**Capstone Project Proposal (Data Analysis Pathway)**

**Project Title:**

**Proposed By:***Lauren Sapp*

**GitHub Repository:**

**Project Overview**

Briefly describe the overall purpose of your project. What are you trying to analyze, explain, or uncover? Why does this matter? Write 2–3 sentences that summarize the big picture.

**Data Sources**

Dataset 1:

* **Dataset Name: Audible Complete Catalog**
* **File Name(s):** Audibly\_Catlog.csv, Audible\_Catlog\_Advanced\_Features.csv
* **Source:** Kaggle
* **Relevant fields**: Book Name (PK), Author, Rating (Out of 5), Number of Reviews, Price
* **Link:** <https://www.kaggle.com/datasets/amritvirsinghx/audible-complete-catalog>

Dataset 2:

* **Dataset Name: Google Books Dataset**
* **File Name(s):** google\_books\_1299.csv, google\_books\_dataset.csv
* **Source**: Kaggle
* **Relevant fields**: Title (PK), Authors, Language, Rating, Voters, Price, Page Count
* **Link**: <https://www.kaggle.com/datasets/bilalyussef/google-books-dataset>

Dataset 3:

* **Dataset Name: Goodreads-books**
* **File Name(s):** Goodreads\_books.csv
* **Source**: Kaggle
* **Relevant fields**: Title (PK), Author, Average Rating (Out of 5), Number of Pages, Ratings Count, Language

**Link**: <https://www.kaggle.com/datasets/jealousleopard/goodreadsbooks>

**Research Objectives**

State your main questions and objectives. You should include:

* **Primary Question(s):** the key relationships or patterns you want to test.
* **Secondary/Exploratory Questions:** additional comparisons or trends you want to explore.

Explain briefly why these objectives are valuable or interesting.

**Data Preparation Approach**

Describe how you will prepare the data for analysis. Be specific:

* How you will align or merge datasets (keys, years, geography).
* How you will address missing values (replacement, dropping rows, etc.).
* How you will address outliers or extreme values.
* What new variables or indices you may create.
* How you will structure your relational database (at minimum two tables from different sources, joined through a common key).

**Current Status**

Summarize what you have already completed. Examples:

* Data acquired and inspected.
* Early cleaning or standardization steps.
* Prototype functions or exploratory plots.
* Evidence that the datasets can be successfully merged.

**Deliverables (Remaining Work)**

* **Required Tasks (must be completed):***(Example: finish cleaning literacy dataset; implement 3 Python functions; build SQLite schema and load tables; produce 3 required visualizations; write README and data dictionary.)*
* **Stretch Goals (optional, if time allows):***(Example: add ACS demographics to analysis; build interactive dashboard; run regression model.)*

**Project Timeline**

* Phase 1: Acquire and clean data, set up database.
* Phase 2: Exploratory analysis, build functions, first draft visuals.
* Phase 3: Deeper analysis, refine visuals, draft report.
* Phase 4: Finalize deliverables, polish repo, record presentation.

**Additional Considerations**

Note any assumptions, limitations, or risks you anticipate. Examples:

* Geographic regions or variables you will exclude (and why).
* Additional datasets you may add if time allows.
* Possible challenges (e.g., very large files, inconsistent variable names, data sparsity).