Visualization of Drug Poisoning Mortality in the U.S. using Dash Plotly

Sameer Jung Subedi[*1], Sandesh Dahal[*2]

Department of Computer Science and Engineering, Kathmandu University
Dhulikhel-4, Kavrepalanchok, Nepal

[1]ss02435318@student.ku.edu.np

[2]sd02431718@student.ku.edu.np

***Abstract*—Data in the world is growing at an unprecedented rate and without any graphical aids, such a large amount of data is difficult to comprehend. Whether simple or complex, the right visualization can bring everyone on the same page, regardless of their level of expertise. In recent years, the cases of deaths by drug poisoning have been rapidly increasing around the world. The United States of America is a major contributor to this statistic. Since 1999, the cases of death by drug poisoning have increased by more than three times by 2015. The goal of this study is to factually represent the findings according to required classes and generate possible solutions to decrease such undesired cases. The selected dataset contains records for drug poisoning mortality in the United States, therefore, steps like data preprocessing, exploratory data analysis, and finally visualization are performed on the dataset in order to obtain and visualize the drug poisoning mortality in the U.S. The visualization is performed using Python's Plotly Dash.**

Keywords—data visualization, data mining, data cleaning, exploratory data analysis, dataset

I. Introduction

Drug overdoses can be accidental or intentional. They occur when a person takes more than the medically recommended dose. However, some people may be more sensitive to certain medications, so the low (more dangerous) end of a drug may be toxic for them; a dose that is still within the range of acceptable medical use may be too much for their bodies to handle. Illicit drugs, used to get high, may be taken in overdose amounts when a person's metabolism cannot detoxify the drug fast enough to avoid unintended side effects. Exposure to chemicals, plants, and other toxic substances that can cause harm are called poisonings. The higher the dose or the longer the exposure, the worse the poisoning. Examples include carbon monoxide poisoning and mushroom poisoning. Alcohol, medications and some herbal remedies can also cause harm when taken too much. The risk of overdose also increases when more than one of these substances are taken at a time, or if the body is not used to taking a substance.

The U.S. government does not track death rates for every drug. However, the National Center for Health Statistics (NCHS) at the Centers for Disease Control and Prevention collects information on deaths involving many of the more commonly used drugs available throughout the years in a searchable database. [1]

II. Technical Background

The dataset obtained from NCHS is complicated yet interesting due to the usability of the attributes and values in many interpretable ways.

The data was obtained from the website of National Center for Health Statistics which is a US based government agency. It provides statistical information to guide actions and policies to improve the public health of the American people. It is also a unit of the Centers for Disease Control and Prevention and a principal agency of the U.S. Federal Statistical System.

The data set contains the records of mortality due to drug poisoning from the US and by using Python programming language and Plotly Dash framework of Python, which is an open source framework for building data visualization interfaces, visualization of the data from the dataset is performed.

A. Dash Plotly

Dash is a framework for building data apps in Python. Dash apps give a point-&-click interface to models, vastly expanding the notion of what's possible in a traditional "dashboard." With Dash apps, data scientists and engineers put complex Python analytics in the hands of decision-makers and operators [2].

B. The NCHS Data Set

The NCHS data set is maintained by the National Center for Health Statistics is a U.S. government agency that provides statistical information as a collection of interactive displays and tools that present NCHS data in a visually engaging format. [3] The data set is also updated frequently, so the data used are up-to-date.

III. Related Work

There are several works related to drug poisoning mortality data mining and data visualization. One of them is a web dashboard created by 'Centers for Disease Control and Prevention'. The dashboard contains figures that present drug overdose death rates at the county level of the United States. The three visualizations depict heat maps, grids, and trend-lines of model-based county estimates of drug overdose mortality beginning in 2003. Hierarchical Bayesian models with spatial and temporal random effects were used here to generate the county-level estimates. [4]

IV. Methods

Initially understanding the data thoroughly is essential before visualizing it. Various methods are performed to pre-process the data for this purpose. Some methods include exploratory data analysis, data wrangling, and data cleaning which were performed in advance. After the data was understood adequately and that the data was obtained in the required format, data visualization was performed using the visualization tools through charts and graphs.

A. Exploratory Data Analysis

The structure of the data was traversed thoroughly. And, upon inspection, it was found that the data set contained a total of 2704 records and 18 columns.The data is then loaded into the datasets with the help of pandas which is a python data analysis library.

```
poison = pd.read_csv(
'nchs-drug-poisoning-mortality-united-states
.csv')
```

The dataset contained various available columns and after referring to the values and their meanings, we considered only a few columns to be enough for our exploration. So we got rid of the unnecessary attributes for the visualization.

```
poison.drop(["Low Confidence Limit for Crude
Rate","Upper Confidence Limit for Crude
Rate", "Age-adjusted Rate", "US Crude
Rate"], axis=1).head()
```

B. Data Wrangling

To make the complex dataset more accessible, easier to analyze, and to remove errors, data wrangling was performed. Data wrangling, sometimes referred to as data munging, is the process of transforming and mapping data from one "raw" data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics [5].

Initially, the data was extracted using ‘Age’ as a primary attribute for the visualizations. But it was found that the attribute contained the value ‘All Age’ along with the specific age groups as well which made the data redundant. So, the data values excluding it were taken.

```
poison = poison[poison['Age']!='All Ages']
```

C. Data Cleaning

Upon further inspection, many missing values were found in the dataset, which might have caused errors in processing the data for visualization. So, the missing values were replaced by zero in the original dataset.

```
poison.fillna(0, inplace=True)
```

Some data contained commas instead of points (decimal commas), so they were converted into points for ease of use.

```
poison['Crude Death Rate'] = poison['Crude
Death Rate'].str.replace(
    ",", ".").astype('float64')
```

D. Data Visualization

After conducting operations on the data set such as EDA, data cleansing, and error elimination, the data was ready to be visualized. Visualization was performed using Plotly Dash, a Python framework.

We visualized the Drug Poisoning Mortality by year of record for each year as we wanted to see how many deaths had occurred from 1999 to 2015 in the U.S. Since we also had the data for the origin(race) of the people, and the age group of the people, the data was grouped accordingly and then the number of deaths were calculated and displayed accordingly.

```
poison1 = poison.groupby(['Year', 'Race and
Hispanic Origin', 'Age'])[
        'Deaths'].sum().reset_index()
```

We also plotted the total deaths by drug poisoning in all years, along with a bar graph for deaths categorized by age groups and a pie chart for deaths categorized by sex. The dataset describes drug poisoning deaths in the U.S. by selected demographic characteristics, and includes age-adjusted death rates for drug poisoning.

V. Results

The main purpose of this study was to better understand the death caused by drugs in the U.S. using the data NCHS provides through visualization techniques. Upon exploring the data and performing all the preliminary analysis and data cleaning on it, the data was finally ready to be visualized. The number of occurrences of death for each year in the states was then displayed and also plotted.

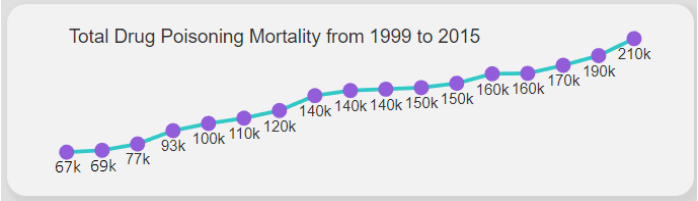

Fig. 1 Line Chart showing total mortality from 1999 to 2015

The figure shows the plot of the total number of drug poisoning deaths from 1999 to 2015 in the U.S. Each of the years are represented by plots and the line progressing in y-axis shows the increase in the number of deaths.

Upon hovering over the plots, the year and the total number of deaths in that particular year is displayed.

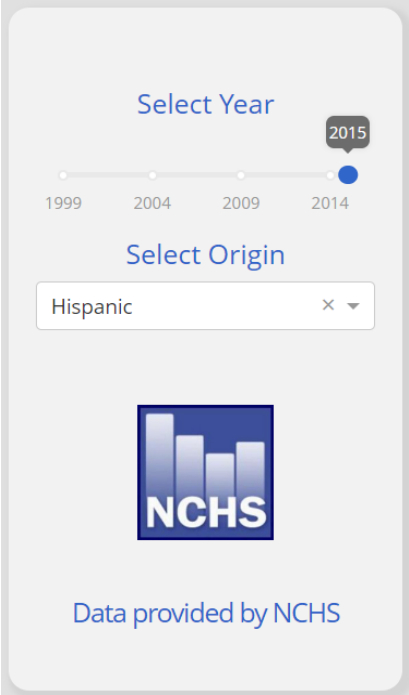

Fig. 2 Slider for year and drop-down menu for origin selection

The year can be picked using the slider provided and the origin/race can be picked using the drop-down menu. This provides easy navigation and selection of required filters to visualize the data as per required.

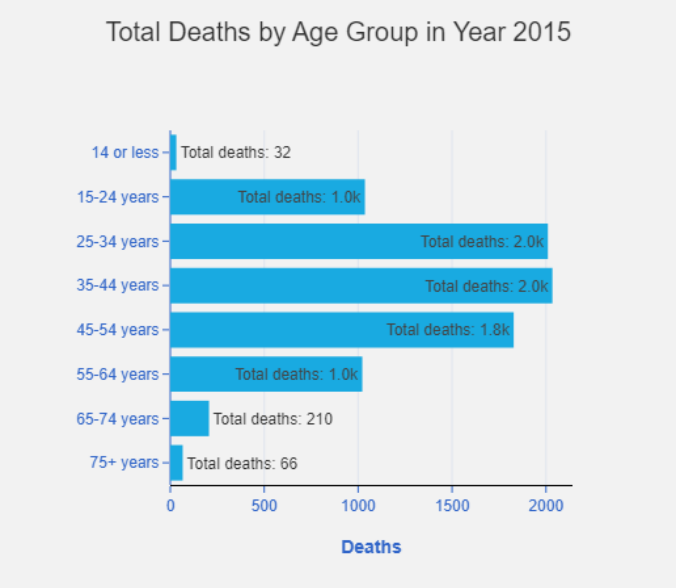

Fig. 3 Bar graph showing total deaths by age group

The bar graph is a chart that presents the provided categorical data with rectangular bars with lengths proportional to the values that it represents and the pie chart is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. [6]

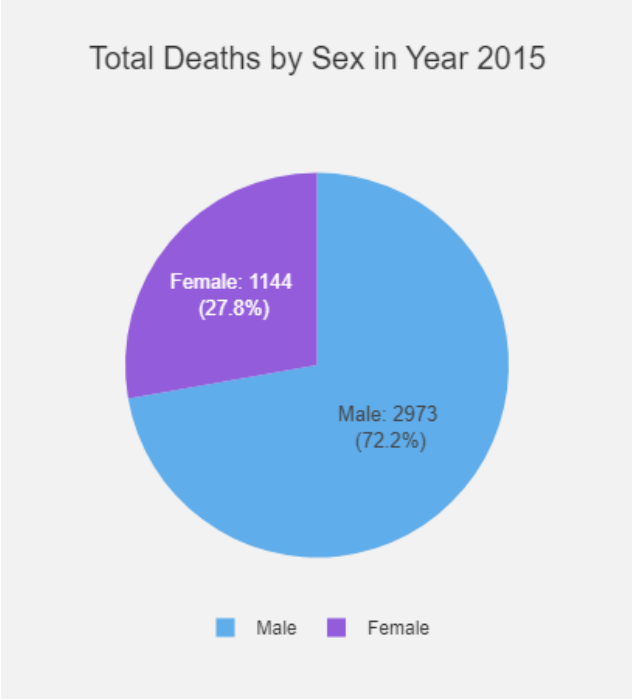

Fig. 4 Pie Chart showing total deaths by sex

The main visualization is done with the help of a bar graph and pie chart. For instance, the total deaths in a particular year are plotted against the age group sorted in ascending order in a horizontal bar graph. Also, the total deaths in a particular year based on sex is shown in a pie chart.

The users have an option to select the needed filters of data by year and by origin/race which dynamically changes the data in the bar graph and pie chart as per the requirement.

VI. Discussion

The NCHS data set had thousands of records and upon cleaning the data and converting them to meet our requirements, we were able to visually represent it. The bar diagram showed that from all the years, the deaths of people of the age group 35-44 years was the highest and rapidly increasing. Upon further research, the information turned out to be true when cross-checked with a recent article from CDC [7].

There were 8234 total drug-involved overdose deaths reported in the U.S. in 2015 (from figures 2 and 4); 72.2% of cases occurred among males. The age group of 35-44 years had the highest deaths with 2034 deaths (from figure 3). These data were accurately depicted with the help of the graphical visualization tools.

Drug related death statistics can help determine the prevalence and severity of addiction within a nation. Death and addiction statistics are important because they show that addiction isn't a personal problem but an issue impacting hundreds of thousands of people. [8]

VII. Conclusion

With the help of data mining methodologies like Exploratory Data Analysis, Data Cleaning, Data Wrangling and Data Visualization, this project has helped us to get better at understanding and working with large datasets and accurately depicting the features and the attributes of our interest as required. The drug mortality dataset initially consisted of unwanted columns and missing data which could hamper the data visualization process. But after performing the data mining works and tasks, we were able to extract the needed and important information and using visualization techniques, the data was presented visually to ensure that even non-technical persons could grasp what was going on and which data corresponded to what filters.

There are many innovative and exciting statistical methods now being developed and applied to substance abuse prevention research data. Standard statistical analyses are often inadequate in substance abuse prevention research because of the special characteristics of the data collected in these studies[9]. With better understanding and collection of data along with well defined and clean data, many people can benefit from its accurate representations.

At last, this work on visualization of drug poisoning mortality in the U.S. is not only limited to the U.S. but can be extended to visualize the mortality rates of any country in the world during any time provided there is enough data to visually represent, hence improving the scalability of the visualization..

Acknowledgements

This project was done for the partial fulfillment of the requirements of the course Data Mining (Course Code : COMP 482) under the guidance and supervision of Dr. Rajani Chulyadyo ma'am, Assistant Professor and CE Program Coordinator, Department of Computer Science and Engineering, Kathmandu University, Dhulikhel, Nepal. We are grateful to Dr. Chulyadyo for her constant assistance and continuous supervision, as well as for supplying us with project-related help and information.

References

[1] "Overdose Death Rates" National Institute on Drug Abuse, 2022. [Online]. Available: https://nida.nih.gov/research-topics/trends-statistics/.

[2] "Dash Plotly: The Premier Data App Platform for Python," Plotly, 2013. [Online]. Available: https://plotly.com/dash/.

[3] "NCHS - CDC" Centers for Disease Control and Prevention, 2022. [Online]. Available: https://www.cdc.gov/nchs/index.htm.

[4] "County-level Drug Overdose Mortality in the United States, 2003-2020", NCHS - Data Visualization Gallery. [Online]. Available: www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality

[5] "Data Wrangling," Wikipedia, 2022. [Online]. Available: https://en.wikipedia.org/wiki/Data_wrangling.

[6] "Bar Diagram and Pie Chart", Unacademy, 2022. [Online]. Available: https://unacademy.com/content/ssc/study-material/mathematics/bar-diagram-and-pie-chart/

[7] "Drug Overdose Deaths", Centers for Disease Control and Prevention, 2022. [Online]. Available: https://www.cdc.gov/nchs/hus/topics/drug-overdose-deaths.htm

[8] "Statistics on Drug Addiction" ArrowPassage, 2022. [Online]. Available: https://www.arrowpassage.com/drug-use-statistics/

[9] "Advances in Statistical Methods for Substance Abuse Prevention Research", National Library of Medicine, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2843515/