Analysis of U.S. Regional Crime Rates

Ziwei Meng, Ao Liu

April 26, 2017

Outline

Goal and Procedure

- Compute the regression model based on the training set and test the accuracy of the model using the test data.
- Based on the model, implement policies that will lead to the reduction of the number of serious crimes in their county.
- Discuss the future improvements of the model.

Data Overview

- **Geographic Data:** Land Area, Geographic Region
- **Demographic Data:** Total population, Percent of population aged 18-34, Percent Bachelor's Degree
- **Economics Data:** Percent Below Poverty Level, Total Personal Income, Per Capita Income

Data Processing

- Check for missing values (and substitute them with mean values)
- Calculate more variables that cater to our needs:
  (1) Population Density $= \frac{Population}{Area}$
  (2) Physician Per 1000 Population $= \frac{Population}{Area}$
  (3) Hospital Beds Per 1000 Population $= \frac{HospitalBeds}{Population/1000}$
  (4) Crime Rate Per 1000 Population $= \frac{Crimes}{Population/1000}$
- Randomly Select 330 rows of data to train the regression model, and the remaining 110 rows are used for testing the accuracy of our model

Heatmap

First we explore the correlation of variables:

Heatmap
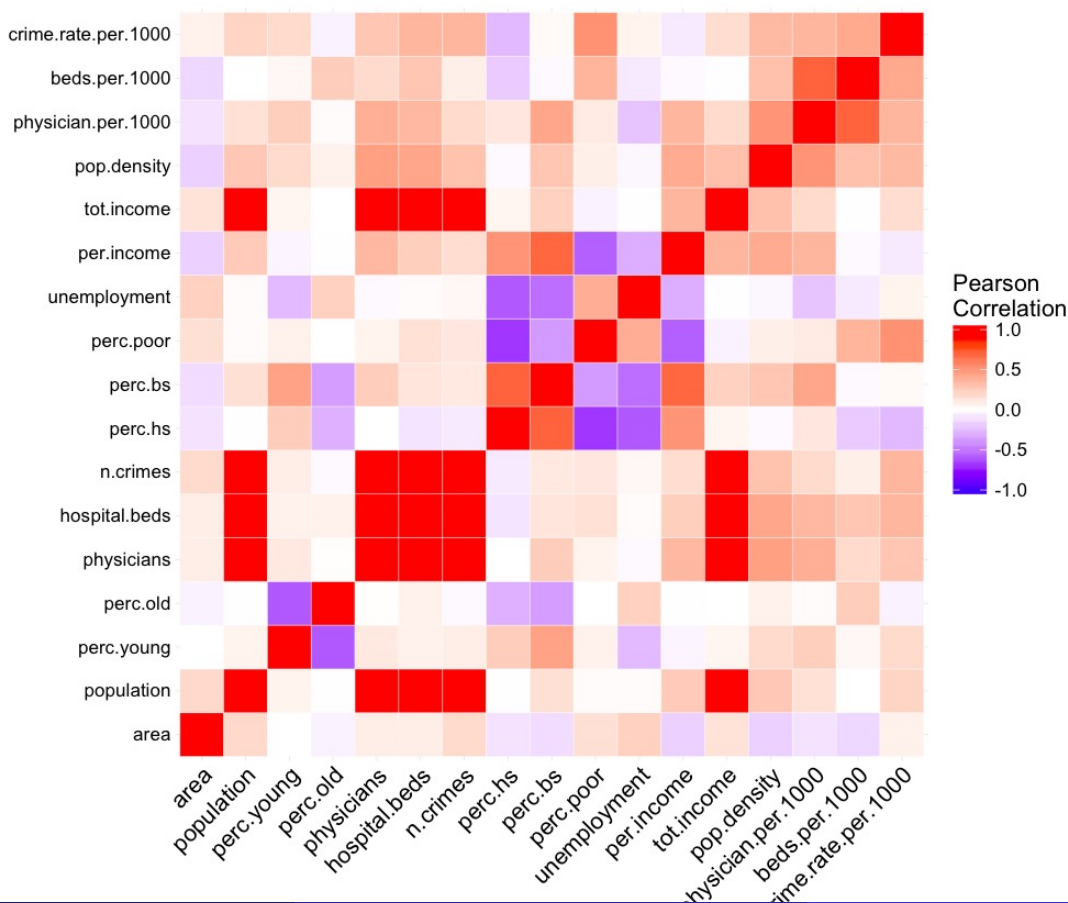
- Given 16 predictor variables, some of them are strongly correlated with each other, which will cause us to get some potentially false conclusion, thus we remove these variables.

- The remaining variables are:
  Area, Percentage of Young People, Percentage of Old People, Percentage of High School, Percentage of Bachelor, Percentage of Poor, Unemployment, Income, Region, Population Density, Physician Per 1000 Population, Beds Per 1000 Population

Regression Model

- Given the fact that crime rate is a value between 0 and 1, using an ordinary linear regression model will affect model's accuracy, in this question we fit the data to **Poisson Regression Model with Offset and Quasi-likelihood**

- Then we do the significant test for each variable, and remove the insignificant variables, then do the regression again.

Outliers

Check outliers
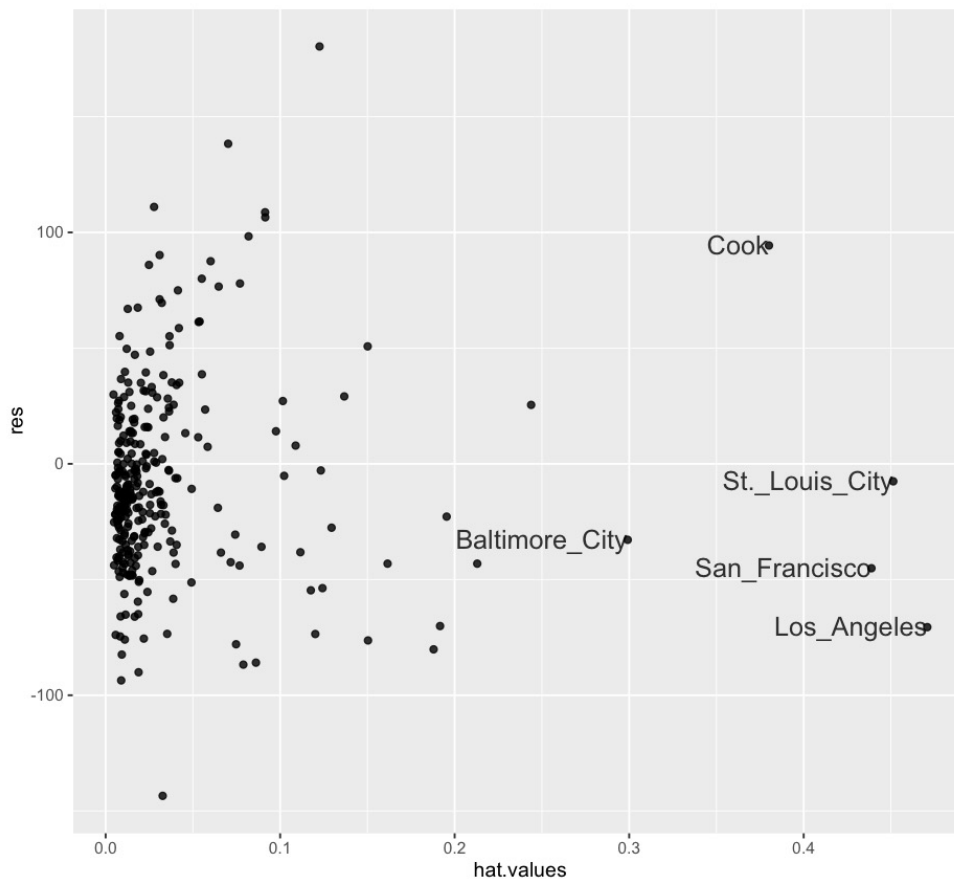
Removal of Insignificant Variables

- Through the resulting output table from Poisson Regression, the following variables are insignificant:
  area, percent of old people, percent of people with high school education.

- After removing the insignificant variables, we build the Poisson Regression Model again using only the most important variables.

Interpretation of Parameters and Visualization

Here we interpret the meaning of each parameters in our model:

- percent young: If we increase the percent on young people by 1 while holding all other variables the same, the crime rate would increase by a multiplicative factor of 1.017847 on average.

- percent poor: If we increase the percent on poor people by 1 while holding all other variables the same, the crime rate would increase by a multiplicative factor of 1.024423 on average

- population density: If we increase the log of the population density by 1 while holding all other variables the same, the crime rate would increase by a multiplicative factor of 1.085662 on average.

Interpretation of Parameters and Visualization

- region: Holding all other variables the same, the crime rate in NC is higher than that in NE by a multiplicative factor of 1.347162 on average, the crime rate in S is higher than that in NE by a multiplicative factor of 1.775532 on average, the crime rate in NC is higher than that in W by a multiplicative factor of on average.

- beds per 1000 population: If we increase the density of beds per 1000 people by 1 while holding all other variables the same, the crime rate would increase by a multiplicative factor of 1.049475 on average.
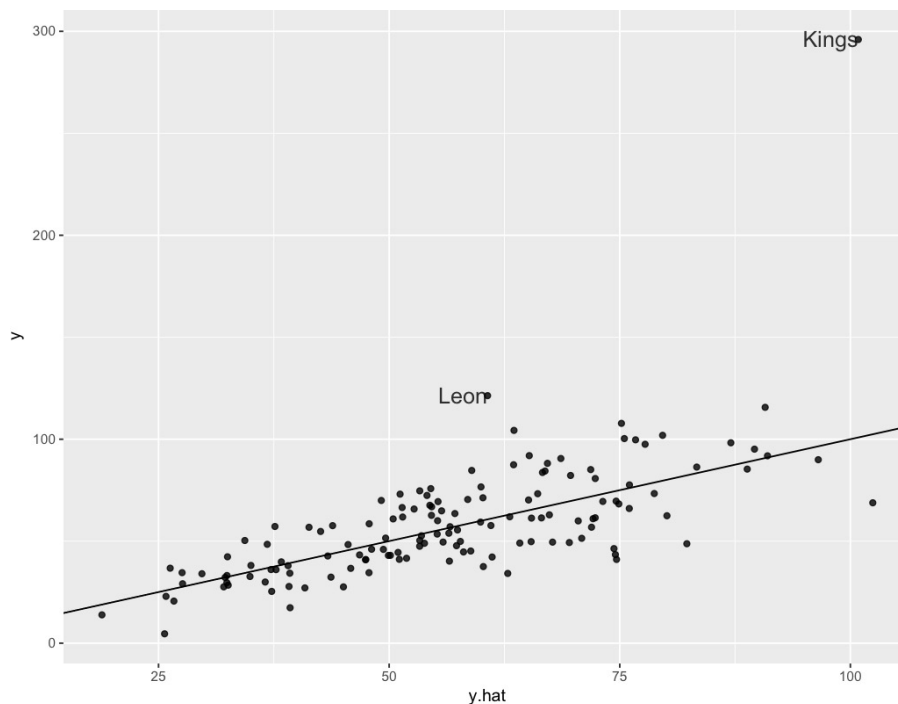
Prediction on Testing Data

- Finally we use the testing data to predict the crime rate of the remaining 110 counties and examine the accuracy of the regression model
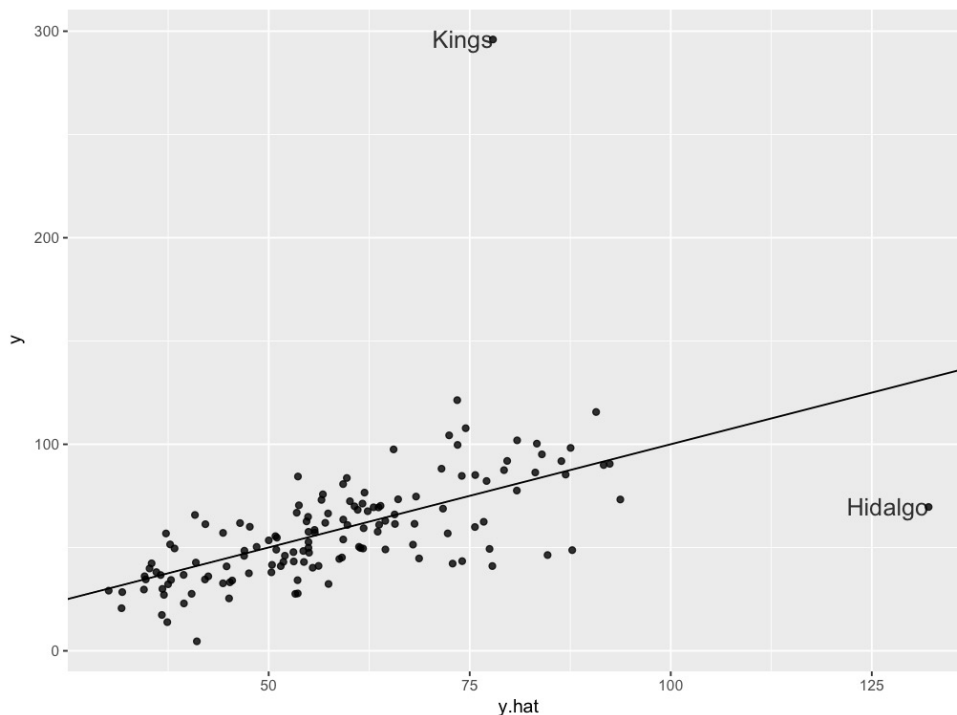
Random Forest/XGboost Model

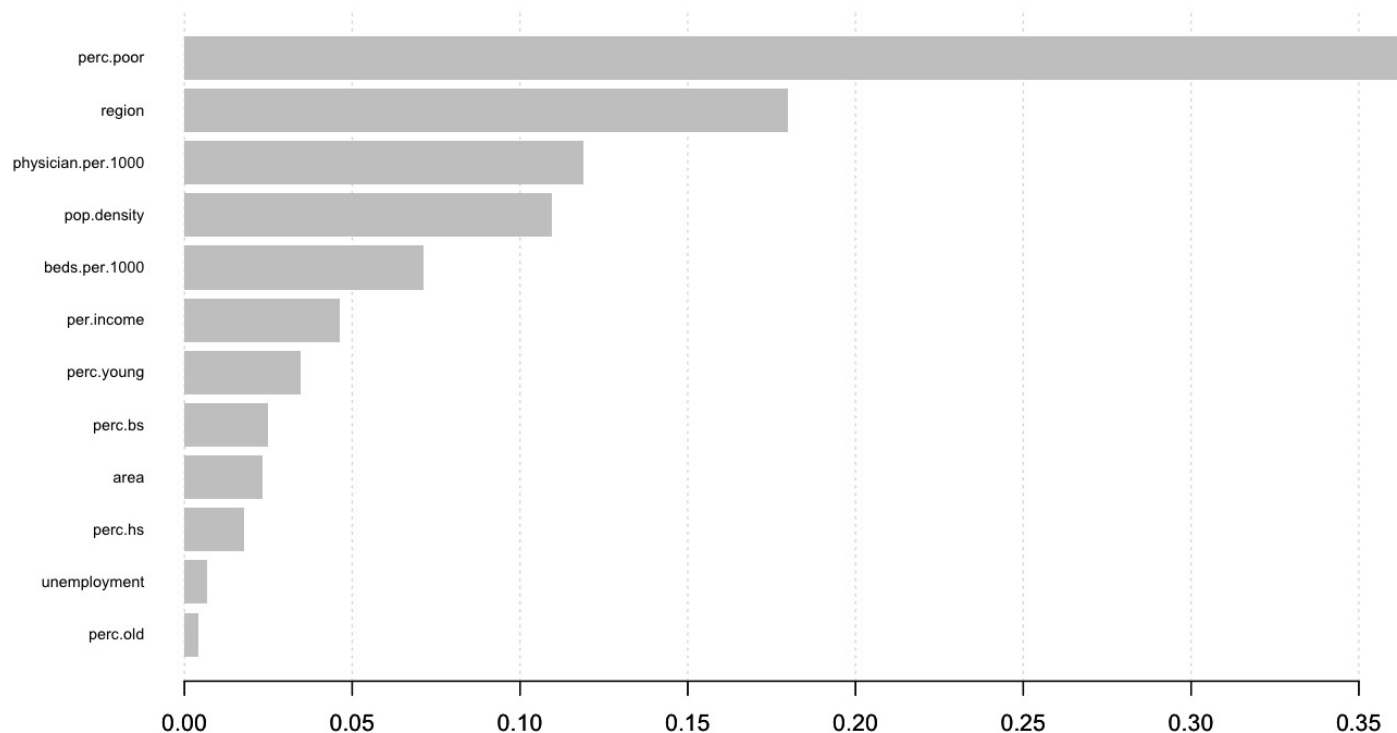To further explore the data, we fit our data into Random Forest/XGboost Model:

- The two models both fit the data well
- Only several points are outliers, which we don't know the exactly reason for their high crime rate
- Since our client - Kings County is also among the several outliers in both models, we have to do further analysis to find out the hidden reason for its high crime rate. Otherwise, our suggestions may not be applicable to Kings County.
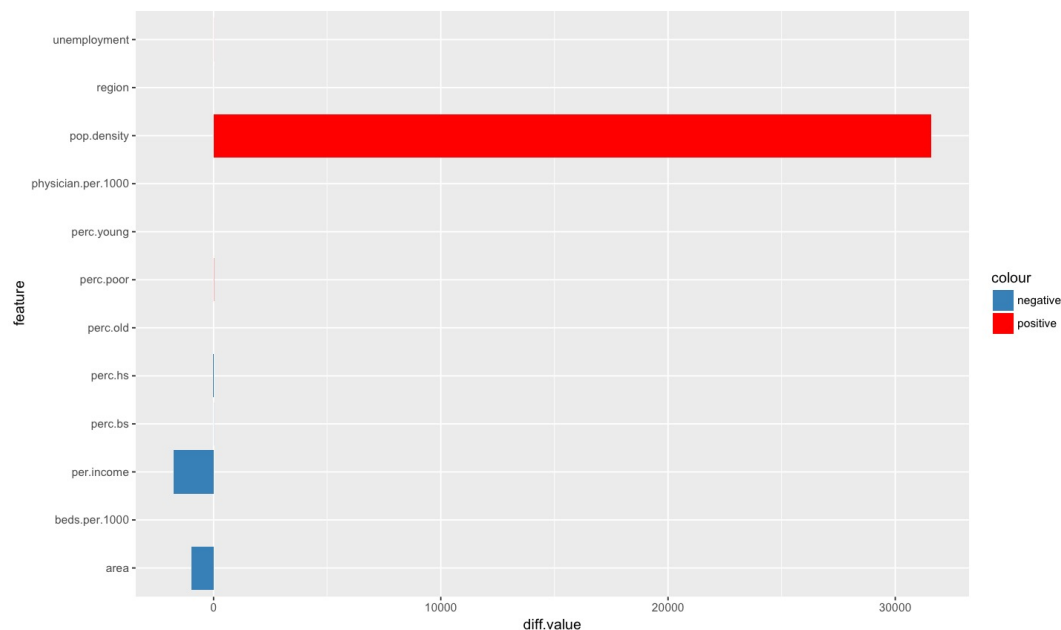
Most Important Variables

Let's see what are the most important variables:

We can see that among the several variables we focus on, Kings county has a significant high population density, which might be one of the reasons why Kings County has a high crime rate.
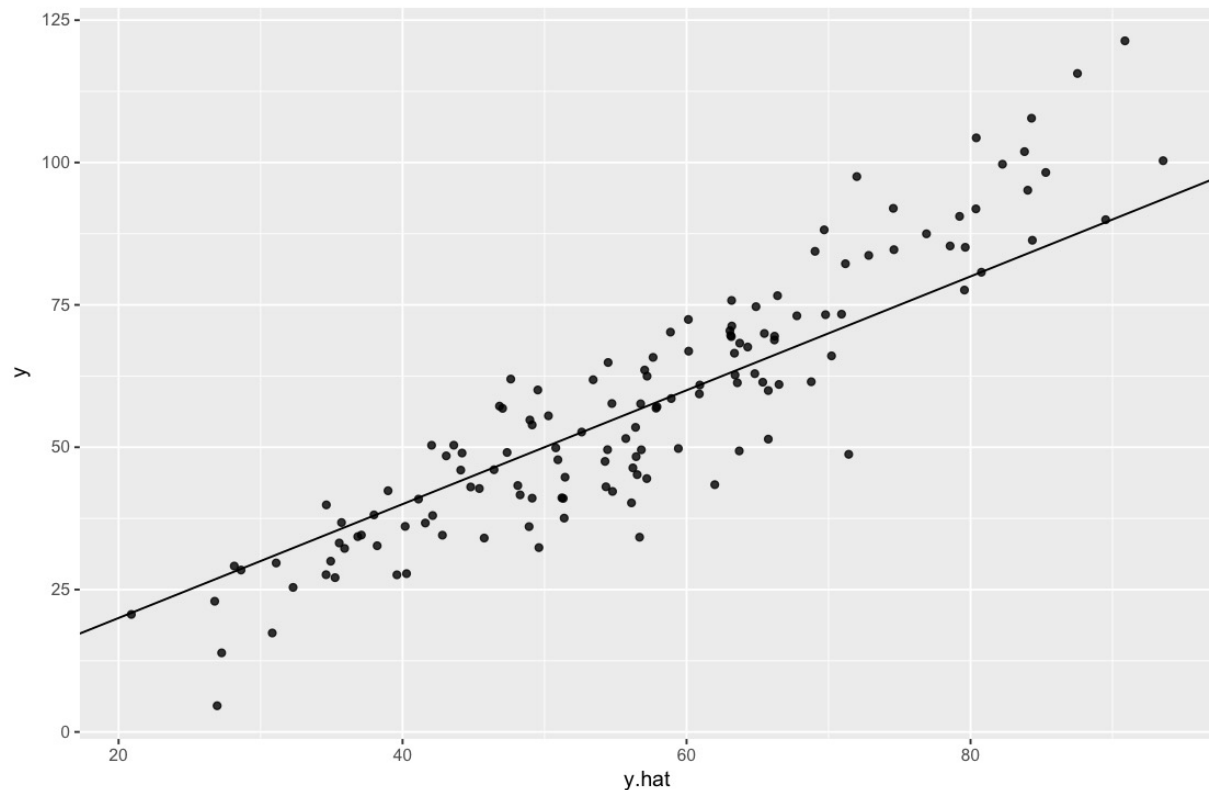
Assuming counties that are adjacent to each other might have more similarities, we see the difference between Kings County and other counties in the State of NY.
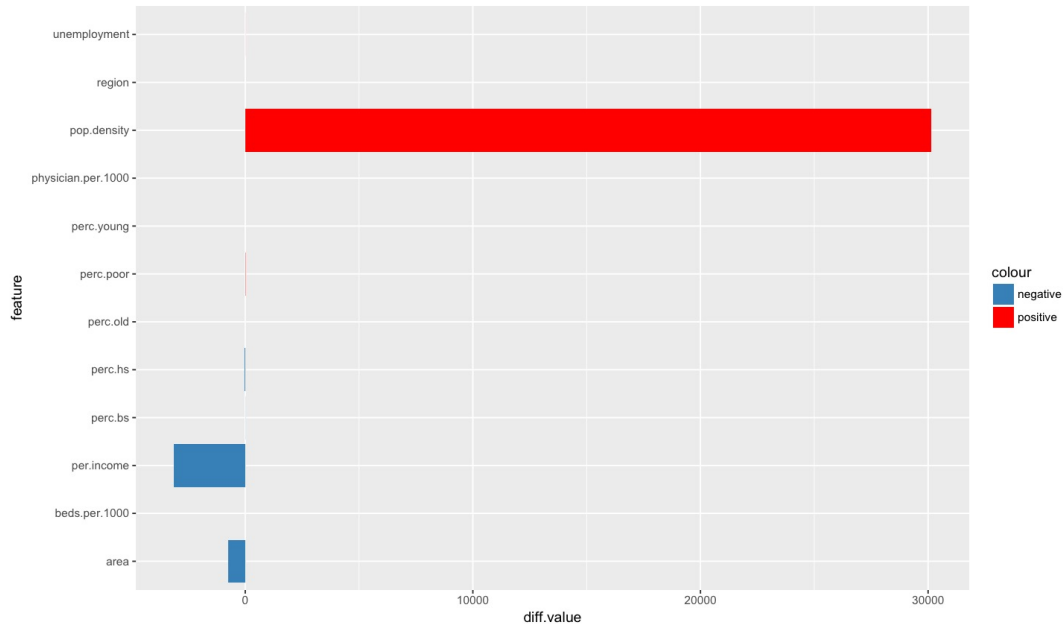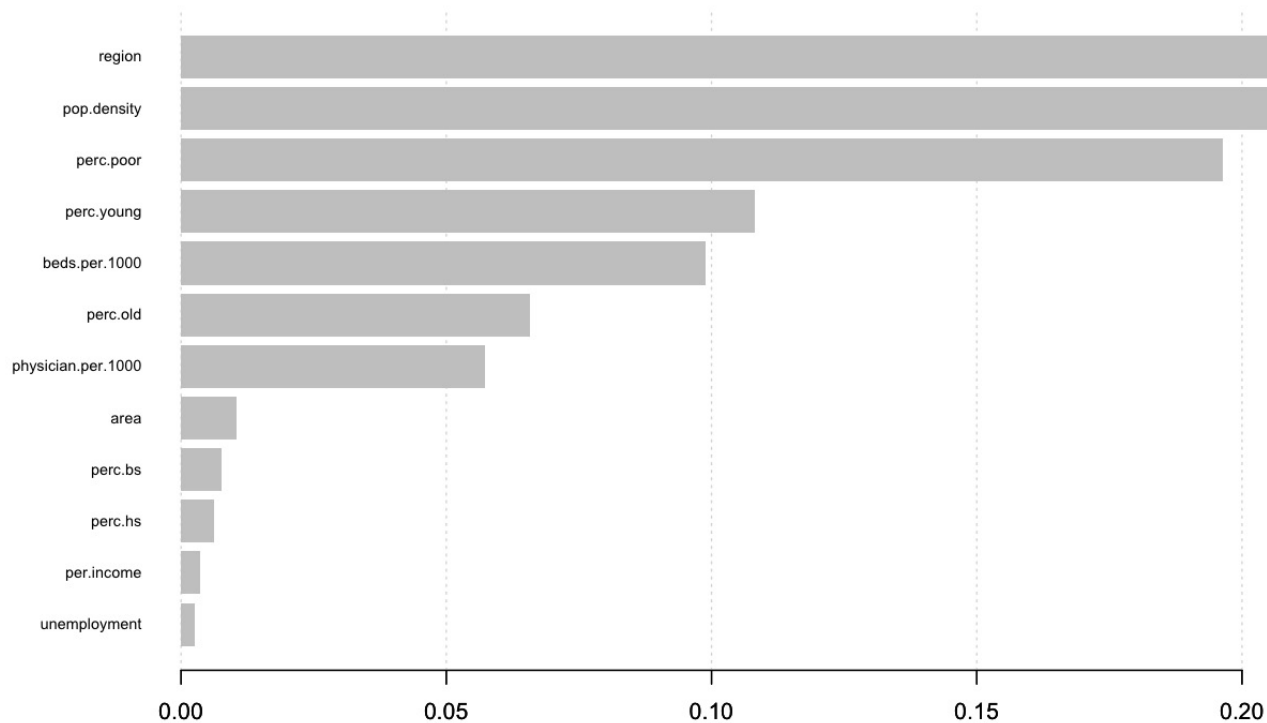

We can see that the biggest difference is still population density. Thus, we analyze population density's impact on crime rate.
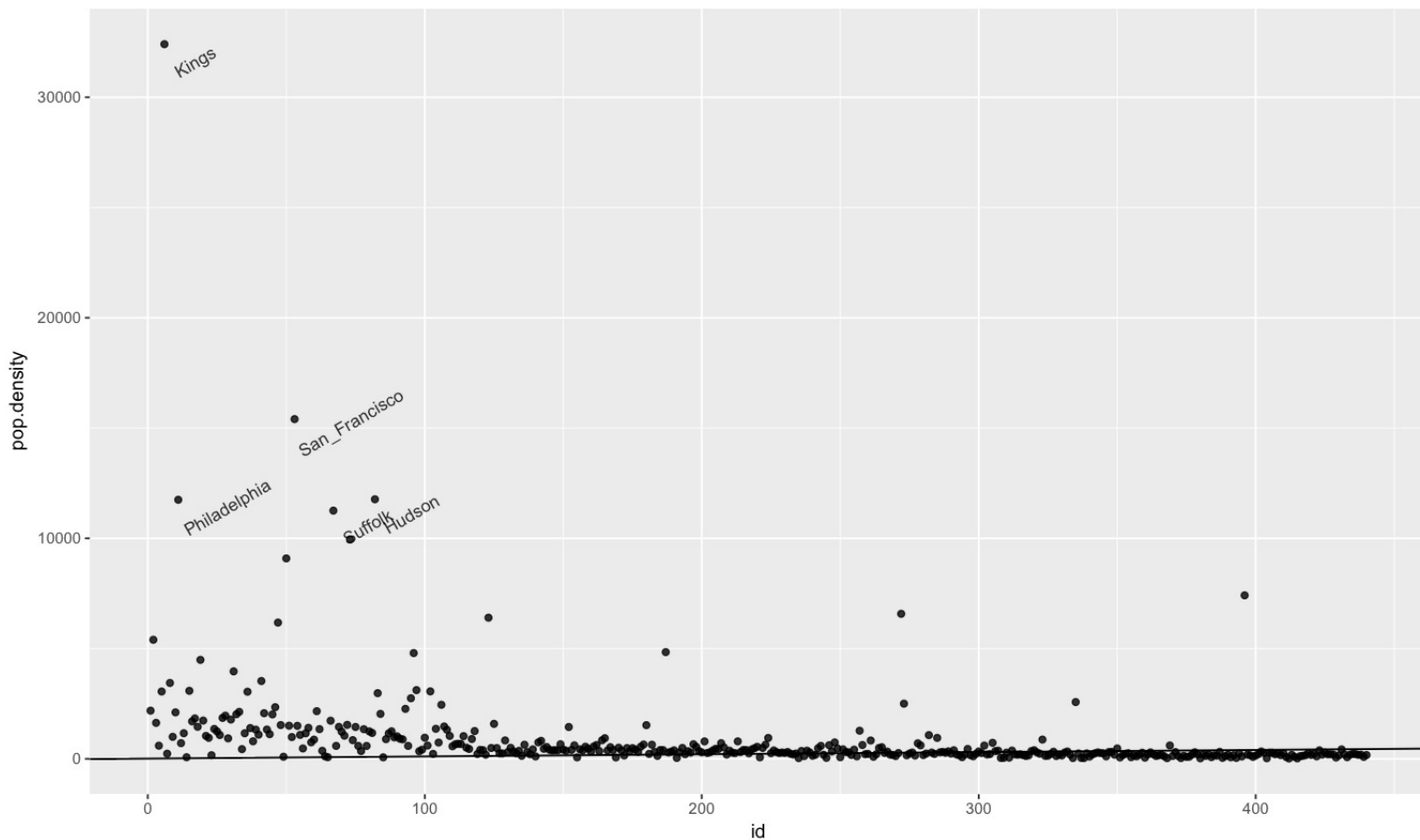
Regression on counties including Kings

Important Features

Population Density

Suggestions...

Based on the value of the parameters, we give the following suggestions to the officials of Kings County:

- 1. Control the population of King's County
- 2. Adopt better policy to raise the level of people's life quality.
- 3.

Improvements. . .

The Regression Model above may be fit for most counties in America, but it doesn't reveal the hidden reasons for the extremely high crime rate in some counties.

Social Economic Reasons

"Crime rates spiked in the 1980s and early 1990s as **the crack epidemic** hit the city."

Crime in New York City - Wikipedia
`http://bit.ly/2oYXTQQ`

Food For Thought

"New York City Crime in the Nineties - The New Yoker"

`http://bit.ly/2os9ZTQ`