Stat 222 Project: Geospatial Modeling of Drug Overdose in the United States

Final Report

Written By Wei Deng

Group Members:
Meera Duggal
Ashlyn Jew
Rhiann Zhang

May 11, 2022

I Exposition

The primary goal of our project is to assess the relationship between aspects of counties in the United States (including demographics, location, incarceration rates, and opioid dispense rates) and their corresponding drug overdose rates using publicly available county-level data. We also added a geospatial component to our model, estimating a county's drug overdose rate based on surrounding counties. Using estimates from this model, we iteratively fill in adjacent missing counties.

With this project, we hope to inform public policy decisions related to drug overdoses at a national level, appealing to law makers and government agencies about the factors that most contribute to drug overdoses. Additionally, we hope to form a robust model that estimates drug overdose rates for counties where the data is not well maintained. Moreover, the model could eventually be transitioned into one that predicts future drug overdose rates.

II Data

i Data Sources

The data we are using are mainly collected and maintained by divisions of the United States government. These include the Center for Disease Control and Prevention [1][2], U.S. Census Bureau [3], and the U.S. Bureau of Labor Statistics [4]. Other data (such as incarceration rates) were collected by independent organizations such as the Vera Institute of Justice [5] and County Health Rankings [7].

ii Data Description

The outcome variable of interest in our data is the drug overdose rate. We labeled this as "Cruder Rate" in our data set, which is a name derived from the "Crude Rate" provided by the CDC. Since for many of the counties, this rate was left blank or labeled "Unreliable", but still retained the raw numbers, we estimated the rate as the number of overdoses per 100,000 people.

The independent variables we collected and merged to our data set include population by gender and age, poverty rate, urbanicity, opioid dispense rate, incarceration rate, and various health statistics like percent of uninsured population and number of primary care physicians per 100k people. All of our data are at the county level, over recent years (mainly 2011–2020). All together, there are approximately 31000 observations in our data set, each with about 44 features.

From Figure 1, we see that we have a lot of missing data for the earlier years, as well as some for the most recent years. As such, we decided to focus our analysis on the years in which we have all of the data present, 2010–2019.

Although we are missing drug overdose rates for many counties for the years of interest, as shown by Figure 2, the counties for which we have data account for more than 60% of the total U.S. population. As such, we can still make some valid inferences about drug overdose rates. We also see that there is a concentration of high drug overdose rates in the mid-west/south area of the country, which suggests that there might be a geospatial relationship for drug overdose rates (i.e. a county surrounded by others with high drug overdose rates has a high drug overdose rate itself).

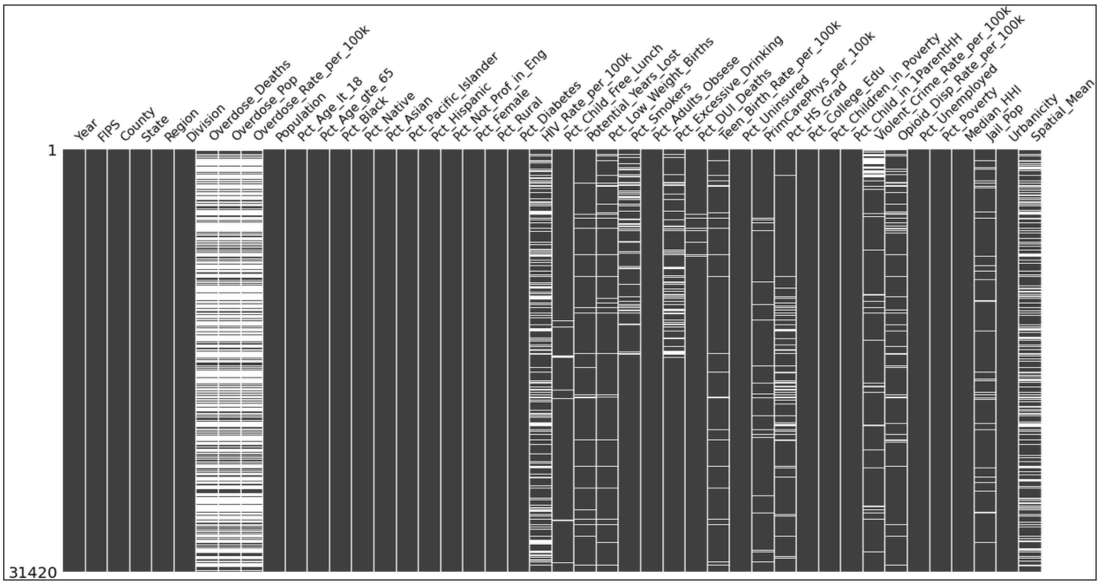


Figure 1: For each variable, whether we have data or not, sorted by year (top is 2011)

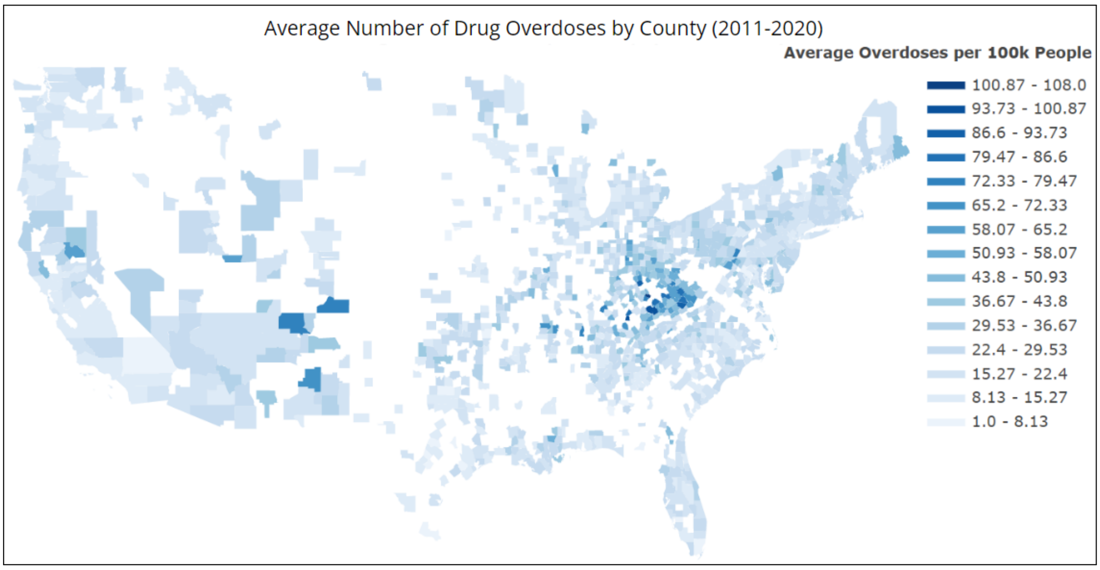


Figure 2: Our dependent variable, drug overdose rate, is missing in many counties

Based on intuition and popularly purported reasons for drug overdose, we decided to focus on county-level variables such as unemployment rate, opioid dispense rate, incarceration rate, and other health-related statistics. In Appendix 1, we compare the the average drug overdose rates in the counties of West Virginia (the state with the highest drug overdose rate by far) with the average values for unemployment, opioid dispensary, and incarceration. Additionally, we see that there appears to be clusters of higher and lower drug overdose rates among the counties, suggesting that there may be a geospatial component to a county's drug overdose rate. We also see that it appears incarceration rate is more correlative of a factor to drug overdose rate compared to unemployment rate and opioid dispense rate. Lastly, these plots illustrate the amount of missing values in some counties in our data set.

Before doing any numeric analysis, we standardized all of our independent variables to put the variables on a similar scale and lessen the impact of outliers.

III Methods

i Part 1 - Aspatial Modeling and Initial Variables

In order to establish a baseline model, we first fitted the model without any consideration of geospatial relationships. This allows us to compare this initial model to models in the future with geospatial components to see if the geospatial components have any significant exploratory power. We also calculate Moran's I for an initial assessment of geospatial relationships in our data to motivate future models. As a note, the analysis done in this part considers only our initial variables without those later added from the County Health Rankings [7] data.

i.1 Aspatial Ordinary Least Squares Regression

Without taking into account the spatial relationship of our data, we fit ordinary least squares regression models of drug overdose rate on subsets of our independent variables. We wished to start with a simple, interpretable model that we could use a baseline with which to compare more complex models later. In Appendix 2, we see the output of the OLS regression of drug overdose rate ("Cruder Rate") on year, number of white males, urbanicity (ordinal variable from 1-4, 4 being urban), unemployment rate, dispense rate, incarceration rate, poverty percentage, and median household income. These are variables that we picked manually, based on our intuition of what factors could have a strong relationship with drug overdose rates based on initial exploratory data analysis. The general OLS regression formula is given by:

$$C_{it} = \alpha + \sum_{k=1}^{p} \beta_k X_{ikt} + \epsilon_{it}$$

where C is the estimated drug overdose rate for the i^{th} observation and t^{th} year, α is the intercept, p is the number of features in our model, β_k is the coefficient associated with the k^{th} feature, X is the set of features, and ϵ is the error term. We used this model primarily as an initial baseline onto which we could build with more predictive features and a geospatial component as well.

i.2 Best Subset Selection

Because we have so many different combinations of features to consider in our OLS regression model, we implemented an exhaustive search for the best subset of features. The metric we used to compare between the different models was AIC. The model with the best performing AIC was the one with the following subset of features: year, unemployment rate, opioid dispense rate, African American male, total male, non-Hispanic male, jail population, incarceration rate, poverty county, and median household income. The results of the OLS regression on this subset of features can be seen in Appendix 3.

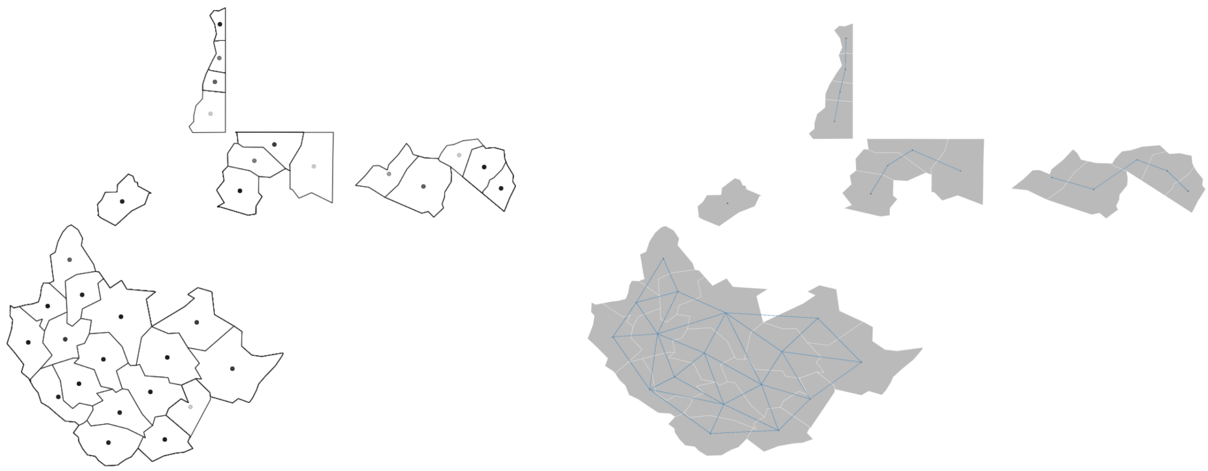
Looking at the chosen subset of features, there seems to be obvious multicollinearity. This is confirmed by looking at the respective VIF (variance inflation factor) scores for each of the features. In the future, we will research methods to reduce our model's multicollinearity. Such methods could include ridge/LASSO regression and principal component analysis.

i.3 Moran's I

In order to determine whether or not there is a spatial relationship between the drug overdose rates of counties, we calculated Moran's I, a measure of spatial autocorrelation, similar to Pearson's correlation coefficient. Moran's I ranges from -1 to 1, with 0 indicating no spatial relationship, 1 indicating positive spatial relationship (counties with similar drug overdose rates are close to each other) and -1 indicating negative spatial relationship (counties with similar drug overdose rates are far from each other). The general formula for Moran's I is given by:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_i z_i^2}$$

where n is the number of observations, z_i is the standardized value of the variable of interest $(x_i - \bar{X})$ at location i, and w_{ij} is the spatial weight between locations i and j (the spatial relationship between points i and j) [6]. With our implementation of Moran's I, we chose to use Queen's weights, meaning that a spatial relationship can be established by one county with all the other counties adjacent to it. We did this by using the GeoPandas library, which allowed us to construct polygons to represent the county boundaries. Then, we could compute a centroid for each polygon (county) that represented the middle. Next, we could use the Queen's weight method to calculate spatial weights for our Moran's I calculation. A visual representation of this process can be seen with West Virginia in Figure 3. Calculating the global Moran's I for all of our available counties, we obtained a value of 0.46. This indicates that there is a moderate positive autocorrelation between county proximity and drug overdose rate (i.e. a county being close to another correlates with drug overdose rate). In Appendix 4, we can see, on a local level, what Moran's I means for West Virginia and how the magnitude of drug overdose rates tend to cluster, leading to higher autocorrelation.



(a) Centroids

(b) Queen's weights

Figure 3: West Virginia Counties

ii Part 2 - Geospatial Modeling and Added Health Variables

In this part, we start to consider geospatial variables in our modeling, given that we have a Moran's I metric that indicates a positive moderate geospatial relationship between drug overdose rates in the United States. We also add more variables to our data from the County Health Rankings [7] source. Some key county-level features added include percent of uninsured individuals and number of primary care physicians per 100k people.

ii.1 Baseline Geospatial OLS Regression Model

In order to assess the effect of a spatial component in our model, we fit an OLS regression model with only a spatial component: spatial mean of drug overdose rates. By looking at certain metrics like adjusted R^2, we can get a sense of how much variance is being explained by just the spatial mean variable.

The spatial mean variable for a specific county is calculated by taking the average drug overdose rate of counties adjacent to it. For instance, if we were to assign the spatial mean variable of Los Angeles county, we would compute the average drug overdose rate of its adjacent counties: Kern county, Ventura county, San Bernardino county, and Orange county.

As seen in Appendix 5, the output from the OLS regression with just the spatial mean variable, we have an adjusted R^2 value of 0.451, which suggests that the spatial mean component already explains a lot of the variation in the data.

ii.2 Backward Stepwise Feature Selection

After confirming that spatial mean is a variable that explains a significant amount of variance in our data, we perform a backward stepwise selection algorithm to determine other features in our data to add to the model that would best explain variance without overfitting. Since we observed from Part 1 that there is significant multicollinearity in our covariates, we decide to use a backward selection method instead of a forward selection method, which is more sensitive to selecting multicollinear variables.

Using 5-fold cross-validation with root mean squared error (RMSE) as our metric in the backward stepwise selection, we find that the best fitting model is one with eight features:

- `Spatial_Mean` - average overdose rate of adjacent counties
- `PrimCarePhys_per_100k` - number of primary care physicians per 100k residents
- `Pct_Uninsured` - percent of uninsured residents residents
- `Pct_Child_in_1ParentHH` - percent of children in 1 parent households
- `Pct_Poverty` - percent of residents in poverty
- `Pct_Black` - percent of Black residents
- `Pct_Age_lt_18` - percent of residents that are less than 18 years old
- `Potential_Years_Lost` - years of potential life lost before age 75 per 100k residents

The output for the model fitted with the backward selected variables are shown in Appendix 6. With the spatial mean variable with the additional 7 selected variables, we are able to achieve an adjusted R^2 value of 0.612.

ii.3 Feature Transformation

After plotting the residuals from the model against each of the variables included as predictors, we observed some instances where the residuals were not randomly dispersed about the horizontal axis, instead they are clustered toward a certain region of the plot.

As such, we see the need to apply data transformations for these variables. The variables we transform are: `Pct_Black`, `PrimCarePhys_per_100k`, and `Spatial_Mean`. As seen in Figure 4, we apply a log transformation to each of these variables.

After log transformation, we see that these specific variables have residuals that are dispersed much more randomly and evenly around the horizontal axis. For `Pct_Black` specifically, since so many of the observations are close to 0, we decide to add 1 to the observations before log transforming.

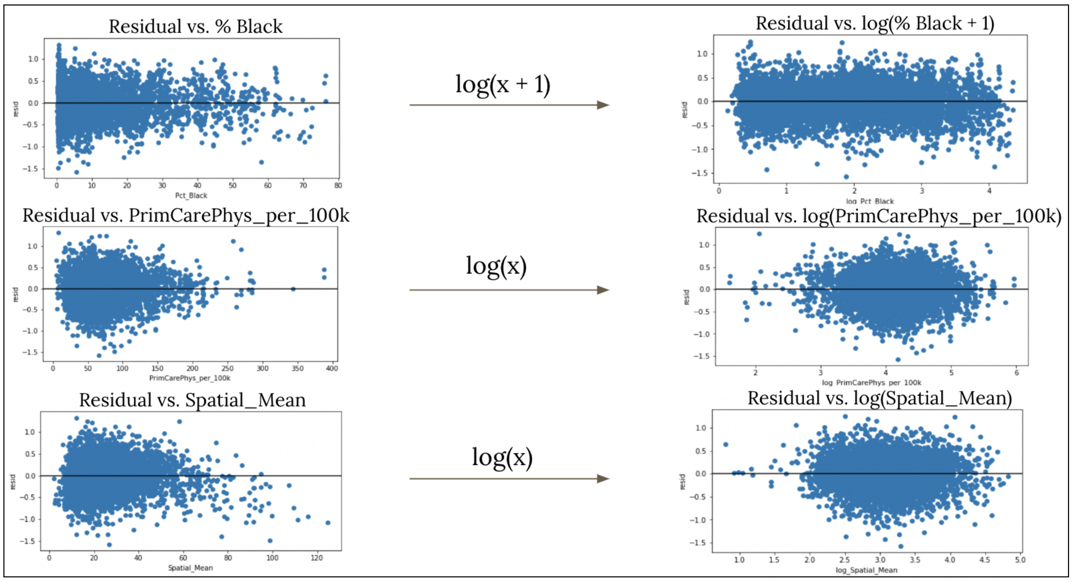


Figure 4: Residual plots before and after log transformation

IV Results

i Initial OLS Regression

For the particular regression shown in Appendix 2 (Aspatial OLS with manually selected features), we see a low p-value for the F-statistic, meaning that it is statistically significant (under 0.05 level of significance) that our model explains more of the variation in drug overdose rate than a model with no features would. Taking a look at the significance of individual features, we see that all of our features seem to be statistically significant except for urbanicity, incarceration rate, and poverty percentage.

After running the regression again with the best selected feature with the output in Appendix 3, we see that under a significance level of 0.05, all of our features are statistically significant. Interestingly, we see some of the same features from our first OLS regression that weren't significant become significant in this new model.

ii Moran's I

In Appendix 4, we can take a closer look at Moran's I at a local level. Looking at the scatterplot specifically, we recognize that the plot is sectioned into four quadrants: high-high (HH), high-low (HL), low-high (LH), and low-low (LL). The number of points in each of these quadrants represents the spatial relationship pairings of each of the counties of West Virginia, and the slope drawn as the best fit line for these points is Moran's I. We see that the slope is positive, meaning that county proximity is positively correlative with drug overdose rates. We can see this illustrated on the two subsequent plots, where the highest concentration of drug overdose rates correspond with the most spatial correlation.

iii Final Model

The OLS regression output of the final model can be seen in Appendix 7. Here, after selecting the best variables with backward selection and transforming some problematic features, we get an increase in explanatory power with an adjusted R^2 value of 0.626 (an increased value from the model with no transformations). We see that, with the very low p-values, all of the selected features are statistically significant at the 0.01 level.

With this final model, we can start to fill in the missing drug overdose rates in our data, starting with those counties adjacent to ones for which we have the complete data.

iv Iterative Estimation

Putting the final model into action, we predict drug overdose rates for counties with missing data for a specific year (using the data of that same year). After fitting the model, the iterative process is as follows:

1. Use the OLS regression model to estimate the drug overdose rate of counties for which we have missing data
2. Compute the spatial mean of drug overdose rate for counties adjacent to at least one other county with known drug overdose rate
3. Repeat steps 1 and 2 until we cannot estimate any more values due to unknown covariates

In other words, for counties whose drug overdose rate is not estimated in the first iteration of the process, their estimated drug overdose rate will be an estimate of estimates. This is due to the new computation of spatial mean (dependent on the estimate of drug overdose rate) with each new iteration.

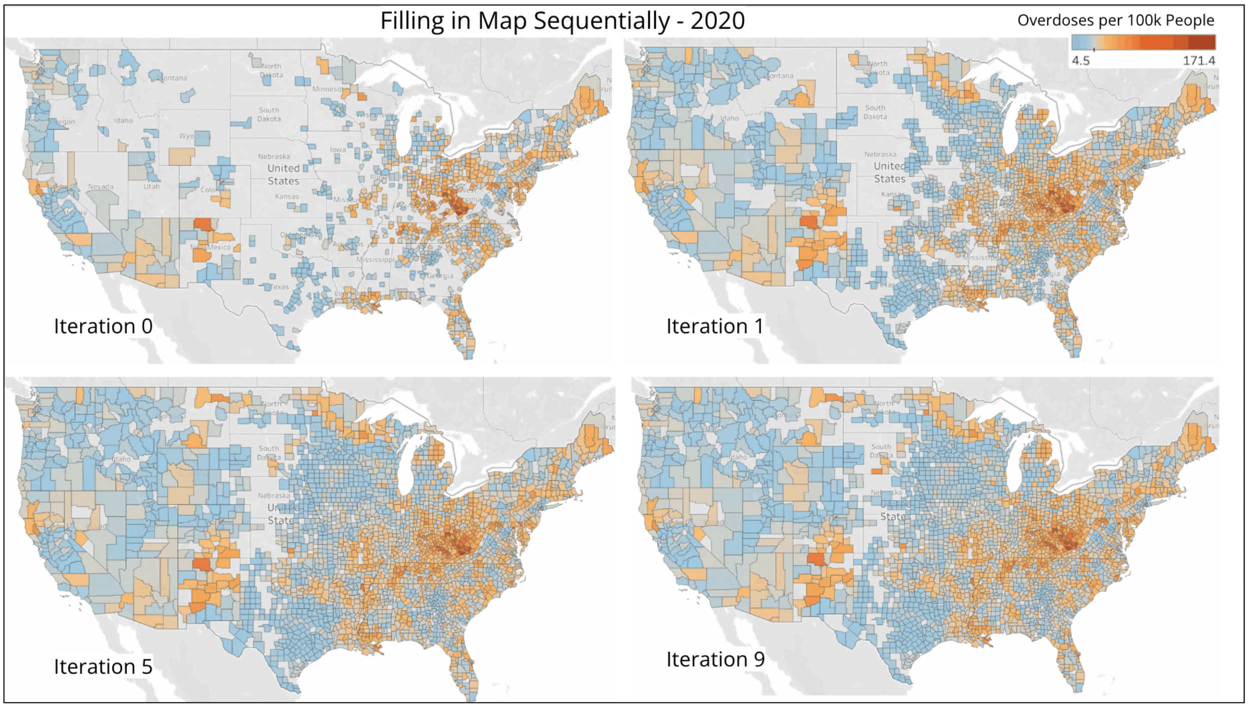


Figure 5: Series of drug overdose rates mapped at the county level, with 4 different iterations

In Figure 5, we see a series of maps indicating which iteration of the process the estimation is at for the year 2020. At iteration 0, we have only the data we collected (and built the model with). With each successive iteration, adjacent counties with missing data are being filled in with new estimated drug overdose rates. Since the geospatial component of the model has such a large impact on the estimate, we see that areas of the map where there are high overdose rates in general will fill in surrounding counties with similarly high drug overdose rate estimates.

At iteration 9, we get to a point where we cannot fill in any more missing values with our fitted model. This is due to the lack of data for certain covariates in the model for those counties.

V Conclusion

Starting at the beginning with a simple, aspatial model for estimating drug overdose rate, we are able to build on the predictive power by introducing a geospatial component and selecting the features with best explanatory power. As a result, the adjusted R^2 value of the first and last models are 0.24 and 0.626, respectively. With each successive method introduced, we see a marked increase in explanatory power in our model.

With the iterative estimation process we introduced, we end with a map of the United States's drug overdose rates with a few missing counties. This map's filled-in drug overdose rates are comparable to those estimated by other researchers.

With this project, we inform the reader about the rising drug epidemic in the United States, specifically which areas are most impacted. Additionally, we detail which aspects of counties are most significantly associated with drug overdose rates in the United States and hope that policy makers can use this information to know what aspects of their counties to address when attempting to combat rising drug overdose rates in the United States.

VI Appendix

Appendix 1: Average values for West Virginia (2010–2019)

Dep. Variable:	Cruder_Rate	**R-squared:**	0.240
Model:	OLS	**Adj. R-squared:**	0.240
Method:	Least Squares	**F-statistic:**	299.8
Date:	Sun, 20 Mar 2022	**Prob (F-statistic):**	0.00
Time:	05:11:24	**Log-Likelihood:**	-9572.3
No. Observations:	7591	**AIC:**	1.916e+04
Df Residuals:	7582	**BIC:**	1.923e+04
Df Model:	8		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-360.5609	11.472	-31.430	0.000	-383.049	-338.073
Year	0.1790	0.006	31.439	0.000	0.168	0.190
WA_MALE	-0.0450	0.011	-4.009	0.000	-0.067	-0.023
Urbanicity	-0.0230	0.013	-1.815	0.069	-0.048	0.002
Unemployment_rate	0.1494	0.015	10.241	0.000	0.121	0.178
Dispense_rate	0.3134	0.014	23.191	0.000	0.287	0.340
Incarceration_Rate_per_100k	0.0088	0.011	0.830	0.406	-0.012	0.029
PovertyPercentage	-0.0275	0.018	-1.572	0.116	-0.062	0.007
MedianHHI	-0.1330	0.020	-6.772	0.000	-0.171	-0.094

Omnibus:	3914.240	**Durbin-Watson:**	1.362
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	41611.534
Skew:	2.240	**Prob(JB):**	0.00
Kurtosis:	13.559	**Cond. No.**	2.36e+06

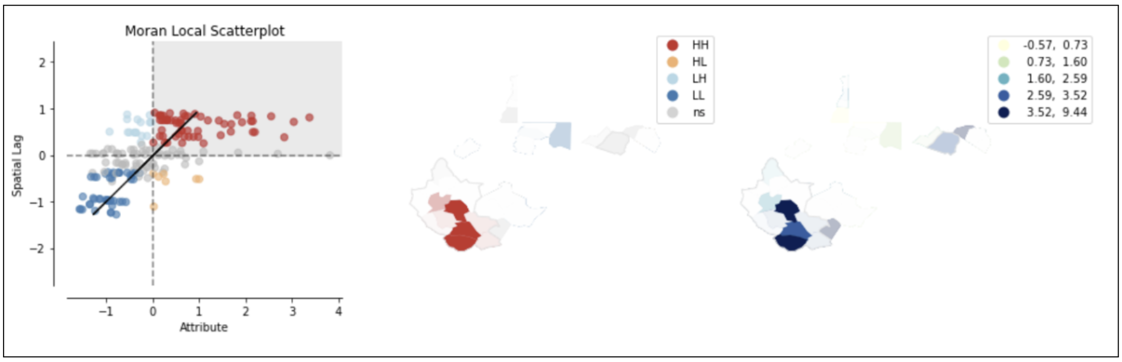
Appendix 2: OLS output of Drug Overdose Rate on manually selected features

Dep. Variable:	Cruder_Rate	**R-squared:**	0.247
Model:	OLS	**Adj. R-squared:**	0.246
Method:	Least Squares	**F-statistic:**	249.0
Date:	Sun, 20 Mar 2022	**Prob (F-statistic):**	0.00
Time:	06:11:30	**Log-Likelihood:**	-9537.4
No. Observations:	7591	**AIC:**	1.910e+04
Df Residuals:	7580	**BIC:**	1.917e+04
Df Model:	10		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P> |t|**	**[0.025**	**0.975]**
Intercept	-367.9896	11.136	-33.045	0.000	-389.819	-346.160
Year	0.1827	0.006	33.057	0.000	0.172	0.194
Unemployment_rate	0.1521	0.014	10.784	0.000	0.124	0.180
Dispense_rate	0.2985	0.014	21.930	0.000	0.272	0.325
AA_MALE	0.0982	0.024	4.073	0.000	0.051	0.145
TOM_MALE	-0.1459	0.037	-3.920	0.000	-0.219	-0.073
NH_MALE	0.2223	0.032	7.024	0.000	0.160	0.284
Jail_Population	-0.0997	0.031	-3.231	0.001	-0.160	-0.039
Incarceration_Rate_per_100k	0.0231	0.011	2.017	0.044	0.001	0.046
PovertyCount	-0.1429	0.033	-4.313	0.000	-0.208	-0.078
MedianHHI	-0.1742	0.015	-11.574	0.000	-0.204	-0.145

Omnibus:	3866.583	**Durbin-Watson:**	1.362
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	40325.914
Skew:	2.211	**Prob(JB):**	0.00
Kurtosis:	13.390	**Cond. No.**	2.30e+06

Appendix 3: OLS output of Drug Overdose Rate on best subset selected features



Appendix 4: Local Moran's I, West Virginia

Spatial lag scatterplot (middle), spatial correlation (middle), drug overdose rate (right)

Dep. Variable:	log_Overdose_Rate_per_100k	**R-squared:**	0.451
Model:	OLS	**Adj. R-squared:**	0.451
Method:	Least Squares	**F-statistic:**	5298.
Date:	Tue, 10 May 2022	**Prob (F-statistic):**	0.00
Time:	17:31:31	**Log-Likelihood:**	-3459.3
No. Observations:	6450	**AIC:**	6923.
Df Residuals:	6448	**BIC:**	6936.
Df Model:	1		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.3415	0.011	212.993	0.000	2.320	2.363
Spatial_Mean	0.0288	0.000	72.791	0.000	0.028	0.030

Omnibus:	48.031	**Durbin-Watson:**	1.987
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	66.653
Skew:	-0.089	**Prob(JB):**	3.36e-15
Kurtosis:	3.465	**Cond. No.**	59.3

Appendix 5: OLS output of Drug Overdose Rate on only Spatial Mean variable

Dep. Variable:	Overdose_Rate_per_100k	**R-squared:**	0.612
Model:	OLS	**Adj. R-squared:**	0.612
Method:	Least Squares	**F-statistic:**	1212.
Date:	Tue, 10 May 2022	**Prob (F-statistic):**	0.00
Time:	17:57:39	**Log-Likelihood:**	-2190.6
No. Observations:	6152	**AIC:**	4399.
Df Residuals:	6143	**BIC:**	4460.
Df Model:	8		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	2.5129	0.053	47.675	0.000	2.410	2.616
Pct_Age_lt_18	-0.0230	0.002	-14.004	0.000	-0.026	-0.020
Pct_Black	-0.0068	0.001	-13.385	0.000	-0.008	-0.006
Potential_Years_Lost	0.0001	3.34e-06	30.084	0.000	9.38e-05	0.000
Pct_Uninsured	-0.0126	0.001	-12.688	0.000	-0.015	-0.011
PrimCarePhys_per_100k	-0.0013	0.000	-10.312	0.000	-0.002	-0.001
Pct_Child_in_1ParentHH	0.0121	0.001	12.552	0.000	0.010	0.014
Pct_Poverty	-0.0149	0.001	-10.462	0.000	-0.018	-0.012
Spatial_Mean	0.0182	0.000	42.733	0.000	0.017	0.019

Omnibus:	76.309	**Durbin-Watson:**	1.988
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	93.320
Skew:	-0.200	**Prob(JB):**	5.44e-21
Kurtosis:	3.451	**Cond. No.**	9.47e+04

Appendix 6: OLS output of Drug Overdose Rate on selected features from backward stepwise selection algorithm

Dep. Variable:	Overdose_Rate_per_100k	**R-squared:**	0.627
Model:	OLS	**Adj. R-squared:**	0.626
Method:	Least Squares	**F-statistic:**	1289.
Date:	Wed, 11 May 2022	**Prob (F-statistic):**	0.00
Time:	01:53:27	**Log-Likelihood:**	-2074.1
No. Observations:	6152	**AIC:**	4166.
Df Residuals:	6143	**BIC:**	4227.
Df Model:	8		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
Intercept	1.6449	0.079	20.789	0.000	1.490	1.800
Pct_Age_lt_18	-0.0206	0.002	-12.856	0.000	-0.024	-0.017
log_Pct_Black	-0.0579	0.005	-10.546	0.000	-0.069	-0.047
Potential_Years_Lost	9.73e-05	3.26e-06	29.806	0.000	9.09e-05	0.000
Pct_Uninsured	-0.0106	0.001	-10.859	0.000	-0.013	-0.009
log_PrimCarePhys_per_100k	-0.0906	0.009	-9.638	0.000	-0.109	-0.072
Pct_Child_in_1ParentHH	0.0079	0.001	8.960	0.000	0.006	0.010
Pct_Poverty	-0.0113	0.001	-8.081	0.000	-0.014	-0.009
log_Spatial_Mean	0.5414	0.012	47.015	0.000	0.519	0.564

Omnibus:	64.709	**Durbin-Watson:**	1.987
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	78.089
Skew:	-0.182	**Prob(JB):**	1.10e-17
Kurtosis:	3.415	**Cond. No.**	1.46e+05

Appendix 7: OLS output of Drug Overdose Rate on selected features after transformation

References

[1] Centers for Disease Control and Prevention, National Center for Health Statistics (2021) *Multiple Cause of Death*. 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.

[2] Centers for Disease Control and Prevention (2021) *U.S. opioid dispensing rate maps*. https://www.cdc.gov/drugoverdose/rxrate-maps/index.html.

[3] U.S. Census Bureau (2021) *County population by characteristics: 2010-2019*. https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html.

[4] U.S. Bureau of Labor Statistics (2021) *Local Area Unemployment Statistics. U.S. Bureau of Labor Statistics*. Retrieved at https://www.bls.gov/lau/#tables.

[5] Institute, Vera (2021) *Vera-Institute/Incarceration-Trends: Incarceration trends dataset and Documentation*. Github. Retrieved at https://github.com/vera-institute/incarceration-trends.

[6] Rey, Arribas-Bel, Wolf (2020) *Geographic Data Science with Python*. Retrieved at https://geographicdata.science/book/intro.html.

[7] County Health Rankings (2020). Retrieved at https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019