Analysis of Opioid Overdose Deaths

Members: Shubika Munot (spm54), Shreyans Gupta (sg609)

Part 1: Introduction and Research Questions

As recent Breaking Bad and Narcos viewers, we were captivated by the dramatized narratives of the drug trade and the complex dynamics that underpin illicit substance markets. The grim reality of opioid overdose deaths demands a rigorous examination that goes beyond the screen, prompting us to shift our focus from fictional portrayals to the real-world factors influencing the disparities in opioid-related mortality rates. According to the Centers for Disease Control and Prevention (CDC), opioids were responsible for over 70% of drug overdose deaths in the United States in recent years. After our initial exploratory data analysis, we found out that North Dakota was the best-performing state with just 2398 opioid-related overdose deaths and West Virginia was the worst with 47416 deaths. Intrigued, we questioned ourselves: Why was West Virginia performing this poorly compared to a state that was 3 times its size? Are there any particular factors that influence the performance of a state in its ability to counteract opioid damage? If so, what factors affect the state's performance more than others?

In this research paper, we will focus on a **singular broad research question**: What are the key contributing factors that could explain the disparity in opioid overdose deaths between high and low-affected areas?

The research question is **relevant** as opioids are the most common drug currently and making revelations about our hypotheses has practical implications for policy decision making and legal interventions. Additionally, pursuing research on factors influencing opioid overdose deaths is pivotal for public health. Our findings can potentially guide resource allocation, and enhance educational initiatives. The correlation with unemployment, education, poverty rates, and net migration highlights broader social and economic implications, potentially catalyzing systemic reforms. This research informs preventive healthcare strategies, and, overall, contributes significantly to addressing the opioid crisis and its societal impact.

Before our EDA, we made **some hypotheses** in the context of opioids:
1. Higher unemployment will have a positive relationship with opioid overdose deaths in regions.
2. Different education levels are not likely associated with elevated opioid overdose death rates.
3. Higher poverty rates have a positive relationship with opioid overdose deaths.

Part 2: Data Sources

Our data were sourced primarily from websites that are primary sources of public health information, guidelines, and resources. Our datasets are:

1. State Overdose Deaths - Dataset on U.S State-wise. Medicare-covered prescribers, opioids, and overdoses to analyze opiate prescriptions and fatalities. A kaggle subset sourced from cms.gov, the dataset gives us relevant information about state-wise distribution.
2. County Overdose Deaths - Dataset describes the number of opioid drug overdose-related deaths for each American county in each state from 2020 to 2022. Sourced from the CDC, this dataset provides relevant information about the country-wise distribution of overdoses.
3. County Wise Predictors - These datasets were sourced from the USDA Economic Research Service. The following datasets provide potential contributing factors for opioid-related deaths by county and thus can be used to answer the research question, **(i) Education** - This dataset describes population values for different educational attributes (eg. education levels, etc) for each American state and county over the following years, 1970, 1990, 2000, 2008-12, and 2017-21.

(ii) Population Estimates - This dataset describes population values for different attributes (eg. net migration, etc) for each American state and county for 2021.
(iii) Poverty Estimates - This dataset describes population values for different economic attributes for each American state and county for 2021.
(iv) Unemployment - This dataset describes population values for different employment attributes (eg. unemployment rate, etc) for each American state and county for the years 2000-2022.

Using the State Overdose Deaths we yielded that West Virginia has the highest opioid-related deaths/population while North Dakota has the lowest. Thus, we selected these two states and decided to focus on their counties as we reasoned that focusing on smaller regions would give us more accurate results.

To be able to use the data we collected concisely and systematically for exploratory and statistical analysis we performed in-depth data wrangling, cleaning, and preparation. To form our main merged dataset we started with the County Overdose Deaths and the 4 County Wise Predictors. The methodology used is as follows:

1. For the County Overdose Deaths dataset
    a. We changed the Overdose Deaths to numeric and dropped all NANs.
    b. We grouped the data by Year, State, and County, and summed up the deaths.
    c. We pivoted the table to making Year the column headers and Overdose Deaths as values and made new tables for county deaths in each year.
2. For all 4 County Wise Predictors tables
    a. We pivoted the tables to make each unique attribute a column with rows as the corresponding values for each county, selected the attributes we wanted from each, dropped the remaining, and ensured consistent column names for State and Area_name
    b. We separated the values into new tables for each year.
3. For both the County Overdose Deaths dataset and all 4 County Wise Predictors tables
    a. We selected tables for the year 2021, dropped all states other than West Virginia and North Dakota, and ensured consistent column names for State and County
    b. We merged these to get the final data table for predictors and deaths for 2021

After data wrangling, the dataset, reduced to 79 counties (73% of the original), maintained a representative ratio of opioid overdose deaths. Originally, North Dakota had 2398 deaths and West Virginia had 47,416 deaths, with a ratio of 0.05. Post-wrangling, the numbers became 693 for North Dakota and 16,719 for West Virginia, keeping the ratio similar at 0.04. Despite the removal of 30 counties, our sample remains reflective of the original dataset.

Part 3: What Modules Are You Using?

For our research project, we found most of our analysis to fall under these three modules:

1. **Module 04: Data Wrangling:** During our exploratory data analysis, we will utilize this module to handle data manipulation and cleaning which will involve dropping columns, converting data types, and aggregating information. We use this model to ensure the successful creation of our final dataset after operating on and merging our 6 initial datasets. Using the concepts from Module 04, we will prepare the data to be used to make visualizations.
2. **Module 03: Visualization:** After our exploratory data analysis, we will implement the *seaborn* library that will help us communicate complex relationships and patterns in the data effectively. Our visualization will range from bar plots, scatterplots, and boxplots that provide an intuitive understanding to advanced correlation and confusion matrices that

assess our model's performance metrics. We will use these visualizations to assess our predictors.

3. **Module 05: Statistical Inference:** During our main statistical analysis, we will use four different techniques from Module 05 which are simple linear regression, bootstrapping, t-test, and logistic regression. We use these techniques to measure the relationship between opioid overdose deaths and predictors such as education level, poverty, migration rate, and household income.

Part 4: Results and Methods

Our full implementation can be accessed in this *git repository*. For our analysis, we decided to pursue individual analysis of each predictor.

(1) Education Level

Individuals were classified into three categories: (i) Bachelor's Degree or higher, (ii) High School Diploma, and (iii) Less than a High School Diploma that took numerical values.

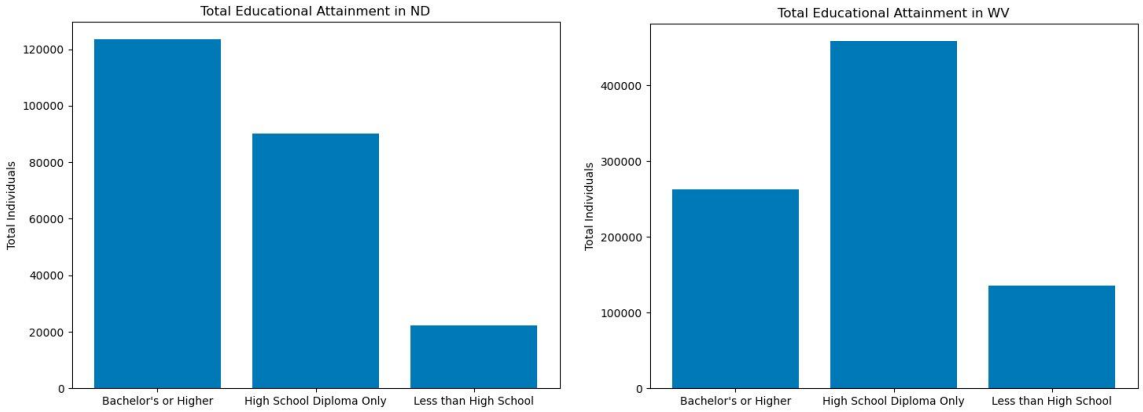

Fig 1.1: Distribution of Population by Education Levels in North Dakota and West Virginia

In our EDA, we noticed that the educational attainment in North Dakota was higher than in West Virginia as the majority in ND obtained a "Bachelor's or Higher" education whereas the majority in WV has a "High School Diploma Only". We reasoned that this discrepancy in educational attainment likely reflected differing socioeconomic statuses, potentially influencing vulnerability to opioid misuse and overdose so we decided to include it in our analysis.

To quantify the strength and direction of the relationship between each educational level and the opioid death cases, we decided to use a simple linear regression. This allowed us to assess which education levels are more strongly associated with overdose deaths and how they compare with each other. Setting X to each education level and Y to the predicted deaths, we obtained the following coefficients for the education level:

	Predictor Category	Coefficient
0	Bachelor's Degree or Higher	-0.006199
1	High School Diploma Only	0.004715
2	Less than a High School Diploma	0.157937

Table 1.1: Results from Simple Linear Regression Predicting Deaths by Educational Attainment

Bachelor's Degree or Higher: The negative coefficient (-0.0062) suggests that an increase in the number of individuals with a bachelor's degree or higher is associated with a slight decrease in drug overdose deaths. However, this effect is relatively small.
High School Diploma Only: The coefficient (0.0047) indicates a small positive association between the number of individuals with only a high school diploma and the number of drug overdose deaths.
Less than a High School Diploma: The positive coefficient (0.1579) is notably larger than the others, suggesting a stronger positive association between the number of individuals with less than a high school diploma and the number of drug overdose deaths.

We computed our model's performance metrics to assess its strength. The **r-squared** value is approximately 0.862 suggesting that about 86.2% of the variability in the number of drug overdose deaths is explained by the model whereas the **MSE value** is approximately 21333.73. In the context of the data which has a mean of 215 deaths and a standard deviation of 395.26 deaths, the MSE was significantly large suggesting that there's still a substantial average squared error in the predictions. This could be due to the scales of the independent variables being quite large compared to the dependent variable.
Overall, the linear regression findings indicate that higher education (e.g., bachelor's degree or higher) is associated with a slight decrease in drug overdose deaths, while lower education levels (e.g., less than a high school diploma) show a stronger positive association.

(2) Unemployment Rate

Unemployment rate is a key indicator of the health of an economy and the labor market that measures the percentage of the labor force that is unemployed and actively seeking employment.

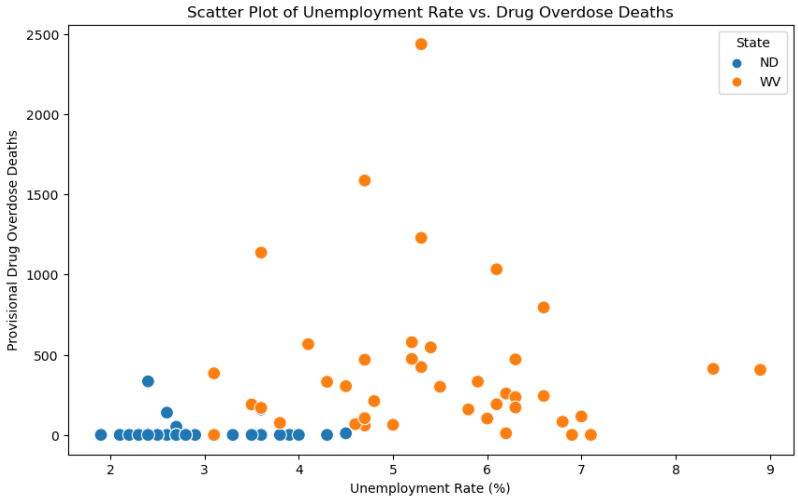

Fig 1.2: Distribution of Unemployment Rates between North Dakota and West Virginia

The scatter plot presents data points for two states, North Dakota (ND) and West Virginia (WV), comparing the unemployment rate to provisional drug overdose deaths. It seems to suggest that as the unemployment rate increases, the number of drug overdose deaths also tends to increase. West Virginia shows particularly higher numbers of deaths across varying unemployment rates compared to North Dakota. The data points for WV are more spread out and generally higher on the graph, indicating a broader range of unemployment rates and a more severe issue with drug overdose deaths than in ND. The observed patterns in the scatter

plot prompt the formulation of a hypothesis to investigate the potential relationship between opioid overdose deaths and unemployment rates. Our hypothesis was framed as below:

(i) **Null Hypothesis (H0)**: There is no difference in unemployment rates between counties with high and low opioid overdose death rates.
(ii) **Alternative Hypothesis (H1)**: There is a significant difference in unemployment rates between counties with high and low opioid overdose death rates.

To test this hypothesis we conducted a t-test. The t-test was employed to examine the potential relationship between unemployment rates and opioid overdose deaths in counties within North Dakota (ND) and West Virginia (WV). This statistical method is suitable for comparing means between two groups, allowing us to discern if there's a significant distinction in unemployment rates between counties with high and low opioid overdose death rates, yielding to it being a potential factor. The t-test analysis yielded the following results:

	Statistic	Value	P-value
0	t_stat_unemployment	6.712007	2.968735e-09

Table 1.2: Results from the t-test on the unemployment rate

The p-value is below 0.05 which means we can reject the null hypothesis. This means that we have sufficient evidence to conclude that there is a significant difference in unemployment rates between counties with high and low opioid overdose death rates. We can conclude that the obtained results showcase a statistically significant difference in unemployment rates between counties characterized by high and low rates of opioid overdose deaths.
We concluded from the t-test that the counties experiencing higher rates of opioid overdose deaths demonstrated notably elevated levels of unemployment.

(3) County-wise Median Household Income

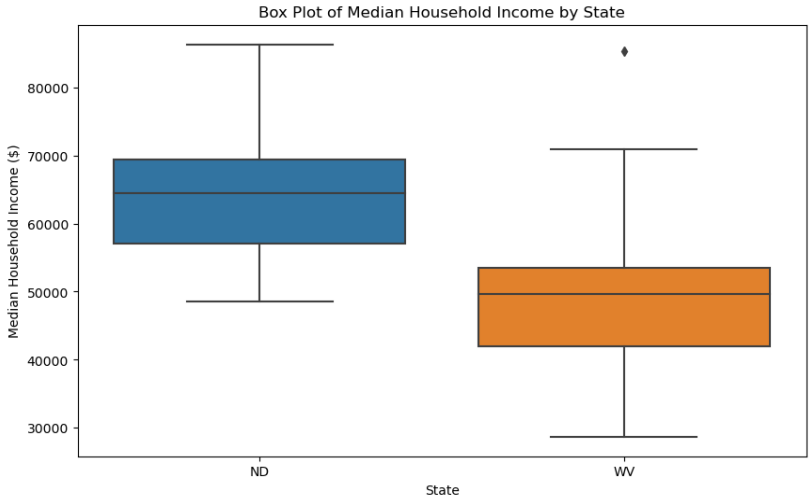

Fig 1.3: Distribution of Median Household Incomes between North Dakota and West Virginia

The box plot displays the distribution of median household incomes between North Dakota (ND) and West Virginia (WV). ND has a higher and narrower income range with its interquartile range (IQR) between $57,049 and $69,440, and a median of $64,454. This suggests a more

consistent income level among middle-class households. In contrast, WV's IQR is between $41,958 and $53,560 with a median of $49,704, indicating a broader spread of household incomes and overall lower earnings. The disparity in median household incomes is clear, with ND households generally earning more than those in WV. We reasoned that a lower median income may correlate with limited access to healthcare and substance abuse treatment resources as well as higher stress levels, potentially increasing vulnerability to opioid addiction and overdose.

Therefore, to measure the impact of Household Income on overdose deaths, we conducted logistic regression because it is a robust statistical method suited for binary outcome variables. In this case, the outcome variable (provisional drug overdose deaths) was transformed into a binary format (high vs. low risk of overdose deaths, based on a defined threshold: median). This transformation allowed us to use logistic regression to explore the relationship between a continuous predictor (median household income) and a binary outcome (overdose deaths). The goal was to understand whether and how median household income might influence the likelihood of a county experiencing high or low rates of opioid overdose deaths.

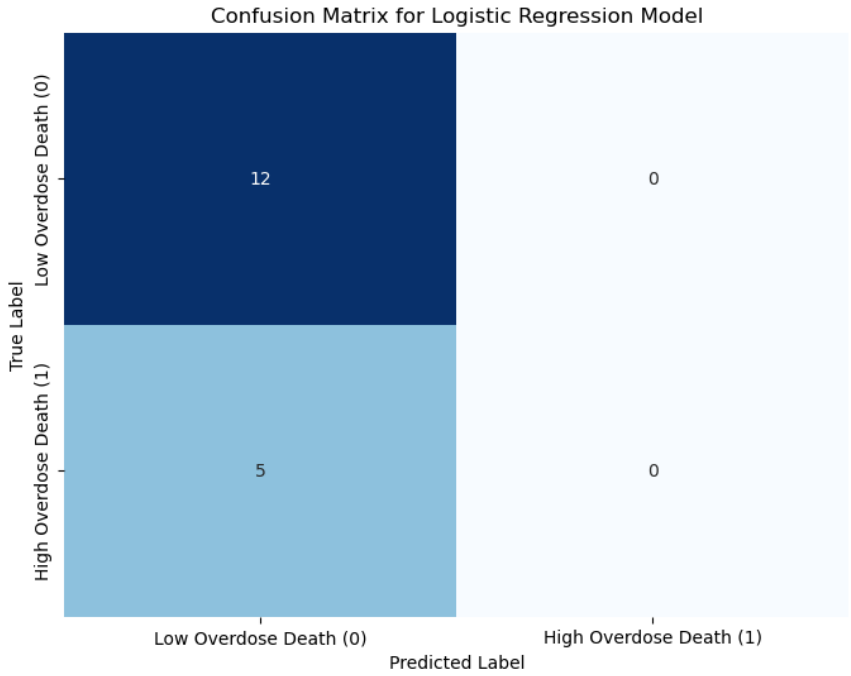

Fig 1.4: Confusion Matrix for Logistic Regression Model

To assess the performance of our model, we built a confusion matrix which highlighted:

- **Precision for Low Overdose Death Rate Counties (71%):** This indicates that when the model predicts a county to have a low overdose death rate, it is correct 71% of the time. However, this relatively high precision might be misleading due to the model's bias towards predicting low overdose death rates.
- **Recall for Low Overdose Death Rate Counties (100%):** The model identified all counties with low overdose death rates correctly. While this might seem positive, it's important to note that this is also due to the model's bias towards predicting low overdose rates.
- **F1-Score for Low Overdose Death Rate Counties (83%):** The F1-score, which is a balance between precision and recall, is relatively high for low overdose death rate

counties. This again reflects the model's effectiveness in identifying low-risk counties but doesn't provide insight into its performance for high-risk counties.

- **Failure to Predict High Overdose Death Rate Counties:** The model's complete failure to predict any high overdose death rate counties (with both precision and recall at 0% for these counties) is a **critical limitation**. This suggests that the model, with median household income as the sole predictor, does not capture the factors that contribute to high overdose death rates. This is a significant concern for the research question, as understanding the factors contributing to high overdose rates is crucial.
- **Overall Accuracy (71%):** While the overall accuracy might seem reasonably high, it's mostly driven by the model's bias towards predicting low overdose rates. In the context of the research question, this accuracy is not very informative, as the key focus is on understanding the disparity in overdose deaths, including both high and low-affected counties.

Overall, the logistic regression model suggests that median household income alone may not be a sufficient predictor for understanding the complex dynamics of opioid overdose deaths.

(4) Net Migration Rate

This statistic measures the difference between the number of people entering the county and the number of people leaving divided by the population. Net migration alters a region's demographic composition and also reshapes the social fabric, affecting support networks, community engagement, and cohesion, which are crucial in individual health behaviors and substance use patterns.

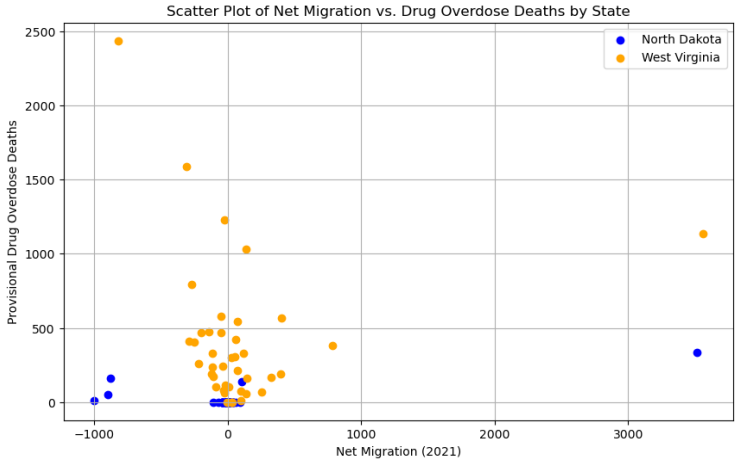

Fig 1.5: Distribution of Net Migration by Two States

Seeing that there were no visible interpretations of the plot, we decided to use bootstrapping to statistically analyze the relationship between net migration and overdose deaths. We opted for bootstrapping since it allows for estimating the distribution of mean by resampling from the data with replacement. This was particularly useful since we were dealing with small datasets and made it ideal for understanding and quantifying relationships in data where clear patterns are not immediately evident, as in the case of the net migration and overdose deaths scatter plot.

	Metric	Value
0	Mean Effect	-163.181543
1	Lower Bound	-318.102469
2	Upper Bound	-13.665432

Table 1.2: Results from Bootstrapping

The bootstrapping analysis on the effect of net migration on provisional drug overdose deaths yielded the following insights:

- **Mean Effect:** The mean effect of approximately -163.18 indicates a negative association, suggesting that counties with higher net migration tend to have fewer drug overdose deaths on average.
- **Confidence Interval:** The 95% confidence interval, ranging from about -318.10 to -13.67, highlights the variability in this association. The negative interval suggests a general negative correlation, but the width indicates significant variation across counties.
- **Interpretation:** The negative association could imply favorable conditions in counties with net migration gain, such as better healthcare access or economic opportunities. Factors like urbanization and demographics might also play a role.

Overall, while the model does indicate a negative association, these findings underscore the need for nuanced consideration of various factors influencing opioid overdose deaths.

Part 5: Limitations and Future Work

Limitations: The research conducted on opioid-related mortality rates, socio-economic factors, and net migration in North Dakota and West Virginia offers valuable insights but also presents several limitations that must be considered for a more comprehensive understanding. Firstly, the scope of education data is limited, with only a few categories of educational attainment considered, and the focus is predominantly on single variables like education levels, unemployment rates, and median household income. This approach may overlook other critical socio-economic factors such as income levels, access to healthcare, cultural attitudes, and the complex interplay of factors influencing opioid overdose deaths. The Linear and logistic regression may oversimplify the intricate nature of drug addiction and its socio-economic determinants, potentially leading to biased predictions and overlooking non-linear relationships. Moreover, the generalizability of findings is questionable due to the unique socio-economic and cultural contexts of the studied states, and data limitations could affect the accuracy and reliability of the results.

Future Work: We could expand the data scope to include more detailed educational and income categories, and improve data quality. Advanced analytical techniques, such as non-linear models or machine learning algorithms, could provide a more nuanced understanding of the socio-economic drivers of opioid addiction. Longitudinal and multivariate studies are necessary to observe temporal changes and understand the interplay of various factors. Broadening the geographic scope of the study by looking at more regions in the US and more countries can enhance the generalizability and depth of the findings. Finally, exploring the impact of specific policies and interventions on opioid mortality rates could offer practical insights for effective public health strategies. These future research directions would not only address current limitations but also enrich the understanding of the socio-economic determinants of opioid overdose deaths. We created a correlation matrix between all predictors to see if there is a future scope to analyze any other predictors in the dataset. Predictor “Civilian Labour Force” shows a strong positive correlation which may be of future interest.

Part 6: Conclusion

Our analysis unveiled key insights into opioid overdose deaths. Lower education levels, particularly less than a high school diploma, showed a significant positive association. Elevated unemployment rates correlated with increased opioid fatalities. Predicting high overdose rates using median household income had limitations. However, a negative association was found between net migration and opioid deaths, suggesting fewer fatalities in counties with higher migration. These nuanced findings contribute valuable perspectives for addressing the opioid crisis and informing policy decisions.