Math 402/403 Final Project Peer Review (2)

Project Authors: Benjamin Busath, Josh Hiatt, Eric Manner

Reviewers: Clark Brown, Sam Cochran, Daniel Swingle

Motivation and Overview of the Data: 6/10 (10 very important points!)

Is the purpose of the project clear? What questions do the project try to answer? How could the objective be clarified?

- It seems that the objective is to find out how fair the justice system currently is and how to improve it, but all the data is from a century ago. It seems like the project has more of a historical focus, and that might need to be better reflected in the objective statement. Then, perhaps a great next step would be to use current data and show how past incarceration rates and lynchings can continue to be used to predict current incarceration rates--drawing the connection to how the past continues to influence how the justice system works in the present. But as it is, it seems that the project really answers questions mostly from the past.

The author should *briefly* review what is already known about the research questions and what techniques others have used to study these questions. How well does the author explain the scope of the project and how it fits into existing research? How could the explanation be improved? Does the author cite any existing research?

- N/A
- You should do this!

Does the author give an overview of the data? Do they evaluate the validity of the data source? Do they defend why this data set is relevant to the project objective? What is missing from the overview of the data?

- They give a pretty good overview of the dataset and how they are using it in their project.

Data Collection and Cleaning: 15/15

The author should explain how the data was collected and include any corresponding code (possibly in an auxiliary file). Is the collection procedure clear? Are there any missing references or potential legal problems with the way the data was gathered?

- The dataset is publicly released census data from 1900-1940, so the data has no legal problems.

The author should explain how the data was cleaned and include any corresponding code (possibly in an auxiliary file). Is the cleaning procedure clear? Does the data need more cleaning, and if so, how? If there is little or no data cleaning, the author should strongly defend why no cleaning is needed.

- The data was cleaned previously for "other research projects here at BYU" (probably in the Record Linking Lab in the Department of Economics). There isn't a lot of detail on that process, but that makes sense. They defend not needing a lot of cleaning because of that.

How does the author handle missing, badly formatted, or incorrect data? Does the author justify their choices of what they removed, edited, reformatted, or left unchanged?

- They drop a couple states from their data because of missing data; because this is old census data, this makes sense and while not optimal is probably the best they can do with what they have.

The author should justify any newly engineered features. Are there any additional features that should be engineered? If there is little or no feature engineering, the author should strongly defend why no feature engineering is needed.

- The feature engineering seemed very reasonable--they added features like incarceration rate and percentage of black police officers, as well as percentage of black judges and attorneys. They addressed the potential issues with this data when they talked about certain issues with census data in general, but these features seem well justified and sensible. Other added features like the control variable for economic well being of counties and residency of police officers seemed solid.

Robustness: 10/10

If the author provided code, evaluate the readability of it throughout the project, especially code for scraping or cleaning. Do they comment their code neatly? Are variable names reasonable? Does every function have an appropriate docstring? Can you tell what the code does just by looking at it? How would you improve the code to make it more readable?

- The code throughout has a lot of comments and is pretty readable. If you are planning to cut out a lot of the code, it would probably be helpful to put it into functions that you can call in the notebook. You could also consider hosting the code in a github repository so readers can find everything if they're interested.

How robust is the cleaning and scraping code? Could the code be easily modified so that it is usable for similar data sets, or so it could handle the sa

me data set but with more data points? Identify parts of the code that are not general enough and make suggestions for improving it.

- It looks like most of the code for cleaning and engineering works with specific data frames with specific column labels, etc, and won't generalize well. This is probably not really an issue, and I guess if they got data from more recent years they would be able to adjust it to have the right labels so the code would continue to work.
- As far as cleaning, it looks like they dropped a lot of null values, and so the ability of their code to generalize would likely depend on the quality of the data they tried to generalize it to.

Data Visualization and Basic Analysis: 10/15
The author should use summary statistics and visualizations to display and describe the data with some depth. What parts of the data are still mysterious or unexplained? How well do the visualizations and analysis contribute to answering the questions from the introduction? Are there any questions that are not addressed adequately?

- The visualizations look great, but it would be nice to have more of a conclusion from the plots. If you can, try to find a way to have the plots tie into your conclusions in the Algorithms and Advanced Analysis section on page 10. Other than that, it looks like there are visualizations for each part of the paper.

Are the visualizations readable? Do they convey the information clearly, and in an aesthetic way? Identify any visualizations that do not make sense, that should be made with a different kind of plot, or that can be simplified or otherwise improved.

- I liked how they used geopandas, which is a perfect plot for this kind of data.
- It's confusing to have the 'coolwarm' color scale for your incarceration rate plots on page 7 unless the shift from red to blue happens at meaningful value. Maybe try using something like you use for all the other scales.
- In general, the visualizations are great but need some interpretation below them. You talk about different features over time by county, and the plots do a great job of mapping that, but it is hard for the reader to really draw meaningful conclusions on their own by comparing one page to the next. You should consider putting some more text in that, in addition to explaining what data each map shows, also tells the reader why that is important.

Evaluate the validity of the analysis and conclusions about the data. Does the author avoid bias? Are any of the conclusions statistically unsound? Point out anything that you find suspicious.

- There isn't much analysis or conclusion making about the data.

- The data and visualizations are there, they just need to add some analysis to explain it to the reader. For example, what the heck is going on for incarceration rates in 1930? How do the changing trends across decades give us information about the overall fairness or unfairness of the justice system? Overall, why is each set of maps important. This section is great, but needs more of this interpretation text.

Learning Algorithms and In-depth Analysis: 5/15

At this checkpoint, this section is not expected. If the authors have provided material for this section, give feedback roughly using the same metrics as above

- The OLS analysis they wrote is good, but you should probably exclude the statsmodels printout, as it is hard to read and could be summarized better elsewhere.
- It would be good to put in results for some of the other algorithms we learned about this semester, even if they didn't get very good results. Currently they only have the OLS analysis.
- You are getting warnings about possible multicollinearity in the OLS output. This could be a good reason to try and make some visualizations that compare data against each other and not just having a single feature displayed by county and do feature reduction. Or come up with a justification for the warnings if they do show in your final submission.
- At the end you say that though the system is supposed to be fair, your results paint a different picture. This is great, but that painting can be made more apparent to the reader if you include more analysis saying what the OLS told you--there wasn't much of that as it is so it was hard for us to understand your conclusions.

Quality of Communication: 12/20 Find and correct grammatical mistakes.

Is the presentation clear and easy to follow? What parts of the project are the most confusing, and why? Make suggestions for improvement.

- Please do some formatting. It will make it a lot more readable. It looks like you have some long sections of text that need to be split up into paragraphs where there is an extra large space, but something may have gone wrong in the formatting of it. Remember to separate paragraphs by a blank line in markdown otherwise it gets appended to the previous text.
- Also the headers will be converted better if you follow the markdown hierarchy. Each top level heading like "Algorithms and Advanced Analysis" should be "# Algorithms and Advanced Analysis" and that will make the numbering in LaTeX make a lot more sense. You could consider some subheadings for organization to help clear up some of your longer sections.

- Remember the 10 page limit--looks like they're gonna grade pretty strictly on that. I imagine a lot of this will be solved if you put some of the code in supplementary files, but be sure that it doesn't go over after you add your analysis.

Identify wordy passages and suggest ways to shorten them. Where does the author use an entire paragraph to say something that could be said in a single sentence? Are there any $50-words that could be replaced with a 50¢-word? Is there any unnecessary repetition?

- At the beginning, there are a lot of long paragraphs that could be broken up to make it more readable.

Ethics: 10 / 10

Did the authors analyze the ethical implications of their research questions, the data gathered, and the analysis performed? Are there privacy or other implications from the collection or use of the data that the authors did not notice?

- The authors did a good analysis of the ethical implications of the data used and the feature engineering done. As they add more analysis, they should make sure to think about any ethical implications of what they conclude.
- For privacy concerns, it may be worth mentioning that the census years that you are using have been made available by the US Census Bureau in keeping with their policy of publishing census records after 70 something years.

Could the results and methods from this project be misused or misunderstood? Did the authors address this possible misuse?

- They addressed the fact that it is a charged issue right now, but did a good job of presenting the data fairly. As long as they continue with this trend in the final analysis section, they are on a good track.

Is there any other feedback about the ethical implications of the project you wish to give?

- N/A

Impressiveness: 10/10

If you are an employer, how much does this project make you want to give the author a job?

- I think it was a pretty impressive project, on an issue that is very important right now. It was impressive that you collaborated a little bit to get data from the professor of the U of Illinois, so that is cool.

How could it be improved to make you more eager to hire the author? Explain.

- To make it more impressive, I would again recommend bolstering your analysis at the end, and talking about other models you tried besides just OLS, even if that is the best model and the one you end up focusing on when you talk about your results. And talk more about your results and implications at the end.
- You could also draw a connection to how looking at historical data has implications for the present situation and can give insight into how to make the justice system fair going forward.
- It is an awesome project, just the analysis and algorithms section need to be pulled together a bit for the final draft.

Other comments:

- You should provide references for where your data came from at the end. If it is Record Linking Lab data you could just attribute it to the lab.

Total Score: 78/105 (90 for this checkpoint)