Predicting Provider Prescriptions

Rika Yatchak

March 5, 2019

The problem:

Given prescription data for providers, can we predict how much of a drug they will prescribe based on their specialty?

Dataset source and description

Kaggle Dataset posted by A.J. Pryor: U.S. Opiate Prescriptions/Overdoses

Original

The dataset contains three files:

1. opioids.csv: A list of drugs classified as opiates.
2. overdoses.csv: The number of overdose rates in U.S. states in 2013 as well as their population.
3. prescriber-info.csv: Information about 25,000 prescribers including their geographic location, specialty, credentials, and information about which drugs they prescribed in 2013.

Geographic Distribution of the Opioid Epidemic

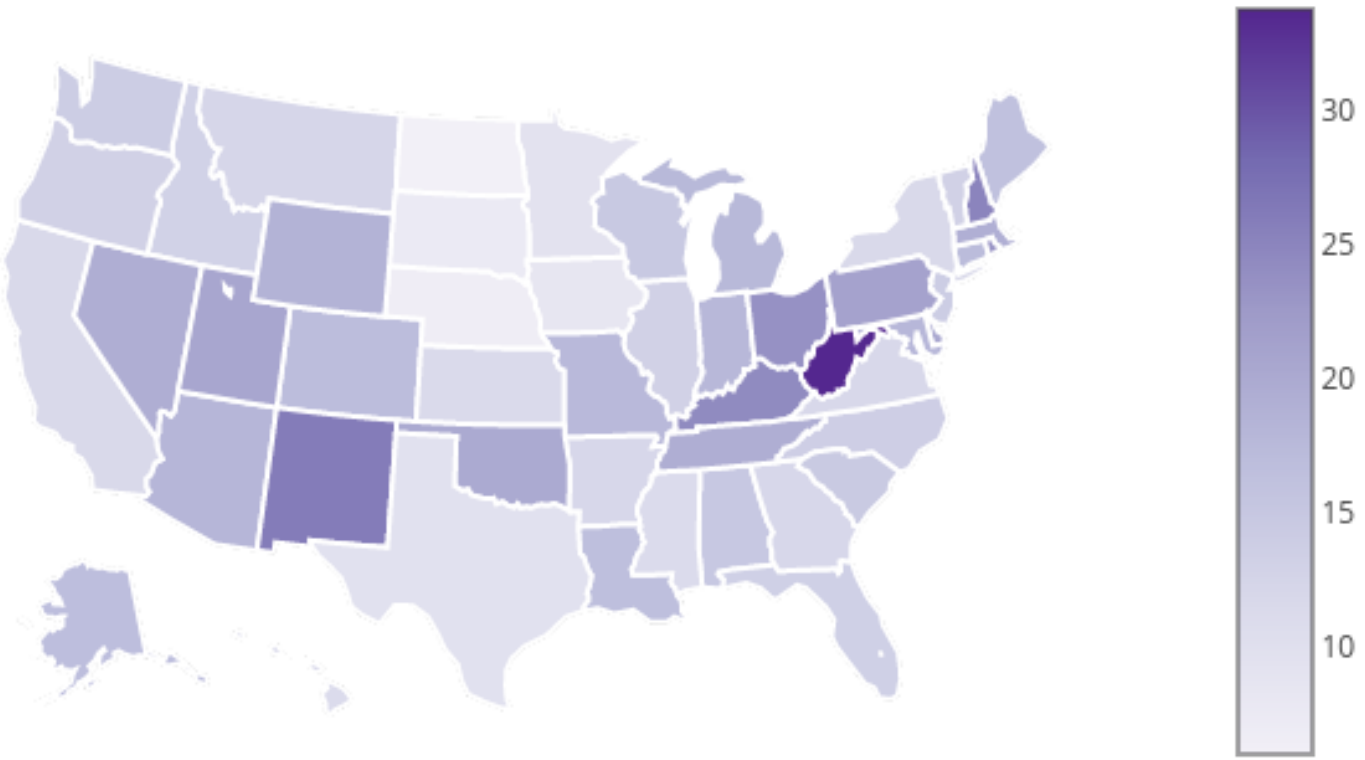
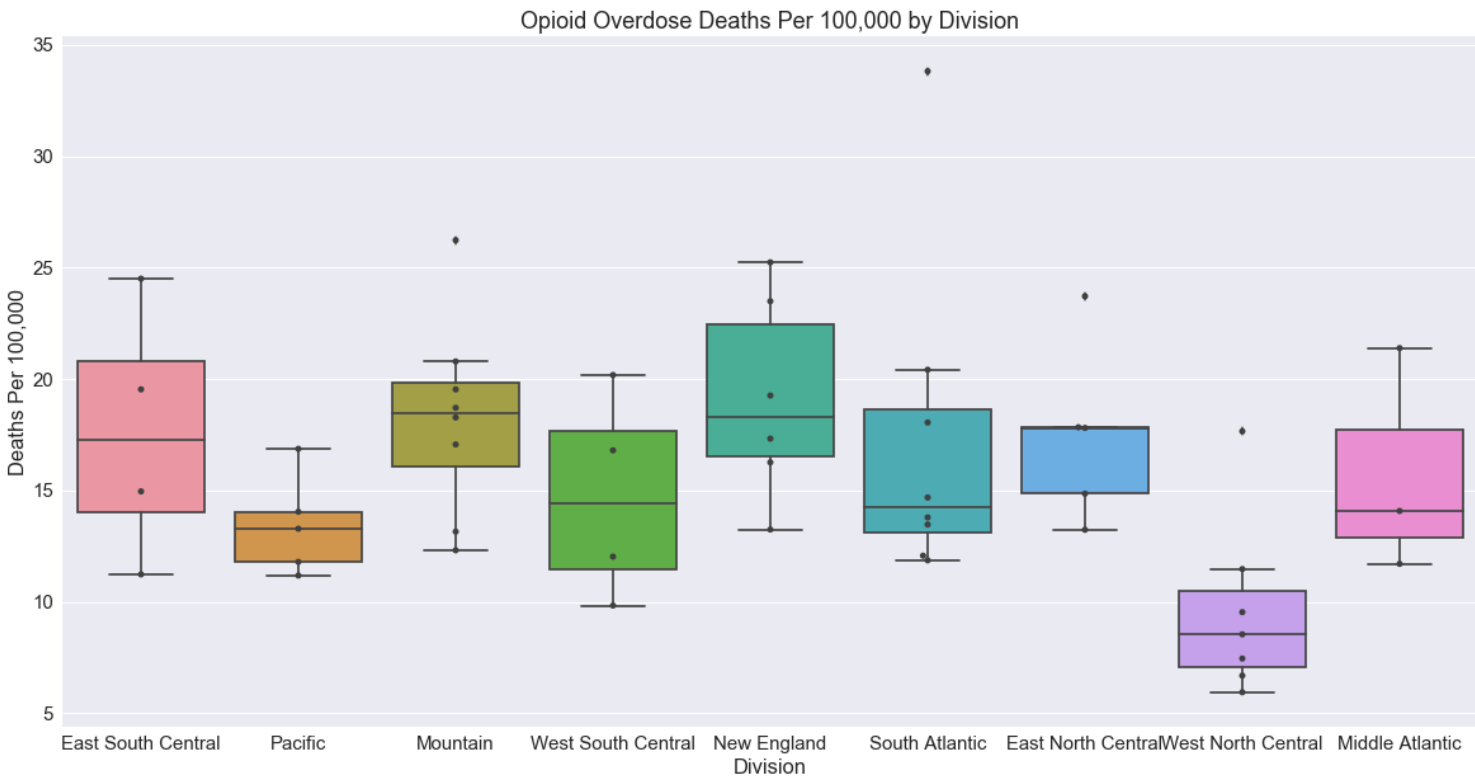


Figure : U.S. opioid overdose deaths per 100,000 by state.

Overdose Death Rate by U.S. Census Division



U.S. Census Divisions

East North Central: Illinois, Indiana, Michigan, Ohio, Wisconsin

East South Central: Alabama, Kentucky, Mississippi, Tennessee

Middle Atlantic: New Jersey, New York, Pennsylvania

Mountain: Arizona, Colorado, Idaho, Montana, Nevada, New Mexico, Utah, Wyoming

New England: Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont

Pacific: Alaska, California, Hawaii, Oregon, Washington

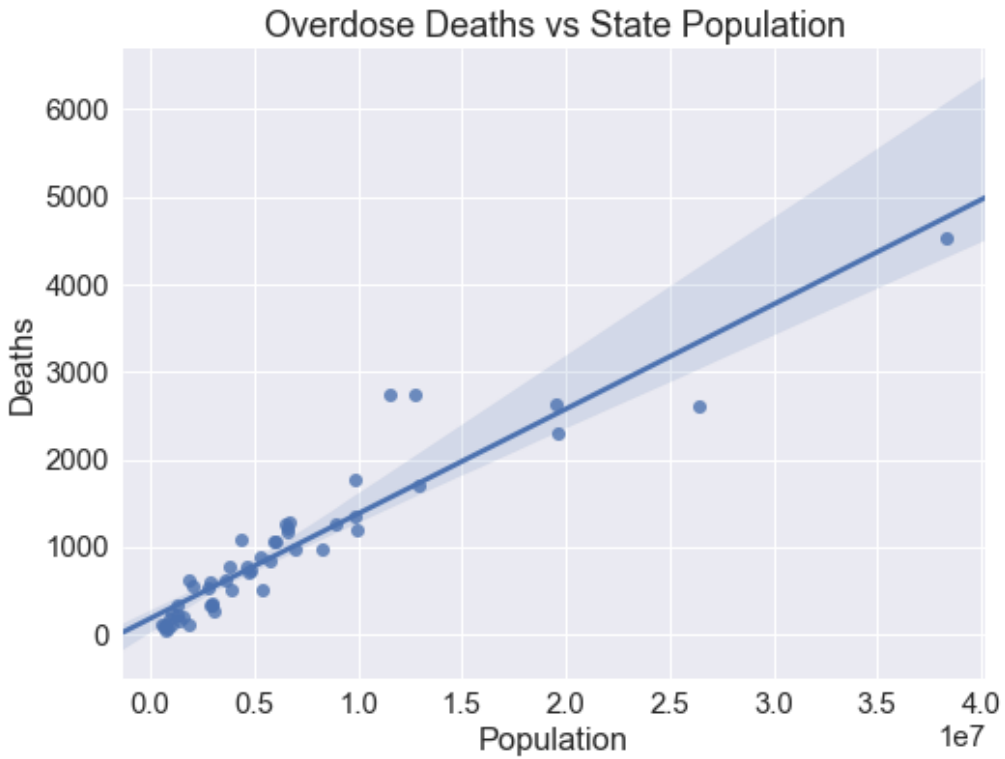
South Atlantic: Delaware, District of Columbia, Florida, Georgia, Maryland, North Carolina, South Carolina, Virginia, West Virginia

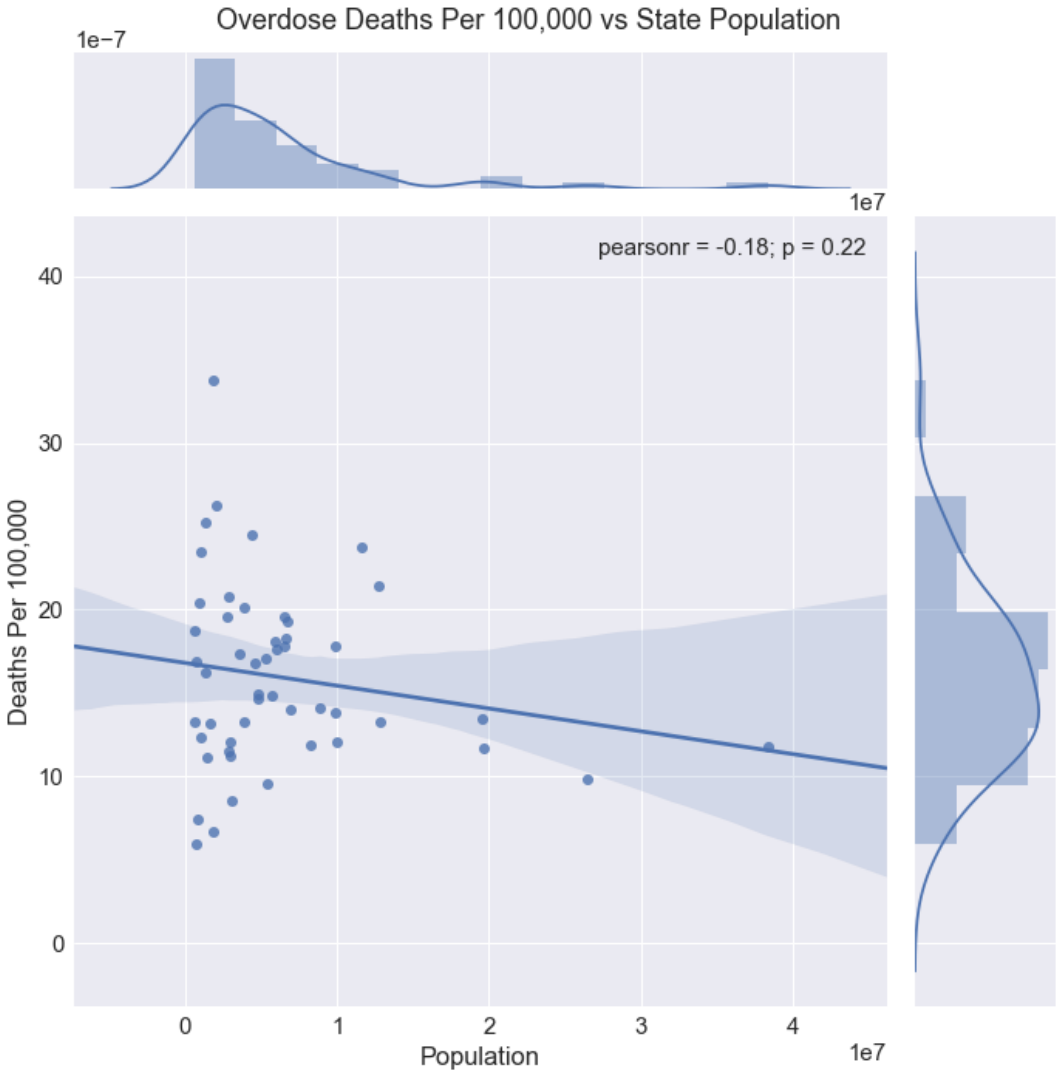
West North Central: Iowa, Kansas, Minnesota, Missouri, Nebraska, North Dakota, South Dakota

West South Central: Arkansas, Louisiana, Oklahoma, Texas

What is an epidemic?

epidemic: a widespread occurence of an infectious disease in a community at a particular time.
Based on the 2013 opioid overdose deaths by state, can we make the case that there is an opioid epidemic?




Overdose Deaths Per 100,000 vs State Population
1e−7
1e7
pearsonr = -0.18; p = 0.22
Deaths Per 100,000
0
10
20
30
40
0
1
2
3
4
Population
1e7

Can providers be distinguished by the drugs they prescribe?

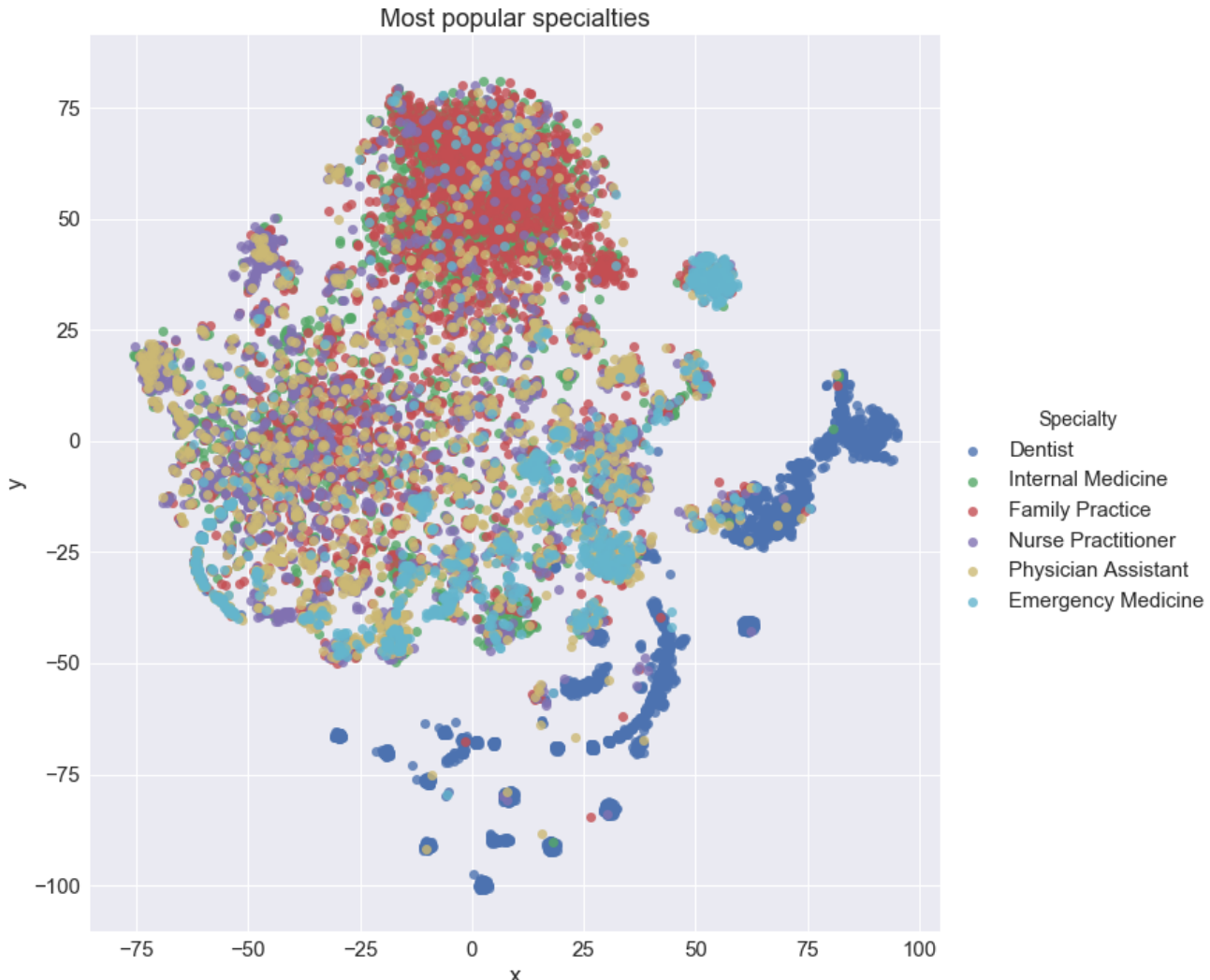
Yes!

Example (Duloxetine HCL prescriptions)

In our sample, **orthopedic surgeons** write **.39** prescriptions for Duloxetine HCL per provider.
Psychiatrists write nearly **41** prescriptions per provider.
The difference in mean Duloxetine HCL prescriptions is statistically significant. Duloxetine is generally prescribed to treat depression and anxiety, as well as chronic pain.

TSNE Visualization: Specialists distinguished by their prescriptions

TSNE Visualization: General practitioners distinguished by their prescriptions



Lisinopril prescriptions based on specialty

Linear regression task: For "popular" specialties, predict Lisinopril prescriptions.

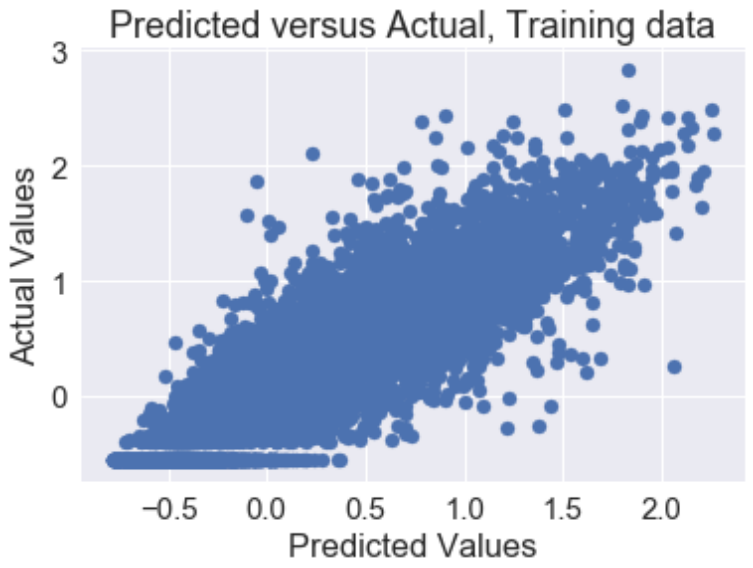
Features: Prescriber specialty, total number of drugs prescribed (from possible 250 in dataset).

Residuals plot exhibited linear correlation. Log transformation of the target variable significantly improved prediction accuracy and residuals plot.

R^2 score: .85

Residuals plot, Predicted vs. Actual plot

(a) Residuals plot



(b) Predicted vs. Actual Plot

Predictions for other drugs

In principle, same strategy should work.
R^2 score was poor for most other drugs.
A log transformation for the target variable did *not* improve prediction accuracy for any other drugs with original R^2 score ¿ .6.

Predicting provider specialty

Task was only performed for a select set of specialties: Psychiatry, Cardiology, Obstetrics/Gynecology, Orthopedic Surgery, Optometry, Opthalmology
Multinomial logistic regression.
Using numbers for all drugs prescribed, predictive accuracy was excellent:

Interesting question for further work: best way to find a smaller subset of drugs without sacrificing too much predictive accuracy.