Drug Abuse EDA

Jillian Warburton and Mary Ebbert

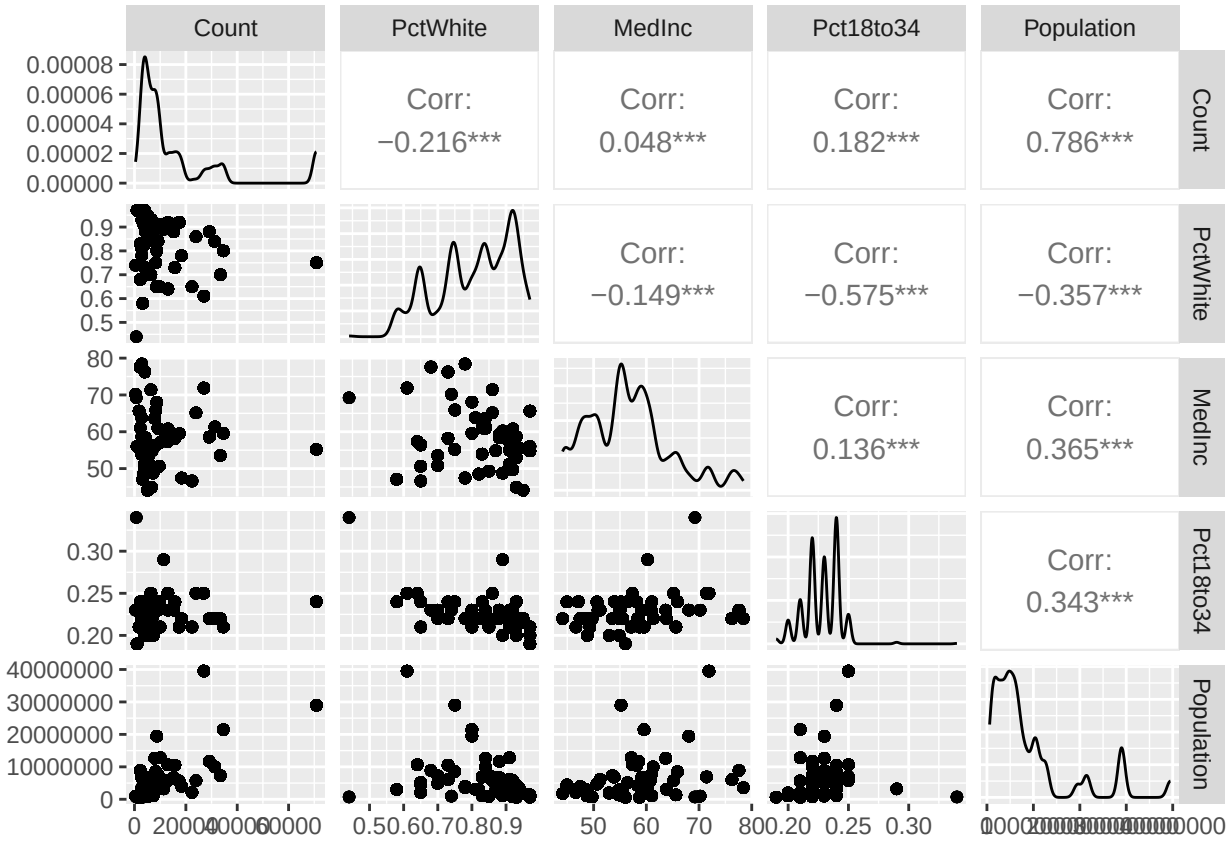
2023-03-31

Purpose of Analysis

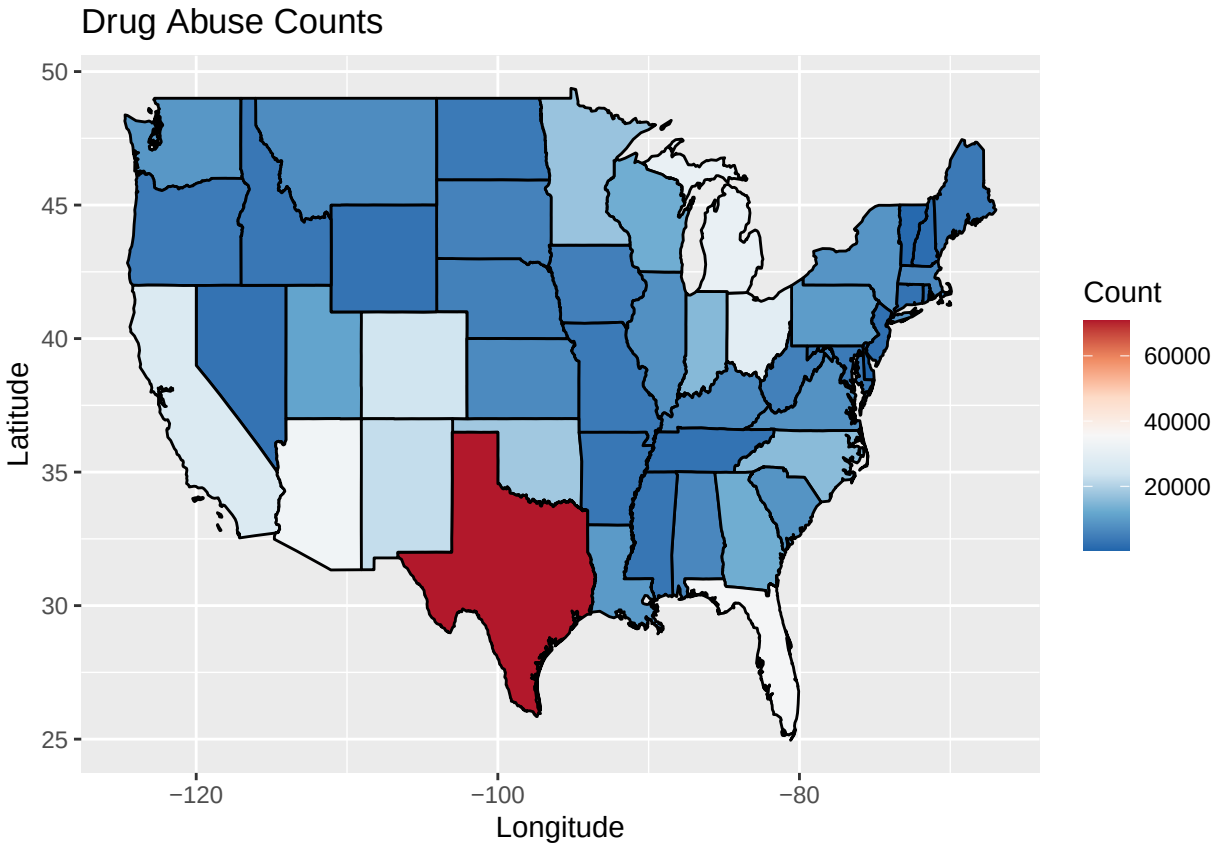
Oxycodone is a Schedule II controlled substance with high abuse potential, due to its status as a narcotic and the potential for overdose death. Our data set contains state names, the population count of the state, the number of drug abuse cases in the state, the percent of the state's population who are between the ages of 18 and 34, the state median income, and the percent of the state's population who are white/Caucasian. We want to know which population demographics are most prone to oxycodone abuse and which states are outliers for abuse rates, after accounting for demographics. Analyzing this data set will allow us to determine which states would most benefit from targeted anti-drug abuse interventions.

Main Features and Patterns

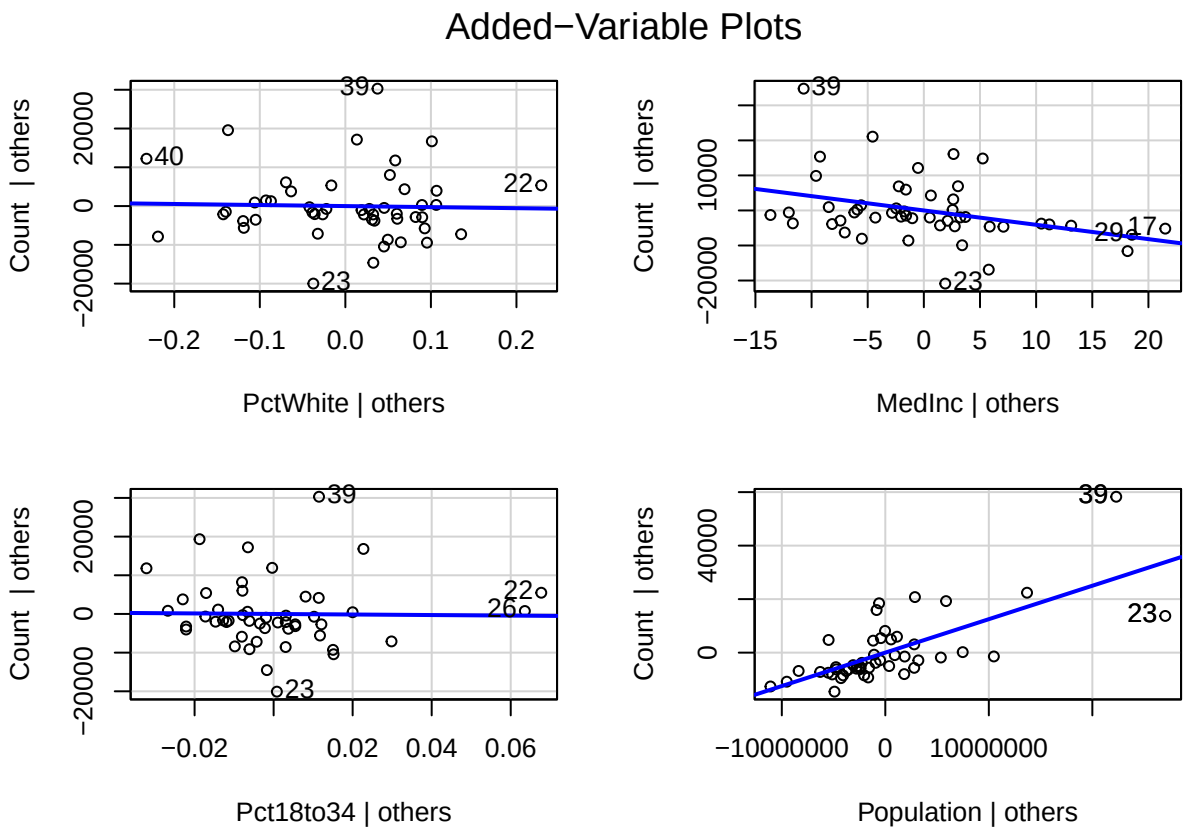
Of our 6 variables in this data set, 1 variable is the nominal and categorical variable of States, while the other 5 variables are the numeric variables of percentages and counts. Below shows a pairs plot of the four numeric variables. We can see that there is a moderate, negative correlation of -0.575 between the percent of the state's population who are between the ages of 18 and 34 and the percent of the state's population who are white/Caucasian. There is also a strong, positive correlation of 0.786 between the population of a state and the drug abuse case counts. The data may require transformations to normalize the data or reduce skew. There is also one outlier in drug abuse count that may affect modeling.



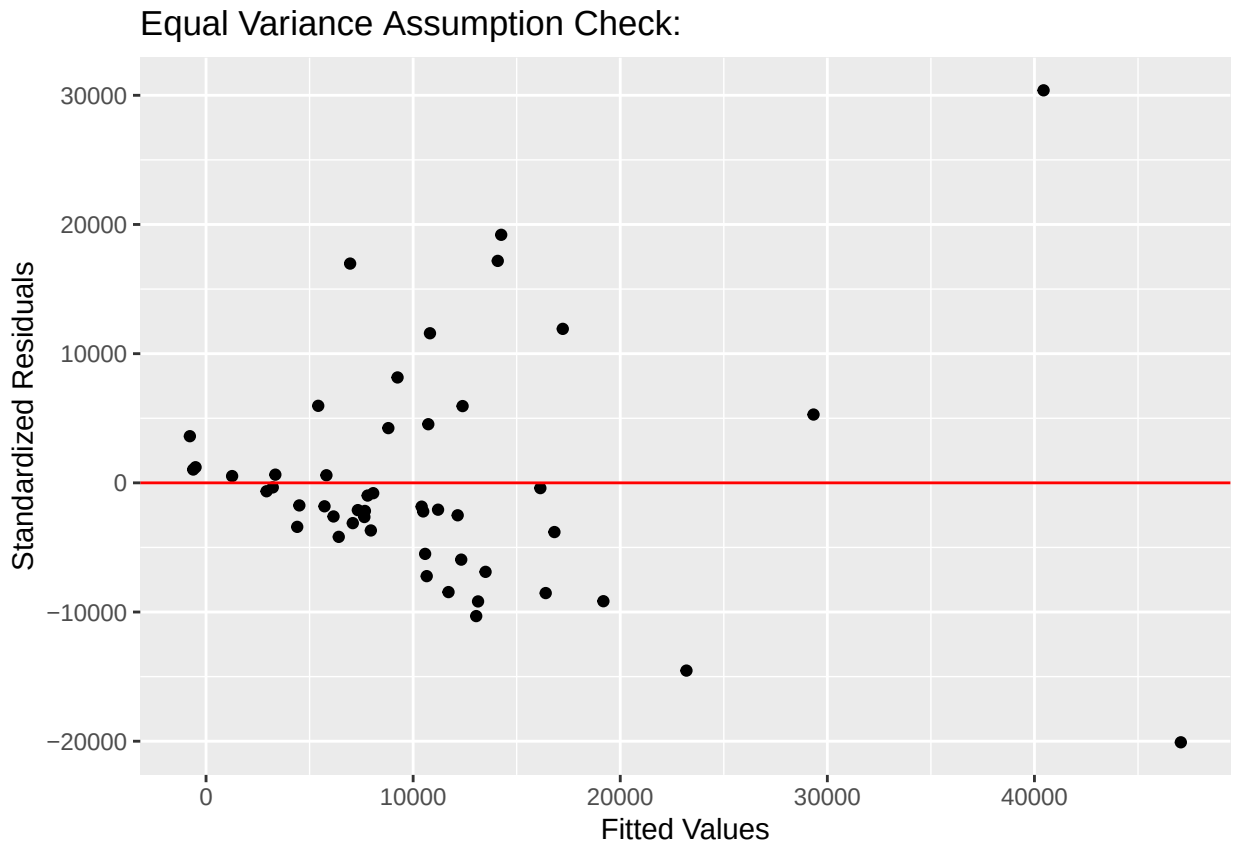
Below is a choropleth map showing which states have which counts of drug abuse. It proves there is at least one outlier in the data set.



Below are linearity plots showing that linearity is okay for this data set.



Finally, below is graph showing the difference between the fitted values and the residuals of a basic linear model, showing that the data set is heteroskedastic. The data's variance will need to be accounted for.



Potential Sources of Correlation Between Observations

The data set is based on drug abuse case counts for states, a spatial location with sets of data related to it, which is areal data. There may also be correlation between other variables, but that will require testing for collinearity.

Potentially Appropriate Statistical Models

With the lack of normality and the heteroskedasticity, a multivariate linear model may be appropriate, after transforming the drug abuse case counts and adjusting for the variance.

Next Steps for Analysis

We need to learn how to deal with areal data. We also need to account for the lack of independence, normality, and heteroskedasticity in this areal dataset.