Baruch College
Zicklin School of Business
Paul H. Chook Department of Information Systems and Statistics
STA9750/OPR9750
Software Tools for Data Analysis
Fall 2023

Final project instruction:
Choose one interesting real data set which requires cleaning and define one real data problem that you want to analysis and find the answer for. It would be better to use one quantitative variable as your y-variable to apply skills we learned through out this semester but it is totally up to you to decide the real data set and real problem you want to work on.

Your written report must be no longer than 10 pages including figures and tables. Make sure to submit the report and relevant material to the blackboard before **Dec 19th (Tue), 11:59pm**. The report **must cover**:

1. Introduction: One page of the description of the data that you selected and present the real problem.
2. Data Cleaning: Use R to clean the data and use the complete data for further steps, e.g., remove all the missing values, combine variables, define new variables, transform variables, etc.
3. Association Analysis:
    - Decide **one variable** as your y-variable that you want to predict.
    - For the other remaining variables (except for the y-variable), select 5 mostly related variables to your y. Generate visuals that summarize the associations where one of the variables is always y-variable, e.g., If you have 6 variables $(x_1, \cdots, x_5,\ y)$ in your data set, you should produce 5 visuals which show the relationships of $(x_1, y), \cdots, (x_5, y)$. Check what variables show interesting/meaningful/strong associations.
    - Do the statistical tests to check whether the associations you observe from all the visuals are statistically significant.
4. Regression Model: Fit regression models that predict your y-variable and assess/interpret the fitted models, e.g., compare/assess different fitted models, interpret the fitted regression models, etc.

5. Other Techniques: You can be creative and do further analysis (beyond what we covered in class) which you feel is helpful to find answers to the real problem. In practice, you would need to learn and use new R packages by yourself and this is the best time to practice that. Please feel free to use other R packages to analyze your data.
6. Conclusion: Based on your analysis, provide the best finding/answer to the real data problem.
7. Appendix I: Summary of what role each team member had for the project.
8. Appendix II (separate files): Provide four things, 1. Real data, 2. R codes you used to produce the results, 3. Presentation Slides, 4. Final report.

Each group will present results to the class. Each group will have **roughly 10 minutes** to present your interesting findings. In your presentation, use as many graphics as you need. Your presentation should focus on:

1. Description of the data.
2. Association analysis.
3. Regression models.
4. Assessments and interpretations of regression models.
5. Other techniques you tried for this project.

Below is the grading rubric:

- Presentation: Clearly introducing the real data, stating the real problem, showing/summarizing results or findings.
- Accuracy: Accuracy in terms of how you clean the data, conduct the association analysis, fit the regression models, interprete/assess the regression models, etc.
- Creativity: Try new techniques and think beyond the traditional way to solve the real data problem. Do not be afraid of trying new things. There are no correct answers. Think freely and try new ways. It may turn out to be better.
- Reproducibility: I will run your R code with your real data and checking whether I can generate exactly same results which are included in your final project.
- Team Work: How (fairly) each member contributed to this final project.

Here are some links you may find a real data set:
Federal Reserve Bank of St. Louis: `https://research.stlouisfed.org/`
Kaggle: `https://www.kaggle.com/datasets`