Stage 2: Peer Feedback

Reviewer: Group 2
Reviewee: Group 3

Introduction

The Objective behind the analysis and model building is clear and has been stated clearly along with the intended audience. The year of dataset is 1987 not 1994

Comprehensive analysis of variables followed by model building with specific research question helps build a strong story

Initial EDA

It would be helpful to see the outputs of the code prior to digging into the narrative especially the data cleaning part

Data Cleaning

- Nice work on finding the duplicates and missing data rows
- Knowing the final record count would help understand the sanity of the raw data and the magnitude of clean up that was required

Plausibility Checks for Variables

- Well thought-out arguments to explain the ‘prbarr’ and ‘prbconv’ over 1.
- Although not sure why prbarr >1 was filtered out and prbconv >1 was not
- Should look at all variables. If some are really meant to be neglected, reasoning to support the same would be helpful
- Besides the linearity, should we also explore the correlation among independent variables too? Should we have very highly correlated predictors, the coefficients will be less precision
- Very interesting explanation for county 115’s unusually low crime rate

Transformation Analysis

- Scatterplotmatrix can be used to look at multicollinearity, in addition to space savings. It would also provide numerical value of correlation, which is easier to comprehend
- No need for explanatory variables to have normality. Independent variable transformation should be based on linearity only. QQplot works well to determine anomalies. Can try histogram too.
- Squared variables are hard to interpret Model building. Why do you use square root of density and square of probability of prison sentence in analysis?

Model Building

- Are you losing details if you aggregate wages into groups? Don’t think we can aggregate all the wages variables into one by using just a simple average calculation. The base numbers of people (denominators) differ across metrics.
- Inclusion of new variables in model 2 and 3 supported by hypothesis stated about causes of crime adds a lot of value to the report
- Titles and labels on charts would help reader understand and link the charts with the narrative better

Selecting Models

- Can you please include comments about the helper functions? For example, how are they relevant to the analysis. It is clear that you will use them, but their purpose is never outlined.
- Good separation of models based on categories. Should include more variables incrementally (Eg. Moving model 4 -> model 2) instead of using final model to show parsimony

The Regression Table

How do you plan to solve the simultaneity issue of the police per capita variable?

The Omitted Variables Discussion

- The report includes a comprehensive list of potentially omitted variables.
- However, next step is to sort through them to identify those that have impacts on the outcomes of the model, and include some discussion of why they would have an impact.
- How we can accurately measure the police methodology and police representation?

Conclusion and Final Result

It would add value to the conclusion if you could identify one of the models as the best approximation of the problem statement and provide a reasonable justification to support the claim. Also, certain recommendation to increase in investment in certain information/ research made in this section would be more valuable with more justification