A Regression Analysis of Socioeconomic Factors Impacting Opioid Deaths in U.S. counties

Carr, A., Euling, K., Handoko, A., Neo, C.Z.

Abstract

This project consists of a detailed analysis of factors potentially influencing the number of opioid-related deaths in US counties. The problem of opioid use in the United States is a well-documented phenomenon. The need for the use of analytic methods to investigate this issue is therefore apparent as information derived from such investigations is of great utility to government policymakers and strategic decision makers in the private sector. This project compares various socioeconomic factors at the county-level in the United States in terms of the predictive utility in identifying the level of opioid-related death using regression methods. The researchers found that the strongest predictors of opioid-related death were social rather than economic, including age, gender, and level of education, and that the proportion of males aged 50 and over was the largest contributor to opioid-related death in U.S. counties. These findings have implications for how government and private services seeking to tackle the problem of opioid-related death rate elect to target their resources.

Keywords: opioid-related death, socioeconomic factors, regression analysis

1 Problem statement

The United States has been experiencing a drug overdose epidemic since the early 1990s. The only cause of death exceeding that of drug overdose among 24–54-year-old people in the U.S. population for the year 2018 was cancer (McGranahan & Parker, 2021). The purpose of the analysis will be to establish which predictor variables are most useful in identifying those US counties that are most susceptible to opioid-related mortality issues. For this reason, the project is intended to produce a quantitative description of the relative explanatory strength of numerous variables using regression methods. It will also seek to produce a suitable regression model that could be used as a predictive tool. Its greater practical purpose would be to assist the decisions of policy makers making use of county-level data, particularly in the sphere of public health.

1.1 Research questions

Primary research question:

- Which socioeconomic factors are most predictive of opioid-related death in U.S. counties?

Secondary research questions:

- How useful are various categories of factors in the prediction of opioid-related deaths at the county level?
- What types of relationship do we more commonly see between opioid-related death and socioeconomic factors?
- How efficient are socioeconomic factors in the prediction of opioid-related death?
- Which factors should be controlled for when predicting the number of opioid-related deaths in a US county?

2 Business justification

The economic cost of opioid use disorder in the U.S. was estimated to be $1,021 billion (USD) for the year 2017, with fatal opioid overdose constituting $550 billion of that figure (Luo et al., 2021). Consequently, to state that there is a significant economic interest in reducing this cost would be to risk understatement. Policy makers in government and in the private healthcare sector must rely on analysts' interpretations of data, particularly geographic data, to inform their decisions when aiming to optimally distribute resources intended to tackle the social issue. In finding methods of resource allocation that are closer to optimal, vast sums of money could be saved or redistributed elsewhere. Further to budgeting, which is a financial aspect of the problem, such

information about opioid-related death and socioeconomic factors at the state level is also relevant to companies, NPOs and government departments seeking to reach vulnerable demographics with messaging (marketing) and the approaches they take to organizing their personnel and resources on the ground (operations).

From a financial standpoint, any investor in a healthcare service provider that focuses on drug-related health problems should be keen to know where investment in such services is liable to prove most profitable. Investment decisions could therefore be at least partially informed by the results of the research. From a marketing standpoint, any provider of such services would benefit from being informed of the relative likelihood of serious opioid-related problems as the basis for the targeting of their campaigns. From an operational standpoint, healthcare service providers must be informed about the potential for incidence of opioid-related death in their geographic region in order to make operational decisions. However, given that this is a question pertaining to human mortality, the private sectors' tangential benefit from any such research is overshadowed by the utility of such information to both federal and state government to inform public spending decisions relating to public health

More concretely, by understanding the factors that correlate with the abuse of opiates, a company or NPO that operates drug rehab facilities would be able to optimize their geographic positioning to ensure maximal coverage. Furthermore, a pharmaceutical company producing and distributing drugs such Naloxone (or equivalents), which is used to prevent opiate overdose, would be able to target their marketing strategy in regions where demand is liable to be highest. Such information as this project aims to provide could also be used by supply chain planners seeking to position and store drugs and equipment used in the treatment of opiate-related disorders such that they minimize holding costs and lead times.

In summary, there appear to be numerous direct business applications to information that aids the understanding of the relationship between the socioeconomic features of US counties and opiate-related death. It is entirely conceivable that numerous types of company within the overarching medical and pharmaceutical industries would be willing to commission such a study in the hope of gaining a competitive advantage.

3 Dataset description

3.1 Data sources and overview

The team originally started with the dataset found in Kaggle, which was a compiled dataset of opioid deaths which incorporated many features [1]. However, there were limitations to the dataset:

1. Its "Total" column did not tally with the "Heroin", "Other" and "Methadone" columns.
2. There were a considerable number of "NA" cells in all the columns.
3. Some columns were lacking what the research team deemed valuable data. For example, the team wanted to find out if age group separated between male and female would shed light on the frequency of opiate-related death.
4. The data definition provided in the Kaggle dataset did not explain what each column meant well enough. For example, it was not clear whether "GDP Total" meant raw data or adjusted for the associated year.
5. The dataset did not provide any data earlier than 2011.

Many of these limitations were due to the withholding of data for certain regions where there were very few relevant medical incidents for the relevant year, in this case opiate-related deaths, in order to protect the anonymity of the individuals affected, as is common practice with a great deal of medical data. Nonetheless, these restrictions imposed significant limitations on the analysis in the view of the researchers.

Consequently, the team elected to query the data from each different source and take on the task of compiling, merging, and cleaning the dataset instead of using the pre-prepared dataset. While this meant a significant amount of extra work, it allowed the team to have much greater control of the features included as well as permitting a more robust modelling process. The new source of data queried was for the period running 2009 to 2019, extending the time horizon of the research by two years, and thereby potentially improving the reliability of the results obtained.

For querying the data, the team had access to US Census data via an API as well as through direct downloads from the same source. Another option available to the research team was web scraping. The team was able to access an API since the US Census Bureau provides an API for the team to call. Using the package "TidyCensus" the team was able to query the datasets directly. However, the API calls only allow for querying one year at a time. Thus, the team had to automate the call for each year, iterating through the relevant years included in the study.

[1] https://www.kaggle.com/datasets/ryanandreweckberg/opioid-crisis-by-interpersonal-relationships

3.2 Data cleaning and preparation

The data compiling and cleaning involved joining the disparate datasets into one main dataset by using the county FIPS code. The team needed to make sure that the datasets were cleaned properly and in the right encoding. The major source of discrepancies was that the "County" column usually contained a string describing the relevant county and state (sometimes given in an abbreviated form). Additionally, the counties were sometimes called municipality/*municipio*. Thus, the main task was to make sure these were consistent across all the requested datasets before joining them together.

The table below illustrates the data sources as well as the scripts which pulled, compiled, and cleaned them. For screenshots of each of these datasets, refer to Appendix A.

Data	Source	Scripts involved
Age Group for Male and Female	American Community Survey (ACS)	AgeGroup.Rmd (query from census data, cleaning and compiling)
Education for Male and Female	American Community Survey (ACS)	Education.Rmd (query from census data, cleaning and compiling)
US GDP (chained to 2012 levels)	U.S. Bureau of Economic Analysis (BEA)	data_cleaning.ipynb
Income Data	U.S. Bureau of Economic Analysis (BEA)	Data_merge.ipynb
Employment, Unemployment and Employment Rate	US Census Data	Clean_Join_Data.Rmd (cleaning and joining dataset)
Opioid Dispensed Data	CDC	Clean_Join_Data.Rmd (cleaning and joining dataset)
Opioid Deaths	CDC Wonder	Clean_Join_Data.Rmd (cleaning and joining dataset)

Figure 1: Data sources

After the data were pulled and joined together, the team normalized the dataset by population (per 100,000). This meant that for each of the relevant features, the values were divided by population before multiplying them by 100,000. This was done because the team observed through preliminary model experiments that these unnormalized features, while contributing well to the model results, were not meaningful in predicting the dependent variable. It is trivial to note that a higher number of opiate-related deaths in a county is correlated with the county having a large population after all.

3.3 Key variables

The response variable (dependent variable) of the model is the total number of known deaths related to opioid use in each US county. As stated earlier, total deaths were also normalized by population per 100,000. Hence, the target variable was renamed to "Deaths_x100000" instead. Hence, in subsequent sections, "Total Deaths" would refer to the normalized deaths values.

A wide range of socio-economic predictor variables (independent variables) have been explored and analyzed, including the educational levels prevalent among different age groups and genders in each county, the number of unemployed people in each county, the unemployment rate of each county, the real GDP of each county, per capita income, and the rate of opiate dispensation in each county. When appropriate, these variables were also normalized per 100000.

3.4 Observations from exploratory data analysis

Firstly, the data for most of the columns appear to be of similar distribution and spread when compared to Total Deaths. The typical feature included in the dataset does not display a linear relationship with Total Deaths. This is illustrated in the figure below. This means that while linear regression is a good starting point to model the data against Total Death, much work transforming the data was needed to ensure that the linear regression model could be improved.

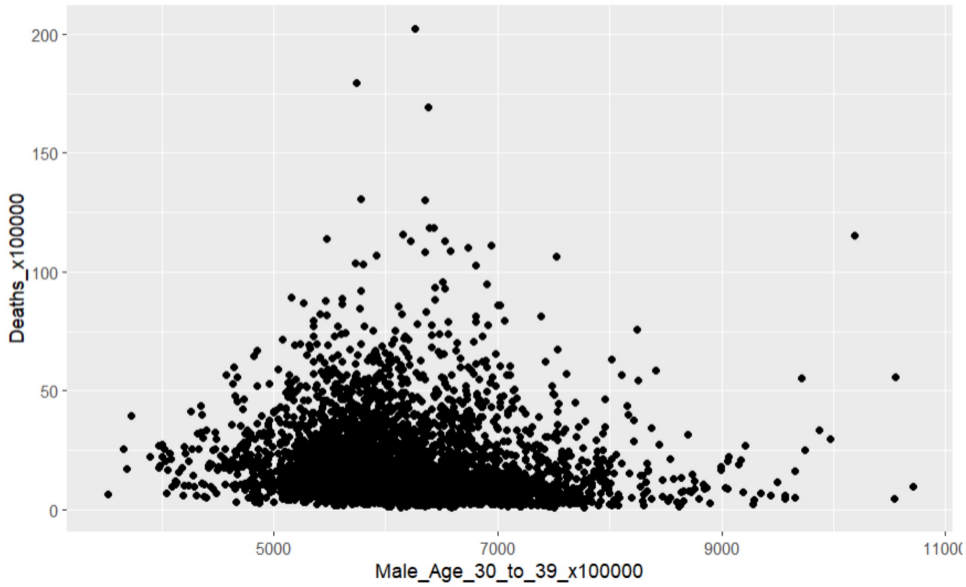


Figure 2: Males aged 30 to 39 per 100000 pop. compared to opioid-related deaths per 100000 pop.

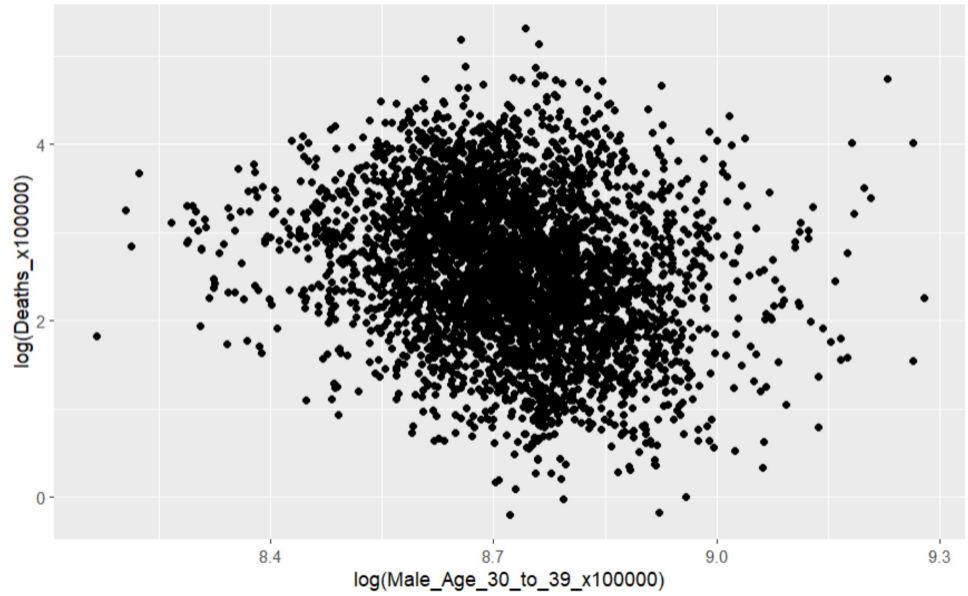


Figure 3: Log of males aged 30 to 39 per 100000 pop. compared to log of opioid-related deaths per 100000 pop.

Using the normalized data, the team explored log-transformation and was able to better establish a linear relation between some of the features and the log of Total Death Rate. One example of this was that there appeared to be positive correlation between log of the rate of high school education and the log of population-adjusted Death Rate (refer to figure 5).

Secondly, a correlation analysis shows that features displayed varying degree of correlation with each other.[2] Therefore, while one of the goals for the researchers to pursue was that of dealing with high degrees of collinearity between some of the features, it was not identified as a great concern for this dataset since many feature pairs showed minimal correlation.

[2]More graphs generated and analyzed can be found in Appendix B.

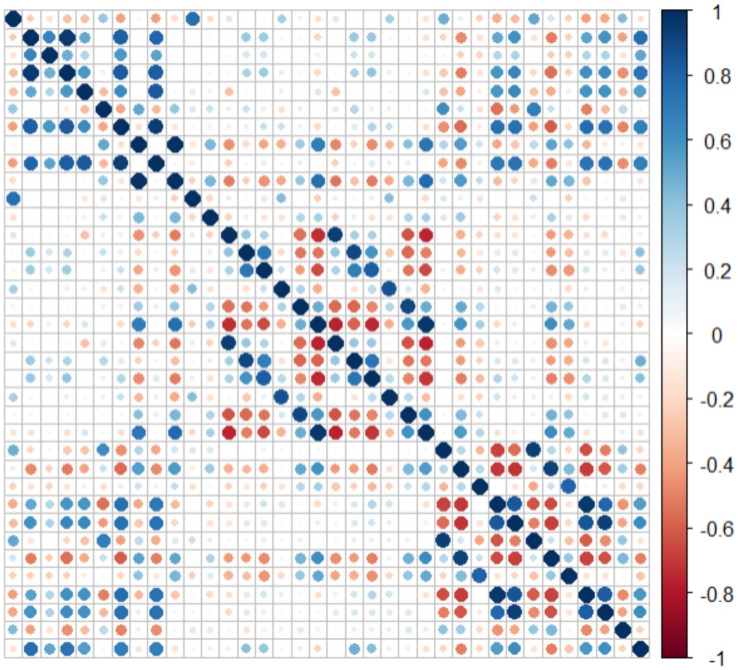


Figure 4: Correlation matrix for complete dataset

3.5 Results of feature engineering

The features which were depended on the model experiments that the team conducted. Aside from normalizing the data as stated earlier, there were three main ways features were engineered for model experimentation.

Firstly, The team did log transformation for some of the factors as well as the Deaths_x100000 feature. Log-transformations on normalized features was also done for some of the model experiments. However, not all factors required log-transformation. The graphs below show the Opioid Dispensing Rate (Disp_Rate_x100000) feature did not need log-transformation.

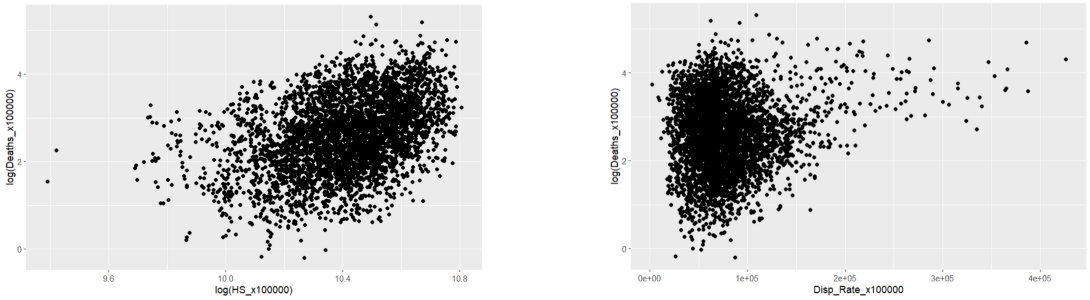


Figure 5:
Left: Log of less than high school per 100000 pop. compared with log of total death per 100000 pop.
Right: Dispensation rate compared with log of Total Death per 100000 pop.

Secondly, some of the features were also combined where plausible. For example, "Male_Age_30_to_39_x100000" and "Female_Ag_30_to_39_x100000" could be summed, becoming "Age_30_to_39_x100000". For normalizing such columns, the additions were done first before normalizing.

Thirdly, the features were also standardized for some of the model experiments. This meant that the features were scaled to ensure that their mean and standard deviation were between 0 and 1. This was done by subtracting each value with the feature's mean and dividing that difference with the feature's standard deviation:

$$x_{i,standardized} = \frac{x_i - \bar{x}}{\sigma_x}$$

4 Modeling

4.1 Models used and model selection process

The primary model used in this research was multiple linear regression. The end goal of the project was to produce a functional linear regression model whose coefficients could be analyzed to determine the explanatory strength of the various factors. Random Forest regression was also explored in this project. The drawback of Random Forest regression is that, while it is possible to use it to ascertain the relative feature importance, it is not able to provide the coefficients of each of the factors. One potential benefit, however, was that it was less susceptiable to bias introduced by feature collinearity.

To evaluate these models, the team chose to use adjusted R^2 for comparison between linear models. This allowed the team to determine the portion of the variance in the dataset that each model was able to explain while also adjusting for any differences in the number of features included (since the introduction of additional features never decreases unadjusted R^2). However, for comparing between the linear models and Random Forest, root mean square error (RMSE) was used instead. This is because adjusted R^2 is applicable to linear models only, while Random Forests models are mostly not linear. Despite this, the team also generated Pseudo-R^2 values for Random Forests as they allowed us to infer how much variation within the dataset the Random Forest model could explain.

In terms of feature importance and selection, the team adopted a 95% confidence level ($\alpha = 0.05$) for linear regression, and %MSE for Random Forest. In terms of confidence level, any feature that hit at least 95% confidence level would be identified as an important feature. For %MSE, the features were ranked from highest to lowest.

4.2 Comparison of models

A baseline model of the Random Forest model (ntree = 100) and Linear Regression were used. This means that all the features, save for "State", "County", "Year", "Labor", "Employed" and "Unemployed", were modeled using both Random Forest and Linear Regression.

The results of the experiments outlined above are shown below. The details of the variables of significance/importance can be found in Appendix C.

Model	RMSE	(Pseudo/Adjusted) R^2
Linear Regression (Baseline)	19.435	0.361
Random Forest (Baseline)	11.110	0.536

Figure 6: Comparison of models

4.3 Hyperparameter tuning, optimization and other experiments

As mentioned above, the experiments conducted were with the normalized dataset, and some experiments were conducted with combined features and log transformation.

For linear regression, variable elimination techniques have been employed in an effort to produce an optimal model for the prediction of opioid-related death. These include forward-, backward- and stepwise-selection. Other than feature selection, the team used feature engineering in the model to determine if these features helped to boost the performance of the model. Further fine-tuning experiments were also conducted using regularization methods such as LASSO, ElasticNet, and Ridge Regression.

For Random Forest Regression, the team was able to tune number of trees (ntrees), node size and maximum number of nodes for the trees.

Models	Feature Selection	Feature Transformation/Engineering (where necessary)	Fine Tuning	Model Comparison
(Multiple) Linear Regression	• Backward • Forward • LASSO	• Log Transformation • Feature Combinations	• ElasticNet/ Ridge Regularisation	• Adjusted R^2 • Root Mean Square Error
Random Forest Regression	• Feature Importance		• Node Size • Maximum number of terminal nodes in trees • Number of trees	• Pseudo-R^2 • Root Mean Square Error

Figure 7: Model Tuning

A total of 76 experiments were conducted in order to find the optimal parameters and features to use for the final model. This did not include prior experiments made with earlier versions of the datasets.[3]

4.4 Model performance and utility

In this section, only the best performing models from Linear Regression and Random Forest will be discussed.

For linear regression, the model with the highest adjusted R^2 value was obtained with the following configuration:

1. Standardized features (z-score normalized)
2. Log-linear transformation features: $\log(Y) = b_0 + b_1x_1 + \ldots + b_nx_n$
3. Lasso / Elastic Net regularization

With this configuration, the model was able to obtain a 0.4434 adjusted R^2 values, which constituted a marked increase over the 0.361 R^2 from the baseline. The summary of the model is below:

```
36 x 1 sparse Matrix of class "dgCMatrix"
                                                s1
(Intercept)                          -3.170408e+00
Unemp_Rate                           -8.153441e-02
Average.earnings.per.job..dollars.   -5.340529e-09
Avg_nonfarm_prop_inc                 -2.785172e-09
Avg_wages_salaries                    .
PerCapita_div_int_rent               -1.097030e-08
PerCapita_inc_maint_benef             2.143404e-04
PerCapita_net_earnings                1.440830e-08
PerCapita_pers_curr_transf_receipts   .
PerCapita_pers_inc                    .
PerCapita_retire_and_other            8.285135e-08
PerCapita_unemp_ins_comp              9.161905e-05
Male_Under_25_x100000                 2.364070e-05
Male_Age_25_to_30_x100000             1.369134e-04
Male_Age_30_to_39_x100000             6.424068e-05
Male_Age_40_to_49_x100000            -2.584744e-05
Male_Age_50_to_59_x100000             3.180069e-04
Male_Age_60_and_abv_x100000           1.539600e-04
Female_Under_25_x100000               6.046777e-05
Female_Age_25_to_30_x100000           9.008010e-05
Female_Age_30_to_39_x100000           1.805107e-04
Female_Age_40_to_49_x100000          -4.177690e-05
Female_Age_50_to_59_x100000          -9.526061e-06
Female_Age_60_and_abv_x100000         .
Male_below_HS_x100000                 9.268045e-06
Male_HS_and_abv_x100000               .
Male_Assoc_Degree_x100000            -4.650407e-04
Male_Bachelor_Degree_x100000          9.827893e-05
Male_Grad_Degree_x100000             -1.252013e-04
Female_below_HS_x100000              -6.583119e-05
Female_HS_x100000                     3.894938e-07
Female_Assoc_Degree_x100000           1.794139e-04
Female_Bachelor_Degree_x100000       -1.775251e-04
Female_Grad_Degree_x100000            1.338686e-04
Disp_Rate_x100000                     2.228869e-06
PerCapita_GDP                        -2.409268e-03
```

Figure 8: Summary of linear regression model

[3]Refer to Appendix C for the results of the experiments.

Given the nature of ElasticNet/LASSO, certain features were suppressed and rendered unused. These were primarily the economic factors as shown above. This was likely the product of collinearity existing between these features.

As for Random Forest, the model with the highest pseudo-R^2 value had the following configuration:

1. Standardized features (z-score normalized)
2. Number of trees: 1000

This yielded the pseudo-R^2 value of 0.559 which was a slight improvement over the baseline Random Forest model. The summary of the feature importance can be found below and are ranked from top to bottom.

	%IncMSE	IncNodePurity
PerCapita_retire_and_other	90.590038	84220.894
Female_Age_50_to_59_x100000	88.959320	29158.909
PerCapita_pers_curr_transf_receipts	67.366115	62562.493
Male_Assoc_Degree_x100000	56.136761	29343.513
Male_Age_50_to_59_x100000	53.354417	27599.429
Male_Age_60_and_abv_x100000	51.614684	36080.767
Unemp_Rate	41.616081	26629.290
Female_Age_60_and_abv_x100000	33.000367	26057.250
Male_Bachelor_Degree_x100000	30.411210	38528.750
PerCapita_unemp_ins_comp	29.029475	26142.482
Male_Under_25_x100000	28.117150	17536.756
Male_HS_and_abv_x100000	26.433653	22021.420
Female_Under_25_x100000	26.195887	16526.082
Male_Age_30_to_39_x100000	19.499174	16312.442
Female_Age_25_to_30_x100000	19.022557	15501.279
Female_Age_30_to_39_x100000	18.876776	15662.582
PerCapita_inc_maint_benef	18.747821	20517.556
Female_HS_x100000	17.591744	14961.568
PerCapita_div_int_rent	16.972741	14983.983
Male_Grad_Degree_x100000	16.824026	12184.615
Female_Bachelor_Degree_x100000	16.294686	15133.766
Female_Assoc_Degree_x100000	14.977299	14799.355
Female_Grad_Degree_x100000	14.875487	11528.278
Male_Age_25_to_30_x100000	13.220080	13962.416
Average.earnings.per.job..dollars.	13.169521	13515.574
Disp_Rate_x100000	12.955296	17063.131
Female_below_HS_x100000	12.033458	13277.373
PerCapita_net_earnings	11.183173	10585.706
Male_below_HS_x100000	11.098534	16096.595
PerCapita_pers_inc	9.298104	8515.374
Avg_wages_salaries	9.167209	11156.242
Avg_nonfarm_prop_inc	8.367770	13063.289
Female_Age_40_to_49_x100000	8.149369	14420.002
PerCapita_GDP	7.140658	11411.041
Male_Age_40_to_49_x100000	6.717888	12438.254

Figure 9: Summary of random forest model feature importance

4.5 Further insights from modeling

Firstly, the rate of opioid dispensation does not appear to be important as previously thought. While a higher rate of opioid dispensation would naturally be assumed to contribute to higher deaths (Dhalla, Persaud & Nuurlink, 2011), this did not seem to be the case, as the Random Forest models registered it far below in terms of feature importance.

Secondly, it seems that Males, compared to Females, contribute more to Total Deaths. Additionally, Males at the higher age group also contribute more to Total Deaths. For linear regression, a county that has a higher proportion of Males aged 50 and above tends to have a higher Total Death rate. In contrast to this, the Linear Regression shows a negative relation for Females Aged between 40 and 50 and the Deaths Rate.

Thirdly, economic factors do not seem to be as important as social factors when modeling the opioid deaths. As stated earlier, the insignificant factors tend to be the economic ones, and the Random Forest feature importance ranks them low as well.

Lastly, it appears that Random Forest would be the model that better fits the data. This is seen prominently with the pseudo-R^2 value of 0.559 compared to the linear model's adjusted R^2 of 0.443, meaning that there was an approximate difference of 0.116 between the two models. Hence, if there was a model that needed to be deployed immediately given the current state of research, the Random Forest model would be chosen.

5 Discussion

5.1 Challenges to interpretation

After optimization yielded the above models based on the adjusted R^2, pseudo-R^2 and %MSE, the research team set about interpreting and analyzing both the final multiple regression model and the large number of unselected models produced in the hope of deriving tentative hypotheses regarding the real-world significance of the findings. It was noted that a large number of economic factors included in the models did show signs of multicollinearity, and this was causing the inflation of some of the coefficients in both positive and negative directions. A further correlated feature pair was that of Graduate degrees holders per 100,000 inhabitants, which was highly correlated with the feature Bachelor degree holders per 100,000 inhabitants.

The team then set about utilizing Variance Inflation Factor analysis (VIF) to strip some of the more correlated feature pairs from the models with the goal of identifying those features that offered the most interpretative value with respect to the research questions. Both a programmatic and visual approach were employed, and under the latter approach, the researchers made inferences about which types of feature were liable to be of little value for interpretation. Many combinations of features were explored, and then the VIF values recorded, in the hope of identifying the overall trends in the data that may shed light on the problem of the opiate-related death in US counties.

The programmatic approach involved removing all features that were identified as correlated using the results returned by a correlation matrix. While this did significantly reduce the VIF scores (all scores were < 4 and all but one were < 3), and in turn caused the coefficients of the linear models to return less inflated values, this approach did not necessarily select the most meaningful feature combinations for interpretation e.g. removing some, but not all data pertaining to education levels.

The visual approach proved itself more fruitful, as it allowed the team to gain insight into which features were likely to be most predictive of opiate-related death at the county level. Sadly, this meant losing a large number of features and accepting a lower adjusted R^2. The standardized dataset also posed a challenge to the interpretation of coefficients, so some experiments were conducted using the raw data to aid interpretation. Finally, it was seen that there was a clear relationship between the proportion of the population in older age brackets and the rate of opioid-related death; however, the age bracket 50-59 and age bracket 60+ showed a high degree of correlation. Consequently, for the purposes of interpretation, these two features were combined in the raw dataset prior to transformation.

5.2 Results of interpretation

The model used for interpretation was selected for the explanatory value each of the features offered as well as the low degree of correlation between the features (all VIF scores were < 3). Each feature pair showed a low degree of correlation and was considered to be representative of a key social or economic factor that the experiments indicated was predictive of opioid-related death. Each feature was also seen to be highly statistically significant, well below the alpha value ($\alpha = 0.05$) selected for the project. The coefficients of the model and their respective VIF scores can be seen below.

```
Call:
lm(formula = log(Deaths_x100000) ~ ., data = data_old)

Residuals:
     Min       1Q   Median       3Q      Max
-2.54091 -0.49559  0.02022  0.48935  2.46172

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)             1.952e+00  1.127e-01  17.320  < 2e-16 ***
Unemp_Rate             -1.205e-01  4.979e-03 -24.202  < 2e-16 ***
aged_50_plus_x100000    4.856e-05  1.921e-06  25.279  < 2e-16 ***
Below_HS_x100000        1.490e-05  4.725e-06   3.152  0.00163 **
Bachelor_Degree_x100000 -6.146e-05  4.851e-06 -12.671  < 2e-16 ***
Disp_Rate_x100000       2.518e-06  3.235e-07   7.783 8.85e-15 ***
PerCapita_GDP           2.660e-03  5.765e-04   4.614 4.07e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6986 on 4209 degrees of freedom
Multiple R-squared:  0.319,     Adjusted R-squared:  0.318
F-statistic: 328.6 on 6 and 4209 DF,  p-value: < 2.2e-16
```

Figure 10: Key coefficients identified

Table 1: VIF scores for key coefficients (log-linear model)

Feature	**VIF Score**	**Coefficient**	$\%\Delta\ (e^{b_1} - 1) \cdot 100$
Unemployment rate	1.564996	-0.1205	-11.35229
Pop. aged 50+	1.115287	0.00004856	0.00486
Pop. below high school education	2.188656	0.0000149	0.00149
Pop. holding Bachelor degree	2.907286	-0.00006146	-0.00615
Opioid dispensation rate	1.438816	0.000002518	0.00025
GDP per capita for county	1.646792	0.002660	0.26635

Based on the above model, the following correlations were identified:

- Opioid-related death is inversely correlated with educational level
- Opioid-related death is positively correlated with opioid dispensation rate
- Opioid-related death is positively correlated with the proportion of the population over the age of 50

The researchers also identified the following unanticipated correlations:

- Opioid-related death appears to be inversely correlated with the unemployment rate
- Opioid-related death appears to be positively correlated with GDP per capita

The relationships shown by both educational level and opioid dispensation rate were anticipated by the researchers in the initial planning stage of the project. However, the relationship between age and opioid-related death was not. Despite this fact, it seems plausible that older people die more frequently from opioid-related causes since research supports the belief that long-term opioid use for medical grounds is more common among older demographics (Campbell et al., 2010).

What is of note regarding the the relationship between opioid dispensation rate and opioid-related death is that this was expected by the researchers to be by far the most useful predictor of opioid-related death rate at the county level, but this is not what was observed in the findings. While positively correlated, this feature did not appear to be more important than other socioeconomic factors. Potentially, one reason for this could be the presence of illegal sources of opioids in different geographical regions, the dispensation of which is not recorded in the data.

The anticipated relationship for educational level is also supported by existing research, which found that opioid usage was more frequent among less educated people in the United States (Ellis et al., 2020). This relationship seems to have a cogent basis since it seems likely that people with a higher levels of educational attainment are more aware of the dangers of long-term opioid usage and that the usage of opioids discourages people from attending academic institutions. Naturally, support for such causative claims would require further research.

The relationships observed for unemployment rate (negative correlation) and GDP per capita (positive correlation) were not anticipated by the researchers and appear counter-intuitive. It was assumed by the researchers that economic deprivation was a leading driver of opioid-related death. Unfortunately, this conclusion cannot be supported by this research. One possible explanation for these findings is the presence of some unknown confounding variable that was not taken into consideration by the researchers. For example, our analysis did not take into account the urban density of each of the counties, and it seems plausible that highly urban areas would, on average, perform better economically, with lower rates of unemployment, but also have a higher rate of opioid usage among the population. Further research would need to be undertaken to test this hypothesis.

Finally, the researchers would like to comment on the supposition, supported by the Random Forest regression model, that the proportion of men in the population was a factor in the rate of opioid-related death. This was not revealed by the analysis of the multiple regression model above, but nonetheless the impact of this feature was indicated by the relative feature importance of the optimized Random Forest model. It is the position of the researchers that this is likely a useful predictive feature, and this is supported by research conducted by The National Institute on Drug Abuse (Overdose Death Rates, 2022). They found that in the U.S. between 2019 and 2020, 70% of drug overdoses involving any type of opioid occurred among men.

5.3 Responding to the research questions

- Which socioeconomic factors are most predictive of opioid-related death in U.S. counties?

From the results of the research, it appears that educational level, age, and gender of the population are most predictive of opioid-related death. Economic factors were not as useful in predicting opioid-related death at the county level as the researchers initially suspected.

- How useful are various categories of factors in the prediction of opioid-related deaths at the county level?

Broad social factors proved to be the most useful for our models. Many of these factors could be categorized as demographic data.

- What types of relationship do we more commonly see between opioid-related death and socioeconomic factors?

As noted earlier in the paper, few of the relationships between the various socioeconomic factors and the rate of the opioid-related death per 100,000 inhabitants of a county were seen to be linear. Log transformation of the response variable yielded better R^2 values. It is therefore possible that there exists a logarithmic relationship between many socioeconomic factors and opioid-related death. This seems plausible, as opioid-related death is liable to disproportionately affect regions with the worst social and economic conditions.

- How efficient are socioeconomic factors in the prediction of opioid-related death?

Given the relatively low R^2 values seen in our models, we must conclude that the factors used in this project were not particularly efficient for the prediction of opioid-related death. Further research would be required to build more efficient models.

- Which factors should be controlled for when predicting the number of opioid-related deaths in a US county?

Population was seen to be a key factor that needed to be controlled for when predicting opioid-related death. Although models that were not adjusted for population showed impressive R^2 values, this established little more than the trivial fact that more populace U.S. counties see more opioid-related deaths per year. Due to the likely presence of unknown confounding variables influencing the results of the experiments, it is also probable that further factors, such as urban density, should be controlled for.

5.4 Suggestions for further research

Unfortunately, our models had many limitations and were therefore able to provide only minor insight into the problem of opiate-related death in U.S. counties. In order to build a better model, it is likely that more socioeconomic factors should be examined and included in the research. This could identify potential confounding variables that may have affected our findings and may go some way towards shedding greater light on problem of opioid-related death in U.S. counties.

Moreover, it is important for researchers attempting to investigate this issue to be aware of the problems that multicollinearity can pose when interpreting the features included in the dataset. Researchers should select their features carefully to avoid including correlated data. This can prove challenging with geographic data since regions that perform poorly in one socioeconomic aspect are likely to perform poorly in another. Ridge, Lasso, and Elastic Net regression are likely to perform better in terms of R^2 for this reason. Additionally, Random Forest regression offers unique benefits over standard multiple regression in that it is better able function despite the presence of multicollinearity. However, all of these methods come at a cost when it comes to the interpretation of coefficients. One potential further means of tackling the problem would be to use Principal Component Analysis. Unfortunately, this was outside the scope of this project.

6 Conclusion

The researchers set out to build both a multiple regression model and a random forest regression model that could be used to predict the rate of opioid-related death in U.S. counties. This was accomplished with an adjusted-R^2 value of 0.443 for the multiple regression model and a pseudo-R^2 value of 0.559 for the Random Forest model after optimization. These models therefore offer some utility in predicting the likely opioid-related death rate per 100,000 inhabitants in U.S. counties. Further analysis of the dataset showed that multicollinearity was a problem impacting the multiple regression model and potentially hindering the reliability of coefficient interpretation. Nevertheless, the researchers were able to identify key trends in the opiate-related death rate at the county level, including the influence of education, age, and gender on the issue.

7 Reference List

Campbell, C. I., Weisner, C., Leresche, L., Ray, G. T., Saunders, K., Sullivan, M. D., Banta-Green, C. J., Merrill, J. O., Silverberg, M. J., Boudreau, D., Satre, D. D., & Von Korff, M. (2010). Age and gender trends in long-term opioid analgesic use for noncancer pain. American journal of public health, 100(12), 2541–2547.

Dhalla, I. A., Persaud, N., & Juurlink, D. N. (2011). Facing up to the prescription opioid crisis. BMJ (Clinical research ed.), 343, d5142.

Ellis, M. S., Kasper, Z. A., & Cicero, T. J. (2020). The impact of opioid use disorder on levels of educational attainment: Perceived benefits and consequences. Drug and alcohol dependence, 206, 107618.

Luo, F., Li, M., & Florence, C. (2021, April 16). State-Level Economic Costs of Opioid Use Disorder and Fatal Opioid Overdose. Morbidity and Mortality Weekly Report, 70(15), 541–546.

McGranahan, D., & Parker, T. (2021, April). The Opioid Epidemic: A Geography in Two Phases ERR-287. U.S. Department of Agriculture, Economic Research Service, 287.

National Institute on Drug Abuse (2022). Overdose Death Rates.
https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates

8 Appendices

8.1 Appendix A: Screenshots of datasets compiled

- US GDP Data by County

GeoFIPS	GeoName	Region	TableName	LineCode	IndustryCla	Description	Unit	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
"00000"	United Stat		CAGDP1	1	...	Real GDP (t	Thousands	1.33E+10	1.35E+10	1.39E+10	1.44E+10	1.49E+10	1.53E+10	1.56E+10	1.56E+10	1.52E+10	1.56E+10	1.59E+10
"00000"	United Stat		CAGDP1	2	...	Chain-type	Quantity in	81.601	82.985	85.305	88.592	91.678	94.229	96.123	96.241	93.739	96.278	97.77
"00000"	United Stat		CAGDP1	3	...	Current-do	Thousands	1.06E+10	1.09E+10	1.15E+10	1.22E+10	1.3E+10	1.38E+10	1.45E+10	1.48E+10	1.45E+10	1.5E+10	1.56E+10
"01000"	Alabama	5	CAGDP1	1	...	Real GDP (t	Thousands	1.57E+08	1.6E+08	1.65E+08	1.77E+08	1.84E+08	1.87E+08	1.89E+08	1.87E+08	1.81E+08	1.85E+08	1.88E+08
"01000"	Alabama	5	CAGDP1	2	...	Chain-type	Quantity in	82.883	84.769	87.26	93.331	97.423	98.957	99.872	98.785	95.488	97.599	99.134
"01000"	Alabama	5	CAGDP1	3	...	Current-do	Thousands	1.23E+08	1.28E+08	1.34E+08	1.48E+08	1.59E+08	1.66E+08	1.73E+08	1.75E+08	1.71E+08	1.77E+08	1.84E+08
"01001"	Autauga, A	5	CAGDP1	1	...	Real GDP (t	Thousands	949800	984039	1008032	1170278	1195319	1277094	1298942	1172510	1209088	1286603	1375813
"01001"	Autauga, A	5	CAGDP1	2	...	Chain-type	Quantity in	59.839	61.996	63.508	73.73	75.307	80.459	81.836	73.87	76.175	81.058	86.679
"01001"	Autauga, A	5	CAGDP1	3	...	Current-do	Thousands	755902	792382	824096	978609	1020385	1130144	1192292	1096667	1179897	1265180	1368637
"01003"	Baldwin, Al	5	CAGDP1	1	...	Real GDP (t	Thousands	4007706	4193288	4426472	4912303	5502688	5673657	5852051	5569570	5328816	5381575	5372338
"01003"	Baldwin, Al	5	CAGDP1	2	...	Chain-type	Quantity in	73.853	77.273	81.57	90.523	101.402	104.553	107.84	102.635	98.198	99.17	99
"01003"	Baldwin, Al	5	CAGDP1	3	...	Current-do	Thousands	3103811	3324970	3580381	4084320	4725286	5049308	5372797	5202606	5095994	5180738	5240401
"01005"	Barbour, Al	5	CAGDP1	1	...	Real GDP (t	Thousands	812751	798398	818206	888479	892271	875189	845113	790132	774354	784043	741762
"01005"	Barbour, Al	5	CAGDP1	2	...	Chain-type	Quantity in	113.864	111.853	114.628	124.473	125.004	122.611	118.397	110.695	108.484	109.842	103.918
"01005"	Barbour, Al	5	CAGDP1	3	...	Current-do	Thousands	638173	633251	660044	738033	748096	745222	742666	712939	721919	742323	720762
"01007"	Bibb, AL	5	CAGDP1	1	...	Real GDP (t	Thousands	292495	296437	309513	324760	322815	325779	346544	344498	337952	364023	372908
"01007"	Bibb, AL	5	CAGDP1	2	...	Chain-type	Quantity in	80.443	81.527	85.124	89.317	88.782	89.597	95.308	94.745	92.945	100.115	102.559
"01007"	Bibb, AL	5	CAGDP1	3	...	Current-do	Thousands	218439	226073	243273	267253	275183	287561	316273	321186	317863	348182	363634
"01009"	Blount, AL	5	CAGDP1	1	...	Real GDP (t	Thousands	810054	814356	839651	863041	887578	856880	850648	853788	843906	839899	808595
"01009"	Blount, AL	5	CAGDP1	2	...	Chain-type	Quantity in	92.104	92.593	95.469	98.129	100.918	97.428	96.72	97.077	95.953	95.497	91.938
"01009"	Blount, AL	5	CAGDP1	3	...	Current-do	Thousands	620207	634119	672579	719321	750838	746487	767929	789254	795974	802649	788735
"01011"	Bullock, AL	5	CAGDP1	1	...	Real GDP (t	Thousands	234911	242874	239075	236364	234789	244146	229268	225788	254445	238863	230337
"01011"	Bullock, AL	5	CAGDP1	2	...	Chain-type	Quantity in	95.536	98.775	97.23	96.127	95.487	99.292	93.241	91.826	103.48	97.144	93.676
"01011"	Bullock, AL	5	CAGDP1	3	...	Current-do	Thousands	173638	182777	185888	190326	189801	200032	197543	200434	229297	221428	222770
"01013"	Butler, AL	5	CAGDP1	1	...	Real GDP (t	Thousands	435433	450276	468458	492437	487731	485727	515248	499219	468208	541838	568182
"01013"	Butler, AL	5	CAGDP1	2	...	Chain-type	Quantity in	78.606	81.286	84.568	88.897	88.047	87.686	93.015	90.121	84.523	97.815	102.571

- Income Levels

State	County	Year	Description	Average earnings per job (dollars)	Average nonfarm proprietors' income	Average wages and salaries	dividends	Per capita interest and rent	Per capita income maintenance benefits	Per capita net earnings	Per capita personal current transfer receipts	Per capita personal income	Per capita retirement and other	Per capita unemployment insurance compensation
Alabama	Autauga	2006		31.212	16.271	29.082		4.646	531.0	20.481	4.881	30.008	4.309	41.0
		2007		30.682	13.878	30.000		5.173	573.0	21.383	5.213	31.769	4.597	43.0
		2008		30.096	10.398	30.792		5.545	581.0	21.918	5.852	33.315	5.189	83.0
		2009		30.638	9.584	31.159		5.376	684.0	21.437	6.178	32.991	5.304	191.0
		2010		31.848	12.076	31.752		5.303	766.0	21.737	6.736	33.776	5.764	205.0

- Education Data

County	State	Year	Male_below_HS	Male_HS	Male_Assoc_Degree	Male_Bachelor_Degree	Male_Grad_Degree	Female_below_HS	Female_HS	Female_Assoc_Degree	Female_Bachelor_Degree	Female_Grad_Degree
Abbeville	South Carolin	2009	2206	3231	690	823	349	1725	4961	816	1090	500
Acadia Parish	Louisiana	2009	5645	8431	577	1097	507	6110	9809	1214	1734	662
Accomack	Virginia	2009	3368	5632	519	1296	967	2967	7686	733	1448	1039
Ada	Idaho	2009	9413	35169	8901	29106	14722	8603	63784	9697	26985	10963
Adair	Iowa	2009	248	1444	239	272	89	168	1894	343	268	84
Adair	Kentucky	2009	1938	2644	199	320	263	1705	3231	295	403	510
Adair	Missouri	2009	692	2718	428	786	618	860	3796	530	789	784
Adair	Oklahoma	2009	1678	3313	172	469	146	1616	4266	252	520	266
Adams	Colorado	2009	26491	50976	10397	18635	7303	23488	69852	10217	18875	7775
Adams	Idaho	2009	160	566	56	174	106	162	823	73	150	77
Adams	Illinois	2009	2616	10305	1395	3046	1513	2330	14766	1787	3581	1519
Adams	Indiana	2009	1777	5232	586	892	418	1741	6548	1085	830	400
Adams	Iowa	2009	147	813	102	137	60	107	1045	109	204	54
Adams	Mississippi	2009	1900	3552	699	950	547	2396	5753	911	1529	1042
Adams	Nebraska	2009	1233	4450	1129	1339	825	1005	6184	1340	1798	808

- Opioid Deaths

State	County	County Code	Year	Multiple Cause of death	Deaths	Population
Alabama	Baldwin County, AL	1003	2006	Methadone	10	168121
Alabama	Jefferson County, AL	1073	2006	Other opioids	18	655893
Alabama	Jefferson County, AL	1073	2006	Methadone	21	655893
Alabama	Jefferson County, AL	1073	2006	Other synthetic narcotics	11	655893
Alabama	Jefferson County, AL	1073	2007	Other opioids	22	655163
Alabama	Jefferson County, AL	1073	2007	Methadone	21	655163
Alabama	Jefferson County, AL	1073	2008	Other opioids	19	656510
Alabama	Jefferson County, AL	1073	2008	Methadone	24	656510
Alabama	Jefferson County, AL	1073	2009	Heroin	11	658441
Alabama	Jefferson County, AL	1073	2009	Other opioids	34	658441
Alabama	Jefferson County, AL	1073	2009	Methadone	19	658441
Alabama	Jefferson County, AL	1073	2010	Other opioids	15	658466
Alabama	Jefferson County, AL	1073	2010	Methadone	20	658466
Alabama	Jefferson County, AL	1073	2010	Other synthetic narcotics	10	658466
Alabama	Jefferson County, AL	1073	2011	Heroin	10	658931
Alabama	Jefferson County, AL	1073	2011	Other opioids	28	658931
Alabama	Jefferson County, AL	1073	2011	Methadone	10	658931
Alabama	Jefferson County, AL	1073	2012	Heroin	28	660009
Alabama	Jefferson County, AL	1073	2012	Other opioids	21	660009

- Opioid Dispensing Rate

id	County	State	County FIPS Code	Opioid Dispensing Rate per 100	Year
0	Aleutians East, AK	AK	2013	–	2006
1	Aleutians West, AK	AK	2016	–	2006
2	Anchorage, AK	AK	2020	71.5	2006
3	Bethel, AK	AK	2050	–	2006
4	Bristol Bay, AK	AK	2060	–	2006
5	Denali, AK	AK	2068	–	2006
6	Dillingham, AK	AK	2070	–	2006
7	Fairbanks North Star, AK	AK	2090	54.7	2006
8	Haines, AK	AK	2100	–	2006
9	Hoonah-Angoon, AK	AK	2105	–	2006
10	Juneau, AK	AK	2110	95.3	2006
11	Kenai Peninsula, AK	AK	2122	89.1	2006
12	Ketchikan Gateway, AK	AK	2130	144.4	2006
13	Kodiak Island, AK	AK	2150	69.4	2006
14	Lake and Peninsula, AK	AK	2164	–	2006
15	Matanuska-Susitna, AK	AK	2170	82.5	2006
16	Nome, AK	AK	2180	–	2006
17	North Slope, AK	AK	2185	–	2006
18	Northwest Arctic, AK	AK	2188	–	2006
19	Petersburg, AK	AK	2195	107.4	2006
20	Prince of Wales-Hyder, AK	AK	2198	–	2006
21	Sitka, AK	AK	2220	–	2006
22	Skagway, AK	AK	2230	–	2006

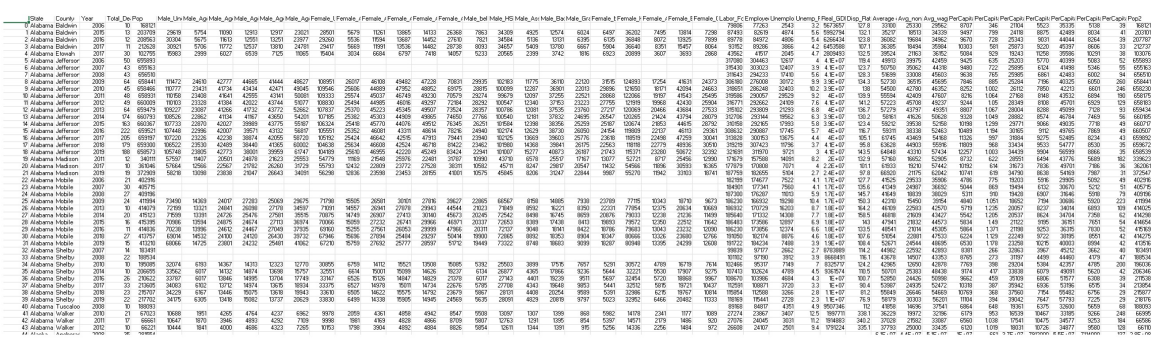
- Employment Data

LAUS_Code	State_FIPS_Code	County_FIPS_Code	County_Name/State	Year	Labor_Force	Employed	Unemployed	Rate	FIPS Clean
CN0100100000000	1	1	Autauga County, AL	2006	24425	23619	806	3.3	1001
CN0100300000000	1	3	Baldwin County, AL	2006	79806	77263	2543	3.2	1003
CN0100500000000	1	5	Barbour County, AL	2006	10713	10110	603	5.6	1005
CN0100700000000	1	7	Bibb County, AL	2006	8858	8489	369	4.2	1007
CN0100900000000	1	9	Blount County, AL	2006	26799	25939	860	3.2	1009
CN0101100000000	1	11	Bullock County, AL	2006	3703	3377	326	8.8	1011
CN0101300000000	1	13	Butler County, AL	2006	9247	8722	525	5.7	1013
CN0101500000000	1	15	Calhoun County, AL	2006	54118	51946	2172	4	1015
CN0101700000000	1	17	Chambers County, AL	2006	15761	14849	912	5.8	1017
CN0101900000000	1	19	Cherokee County, AL	2006	11864	11367	497	4.2	1019
CN0102100000000	1	21	Chilton County, AL	2006	19792	19087	705	3.6	1021
CN0102300000000	1	23	Choctaw County, AL	2006	5335	5038	297	5.6	1023
CN0102500000000	1	25	Clarke County, AL	2006	10268	9636	632	6.2	1025
CN0102700000000	1	27	Clay County, AL	2006	6189	5911	278	4.5	1027
CN0102900000000	1	29	Cleburne County, AL	2006	6646	6415	231	3.5	1029
CN0103100000000	1	31	Coffee County, AL	2006	20712	19945	767	3.7	1031
CN0103300000000	1	33	Colbert County, AL	2006	25331	24162	1169	4.6	1033
CN0103500000000	1	35	Conecuh County, AL	2006	5127	4840	287	5.6	1035
CN0103700000000	1	37	Coosa County, AL	2006	4797	4531	266	5.5	1037
CN0103900000000	1	39	Covington County, AL	2006	17213	16553	660	3.8	1039
CN0104100000000	1	41	Crenshaw County, AL	2006	6602	6341	261	4	1041

- GDP Data

State	County	Year	Real GDP (thousands of chained 2012 dollars)
Alabama	Autauga	2001	949800
Alabama	Baldwin	2001	4007706
Alabama	Barbour	2001	812751
Alabama	Bibb	2001	292495
Alabama	Blount	2001	810054
Alabama	Bullock	2001	234911
Alabama	Butler	2001	435433
Alabama	Calhoun	2001	3487094
Alabama	Chambers	2001	841042
Alabama	Cherokee	2001	384513
Alabama	Chilton	2001	638533
Alabama	Choctaw	2001	640561
Alabama	Clarke	2001	672545
Alabama	Clay	2001	317661
Alabama	Cleburne	2001	329499
Alabama	Coffee	2001	1007082
Alabama	Colbert	2001	1920841
Alabama	Conecuh	2001	299986
Alabama	Coosa	2001	162956
Alabama	Covington	2001	988519
Alabama	Crenshaw	2001	266412
Alabama	Cullman	2001	2024681
Alabama	Dale	2001	1974020
Alabama	Dallas	2001	1263245
Alabama	DeKalb	2001	1665761
Alabama	Elmore	2001	1100289
Alabama	Escambia	2001	1160030
Alabama	Etowah	2001	2638640

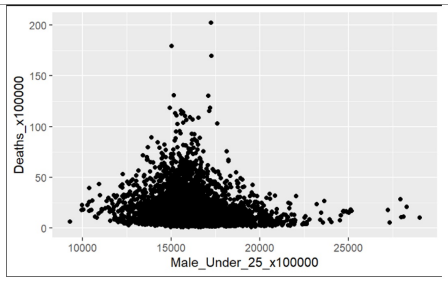
- Normalized Dataset

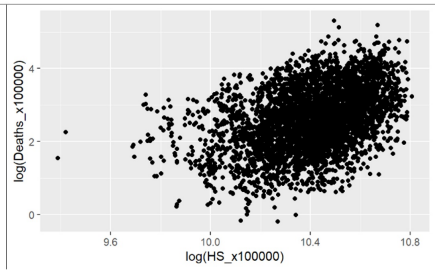
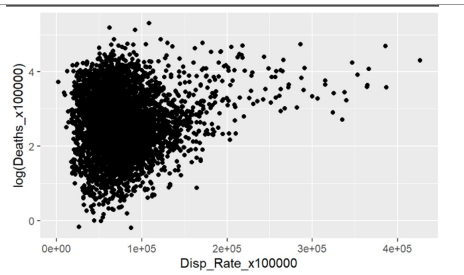


8.2 Appendix B: Plots and Figures of Exploratory Data Analysis

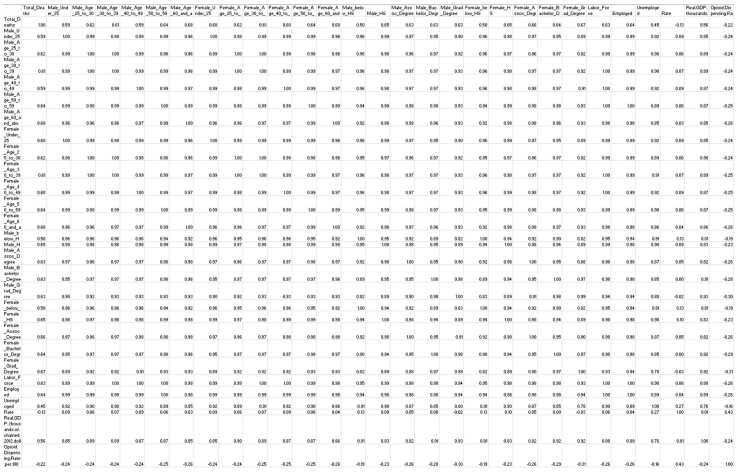
For more plots, refer to EDA_Final.html file in the Team-32 repository

- Scatterplots

X	Y	Plot
Males Aged 25 and below	Total Deaths	

Log(HS_x100000)	Log(Deaths_x100000)	
Opioid Dispensing Rate per 100000	Log(Total_Deaths)	

- Correlation Table



8.3 Appendix C: Model Experiment Results

Linear Regression (Normal Data)

S/N	Parameters	Adj. R^2	RMSE
1	Baseline	0.3609	19.4354
2	Log-Lin transformation	0.4280	0.6442
3	Lin-Log transformation	0.3650	850112609
4	Log-Log transformation	0.4280	26170000
5	Log-Lin transformation + Backward Step	0.4384	0.6541
6	Log-Lin transformation + Forward Step	0.4399	0.6527
7	Log-Lin transformation + Ridge	0.4340	0.6462
8	Log-Lin transformation + Lasso	0.4434	0.6501
9	Log-Lin transformation + Elastic Net	0.4434	0.6501

Linear Regression (Combine Data)

S/N	Parameters	Adj. R^2	RMSE
1	Baseline	0.3353	19.1683
2	Log-Lin transformation	0.4040	0.6514
3	Lin-Log transformation	0.2803	21.6492
4	Log-Log transformation	0.3350	0.7110
5	Log-Log transformation + Backward Step	0.4079	0.6607
6	Log-Log transformation + Forward Step	0.4119	0.6578
7	Log-Log transformation + Ridge	0.4097	0.6552
8	Log-Log transformation + Lasso	0.4139	0.6565
9	Log-Log transformation + Elastic Net	0.4139	0.6565

Linear Regression (Normal Data) - Standardized

S/N	Parameters	Adj. R^2	RMSE
1	Baseline	0.3609	19.4354
2	Log-Lin transformation	0.4280	0.6442
3	Lin-Log transformation	0.3053	20.2019
4	Log-Log transformation	0.3319	0.6855
5	Log-Lin transformation + Backward Step	0.4384	0.6541
6	Log-Lin transformation + Forward Step	0.4399	0.6527
7	Log-Lin transformation + Ridge	0.4340	0.6462
8	Log-Lin transformation + Lasso	0.4434	0.6501
9	Log-Lin transformation + Elastic Net	0.4434	0.6501

Linear Regression (Combine Data) - Standardized

S/N	Parameters	Adj. R^2	RMSE
1	Baseline	0.3353	19.1683
2	Log-Lin transformation	0.4040	0.6514
3	Lin-Log transformation	0.2803	21.6492
4	Log-Log transformation	0.3350	0.7110
5	Log-Log transformation + Backward Step	0.4079	0.6607
6	Log-Log transformation + Forward Step	0.4119	0.6578
7	Log-Log transformation + Ridge	0.4097	0.6552
8	Log-Log transformation + Lasso	0.4139	0.6565
9	Log-Log transformation + Elastic Net	0.4139	0.6565

Random Forest (Normal Data)

S/N	Parameters	RSQ	RMSE
1	Baseline	0.5364025	11.11002
2	ntree = 200	0.5411327	11.05319
3	ntree = 1000	0.5476726	10.97415
4	ntree = 200 nodesize = 10	0.5354913	11.12093
5	ntree = 200 nodesize = 20	0.5176722	11.33223
6	ntree = 1000 nodesize = 10	0.4919352	0.0122907
7	ntree = 1000 nodesize = 20	0.4628682	0.0126374
8	ntree = 1000 nodesize = 100	0.4457887	12.14736
9	ntree = 1000 maxnodes = 1000	0.5476726	10.97415
10	ntree = 1000 maxnodes = 300	0.5150577	11.9047
Combined Gender over Age group and Education			
11	Baseline	0.5146878	12.0202
12	ntree = 200	0.5263407	11.87502
13	ntree = 1000	0.5348184	11.76827
14	ntree = 200 nodesize = 10	0.5191726	11.96454
15	ntree = 200 nodesize = 20	0.4985773	12.21809
16	ntree = 1000 nodesize = 10	0.4919352	0.0122907
17	ntree = 1000 nodesize = 20	0.4628682	0.0126374
18	ntree = 1000 nodesize = 100	0.4282345	13.04699

19	ntree = 1000 maxnodes = 1000	0.5348184	11.76827
20	ntree = 1000 maxnodes = 300	0.5039695	12.15221

Random Forest (Normal Data) - Standardized

S/N	Parameters	RSQ	RMSE
1	Baseline	0.5545448	11.43049
2	ntree = 200	0.558487	11.3798
3	ntree = 1000	0.5591062	11.37182
4	ntree = 200 nodesize = 10	0.5437471	11.5682
5	ntree = 200 nodesize = 20	0.5220141	11.84051
6	ntree = 1000 nodesize = 10	0.4919352	0.0122907
7	ntree = 1000 nodesize = 20	0.4628682	0.0126374
8	ntree = 1000 nodesize = 100	0.4439942	12.77034
9	ntree = 1000 maxnodes = 1000	0.5591062	11.37182
10	ntree = 1000 maxnodes = 300	0.5288619	11.75539
Combined Gender over Age group and Education			
11	Baseline	0.5152334	11.76785
12	ntree = 200	0.5281687	11.60978
13	ntree = 1000	0.5362996	11.50931
14	ntree = 200 nodesize = 10	0.5178644	11.73587
15	ntree = 200 nodesize = 20	0.5012989	11.93578
16	ntree = 1000 nodesize = 10	0.4919352	0.0122907
17	ntree = 1000 nodesize = 20	0.4628682	0.0126374
18	ntree = 1000 nodesize = 100	0.4309486	12.74989
19	ntree = 1000 maxnodes = 1000	0.5362996	11.50931
20	ntree = 1000 maxnodes = 300	0.5039031	11.90458