Introduction

In an ideal world, students would have access to "real" or "real world" data for assignments and projects. However, real data is difficult to acquire and very messy. Sometimes it doesn't meet the requirements of the particular skills that instructors hope to help students develop.

Effective use of AI can provide realistic data that replicates the experience of working with real data, while also allowing customization and easy updates each semester to provide a fresh experience.

Anecdotally, as of 2025, the five best generative AI tools for working with quantitative datasets, especially with a detailed prompt as outlined here, are (in my order of preference) Claude, JuliusAI, ChatGPT, Gemini, and CoPilot.

Summary of the DATASET Framework

- D - Define the Persona & Learning Objectives (Steps 1-2)
- A - Articulate the Assignment Context (Step 3)
- T - Tailor the High-Level Structure (Step 4)
- A - Add Variable Details & Nuances (Step 5)
- S - Specify Inter-Variable Relationships (Step 6)
- E - Establish the Output Format (Step 7)
- T - Task the AI with Assignment Text (Step 8)

(Note: This acronym was generated by Gemini 2.5 Pro, unprompted, when I asked for feedback on my detailed framework.)

Details: Developing AI Prompts for Generating Datasets for Assignments

Note: I strongly suggest to build your complete prompt before submitting. AI tools generate more coherent datasets from comprehensive initial prompts than from iterative refinements. If the output needs adjustment, start a new chat with a revised complete prompt rather than requesting modifications.

- Step 1: Give the AI a "persona"
    - For a more realistic dataset, tell the AI they are an industry professional (e.g., "You are an expert Power BI analyst.").

    - For a more textbook-like dataset for basic practice, tell the AI they are a professor.
- Step 2: Define Learning Objectives *(note: AI performs better when objectives are phrased as actions, not topics)*
    - Determine what analytical skills or concepts students should practice and what tools they should use to do so
    - Determine the cognitive level (e.g., Bloom's taxonomy: understanding vs. evaluating vs. creating)
    - Identify the complexity level appropriate for the students/course
    - Identify the scope (e.g., small enough for classroom demo or larger for assignments or big projects)
- Step 3: Define the assignment that the dataset is to accompany.
    - If the assignment is already written, upload or copy/paste the assignment as part of this prompt.
    - If the assignment does not yet exist, include information about the assignment as part of the prompt. Determine and state the following:
        - the type of assignment (e.g., case study, practice problem, etc.)
        - the level (introductory undergraduate, core/major undergraduate, graduate level, etc.)
        - For example, "Provide a case study in the Harvard Case Study style for an undergraduate senior-level course.")
- Step 4: Define the High Level Data Structure and Format
    - Determine the number of variables (columns) needed and the name of each
        - **Note**: If this is unknown, then write a preliminary prompt for the AI tool asking for ideas. For example, "in a dataset from [context], what would be typical variables? The context could be different industries (e.g., healthcare, non-profit, etc.). If you are not sure the context of the data you would like to use, write another preliminary prompt (e.g., "In an assignment where students learn [x] by working with a dataset, what type of data would be most engaging or interesting, i.e., what topic, industry, or context would the data ideally originate from"?
    - Determine the number of records (rows) needed (i.e., sample size)
- Step 5: Define Details for Each Variable (if known)
    - Define the range and distribution
    - Determine the data type: text, number, dates, locations, etc.
        - For text, determine formatting details such as length, capitalization
        - For numbers, determine format such as number of decimal points, currency symbols, etc.

        - For dates, determine correct formatting, and specific time periods and whether seasonal patterns should exist
    - Determine if outliers should be included and to what extent
    - Decide if missing values should be incorporated
    - Decide if there should be data entry errors, typos, or formatting inconsistencies if students need to practice cleaning messy data
    - Consider issues of diversity, bias, and ethics. With sensitive variables, specify whether the data should reflect balanced representation or real-world skew.
- Step 6: Define Relationships between Variables
    - Should some variables explicitly be correlated?
    - Should values of one variable depend on another? (e.g., costs should be less than prices)
    - If desired, specify what insights students *should* discover and/or any red herrings or non-significant relationships to include for critical thinking
    - Add a note to "Review that all relationships make narrative sense and validate internal consistency. The dataset should not contain any values that are logically impossible" (Sometimes AI produces logically inconsistent data if this is not explicitly reinforced.)
- Step 7: Define the Output
    - Choose a file format (e.g., csv, Excel, json, SQL database, etc.) for the data
    - Ask the AI to include a brief data dictionary explaining each variable.
    - If desired, ask for a narrative introduction to accompany the data.
    - Reiterate in the prompt that the dataset should feel realistic
    - If the number of rows you desire is large, ask the AI to provide the first 10 rows as a preview before generating the full dataset.
- Step 8: Provide additional details about instructions that should accompany the dataset.
    - This should provide additional detail beyond what was listed in Step 3 above, e.g., desired style, length, and formatting of the assignment. For example, "Generate a draft of the assignment with one paragraph under the heading 'Purpose,' 3-5 paragraphs under the heading 'Task,' and one paragraph under the heading 'Criteria'."

Reusability

Save a copy of the final prompt that you use, so that you can use the prompt each semester (with slight modifications as needed) to have a fresh dataset each time.

Template and Examples

AI Dataset Generation Prompt Template for Educators

Instructions: Fill in each blank or bracketed section below. When complete, delete any remaining blue/Italic text, including headers, then copy the entire remaining text into your AI tool to generate the dataset and any accompanying materials.

Step 1. Define the AI's Role (Persona) (note: defining as an industry professional leads to more realistic data than when you define the AI role as a professor)

You are a [industry professional / professor / data analyst / domain expert] specializing in [industry or field, e.g., healthcare analytics, marketing, finance, operations]. Your task is to generate a realistic dataset suitable for [type of class or assignment, e.g., a graduate-level analytics project, an introductory Excel lab, etc.].

Step 2. Define Learning Objectives

The dataset should allow students to practice [specific analytical skills or actions, e.g., performing regression analysis, creating dashboards, applying data cleaning, building predictive models]. The objective is to help students [describe cognitive goal: understand, analyze, evaluate, create, etc.] using [specific tools or methods: Excel, Power BI, SQL, Python, etc.]. The dataset should be designed for [complexity level: introductory, intermediate, advanced] and appropriate for [student level: undergraduate, graduate, etc.].

Step 3. Define the Assignment Context

This dataset supports a [case study / practice exercise / project / lab / simulation] for [course level, e.g., undergraduate core course].

> *If available, paste the assignment text:* The assignment instructions are as follows: [Insert assignment text here]

If not, describe the assignment: Students will [describe what students will do with the data, e.g., analyze customer churn, visualize sales patterns, forecast revenue, identify operational inefficiencies].

The dataset and materials should align with this assignment and be suitable for [scope: classroom demo / full assignment / multi-week project].

Step 4. Define Data Structure and Format (repeat as needed if you want datasets with multiple tables)

Create a dataset with [number] variables (columns) and [number] records (rows). [The variables should include [list variable names if known]. [Specify how tables are related if you are generating multiple tables.]

If unsure about variables or you would like additional variables, state: Please suggest [additional] typical variables for a dataset in [context or industry] that support [the learning objectives]. *[Note: Instead of a paragraph here, this could be a separate prompt that you run before you finish writing this prompt.]*

Step 5. Define Variable Details (if known)

For each known variable (or the key ones):
- Name: [variable name]
- Type: [numeric / text / date / categorical / boolean]
- Range or categories: [specify values or range if known]
- Formatting rules: [e.g., currency symbol, date format, capitalization]
- Outliers or missing values: [yes/no, with explanation]
- Data quality issues: [include typos / formatting inconsistencies / none]
- Ethical/diversity considerations: [balanced representation / real-world skew / N/A]

Step 6. Define Relationships Between Variables (note: each of the categories in this step is optional; fill out if known or desired; the last statement is required)

- Correlated variables: [list or describe correlations, e.g., “sales increase with advertising spend”]
- Dependencies: [describe dependencies, e.g., “cost < price”]
- Intended insights: [what patterns students should find, e.g., “seasonal demand peaks”]
- Non-significant relationships: [any distractors or red herrings]

Ensure: All values must be logically consistent and realistic. Validate internal consistency. The dataset should not contain any values that are logically impossible.

Step 7. Define the Output Deliverables

Generate the following:
1. Dataset file in [CSV/Excel/etc.] format.
2. Data dictionary explaining each variable (include units and definitions).
3. Optional narrative introduction (~1–2 paragraphs) describing the scenario, industry, and dataset context.
Again, the dataset should feel realistic and internally coherent.
If the requested dataset is large: Preview the first 10 rows before generating the full dataset.

Step 8. Define Assignment Instructions to Accompany Dataset (Optional)

Include brief assignment text for students:
- Purpose: [one paragraph summarizing learning intent]
- Task: [3–5 paragraphs describing what students must do using the dataset]
- Criteria: [one paragraph describing evaluation criteria or deliverables]
If relevant, specify: Use the same writing style or structure as a [case study / lab / applied project / report].

Final Instruction to AI

Generate the dataset and accompanying materials according to the specifications above. Ensure all values, relationships, and formatting are internally consistent and realistic. Output the dataset in [chosen format] and include the data dictionary and scenario narrative.