Ashlyn Jew
STAT 222

Drug Overdoses in the United States

1 Problem Overview

1.1 Motivation

The drug epidemic in the United States has been a growing concern in the past years. We are currently at the peak of drug overdoses where the number of deaths caused by drug poisoning tower over the number deaths by car accidents, guns, and HIV [1]. To understand this devastating and widespread problem, we will analyze the potential factors contributing to the drug epidemic.

1.2 Problem Statement and Goals

Through this project, we aim to answer the following question: "Which characteristics of U.S. counties can best explain drug overdose rates?" By determining these characteristics, we will be able to estimate drug overdose rates and grant policy makers insight as to what factors to focus on as well as which communities should be addressed more urgently when trying to counteract the drug overdose epidemic.

2 Data Description

2.1 Original Data

The county-level drug overdose mortality data was extracted from the Multiple Cause of Death (Final) Database on CDC Wonder. The database contains data collected by the National Center for Health Statistics and is based on information on death certificates filed in the United States [2]. We filtered for data with death codes defined by the CDC as death by drug poisoning.

Additional data on health factors [4] (e.g. % smokers, % uninsured, # of primary care physicians), opioid dispensing rates [3], demographics (e.g. % college graduates, # of children in single parent households) [4], unemployment rates [5], jail population sizes [6], poverty rates, and median household income [7] was retrieved by our team. We also added geospatial variables to our data: average overdose death rates of adjacent counties for each year and geometry data from county shapefiles.

2.2 Data Description

For this report, we are focusing on data from 2011 to 2020 since our data is more complete for those years (Figure 1). We have 31420 observations and 44 variables in our 2011-2020 dataset. Each observation represents a county in a certain year. Each variable is also at a county-level granularity.

Our dataset is missing data for many counties each year. However, the counties present in the data represent over 50% of the total U.S. population. As mentioned above, we aim to fill in the missing data with our explanatory model.

2.3 Variables of Interest

Since all the data and variables retrieved were self-curated, all variables are considered important potential factors contributing to the overdose death rates, albeit we expect our model to tell us which variables are more important than others. Table 1 details a list of our variables of interest and their hypothesized relationship to drug overdose death rates.

In our initial EDA, we have noticed that there were clusters of higher overdose death rates in certain areas of the U.S. In Figure 2 below, West Virginia and the surrounding Appalachian region seem to have a cluster of higher overdose death rates. Thus, we hypothesize that there are spatial autocorrelations and spillover effects between counties. So we have included Spatial_Max (maximum drug overdose death rate of adjacent counties) as a variable of interest.

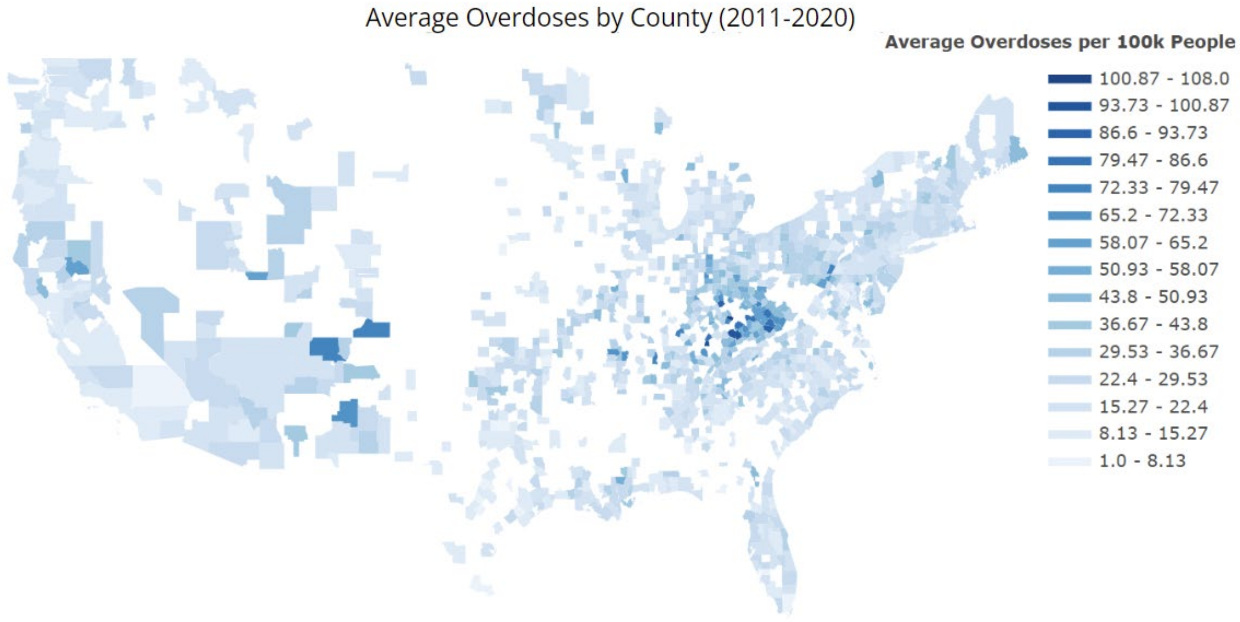


Figure 2: Average Overdoses per 100k People by County (2011 - 2020)

3 Methods

3.1 Geospatial Components

To support our idea of adding spatial components to our model, we calculated Moran's I to determine if there was a spatial autocorrelation between the U.S. counties. Moran's I is computed as follows:

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

where $z_i = (x_i - \bar{x})$ is the deviation of an attribute (i.e. our variables) for county i from its mean and $w_{i,j}$ is the spatial weight between county i and j, and n is the total number of counties [8].

The spatial weight was calculated using the notion of Queen contiguity, which determines weights depending on if they share a common vertex, and hence considered neighboring [8]. The Moran's I we computed was approximately 0.46085, which indicates that there is a positive spatial autocorrelation. Thus, we have evidence that there is some sort of spatial clustering of our variables.

Next, we look at the local Moran's I for West Virginia, which is one of the states with the most obvious clustering of drug overdose deaths. In the Moran Local Scatterplot (Figure 3 left plot), the upper right and lower left quadrants indicate a positive spatial autocorrelation (i.e. similar cluster of counties) for drug overdose death rates. The lower right and upper left quadrants indicate negative spatial autocorrelation (i.e. counties are different from their neighbors) for drug overdose death rates. The slope indicates the local Moran's I for West Virginia. Since we see more points in the upper right and lower left quadrants, there is likely stronger positive spatial autocorrelation in West Virginia. In the right plot, the red indicates High-High local autocorrelation, where a county with higher drug overdose death rate has neighbors that also have higher drug overdose death rates.

With this spatial autocorrelation analysis, we see that there is evidence of positive spatial correlation of drug overdose rates in our counties. Thus, we will be including a spatial component within our model.

3.2 Statistical Model

We fit OLS models with the following formula:

$$y_{it} = \beta_0 + \sum_{k=1}^{p} \beta_k x_{itk} + \epsilon_{it}$$

where y_{it} is the drug overdose death rate at county i for year t, β_0 is the global intercept coefficient, x_{itk} is the k-th explanatory variable at county i for year t, β_k is the global k-th local regression coefficient for the k-th explanatory variable, ϵ_{it} is the random error term associated with county i for year t, and i goes from 1 to n (number of observed counties) [9]. Model coefficients were computed using all data throughout all years, hence they are considered global variables. However, when estimating our overdose rates, we input data for each county subsetted by year.

To begin, we fit an OLS regression model with logged drug overdose death rate ('log_Overdose_Rate_per_100k') as the dependent variable and 'Spatial_Mean' as the independent variable. We transformed the drug overdose death rate because it was right-skewed, and logging it made it more normally distributed (Figure 4). We used this baseline model to see how much variability our geospatial component explained on its own. As we can see from the detailed results of the baseline model in Table 2, the adjusted R^2 is approximately 0.451. This indicates that the model already explains a considerable amount of the variation in our data. Using a train-test split, we obtained a normalized RMSE of 0.0938 (normalized RMSE ranges from 0 to 1, with values closer to 0 being a better fit), which suggests that this baseline model is already a pretty good fit.

Now that we have created our baseline model, we utilized backwards stepwise feature selection to determine which of our features can most explain drug overdose death rates. We chose to use backwards feature selection instead of forward feature selection because we know that our features are collinear, and backwards selection is better equipped to reduce multicollinear features [10].

4 Results

4.1 Backwards Selection Model Output

Using 5-fold cross validation and evaluating the training and testing RMSE to determine the best model, we found that the best model was as follows:

Overdose_Rate_per_100k ~

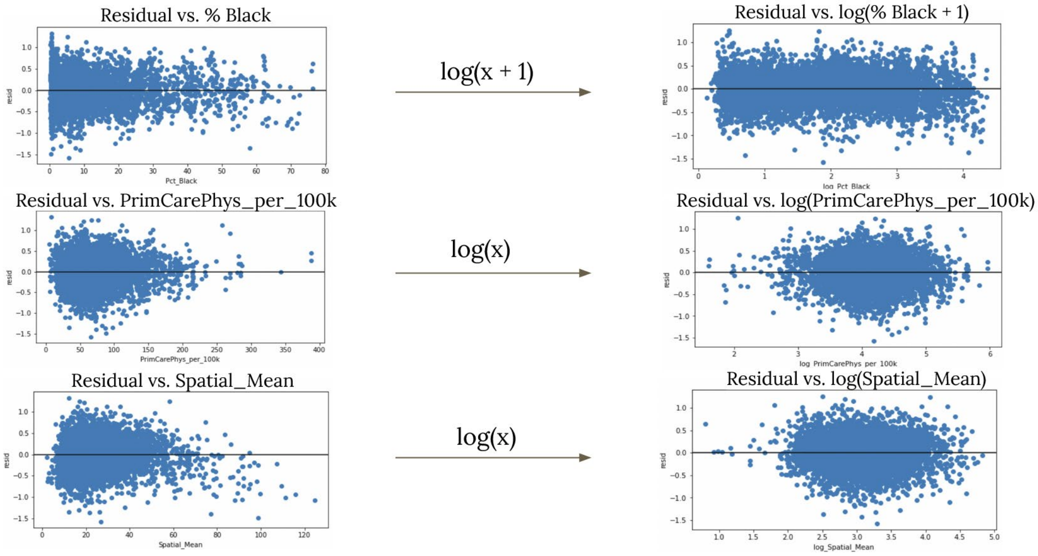
- Pct_Age_lt_18 + (% of residents that are less than 18 years old)
- Pct_Black + (% of Black residents)
- Potential_Years_Lost + (Potential yrs lost before age 75 per 100k residents)
- Pct_Uninsured + (% of uninsured residents)
- PrimCarePhys_per_100k + (# of primary care physicians per 100k residents)
- Pct_Child_in_1ParentHH + (% of children in 1 parent households)
- Pct_Poverty + (% of residents in poverty)
- Spatial_Mean (Average overdose rate of adjacent counties)

The detailed OLS output table is shown in Table 3.

The model's adjusted R^2 improved from the baseline model's 0.46 to 0.613 and the AIC decreased from 6125 to 4221. This indicates that the additional features are a good addition to our model when comparing it to the baseline model.

4.2 Transformations

After fitting this model, we analyzed the residual vs. predictor plots to determine any variable transformations that we had to make.


Residual vs. % Black

Residual vs. log(% Black + 1)

Residual vs. PrimCarePhys_per_100k

Residual vs. log(PrimCarePhys_per_100k)

Residual vs. Spatial_Mean

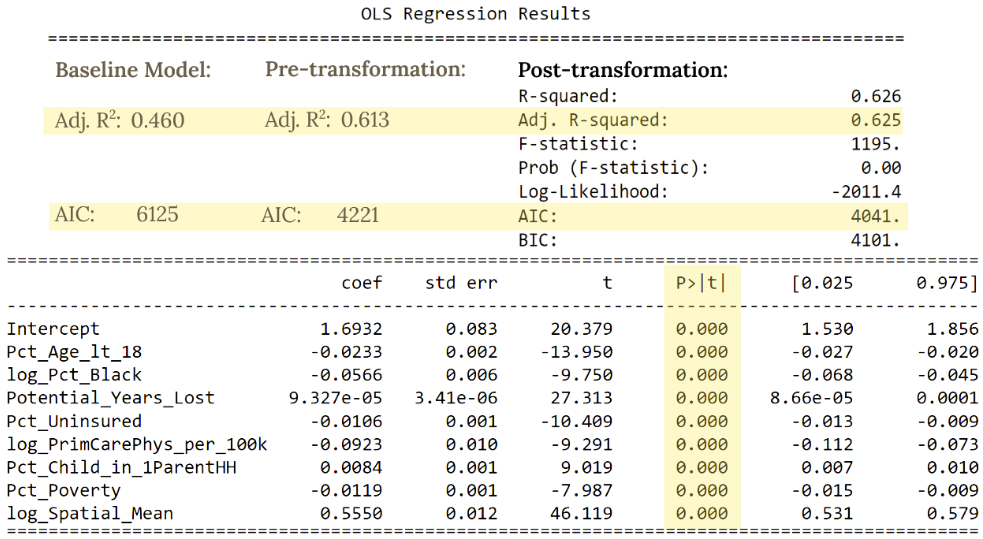
Residual vs. log(Spatial_Mean)

Figure 5: Residual Plot Transformations

Looking at the residual plots, we noted that the residuals were clustered to the left, so we decided to log-transform the ‘Pct_Black’, ‘PrimCarePhys_per_100k’, and ‘Spatial_Mean’ variables. We transformed ‘Pct_Black’ with log(‘Pct_Black’ + 1) because many of the ‘Pct_Black’ values were close to 0. After the transformation, we saw that the residuals were more randomly scattered around the horizontal 0 line.

4.3 Final Model

Using the transformed variables, we fit our final model:



```
                                OLS Regression Results
==============================================================================
 Baseline Model:        Pre-transformation:       Post-transformation:
                                                  R-squared:                       0.626
 Adj. R²:  0.460        Adj. R²:  0.613           Adj. R-squared:                  0.625
                                                  F-statistic:                     1195.
                                                  Prob (F-statistic):               0.00
                                                  Log-Likelihood:                -2011.4
 AIC:      6125         AIC:      4221            AIC:                             4041.
                                                  BIC:                             4101.
=================================================================================
                              coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------
Intercept                   1.6932      0.083     20.379      0.000       1.530       1.856
Pct_Age_lt_18              -0.0233      0.002    -13.950      0.000      -0.027      -0.020
log_Pct_Black              -0.0566      0.006     -9.750      0.000      -0.068      -0.045
Potential_Years_Lost     9.327e-05   3.41e-06     27.313      0.000    8.66e-05      0.0001
Pct_Uninsured              -0.0106      0.001    -10.409      0.000      -0.013      -0.009
log_PrimCarePhys_per_100k  -0.0923      0.010     -9.291      0.000      -0.112      -0.073
Pct_Child_in_1ParentHH      0.0084      0.001      9.019      0.000       0.007       0.010
Pct_Poverty                -0.0119      0.001     -7.987      0.000      -0.015      -0.009
log_Spatial_Mean            0.5550      0.012     46.119      0.000       0.531       0.579
=================================================================================
```

Table 4: Detailed OLS Output of Final Model

Table 4 shows the detailed output of our final model: the top half shows comparison of the 3 different models (Baseline, Pre-Transformation, and Post-Transformation/Final); the bottom half is the model estimates.

4.4 Model Interpretation

Looking at the top half of Table 4, we can compare the adjusted R^2 values and the AIC scores. For every consecutive model, our adjusted R^2 increased slightly (0.46 to 0.613 to 0.625) and our AIC score decreased slightly (6125 to 4221 to 4041), meaning that each subsequent model had improved performance. In addition, the normalized test RMSE for this model is 0.07887, which is lower than our baseline model's 0.0938 normalized RMSE.

Now looking at the bottom half of Table 4, we see that all the p-values are approximately 0, which means that all the variables are significant at a 0.05 significance level. Some of the model coefficient signs do not follow the hypothesized relationships that we stated in Table 1. In particular, having more people in poverty and uninsured actually decreases the drug overdose rates in our model. These unintuitive relationships may be due to multicollinearity still present within our model, since multicollinearity may lead the coefficients to change signs [11]. However, the goal of our project is to determine which features are significant rather than how the features are associated with the drug overdose rates.

4.5 Residual Analysis

To check if our model accounted for the spatial component in our data, we plot our residuals on the U.S. map.

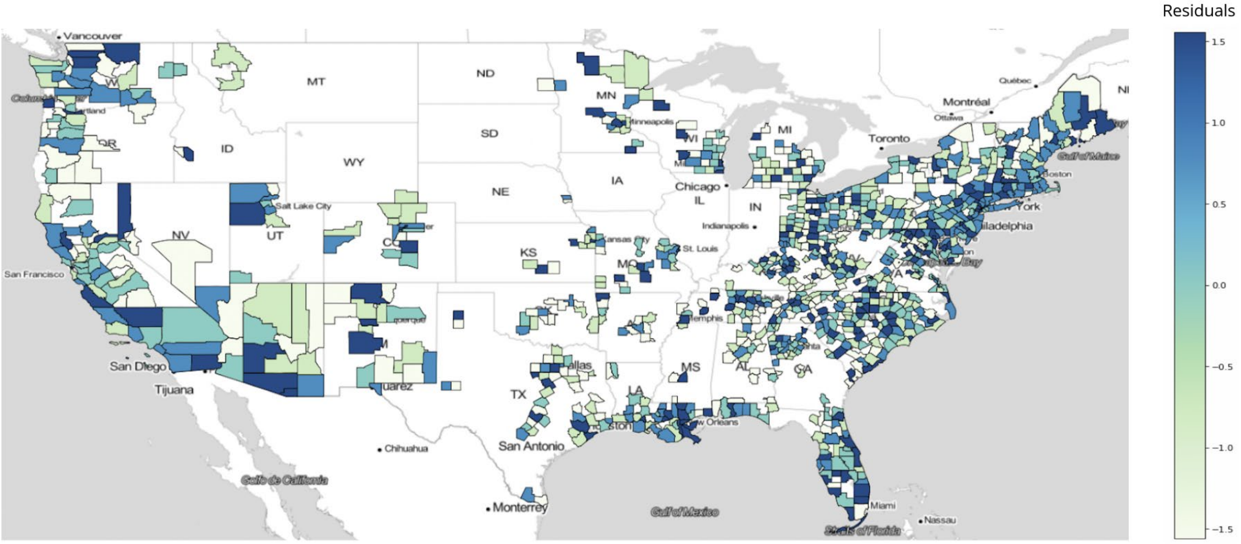


Figure 6: Residual Map for 2020

Above is our residuals plotted on the U.S. map for our 2020 model. We see that the residuals are randomly scattered across the different counties. The Moran's I for the residuals is approximately -0.098 and thus close to 0, which confirms that the spatial autocorrelation is no longer significant. Thus, our model has successfully accounted for the spatial component in our data. We see similar results for models across different years (2011-2020).

4.6 Model Estimates on the U.S. Map

With our final model, we are now able to estimate drug overdose rates and fill in the missing drug overdose data. To do this, we need to estimate drug overdose rates recursively. The procedure that we used is as follows:

1. Use model to estimate missing overdose rates for counties that are adjacent to those that have non-null 'Spatial_Mean' data
2. With these new estimates of the adjacent missing county overdose rates, calculate a new 'Spatial_Mean'
3. Use model again to estimate county overdose rates that are still missing
4. Repeat until we can no longer estimate missing overdose rates

With this procedure, we are estimating overdose rates for the counties adjacent to those for which we have the raw data. Then we are further estimating those counties' adjacent counties and so on. We are not able to estimate every county's overdose rate because they have missing data for other covariates in our model.

An example of our model estimating procedure for 2020 is shown below. Iteration 0 is the raw observed overdose rates from the CDC and Iteration 9 is the estimates produced using our recursive estimation procedure.

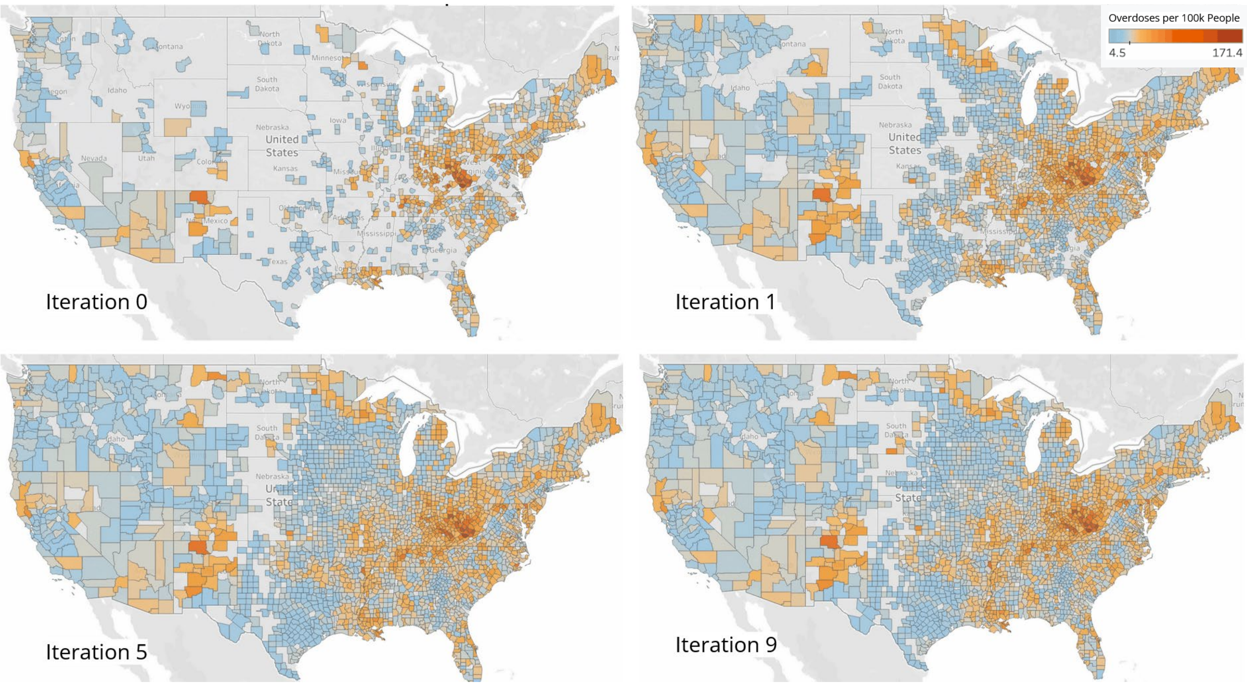


Figure 6: Residual Map for 2020

Visually, these estimates are similar to what we expect the drug overdose choropleth to look like in 2020. We expect many areas to be orange, indicating higher overdose rates in 2020. We also expect the clusters of orange that we see, especially in the Appalachian region. We are able to repeat this procedure for all years and get a fuller picture of what drug overdose rates look like across the U.S. this past decade.

5 Conclusions

5.1 Summary

Through our geospatial correlation analysis and subsequent modeling, we have determined that spatial components have the best explanatory power among the different features we selected as potential factors contributing to the drug overdose death rates in the United States. We have also filled out the map of the U.S. with estimates using our model recursively.

5.2 Conclusion

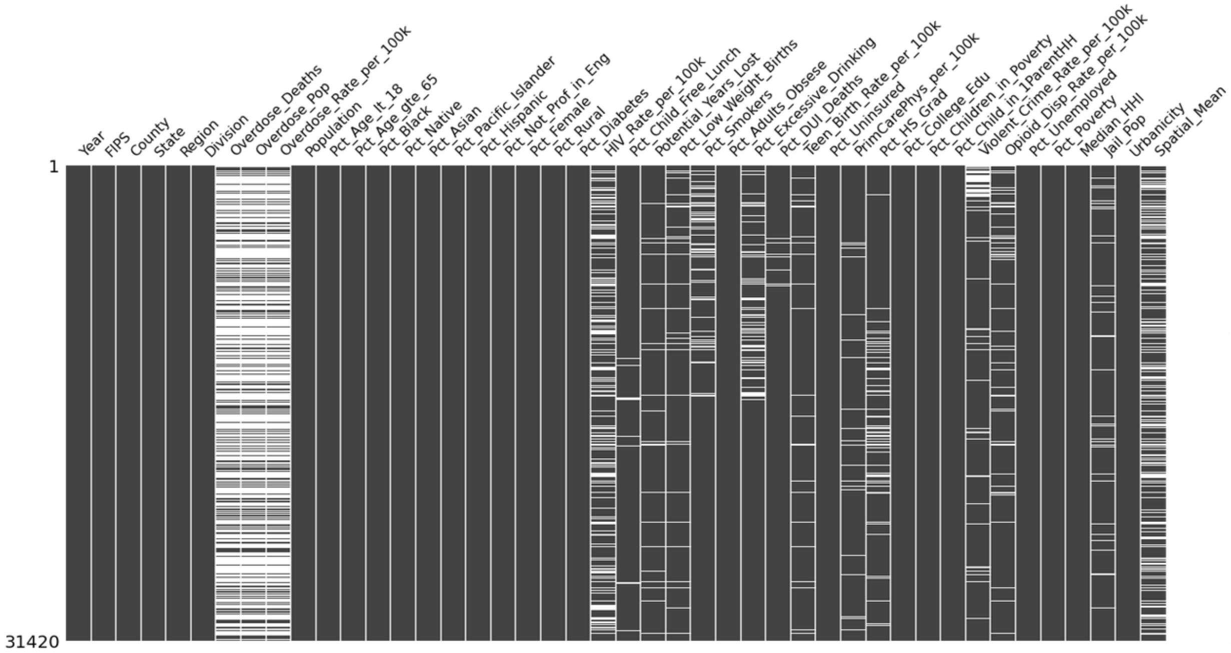
With these new estimates, we hope that policy makers can get more insight for those areas that had missing overdose rates. By considering numerous health and demographic factors, and using backwards selection to narrow down the most impactful features, we were able to determine what factors to focus on as well as which communities should be addressed more urgently when trying to counteract the drug overdose epidemic.

We found that policy makers should target counties with severe overdose rates and their surrounding counties because counties with high drug overdose rates tend to have a spillover to nearby counties which creates county clusters. Policy makers should also aim to improve health care availability because some of the significant variables in our model include the percent of residents uninsured (‘%_Uninsured’) and the number of primary care physicians available per 100,000 residents (‘PrimCarePhys_per_100k’). Policy makers should also focus on these demographics: youth less than 18 years old, the black population, those in poverty, and single parent households. With these factors in mind, policy makers can promote and encourage action in these communities to reduce drug overdose rates.

The drug overdose epidemic is a very complex societal issue that is being researched at many institutions around the world. Thus, we understand that there is still room for improvement for our model. For example, we may be leaving out predictors that are important but were not included because we did not procure those variables in our data collection. We may also get better estimates using other models such as weighted least squares that have population and spatial weights, or a time series model that takes into account different yearly trends. Despite these shortcomings, our model adequately takes into account the spatial component of drug overdoses. With further research, we hope that researchers, analysts, and policy makers alike can solve the drug epidemic in the U.S. soon.

Appendix

Figure 1. Completeness of data.



Black fill indicates data that is present and white indicates missing data.

Table 1. Variables of Interest and Hypothesized Relationship to Drug Overdose Death Rates

Variable Name	**Variable Description** **(all variables are at county-level)**	**Hypothesized relationship to drug overdose death rates**
Overdose_Rate_per_100k	Number of drug overdose deaths per 100,000 persons	
Opioid_Disp_Rate_per_100k	Retail opioid prescriptions dispensed per 100,000 persons	Opioid_Disp_Rate_per_100k *increase* → Overdose_Rate_per_100k *increase*
Pct_Unemployed	Percent of unemployed residents	Unemployment_rate *increase* → Overdose_Rate_per_100k *increase*
Pct_Uninsured	Percent of uninsured residents	Pct_Uninsured *increase* → Overdose_Rate_per_100k *increase*

Pct_Poverty	Estimated percent of people of all ages in poverty	Pct_Poverty*increase* → Overdose_Rate_per_100k *increase*
Median_HHI	Estimate of median household income	Median_HHI*increase* → Overdose_Rate_per_100k *decrease*
Pct_Age_lt_18 Pct_Age_gte_65	Percentage of county residents < 18 years old and percentage of county resident > 65 years old	Pct_Age_*_* *increase* → Cruder_Rate *increase*
Jail_Pop	Jail population count	Jail_Pop *increase* → Overdose_Rate_per_100k *increase*
Spatial_Mean	Mean Cruder_Rate of adjacent counties	Spatial_Mean *increase* → Overdose_Rate_per_100k *increase*

Figure 3. Moran Local Plots for West Virginia

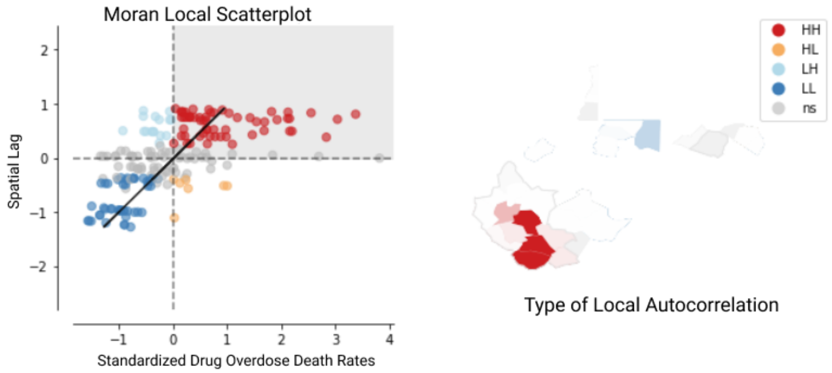


Figure 4. Overdose_Rate_per_100k Transformation

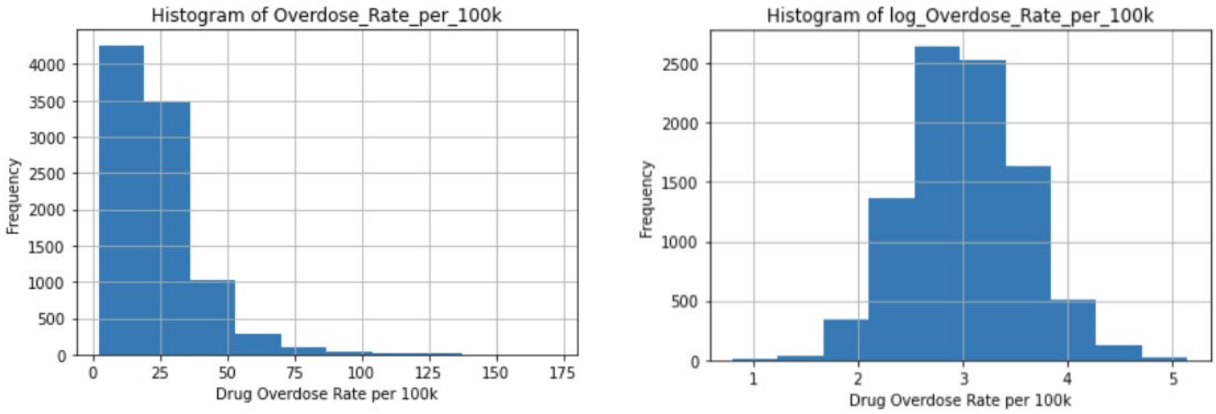
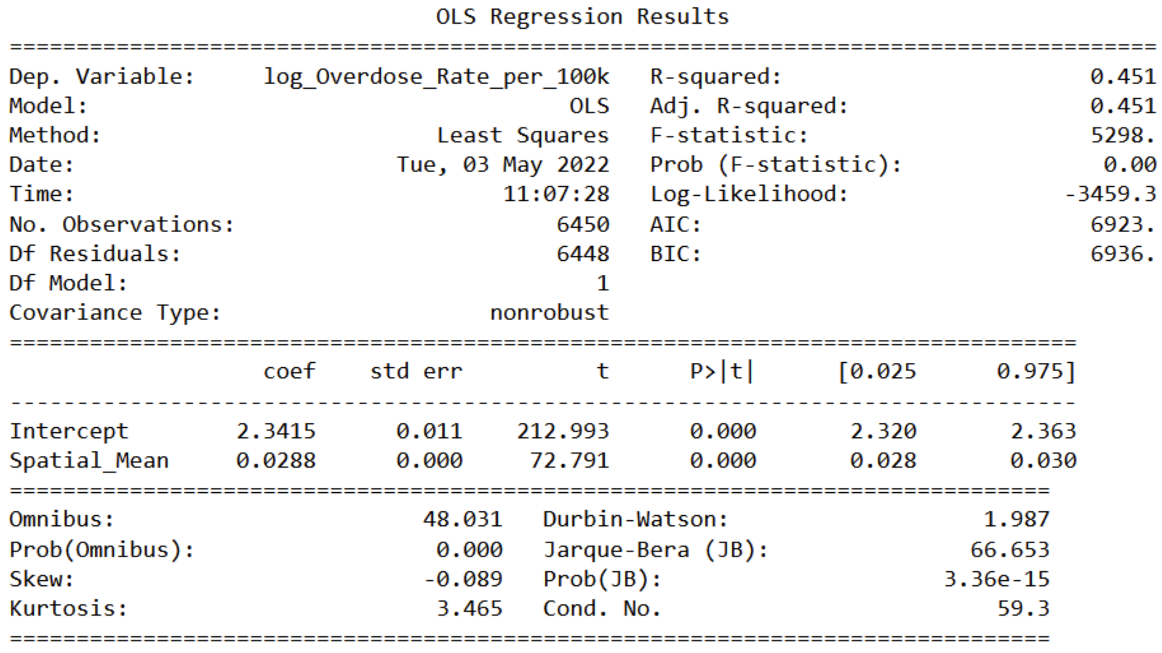
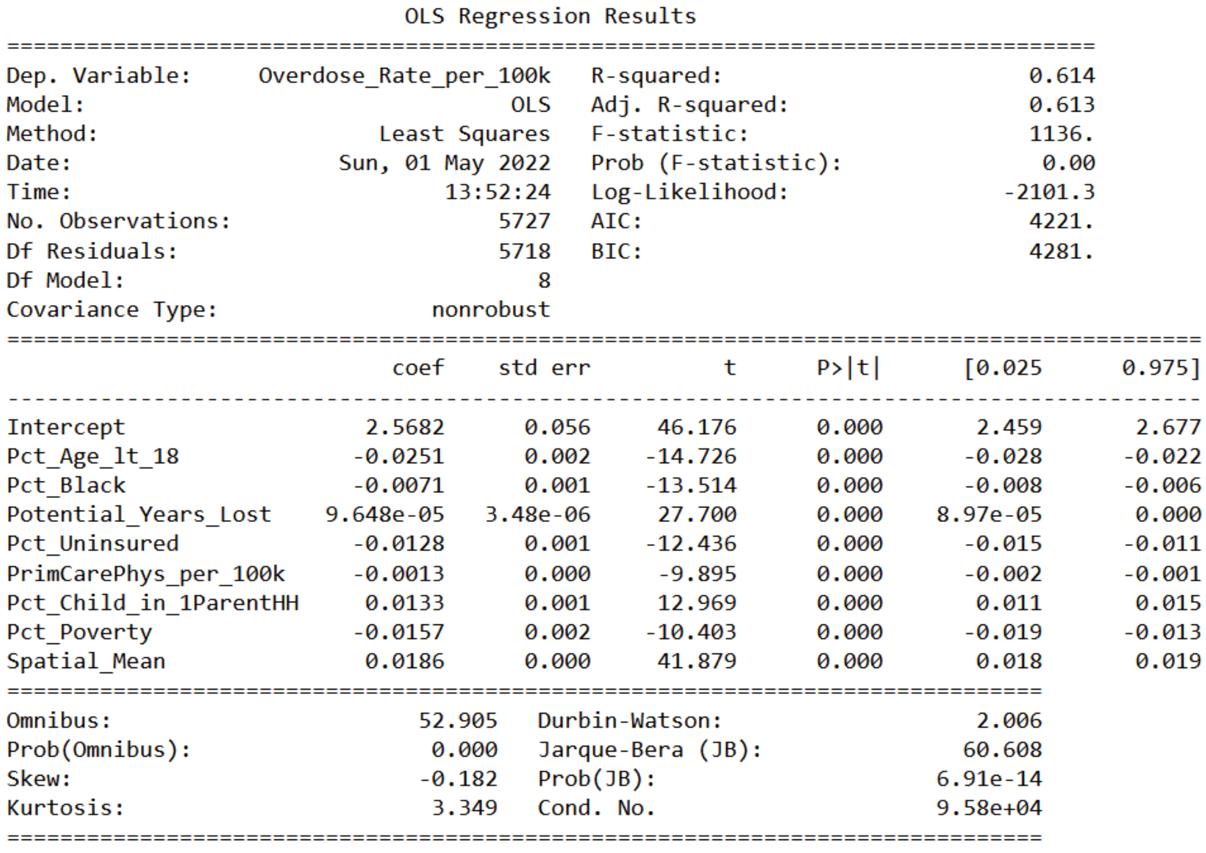


Table 2. Detailed OLS Output of Baseline Model



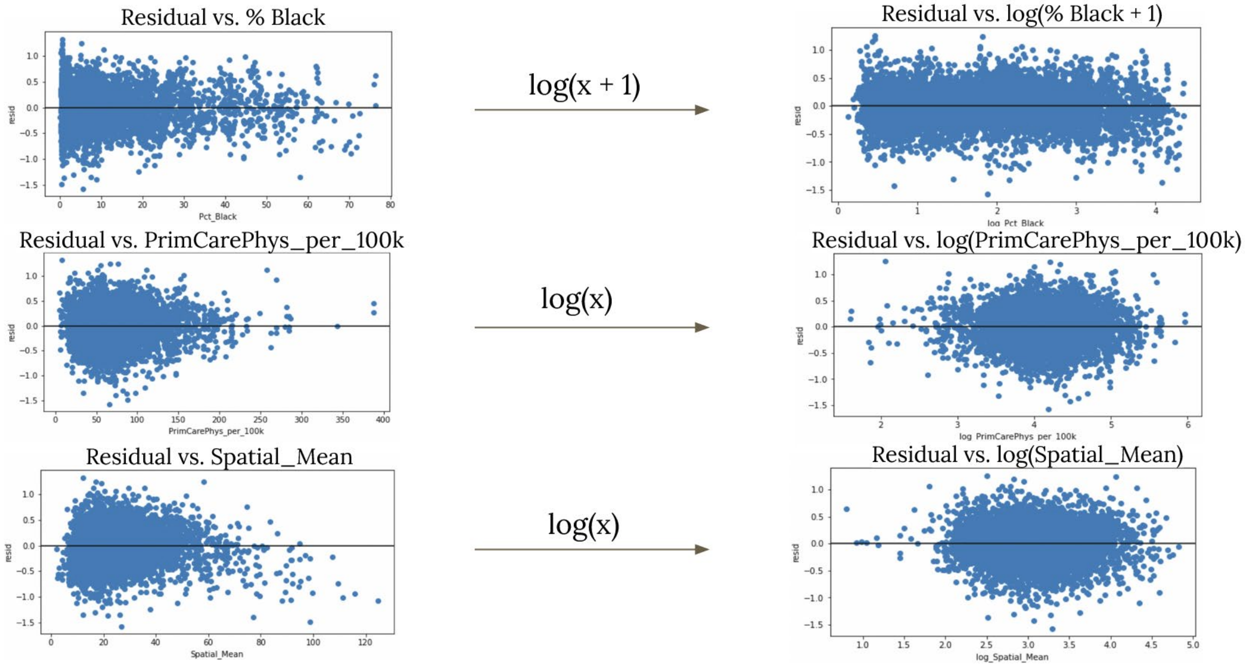
```
                               OLS Regression Results
====================================================================================
Dep. Variable:    log_Overdose_Rate_per_100k   R-squared:                      0.451
Model:                                   OLS   Adj. R-squared:                 0.451
Method:                        Least Squares   F-statistic:                    5298.
Date:                       Tue, 03 May 2022   Prob (F-statistic):              0.00
Time:                               11:07:28   Log-Likelihood:               -3459.3
No. Observations:                       6450   AIC:                            6923.
Df Residuals:                           6448   BIC:                            6936.
Df Model:                                  1
Covariance Type:                   nonrobust
==============================================================================
                  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
============================================================================
```

Table 3. Detailed OLS Output of Backwards Selection Model
(Note: Overdose_Rate_per_100k is logged)



```
                             OLS Regression Results
================================================================================
Dep. Variable:    Overdose_Rate_per_100k   R-squared:                      0.614
Model:                               OLS   Adj. R-squared:                 0.613
Method:                    Least Squares   F-statistic:                    1136.
Date:                   Sun, 01 May 2022   Prob (F-statistic):              0.00
Time:                           13:52:24   Log-Likelihood:               -2101.3
No. Observations:                   5727   AIC:                            4221.
Df Residuals:                       5718   BIC:                            4281.
Df Model:                              8
Covariance Type:               nonrobust
========================================================================================
                            coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------
Intercept                 2.5682      0.056     46.176      0.000       2.459       2.677
Pct_Age_lt_18            -0.0251      0.002    -14.726      0.000      -0.028      -0.022
Pct_Black                -0.0071      0.001    -13.514      0.000      -0.008      -0.006
Potential_Years_Lost   9.648e-05   3.48e-06     27.700      0.000    8.97e-05       0.000
Pct_Uninsured            -0.0128      0.001    -12.436      0.000      -0.015      -0.011
PrimCarePhys_per_100k    -0.0013      0.000     -9.895      0.000      -0.002      -0.001
Pct_Child_in_1ParentHH    0.0133      0.001     12.969      0.000       0.011       0.015
Pct_Poverty              -0.0157      0.002    -10.403      0.000      -0.019      -0.013
Spatial_Mean              0.0186      0.000     41.879      0.000       0.018       0.019
============================================================================
Omnibus:                      52.905   Durbin-Watson:                  2.006
Prob(Omnibus):                 0.000   Jarque-Bera (JB):              60.608
Skew:                         -0.182   Prob(JB):                    6.91e-14
Kurtosis:                      3.349   Cond. No.                    9.58e+04
============================================================================
```

Figure 5. Residual Plot Transformations



Residual vs. % Black

Residual vs. log(% Black + 1)

Residual vs. PrimCarePhys_per_100k

Residual vs. log(PrimCarePhys_per_100k)

Residual vs. Spatial_Mean

Residual vs. log(Spatial_Mean)

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. 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
3. 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
4. Rankings Data & Documentation: National Data & Documentation: 2010-2020. County Health Rankings & Roadmaps. (n.d.). Retrieved May 10, 2022, from https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation/national-data-documentation-2010-2019
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. 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
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. Rey, S. J., Arribas-Bel,, D., & Wolf, L. J. (2020). Geographic Data Science with python. Global Spatial Autocorrelation - Geographic Data Science with Python. Retrieved March 20, 2022, from https://geographicdata.science/book/notebooks/06_spatial_autocorrelation.html?highlight=moran+s+i#continuous-case-moran-plot-and-moran-s-i, https://geographicdata.science/book/notebooks/04_spatial_weights.html#contiguity-weights

9. Oshan, T., Li, Z., Kang, W., Wolf, L., & Fotheringham, A. (2019). mgwr: A Python Implementation of Multiscale Geographically Weighted Regression for Investigating Process Spatial Heterogeneity and Scale. ISPRS International Journal of Geo-Information, 8(6), 269. https://doi.org/10.3390/ijgi8060269
10. Choueiry, G. (n.d.). Understand forward and backward stepwise regression. Quantifying Health. Retrieved May 10, 2022, from https://quantifyinghealth.com/stepwise-selection/
11. Frost, J. (2021, September 24). Multicollinearity in regression analysis: Problems, detection, and solutions. Statistics By Jim. Retrieved May 10, 2022, from https://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/