DAS732: Data Visualization Assignment-3 Report
Team name: Kurtosis

Sourav M Dileep
IMT2022033
Sourav.Dileep@iiitb.ac.in

Soham Pawar
IMT2022127
Soham.Pawar127@iiitb.ac.in

Siddharth Ayathu
IMT2022517
Siddharth.Ayathu@iiitb.ac.in

DATASET AND DATA MINING OPTIONS

We continue to use the same dataset as before i.e the Kaggle Drug Overdose Deaths dataset. The dataset originally describes 5105 cases of drug overdose deaths which took place in Connecticut between years 2012 and 2018. This includes the location, sex, race and other important factors for the cases. To get a detailed description of each column, please refer to the Appendix. The Appendix contains details about the original dataset before preprocessing. The preprocessing which was done for A1 is also mentioned in the appendix.

We also used a supplement dataset, The Drug Overdose Deaths in the year 2019. It has the same columns as the original dataset, but for the data in the year 2019. This supplement data was combined with the original dataset. Further preprocessing was done on the combined dataset, where new columns required for A3 were added. These changes are listed below:

1) **no_of_drugs:** This column represents the number of drugs which the person has taken throughout his life.
2) **num_of_deaths_in_county:** This column represents the number of deaths which have occurred in the county where the person has died.
3) **num_of_injuries_in_race:** This column represents the number of injuries which have happened from the race of the person.
4) **CityResidentCount:** This column has the total number of people who live in the same resident city as the person in study.
5) **Cluster:** This column signifies the cluster to which each person belongs to, depending on the criteria we fix.
6) **DeathsPerCity:** This column indicates the number of deaths which occurred in a specific city. If two or more rows have the same DeathCity, they will have the same value of DeathsPerCity.
7) **ResidentsPerCounty:** This column, similar to CityResidentCount has the total number of people who are residents in the same city as the person in study(the person in the row in question).

We have also removed columns such as 'DeathCityGeo', 'ResidenceCityGeo' and 'InjuryCityGeo' which were created for A1, but serve no longer use in A3.

MEMBER WISE CONTRIBUTIONS

The tasks were initially discussed over a meeting and we came up with the tasks together. For the visualizations, dashboards, and stories, the following distribution was followed:

1) Sourav did the visualisations and stories for Loop 1 and Loop 2. The treemap visualisation in Loop 4 and its corresponding inferences were also done by him.
2) Siddharth did the visualisations and stories for Loop 3 and he also did the analysis of resident count vs death count per county.
3) Soham did the preprocessing which was needed for A3 as well as the visualisations, dashboards and stories for Loop 4.

VISUAL ANALYTICS WORKFLOW

As mentioned, we will be using the Visual Analytics Workflow as proposed by Keim et al. [8]. This workflow is represented as a flow diagram shown in Fig. 1. We will be referring to this figure subsequently throughout the report.

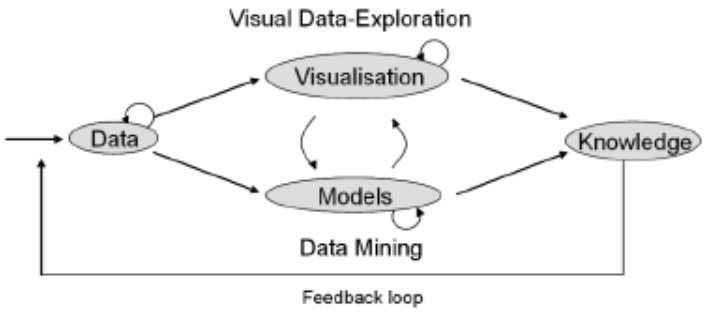

Fig. 1: Visual Analytics Workflow

The following steps are defined as part of the workflow:

- **Data:** Changes made to data, e.g., Data Transformations
- **Visualization:** Plots or figures made using the data, e.g., histograms
- **Models:** Models constructed on the available data, e.g., KMeans clustering
- **Knowledge:** Inferences or observations drawn from the visualizations and/or models

SUBMISSION STRUCTURE

We have submitted a SharePoint link, which upon opening, you will see three more directories, namely:

- **Code:** This includes all the codes and tableau books which we needed for creating the visualisations. It also consists of the models as well as an HTML file consisting of an interactive treemap visualisation.
- **Dataset:** This includes the original dataset, the supplement dataset, the combined dataset and the preprocessed dataset.
- **Images:** This folder includes all the figures which are present in the report, named according to the figure number in this report.

- **A3**$_{R}eport : This is our report for A3.$

I. ANALYSIS ON INJURIES, DEATHS AND COUNTIES

A. Loop1 - Assignment1

The analysis of injuries, deaths, and counties was started in Assignment 1. This can be seen in Task 2 (mentioned in the Appendix). The following steps were followed:

- **Data:** Additional columns, such as num_of_deaths_in_county, were created by aggregating the total number of drug overdose deaths for each county. Similarly, the num_of_injuries_in_race column was derived by summing up the injuries for each racial group.
- **Visualizations:** Multiple visualizations were created, including Pie Charts, Tree Maps, and Histograms, to understand the distribution and patterns in the data.
- **Models:** No predictive Xor statistical model was used in this assignment as the focus was on data preparation and exploration.
- **Knowledge:** Several conclusions and possible reasons for anomalies and patterns in the data were provided. Examples include identifying correlations between drug overdose deaths and injuries, and other demographic and geographic attributes.

Although the content in Assignment-1 can be unrolled into multiple loops, we conclude the work regarding the same as the first loop for the sake of brevity.

B. Loop2 - Deciding Specific Counties to Perform Study On

The study on injuries, deaths, and counties continued with a focus on selecting specific counties for detailed analysis. The following steps were followed:

- **Data:** Additional columns, such as DeathsPerCounty and InjuriesPerCounty, were created by aggregating data at the county level to calculate the total deaths and injuries in each county.
- **Visualization1:** To gain an initial overview, a Donut Chart was created using Tableau to visualize the composition of county types, as shown in Fig. 2.

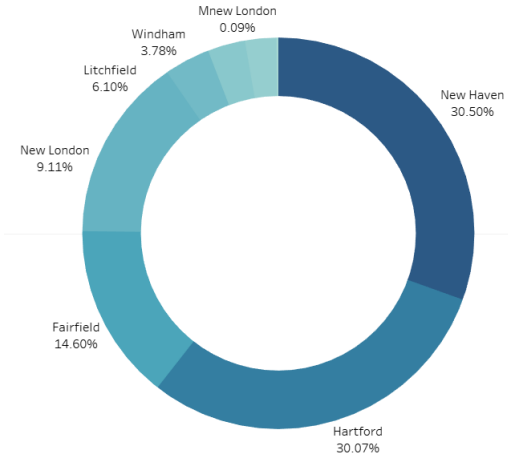

Fig. 2: Donut chart illustrating the distribution of data across counties, highlighting New Haven and Hartford as the largest contributors, followed by Fairfield , with smaller shares from other counties.

- **Knowledge1:** The donut chart illustrates the proportional distribution of data across various categories, each likely representing regions or counties (e.g., New Haven, Hartford, Fairfield, etc.). Notably, New Haven and Hartford dominate the dataset, together accounting for over 60% (30.50% and 30.07%, respectively) of the total. Fairfield follows with 14.60%, while New London and Litchfield contribute smaller shares (9.11% and 6.10%). Windham and New London represent minor contributions, at 3.78% and 0.09%, respectively. The chart emphasizes the significant disparity in distribution, with New Haven and Hartford being the most impactful, while the other regions hold less prominence.
- **Visualization2**: A dumbell plot has been created as shown in Fig. 3. to visualize the maximum and minimum values range of deaths in each county.

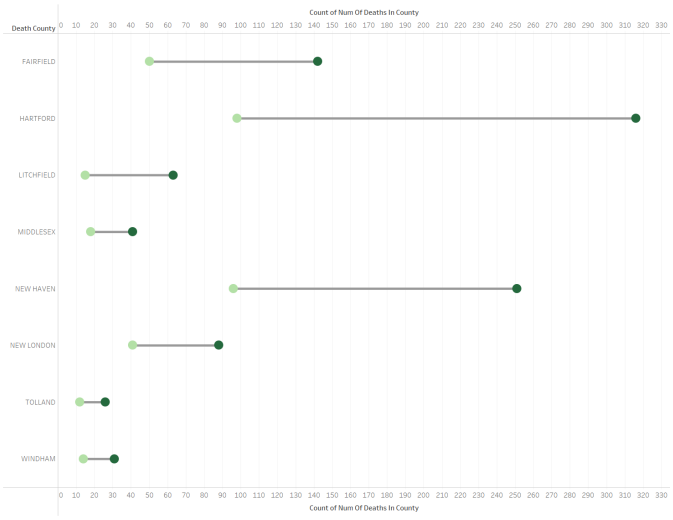

Fig. 3: Dumbbell chart showing the distribution of death counts across counties, with Hartford and New Haven leading significantly, followed by Fairfield, and smaller contributions from other counties.

- **Knowledge2**: We compare the number of deaths across

different counties. Each county is represented by a horizontal line with two markers, indicating the range or count of deaths at specific points. Hartford and New Haven counties exhibit the highest death counts, exceeding 300 and 250 respectively, signifying their prominent roles in the dataset. Fairfield follows with counts near 150, while other counties such as Litchfield, New London, Middlesex, Tolland, and Windham show significantly lower counts, generally below 100. The visualization highlights stark disparities, with Hartford and New Haven dominating while smaller counties contribute minimally.

- **Data3:** Additional columns, such as num_of_deaths_in_county, were created by aggregating the total number of drug overdose deaths for each county. The data was filtered to include only the resident county.
- **Visualization3:** A hierarchical tree map was created with the layers County, City and Race in order. Each layer is visualised based on the number of deaths which happened in that element of that layer. Here, in the above figure, you can see the treemap distributed by counties according to the number of deaths which occurred in that region. This is an interactive treemap, where on clicking, it will take you into the City vs Number of Deaths treemap. This will be discussed in the next visualisation.

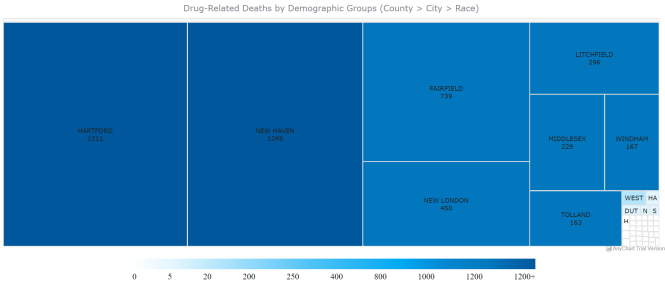

Fig. 4: A nested squared layout of a treemap denoting the drug-related deaths with the layers of County ¿ City ¿ Race

- **Knowledge3:** We had already discussed how the economically developed regions tend to have more drug usage and have more death cases than underdeveloped regions in A1. This inference is further confirmed by this treemap visualilsation. Here you can see Hartford, New Haven and Fairfield take about ¾ th of the space in this visualisation, which denotes that three-fourths of the deaths which occurred happened in these regions. Moreover we had already proved in A1 that these are more economically stable than the other regions which are present towards the bottom right corner of the visualisation. The common knowledge that drug usage is more common in cities than in rural areas can be seen in this knowledge that we inferred. Hartford, New Haven and Fairfield are the three main urban centres in our dataset. We can also see the wide disparity between urban and rural areas. The number of deaths are at both extremes for urban and for rural. While the number of deaths are at a really high level in urban areas, the number of deaths in economically weaker regions are at a really low level.
- **Data4:** Additional columns, such as num_of_deaths_in_county, were created by aggregating the total number of drug overdose deaths for each county. The data was filtered to include only the resident county.
- **Visualization4:** This is the next layer of the hierarchical treemap which was mentioned above. This shows the treemap visualisation distributed by the cities of Hartford county according the number of deaths which occurred there. The darker shade and size is directly correlated to the number of deaths of each city.

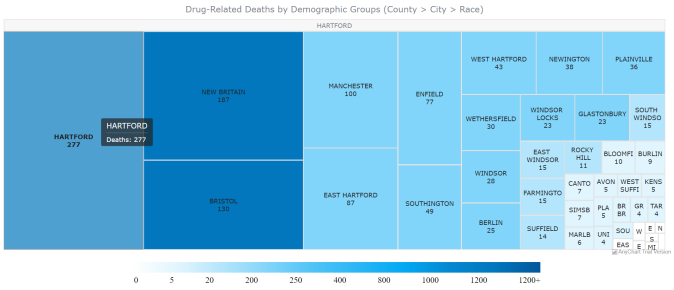

Fig. 5: The inner layer of the previous treemap you get once you click Hartford.

- **Knowledge4:** Once again, the knowledge which we inferred above can be seen in this visualisation as well. Hartford, New Britain, Bristol, Manchester and East Hartford constitute the well-known, populous and economically stronger cities of Hartford county. By looking into this visualisation, you can once again see that these are the cities constituting half the area of this treemap visualisation. While all the rest of the cities constitute the other half. This again proves the disparity of the regions in economic criteria as well as drug usage.

C. Loop3 - A Look at Most Widely Used Drugs and Counties with Highest Death Count

Building on the analytics from Loop 2, the analysis became more focused by studying specific categories. The counties of **Hartford** and **New Haven** were selected for detailed examination due to their highest death counts, making them of particular interest for further study.

- **Data1:** The data was filtered to include only the age and num_of_deaths_in_county columns for the counties of Hartford and New Haven.
- **Visualization1:** A juxtaposed visualization was created consisting of two violin plots: one for Hartford and the other for New Haven. Both plots share the same y-axis for easy comparison. Hartford is represented in orange, and New Haven in blue. Corresponding box plots were also overlaid on each violin plot for additional statistical context. This visualization can be seen in Fig. 6.

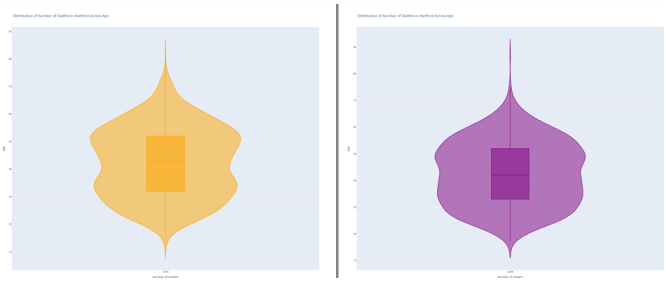

Fig. 6: Violin plots depicting the distribution of death counts across age groups in Hartford, showing density concentrations and variability in the number of deaths for each age range.

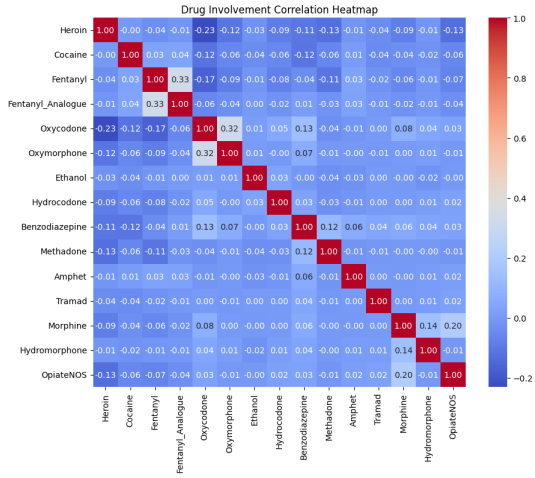

Fig. 7: Correlation heatmap regarding the co-existance of drugs in each of the drug cases.

- **Knowledge1:** We see two violin plots comparing the distribution of the number of deaths across different age groups in Hartford. The violin plot on the left (orange) and the one on the right (purple) highlight the density and spread of deaths across ages. Both plots showcase similar symmetric distributions, with the central box representing the interquartile range (IQR), indicating where the majority of data lies. The peaks and tails of the violins provide insights into higher or lower densities of death counts at specific age ranges. These visualizations help identify which age groups experience more concentrated death counts and the overall variability in the data.

To further our understanding, we analyzed the relationship between the involvement of different drugs and their role in causing deaths.

- **Data2:** The dataset was filtered to include columns representing specific drugs involved in overdoses and their corresponding death counts.
- **Visualization2:** A heatmap was created (Fig. **??**) to illustrate the co-occurrence of different drugs in overdose cases and their association with death counts. This visualization highlights the combinations of drugs that are most frequently involved in fatal cases.
- **Knowledge2:** The correlation heatmap follows the **cool-warm** coloring scheme. Each cell here represents the possibility of the involvement of the combination of drugs in each of the drug cases, which have happened. As an example, across all the drug cases, if oxycodone was involved, then there was a relatively high chance that oxymorphone was also involved.
- **Data3:** The num_of_deaths_in_county column was used to filter the number of deaths caused by heroin across all counties.
- **Visualization3:** Another juxtaposed visualization was created, featuring two violin plots: one for deaths caused by heroin and the other for deaths caused by opioids. These plots also share the same y-axis. Heroin is represented in orange, and opioids in blue. Corresponding box plots were overlaid for added detail. This visualization can be seen in Fig. 8.

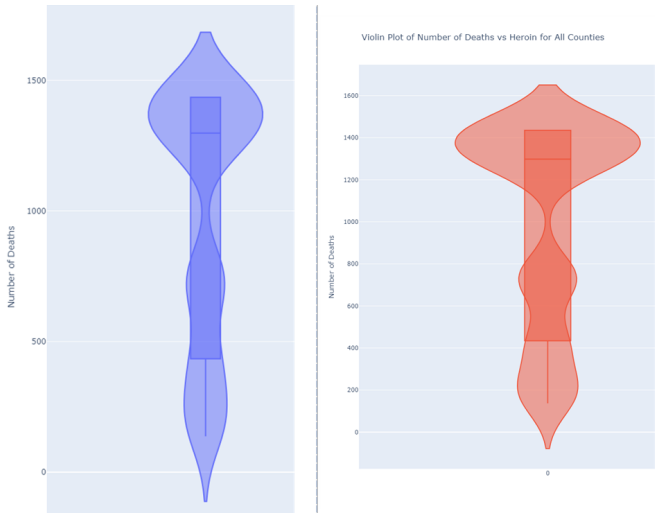

Fig. 8: Violin plots comparing the distribution of heroin-related deaths (blue) to overall deaths (red) across all counties, highlighting distinct density patterns and concentration ranges.

- **Knowledge3:**The two violin plots, with the left plot (blue) represents the distribution of deaths related to heroin across all counties, and the right plot (red) showing the overall distribution of deaths for all causes. The blue violin plot indicates a narrower distribution, with fewer high-density peaks, suggesting that heroin-related deaths are concentrated within a limited range. In contrast, the red violin plot has a broader distribution with a higher density in certain ranges, reflecting greater variability in overall deaths. Both plots emphasize differences in patterns between heroin-related deaths and general mortality.

D. Loop4 - A Closer Look at Number of Drugs Overused

This analysis aimed to investigate whether an overdose involving a higher number of drugs leads to a greater number of deaths or if it is equally likely to result in the same death count when fewer than three drugs are involved.

- **Data1:** The column number_of_drugs was used, which represents the number of drugs overdosed per person.

- **Visualization1:** To visualize the difference in the number of deaths across all counties, a Parallel Coordinate Plot was created with limited attributes. This visualization can be seen in Fig. 9

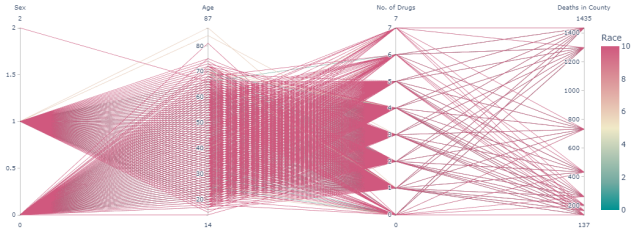

Fig. 9: Parallel coordinates plot showing relationships between sex, age, number of drugs, county deaths, and race, highlighting trends and correlations in fatalities across multiple dimensions.

- **Knowledge1:** Each line represents a data point, connecting values across these dimensions. The density and clustering of lines provide insights into trends and patterns. For example, there appears to be a high concentration of deaths within certain age ranges (40–70) and with specific numbers of drugs (2–5), potentially signifying correlations between age, drug use, and fatalities. The color gradient reflects race, with darker shades representing higher race indices, allowing for the observation of any racial disparities or patterns. This visualization aids in understanding the complex interplay of factors influencing deaths across counties.
  Using interactions such as brushing, cases involving 4 to 7 drugs were selected. The analysis revealed that the number of deaths due to the overdose of 4–7 drugs was evenly spread compared to overdoses involving 1–3 drugs. This highlights the fact that even the overdose of a single drug is extremely harmful.

We thought it would be useful to explore the involvement of different drugs in causing deaths across various races.

- **Data2:** The dataset was filtered to include columns representing specific drugs involved in overdoses, the corresponding death counts, and the race of individuals.
- **Visualization2:** A correlation plot was created to represent drug involvement across different races in Fig. 10
- **Knowledge2:** Notable correlations were seen such as Benzodiazepine usage among Black individuals and Oxycodone usage among White individuals, while most correlations remain low or negligible.

Since whites have the majority no.of deaths , we wanted to see the distribution of their drug overdose across the counties in Fig. 11

The cartograph the distribution of death cases for the subjects belonging to the white race over the region of USA. We choose a white dress because it was seen that the most number of death cases happened for this group. This dot distribution has been clustered using Gaussian Mixer Models or GMM into 46 clusters, which is the total number of counties

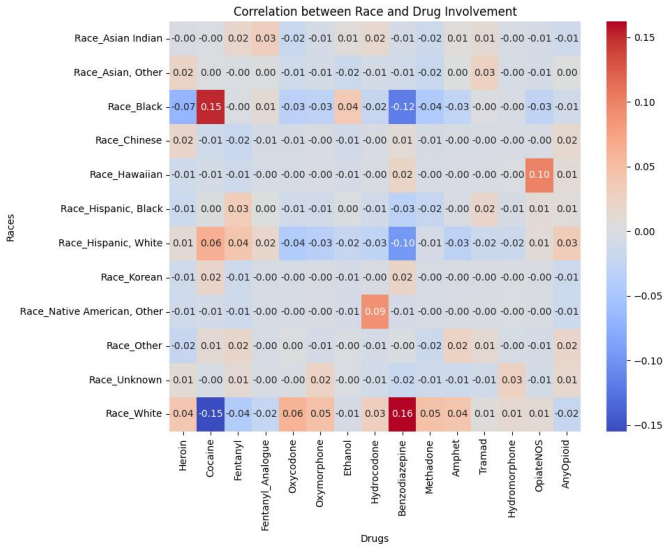

Fig. 10: Illustrates the associations between different races and their involvement with specific substances

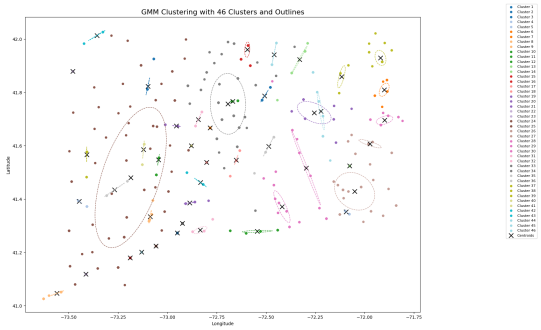

Fig. 11: GMM clustering based grouping of death cases for white race across counties.

in USA. The plot itself is flipped and can be aligned with the map of USA after simple rotations. From these clusters we can see that the county Connecticut falls in the largest cluster, which is formed.

- **Data3:** Already existing columns were used. The dataset was filtered to the value of deaths by race and by location
- **Visualization3:** This treemap is a nested treemap with the layers Race, Location in order. Race determines the race of the person and Location shows the place of death. The darker shade and size is directly correlated to the number of deaths of each city.
- **Knowledge3:**The data indicates that white individuals are more prominent in drug usage compared to black individuals. Contrary to the common misconception that black people are more associated with drug use, our dataset—and specifically Fig. 3—shows that white and Hispanic white individuals constitute approximately three-quarters of the cases. As previously established in the Assignment-1 (A1) report, the stereotype often depicted in media about drug association with black individuals is inaccurate. Additionally, the data shows

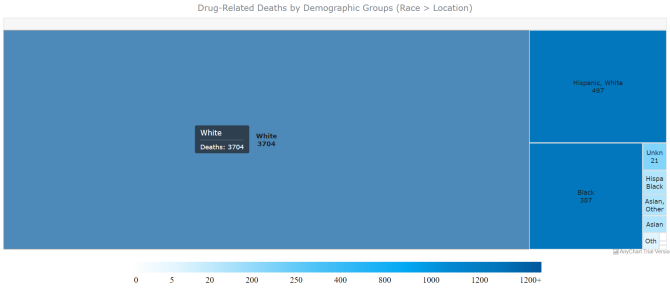

Fig. 12: A nested treemap denoting the distributions of cases in Race and Location nested. .

that other racial groups, such as Chinese, Indian, and Hawaiian, together make up only about 1% of the cases, further highlighting the racial distribution of drug use in this dataset.

- **Data4:** Already existing columns were used. The dataset was filtered to the value of deaths by race and by location
- **Visualization4:** This treemap is a nested treemap with the layers Race, Location in order. Race determines the race of the person and Location shows the place of death. The darker shade and size is directly correlated to the number of deaths of each city. In the first figure, you can see the treemap visualisation distributed based on location of death of the white people. In the second figure you can see the same for the location of death of black coloured people.

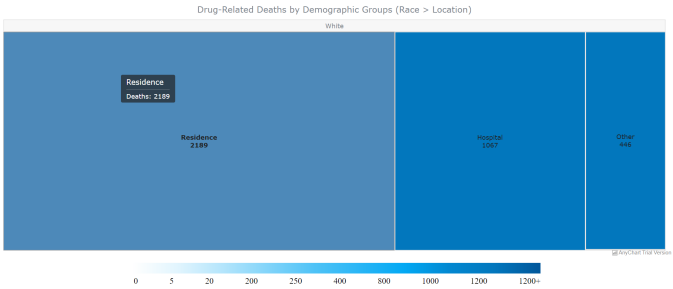

Fig. 13: The inner-most treemap of Fig.6 which displays the distribution of Death location of the white people.

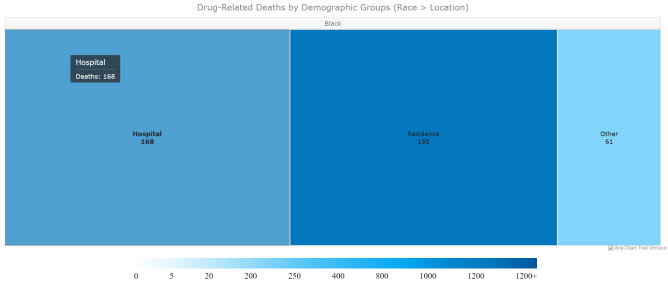

Fig. 14: The inner-most treemap of Fig.6 which displays the distribution of Death location of the white people.

- **Knowledge4:** Racial Disparities in Police Encounters and Hospital Admissions: Data suggests that black individuals are more frequently caught by the police and admitted to hospitals for drug-related incidents than white individuals. This inference was previously established in the Assignment-1 (A1) report, and here we find additional evidence to support it. As discussed above, it is a misconception that black individuals use drugs more frequently than other races. This disparity may stem from the fact that most drug usage cases are reported when individuals are apprehended by the police, directly linking it to the location of death data. Fig. 13 illustrates the distribution of death locations among white individuals. In this figure, nearly half of the white population's deaths occurred in their residences, while only about one-third were reported in hospitals. This aligns with the observation that white individuals are less frequently apprehended for drug-related incidents. Fig. 14 displays the distribution of death locations among black individuals. Here, the numbers of individuals who died in their residences and those who died in hospitals are nearly equal. This further supports our inference that black individuals are more likely to be caught by police, reinforcing the misconception about drug usage prevalence in this group.
  - **Data5:** The columns Injuries_per_race and number_of_drugs (representing the number of drugs overdosed per person) were used for this analysis.
  - **Visualization5:** To visualize the differences in the number of injuries across races, a Parallel Coordinate Plot was created with limited attributes. This visualization can be seen in Fig. 15.

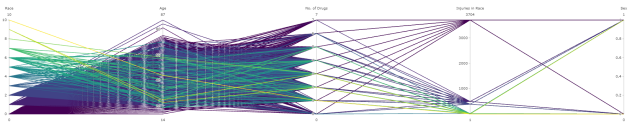

Fig. 15: Parallel coordinates plot displaying relationships between race, age, drug usage, injuries, and sex, highlighting patterns and disparities in the dataset.

  - **Knowledge5:** Each line represents a unique data point, connecting values across these dimensions. Key insights include clustering of age values between 20 and 60, suggesting a significant focus of injuries within this range. The number of drugs involved varies between 0 and 7, with lines showing interactions between higher drug usage and the extent of injuries. Additionally, the plot highlights variations across races and their relationship to injuries, suggesting disparities or trends. The sex variable indicates a balance or disparity in injury occurrences among genders. This multidimensional visualization helps in identifying correlations and patterns across diverse factors influencing injuries.
    The analysis revealed the following key insights:
    * White and Hispanic Whites of all ages have experienced injuries from drug overdoses.

* In other racial groups, injuries due to drug overdose are predominantly observed in younger populations (less than 45 years old).
* Whites are the only race to experience deaths involving overdoses of more than five different drugs.

II. Analysis of Resident Count vs Death Count per County.

The final addition to the dataset involved creating columns for `CityResidentCount` and `Residents per County`.

This analysis aimed to explore the relationship between the number of residents in a county and the number of deaths due to drug overdose.

Initially, we hypothesized that the number of deaths due to drug overdose in a county is directly proportional to the number of residents in the county. This hypothesis is visualized and evaluated using Fig. 16.

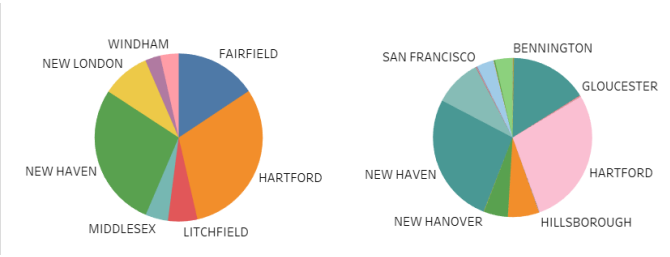

Fig. 16: Comparison of deaths per county (left) and resident population per county (right), highlighting disparities and notable contributions from counties such as Hartford and New Haven in both metrics.

On the left, counties such as Fairfield, Hartford, and New Haven dominate the distribution, reflecting higher instances of fatalities in these regions. In contrast, the right chart illustrates the proportional resident population across counties, with New Haven and Hartford again displaying significant proportions alongside counties like Bennington and Gloucester. This visual comparison highlights potential disparities between fatality counts and population size, suggesting factors such as population density, healthcare access, or varying risk factors may influence the observed patterns in fatalities.

The plot in Fig. 17 represents the number of drug death cases where the counties on the y-axis represent the residency county of the drug death case. Similarly, the counties on the x-axis represent the counties where the death happened for each case. Thus, the intersection of each row and column represents the number of death cases in the county represented by the corresponding column, where the subject resided in the county represented by the corresponding row. From the plot, we can see that the most number of death cases have happened in Hartford where the subjects also reside in Hartford.

The scatter plot illustrates the relationship between the number of deaths per city and the median resident count

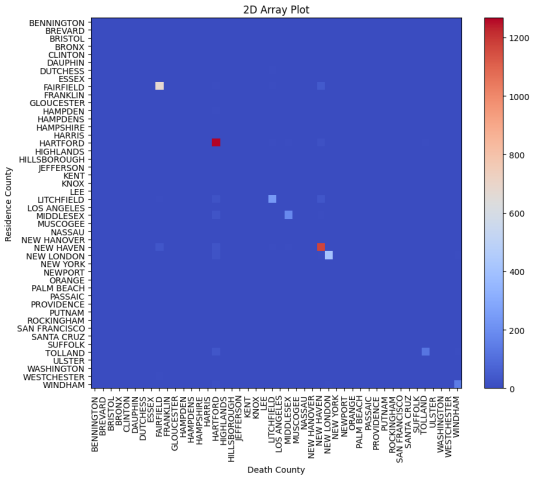

Fig. 17: A 2D matrix visualization rerpesenting the residence city and death city for each case

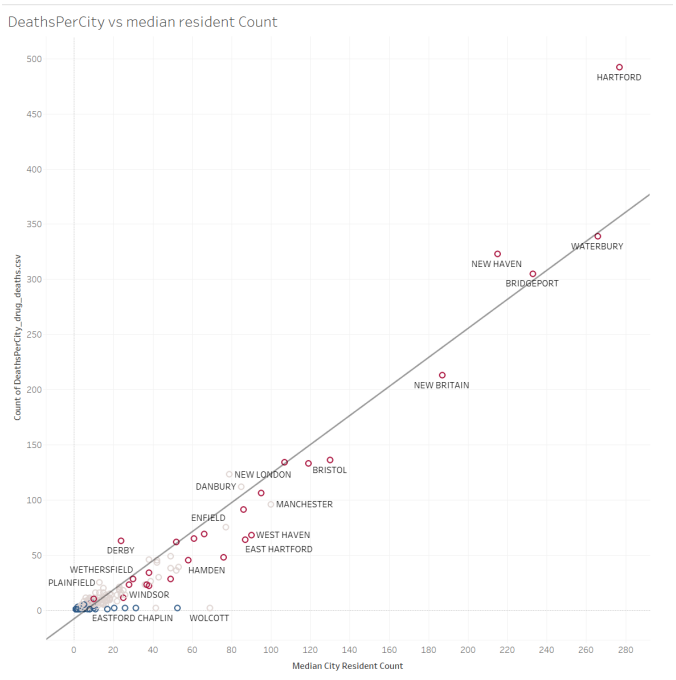

Fig. 18: Scatter Plot of Deaths per City vs. Median Resident Count Highlighting the Positive Correlation and Outliers.

across various cities. A strong positive correlation is evident, as cities with higher median resident counts generally report a greater number of deaths. Prominent cities like Hartford, Waterbury, and New Haven are notable outliers, positioned significantly above the regression line, indicating a disproportionately higher count of deaths compared to their median resident numbers. This trend suggests that population density or size might be a significant factor influencing the count of deaths, though other socio-economic or health-related variables could also play a role. Cities with lower median resident counts cluster near the origin, reflecting fewer deaths.

based on the relationship between drug overdose deaths per city and the median resident count. These clusters can be associated with varying economic categories of the cities, shedding light on the interplay between economic disparity and drug-related fatalities.

Blue Cluster (Low Population, Low Deaths): This cluster comprises small cities or rural towns with lower median resident counts and fewer drug overdose deaths. These areas likely represent economically disadvantaged regions with limited access to healthcare services or drug rehabilitation facilities. The relatively low number of reported deaths may stem from smaller populations or underreporting due to limited infrastructure for accurate record-keeping.

Gray Cluster (Moderate Population, Moderate Deaths): Cities in this cluster display moderate levels of both population and drug-related deaths. These are likely mid-sized cities or suburban areas with middle-income residents. The deaths in these regions could reflect a combination of socio-economic pressures, such as limited employment opportunities or moderate healthcare access, which may contribute to drug misuse and overdose incidents.

Red Cluster (High Population, High Deaths): The red cluster includes cities like Hartford, New Haven, and Waterbury, characterized by high resident counts and significantly elevated drug overdose deaths. These cities often represent urban centers with diverse economic conditions. While urban areas typically offer better healthcare access, they also face heightened challenges, including economic inequality, unemployment, and the prevalence of drug trafficking. These factors can exacerbate drug misuse among economically vulnerable populations, leading to higher overdose fatalities.

Economic Disparity and Drug Overdose Deaths: The plot underscores how economic disparity influences drug overdose deaths. Urban centers (red cluster) often exhibit stark economic divides, with low-income communities disproportionately affected by substance abuse due to stressors like unemployment, housing instability, and lack of affordable treatment options. Conversely, rural areas (blue cluster) face challenges such as inadequate healthcare infrastructure and social isolation, which can also contribute to substance misuse but result in lower absolute numbers due to smaller populations. Mid-sized cities (gray cluster) fall between these extremes, reflecting a mix of urban and rural dynamics.

APPENDIX

A. Dataset

In this assignment, we study the Kaggle Drug Overdose Deaths dataset. The dataset originally describes 5105 cases of drug overdose deaths which took place in Connecticut between years 2012 and 2018. This includes the location, sex, race and other important factors for the cases. We try to find the trends in the data, among the different data fields using various visualizations. The fields present in the dataset are:

1) **ID**: This column's values are used to uniquely identify the cases (represented by rows).
2) **Date**: Date when the incident took place. The type of incident being described by column 'DateType'.
3) **DateType**: Type of incident; 1 for Date Reported and 0 for Date of Death.
4) **Age**: A float value representing the age of the subject.
5) **Sex**: String value representing the sex of the subject.
6) **Race**: String value representing the race of the subject.
7) **ResidenceCity**: String value representing the residence city of the subject.
8) **ResidenceCounty**: String value representing the residence county of the subject.
9) **ResidenceState**: String value representing the residence state of the subject.
10) **DeathCity**: String value representing the death city of the subject.
11) **DeathCounty**: String value representing the death county of the subject.
12) **Location**: The location of the reported incident. These are general locations .eg. Hospital, Residence, etc.
13) **LocationifOther**: A more detailed location value for the items with value 'Other' in column 'Location'.
14) **DescriptionofInjury**: Brief description of the cause of injury .eg. inhalation, substance abuse, etc.
15) **InjuryPlace**: The type of location where the subject was injured .eg. residence, hotel, etc.
16) **InjuryCity**: String value representing the city where the subject was injured.
17) **InjuryCounty**: String value representing the county where the subject was injured.
18) **InjuryState**: String value representing the state where the subject was injured.
19) **COD (Cause of Death)**: The values specify a detailed description of the cause of death of the subject.
20) **OtherSignificantFactors**: This field mostly has null values. The non-null values add to the description of Cause of Death values for the subject(s).
21) **Heroin**: Binary value representing if Heroin was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
22) **Cocaine**: Binary value representing if Cocaine was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
23) **Fentanyl**: Value representing if Fentanyl was involved in the incident; 1 representing if the drug was present, 0 if it wasn't. Some values are about the form in which the drug was consumed.

24) **Fentanyl_Analogue**: Binary value representing if any Fentanyl analogue was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
25) **Oxycodone**: Binary value representing if Oxycodone was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
26) **Oxymorphone**: Binary value representing if Oxymorphone was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
27) **Ethanol**: Binary value representing if Ethanol was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
28) **Hydrocodone**: Binary value representing if Hydrocodone was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
29) **Benzodiazepine**: Binary value representing if Benzodiazepine was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
30) **Methadone**: Binary value representing if Methadone was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
31) **Amphet**: Binary value representing if Amphet was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
32) **Tramad**: Binary value representing if Tramad was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
33) **Morhpine_NotHeroin**: Binary value representing if Morphine was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
34) **Hydromorphone**: Binary value representing if Hydromorphone was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
35) **Other**: This column has text values which are names of drugs used other than the ones which have separate columns for them.
36) **OpiateNOS**: Binary value representing if an Opiod, not otherwise specified, was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
37) **AnyOpioid**: Binary value representing if any Opiod was involved in the incident; 1 representing if the drug was present, 0 if it wasn't.
38) **MannerofDeath**: This column's value represents the way in which the death happened .eg. if the death was accidental or natural, etc.
39) **DeathCityGeo**: The values of this column contain the Latitude and Longitude values of the location of death of the subject.
40) **ResidenceCityGeo**: The values of this column contain the Latitude and Longitude values of the location of residence of the subject.
41) **InjuryCityGeo**: The values of this column contain the Latitude and Longitude values of the location of injury of the subject.

B. Data Pre-Processing

1) The columns 'DeathCityGeo', 'ResidenceCityGeo' and 'InjuryCityGeo' contain the coordinates of the incident. Every cell contains the coordinates along with the location names or codes. This is present in string format and is formatted into two columns; for latitude and longitude, thus six new columns in total.
2) The COD (Cause of Death) column contains string-based description of the cause of death. This column required some language-based processing for extracting important information for the drugs involved. This was done by first removing 'stop' words from the sentences and then performing `NER (Named Entity Recognition)` on the sentences. The words of interest had the tags 'NN', 'JJ' and 'NNS'. The 'NN' and 'NNS' tags are used to identify Noun words and 'JJ' for Adjective words. This was done using the `spacy` library.
   It was expected that the words with 'NN' and 'NNS' tags would be the drug names. However, these still contained some un-related words concatenated with the drug names. For filtering the data further, the `scispacy` library was used for identifying drug names as this library is used for bio-medical words processing. After this, a list of words was formed with only drug names, across all the rows in the original dataset.
   After this, these drugs were categorized into six categories: Amphetamines, Benzodiazepines, Opiates, Barbiturates, Antidepressants and Antihistamines. These are used in different visualizations later.
3) Data pre-processing involved handling missing values based on specific thresholds. Columns with 40% or more null values were removed entirely to eliminate features with excessive missing data, ensuring cleaner and more reliable insights. For columns with 1% or fewer null values, we opted to remove the corresponding rows to retain the feature while minimizing the loss of data. This approach balanced the need to preserve important features while maintaining data quality for analysis.
4) Removed columns 'DateType' and 'ID' which were not related to the context of the data. The column 'DateType' represents the type of incident when it was reported on that date; if it is the date it was reported or if it is the date of death of the subject. The column 'ID' contains values for unique representation of the cases in the dataset.
5) For the columns 'DeathCounty', 'DeathCity', 'ResidenceCounty' and 'ResidenceCity', data imputation was done in case of missing values. This was done based on the frequency of values across other cells for that column. For instance, if for the value 'STRATFORD' in 'DeathCity', the value 'FAIR-

FIELD' has occurred in 'DeathCounty' for seven times for other cells and a new cell is now encountered in this column, then the value 'FAIRFIELD'is used.

6) To create the area plot of drug overdose cases by year for each drug, we started with dates formatted as "dd/mm/yyyy." Since we only needed the year, we extracted it by taking the year part of the date string. Then, for each record in the dataset, we checked if the year fell within the range of 2012 to 2018. If it did, we incremented the count of overdose cases for that particular drug in that specific year by 1. This way, we accumulated the total number of overdose cases for each drug over the years. Finally, the organized data allowed us to generate an area plot to visually represent how the overdose cases for different drugs evolved year by year.

C. Objectives

The objective of this analysis is to gain deeper insights into the deaths and injuries of the people because of drug overdose and how it is related to other factors. Specifically, we aim to understand the correlation with the following factors:

1) Deaths, Year and Age based analysis
2) Deaths, Location and types of Drugs based analysis
3) Deaths, Race and Sex based analysis

D. Data Stories

1) Deaths, Year and Age based analysis: Hypothesis 1: Younger people are more correlated to drug overdose deaths as compared to older people. The idea behind this hypothesis is that younger individuals tend to consume more drugs, leading to higher overdose death counts. Since younger people, defined as those under 30 years of age, are often more involved in risky behaviors, we expect them to have a greater share of drug overdose deaths.

Research also suggests that while younger people may use substances like cocaine or ethanol more often, these drugs are often less lethal compared to harder drugs like opioids. This could explain why overdose death rates for younger individuals may sometimes be lower, despite high rates of consumption [1].

This hypothesis is verified visually, as scatter plots indicate that younger people have fewer deaths compared to older age groups, and drugs like cocaine and ethanol have led to fewer deaths overall.

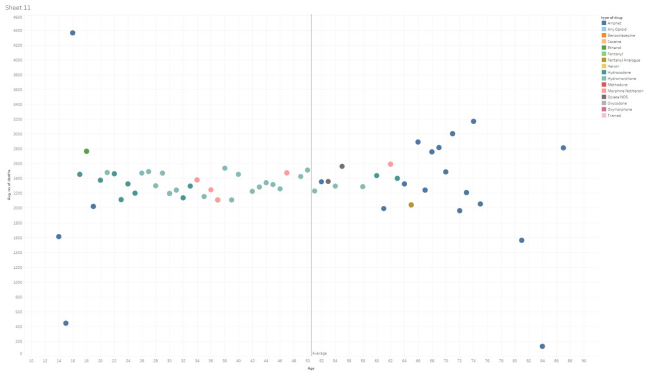

Fig. 19: Scatter plot of average number of deaths by age, with drug categories color-coded and an age average line

This scatter plot displays the average number of deaths versus age, with drug categories represented by color-coded dots. An average age line is included for reference. The plot highlights that Amphetamines (Amphet) were the most lethal, contributing to the highest number of deaths compared to other drugs. This visualization allows for the assessment of drug-related mortality across different ages, with Amphetamines showing a particularly significant impact.

We can verify that younger people tend to take less lethal substances like alcohol, cigarettes, and marijuana, according to sources such as Addiction Center (https://www.addictioncenter.com/addiction/young-adults/). This aligns with the dataset, where the plot demonstrates that individuals below the age of 25 have died from the intake of fewer drugs, supporting the conclusion that younger populations tend to engage with less harmful substances.

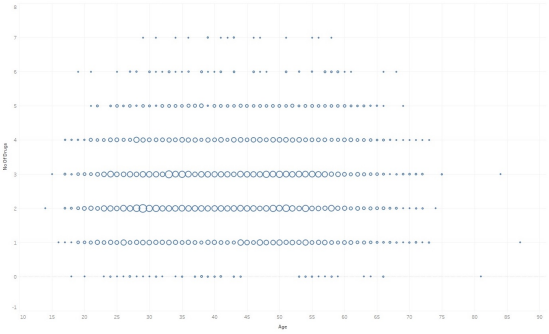

Fig. 20: Circle plot showing the age of drug users and the number of drugs used

This circle plot visualizes the relationship between the age of drug users and the number of drugs they use. Each circle's position reflects a specific age and the number of drugs consumed, while the radius of each circle represents the number of users for that age and drug combination. Larger circles indicate higher numbers of users, helping to identify the most common age groups and their corresponding drug usage patterns.

Another reason supporting this hypothesis is that younger people generally have healthier bodies compared to older individuals, which may contribute to fewer deaths despite drug

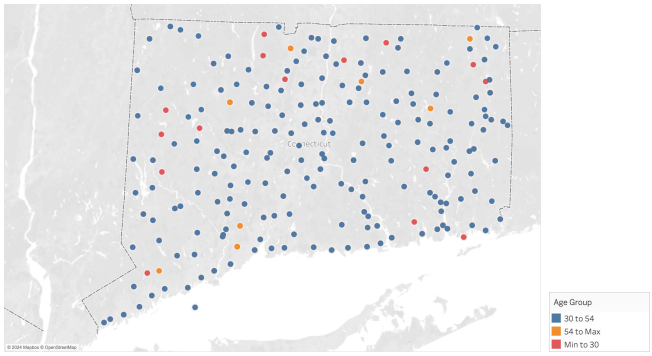

Fig. 21: Geographical Distribution of Drug Overdose Deaths by Age Group in Connecticut

use. Younger bodies may have a greater ability to metabolize and recover from the substances consumed, leading to a lower fatality rate from drug misuse.

This cartograph represents the geographic distribution of drug overdose deaths across Connecticut, segmented by age groups. Three age categories are used: individuals aged under 30 (in red), aged 30 to 54 (in blue), and aged 54 or above (in orange). Each point marks the location of a recorded death, with color coding indicating the corresponding age group. The map provides insight into how drug-related deaths are distributed across the state, with the majority of cases falling in the 30 to 54 age group.

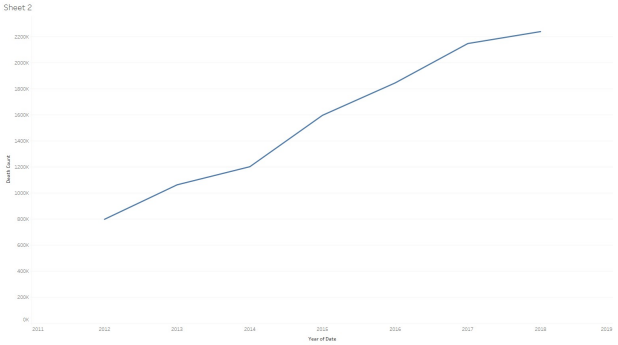

Fig. 22: Line plot showing the irregular increase in deaths from 2012 to 2018.

The data reveals that the number of deaths has increased irregularly from 2012 to 2018. This trend is consistent with the dataset, which shows fluctuations in death rates over this period. The irregular pattern indicates variability in the number of deaths year by year, aligning with the observed data trends. Now, let's explore how the number of drug overdose deaths has changed for each specific drug over the years, building on the overall trend we've just seen. This will help us better understand how each drug contributes to the total number of deaths.

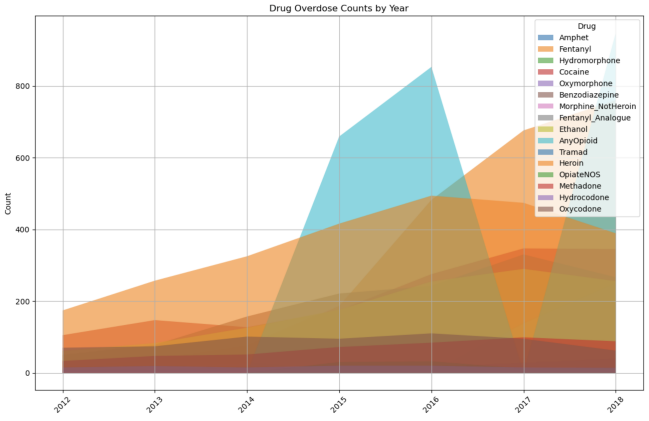

Fig. 23: Drug overdose counts by Year for various substances (2012–2018)

This area plot visualizes the increasing number of drug overdose deaths over the years 2012 to 2018, categorized by the type of drug involved. The x-axis represents the years, while the y-axis shows the total number of overdose counts for each drug, represented by different colors. Notable trends include the sharp rise in fentanyl-related overdoses (shown in blue) around 2015, which continues to increase through 2017. Other opioids like heroin, oxycodone, and methadone are also represented, although fentanyl clearly dominates the latter part of the timeline. The chart suggests a significant shift towards more potent opioids driving overdose death counts in recent years.

2) Deaths, Location and types of Drugs based analysis: There are 6 columns in the dataset that provide insights into residence, death location, and injury place. These columns are ResidenceCity, ResidenceCounty, DeathCity, DeathCounty, Location (of death), and InjuryPlace. All the locations in the filtered dataset fall within 8 counties, all of which are located in the state of Connecticut (CT), USA. from figure 15, it is clear that most of the injuries have taken place in residences, but low average amount of drug taken per person who died.

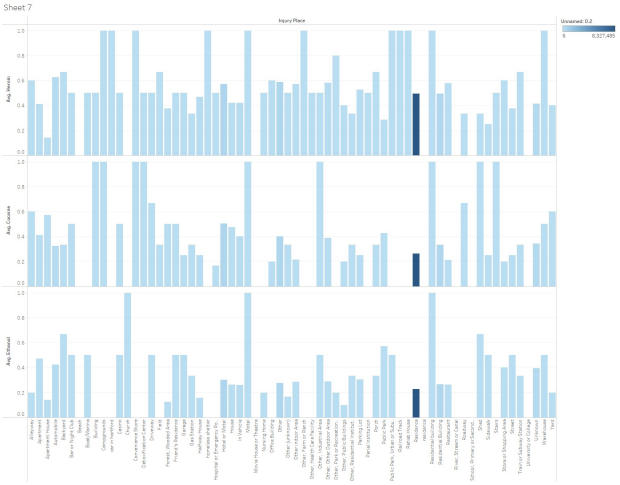

Fig. 24: Bar graph of drug-related deaths by location and drug type

This bar graph compares the number of deaths from the three most widely used drugs across various locations. The depth of the blue color indicates the death count, with a deeper

blue representing higher counts. Residences show the highest number of deaths, significantly outnumbering other locations, though the average level of drug abuse per person is lower in these areas compared to others. This visualization highlights the correlation between drug usage and death locations, emphasizing the prominence of residential deaths. From Figure 15, we observe that the majority of injuries due to drug abuse have occurred in residences, despite the fact that the average quantity of drugs consumed per injured person in this setting is lower than anticipated and also lower compared to other venues. The following plot highlights that a greater variety of drugs has been consumed in residences, which reduces the average quantity of any single drug taken at home. This suggests that the presence of multiple drug types might contribute to the pattern of injury, even though individual drug consumption levels remain modest.

Fig. 25: Circle plot of injury locations with circle size representing different drug types used.

The circle plot displays the locations where injuries have occurred. In this plot, the size of each circle corresponds to the number of different types of drugs used at each location, rather than the total quantity of drugs used. This visualization helps to understand the diversity of drug types involved in injuries across various locations.

Hypothesis 2: It is expected that individuals from wealthier areas are more prone to engaging in substance abuse, and consuming more drugs results in a higher likelihood of death due to overdose.

Research [5] has indicated that affluent communities tend to experience elevated rates of substance abuse, particularly among adolescents and young adults. This is largely attributed to factors such as heightened academic pressure in elite schools, where students are often held to high expectations of achievement. Stress and competition in these environments can lead to the misuse of substances like ethanol, marijuana, and stimulants (e.g., Adderall and cocaine), which are more accessible to wealthy teens due to their greater disposable income.Additionally, easier access to drugs and alcohol, along with social influences where substance use may be normalized within certain peer groups, contributes to the trend. Wealthier teens may engage in drug use as part of a cultural or social dynamic within their communities, which can lead to higher rates of drug abuse and overdose fatalities.

Hartford and New Haven contain areas of both wealth and poverty, reflecting the broader economic disparities seen in Connecticut. Hartford is often associated with affluence due to high-earning neighborhoods and the presence of wealth. Greater Hartford ranks high in terms of the percentage of top-earning households, with 26% of these households living in predominantly affluent. Similarly, in New Haven [6], wealth is concentrated in certain neighborhoods. The top earners in the New Haven-Milford area have an average income of $282,000, placing it among the top 20 regions in the country for high-income households.

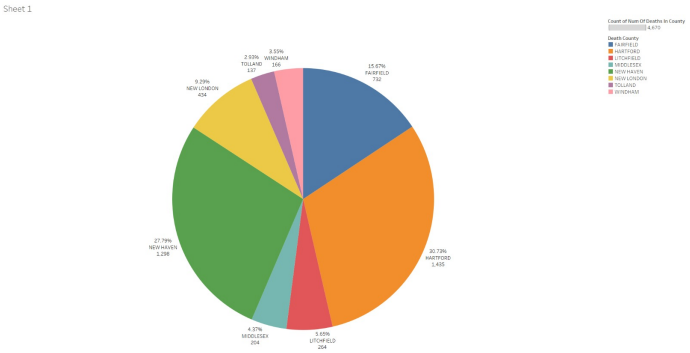

Fig. 26: Pie chart of death distribution by county.

This pie chart represents the distribution of death counts by county in Connecticut. The largest segment is Hartford County, accounting for 30.73% of the deaths, followed closely by New Haven County with 27.79%. Fairfield County holds 15.67%. Other counties like New London, Middlesex, Litchfield, Windham, and Tolland make up the remaining portions with smaller percentages. This chart visually emphasizes that the majority of deaths are concentrated in Hartford, New Haven, and Fairfield counties.

Thus we can see from fig 16 that Hartford, New Haven and Fairfield , the three richest counties in Connecticut have the most drug overdose deaths.

From the below tree map we can also see that the same counties have highest no.of drugs consumed as well.

Fig. 27: Treemap of deaths by county and city, with cell size representing death count and shade indicating number of drugs used.

The treemap visualizes the number of deaths by county and city, with each cell representing a specific location. The size of each cell corresponds to the number of deaths in that area, while the color shade indicates the number of different drugs used. Darker shades of color are associated with larger cells, suggesting that higher drug use correlates with a greater number of deaths. This visualization highlights the relationship between drug use and mortality across various locations.

The analysis of deaths caused by drug overdose reveals a striking trend in the types of substances contributing to fatal incidents. The accompanying pie chart visually represents the distribution of drug-related deaths over the years, segmented by the specific drugs involved.

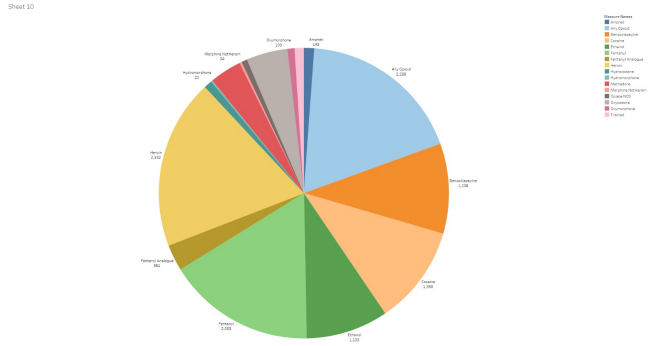

Fig. 28: Pie chart showing the distribution of deaths by drug type.

It is clear from the data that Heroin and Opioids are the predominant contributors to overdose deaths, significantly outpacing other drugs in the dataset. This indicates the severe impact of the opioid crisis, where synthetic opioids like fentanyl, as well as natural and semi-synthetic opioids such as heroin, have become the primary cause of fatalities. This chart underscores the critical public health challenge posed by these drugs, illustrating that efforts to combat overdose deaths must focus heavily on opioids. The visualization also highlights the disproportionate rate of deaths caused by heroin and opioids when compared to substances like cocaine, ethanol, or marijuana. Despite the widespread use of these other drugs, their contribution to the overall death toll is comparatively smaller. By illustrating the significant share of overdose deaths attributed to these substances, this data provides essential insights into the areas where public health interventions and policies need to be directed, particularly in addressing the opioid epidemic. This calls for continued education, prevention strategies, and access to treatment, specifically targeting opioid and heroin abuse, to reduce fatalities and mitigate the ongoing crisis.

One way to categorize drugs is based on the purpose of intake or their intended effects:

The histogram displays the counts of various substances involved in drug overdose cases, highlighting their

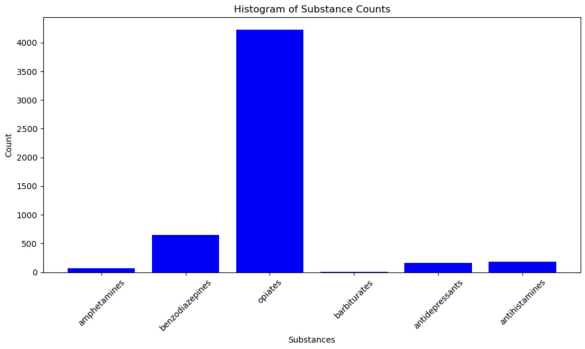

Fig. 29: Histogram of substance involvement in drug overdose cases

frequency of occurrence. The x-axis lists different substance categories, including amphetamines, benzodiazepines, opiates, barbiturates, antidepressants, and antihistamines, while the y-axis represents the count of cases in which these substances were present. The overwhelming majority of cases involve opiates, as seen by the tall bar dominating the graph, followed by benzodiazepines. Other substances, such as amphetamines, antidepressants, antihistamines, and barbiturates, are far less common in the data.

We can also categorize drugs based on their means of consumption. There are 4 major modes of drug intake by a human, these are:

1.) ORAL - This involves taking drugs through the mouth in the form of pills, liquids, or capsules.

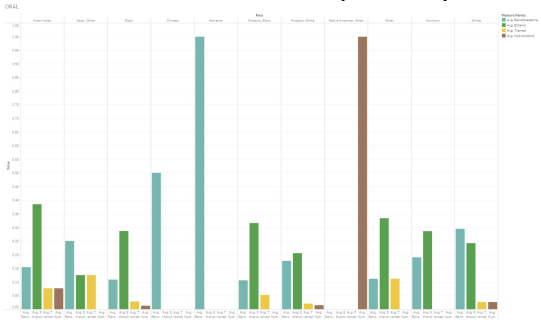

Fig. 30: Bar plot of average oral drug consumption by race.

The bar plot displays the average consumption of four types of oral drugs—Benzodiazepine, Ethanol, Tramad, and Hydrocodone—per person across different racial groups. The data indicates that all Native Americans who died due to drug misuse had used Hydrocodone, while all Hawaiians who died had used Benzodiazepine. Asian Indian shows a notable average for the consumption of Hydrocodone (green bar), with lower averages for Tramad (yellow bar) and Ethanol (orange bar)and Benzodiazepine (blue bar) usage is not significant.

Among Asian,Other Hydrocodone is also the most consumed drug, though the average values are relatively low compared to other races and Ethano, Tramad have lower yet comparable levels.

For Blacks Hydrocodone is again the most prominent

drug .Tramad and Benzodiazepine consumption are also observed at moderate levels, with Ethanol having minimal representation.

For the Chinese, the average value for Benzodiazepine consumption is exceptionally high in this group compared to other drugs. This indicates a significant prevalence of this drug within the Chinese group. Other drugs, such as Hydrocodone, Ethanol, and Tramad, are either non-existent or very low in comparison.

For Hawaiians Hydrocodone (brown bar) average consumption is significantly high for this group, representing a dominant drug in this population, they do not have significant consumption of other drugs.

For Hispanic, White Hydrocodone has the highest average consumption, followed by Benzodiazepine. Tramad and Ethanol have moderate and lower consumption rates, respectively.

In the group of Native American, Hydrocodone consumption is overwhelmingly dominant, with no significant averages in other drugs.

The most consumed drug amongst Whites is Hydrocodone, followed by Benzodiazepine and Tramad. Ethanol consumption is present but at a lower average value compared to other drugs.

2.)Injection - Drugs administered via needle, either intravenously, intramuscularly, or subcutaneously.

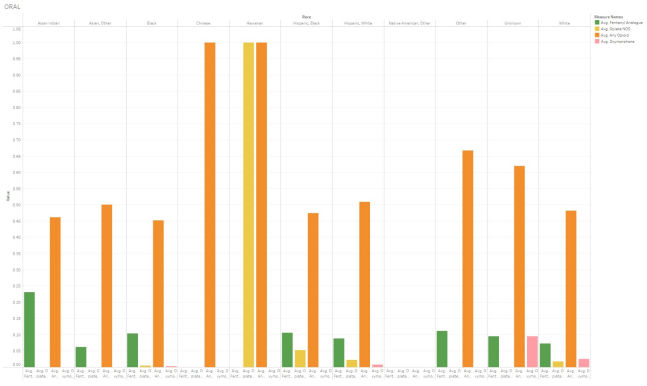

Fig. 31: Bar plot of average injection drug use by race.

The bar chart you provided shows the average oral drug consumption across different racial groups, with specific drugs like fentanyl analogues, oxycodone NOS, and oxymorphone highlighted.

Chinese and Hawaiian populations exhibit the highest average consumption of oral drugs, particularly oxycodone NOS.

Native American and Black individuals also show significant use of oral drugs, but at lower levels compared to the Chinese and Hawaiian populations.

The orange bars, representing oxycodone NOS, are predominant across most racial categories, indicating it is the most frequently consumed oral drug in this dataset.

There are smaller, but noticeable, consumption levels of any opioid and fentanyl analogues, represented by green and yellow bars respectively.

3.)Intranasal (Nasal) - Drugs snorted through the nose, allowing absorption through the nasal mucosa, which can produce a rapid onset of effects as the drug enters the bloodstream quickly.

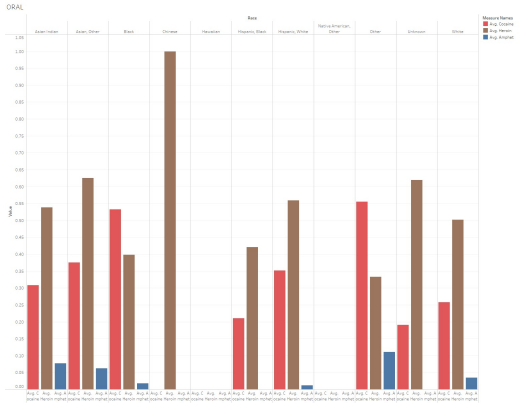

Fig. 32: Bar plot of intranasal drug use by race.

For Asian Indian and Others Heroin is the predominant drug consumed, with notable amounts of cocaine. Amphet consumption is minimal or non-existent in both subgroups.

For Blacks both heroin and cocaine are used in significant amounts. Amphet usage remains quite low, almost negligible compared to the other drugs.

For the Chinese, Heroin consumption is very high, making it the most consumed drug in this group. Cocaine and amphet usage are minimal or absent in this group.

For Hispanic Blacks and Hispanic Whites, Heroin is consumed in the highest quantities, followed by moderate levels of cocaine. Amphet is present, but its usage is quite low across both groups.

For Whites the most significant consumption is of heroin, followed by a considerable amount of cocaine. Amphet usage is lower but still present compared to other racial groups.

4.)Buccal: Drugs placed between the gum and cheek (buccal area) or under the tongue (sublingual) to be absorbed directly into the bloodstream through the oral mucosa.

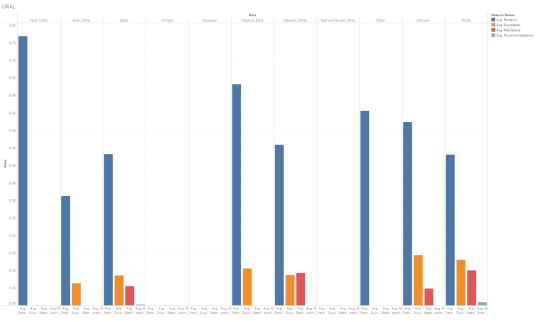

Fig. 33: Bar plot of buccal drug use by race.

Asian Indian: Fentanyl appears to be the predominant drug used, with its usage much higher than other substances. Oxycodone and Methadone have similar, though much lower, values compared to Fentanyl.

Asian, Other: There is a notable consumption of Fentanyl,

which dominates usage compared to the other drugs in this group. However, Oxycodone and Methadone have some, but lower, reported use.
Black: Fentanyl is once again the most consumed drug, though the values are lower compared to the previous groups. Methadone and Oxycodone are consumed in smaller, but still noticeable amounts.
Hispanic, Black: The consumption of Fentanyl remains significant, with much smaller amounts of Oxycodone and Methadone reported. Morphine usage in this group is minimal.
White: Fentanyl leads the chart again, being the most consumed drug. The consumption of Oxycodone and Methadone is slightly higher compared to the other racial groups, but still significantly less than Fentanyl. Morphine consumption is minor.
Therefore, Fentanyl stands out as the most prevalent drug in terms of average use across all racial groups. Oxycodone and Methadone also show some usage but in smaller quantities. Morphine has the least amount of usage across the dataset, regardless of the racial group.

3) Deaths, Race and Sex based analysis: The columns sex,race are very important to learn realistic insights about the deaths due to drug overdose.
Hypothesis 3: It is a common misconception that Black people use more drugs than other racial groups. In fact, drug use rates among Black and white Americans are very similar. However, Black Americans are significantly more likely to be arrested for drug-related offenses—about 2.7 times more likely than white Americans, despite similar usage rates [7].
But from the below figure, we can see a contradiction to this stereotype. The data actually suggests that more deaths due to drug overdoses occur among white individuals than Black individuals. This discrepancy could stem from several factors. One possibility is that the opioid crisis, which has disproportionately impacted white, suburban, and rural communities, plays a significant role in these overdose statistics. Furthermore, socioeconomic factors and disparities in healthcare access may lead to higher overdose mortality among white populations. This stands in contrast to public perception, illustrating that drug-related issues affect all racial groups and that overdose deaths, particularly in recent years, have been more prevalent in white populations due to the widespread availability and misuse of prescription opioids, fentanyl, and other synthetic drugs.
Additionally, while Black Americans face higher rates of incarceration for drug offenses, they are not necessarily engaging in higher levels of drug use or experiencing higher levels of overdose deaths compared to other groups. This challenges the traditional stereotype and highlights the need to address drug abuse and overdose as a broader societal issue, rather than one confined to any single racial group.

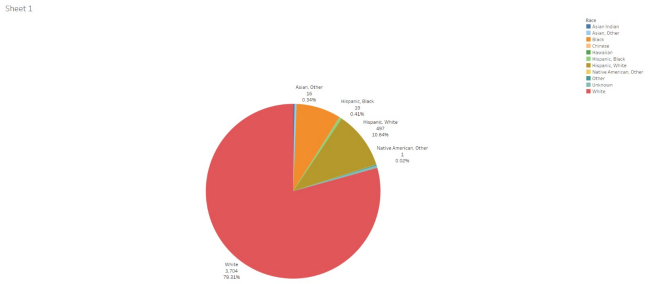

Fig. 34: Pie chart showing the distribution of drug-related injuries by race.
The pie chart illustrates the distribution of drug-related injuries across different racial groups. It reveals that White individuals constitute the majority of these injuries, accounting for 79% of the total. In contrast, racial groups such as Chinese, Native Americans, Hawaiians, Asians, Asian Indians, and others each represent less than 1% of the total injuries. This chart effectively highlights the significant disparity in drug-related injuries among different races.
We can reconfirm this observation by analyzing the trends in the below line graph, which further supports our hypothesis. The graph shows a clear distinction, with white individuals accounting for a larger proportion of deaths due to drug overdose compared to Black individuals. This directly contradicts the stereotype that Black people use more drugs or are more heavily involved in substance abuse.

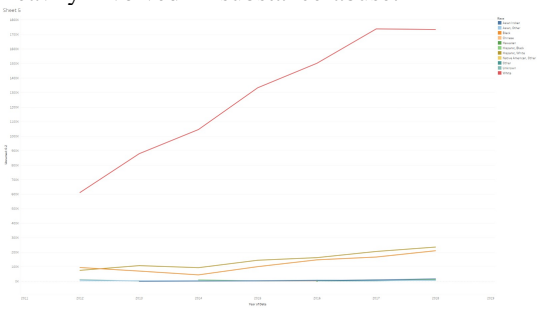

Fig. 35: Line plot showing yearly drug-related deaths by race, highlighting trends and comparisons.
The analysis shows a significantly higher number of drug-related deaths among White individuals compared to other racial groups across all years. The data indicates a sharp increase in deaths among Whites each year. Notably, the number of deaths among Whites remained consistent between 2017 and 2018. Additionally, the death counts for Hispanic and Black individuals are relatively close to each other throughout the years, indicating similar trends for these racial groups.
Therefore, the hypothesis—"It is a common misconception that Black people do more drugs than other racial groups"—has been disproven. The data shows that white people, particularly in the context of the opioid crisis, suffer more from drug overdose fatalities. This underscores the need to reshape public perceptions and address substance abuse as a complex issue affecting various demographics, rather than

reinforcing harmful racial stereotypes.

Next, we observe an even distribution of deaths within the white population across the state of Connecticut due to drug overdose. This suggests that, unlike certain racial or ethnic groups where overdose deaths may be more localized or concentrated in specific regions, the issue of substance abuse and its fatal consequences affect the white population more uniformly throughout the state. This even spread indicates that drug misuse is a widespread problem across various communities, irrespective of location, further reinforcing the narrative that substance abuse transcends racial and geographical boundaries.

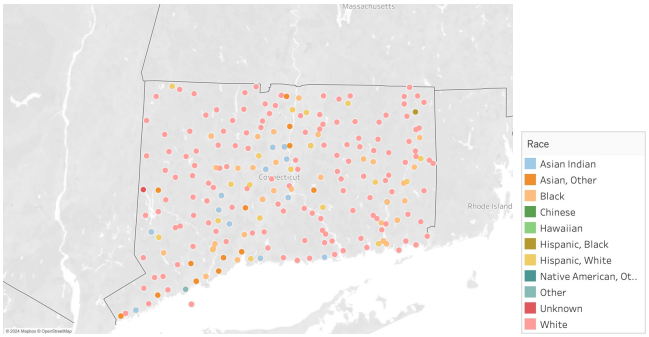

Fig. 36: Cartograph representing the race of death case subjects and their geographic distribution across the state

The cartograph in the report provides a visual representation of the racial breakdown of drug overdose deaths across Connecticut, linking each case to its geographic location. This visualization offers insight into the spatial and demographic distribution of fatalities, highlighting potential disparities among different racial groups. Each race is represented by distinct colors or symbols, allowing for easy identification of patterns across the state. Notably, the map may reveal geographic clusters where certain racial groups experience a higher concentration of overdose deaths, offering clues to regional disparities in health outcomes and access to services. The visualization underscores differences in urban and rural impacts, helping to inform public health efforts aimed at addressing these disparities in overdose fatalities.

Next, we observe the distribution of deaths due to drug overdose between males and females in the state of Connecticut. Data indicates that the number of male overdose deaths is generally higher than that of females, a trend commonly seen in many regions. This disparity can be attributed to several factors, including differences in drug access, societal pressures, and risk-taking behavior that may influence substance use patterns between genders. However, while males show higher fatality rates, it is important to note that female overdose deaths are also significant and have been steadily increasing, emphasizing that drug misuse impacts both genders deeply.

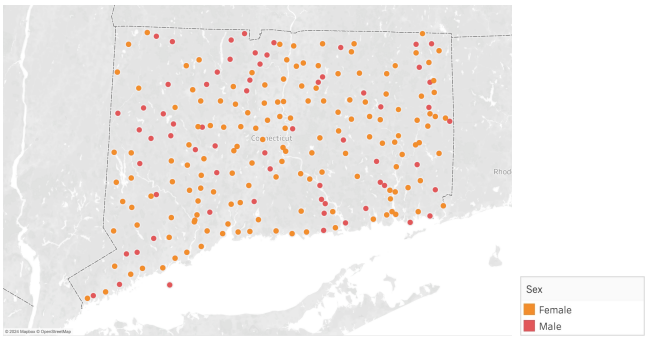

Fig. 37: Drug overdose deaths in Connecticut by sex

The map provides a detailed view of drug overdose deaths across Connecticut, with each point representing a case and color-coded by the sex of the victim—orange for females and red for males. The geographic distribution reveals significant clusters in southern Connecticut along the coast, as well as in central and eastern parts of the state. This visualization highlights potential regional and gender-based disparities, offering insights into the spatial patterns of overdose fatalities. The clear color distinction between male and female cases allows for a quick demographic analysis, helping public health officials identify high-impact areas for targeted interventions.

Finally, we examine how many different kinds of drugs are consumed by each age group, separated by gender.

Gender-wise, males across most age groups tend to consume a greater variety of drugs compared to females. This aligns with the higher incidence of drug-related deaths observed among men. However, in certain age brackets, particularly in younger cohorts, female drug consumption patterns have become more diverse, signaling a shifting trend in substance use.

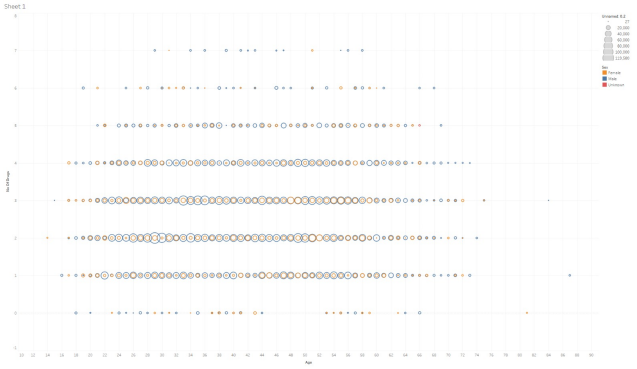

Fig. 38: Circle plot illustrating drug use and gender distribution in drug-related deaths by age group

The circle plot visualizes the number of drugs used by individuals who died across different age groups. Each circle in the plot corresponds to an age group, and the size of the circle indicates the number of deaths in that age group involving a specific number of drugs. Most deaths occur among individuals who used 2 or 3 drugs. The plot also highlights a gender distinction: blue circles represent males, and orange circles represent females. In many age groups, the blue circles are larger and form the outer ring around smaller, concentric orange circles, indicating that more males died

from drug use compared to females. This pattern suggests that male deaths are more prevalent in most age groups where drug use is involved.

E. Member Wise Contributions

After a detailed discussion about the dataset and understanding the context, we came up with the objectives mentioned in Section III. For the visualizations, objectives and hypotheses, the following distribution was followed:

1) Objective 1 (Deaths, Year and Age based analysis) was primarily done by Soham. Siddharth and Sourav suggested a few relations among the visualizations which were appropriately valid and necessary for the hypothesis. Soham also did the pre-processing of data for the COD column, which required concepts from NLP (Natural Language Processing) to extract useful features from the text. He also did the processing for generating geolocation coordinate columns for the cartographs and 'Date' formatting for the Area plot.
2) Objective 2 (Deaths, Location and types of Drugs based analysis) was primarily done by Siddharth. Soham and Sourav suggested a few relations among the visualizations which help relate the visualizations under the given hypotheses. Siddharth also helped plot several bar graphs and pie charts for the visualizations of Objective 2. He also helped with data cleaning, employing important rules for dropping rows and columns which helped in plotting many important visualizations.
3) Objective 3 (Deaths, Race and Sex based analysis) was primarily done by Sourav. Siddharth and Soham suggested a few relations among the visualizations and were appropriately connected under the hypothesis. Sourav also helped in forming many accurate hypotheses used in the report for all the Objectives. The hypotheses suggested and corrected by him made the visualizations more understandable and the inference easy to see.

REFERENCES

[1] Juan M Dominguez Ph.D., Psychology Today, "Cocaine Increases Risky Behaviors, Depending on Your Age ". Link
[2] Drug Policy Facts, "Young People and Substance Use". Link
[3] Addiction Center, "Young Adults". Link
[4] EUDA: European Union Drugs Agency, "Young people and drugs". Link
[5] Biomedcentral"adolescents with high socioeconomic status more likely to engage in drugs". Link
[6] NewHavenIndependent"New Haven and Hartford are affluent counties". Link
[7] Hamiltonproject"arresting of blacks vs whites for drug abuse". Link
[8] Keim, D., Andrienko, G., Fekete, J. D., Görg, C., Kohlhammer, J., and Melançon, G. (2008). Visual analytics: Definition, process, and challenges (pp. 154–175). Springer Berlin Heidelberg.