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

**Valery Lynn, MS**
**Data Science Career Track**

# Problem:

- Centers for Disease Control and Prevention (CDC) reported more than 66% of drug related deaths involved an opioid.
- On October 26, 2017 the opioid crisis was officially declared a national Public Health Emergency under federal law.
- The economic burden is estimated to be more than $78 billion a year.

# County level services in need:

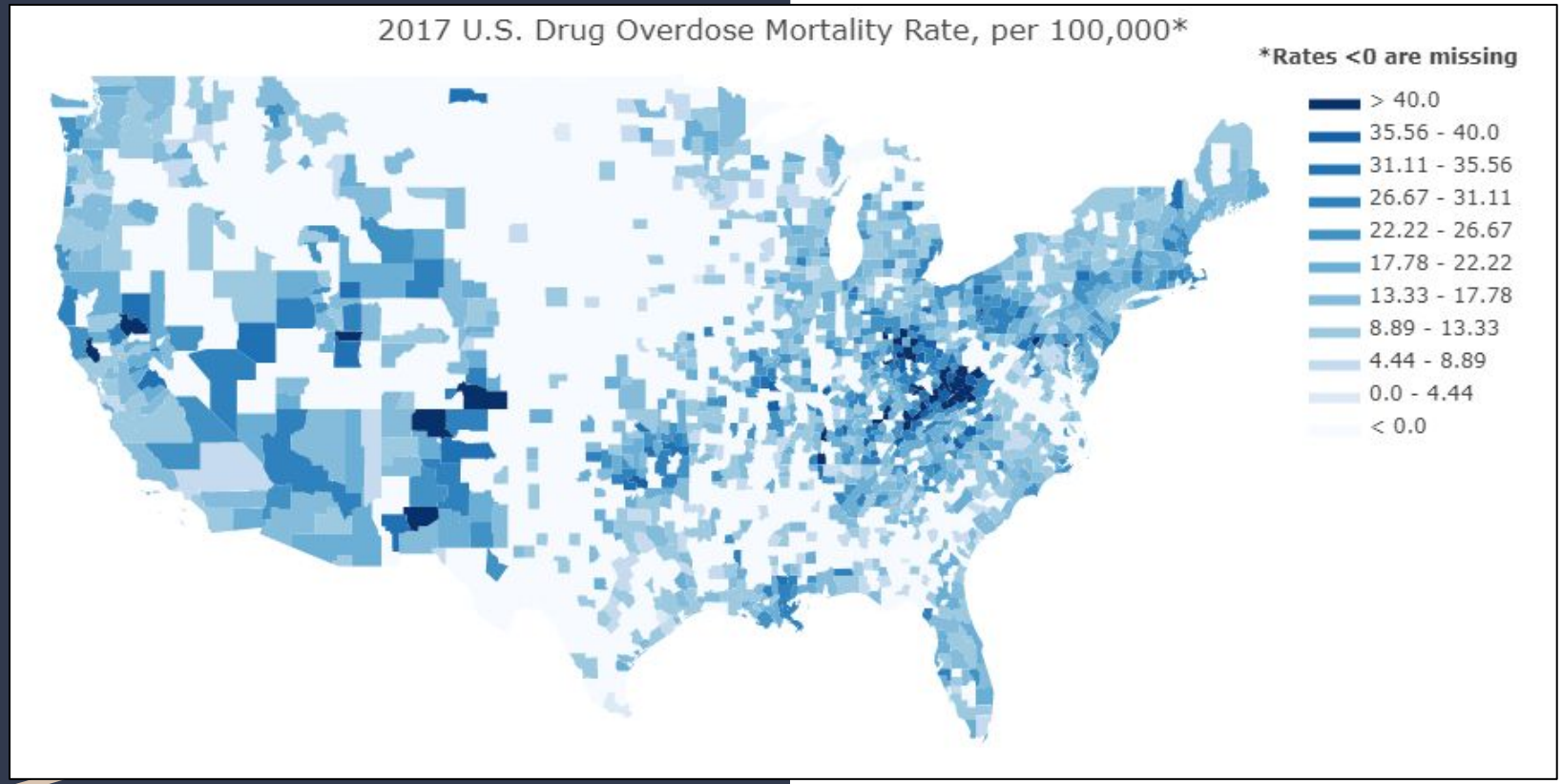
- Hospitals, crisis centers, and local planning boards are in need of predictive models to inform planning, preparation, and resource allocation.
- This model can be used to better estimate the needs of the county to address this crisis.
- Examples include:
  - how many overdose kits hospitals need to have on stock
  - how many full-time crisis prevention professionals to employ
  - how many drug rehabilitation centers need to be established or funded, etc.

# The Data (County Level):

- The County Health Rankings dataset:
  - A collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute.
- Built predominantly from the following:
  - The Behavioral Risk Factor Surveillance System (BRFSS),
  - The National Center for Health Statistics,
  - The CDC WONDER mortality data.

# The Data Story:

- My first task was to visualize the incidence of drug overdose mortality:
  - Created an interactive choropleth (heat) map of the U.S
  - Counties shaded to represent drug overdose mortality rates.




# The Problem: Too many Missing values

# 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?

# Major Findings from EDA:

1) 19 variables correlated with drug overdose mortality.

2) Positive correlations were measures of:
   a) Economic and social status
   b) Environmental stressors
   c) Physical and mental health

3) Three were negatively correlated with drug overdose mortality:
   a) Excessive drinking
   b) Household income
   c) Having at least some college education

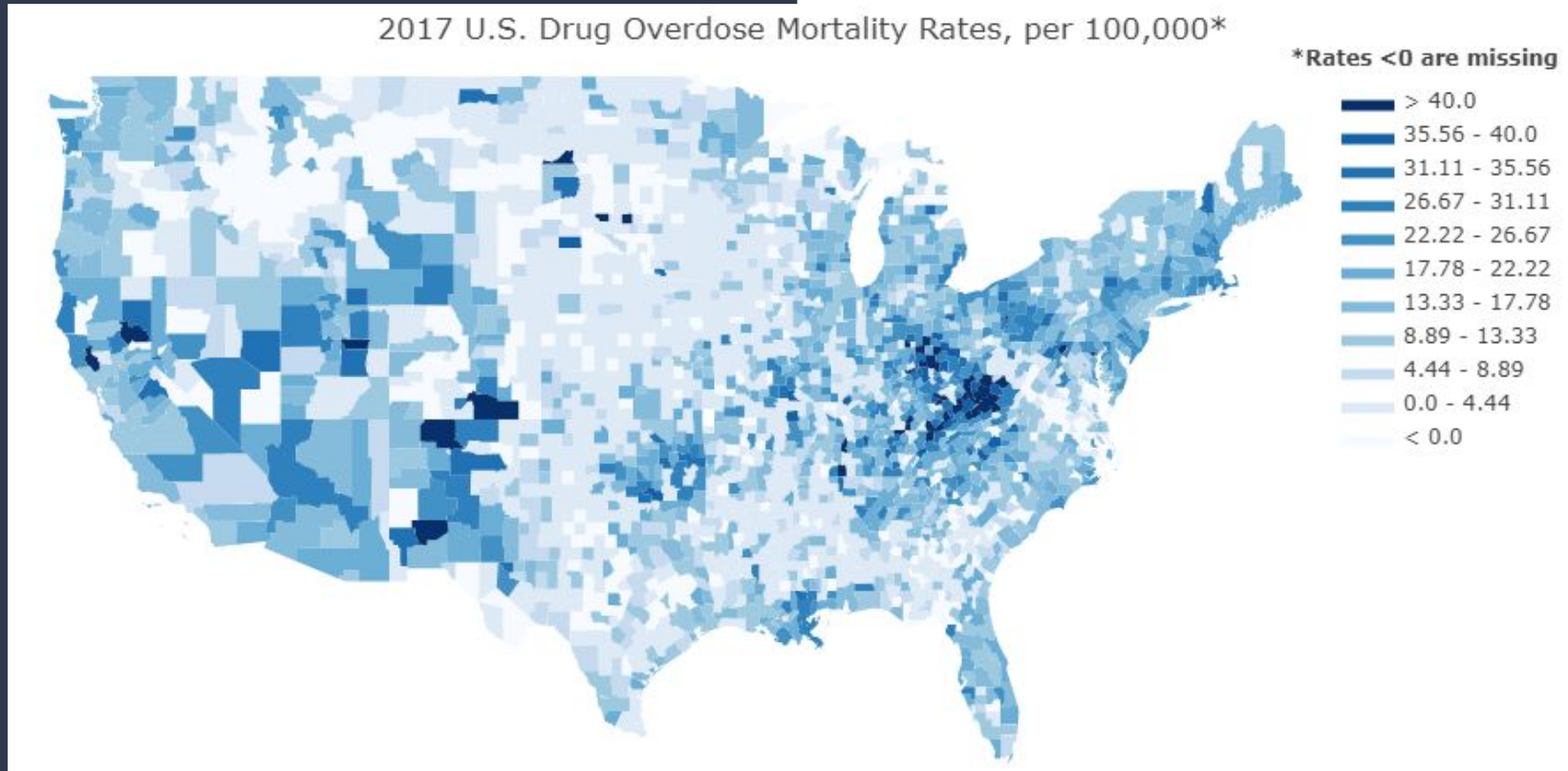
# Predictive Models Tested:

| # | Regression Type | Transformation | Number of Features |
|---|---|---|---|
| 1) | Linear Regression (OLS) | None | Full Model |
| 2) | Linear Regression (OLS) | None | 35 |
| 3) | Ridge Regression | None | Full Model |
| 4) | Linear Regression (OLS) | Log* | Full Model |
| 5) | Linear Regression (OLS) | Log* | 35 |
| 6) | Ridge Regression | Log* | Full Model |

# Model Selected:

- Linear Regression (OLS) - Step reduced, log-transformed

| # | Regression Type | Adjusted $R^2$ and Root Mean Square Error (RMSE) | Average 10-fold Cross-Validation Score (CV) |
|---|---|---|---|
| 1) | Linear Regression (OLS) - Full, untransformed | **Adjusted $R^2$**: 0.5923868474405766 **RMSE**: 0.7256769706179685 | **CV:** 0.3582203997916027 |
| 2) | Linear Regression (OLS) - Step reduced, untransformed | **Adjusted $R^2$**: 0.5739796408675106 **RMSE**: 0.7100274761313532 | **CV:** 0.4239415366434971 |
| 3) | Ridge Regression - untransformed | **Adjusted $R^2$**: 0.5736075788576231 **RMSE**: 0.7210173693911791 | **CV:** 0.3837487217984189 |
| 4) | Linear Regression (OLS) - Full, log-transformed | **Adjusted $R^2$**: 0.5809871211269844 **RMSE**: 0.6727687906026559 | **CV:** 0.391952652638861 |
| 5) | Linear Regression (OLS) - Step reduced, log-transformed | **Adjusted $R^2$**: 0.56778520526356 **RMSE**: 0.6490841781141685 | **CV:** 0.4530938467232424 |
| 6) | Ridge Regression - full, log-transformed | **Adjusted $R^2$**: 0.5685524777888835 **RMSE**: 0.6680830771598685 | **CV:** 0.414709583385758 |




# The Solution: Complete Map with Predicted Values