Data Set Description

In order to construct a more meaningful model, we decided that we would need to obtain more data beyond the information contained in the initial data set described in the project proposal. By combining a series of data sets on Country and Year, we now have one large set with numerous features from which we can construct a model. Our combined dataset currently contains the following variables:

- Features:
    - Country (multiple sources): 136 included in aggregated dataset
    - Region and Sub Region (source: United Nations Office on Drug and Crimes)
    - Kilograms of Drugs Seized for 11 different Drug Groups and Total (source: United Nations Office on Drug and Crimes)
    - GDP per capita (source: World Bank)
    - Population (source: World Bank)
    - Year (multiple sources): 1990-2016
    - Type of government (multiple sources)
        - Data for 30 European countries' laws regarding drug possession, use, and supply punishments as of 2015.
- Outcome Variables:
    - Direct Deaths from Drug Use Disorders/ Overdoses (source: OurWorldinData)
        - Split by alcohol and illicit drugs (we are using the sum of the two)
    - Share of disease burden attributed to substance use disorders (source: OurWorldinData)
        - Measured in Disability-Adjusted Life Years (DALYs) which considers both death rate and years lived without a disability/ health burden

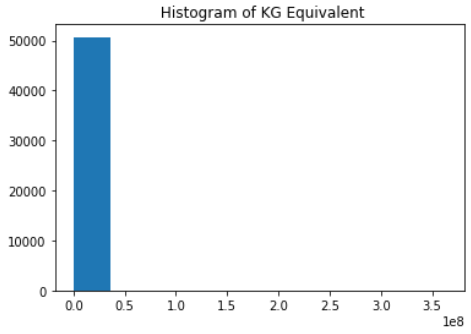
Due to merging multiple data sets, there are a few observations that contain missing values in one or more features. However, there are only 136 total observations out of 3672 in the training set that contain NA values (3.7%), so our best option might just be to remove them.

Testing Model Effectiveness

There are a few factors that we need to be cognizant of in order to prevent overfitting. Creating a dummy variable for each country would lead to over 130 dummy variables, so we might want to pare down the number of countries to those which have a more detailed drug policy, or just replace countries with a coded drug policy variable. This would have the added benefit of making the model more interpretable. To ensure that we are not overfitting or underfitting our models, as well as to establish model accuracy, we will check the training and testing error using mean squared error.

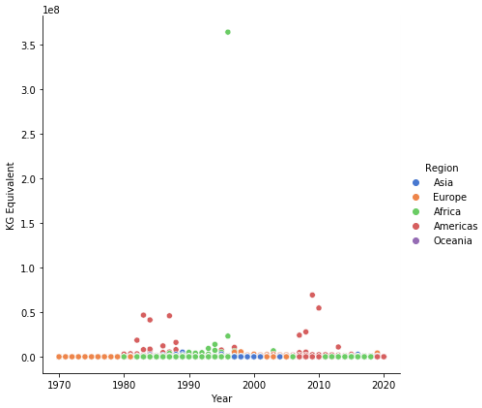
Exploratory Data Analysis

While plotting our variables, we noticed that the histogram from drug seizures (KG Equivalent) had very odd bounds. We then plotted Year by KG Equivalent, separated by Region, to see if we could get a better idea of



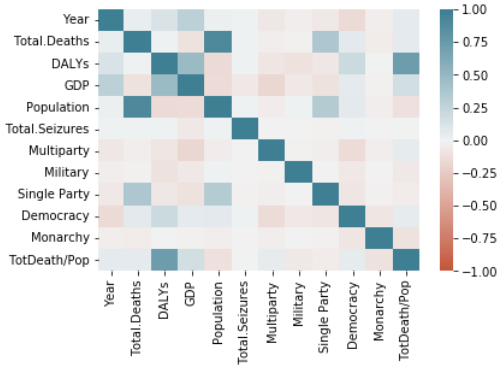
what was causing the bounds to be as wide as they were. The plot shows that there is an extreme value in the Africa region in the year 1996. To fix this, we replaced the value with the average of all of the other values with the same features (Poppy plants in Egypt). We also checked online to see if there was a massive opioid drug bust in Egypt in 1996, but found no record of such an occurrence, so we can assume that this is most likely an error.

We hypothesized that the wide bounds might be due to the fact that certain drugs are seized more often than others, leading to a skewed distribution overall. To ameliorate this, we looked at histograms for each Drug Group's seizure amounts. These still have very skewed distributions. Applying a base 10 log transformation with a small addition to account for 0 value does not help the skewness very much. This is probably because a large portion of the data is 0 (no drugs of that type were seized during that year in the specific country). We will have to keep this issue in mind as we continue working on the project.



Feature Transformations

We encoded the Type of Government feature using a one-hot encoding. We are tentatively considering transforming GDP, population, total deaths, total seizures, and DALYs using a log transformation to obtain more normal distributions. Lastly, we will consider incorporating a lag effect with Seizures, under the hypothesis that high seizures in one year will dissuade people from using drugs in the following year. Further analysis with future models will help us make these decisions.



Correlation Matrix

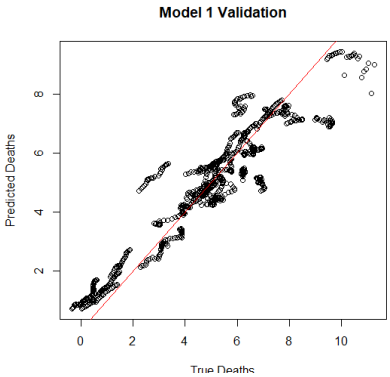
It appears that most of the features are not highly correlated with each other. DALYs has a slight positive correlation with GDP. Government type does not seem to have any strong relationships with any of the outcome variables so we will not include it in our initial models. Total Seizures has particularly very little correlation with any of the other variables, and as a result will likely be the first feature we drop from our models efforts to prevent overfitting.

Preliminary Supervised Models

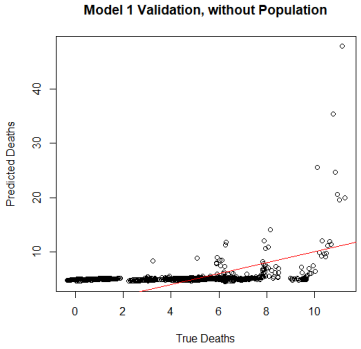
Model 1: log(Total Deaths) ~ log(Population) + log(GDP) + Drug Seizures

After reviewing the data and assessing which features might be important to include in our final model, we ran a preliminary linear regression with select features on log Total Deaths. While we

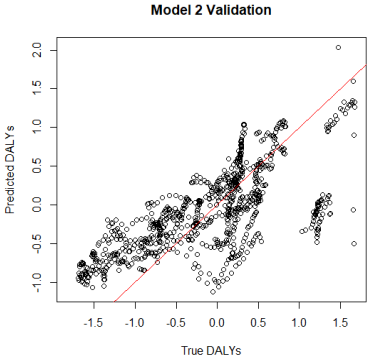
will probably want to create a time-series model using our time variable, we initially wanted to get a very basic overview of the linear effects of the features at our disposal.



We then used this model to make predictions on the validation set. As can be seen in the included figure of Model 1 Validation, the model is somewhat accurate (the red line is y=x). However, when we remove Population as a factor, the model is no longer accurate, which means that most, if not all, of the predictive power of the first model is coming from Population (figure Model 1 Validation, without Population). This makes sense, because as the population of a country increases, the total number of deaths would likely increase as well. Due to this, we posit that Total Deaths is not a viable outcome variable to use to assess the predictor effects, and we will instead use Drug Burden (DALYs) as our outcome variable.



Model 2: log(DALYs) ~ log(GDP) + log(Population) + Drugs Seized
Next, we ran a preliminary linear regression of Year, log(GDP), log(Population), and each of the drugs seized values on log(DALYs). This appears to make better predictions, as shown in the included figure of Model 2 Validation.



Exploring Effect of Country's Drug Possession Laws:
Initially, we wanted to run a preliminary analysis on the possible effect of specific drug policies as a feature. However, because we have to hand-code the policies for all of the countries, we decided instead to hand-code 30 countries and build a basic model to investigate the policy effect. Due to the fact that this data subset is small, we also opted not to validate the model, but just note the significance of the beta values for each policy category as a very general measure of feature importance. It appears that most of the policy categories are significant in predicting Drug Burden (DALYs) variation when GDP, population and Drug Seizures are included, so we will finish hand-coding all of the countries (for which we have policy data) and add that feature to our total model.

Future Work
We plan on finding and incorporating more features into our model, as well as expanding our model complexity beyond basic linear regression. Our dataset "EuropeanDrugLaws" contains columns indicating if a country has an alternative option to receiving jail time, such as entering a rehabilitation center. This information could be included in the form of an ordinal or binary feature. We also have obtained a dataset on mental health that contains data separated by year and country, which might provide insight to what contributes to a country's potential drug problem. We would like to include information about the proportion of people in prison who have been convicted of drug-related offenses. Additionally, we are anticipating building some form of a time series model to better utilize the Year data.