Analysis of U.S. Regional Crime Rates

Ziwei Meng, Ao Liu

April 25, 2017

Outline

1. Overview
   - Goal and Procedure
2. Model Building
   - Data Overview
   - Data Processing
   - Heatmap
   - Regression Model
   - Random Forest/XGboost Model
   - Interpretation of Parameters and Visualization
3. Suggestions and Improvements

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

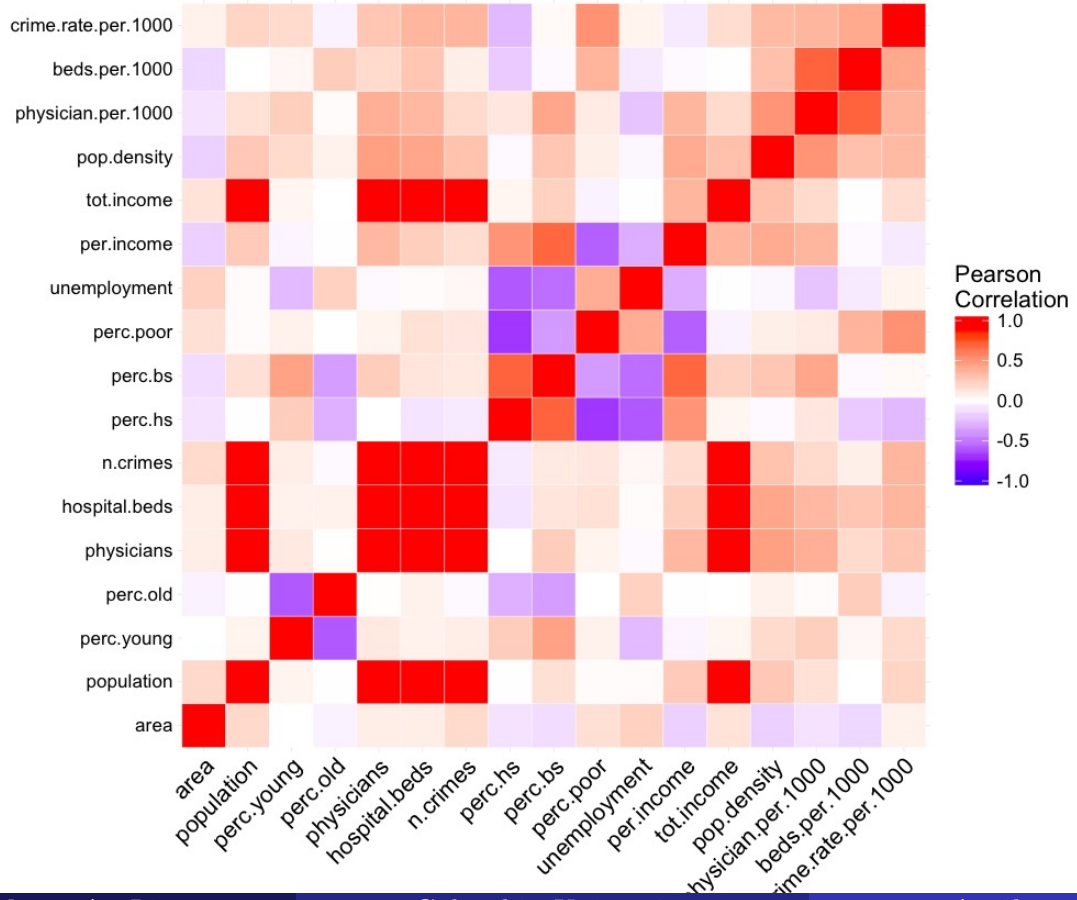
- **Geographic Data:** Land Area, Geographic Region
- **Demographic Data:** Total population, Percent of population aged 18-34, Percent Bachelor's Degree
- **Economics Data:** Percent Below Poverty Level, Total Personal Income, Per Capita Income

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

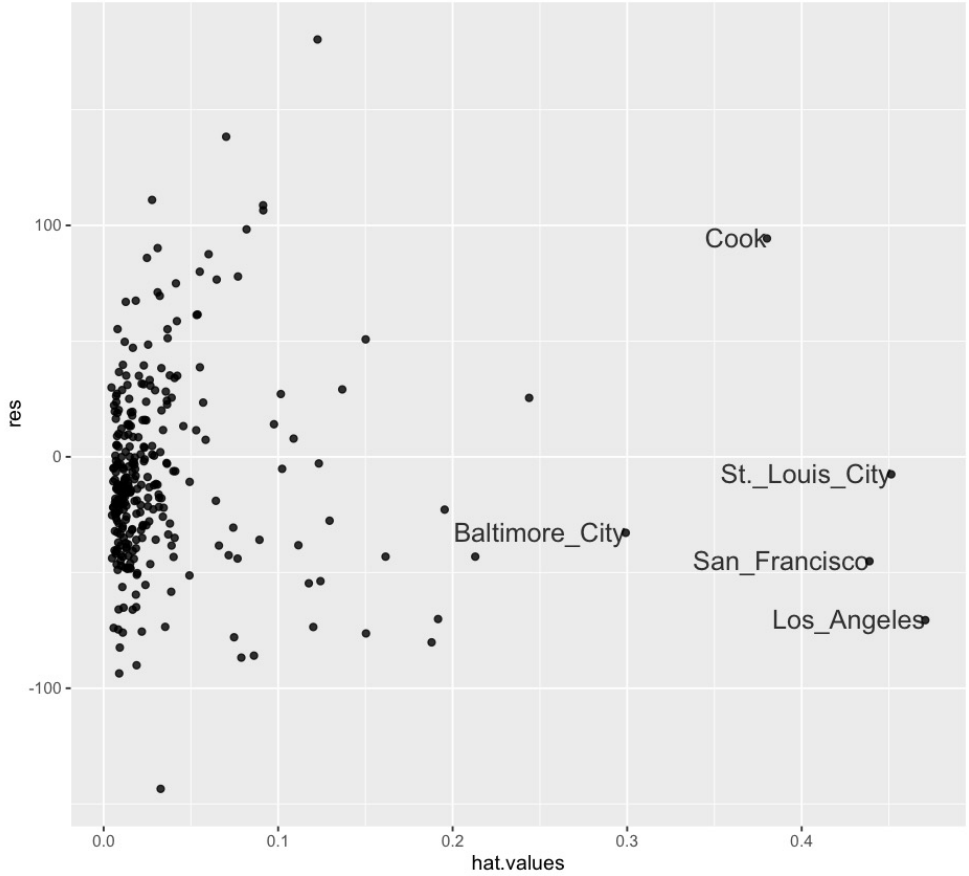
- Given 16 predictor variables, some of them are strongly correlated with each other, which will cause us to get some potentially false conclusion, thus we remove these variables.
- The remaining variables are:
  Area, *Percentage of Young People*, *Percentage of Old People*, *Percentage of High School*, *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 **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.

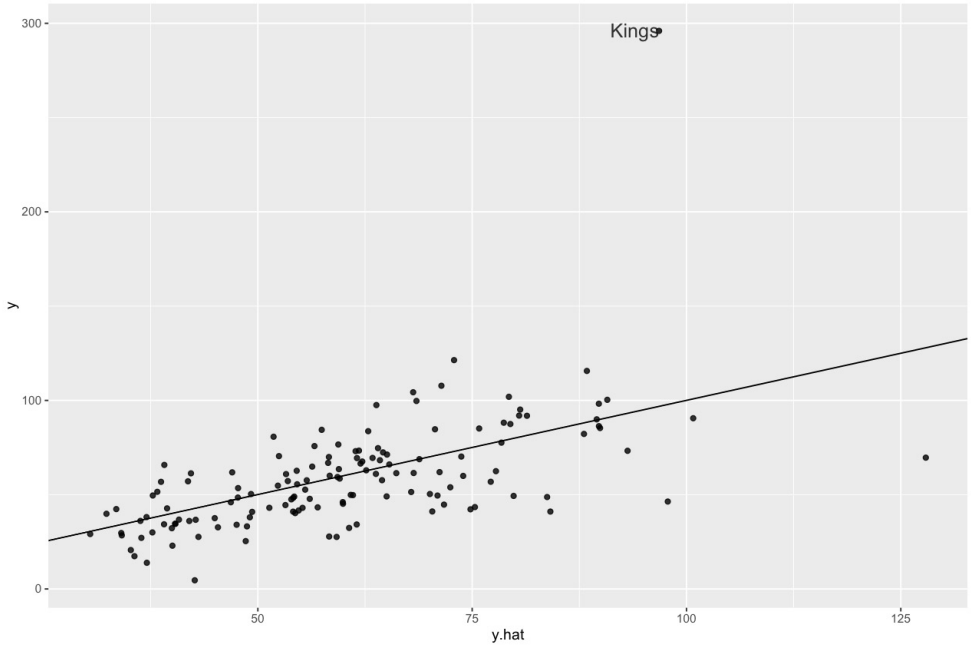
Outliers

Check outliers



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:

Interpretation of Parameters and Visualization

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

Suggestions. . .

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

- 1
- 2
- 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`