MATH113/CAIS105: Intro to Data Science

Fall 2023

Mini-Project 03

Learning Objective: Analyze a topic of interest with multiple data tables

Part 1

Work in groups of 2-3. Pick from one of the following topics for your analysis. Each option includes at least 2 datasets, and you are welcome to add more if you'd like.

- Olympics: https://www.kaggle.com/datasets/heesoo37/120-years-of-olympic-history-athletes-and-results
- NYC Test Scores and Demographics: https://catalog.data.gov/dataset/2012-sat-results, https://catalog.data.gov/dataset/2017-18-2021-22-demographic-snapshot
- Boston Property Assessments: https://data.boston.gov/dataset/property-assessment
- IMDb Datasets: https://developer.imdb.com/non-commercial-datasets/

Your analysis should consist of questions that you can answer with your data (ex., "Do SSA counts of births match Census counts of births?"). At a minimum, your analysis must include:

- Pivoting longer or wider (or both) to create tidy data tables
- Joining two (or more) data tables

You should complete your analysis in an R Markdown file. Be sure to include text chunks in your file where you write out the questions you're investigating and the answers you find from your analysis. In the file, label the code chunks where you completed the requirements above.

You will give a 5 minute presentation of your findings. You presentation must include:

- The questions you asked and their answers
- At least one visualization
- The type of pivot(s) you used to tidy your data and why
- The type of join(s) you used and why

Submission

Save your R Markdown as a PDF and submit as a group through PLATO.