Zach Merritt Review of Group 1's ***Excellent*** Report (Alan Tan & Marry Boardman)

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.
    a. The introduction begins on page 3. I'd recommend moving the introduction to the beginning so that you don't jump directly into the code.
    b. The introduction does an excellent job laying out the group's hypotheses. The analysis is well motivated as they connect their hypotheses to the larger mission of eventual policy recommendations.
    c. I would probably refrain from using the words "hope" and stick more to hypotheses and expectations.

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 datapoints 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?
    a. The team identified the missing data and properly identified it. As a general note, it's better to remove data using actual criteria instead of identifying the missing data and using indexing. If, for some reason, the CSV you are loading in has its order changed by a team member, than your code would get rid of the wrong data.
    b. The team's overall treatment of the data is precise and informative.
    c. They continually connect their EDA to the larger questions at hand, which is helpful.
    d. I'd be careful of giving too much weight to outliers in your explanations. For example, in the discussion of crime rate, you discuss the outlier for county #115. You state that you might be able to learn something from this county. However, this is only one record and it's possible that this county's low crime rate is by chance. You don't want to cherry-pick data and try and learn from it.
    e. Note: on page 6 your code goes off the page. Try and make sure you enter the next line when one line of code is getting long.
        i. This happens throughout the report.
    f. I'm not sure the scatter plot matrix on page 16 is helpful. I think it's just too much information. I'd limit the grid to only show the distribution of the economic variables and then show only a few correlation plots of interest.

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 interperability? Are transformations used to expose linear relationships in scatterplots? Is there enough explanation in the text to understand the meaning of each visualization?
    a. The group does an excellent job discussing their prospective models throughout their EDA. They use linear models for EDA, which I thought was great.
    b. The group discusses independent variables in groups, which helps avoid just dumping out a bunch of plots.
    c. The group approaches the independent variables and models them with crime rate. They try a transformation of the variable and see how R^2 is affected. I am not knowledgeable enough to comment on this and I actually did things kind of similar in my lab. I was concerned that this may fall under the umbrella of P hacking. I'm not sure. I'll be sure to ask in class.
    d. The discussion on multicollinearity is too short and oversimplified. There are variables that have a high correlation with one another that I think are worth mentioning.

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?
    a. Yes, the regression table is formatted properly and very informative.
    b. I'd discuss the exact findings of the model in more detail. I'm incredibly guilty of this in my report as well. I think both of us spend a disproportionate amount of time and space on EDA. Upon further reflection of my own work, the lab prompt, and course content, I feel it is much more important to focus on the model and the results of the model.
    c. The group discusses the implications of the model in broad terms but does not offer practical facts about what the model means (an increase in X leads to an increase in Y by Z).

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?
    a. I did not see any discussion of omitted variables. The most obvious one that I thought of was income.

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?
    a. The group only compares the success of the models using R^2. However, R^2 in unfairly affected by the number of variables in your model. I'd recommend adding AIC measures to more accurately compare your models.
    b. I didn't understand this sentence: "This was not a surprise, although it did not work as independent variable as we hoped."
    c. I'd discuss the variables at greater length (see #4 for more on this).
    d. One note on this: "For example, we expected police per capita to be a cause of dropping crime, but it turns out it was more (likely) a result of higher crime rates."
        i. I've read a little bit about this topic in the past and there's actually an interesting and horrible reason for this. More police will lead to police *catching* more crime, not an increase in crime. So, we send more police to minority areas and, while actual crime may be happening at similar rates, the crime rate in minority areas will go up. This actually has a snowball effect: as more arrests will lead towards more fathers/mothers being taken away from families, which will lead to children growing up in worse home lives, which leads to more crime, which leads to the city sending more police, which leads to more crime and so-on.
        ii. This is what makes this analysis so difficult. Everything is connected.

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?
    a. Anything that I found, I discussed above.

I thought this report was excellent and extremely organized. I think the process, models, and conclusion were all well founded and insightful.
Great work! I look forward to reading the final report.

Thanks,
Zach