DATS 6101: Introduction to Data Science
Midterm Project Outline (Fall 2025)

Description and Purpose

The goal of this midterm project is to gain a deeper understanding of the initial stages of data-focused research. This involves conducting background investigation, developing SMART questions, performing Exploratory Data Analysis (EDA), and applying statistical testing to address these questions.

Each team will select a research topic and question. Rather than collecting new data, teams will use existing datasets found online or from other sources. Datasets must include at least 3,000 observations (rows of data).

Teams must submit a topic proposal one week before the presentation is due. Feedback will be provided by classmates, the TA, and the instructor within two days of the presentation. Teams can then revise their final submission. A research paper of no more than 5,000 words (excluding TOC, figure captions, etc.) will be submitted in HTML format. Visuals like graphs do not count towards the word limit.

Project Details:

I. Topic Proposal

Due: October 15th (Wednesday, End of Day)
Submit a 150–200-word proposal on Blackboard, including:

- **a)** The research topic
- **b)** The SMART question(s) of your research (can be revised later)
- **c)** The source(s) of your dataset(s) and approximate number of observations
- **d)** The link to your team's GitHub repository

II. SMART Question Development

Develop a research-driven SMART question focused on a dataset from R or an online source. The dataset must have at least 3,000 observations (rows).

III. Exploratory Data Analysis (EDA)

Due: October 22nd (Wednesday, End of Day)
Submit an R Markdown file, knitted into HTML, which includes:

- R code, explanations, and rationale for your EDA
- A technical overview of your analysis, including:
  - Dataset summary
  - Descriptive statistics
  - Graphical representations of the data
  - [If applicable] Variance/SD measures
  - [If applicable] Normality tests
  - [If applicable] Initial statistical tests (e.g., correlation, Chi-Square, ANOVA, Z-test, T-test, etc.)

Also, submit your dataset or provide the URL for the data source.

IV. Presentation

Due: October 22nd (Wednesday, Class Time)
Prepare a 10–15-minute team presentation summarizing the initial stages of your data science project. The presentation should effectively communicate your findings to the class.

V. Research Paper

Due: October 29th (Wednesday, End of Day)
Submit a 10-page summary (max 4,000 words, excluding visuals) of your research and EDA process. Use R Markdown and knit it into HTML. Include elements from previous parts of the project (e.g., graphs, results). Address the following:

- What does the dataset reveal?
- Dataset limitations
- Data collection methods
- Previous analyses on related topics
- How gathered research contributed to question development
- Additional information that could enhance the analysis
- Changes to your question after EDA
- Preliminary insights based on EDA
- References (APA style preferred)

Grading Breakdown:

1. **Topic Proposal**: 5%
2. **SMART Question Development** (see Part V): Included in summary grading
3. **R Markdown (RMD) & Technical Analysis**: 25%
4. **Presentations**:
    - Individual presentation score: 25%
    - Team presentation slides: 10%
5. **Research Summary (Part V)**: 25%
6. **Git Usage**: 10% (graded individually based on activity in the repository)

Note:

- Grades for parts I, II, III, IV (partially), and V are team-based.
- The instructor reserves the right to assign different grades to team members if there's evidence of unequal contribution.
- A peer evaluation form will be submitted individually by all students after project completion.