Written Report - Final Project

Isabella Villanueva

December 6, 2024

Introduction

In 2023, the opioid crisis was declared a matter of "national health emergency". But the American opioid crisis has been a major public health issue long before 2023. Marked by a significant rise in drug poisoning deaths over the past few decades, this epidemic began in the late 1990s when pharmaceutical companies reassured the doctors and prescribing health workers that opioid pain relievers were not highly addictive, leading to widespread over-prescriptions to treat those in pain. Because of the misinformation about the addictive nature of these prescription opioids, such as OxyContin and Vicodin, as well as their increased availability, misuse was inevitable. Over time, patients who became addicted to prescription opioids often transitioned to cheaper, more accessible alternatives like heroin and synthetic opioids such as fentanyl.

By the 2010s, the crisis had escalated. The Centers for Disease Control and Prevention (CDC) identified three distinct waves of opioid-related deaths:

1. **1999–2010**— Marks the genesis of the opioid crisis with rising deaths from prescription opioids
2. **2010–2013** — A surge in heroin use and overdose deaths as access to prescription opioids became restricted
3. **2013- Present** — The rapid rise of deaths due to illicitly manufactured fentanyl (synthetic opioid overdose), which is far more potent than heroin or prescription opioids. Yet fentanyl is exponentially more dangerous because of its intense potency in small quantities – making it lethal.

National public health agencies, like the CDC, are able to report the numbers of deaths in the United States because of the death certificate reporting process, which involves the use of International Classification of Diseases, Tenth Revision (ICD–10) codes. "Drug-poisoning deaths are defined as having ICD–10 underlying cause-of-death codes X40–X44 (unintentional), X60–X64 (suicide), X85 (homicide), or Y10–Y14 (undetermined intent)" (CDC). According to National Library of Medicine article "Defining indicators for drug overdose emergency department visits and hospitalisations in ICD-10-CM coded discharge data", the diagnosis code being used for drug poisoning cases would begin with the letter T, "drug poisoning T-codes" indicate "poisoning by unspecified drugs, medicaments and biological substances, accidental (unintentional), initial encounter" (Vivolo-Kantor et al.).

In the dataset I have chosen: "Drug Overdose Death Rates by Drug Type, Sex, Age, Race and Hispanic Origin in the United States (1999 – 2016)", the CDC reports 2862 observations (rows) of 19 variables (columns), such as crude and age-related death rates related to drug poisoning, the states reporting these deaths, and the demographic data from each individual – involving salient identity factors like race and Hispanic origin, sex, and age. These variables will be vital to exploring which US region has seen the largest influx of drug poisoning mortality rates in comparison to the US crude rate of drug poisoning mortality.

Questions of Interest:

1. Over time (1999-2016) and in consideration of all ages and all origins, which regions of the United States have had the greatest increase of drug poisoning death rates?

2. Throughout the United States, is there a strong association with Hispanic origin and drug mortality?
3. How do drug poisoning death rates compare between men and women nationally from 1999-2016?

Methods

How and Where the Data Were Acquired

This public access dataset was created by the National Center for Health Statistics (NCHS) in December 2017, and available on the CDC Data website, linked on my website's About page. Among the NCHS content on this database, a plethora of drug poisoning mortality datasets and visualization tools are available to assess the opioid crisis under different lenses. For this dataset, multiple variables of interest were included like sex, age, race and Hispanic origin, and region. This made this specific dataset the most attractive because of its amount of information available.

Cleaning and Wrangling the Data

Understanding the NA Values

In cleaning the dataset, checking for NA values is an important step. Using the two functions: `colSums(is.na(drug_mortality))`, there is an equal amount of NA values across the variables named Age-adjusted Rate, Standard Error for Age-adjusted Rate, Lower Confidence Limit for Age-adjusted Rate, and Upper Confidence Limit for Age-adjusted Rate. The age-adjusted rates are missing for many entries, and the decision to filter them out or use an alternative rate (like crude rates) must be made.

This equal amount of 1728 NA values is due to how age-adjusted rates cannot be calculated without the known weight of each age group within a population each year. When assessing the dataset, the NA values only occur when matched with the national data (when the State variable value is equal to "United States") as well as when the Age Group variable is a numbered age range other than the value "All Ages". Because all of the NA values are related to the age-adjusted rates, this is not an issue for using the data. Deleting the data with NA values would be a mistake as it augments our understanding of the true data, and imputing the data to be the mode or mean of similar characteristics is not possible because it would not be accurate to what the variables measure.

Crude death rates will be used instead when considering specific states and specific age ranges, gender, or ethnicity and when considering the national data, the variable US Crude Rates will be used.

Removing Irrelevant Variables

Next in cleaning the data, I opted to remove two columns (variables) because they were not relevant to the data being analyzed. The Unit and State Crude Rate in Range variables will not be necessary in this project because of our intended research question and scope of our data analysis. The original dataset's Unit variable has only one value of "per 100,000 population". This does not fill any gaps of information that another variable cannot inform, and removing this variable is not a loss in the greater scope of the data. The variable State Crude Rate in Range is not relevant to the question of interest as we are considering regions in the United States instead of individual states as well as the comparison to the US Crude Rate, the State Crude Rate in Range variable is also repetitive and contains excess information.

After cleaning the data and using the functions: `str(drug_mortality[sapply(drug_mortality, is.character)])`, six variables were loaded in as `chr` (character or text) variables: State, Sex, Age Group, Race and Hispanic Origin, State Crude Rate in Range, and Unit. The last two variables have been removed. Using the functions: `str(drug_mortality[sapply(drug_mortality, is.double)])`, 13 variables were loaded in as `dbl` (double or numeric with decimal) variables: Year, Deaths, Population, Crude Death Rate, Standard Error for Crude Rate, Lower Confidence Limit for Crude Rate, Upper Confidence Limit for Crude Rate, Age-adjusted Rate, Standard Error for Age-adjusted Rate, Lower Confidence Limit for Age-adjusted Rate, Upper Confidence Limit for Age-adjusted Rate, US Crude Rate, and US Age-adjusted Rate.

Seeing that the Sex, Race and Hispanic Origin, and Age Group variables are character or text variables and have repeat values, I decided to create these variables as factors where repeated values are then organized into categories (i.e. Sex the values: "Both Sexes", "Female", and "Male"). This will make calling on these variables in the future easier, like in the instance of comparing drug mortality rates between the male and female sexes. Using the function: `str(drug_mortality)` , I was able to authenticate the conversion of variables to categorical variables where the variables values were listed by each unique value (i.e. the Sex variable should only have 3 values if done correctly).

In the case of the variable "Race and Hispanic Origin" (RHO), values like All Races - All Origins, Non-Hispanic Black, Non-Hispanic White, and Hispanic were all present in this column. To combat the excess of wordings to get information about this variable, I opted to create a new categorical variable (new column) named Hispanic and assign values whether the individual is of Hispanic origin or not. When the RHO column indicated only the word "Hispanic" the value would be 1, while the value would be 0 if the RHO column indicated "Non-Hispanic" along with their race, or if it indicated "All Races - All Origins".

Representative States for U.S. Regions

Our research question of interest asks what regions have had the greatest increase in drug poisoning mortality, I divided all 50 states (and the District of Columbia) into their respective US regions according the how the CDC defines the geographic regions in the United States. In order to gain a more comprehensive look at the data presented, and accurately create a visualization of the data, the use of all 50 states (and additional territory of the nation's capital, Washington D.C.) is necessary.

In the `State` variable, "United States" is also included as a value which indicates that data of those rows are national data. This points to the US Crude Rates, which will be useful when comparing the regions to the national rates.

#Add more

Tools Used for Data Exploration

Results

Average Drug Poisoning Death Rate by US Region Compared to the National Average (1999-2016)

Analysis for graph above

This graph categorizes all 50 states and the District of Columbia into the four US regions. With this, we can properly assess the mortality rates of each region and can conclude that the Northeast region has had the greatest increase in mortality rates, as indicated by the steep slope of its line (large increase over smaller period of time) as well as the magnitude of the mean death rate value (nearing 40 deaths per 100,000 at the peak in the year 2016). When ranking the steepness of linear slopes of each region, the South has the second steepest slope, then the National mortality rate, then the Midwest, and finally the West. The West region is ranked last in this ranking, though in the graph it has the third highest mortality rate of about 17 deaths per 100,000 persons, its slope plateaus around the year 2011-2012, and steadily rises as seen in the faceted line plot. (fix national data explanation)

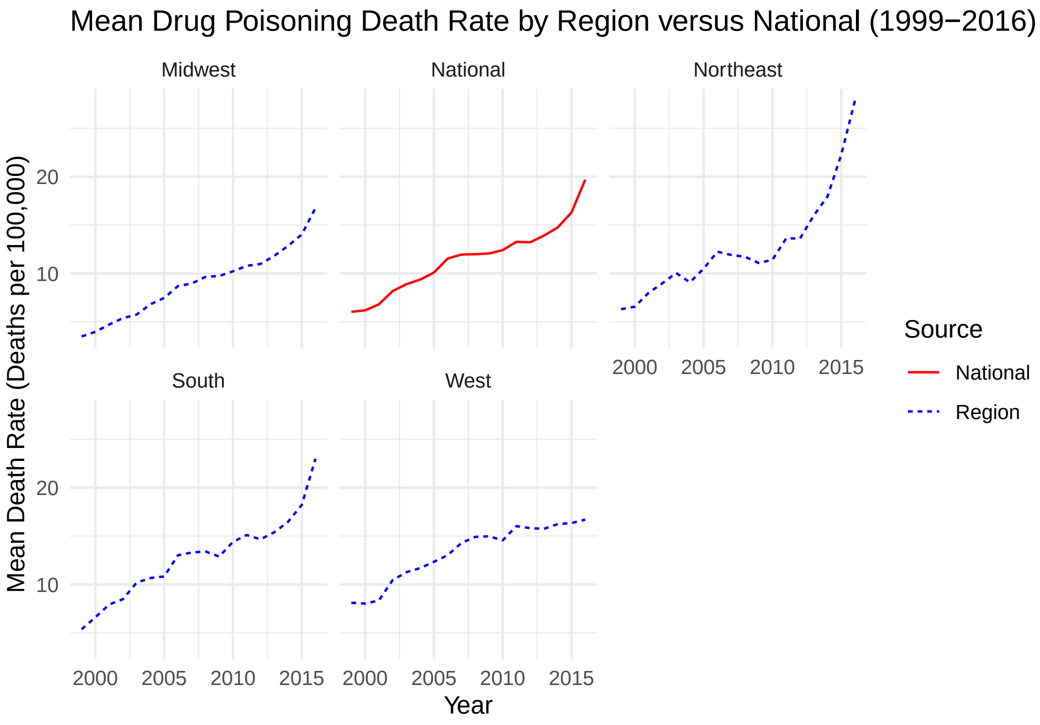

Figure 1: Faceted Line Plot by Region vs National

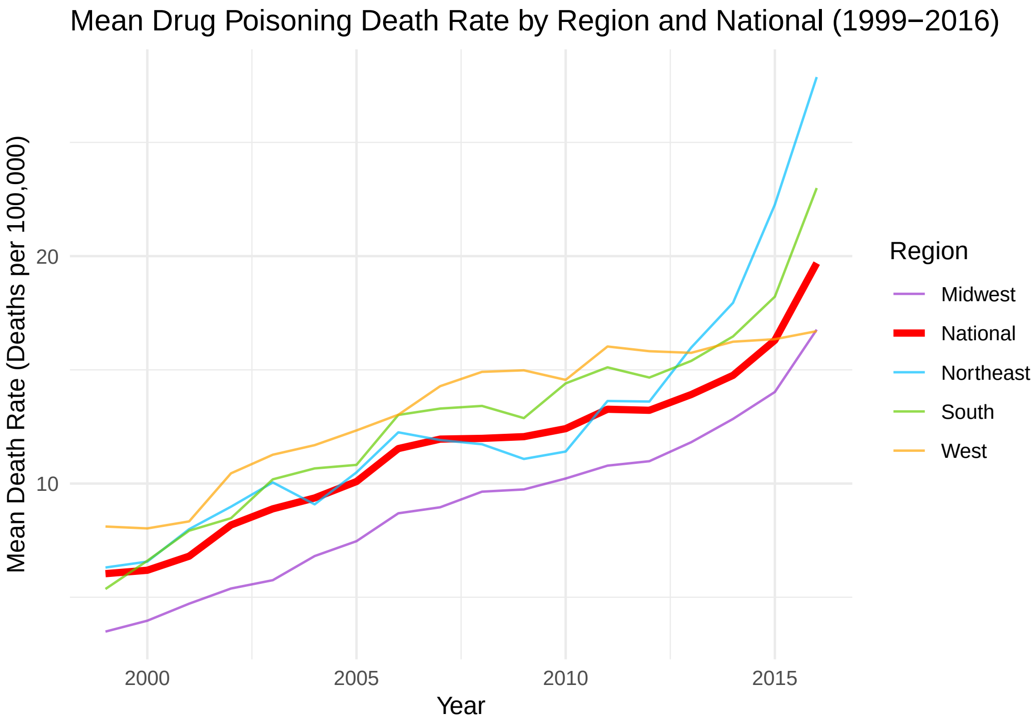

Figure 2: Line Plot of US Regions Compared to National Average Drug Mortality

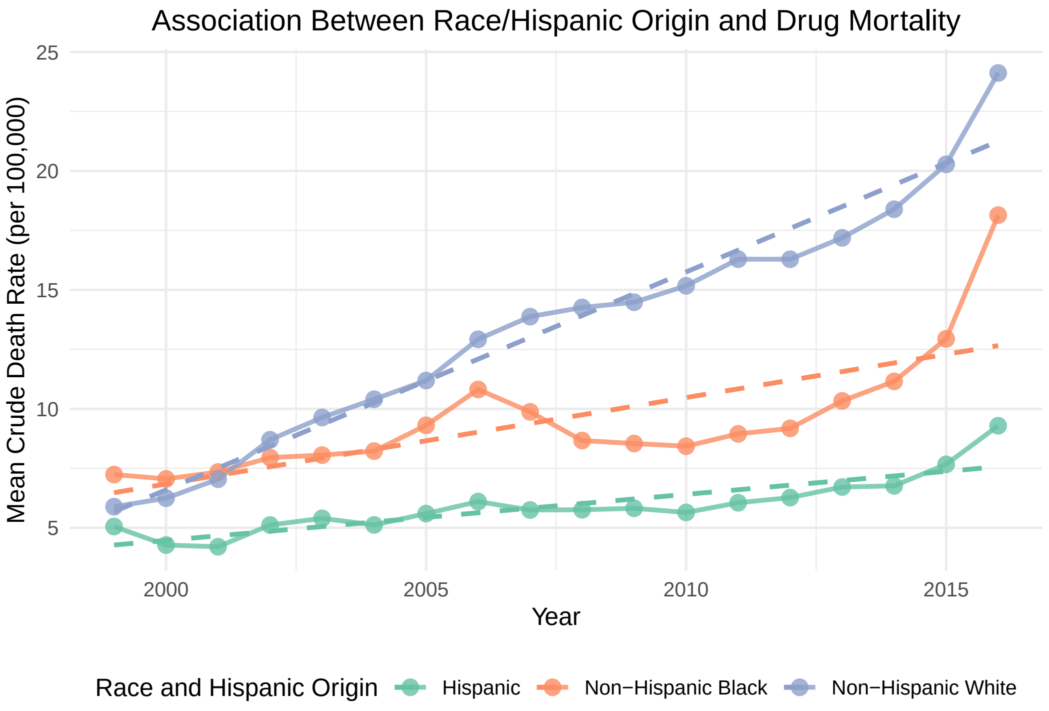

Figure 3: Line Plot with Dashed Regression Lines of Race/Ethnicity

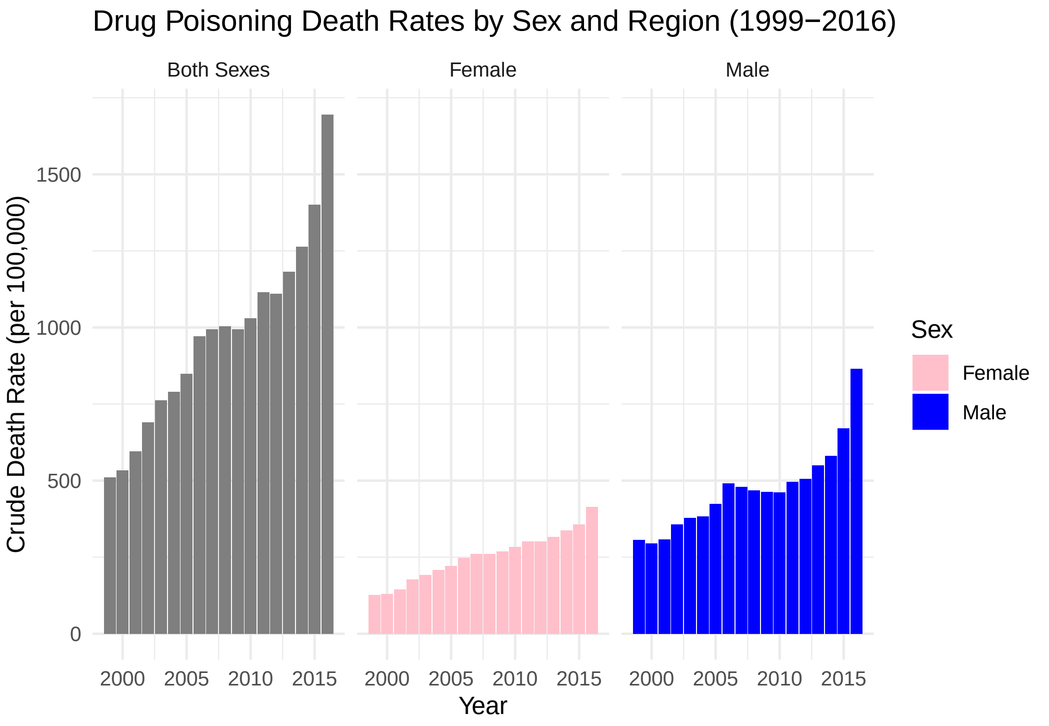

Figure 4: Faceted Barplot by Sex

Conclusion

The following graphs were created to visualize the numerical data of the average drug poisoning mortality rates in each US region and compare these rates to the national average. When considering the research question of interest, the US region that has the greatest increase in drug poisoning mortality in comparison to all US four regions and the national average was **the Northeast region**, clearly visualized by the second line plot graphing “Average Drug Poisoning Death Rate by US Region (1999-2016)”.

Potential Factors Behind Observed Trends

Public health responses to the declaration of a national health emergency have affected populations, yet some regions of the United States may have more ease of accessing health care arising from policy like increased regulation of opioid prescriptions, efforts to expand access to treatment for opioid use disorder, and the distribution of naloxone — a life-saving medication that can reverse opioid overdoses. This may explain the plateau of the West region’s drug poisoning mortality rates, while Northeast and South rates continued to skyrocket – as these regions are politically more diverse and unlikely to resolve policy easily. Despite these efforts, drug poisoning deaths have continued to rise in the last decade, exacerbated by the increasing presence of synthetic opioids like fentanyl noted as the “third wave” of the opioid epidemic.

The opioid crisis undoubtedly disproportionately affects certain demographics, with higher rates among middle-aged adults, non-Hispanic whites, and men. It would be worth exploring the correlation rates between these demographics, such as sex, Hispanic origin, and age, with the newfound knowledge gained from this data analysis of US regions and death mortality rates.