PDA Reflection

Liangkang Wang

Reflection

Code Reproducibility

I made missing heatmap for each projects. Now, I wrote a Utils.R file, which can be sourced to store some common useful functions. In the three projects, all of them suffer from missing data problem. I have relocated the tran_brier function from project 3.

Project 1

Project 1 was an exploratory data analysis, an essential step in my journey to understand and interpret complex datasets. This project was not just about analyzing data but also about refining my approach to data presentation and interpretation. In the initial submission, I encountered several challenges, particularly in formatting and data processing, which I addressed in the updated version. This process was instrumental in enhancing my skills and understanding the nuances of effective data analysis.

Addressing Formatting Errors

In the first iteration of the project, I noticed some formatting errors, including inaccuracies in the title of the project report and the subtitles for each section. These errors were corrected in the revised version, underscoring the importance of attention to detail in presenting analytical work. This modification taught me the significance of precision in reporting, as it directly impacts the clarity and professionalism of the output.

Enhancing the Introduction with Data Collection Details

I realized the omission of data collection information in the introduction of the original report. By adding these details, the report gained context and depth, making it more informative and comprehensive. This addition was a valuable lesson in the importance of complete transparency in data analysis, fostering greater understanding and reproducibility.

Improving Report Compactness and Clarity

Initially, the investigation of variable distributions involved creating five separate plots, leading to an unnecessarily lengthy report. By combining these plots, I made the report more concise and reader-friendly. This modification highlighted the importance of effective data visualization techniques in conveying information succinctly.

Detailed SDP Variable Preprocessing

In the section on Smoking During Pregnancy (SDP) variables preprocessing, I enriched the report by detailing the transformation processes of the original data and adding exact correlation values of specific variables. This enhancement not only improved the report's accuracy but also its instructional value. It taught me the critical role of detailed methodology in research, allowing for better reproducibility and credibility.

Deeper Insights from Existing Analyses

In the revised conclusion, I delved deeper into the results of the existing regression models. This involved a more nuanced interpretation of the data, looking beyond the basic statistical outcomes to consider the broader implications of these findings. I explored how the specific effects of SDP and ETS, as indicated by the regression models, might translate into real-world impacts on adolescent behavior.

In summary, the revisions and enhancements made to this project were not just about correcting errors but were a profound learning experience. They emphasized the importance of thoroughness, clarity, and advanced analytical techniques in data analysis. Each modification brought a new level of understanding, from the importance of detailed reporting to the complexities of interpreting interactions within data. This project has been instrumental in advancing my skills and will undoubtedly influence my approach to future data analysis endeavors.

Project 2

Refining the Abstract

To provide a clearer and more direct understanding of our study's objectives, I have revised the abstract to include a more concise depiction of our outcome variables. This enhancement ensures that our aims and the focus of our research are immediately clear to readers, setting a well-defined context right from the outset.

Enriching the Introduction with References

In the introduction, I have incorporated a reference to a relevant paper, thereby enriching this section with additional scholarly context. This addition not only bolsters the introduction with more comprehensive information but also anchors our study within the existing body of research, providing readers with a broader academic perspective.

Clarifying the Creation of Composite Outcomes

To aid reader comprehension, I have added an explanation in the 'Outcome Variable' section of the data description about why we combined two outcome variables into one composite outcome. This clarification is crucial for readers to understand the rationale behind our methodological choices, enhancing the transparency and replicability of our research.

Addressing Missing Values and MAR Assumption

In the section on handling missing values, I addressed the Missing At Random (MAR) assumption. This addition provides a deeper insight into our approach to dealing with incomplete data, highlighting the statistical considerations and assumptions that underpin our analysis.

Incorporating Random Effect Model

A significant enhancement in this project is the addition of a random effect model for the variable "center." This involved integrating the relevant code and formula into the report, alongside the existing analysis. I noted that the lme4 package, used for this model, does not support lasso regression directly. Consequently, while we utilized mixed models with stepwise selection, lasso regression was not feasible in this context. The final report now includes the output from the mixed model, specifically focusing on the random effect of the center, without extensively altering the report's structure.

Rounding Evaluation Metrics

I refined the presentation of our evaluation metrics by rounding them to three decimal places. This change makes the data more reader-friendly and easier to interpret, avoiding the confusion that can arise from overly lengthy numerical values.

Integrating Model Coefficients into the Main Report

Finally, I moved the coefficients of the model from the appendix to the main body of the report. Discussing these coefficients within the text allows for a more integrated and comprehensive analysis, enabling us to delve deeper into the implications of our findings and how they contribute to the broader context of our research field.

Project 3

Incorporating Essential Citations in the Introduction

To uphold academic integrity and provide proper context, I have added citations for the Framingham and NHANES datasets in the introduction. This not only addresses potential plagiarism concerns but also tell readers about the origin and significance of these datasets. Recognizing the importance of proper attribution, especially for datasets that are not universally known, reinforces the credibility and scholarly value of our work.

Refining the Abstract

The abstract has been revised to clearly articulate the comparison objectives of the study. I included a detailed description of both simulation-based and data-based methods, enhancing the abstract's effectiveness in conveying the primary focus and methodologies of our report. This revision ensures that readers can quickly grasp the essence of our study, its methods, and its intended comparisons, right from the beginning.

Adjusting the Missing Data Heat Map

In the missing data heat map section, I removed variables like SEQN, SYSBP_UT, and SYSBP_T, as they were not part of the original NHANES dataset. This modification ensures the accuracy and relevance of our visual representation of missing data, aligning it more closely with the dataset's actual structure.

Elaborating on ROC-AUC Analysis

Recognizing the need for clarity, I added explanations of the AUC (Area Under the Curve) and ROC (Receiver Operating Characteristic) formulae in the report. This addition is particularly important in the section where we assess the original model performance using the Framingham dataset. By explaining these concepts, I aimed to enhance readers' understanding of how we evaluated model performance.

Streamlining Presentation of Results

In our analysis of ROC-AUC, I rounded the minimal numbers in the plot for better readability and added a table to present transportation Brier scores, replacing the direct R code outputs. This change makes the presentation of our results more accessible and interpretable, focusing on conciseness and clarity.

Revising the Correlation Matrix Plot

The correlation matrix plot has been updated by removing SYSBP_T and SYSBP_UT. This revision aligns the plot more closely with the data used to generate the simulation, ensuring that our visualizations accurately reflect the variables under consideration in our study.