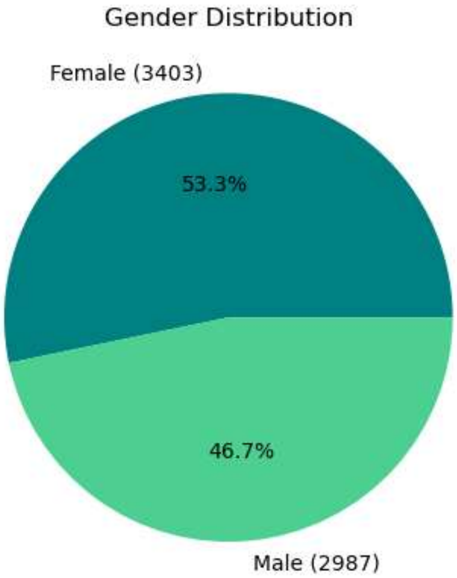

Figure 10: Distribution of Gender for Suicide Death Rates in the United states.

Another pie chart was used to illustrate the distribution of drug overdose deaths among different racial groups in the United States as showed in figure 11.

The evidence presented shows that there is a correlation between the three-domain used in this analysis, as well as revealing distinct patterns indicating relationships.

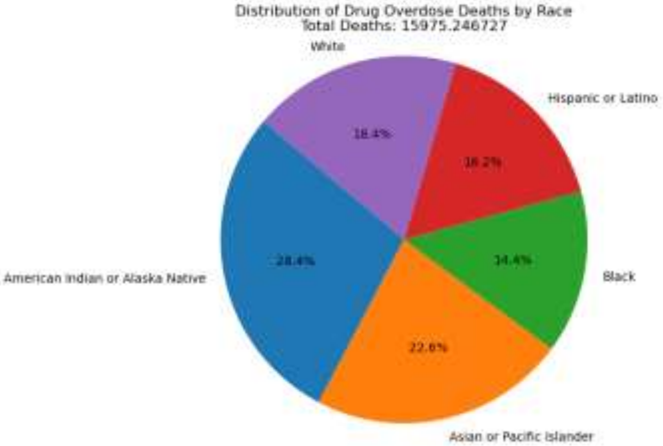

Figure 11: Distribution of Gender for Suicide Death Rates in the United states.

V. CONCLUSIONS AND FUTURE WORK

In summary, the studies reveal hospital and outpatient visit has an inversely proportional relationship with the suicide rate and death rate. Further visualizations also reveal certain races and age groups visit these health facilities less frequently and this affects the suicide and drug overdose rate. Gender targeted solutions are also suggested as there exists a gender-based relationship to deaths by suicides and drug abuse. This would require the Government, health Organization and Policy makers to address the disparities which exist with respect to substance abuse, mental health disorders and unequal access to healthcare. The limitation of this research paper is that it lacks the use of secondary data sources would have had more impact on healthcare investigations. For future improvements, research should encompass the integrating supplementary data sources,

VI. Bibliography

[1] S. K. S. K. H. A. S. G.-M. I. A. R. a. N. M. Nazir, "A comprehensive analysis of healthcare big data management, analytics and scientific programming," IEEE Access, 8, pp.95714-95733., 2020.

[2] T. J. B. I. L. P. a. T. G. R. Peter J. Na, "National trends of suicidal ideation and mental health services use among US adults eith opioid use disorder, 2009 - 2020," Elsevier Ltd. , USA, 2022.

[3] S. C. Curtin, "State Suicide Rates Among Adolescents and Young Adults Aged 10 -24: United State 2000 to 2018," National Vital Statastics Reports; V. 69, NO.11, 2020.

[4] M. W. a. A. M. M. Holly Hadegaard, "Drug Overdose Deaths in the United States, 1999 - 2018," NCHS Data Brief, no 356. National Center of Heath Statistics., 2020.

[5] D. P. e. al, "Predicting state level suicide fatalities in the United Death with real-time data and machine learning.," NPG Mental Health Research, vol. 3, 2024.

[6] D. A. a. M. M. A. Conner, "Suicide case-fatality rates in the United States, 2007 to 2014 a nationwide population-based study," Ann Intern Med, vol. 171, no. 12, pp. 885–895, 2019.

[7] D. C. e. al, "Development of a Machine Learning Model Using Multiple, Heterogeneous Data Sources to Estimate Weekly US Suicide Fatalities," JAMA Netw Open, vol. 3, no. 12, 2022.

[8] W. Mckinney, "Data Structure And Algorithms in Python," Packt Publishing Ltd. , 2010.

[9] Y. Z. B. L. Y. T. Mingke Yang, "On code Reuse from StackOverflow: An Exploratory Study on Jupyter Notebook," Softw Pract Exper , China , 2023.

[10] M. I. A. A. Linggis Galih, "Performance analysis of Neo4j, MongoDB, postgreSQL, on 2019 National Election Big data Managment Database," National Conference of Science in Technology, 2022.