The US Opioid Epidemic:

Predicting Mortality From Community Characteristics

by Glynis Daniels

Executive Summary

The United States is currently in the midst of an opioid crisis that is causing record numbers of overdose deaths. This epidemic is not equally felt in all areas of the country, however. In fact the distribution of overdose deaths is highly skewed, with some places experiencing exponential increases and others barely impacted. Efforts to address the problem could be made more effective by predicting which regions of the country will be most impacted and by understanding what socioeconomic characteristics are associated with high rates of addiction. This project utilizes data on drug overdose mortality from NCHS combined with socioeconomic data from the US Census to build a predictive model. Several modeling techniques are tested using cross-validation. After hyperparameter tuning, the best observed model is a nonlinear support vector regression machine. This model achieves an R-square of 0.60.

The nonlinear support vector machine model operates as a 'black box', and cannot assess relative feature importance. But descriptive information from predicted mortality rates appears to be generally consistent with bivariate analysis. Broadly speaking, it appears that higher rates of drug overdose mortality are associated with markers of lower economic status (such as unemployment, poverty, reliance on public assistance income, use of public insurance) as well as weakened social support systems (such as non-family households, adults previously married). A key recommendation from this study is that these communities should be targeted for both addiction recovery services and preventative outreach and education, due their high vulnerability to this epidemic.

As new Census data becomes available this model can be used to generate updated predictions to guide intervention efforts. But one of the shortcomings of the model is that it fails to predict some of the highest rates of drug mortality that have been seen in the US. It may be possible to improve the model by adding additional features.

1. Introduction: The US Opioid Epidemic¶

In 2017 the United States reached a record high[1] of 70,237 drug overdose deaths, according to recent data from the Department of Health and Human Services[2]. This represents a 9.6% increase[3] in age-adjusted deaths in just one year. About 68% of these deaths are opioid-related.

In response to the severity of the crisis, in October 2018 the US Congress passed a landmark bill[4] to address the problem on several fronts. Among the approaches is to expand the use of 'comprehensive recovery centers' that provide an array of services to recovering addicts. Making the best use of these new funds, as well as state and local resources, can be aided by the ability to predict which locations around the US will have the greatest need for these services. It may also be helpful to understand what local socioeconomic characteristics, if any, are most associated with high levels of drug overdose deaths. Addressing these two needs is the focus of this report.

1.1 Measuring the Opioid Crisis¶

To investigate these questions two types of data are needed: 1) data on the extent of the opioid crisis in different parts of the US, and 2) socioeconomic data for those same regions.

The overdose mortality rate is just one possible measure of the opioid crisis. It would be preferable to use a broader indicator, such as addiction prevalence, so as to focus on more common and earlier stages of the problem. Mortality data has the advantage of being collected consistently across the country, which is a key requirement. The primary source for mortality data in the US is the CDC WONDER[5] system, which is an interactive site that creates data extracts on the fly. (There is currently no API for this interactive system.) Data can be broken down by cause of death, population, year, and geographic unit.

The limitation to CDC WONDER mortality data, however, is that small numbers of deaths (<10) are suppressed by the system, resulting in missing data. For annual data from geographic areas smaller

1 See https://www.nytimes.com/interactive/2018/11/29/upshot/fentanyl-drug-overdose-deaths.html

2 See https://www.cdc.gov/drugoverdose/epidemic/index.html

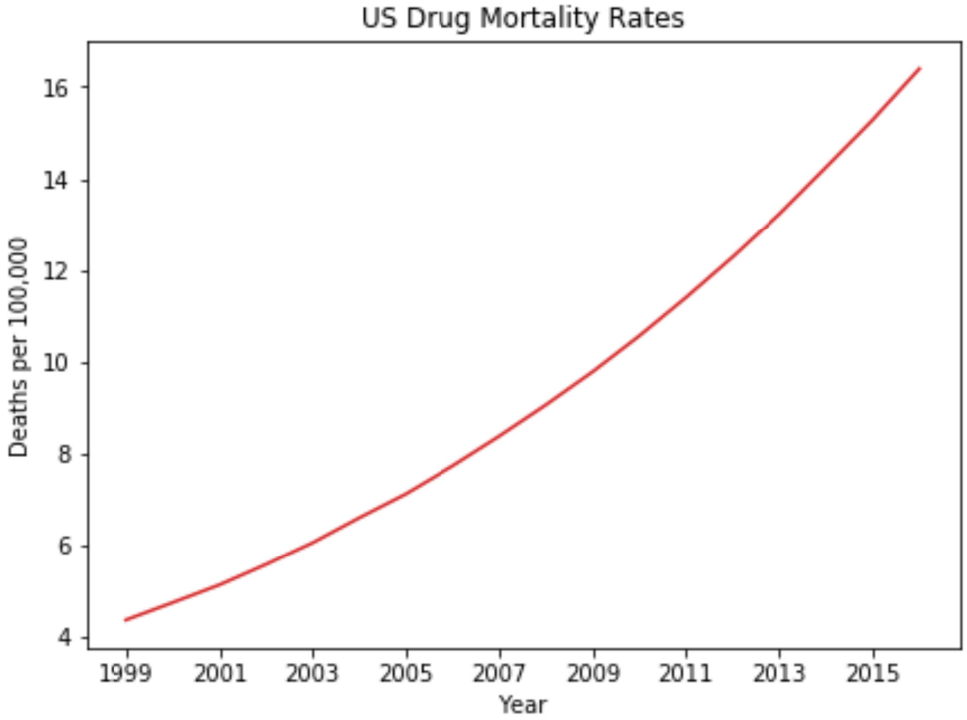
3 See https://www.cdc.gov/drugoverdose/data/statedeaths.html

4 See https://www.usatoday.com/story/news/politics/2018/10/24/donald-trump-opioids-bill-includes-changes-trafficking-treatment/1752329002/

5 See https://wonder.cdc.gov/

than the state level, the result is an unacceptable amount of missing data.

Another option is to use a smoothed mortality rate data set[6] available from the National Center for Health Statistics (NCHS) that provides annual county-level age-adjusted drug poisoning rates in fairly narrow bins. The rate is measured in deaths per 100,000 and then binned by 2's: i.e., <2, 2-3.9, 4-5.9, up to 30+. The midpoints of the ranges can be used to approximate a continuous variable. Data is available for all US counties for the period 1999-2016.

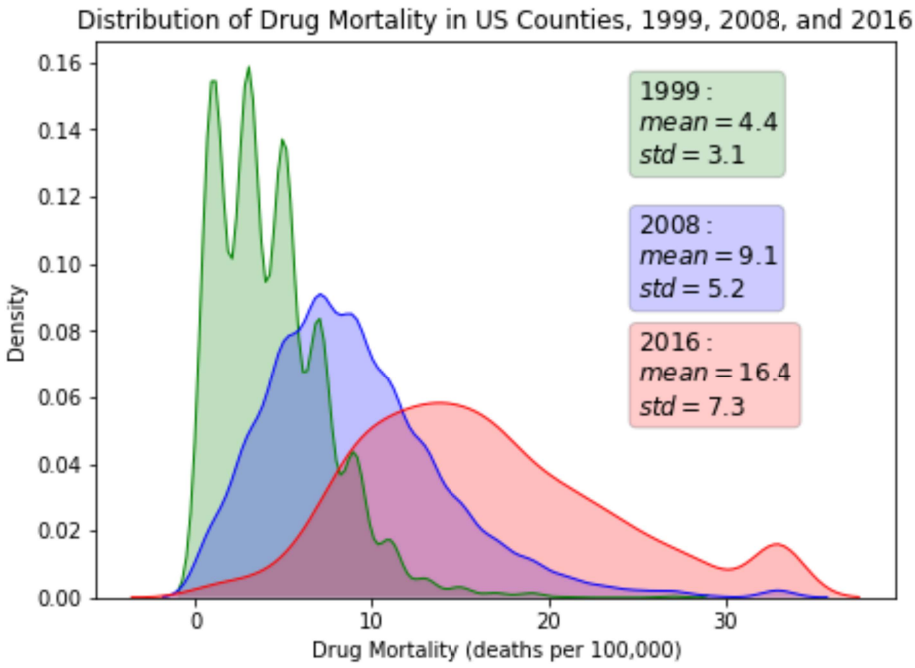


This trend line show how quickly drug-related deaths have increased from 1999-2016 in the US. It is important to note that the NCHS data is not broken down by type of drug involved. Other sources, such as this report from the National Institute on Drug Abuse[7], indicate the rise in deaths is primarily due to opioid related drugs (including heroin, natural opioids, methadone, fentanyl, and other synthetic opioids). Disturbingly, the NIDA report provides national estimates for the years 2016-17 that show continued acceleration of opioid deaths. In particular, deaths attributed to fentanyl rose from near 10,000 in 2015 to over 29,000 in 2017.¶

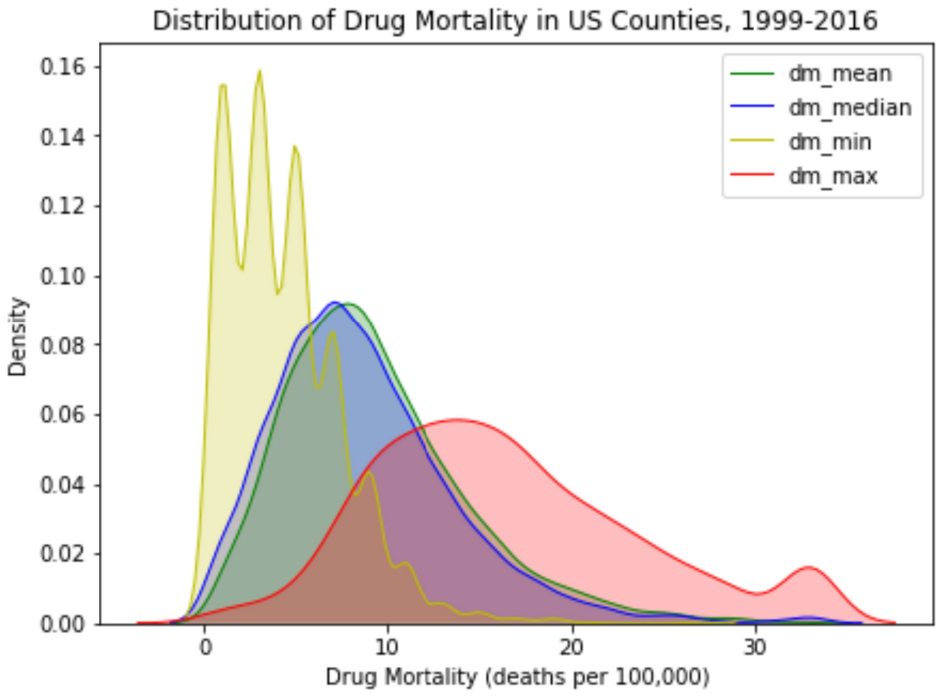
6 See https://data.cdc.gov/NCHS/NCHS-Drug-Poisoning-Mortality-by-County-United-Sta/pbkm-d27e

7 See https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates

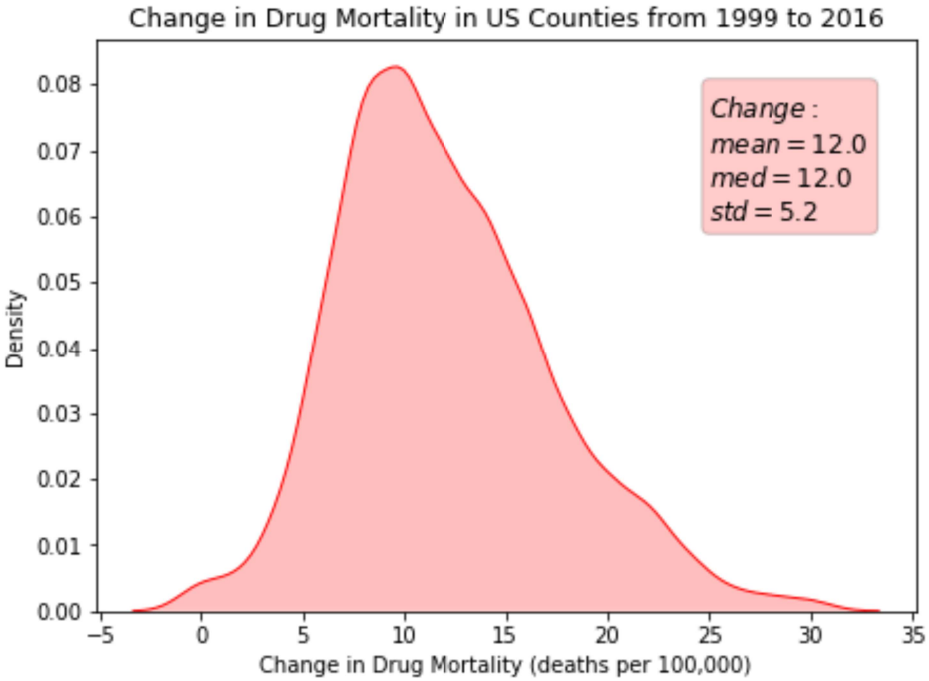
The distribution of drug-related deaths has changed form over time and become more normal as deaths have increased. (Note that the tail end of the 2016 distribution is the result of top-coding in the input data).¶



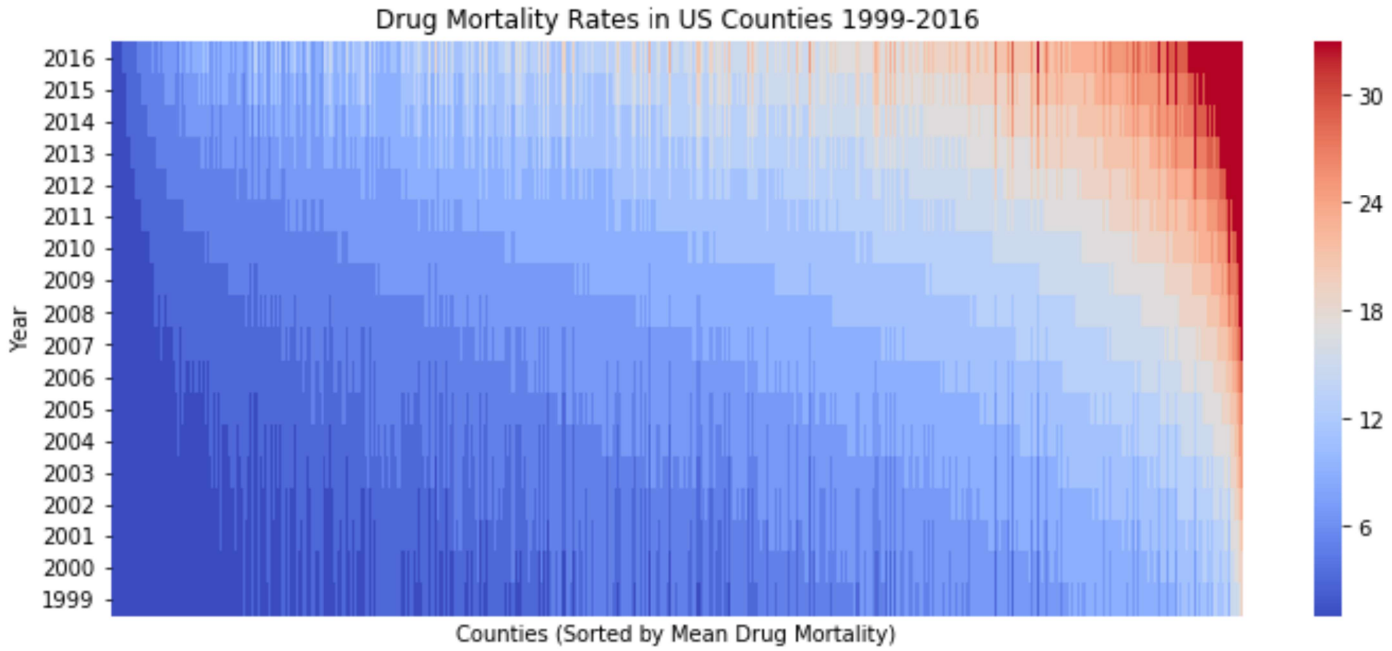
The chart below depicts drug mortality aggregated across years for all US counties. We can see that most counties have a minimum drug mortality rate below 10/100,000. Only a minority of counties have a maximum mortality rate below 10.¶



From 1999—2016, over half of US counties saw an increase in drug-related mortality of 12 or more deaths per 100,000 population.¶



Counties in the heatmap below are sorted left to right by their average drug mortality rate. We can see in this graphic that nearly all US counties saw an increase in drug mortality.



2. Predictor Variables¶

What characteristics of US counties might be predictive of drug-related mortality rates? My approach is heavily influenced by the recent work of Ann Case and Angus Deaton, two Princeton economists who became interested in opioid deaths in 2015 after noticing the shocking decline[8] in life expectancy for middle-aged white Americans. In a spring 2017 Brookings Institute article[9], they investigate on this trend through exploratory research. They develop a broad hypothesis that increases in "deaths of despair" (defined as drug overdose, alcohol-related deaths, and suicide) are one component of a larger trend:

> *We propose a preliminary but plausible story in which cumulative disadvantage from one birth cohort to the next—in the labor market, in marriage and child outcomes, and in health—is triggered by progressively worsening labor market opportunities at the time of entry for whites with low levels of education. This account, which fits much of the data, has the profoundly negative implication that policies—even ones that successfully improve earnings and jobs, or redistribute income—will take many years to reverse the increase in mortality and morbidity, and that those in midlife now are likely to do worse in old age than the current elderly.*

In a Washington Post[10] opinion piece they state:

> *Opioids are like guns handed out in a suicide ward; they have certainly made the total epidemic much worse, but they are not the cause of the underlying depression. We suspect that deaths of despair among those without a university degree are primarily the result of a 40-year stagnation of median real wages and a long-term decline in the number of well-paying jobs for those without a bachelor's degree. Falling labor force participation, sluggish wage growth, and associated dysfunctional marriage and child-rearing patterns have undermined the meaning of working people's lives as well.*

From the work of Case and Deaton we can develop a preliminary list of potential correlates with drug-related mortality:

Demographics¶

- Age
- Race
- Hispanic ethnicity

8 http://www.pnas.org/content/112/49/15078

9 https://www.brookings.edu/past-bpea-editions/

10 https://www.washingtonpost.com/opinions/the-truth-about-deaths-of-despair/2017/09/12/15aa6212-8459-11e7-902a-2a9f2d808496_story.html?utm_term=.c3c7ce910d18

- Sex
- Marital status
- Children's living arrangements

Economic indicators

- Income
- Poverty rates
- Public assistance
- Unemployment
- Labor force participation
- Education
- Health insurance

Many of the above indicators can be derived from the American Community Survey from the US Census Bureau. The 5-year American Community Survey (ACS) will be used. The data is representative of a 5-year period, in this case 2011-2015. There are also 3- and 1-year data files available, but those datasets do not contain estimates for places with fewer than 20,000 or 65,000 residents respectively. (This means not all counties are represented in those datasets, because many counties have fewer than 20,000 residents.)

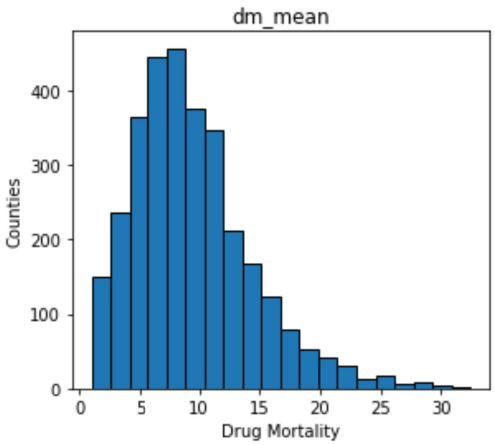
ACS data were accessed through the Census Bureau's API. Thirteen different census tables were downloaded at the county level of geography. In the NCHS drug mortality data several counties in Alaska and Colorado had been combined, so it was necessary to aggregate the counts for these counties in each Census table prior to calculating rates. About 69 county-level socioeconomic features were created from the Census data, although about 23 of these are detailed age breakdowns that were not eventually included in the model.

In addition to the Census ACS data, I have also included data on urban/rural status provided by NCHS.

3. Exploratory Analysis¶

3.1 Univariate Distributions

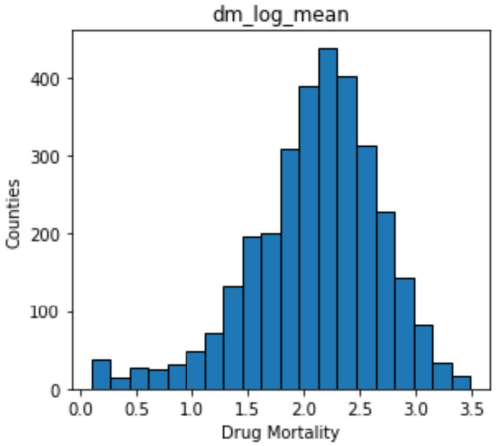
Preliminary explorations of the NCHS data show that its limitations will make it difficult to model in raw form. These limitations include the binning (creating loss of precision) and the top-coding or ceiling effect that caps the mortality rate at 30+/100,000. One alternative is to use the mean drug mortality over the 17 year observation period as an overall indicator of the extent to which the opioid epidemic had affected the county. No US counties have seen a decline in drug mortality, although a few have managed to remain at low levels.



Univariates

mean = 9.3
median = 8.6
std = 4.9
minimum = 1.0
maximum = 32.4
skew = 1.0

The distribution of mean drug mortality across counties is skewed to the right. It probably would be more skewed right if not for the top-coding. Taking the log of the mean does make the distribution more normal, but it takes on a left skew.



Univariates

mean = 2.1
median = 2.2
std = 0.6
minimum = 0.1
maximum = 3.5
skew = -0.8

Density plots were created for all numeric variables (not shown), which includes all Census measures. Many of these distributions are not normal, but taking the log has improved many of them. The variable urban_2013 is an ordinal measure of urbanization, which can also be treated as a categorical variable (1=city, 6=rural). Nearly half of US counties are rural and only 68/3135 (2%) are large cities. However, over 30% of the population lives in large cities while only about 6% live in the most rural areas. .¶

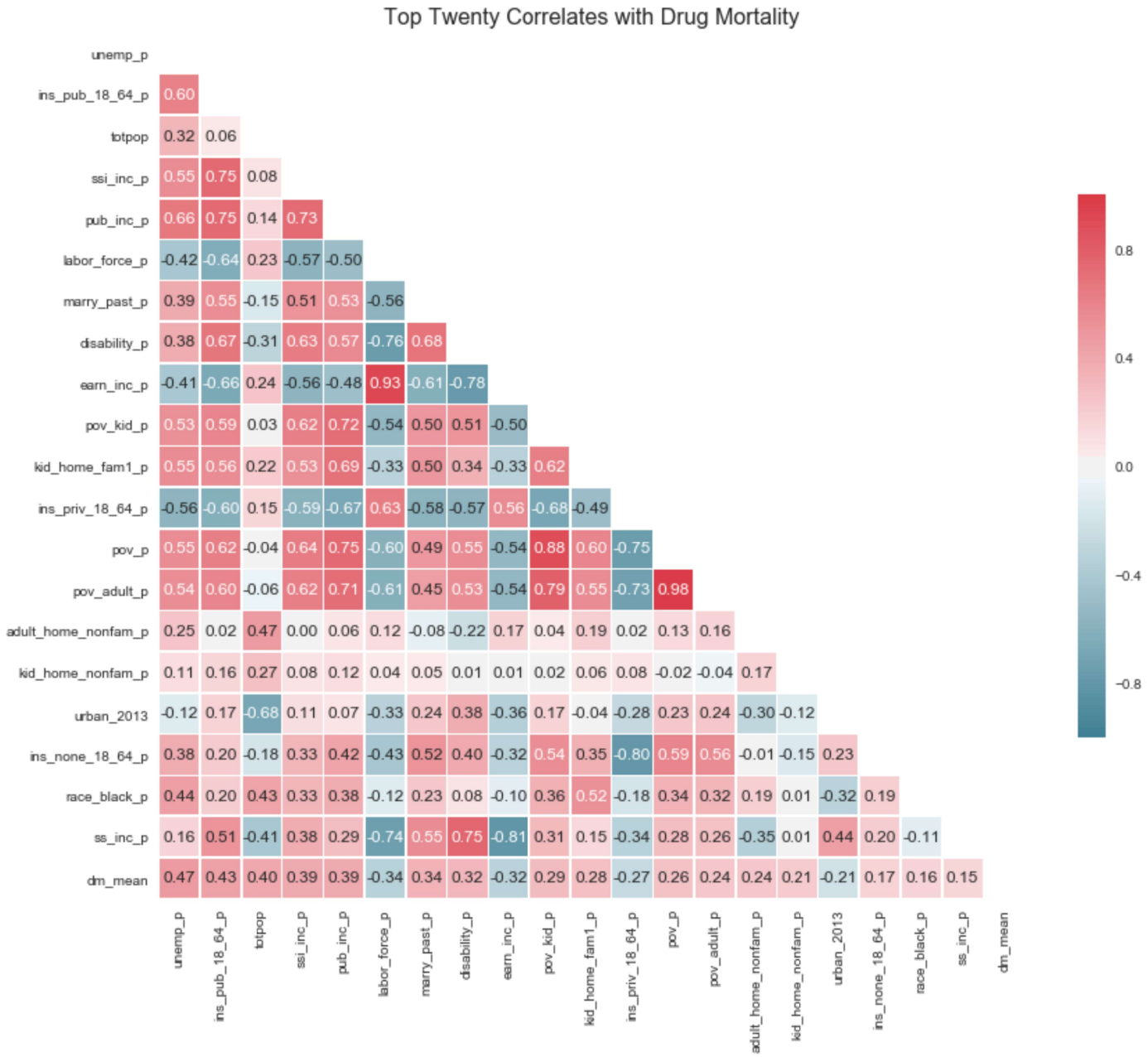
3.2 Correlations Among Numeric Variables¶

The table below lists the 20 features that have the highest correlations with drug mortality. We can see that the county-level factors associated with drug mortality in some ways echo the 'deaths of despair' framework that emerged from the exploratory work of Case and Deaton. Their work focuses on the characteristics of *people* with high drug mortality rates, whereas this county-level analysis uses data on communities. At the community level the strongest relationships seem to be with facets of economic disadvantage and stagnation, including rates of unemployment, use of public assistance, lack of private insurance, and poverty. We can also see an association between drug mortality and indicators of weakened social support, including rates of adults and children living in non-family households, adults who are previously married (i.e., divorced or widowed), and physical disabilities. Finally, we see a small positive relationship between drug mortality and the percent Black, which is surprising considering that race-specific death rates make clear that the opioid epidemic has had far greater impact on whites in the US (see Case and Deaton above). I suspect that this is likely to be a spurious correlation caused by the fact that more urbanized counties in the US have higher proportions of Black residents.

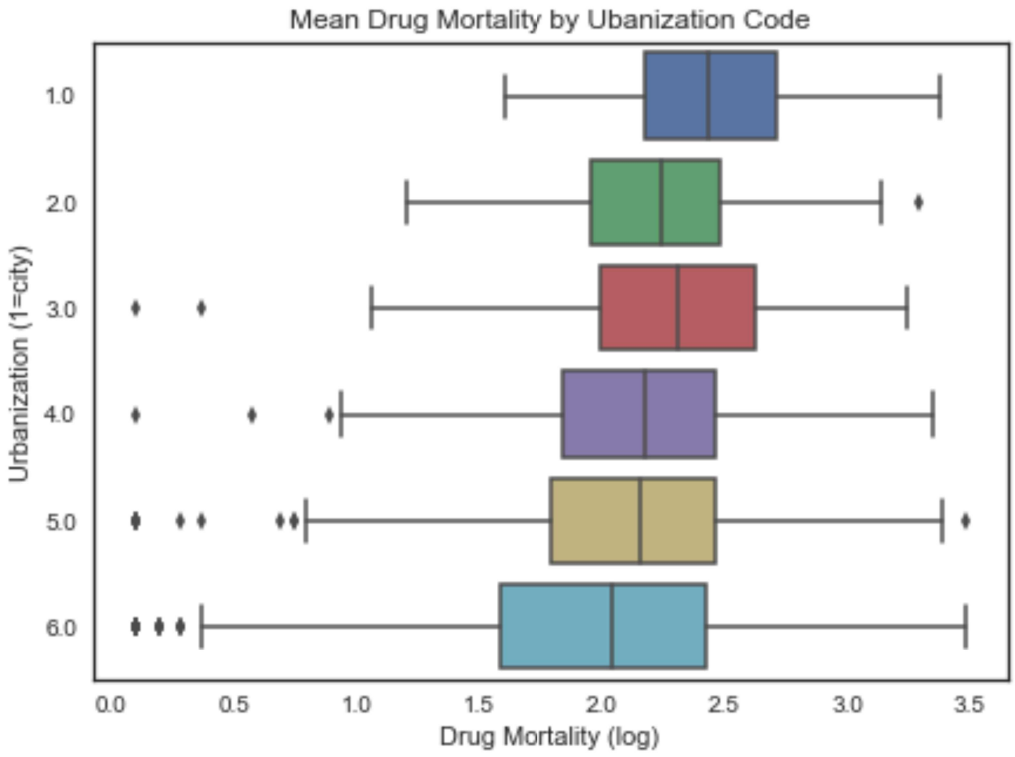
Correlation	Feature Name	Description
0.474638	unemp_p	Unemployment rate
0.431599	ins_pub_18_64_p	Percent of adults with public insurance
0.400171	totpop	Population
0.387905	ssi_inc_p	Percent of households with Supplemental Security Income
0.385667	pub_inc_p	Percent of households with public assistance
-0.343611	labor_force_p	Labor force participation rate
0.340851	marry_past_p	Percent of adults married in the past
0.318582	disability_p	Percent of adults with a disability
-0.318407	earn_inc_p	Percent of households with earned income

Correlation	Feature Name	Description
0.291051	pov_kid_p	Child poverty rate
0.278506	kid_home_fam1_p	Percent of children living in single-parent households
-0.272750	ins_priv_18_64_p	Percent of adults with private insurance
0.260372	pov_p	Poverty rate (overall)
0.241284	pov_adult_p	Poverty rate (adults)
0.236611	adult_home_nonfam_p	Percent of adults living in non-family households
0.208037	kid_home_nonfam_p	Percent of children living in non-family households
-0.205553	urban_2013	Urbanization category (1=city thru 6=rural)
0.166502	ins_none_18_64_p	Percent of adults with no insurance
0.162094	race_black_p	Percent Black
0.150648	ss_inc_p	Percent of households with Social Security income

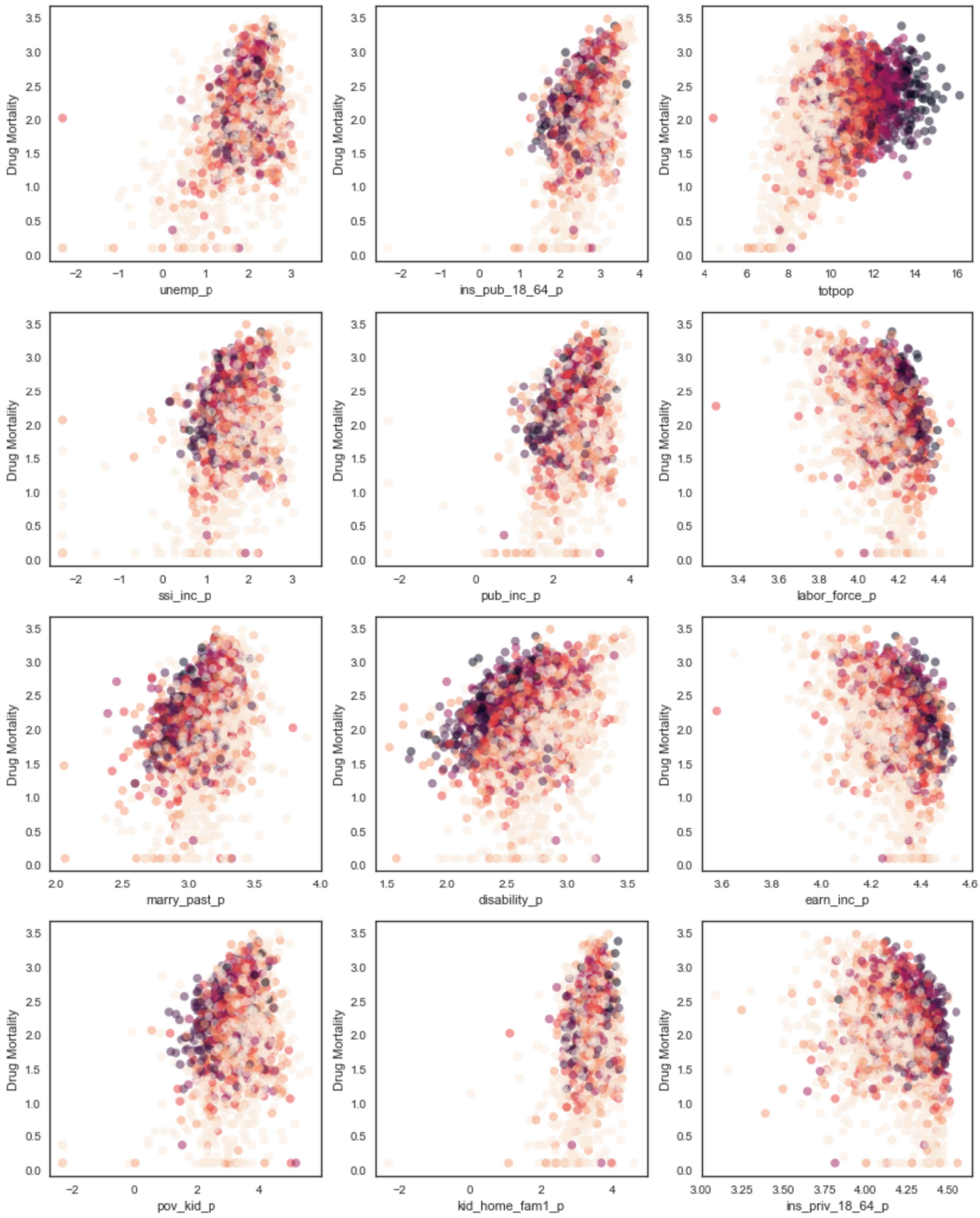
The matrix below shows correlations between mean drug mortality and the top 20 correlates. The correlations with drug mortality can be seen along the bottom row. Many of these predictor variables are also highly correlated with each other.

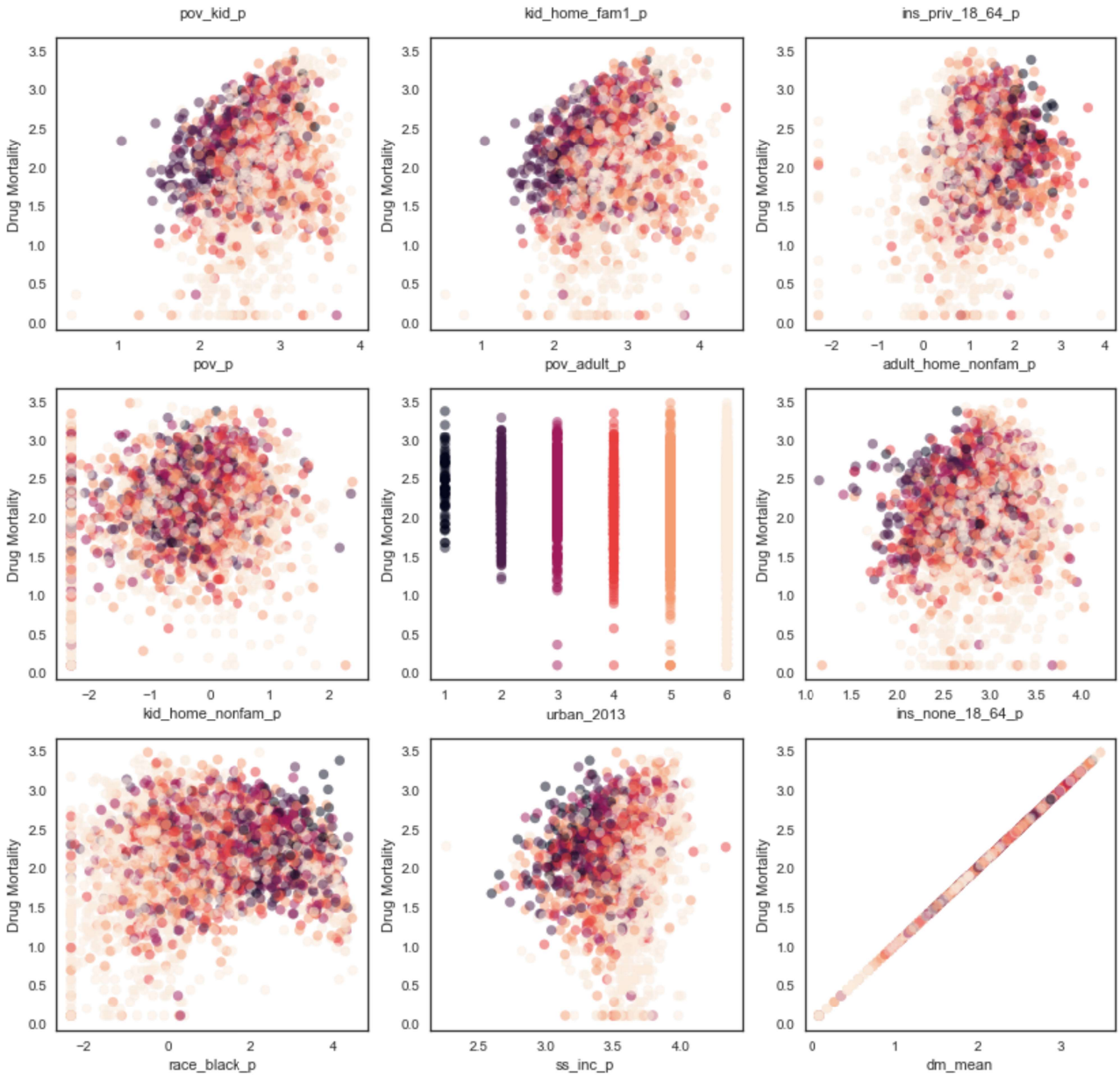


In the box plot below we can see differences in the distribution of mean drug mortality according to level of urbanization. While the mean does seem to vary in a not-quite-linear way across categories, what is most noticeable is the much greater variance of drug mortality among rural counties. Rural counties are home to both the highest and lowest mean drug mortality rates in the US. T



The scatterplots below, which depict the top 20 correlates of drug mortality, have the urbanization code added to the color axis of the plot (1=city/dark thru 6=rural/light), which reveals some interesting patterns and possible interactions. For example, for the features measuring disability status, poverty, and insurance status, it appears that there is a higher baseline for drug mortality in urban counties and possibly a stronger connection between these variables in urban counties as compared to rural counties. For the percent Black in a county's population, we see a possibly non-linear relationship with drug mortality in which urban counties tend to have higher Black populations and also higher (and less variable) drug mortality rates. These interactions suggest that linear models may not be able to fit this data well.


Drug Mortality
unemp_p
ins_pub_18_64_p
totpop
ssi_inc_p
pub_inc_p
labor_force_p
marry_past_p
disability_p
earn_inc_p
pov_kid_p
kid_home_fam1_p
ins_priv_18_64_p


pov_kid_p
kid_home_fam1_p
ins_priv_18_64_p
Drug Mortality
pov_p
pov_adult_p
adult_home_nonfam_p
kid_home_nonfam_p
urban_2013
ins_none_18_64_p
race_black_p
ss_inc_p
dm_mean

4. Data Modelling¶

After preparation and exploration, the data were used to build a model to predict county-level drug mortality rates from data on county economic and social characteristics.

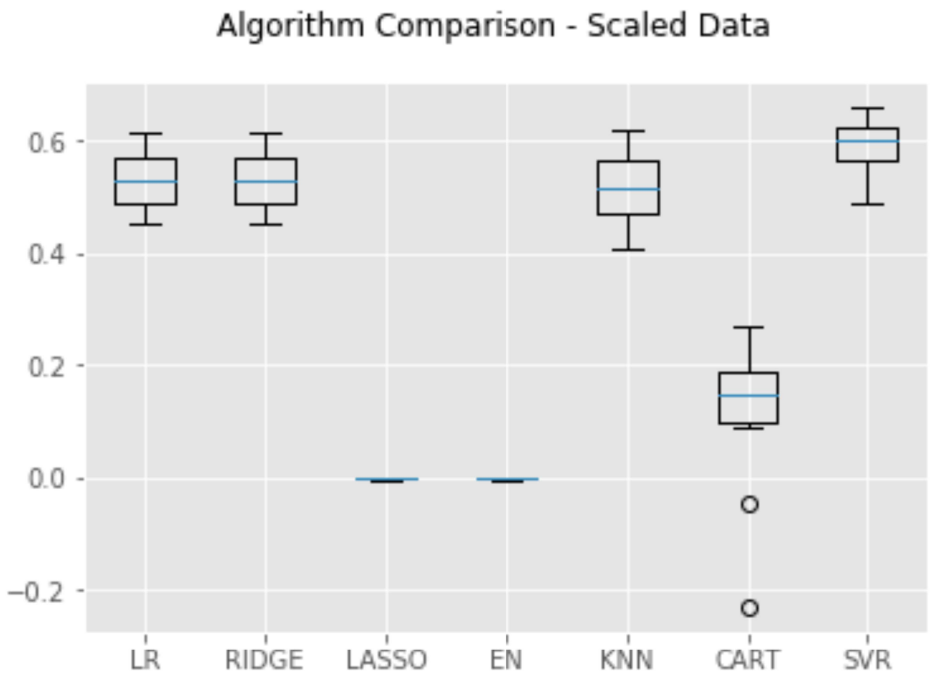
4.1 Prepare Data¶

First, all 46 features were log-transformed and several regression algorithms from Scikit-learn were tested for suitability using default hyperparameters. A Scikit-learn pipeline was built to standardize all the numeric variables and to one-hot encode the measure of urbanization (a 1-6 scale that can be treated as nominal or ordinal).

4.2 Evaluate several model types¶

Several regression techniques were evaluated using default parameters:

- linear regression
- ridge regression
- LASSO regression
- elastic net regression
- K nearest neighbors regression
- decision tree regression



Models were scored using R-square. Support vector regression, K nearest neighbors, and linear or ridge regression appear to be the most promising models. Next I will tune the KNN and SVR models to see if they can be further improved.

4.3 Tune most promising ML algorithms¶

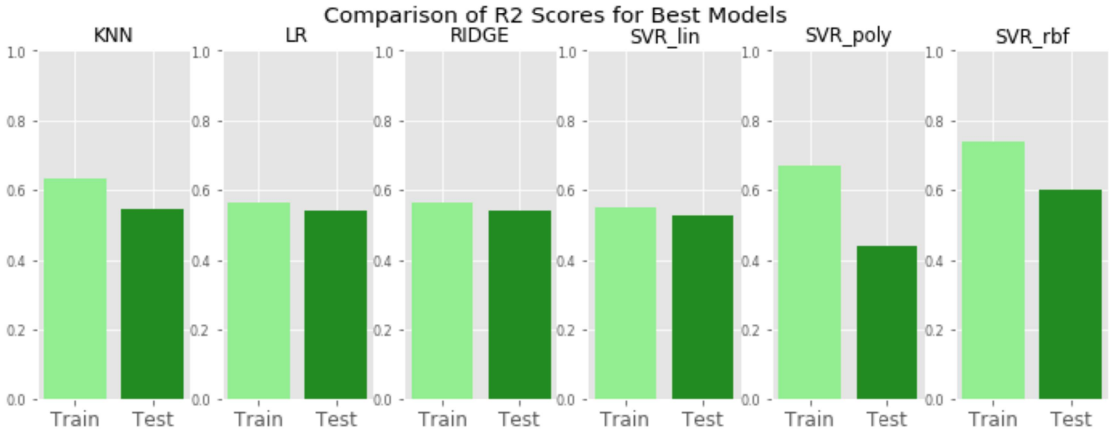
The next step is to tune the KNN and SVR models to see if they can be further improved. The KNN model was tuned for the number of neighbors. The SVR model was tested using 3 different kernels (rbf, linear, and polynomial) and the C and gamma hyperparameters as appropriate. The highest scoring models resulting from the grid searches were:

Model	R-square Score	Hyperparameters
KNN Regression	0.525	n_neighbors=9
SVR – rbf kernel	0.604	C=5.0, gamma=0.005
SVR – linear kernel	0.508	C=0.1
SVR – polynomial kernel	0.535	C=0.1, degree=2, gamma=0.05

The highest R2 score among all tested models is 0.604, from the SVR model with an rbf kernel using C=1.0 and gamma = 0.005. Several models yield similar results, however, so I will apply the best of each model type to the test set for the final determination.

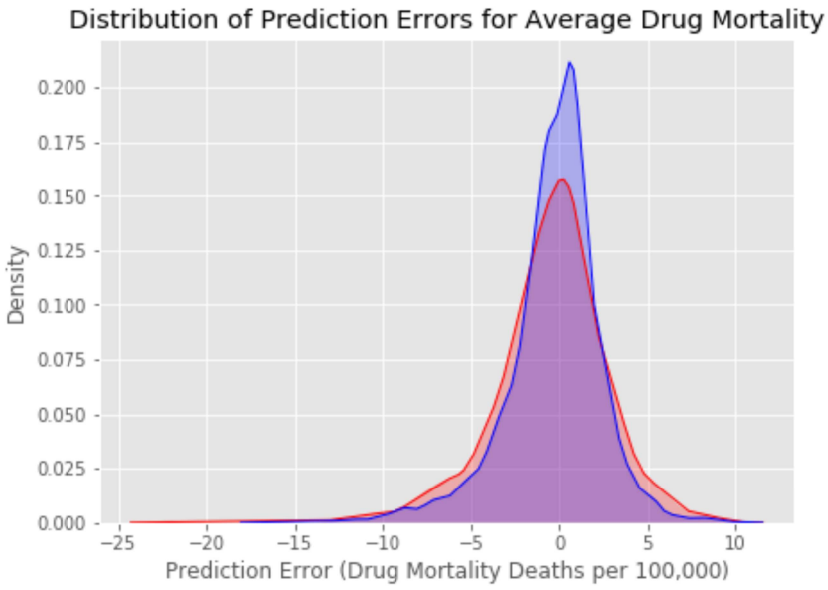
4.4 Validate best models against hold-out data¶

This chart demonstrates that the support vector regression model using the radial basis function kernel produces the highest R-square scores. And although the test score is noticeably lower than that for the training data, the R2 score for the test data using the SVR_rbf model still outperforms the other models.



Now that we have our best-fit model, we can examine the errors it produces to look for insight as to how to improve the model.

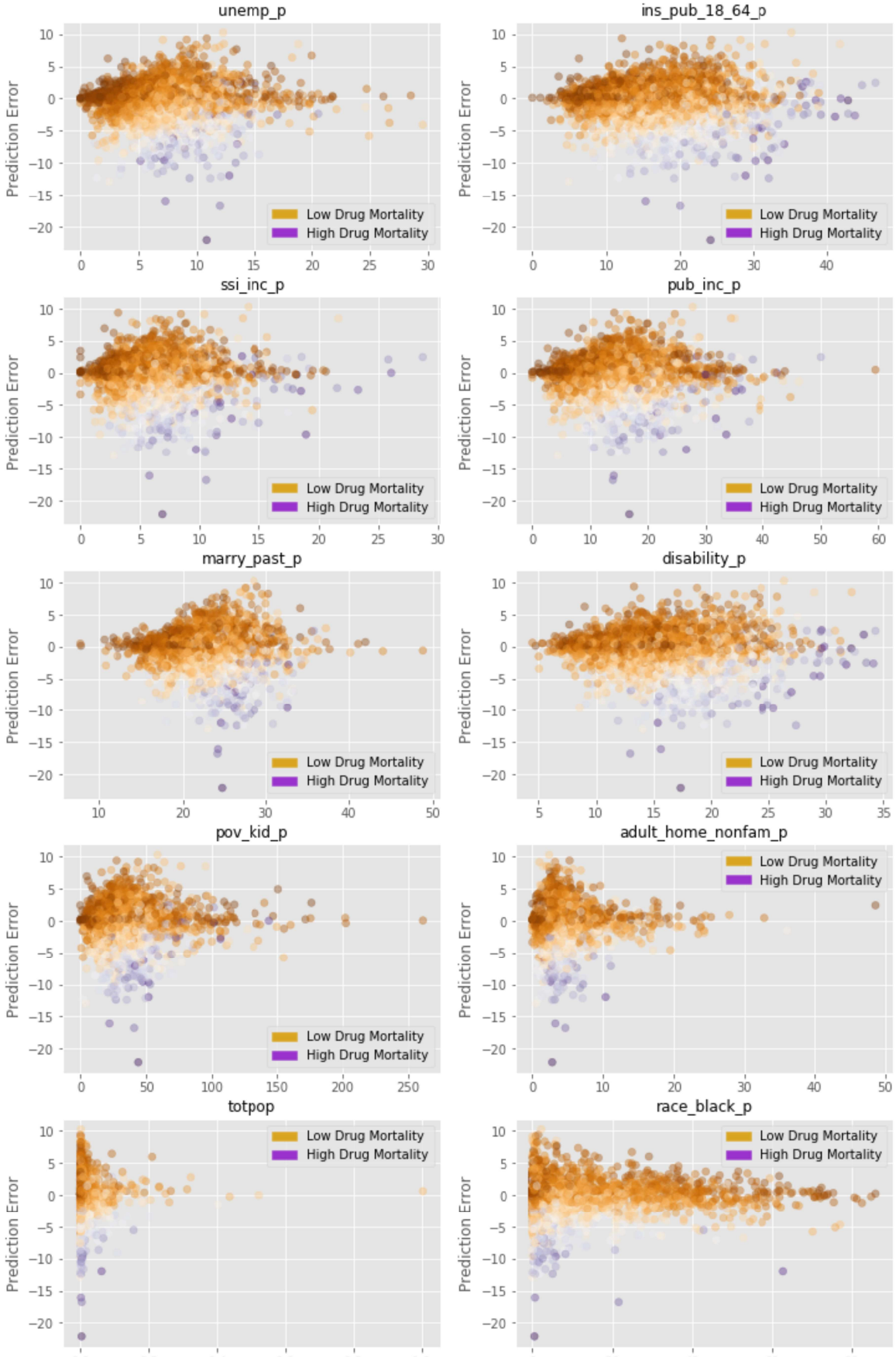
The density plot below shows that the prediction errors for both the train and test sets are skewed to the left. This means there are a few counties for which the model fails to predict high levels of average drug mortality. We can identify those counties and explore some features.



Counties with Highest Prediction Error

drug_mort_pred	drug_mort_actual	pred_error	county	State	urban_2013
10.440407	32.544444	-22.104037	Rio Arriba County, NM	New Mexico	5.0
10.218112	26.988889	-16.770777	St. Bernard Parish, LA	Louisiana	2.0
11.911121	27.988889	-16.077767	Carbon County, UT	Utah	5.0
2.487417	15.433333	-12.945917	Daniels County, MT	Montana	6.0
11.026273	23.433333	-12.407061	Starke County, IN	Indiana	6.0
12.938802	25.322222	-12.383420	Summers County, WV	West Virginia	6.0
5.085411	17.433333	-12.347922	Baylor County, TX	Texas	6.0
10.169477	22.322222	-12.152745	Mineral County, MT	Montana	6.0
17.343736	29.322222	-11.978486	Baltimore city, MD	Maryland	1.0
7.184044	18.988889	-11.804845	Bear Lake County, ID	Idaho	6.0
11.333568	22.988889	-11.655321	Campbell County, KY	Kentucky	2.0
12.272202	23.544444	-11.272242	Pulaski County, VA	Virginia	4.0
8.373072	19.322222	-10.949150	Eddy County, NM	New Mexico	5.0
10.866047	21.766667	-10.900620	Murray County, OK	Oklahoma	6.0
7.077401	17.766667	-10.689265	Young County, TX	Texas	6.0
14.427323	24.988889	-10.561566	Cherokee County, NC	North Carolina	6.0
12.526951	22.988889	-10.461938	Scott County, IN	Indiana	2.0
11.331458	21.766667	-10.435209	Pawnee County, OK	Oklahoma	3.0
9.489536	19.766667	-10.277130	Brooke County, WV	West Virginia	4.0
11.937496	22.211111	-10.273615	Clark County, KY	Kentucky	3.0

Prediction Error of Model with Actual Average Drug Mortality and Selected Features in US Counties



The scatterplots above map the model's prediction error against several predictive features. There do not appear to be any relationships between these features and the prediction error, because they have been accounted for in the model. The orange counties indicate low actual average drug mortality rates while the purple represent high mortality. For many of the features, the underestimated counties (purple) are spread evenly across the x-axis. For others though—such as total population, child poverty, adults in non-family homes, and percent Black—the underestimated counties do seem to be located in a narrow range along the x-axis. But many other counties are also in these bands, and the model cannot distinguish them. Additional features not currently present in the model are probably needed to achieve better predictions.

5. Discussion¶

This data science project has attempted to build a model to predict average drug mortality rates for US counties using a variety of socioeconomic features. After testing and tuning several machine learning algorithms, the best fitting model is a nonlinear support vector regression model. The R-square for this model is 0.60. R-square is sometimes interpreted as the percent of the variance of the outcome variable that can be explained by the input features (or independent variables). Thus this model 'explains' somewhat more than half of the variation across counties in average drug mortality rates. Obviously we would like to achieve a greater degree of accuracy, but this is actually a fairly good result for social science research.

Nonlinear relationships were observed in the exploratory analysis, so it not surprising that an rbf kernel provided a better fit than any of the linear models. Unfortunately, the support vector machine model operates as a 'black box' and cannot provide any information about relative feature importance. For this use case we would ideally like to be able to describe the socioeconomic factors most associated with higher drug mortality rates. Understanding these associations can be helpful in planning for future placement and allocation of resources to fight the opioid epidemic. Our best clues come from the bivariate analysis, which suggested an association between higher drug mortality rates and indicators of lower economic well-being, such as unemployment, reliance on public assistance and public insurance, poverty, and disability rates. These findings can be used to provide guidance to policy makers in directing resources towards communities most likely to experience significant levels of drug mortality.

This model could also be used to generate predicted drug mortality rates for later years by exposing the model to updated socioeconomic data. But one of the shortcomings of the model is that it fails to predict some of the highest rates of drug mortality that have been seen in the US. It may be possible to improve the model by adding additional features. Some features suggested by previous research that might be tried in future modeling attempts include:

- housing data such as median housing values and housing tenure data
- commercial activity or employment data, broken down by industry
- death rates from other despair-related causes, such as suicide
- commuting patterns
- availability of addiction recovery and support services

Recommendations

While the opioid epidemic is certainly a national crisis, it is not equally felt in all areas of the country. In fact the distribution of overdose deaths is highly skewed, with some places experiencing exponential increases and others barely impacted.

Efforts to intervene in the opioid epidemic should be tailored to local conditions. For example, the US Congress passed a law in late 2018 that provides resources for a variety of programs, which are currently awaiting appropriations. Predictive modeling can assist with planning for effective interventions:

- Addiction recovery and physician/first-responder training are most needed in counties with high predicted mortality.
- Locales with less critical levels of addiction can be targeted with education and outreach programs to help prevent the spread of the epidemic.
- Communities with high levels of poverty, low employment, and high reliance on public assistance including medical insurance are disparately impacted by this crisis. Efforts to address these complicating (and possibly causal) factors in these communities may help to stem the growing death rates by reducing addiction rates.