SSC 442

Final Project

Group 13

For our project, we analyze data of hate-crime from the website (<https://raw.githubusercontent.com/fivethirtyeight/data/master/hate-crimes/hate_crimes.csv>), and try to find best-fit models to interpret correlation between hate-crimes-per-100k-splc and several variables (“median-household-income”, “share-unemployed-seasonal”, "share-population-in-metro-areas", "share-population-with-high-school-degree", "share-non-citizen", "share-white-poverty", "gini-index", "share-non-white", "share-voters-voted-trump", "hate-crimes-per-100k-splc”, "avg-hatecrimes-per-100k-fbi”).

First, I clean the data and change NA data to 0 in order to make it easy for me to do analysis in the following steps. For part 1 of the code, I create 11 models to find which model could fit the data perfectly by using the function of RMSE, which is a great measure to know error between actual data and predicted data (show whether the model fits the data well or not). From the result of summary of the simple model (use all variables as independent variable except for hate-crimes-per-100k-splc), we can know that the variable “share-voters-voted-trump” and “avg-hate-crimes- per-100k-fbi” are statistically significant while t-value of both variables are greater than 2 (as we can see in the picture below). According to the summary of the simple model, we can see that “share-voters-voted-trump” is negatively related with hate-crimes, and with one unite increase of “share-voters-voted-trump”, there would be 0.96 decrease in hate-crimes as expected. The variable " avg-hate-crimes- per-100k-fbi" is positively related with hate-crimes, and with one unite increase in “avg-hate-crimes- per-100k-fbi”, there would be 0.07 increase in hate-crimes as expected. Then I create 11 models and calculate rmse of each model, and I plot the results along with the complexity of the model (see the second picture). For the first plot of complexity verses rmse, I was expecting for a downward sloping points, but the result is not the same as what I thought. And for the first model, I was expecting for a low rmse value since the independent variables of the model are “share-voters-voted-trump” and “avg-hate-crimes- per-100k-fbi”, of which the coefficients are statistically significant compared with other variables. Although there are only two results of variables are statistically significant, the best-fit model is not the model only contains these two variables as independent variables but the model using all variables except for hate-crimes-per-100k-splc (by comparing rmse of each model).

A screenshot of text

Description automatically generated

A screenshot of a cell phone

Description automatically generated

For part 2, I try to use rpart package to plot tree plot of the data to show the correlation between hate-crimes-per-100k-splc and other variables. First of all, since the best-fit model that I get from previous work is to use all else variables as independent variables except for hate-crimes-per-100k-splc, I continually use this model to plot tree plot. I use plotcp firstly to see the cross-validation results (since plotcp can provide a graphical representation to the cross-validated error summary. And the cp values are plotted against the geometric mean to depict the deviation until the minimum value is reached). From the plot below, we can see that the line barely changes as cp value changes, which shows low deviation of the data of dependent variable used in the best-fit model above.

A screenshot of a social media post

Description automatically generated

For the picture below, I plot the regression tree plot for hate-crimes-per-100k-splc, which shows correlation between hate-crimes-per-100k-splc between “share-voters-voted-trump” and “avg-hate-crimes- per-100k-fbi” at each node.

A picture containing screenshot

Description automatically generated