# W203 Lab 3 Peer Review

**Beau Kramer, Rory Liu**

### Introduction

We thought the introduction section was very clear and well-laid out. The team clearly explained the motivation behind the study, as well as the assumptions they made and the approach they plan to take. The section is a nice overview of the report.

### Initial EDA

In the data cleaning section, the team correctly identified the 4 anomalies (i.e. NA rows, prbconv coded as factor, unnecessary year variable and the duplicative record). The measures taken to correct these anomalies are well reasoned and justified.

The team then proceeded to remove a few outlier data points that have high leverage. We do not believe that the removal of these data points is justified by the reasoning provided. Having a lot of influence / leverage, and being far very all other data points are not sufficient to justify data removal. For example, if the authors can show that these data points are erroneous or invalid then their removal seems justified.

Additionally, as a reason for variable transformation, the team stated that a non-normal residual distribution will make inference from the model impossible. This is not necessarily true. Since the mode has large enough sample (n>30), they should be able to rely on central limit theorem.

Generally, the team's exploratory analysis on dependent variables are systematic and structured. We found most of the analysis compelling and transformations reasonable. We do have a few comments:

1. It was not very clear to us why the team decided to do exploratory analysis on some variables and not others. For example variables like "west", "central" were not investigated. This could be fixed with some explanation for the selection up front.
2. In the section for population density, the author removed 1 data point because it had a very small value. This may not be fully justified - the value was small, but not out the realm of possibility. Also, note that the maximum density in the data is 8 (ppl per square mile), which is also far too low for any typical city. The authors may need to consider that this entire variable may be invalid as given in the data.
3. In the section for "percent minority" and “percent young male”, it was not very clear why the team wanted to take the square root of the variables. The transformation seem unnecessary, and the interpretation of the transformed variable is rather unclear. As mentioned above, we did not believe that all variables with non-normal distributions need to be transformed, since our sample size is quite large. If they wish to continue with the square root then they should offer some interpretation of the transformed variable.
4. The correlation analysis section should be numbered 4 (not 5)

**Model building process**
We thought the variable table explaining which variables are used in which model is very clear and informative. The only variable that's a little unclear is "wfir", where the author did not comment on why this is selected for model 3 (among all the other wage variables). Additionally, briefly stating what the three models are at the start may help readers understand the flow of the report.

**Regression table**
The regression tables are clear, and the explanation on R2 and AIC are also clear. To improve, the team should add some comments on practical and statistical significance of the key dependent variables.

**Omitted variable discussion**
Generally we think the team did a good job on how the omitted variables are expected to correlate with crime rate. However, the reasoning behind the direction of the bias (and the interpretation) is less clear.

For example, the team argued that education has a positive bias on taxpc. However, the team stated that education has a negative correlation with crime (beta<0) and that education and taxpc would be correlated (delta>0). This means without considering other variables, education should have a negative, instead of positive bias on taxpc. For similar reasoning, we disagree with the bias direction for neighborhood watch and family cohesiveness on taxpc.

Also, education, neighborhood watch and family cohesiveness do not seem to have a readily apparent relationship with prbarr, prbconv, prbpris, and avgsen. If the authors believe such relationships exist, some more explanation here could be helpful.

**Conclusion**
The conclusion section of the report is very well presented. The team laid out clear implications the analysis has on policy - good job!