

1. *Introduction. As you understand it, what is the motivation for this team's report? Does the introduction as written make the motivation easy to understand? Is the analysis well-motivated? Note that we're not necessarily expecting a long introduction. Even a single paragraph is probably enough for most reports.*

What's the "so what"? Even just a sentence can help engage the reader. Also, instead of "believe", a stronger word might be "hypothesize". And, what key socio-economic variables, and why?

Good, clear research questions. However, for your second one, gear the question to your audience (I think it's a political party/candidate?).

2. *The Initial EDA. Is the EDA presented in a systematic and transparent way? Did the team notice any anomalous values? Is there a sufficient justification for any data points that are removed? Did the report note any coding features that affect the meaning of variables (e.g. top-coding or bottom-coding)? Can you identify anything the team could do to improve its understanding or treatment of the data?*

Reading in the crime dataset: Clear, easy to read. Consider changing to past tense (to describe what you already did.)

Processing: Very clear! One thing, in the lectures, it also said to consider not performing a log transformation if theory doesn't support it. Because percentage change in crime rates could make sense, half a sentence addressing this consideration could help cover all bases. Also, converting prbconv to a vector makes sense.

The Urban, Central West variables have some encoding issues, lots of counties are labeled "0" for all three variables, and some counties has more than one variable labeled "1" (i.e. both central and urban are "1"). The EDA did not analyze this and its implications.

Correlation matrix is a good way to go (I prefer it over a scatterplot matrix sometimes). However, since the professors use scatterplot matrix a lot, it might be a good idea to ask if this is necessary, so you don't lose points. With your correlation matrix, consider not just highly correlated variables, but also those whose correlations are statistically significant. (They probably are, but it's good to check just in case.) This is also a good place to briefly discuss the "meaning" of these correlations, are they supported by some theories, or they reveal some hidden relationships that was a "surprise".

For your univariate analysis, each step/plot needs some text telling us what you did, why you included it, and what this means. Summary statistics, in addition to the histograms, could help. For instance, they're a more objective way of determining skew than eyeballing a graph.

3. *The Model Building Process. Overall, is each step in the model building process supported by EDA? Is the outcome variable (or variables) appropriate? Did the team consider available variable transformations and select them with an eye towards model plausibility and interpretability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?*

The quasi Forward Stepwise Regression approach is a good way to do it...just make sure the instructors don't see this as incorporating a method not introduced in the class.

Consider moving your scatterplots to this section, as it's asked for here. Same with your log transformations (which do make sense to do).

For your models, more interpretation is needed. What does this mean, and why would a non-statistician audience member (i.e. the political candidate) care?

The plot is good and makes sense to do for your final draft. Make sure you do that for each version of your model when you test the CLM assumptions.

The outliers leverage was shown, it could be more informative to pick out the actual outliers and do further analysis on them. The outlier can be a very strong political campaign tool used as either model, or focus.

The choice of variables seems based heavily on correlation, there could be more meaning discussion, of the correlation, whether there is causal-effect relationship, an expected influence, a hidden/omitted variable linking them, or just unexpected, surprise finding?

4. *The Regression Table. Are the model specifications properly chosen to outline the boundary of reasonable choices? Is it easy to find key coefficients in the regression table? Does the text include a discussion of practical significance for key effects?*

The regression table was clear and made it easy to find the coefficients.

However, the labels could have used description instead of the column name like "prbarr".

Also it was in the conclusion at the very end, instead of immediately following model building, and needs a discussion of practical significance (not just statistically significant). Especially since this is for a political campaign, the discussion needs to be as free from jargon as possible. Assume your candidate will read/use it and has no background in statistics.

5. *The Omitted Variables Discussion. Did the report miss any important sources of omitted variable bias? For each omitted variable, is there a complete discussion of the direction of bias? Are the estimated directions of bias correct? Does the team consider possible proxy variables, and if so do you find these choices plausible? Is the discussion of omitted variables linked back to the presentation of main results? In other words, does the team adequately re-evaluate their estimated effects in light of the sources of bias?*

No omitted variables discussion was provided. The activity we did in class the week before spring break is a great place to start for this. Also look for zero condition mean, homoskedasticity, and normal error term violations. These can often highlight where omitted variables might lie. I suggest checking the CLM assumptions first, and using that to inform your omitted variable bias discussion.

There was a brief mention of a lack of any omitted variables in the conclusion. However, since the assignment specifically asks for a discussion, provide one, even if you have to get a bit creative.

6. *Conclusion. Does the conclusion address the big-picture concerns that would be at the center of a political campaign? Does it raise interesting points beyond numerical estimates? Does it place relevant context around the results?*

The discussion of your modeling included in your conclusion should go elsewhere, and instead add more of the “so what”. Your (non-technical) candidate should be able to skip straight to your conclusions and use it for a speech on proposed policy. Assume this person is completely innumerate and numbers freak him/her out. How would you communicate this?

The way it’s formed is currently not “actionable” campaign advice per se. From both political talking points, and policy stance perspective, what can be done for a political campaign to benefit from this research? For example, tougher laws, invest in local prosecution, and how is “west part of the state” affect campaign? (invest additional resources to the west? Target these ridings, or avoid? )

One thing I do, is ask a non-technical friend/family member for a sanity check (often it’s my mom). If her eyes glaze over and/or if she’s just polite, I’m being either too technical or not interesting. However, if she’s engaged and/or asks good follow-up questions, then I’m on the right track. Hopefully this will help you!

*7. Throughout the report, do you find any errors, faulty logic, unclear or unpersuasive writing, or other elements that leave you less convinced by the conclusions?*

Nothing major. The one minor issue is the west-central-urban coding, there were counties labeled as both west and urban for example, and majority of the counties are set “0” for all 3 regions. If this was noted, the variable choosing process might have considered twice on how to use “west” as one of the independent variables.

The issues with the writing that I did find are addressed above: a mix of not really there and/or a very technical and brief way of discussing results. It would be helpful to include some causality analysis, for example higher-government employee wages were positively related to higher crime rates, its interpretation, implication to policy could be discussed further.