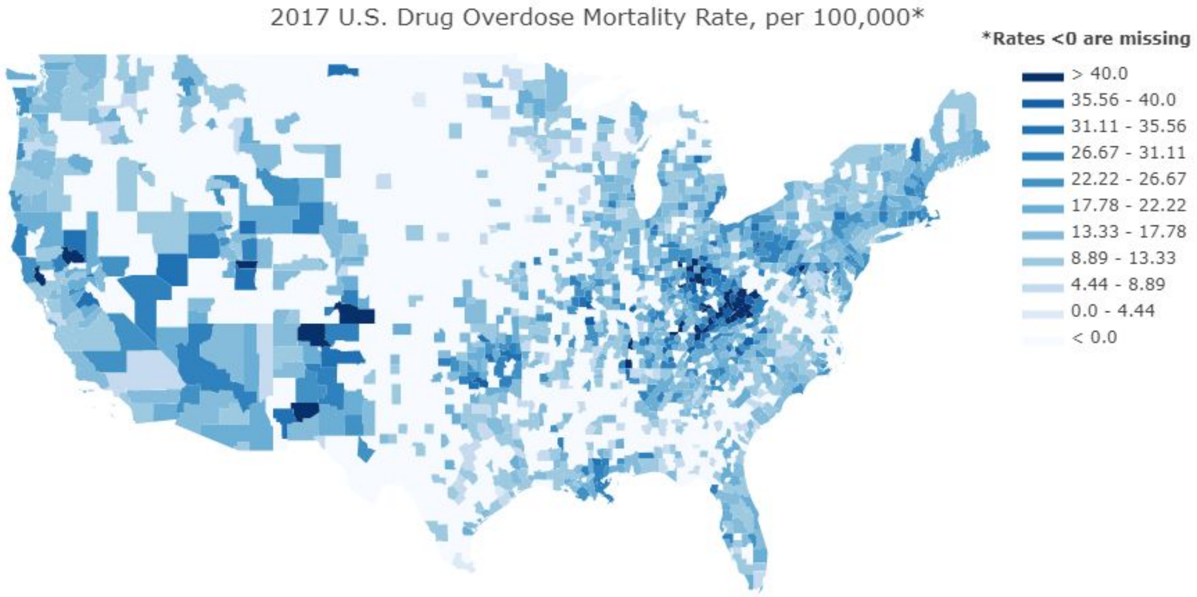
Capstone 1 EDA: Valery Lynn

Title: Predicting drug overdose mortality rates by county level in the U.S.

The purpose of this study is to determine what key variables play a role in predicting drug overdose mortality rates, and how those may vary across counties in the U.S. The dataset contains one response variable - drug overdose mortality rates, and 72 potential predictor variables. My first task was to visualize the incidence of drug overdose mortality by creating an interactive cloropleth map of the U.S. with counties shaded to represent drug overdose mortality rates.



It was obvious that a large portion of the map was unshaded due to missing values for drug overdose mortality rates. Upon inspection there was more than half of all counties with missing values. It was essential therefore to create a regression analysis to estimate (predict) the values for the missing counties. This study will answer two framing questions:

1. How can we best estimate missing drug overdose mortality rates using supervised machine learning algorithms?
2. What are the principal predictors for drug overdose mortality rates?

The best way to estimate a predicted quantity from a set of data is using some type of regression analysis. I started by considering a linear model because linear models can capture a great deal of data relationships. The goal of linear modeling is to use the simplest model that captures the most variation. My next task was to begin shrinking my data by finding only those predictors that have an influence on the response variable. To do this I looked at how each predictor correlated to the response variable. The Pearson-r correlation test assumes that variables are normally distributed. To get the most accurate results I needed to check the shapes of the distributions for each variable and use transformations to get them to a normal distribution if possible.

I started by examining the distributions of all the variables by plotting their kernel density curves and probability plots to check for normality. It appeared that many of them were normal or log-normal, including the response variable. Taking the log of a log-normal distribution leaves a normal distribution. I did this and ran the probability plots again to check for normality. Not all of the variables could be transformed this way, but there are other methods later that can be used.

I ran pearson-r correlation tests on the response vs all other variables and selected any that were +/- 0.3 or higher. This left 20 variables that demonstrated considerable relationships with the response variable. I will use these variables for the linear modeling.

It is important to note that it is possible to transform a response variable for linear modeling but it is usually better not to. It greatly reduces the ability to make predictions because there is no straightforward way to scale back to the original (back-transformed) values that can be interpreted on the original scale. You cannot compare regression coefficients in models that have transformations performed on the response variable(Faraway, 2014).

I will use the 20 identified predictors and the response variable in their non-transformed states to being linear modeling. The next stage will be to perform variable tests of the model, including PCA/PCR to determine the best model for prediction. I will train the final model on this data and use it to predict the missing values and create a complete map of overdose rates. This model will also provide insights into which variables are most important in predicting overdose rates.

Finally, I will use an unsupervised learning model to group the counties to determine if there are differences of interest between the groups. This will also shed light on the strongest indicators for increased drug mortality rates.

References

FARAWAY, J. J. (2014). *Linear Models With R, Second Edition*. Taylor & Francis.