CAPSTONE PROJECT:

Analyzing Drug Overdoses in the United States

Individual Write-Up
by Rhiann Zhang

Team Members: Ashlyn Jew, Meera Duggal, Wei Deng

March 20, 2022

Analyzing Drug Overdoses in the United States

I Problem Description

Throughout the years, there has been a dramatic increase in drug overdoses. We are currently at the point where overdoses are the cause of more deaths than car accidents, guns, or HIV [1]. Since 1999, there have been about 1 million fatal overdoses in the United States and just within the past year, over 100,000 have lost their lives due to overdose [2]. It is very clear that the drug epidemic is a widespread societal problem that needs to be properly researched and addressed.

To do so, our team will be analyzing the number of fatal overdoses in United States counties throughout the last 20 years and exploring the question: How do different aspects of counties (e.g. location, demographics, and opioid dispense rates) relate to drug overdose rates? We also hope to estimate overdose rates in the counties that are missing data in our main overdose dataset. The successful completion of these tasks will allow us to provide policymakers with valuable information about the areas of the U.S. that are not easily observed simply with the raw overdose counts as well as insight as to how to more effectively address and counteract this epidemic.

II Data description

Data Generation

Our team obtained our drug overdose data from the CDC Wonder Search website by requesting data that specifically pertains to overdoses using their "underlying cause of death" codes [3]. The CDC collected this data through the Vital Statistics Cooperative to provide health departments and the general public with open access to detailed information that is beneficial in public health research and decision making.

Variables of Interest

Our primary overdose dataset provides variables of interest such as year, county, overdose death count, and county population. We used these features to calculate the `Overdose Rate` which is the number of overdoses per 100,000 people. This will act as our main variable of interest. In this dataset, we observe an inconsistent number of counties each year. However, the number of observed counties does increase throughout the years (Figure 1) and we have also confirmed that each year of data includes more than 50% of the United States population, despite missing a significant amount of counties. We have also gathered additional datasets detailing various factors that we may relate to overdose rates (opioid dispensary rates [4], unemployment rates [5], ethnicity [6], poverty [7], median household incomes [7], and incarceration rates [8]). We have acquired all of our data by county and will be analyzing the relationship between our

chosen features at this granular level. Unfortunately, we were unable to locate accurate and robust data for some of our additional variables on a county level before the year 2010. Hence, our analysis will focus on the years 2010 to 2019 in order to minimize the number of missing values while maintaining the wide time range of a decade (Figure 2). Currently, our dataset includes 8835 rows and 41 columns.

Hypothesized Relationships

Our team hypothesized that socioeconomic status would be closely connected to overdose rates. Hence, we included variables such as poverty and unemployment rates, assuming that overdose rates would be higher in counties where these factors were higher. We included median household income as well, presuming that counties with a lower median income would correspond with counties with higher overdose rates. We also surmised a possible association between higher opioid dispensary rates and higher overdose rates. Additionally, we explored incarceration rates, supposing that an increased crime rate would correlate to higher drug use and therefore, a higher overdose rate. We considered the possibility that minority groups would be disproportionately affected by the epidemic and included the ethnicity and sex demographics of each county as well. Our team also postulated that the urbanicity and total population of a county would be positively correlated with drug overdoses. Due to the clustering of higher overdose rates in different areas of the United States which can be clearly seen in Figure 3, we postulated that the geographical location of each county and their relative positions with each other would be a strong indicator in overdose rates.

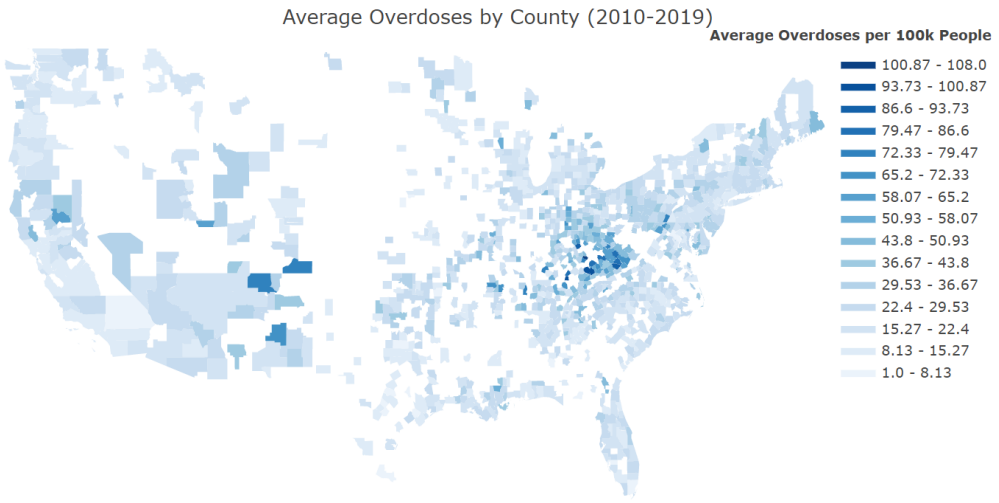


Figure 3. Average Overdose per 100k People by United States Counties (2010 - 2019)

So for each county, we calculated the average overdose rate of its adjacent counties during that year - which we labeled `spatmean` - as well as the maximum overdose rate of its adjacent

counties - which we label ‘spatmax’ - believing that higher overdose rates in neighboring counties would correspond to a higher overdose rate in the focal county.

III Methods

Exhaustive Feature Selection

Of course, many of the variables that we are considering also have relationships with each other. To help account for this multicollinearity, we ran an exhaustive feature selection to determine which of our additional demographic features would best describe our overdose variable. This resulting subset of features: Year, Unemployment_rate, Dispense_rate (Number of Opioid Prescriptions per Capita), AA_MALE (Number of Males that identify exclusively as Asian American), TOM_MALE (Number of Males that identify with two or more ethnicities, NH_MALE (Number of Males that identify as non-Hispanic), Jail Population, Incarceration Rate per 100k People, PoveryCount (Number of people below the poverty threshold), MedianHHI (Median Household Income), will be the ones we use in our first round of preliminary modeling.

Moran’s I

We also utilized Moran’s I to determine the significance of the spatial component in our data. Moran’s I can be calculated with

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \sum_{i=1}^{n} z_i^2}$$

such that n is the total number of features, $w_{i,j}$ is the spatial weight between the i-th and j-th feature, and z_i is an attribute’s deviation from its mean $(x_i - \bar{X})$ for feature i [9]. Our Moran’s I is about 0.4608 for our United States overdose data, indicating a strong positive spatial autocorrelation relationship between the counties. Hence, we ensured that our naive spatial components: spatmax and spatmean (explained in Section 2), were taken into account in our first round of modeling.

OLS Regression

For our preliminary round of modeling, we used a general ordinary least squares (OLS) regression which is given by

$$C_i = \alpha + \sum_{k=1}^{p} \beta_k X_{ik} + \epsilon_i$$

such that C is our estimated response variable, X is our set of features, p is the number of features included in the model, β_k is the coefficient that describes the relationship between the k -th feature and our response variable, α is the intercept, and ϵ is the error term.

In order to systematically investigate our spatial components, we implemented variations of the OLS model: "Cruder_Rate ~ Year + Unemployment_rate + Dispense_rate + AA_MALE + TOM_MALE + NH_MALE + Jail_Population + PovertyCount + MedianHHI + spatmax + spatmean". We began by running a model that only included the best subset of features determined by our exhaustive selection. We then built upon this initial aspatial model by implementing two additional models: one including spatmax and another including spatmean. Afterwards, we implemented the model explicitly written out above, which included our best subset of features and both of our naive spatial components.

IV Results

In Table 1, we compare the essential findings from each of our models. Notably, in our first aspatial model (labeled Best Subset (BSS) OLS), all of our covariates are statistically significant, which may be due to our large number of observations. When including spatmax,

	Best Subset (BSS) OLS		BSS OLS w/ spatmax		BSS OLS w/ spatmean		BSS OLS w/ spatmax & spatmean	
	Estimate	P-value	Estimate	P-value	Estimate	P-value	Estimate	P-value
Intercept	-367.99	0.0	-2593.707	0.0	-2080.257	0.0	-2083.644	0.000
Year	0.183	0.0	1.299	0.0	1.044	0.0	1.045	0.000
Unemployment_rate	0.152	0.0	0.958	0.0	1.161	0.0	1.148	0.000
Dispense_rate	0.299	0.0	2.066	0.0	1.56	0.0	1.566	0.000
AA_MALE	0.098	0.0	0.911	0.001	0.387	0.157	0.403	0.142
TOM_MALE	-0.146	0.0	-1.12	0.011	-0.532	0.208	-0.547	0.196
NH_MALE	0.222	0.0	2.034	0.0	1.941	0.0	1.938	0.000
Jail_Population	-0.1	0.001	-0.708	0.025	-0.556	0.068	-0.559	0.066
Incarceration_Rate_per_100k	0.023	0.044	0.712	0.0	0.73	0.0	0.731	0.000
PovertyCount	-0.143	0.0	-1.503	0.0	-1.357	0.0	-1.359	0.000
MedianHHI	-0.174	0.0	-2.521	0.0	-1.81	0.0	-1.834	0.000
spatmax	N/A	N/A	7.351	0.0	N/A	N/A	0.315	0.330
spatmean	N/A	N/A	N/A	N/A	8.237	0.0	7.935	0.000

Table 1. Comparing OLS Model Results

Note: Best Subset (BSS) is an abbreviation for the best subset of features found through our exhaustive feature selection.

some of the other features' p-values increase which suggests that spatmax is able to explain more of the variation in our model than the aspatial subset of features. However, after adding spatmean, the p-values of AA_MALE, TOM_MALE, and Jail_Population are all above the standard critical level of 0.05, indicating that they are no longer significant in our model. Additionally, when considering both of our spatial components, spatmax is considered insignificant as well. This implies that spatmean is even more adept in explaining the variability

in our model than spatmax is. Although we performed an exhaustive selection, our features' statistical significance vary inconsistently throughout our different models. This may be largely due to the multicollinearity that still remains between our features, which can be seen in Table 2 where the Variance Inflation Factor (VIF) is high (>10) for some variables. Therefore, taking into account the limitations of our current data, we may answer our main question by saying: Year, Unemployment_rate, Dispense_rate, NH_MALE, PovertyCount, MedianHHI, and spatmean are the features that are most indicative of overdose rates in the United States.

When assessing which of our models would best estimate our data, we observed that our aspatial model had the best performing AIC (1.910e+04 compared to the spatial models that had all had an AIC around 5.205e+04). However, it also had the lowest R-squared value (0.247), indicating that the covariates used in this model did not explain the variance in our overdose rates as well as our spatial models did, which all had an R-squared value around 0.500.

V Conclusions

From this preliminary round of modeling, our team has been able to establish a foundation of simple models to build upon. We have not only verified that our additional variables are significant in explaining the variation of overdose rates in United States counties but also strongly confirmed that the spatial component of our data will be an essential factor in our model.

Challenges and Future Analysis

Our initial models all utilized an OLS fit and assumed that our features can explain our response variable linearly. To explore the validity of this assumption, we will be conducting further analysis on our residuals and outliers. We will examine the skewness within our data and appropriate account for it by utilizing logs and other data transformations. We also intend to explore geographically weighted regression models and evolve the naive spatial component of our model by utilizing the Queen's contiguity weight matrix which was found in our calculation of the Moran's I. This will allow us to use spatially weighted values, rather than simply using a raw count. Since our data spans a decade, we plan to incorporate spatial-temporal modeling in our analysis as well. Furthermore, to address the remaining multicollinearity discussed above, we will be including more covariates and conducting additional feature selection such as ridge regression or principal component analysis.

Appendix

Figure 1. Number of Observations included in Overdose Dataset by Year

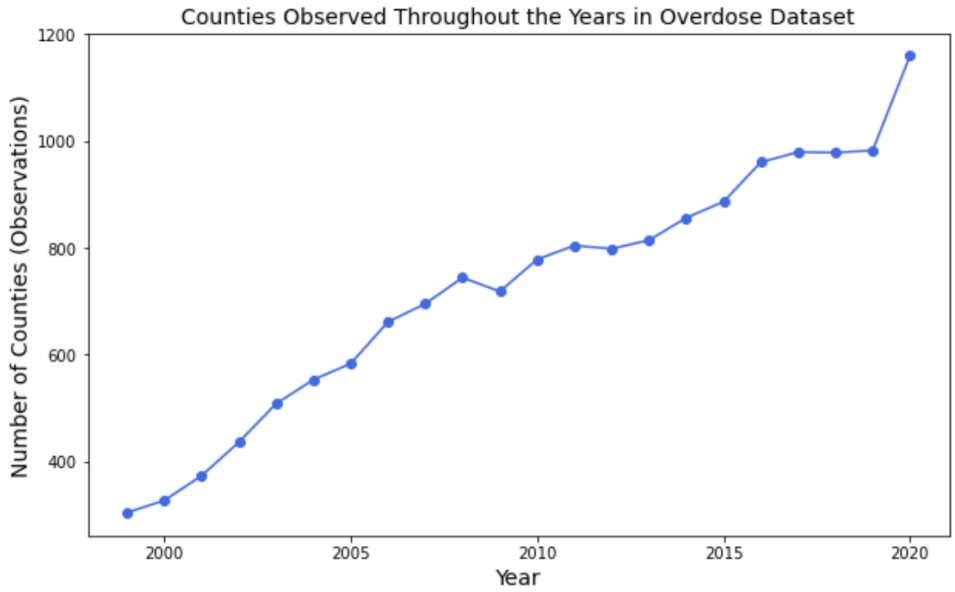
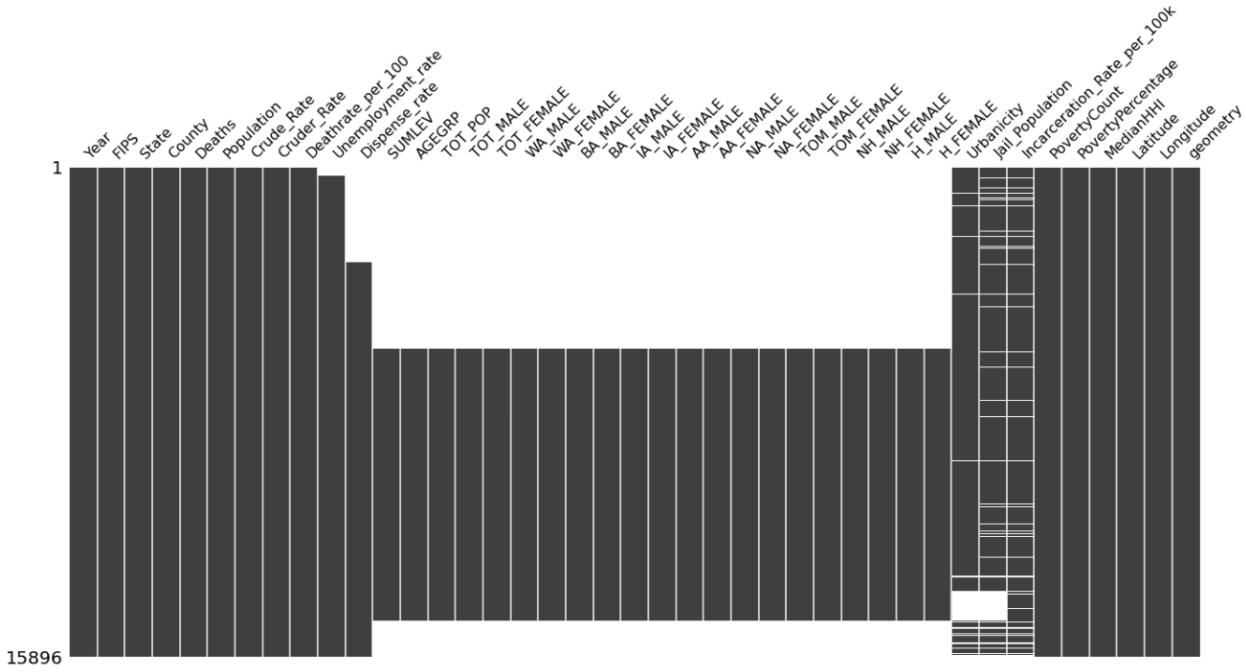


Figure 2. Visualizing Missing Data Throughout All Features



Note: This details the missing values of our data frame after merging our additional covariates with our available overdose data. Black represents the available data. White represents the missing values. 15896 indicates the number of rows in our overall data frame.

Table 2. Variance Inflation Factor (VIF) of Covariates found from Exhaustive Feature Selection

	feature	VIF
0	Year	68.784791
1	Unemployment_rate	8.954940
2	Dispense_rate	10.116277
3	AA_MALE	5.854949
4	TOM_MALE	14.211463
5	NH_MALE	14.546340
6	Jail Population	9.974360
7	Incarceration Rate per 100k	4.811989
8	PovertyCount	13.589621
9	MedianHHI	31.760316

Note: AA_MALE: Number of Males that identify exclusively as Asian American
TOM_MALE: Number of Males that identify with two or more ethnicities
NH_MALE: Number of Males that identify as non-Hispanic
MedianHHI: Median House-Hold Income

References

1. Katz, J. (2017, April 14). You draw it: Just how bad is the drug overdose epidemic? The New York Times. Retrieved March 18, 2022, from https://www.nytimes.com/interactive/2017/04/14/upshot/drug-overdose-epidemic-you-draw-it.html
2. Drug overdose death statistics [2022]: Opioids, fentanyl & more. NCDAS. (2022, February 8). Retrieved March 20, 2022, from https://drugabusestatistics.org/drug-overdose-deaths/
3. Centers for Disease Control and Prevention, National Center for Health Statistics. Multiple Cause of Death, 1999-2020 on CDC WONDER Online Database, released in 2021. Data are from the Multiple Cause of Death Files, 1999-2020, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics

Cooperative Program. Accessed at http://wonder.cdc.gov/mcd-icd10.html on Feb 16, 2022 1:22:58 AM

4. Centers for Disease Control and Prevention. (2021, November 10). U.S. opioid dispensing rate maps. Centers for Disease Control and Prevention. Retrieved March 18, 2022, from https://www.cdc.gov/drugoverdose/rxrate-maps/index.html
5. U.S. Bureau of Labor Statistics. (n.d.). Local Area Unemployment Statistics. U.S. Bureau of Labor Statistics. Retrieved March 18, 2022, from https://www.bls.gov/lau/#tables
6. U.S. Census Bureau. (2021, October 8). County population by characteristics: 2010-2019. Census.gov. Retrieved March 18, 2022, from https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html
7. U.S. Census Bureau. (2021, October 8). Small Area Income and Poverty Estimates (SAIPE) Program Datasets. Census.gov. Retrieved March 18, 2022, from https://www.census.gov/programs-surveys/saipe/data/datasets.html
8. Institute, V. (n.d.). Vera-Institute/Incarceration-Trends: Incarceration trends dataset and Documentation. GitHub. Retrieved March 18, 2022, from https://github.com/vera-institute/incarceration-trends
9. Rey, Arribas-Bel, Wolf (2020) Geographic Data Science with Python. Retrieved at https://geographicdata.science/book/intro.html.