Capstone Project 1: Milestone Report - Valery Lynn

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

Problem: According to data collected by the Centers for Disease Control and Prevention (CDC) there were more than 63,000 drug overdose deaths in 2016. More than 66% of those involved an opioid. On October 26, 2017 the opioid crisis was officially declared a national Public Health Emergency under federal law. The economic burden of prescription opioid misuse is estimated by the CDC to be more than $78 billion a year.

County level services such as hospitals, crisis centers, and local planning boards are in need of predictive models to inform planning, preparation, and resource allocation. This model can be used to better estimate the needs of the county to address this crisis. Examples include how many overdose kits hospitals need to have on stock, how many full-time crisis prevention professionals to employ, how many drug rehabilitation centers need to be established or funded, etc. Local employers are also stakeholders in this crisis. Reports suggest that companies in regions of high opioid usage are unable to maintain employees that can pass prerequisite drug tests for employment. This model can be used as a trends indicator for able local workers.

Knowing what key variables play a role in predicting drug overdoses, and how those may vary across counties having different rates of overdose, can help counties invest wisely in therapeutic resources. Examples would be that a set of counties having particular predicted conditions would require a different set of responses than a set of counties with another set of predicted conditions, and a model could optimize (or customize) those responses. An example of this would be that counties with high overdose rates would see better results by increasing the number of mental health doctors while counties with moderate overdose rates would see better results by encouraging more job growth. Or, that the coefficients of the same set of variables would differ, requiring differing intensities of responses. For example, a new group home vs a new mental hospital. I will explore different machine learning algorithms to attend to different planning needs.

The above accounts for responsive planning and action, but a model like this can be useful for preventative action as well. Policy decisions can be informed by models such as this where key predictors can point to strategic areas regarding investment in infrastructure for education, social services, and economic growth, should those sectors prove to have a positive outcome on reducing drug overdose mortality.

Finally, nearly half of the counties in the U.S. have missing data for drug overdose mortality in the publicly available databases. This gives an incomplete picture of the crisis at a national level. This model will be used to estimate those missing values for a more accurate (complete) visualization.

Acquiring Data

County level data of drug overdose mortality rates and social, economic, and demographic indicators are available for this model from the The County Health Rankings dataset, a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute.This database was built predominantly from the following: The Behavioral Risk Factor Surveillance System (BRFSS), the National Center for Health Statistics, and the CDC WONDER mortality data.

The database for the rankings is available for downloading as an Excel spreadsheet at: http://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation

For a full account of how data for each variable was attained see: http://www.countyhealthrankings.org/sites/default/files/resources/2017_Measures_DataSourcesYears.pdf

Variables are all reported as percentages, rates per 100,000, ratios, or similar population adjusted measures. After data cleaning, they are all of numerical type float or integer. Output data will be in the form of a float for regression analyses, and character for classification and clustering algorithms.

The following is a table with sample input and output variables and their selected summary statistics.

Title: Sample input and output variables with summary statistics.

Variable Name	Description	Min Value	Max Value	Mean Value
Output: Drug Overdose Mortality	Deaths per 100,000 population.	3	93	18.16
Input: Mentally Unhealthy Days	**Health Outcomes:** average number of mentally unhealthy days reported in the past 30 days.	2.3	5.8	3.78
Input: High School Graduation Rate	**Economic Environment:** percentage of ninth-grade cohort that graduates in four years.	30	100	86.28
Input: Housing Problems	**Physical Environment:** percentage of households with at least 1 of 4 housing problems: overcrowding, high housing costs, or lack of kitchen or plumbing facilities.	1	62	14.47

This spreadsheet contained both raw and ranked data. I am interested in raw data for the analysis so I saved two tabs (Excel worksheets) of data as *.csv files: ‘Ranked Measure Data’ and ‘Additional Measure Data’. These needed to be joined using Pandas merge function. Before joining the dataframes, I removed columns from the csv files that contained counts (rather than rates), confidence intervals, and any other columns containing calculated measures.

The data were contained in two separate csv files. These were merged with an 'inner join' method. While I could perform the join on one column, 'FIPS' (the Federal Information Processing Standards code (FIPS) which uniquely identifies counties and county equivalents in the United States), I also joined on 'State' and 'County' names in the event that the county code was erroneously entered. This assured that the join would be correct. I chose an inner join because I wanted to avoid missing data as much as possible and I didn’t want any duplication in the ‘State’ and ‘County’ columns.

The dataframe with additional measures initially had 3136 rows and 38 columns. Each row is a county in the U.S. The dataframe with rank measures initially had 3136 rows and 39 columns. After the inner join, the dataframe had 3134 rows and 74 columns. The reduction of 2 rows was due to discrepancies between the dataframes (so they were dropped in the join) and the reduction of 3 columns took into account that the join was executed on 3 columns. This left values reported as rates, percentages or ratios as functions of county population. The predicted variable (y-variable) is drug overdose mortality as a rate (per 100,000). Approximately half of the counties did not report this rate. I removed all the rows (counties) that had a missing values for drug mortality rate, the y-variable. These will be estimated by the model. After dropping all missing values for the y-variable, there were 1623 rows and 74 columns. Finally, I converted the ‘FIPS’ column to strings because the codes are categorical and cannot be treated as numeric.

There were four columns ('PCP Ratio', 'MHP Ratio', 'Other PCP Ratio', ‘Dentist Ratio’) that were reported as ratios in the format ####:##:## or #####:#. I did a string split on the first ':', then converted the columns from strings to numeric (float64).

I chose to impute missing values using state medians after an analysis of the data. More than ⅓ of the variables had differences between the mean and median greater than 1, with some differences that were quite large. This suggests that for some variables, the distributions are skewed, or not normal, and therefore the median is a more robust measure of center. I chose to impute with state medians rather than column medians (U.S. means) to give more predictive power to the model by preserving as much of the variability of the feature variables as possible.

To do this, I created a new dataframe grouping by state and calculating medians for each variable. This dataframe had 51 rows, one for each state. I created a dataframe with just the column of state names from the original dataframe (1623 rows, 1 column). I did an outer join of these two dataframes, creating a new dataframe with 1623 rows and 73 columns called df_meds. When I created the dataframe with state medians it ignored the ‘FIPS’ and ‘County’ columns as they were not numeric. These were later inserted back into the joined medians (df_meds) dataframe. I checked for missing values after this join with the following command:

```
df_filled[df_filled.isnull().any(axis=1)]
```

It returned four rows that still had missing values. This was because South Dakota (with only 3 counties included in the dataframe after the missing y-variable rows were dropped) had no values reported for one or more of the feature variables and therefore had no state medians for that variable(s). Likewise, the District of Columbia (1 row) had missing values. For those, I used column medians to fill missing data with U.S. means across all feature variables.

To impute the missing values in the original dataframe, I combined the original dataframe with the means dataframe with the following command:

```
df = df.combine_first(df_means)
```

I checked for missing values and there were none.

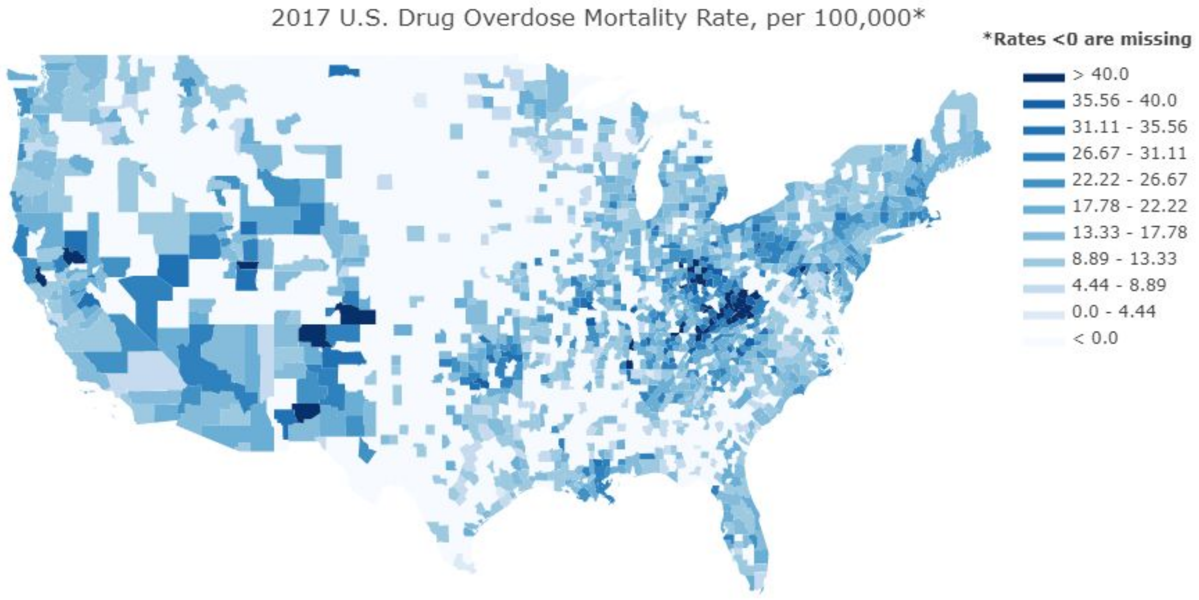
Finally, I plotted boxplots and probability plots to check for outliers. Boxplots provide good visuals for detecting outliers and probability plots check if outliers belong to the distribution. Many of the variables had outliers that were shown by the boxplots and the probability plots. In addition, the plots showed that not all of the distributions were normal. After performing a principal component analysis (PCA), many of these feature variables may be eliminated before fitting a model. It is reasonable to wait before addressing these outliers until a preliminary attempt has been made to fit a model. If there is difficulty fitting a model, the first thing I will attend to is removing any data points that appear to belong to a different distribution (severe outliers).

The Data Story

The dataframe was now complete and I sent it to a csv file for EDA and preprocessing. 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:

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.

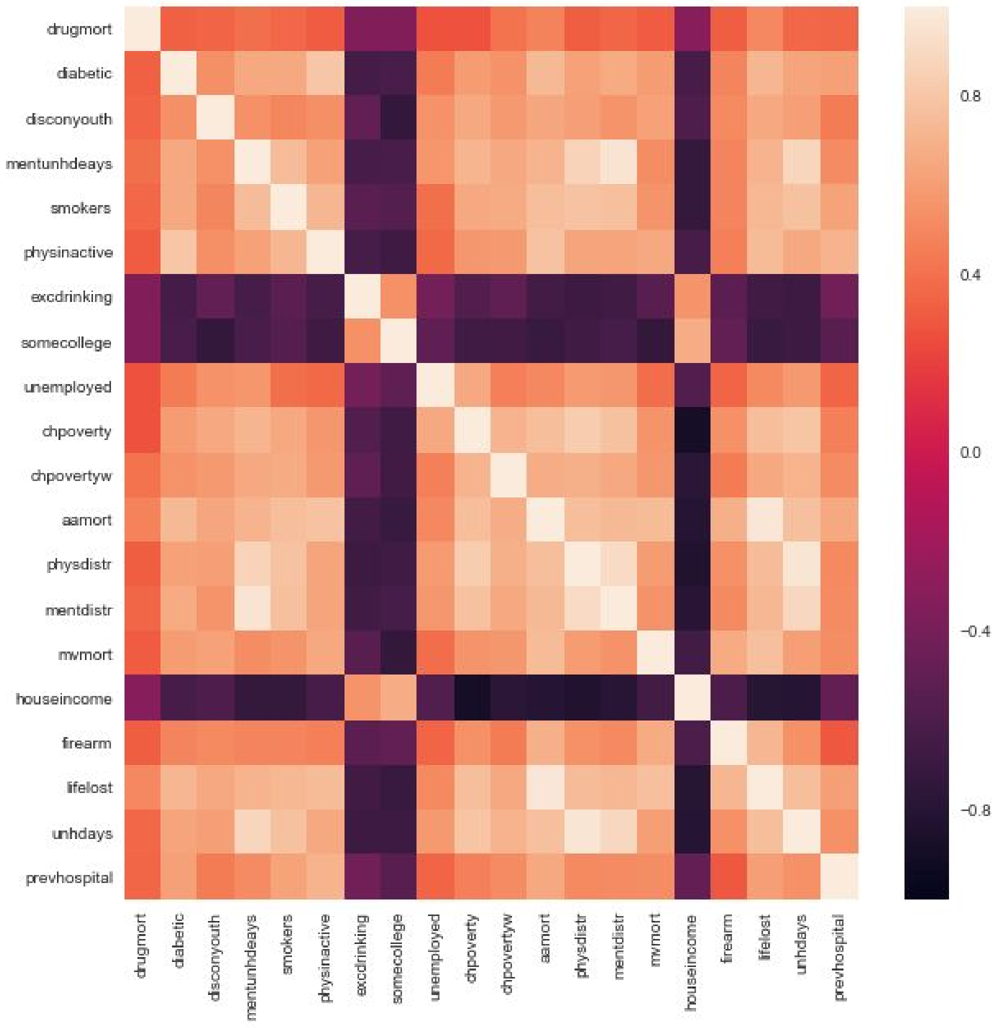
I started by examining the distributions of all the variables by plotting their kernel density curves and probability plots to check for normality. It appeared that many of them were normal or log-normal, including the response variable. Taking the log of a log-normal distribution leaves a normal distribution. 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. A description of these variables, along with their Pearson-r coefficients and p-values can be found in Table 1 of Appendix 1.

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).

Major Findings

The 19 variables correlated with drug overdose mortality begin to tell a story about counties as drug mortality rates rise. However, there are a few things to note before drawing conclusions. Several of these variables are likely collinear. Meaning that predictors are correlated to each other. For example, poverty rates will decrease as median income rises. It may not be necessary to include both of these as they are similar measures. Several variables are measures of physical health (diabetic, smokers, physinactive, physdistr) and are likely to be correlated. A test for collinearity will be conducted during preprocessing. Below is heatmap plot of the correlation matrix:



There are three predictors that are negatively correlated with drug overdose mortality: excessive drinking, household income, and some college education. Income is usually positively correlated to college attainment so it is not surprising that they would both have the same direction in the effect. I am curious to see that increased excessive drinking would be associated with fewer drug overdose deaths. I will explore this is greater detail during modeling. I am resisting the urge at this stage to put meaning on these findings until further statistical analysis has been conducted.

Next Steps

I will use the 20 identified predictors and the response variable in their non-transformed states to begin 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.

Appendix 1

Table 1. Variable names and descriptions with Pearson-r correlation coefficients and p-values.

Name	Description	Pearson-r	p-value
drugmort	Drug overdose mortality rate	1.0	0.0
diabetic	Percentage of population that is diabetic	0.3249	3.1650e-41
disconyouth	Percentage of teenagers and young adults between the ages of 16 and 24 who are neither working nor in school.	0.3662	1.1146e-52
mentunhdeays	Average number of reported mentally unhealthy days per month	0.4029	2.1010e-64

smokers	Percentage of adults that reported currently smoking	0.3915	1.2912e-60
physinactive	Percentage of adults that report no leisure-time physical activity	0.3168	3.4893e-39
excdrinking	Percentage of adults that report excessive drinking	-0.3687	1.8666e-53
somecollege	Percentage of adults age 25-44 with some post-secondary education	-0.3381	1.0696e-44
unemployed	Percentage of population ages 16+ unemployed and looking for work	0.3077	5.9276e-37
chpoverty	Percentage of children (under age 18) living in poverty	0.3030	7.8189e-36
chpovertyw	Percentage of white children (under age 18) living in poverty	0.4322	7.2575e-75
aamort	Premature age-adjusted mortality	0.5067	1.3829e-106
physdistr	Frequent physical distress (measured through the Behavioral Risk Factor Surveillance System)	0.3659	1.3131e-52
mentdistr	Frequent mental distress (measured through the Behavioral Risk Factor Surveillance System)	0.3711	3.6176e-54
mvmort	Motor vehicle crash mortality rate	0.3289	2.9953e-42
houseincome	Median household income	-0.3037	5.5081e-36
firearm	Firearm Fatalities Rate	0.3429	5.2499e-46
lifelost	Years of potential life lost rate	0.5277	4.4739e-117
unhdays	Average number of reported physically unhealthy days per month	0.3996	2.7472e-63

prevhospital	Preventable hospital stay rate (Discharges for ambulatory care sensitive conditions/Medicare enrollees * 1,000)	0.4291	1.0011e-73