CMPU_250 Peer Review
4/18/25
Group 6

Feedback for Group 5

Introduction

The *Introduction* section follows a clear and cohesive structure, with well-defined subsections that help the reader grasp the context effectively. The mention of the research gap enhances the credibility of your study. To strengthen the introduction further, consider incorporating visualizations—particularly for claims such as "...linked to increased default rates" or data on the number of students relying on federal loans. These visual aids can also be used consistently throughout the paper to support your personal findings.

Also, the second paragraph of the introduction largely rephrases the ideas presented in the first. You might consider condensing it or using that space to introduce additional background information.

In the *Research Questions* subsection, the first question could be tightened by removing the clause after the comma, as it doesn't add much value. For the second question, where you include "...by HBCU status" in parentheses, consider specifying what the different statuses are to make it clearer for readers unfamiliar with the terminology.

In the *Expected Outcomes* section, you could strengthen your hypotheses by referencing more literature or specific findings that align with your expectations. Avoid repeating content already covered in the introduction—use this section to deepen the theoretical or empirical support for your anticipated results.

Data

The Data section describes the source of the dataset well and the cleaning process is well-detailed, such as eliminating variables and changing variable names to something easier to understand. I think some improvements could be made, such as the handling of missing values (imputation or exclusion) are not fully clarified, which could affect reproducibility. Additionally, while the paper notes downloading and extracting zip files as CSV and using dataset documentation, I do not think these steps are necessary to be mentioned in the paper. The renaming of vague variable names (BBRR4_FED_PELL_DFLT to Default Rate) is good, but referencing variables by their original, technical names in the text makes it harder for us to understand. I would just use the simplified, easier to understand variable name in the paper. I'm curious if there were variables in the data on the average scholarships given to students or endowment, especially because endowment was being correlated or a possible explanatory? Including such variables, if available, could strengthen the analysis by directly linking financial resources to loan outcomes. Overall, the data preparation is very good, we just had a couple of comments to make it easier for readers to understand your paper.

Methods

The methods section lays out the steps taken to answer each research question clearly, but there are a few things that could improve readability.The formulas on pages 6 and 7 would benefit from better formatting. Right now, they're hard to follow, and readers don't want to have to reference the key in order to understand what is happening. There's also a mention of different plots (like box plots, bar plots, scatter plots, etc.), but none of them are actually shown in the paper. It might make more sense to briefly mention that visualizations were used to get a sense of the data and then focus on the conclusions that were tested using statistical methods. For example, instead of listing all the plots made, you could say something like, "Exploratory plots suggested some trends, which we then tested using t-tests and regression." That would help keep the section more focused on the analysis. Alternatively, you could just include the graphs in the paper.

Results

The results are well framed with a succinct summary of the questions being answered. In section 7.1, the application of the t-test and interpretation of the results are logically sound. In section 7.2, you present a contrast between the relatively low difference in median debt at HBCUs ($1600) and the fact that students at HBCUs experience much higher default rates. Given the explanation that follows, it doesn't seem necessary to make a disclaimer about these values – as you say, there are numerous factors besides the actual amount of debt that could contribute to default. In section 7.3, you do a good job of explaining the meaning behind the figures you are presenting. However, since you describe many of the numbers as marginally significant, it might help the reader to add a definitive concluding sentence about the extent to which these calculations proved or disproved your hypothesis (which you can then elaborate further in the discussion section). In section 7.4, I would consider rephrasing "individual-level data" to be "individual data points" or "individual observations." As mentioned the other day, it might be helpful to be clear about how you're interpreting synthetic data. Does the relationship between high GPA and low default rate come from real evidence? The conclusion to this section is well-written and concise, and I like how you call out the possibility of other factors and support disparities.

Discussion

Overall, it would be more fruitful if your discussion focused more on real world impacts/claims that can be derived from your results. The second paragraph seems a bit repetitive as it restates the results, instead it would be interesting if you guys related it back to the hypothesis and tried to quantify or explain how these values should be interpreted in the context of college admissions by looking at how big the differences are, how skeptical we should be, what might be causing the overfitting, and how to fix these model issues. The third paragraph is a little unclear because while it mentions model fit improving while HBCU becoming not statistically significant, I am not entirely clear on the models that are being referenced nor what this means

for your hypotheses. Would also be interesting to add ethical considerations behind the student-level data when mentioning limitations in data. Lastly, during your edits to your prelim analysis would be interesting to take a look at endowment size especially because it seems included in your dataset and if HBCUs overall have less endowment than non-HBCUs it seems that HBCUs could be a proxy for endowment or school financial status. There is also a latex formatting issue on page 11, where the italicization inadvertently created a new line.