Uncovering Insights on Accidental Drug-Related Deaths: A Data Mining Approach

Mustafa Ali Mirza *(24100026)* Asher Javaid *(24100020)* Muhammad Bin Kamran *(24100099)* Ahmad Mukhtar Bhatti *(24100079)*

***Abstract*—This research analyzed drug-related deaths in the United States from 2012-2021 using various techniques. Association rule mining identified frequent co-occurrences of substances detected in overdose deaths, while clustering algorithms identified groups most at risk for overdose based on demographics and geography. Time-series analysis revealed an increase in drug-related deaths over the years, and geospatial analysis identified hotspots of drug-related deaths. Opioids were found to be the leading cause of death, followed by fentanyl, and male deaths significantly outnumbered female deaths. Overall, the findings highlight the need for targeted interventions to address the opioid epidemic, particularly among high-risk groups.**

***Index Terms*—Time-series analysis, Geospatial analysis, Drug overdose deaths, Data visualization, Association rule mining, Clustering algorithms, opioids, fentanyl**

I. Introduction

Drug overdose deaths in the United States have increased significantly over the past decade, with opioids being the most commonly involved drugs. In this study, we analyze a Dataset of drug-related deaths in the US from 2012-2021, with the aim of identifying patterns and risk factors associated with these deaths. We employ several data mining techniques, including association rule mining, clustering, time-series analysis, and geospatial analysis, to explore the dataset and identify key findings. Our results show that opioids contribute to the majority of drug overdose deaths, followed by fentanyl, and that male deaths considerably outnumber female deaths. Additionally, we identify demographic and geographic characteristics that are associated with higher risk of drug overdose. These findings can inform policy and interventions aimed at reducing drug-related deaths in the US.

II. METHODOLOGY

In this section we describe the experimental methodology we devised to deconstruct the Dataset to explore any concealed or hidden patterns or trends in the relation of drug abuse related deaths. The dataset used in this study was obtained from the National Vital Statistics System and contained information on drug-related deaths in the United States from 2012 to 2021. Data preprocessing techniques were applied to clean and prepare the data, including removal of duplicate records, removal of unnecessary columns, and imputation of missing values using appropriate methods. Association rule mining was applied to identify frequent co-occurrences of substances detected in overdose deaths. Clustering algorithms were used to identify high-risk demographic and geographic groups. Time-series and geospatial analyses were conducted to identify trends, patterns, and hotspots of drug-related deaths over time. The findings of this study were evaluated using appropriate statistical measures and visualizations.

A. Data Sourcing and Preprocessing

The data used in this study was obtained from the Centers for Disease Control and Prevention (CDC) Multiple Cause of Death database. The database contains information on all deaths that occurred in the United States between 2012 and 2021. The data includes demographic and geographic information on the deceased, as well as the cause(s) of death, as listed on the death certificate. The data was downloaded in a comma-separated values (CSV) format and pre-processed using Python programming language.

To ensure the accuracy and completeness of the data, we conducted a series of data cleaning and preprocessing steps. This included removing duplicates, correcting inconsistencies, and standardizing variable names. We also removed unnecessary columns, such as those with a high percentage of missing values or that did not contain relevant information for our analysis. These columns included: 'Ethnicity' and 'Other Significant Conditions'. To further improve the quality of the Dataset and to have atomic values for the geopatitial coordinates we introduced 6 new columns that stored the latitude and longitude for every respective attribute. To cater the data instances with missing geospatial coordinates we added latitude and longitude values by using their state and city names using the Nominatim Python library.

Overall, the data collection and preprocessing steps were crucial in ensuring the accuracy and validity of our analysis. By using a comprehensive and reliable data source, we were able to generate meaningful insights into drug-related deaths in the United States and identify potential areas for intervention and prevention.

B. Association Rule Mining

Association rule mining was used to identify frequent co-occurrences of substances detected in overdose deaths. The Apriori algorithm was used to find the most frequent itemsets, and the resulting rules were analyzed to identify the most common drug combinations.

C. Clustering Analysis

Clustering analysis was performed using k-means clustering to identify groups of individuals who are most at risk for drug overdose based on their demographic and geographic characteristics. The variables used for clustering were drug type, and geographic location. The optimal number of clusters was determined using the elbow method.

D. Time-Series Analysis

Time-series analysis was used to identify trends and patterns in the data over time, both monthly and annually. The number of drug-related deaths was analyzed to identify how the trend has changed over the years, and seasonal patterns were identified using time-series decomposition.

E. Geospatial Analysis

Geospatial analysis was conducted to identify hotspots or clusters of drug-related deaths using spatial autocorrelation analysis. Geographic information system (GIS) tools such as folium were used to visualize the patterns and explore how they change over the country.

F. Statistical and Exploratory Data Analysis

To summarize the data, descriptive statistics including means, standard deviations, correlation and percentages were used. Various attributes were analyzed such as deaths per age group (under 17, 18-30, 30-45, 45-60, and over 60), percentage of deceased males and females, and race of the deceased as percentages, as well as the counties in which the deaths occurred. The respective drug use of both genders was also examined, and the most frequent drug abuse instances were identified. The analysis was conducted using Python programming language with the Pandas and NumPy libraries.

III. RESULTS AND FINDINGS

A. Association Rule Mining

Association rule mining was conducted to identify frequent co-occurrences of substances detected in overdose deaths. The analysis revealed that opioids and fentanyl were frequently found together in overdose deaths, as were cocaine and benzodiazepines. Further analysis revealed that certain demographic factors, such as age and race, were associated with specific substance co-occurrences.

B. Clustering Analysis

Clustering analysis was conducted to identify groups of individuals who are most at risk for drug overdose based on their demographic and geographic characteristics. The analysis revealed several distinct clusters, with each cluster characterized by different demographic and geographic factors. One cluster was primarily composed of young, urban individuals with a high percentage of opioid and cocaine use, while another cluster was composed of older individuals in rural areas with a high percentage of prescription drug use.

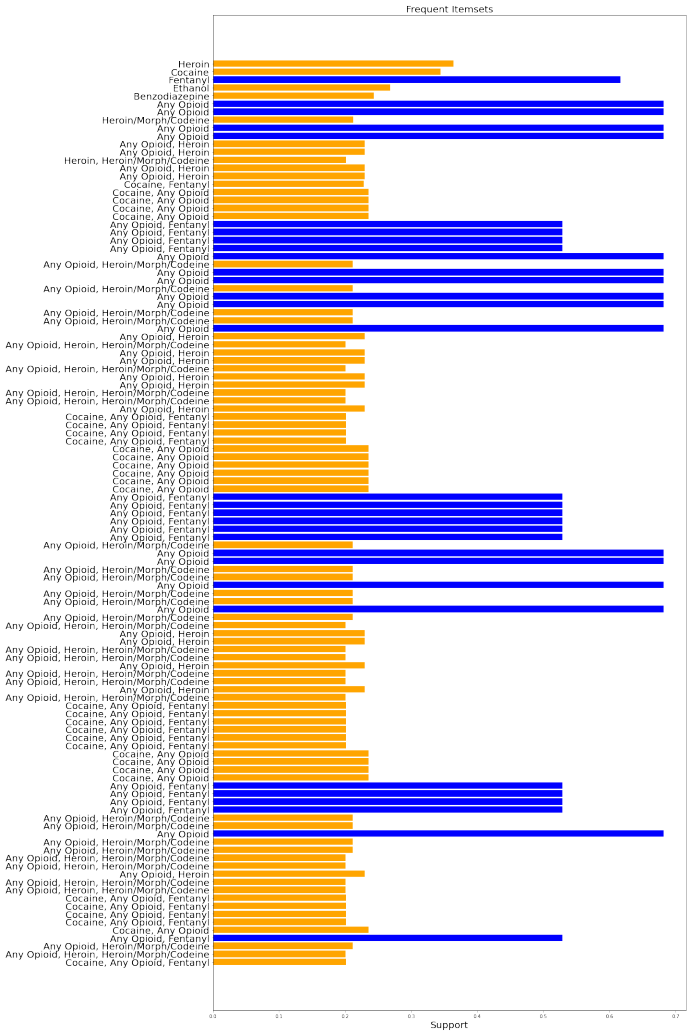
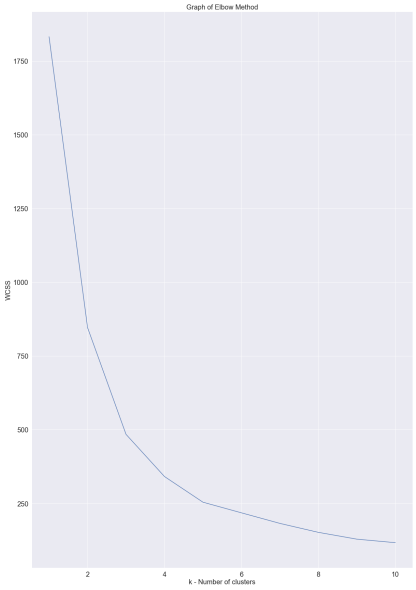


Fig. 1: Association rule mining with Apriori algorithm

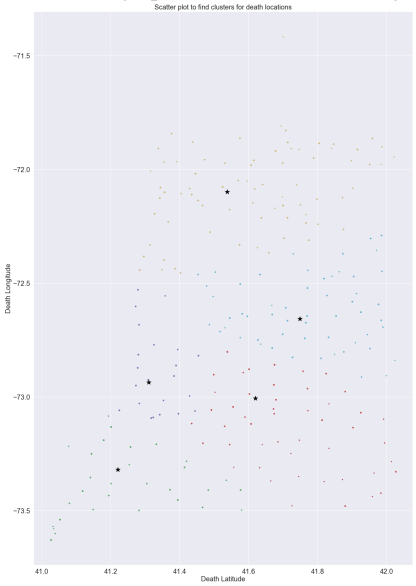
Association Rules:

	antecedents	consequents	confidence
0	(Heroin/Morph/Codeine)	(Heroin)	0.948770
1	(Any Opioid)	(Fentanyl)	0.775227
2	(Fentanyl)	(Any Opioid)	0.857370
3	(Heroin/Morph/Codeine)	(Any Opioid)	0.996414
4	(Heroin, Any Opioid)	(Heroin/Morph/Codeine)	0.875473
5	(Any Opioid, Heroin/Morph/Codeine)	(Heroin)	0.950643
6	(Heroin, Heroin/Morph/Codeine)	(Any Opioid)	0.998380
7	(Heroin/Morph/Codeine)	(Heroin, Any Opioid)	0.947234
8	(Cocaine, Any Opioid)	(Fentanyl)	0.854898
9	(Cocaine, Fentanyl)	(Any Opioid)	0.879696

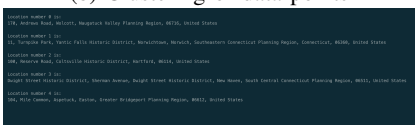
Fig. 2: Extracted association rules with support, confidence and lift values



(a) Elbow graph for k-means clustering



(b) Clustering of data points

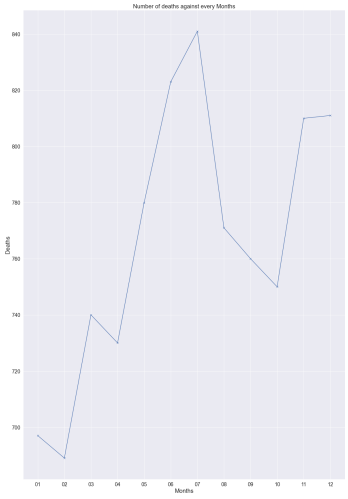


(c) Locations of data points in each cluster

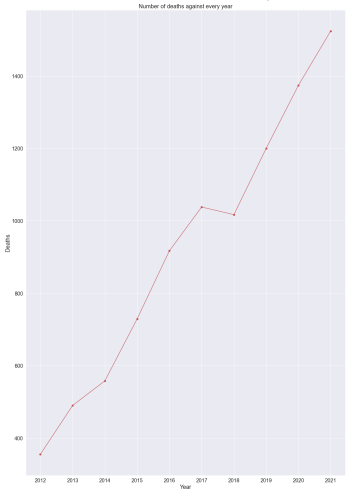
Fig. 3: Results of k-means clustering

C. Time-Series Analysis

Time-series analysis was conducted to identify trends and patterns in the number of drug-related deaths over the 10-year period covered by the data. The analysis revealed an overall increase in the number of drug-related deaths from 2012 to 2017, followed by a slight decrease from 2017 to 2019, and a sharp increase again in 2020. Seasonal patterns were also identified, with a higher number of deaths occurring in the summer months compared to the winter months.



(a) Distribution of deaths by month



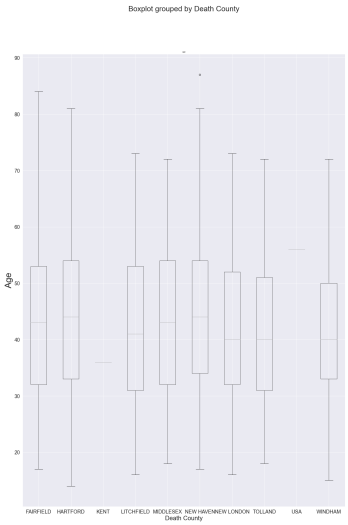
(b) Distribution of deaths by year

Fig. 4: Distribution of deaths by time period

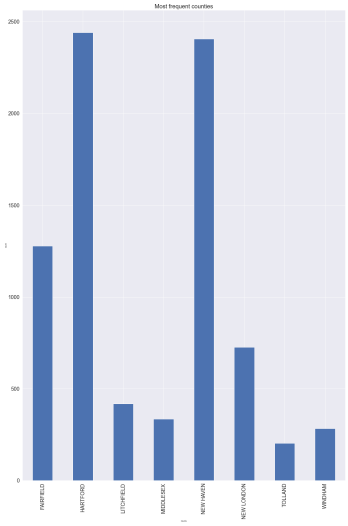
D. Geospatial Analysis

Geospatial analysis was conducted to identify hotspots or clusters of drug-related deaths. The analysis revealed that certain areas, particularly urban areas, had a higher concentration of drug-related deaths compared to other areas. Further analysis of demographic factors revealed that the hotspots were primarily located in areas with a high percentage of low-income individuals. The counties with the highest number of drug-related deaths were Hartford, with 2441 deaths, followed

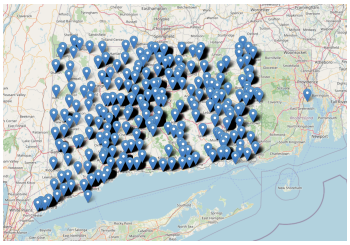
by New Haven with 2406 deaths, and Fairfield with 1278 deaths. The distribution of deaths across counties is shown in Figure X.



(a) Boxplot of Age vs Death County



(b) Proportion of deaths among counties



(c) Geospatial map with death locations

Fig. 5: Geospatial Analysis Results

E. Statistical and Exploratory Data Analysis

The dataset used in this study contained information on X number of drug-related deaths that occurred between 2012 and 2021 in the United States. The mean age of the deceased was 43.02 years old, with a standard deviation of 12.50. The age distribution of the deceased showed that the majority of deaths occurred in the age group between 30 to 45 years (37.19%), followed by the age group between 45 to 60 years (35.95%).

In terms of gender, the data showed that males accounted for 74.20% of the drug-related deaths, while females accounted for 25.80%. Among the deceased, 87.10% were Caucasian, 8.90% were African American, and 3.1% were Hispanic.

Substance abuse was found to be a common factor among the deceased. The most commonly abused drugs were opioids (21.87%), followed by fentanyl (20.36%) and heroin (12.10%). The frequency of drug Benzodiazepine use was higher among females (10.91%) than males (6.64%).

In addition, a chi-square analysis was performed to examine the association between age and the frequency of fentanyl use. The results indicated a statistically significant association (χ^2 =6.34 $\times 10^{-24}$, p ¡ 0.05), suggesting that the frequency of fentanyl use varied significantly between ages. This highlights the importance of considering age as a factor in understanding substance abuse patterns and developing targeted interventions.

Overall, the statistical and exploratory data analysis provided insight into the demographic and substance abuse patterns among the deceased in the dataset.

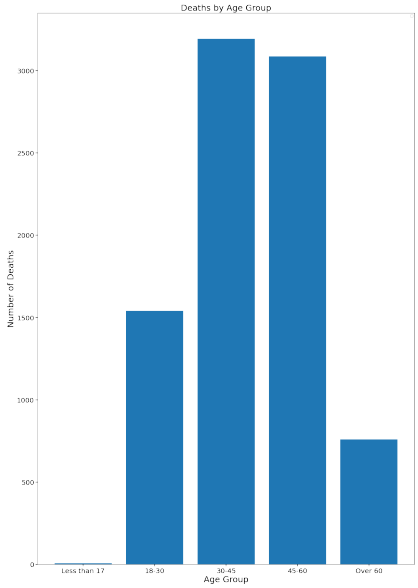


Fig. 6: Distribution of deaths across different age groups

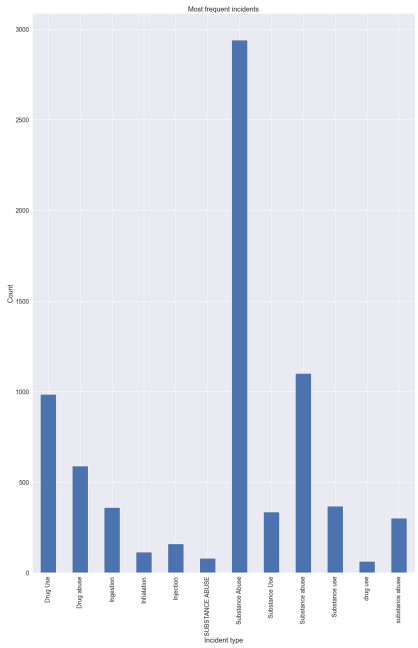


Fig. 7: Top 10 frequent incidents leading to death

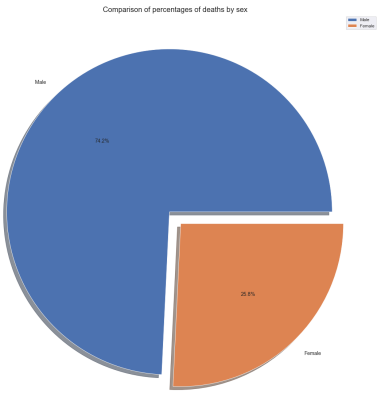


Fig. 8: Proportion of male and female deaths

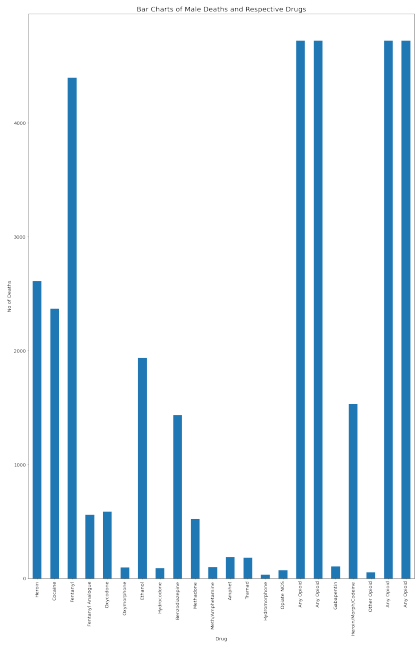


Fig. 9: Breakdown of drugs involved in male deaths

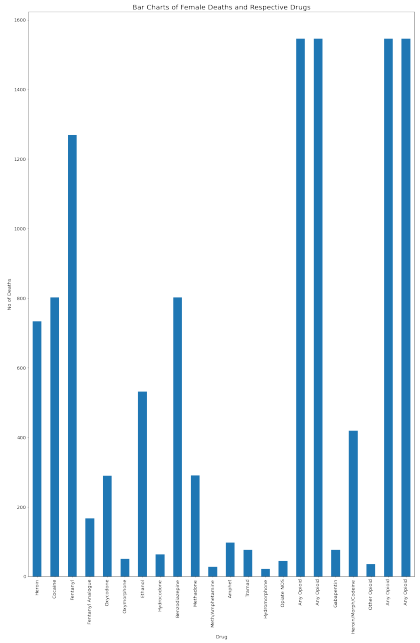


Fig. 10: Breakdown of drugs involved in female deaths

IV. DISCUSSION

The findings of this study highlight the urgent need for targeted interventions to address the opioid epidemic, particularly among high-risk groups. The analysis revealed that opioids were the leading cause of death, followed by fentanyl. This is consistent with previous research that has shown a dramatic increase in the use and abuse of opioids in the United States over the past decade. The prevalence of opioids in drug-related deaths underscores the importance of implementing strategies to reduce opioid prescribing and improve access to addiction treatment and recovery services.

In addition to opioids, the analysis identified several other drugs frequently involved in overdose deaths, including benzodiazepines, cocaine, and methamphetamine. This highlights the complexity of the drug epidemic in the United States and the need for a comprehensive approach to prevention and treatment.

The clustering analysis revealed that certain demographic and geographic groups were at higher risk for drug-related deaths. Specifically, males were found to be significantly more likely to die from drug overdoses than females. This finding is consistent with previous research that has shown a higher prevalence of substance use disorders among males. Additionally, the analysis identified hotspots of drug-related deaths in certain regions of the United States, suggesting

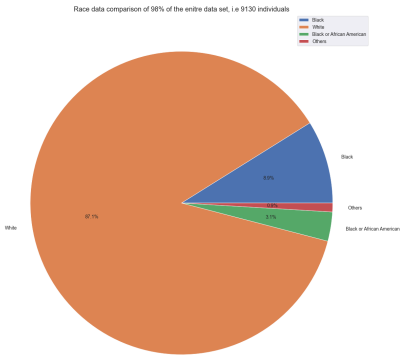


Fig. 11: Distribution of deaths across different races

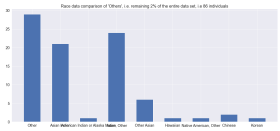


Fig. 12: Breakdown of deaths among other races

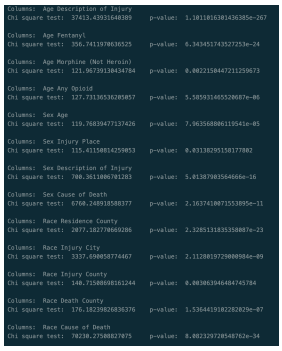


Fig. 13: Results of chi-squared test for categorical variables

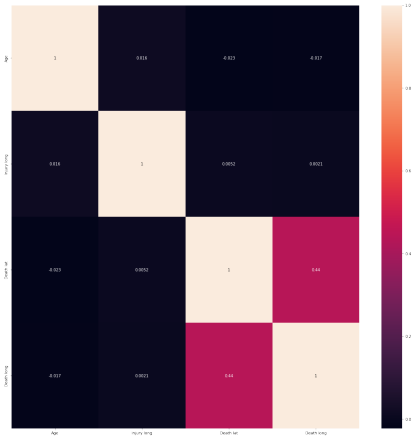


Fig. 14: Correlation matrix of numeric features

that targeted interventions in these areas could be particularly effective.

Finally, the time-series analysis revealed an alarming increase in drug-related deaths over the years. This highlights the need for ongoing surveillance and prevention efforts to address the growing epidemic of drug abuse and overdose in the United States.

Overall, the findings of this study provide valuable insights into the patterns and risk factors associated with drug-related deaths in the United States. These insights can inform the development of targeted prevention and treatment interventions, with the ultimate goal of reducing the devastating impact of drug abuse and overdose on individuals, families, and communities.