Milestone 2: Data Collection - To-Do List

This checklist outlines the detailed steps and deliverables for our team to successfully complete Milestone 2: Data Collection. Use this document to track our progress and ensure all requirements are met by the June 30th deadline.

Detailed Next Steps for Your Team

1. Team Meeting & Review (Early June 17 - June 19)

- [ ] Review Workshop Concepts: As a team, thoroughly discuss the `Key Concepts from "Modeling the World Through Data" Workshop (June 17).pdf` document. Ensure everyone understands the principles of data modeling, the importance of critiquing data, and the distinction between data and knowledge.
- [ ] Review Data Selection Guide: Go through the `Structured Approach for Evaluating and Selecting Datasets (Milestone 2).pdf` document together. Understand the criteria for evaluating datasets and the proposed data selection matrix.
- [ ] Align on Strategy: Discuss and agree upon the team's strategy for approaching data collection, including roles and responsibilities for evaluating the identified datasets.

2. Dataset Evaluation & Selection (June 19 - June 23)

- [ ] Deep Dive into Candidate Datasets: Assign team members to conduct a detailed review of the most promising datasets identified in Milestone 1 (Kaggle: Online Course Student Engagement Metrics, SED, GitHub: eLearning Datasets, edX Open Learning Dataset). Follow the criteria outlined in the `Structured Approach for Evaluating and Selecting Datasets (Milestone 2).pdf`.
- [ ] Fill the Data Selection Matrix: Collaboratively fill out the data selection matrix (or a similar comparison tool) to systematically compare the datasets based on relevance, quality, limitations, and practical considerations.
- [ ] Decision and Justification: As a team, decide on the primary dataset(s) you will use. Document our decision and provide a clear justification for our choice, explaining why it is the best fit for our actionable research question and how it aligns with the principles of data modeling and critique.

3. Data Acquisition & Initial Exploration (June 23 - June 25)

- [ ] Acquire Data: Download the selected dataset(s) to your local development environment.
- [ ] Load Data: Use appropriate programming tools (e.g., Python with Pandas) to load the data and perform initial checks to ensure it loads correctly.
- [ ] Preliminary Data Exploration: Conduct a preliminary exploration of the data. This involves:
  - [ ] Checking data types and basic statistics (e.g., `df.info()`, `df.describe()`).
  - [ ] Identifying the number of rows and columns.
  - [ ] Looking for obvious missing values or inconsistencies.
  - [ ] Visualizing basic distributions of key variables (e.g., histograms, bar charts) to get a feel for the data.
  - [ ] This step is about understanding the data's structure and content, not yet deep analysis.

4. Data Cleaning & Preprocessing (June 25 - June 28)

- [ ] Address Missing Values: Decide on a strategy for handling missing data (e.g., imputation, removal of rows/columns). Document our decisions.
- [ ] Handle Outliers: Identify and address any outliers that might skew our analysis.
- [ ] Data Transformation: Perform necessary transformations (e.g., converting data types, normalizing/scaling numerical features, encoding categorical variables).
- [ ] Feature Engineering (Initial): Based on our research question, consider if any new features can be created from existing ones to better represent student interaction patterns or performance.
- [ ] Document Cleaning Steps: Keep a detailed record of all cleaning and preprocessing steps in our code (e.g., in a Jupyter Notebook or Python script) and in our documentation.

5. Data Documentation (Ongoing throughout Milestone 2, Finalized by June 29)

- [ ] Data Dictionary: Create a comprehensive data dictionary for our chosen dataset. This should describe each variable, its data type, its meaning, and any units or categories.
- [ ] Data Source & Collection Methodology: Document where the data came from, how it was collected (if known), and any relevant details about its origin.
- [ ] Cleaning Log: Maintain a log of all data cleaning and preprocessing steps, explaining the rationale behind each decision.
- [ ] Ethical Considerations: Reiterate any ethical considerations specific to our chosen dataset and how our team plans to address them.

6. Data Hosting (June 29 - June 30)

- [ ] Repository Integration: Ensure our cleaned and documented dataset (or a representative sample if the full dataset is too large for GitHub) is properly stored in our GitHub repository. If the dataset is very large, provide clear instructions or a script for how it can be accessed or downloaded.
- [ ] Version Control: Use Git to track changes to our data and code throughout this process.

Deliverables for Milestone 2 (Due June 30)

- [ ] Decide how to model your problem domain in data: This will be reflected in our chosen dataset and the features you select/engineer.
- [ ] Decide what data is relevant to our research question: Documented in our data selection justification.
- [ ] Collect, clean, document, and host a data set to study: This is the core deliverable, encompassing the cleaned dataset, its documentation (data dictionary, cleaning log), and its presence in our repository.
- [ ] Maintain our planning documents: Continue updating group norms, learning goals, constraints, and communication plan.
- [ ] A retrospective for this milestone: Conduct a team retrospective at the end of Milestone 2.
- [ ] A labeled Git tag for this milestone: Create a Git tag (e.g., `milestone2-submission`) before the deadline.
- [ ] Complete the milestone survey: All team members complete the survey.