First part of our statistical process was to combine our EDA results together to come out with our data cleaning conclusions. This included data cleaning, predictor transformations and variable selection.

First on data cleaning, we decided to remove all missing values before model building which includes 52 missing values of medIncome and 21 missing values of pctEmploy. Furthermore, there were some entries in ownHousMed, rentMed, ownHousQrange and rentQrange which appears to be entered in place of a missing value which we removed as well. With this a total of 92 data points were removed.
With State and region variable, we dropped entries that had State variable being empty and combined Pacific region which only had 3 entries with the West region as it make most sense geographically.
After which, we did a series of transformation on our predictor variables that included Log2 Log2+1, Power, Root transformation. We also made categorical transformation to pctUrban where pctUrban greater than 85 were considered Urban.

On Variable selection, we concluded from our EDA predictor variables that were largely correlated, like pctKids2Par, rentMed, pctCollGrad and removed them. This solves the problem of collinearity between our predictor variables and we also checked it with Variance Inflation Vector to made sure that the problem of collinearity was solved. In the end, we are left with 12 variables for model building.

For model building, we picked Least Absolute Shrinkage and Selection Operator (LASSO) regression as our choice of model for two reasons.
First reason is that LASSO regression performs variable selection, this means that only variables that help us predict our outcome variables will be use in the model and the rest will be removed with their coefficients being zero. This reduces the number of predictors and allows our model have stronger explanatory power.
The second reason is that LASSO regression uses penalized likelihood strategy. This allows our model to not be overfitted to our training data and will perform as well with future data. However, this will also result in our model having higher MSE with our train data compared to a normal linear model but we believe that it is a worthy trade-off between bias and variance.

Compared to the other two model that can be used, Ridge regression and stepwise regression with AIC/BIC, they lack some benefits that LASSO regression can provide. For example Ridge regression does not perform variable selection and so does not help reduce the complexity of the model. On the other hand, stepwise regression is often guilty of overfitting the model to the training data, hence model built with stepwise regression will not work as well for future data compared to LASSO. Hence, we picked LASSO regression as our choice of model.

With our model built, we can analysis the cause of crimes and now we will move on to Oliver to explain the results of our model.