Lab2 Report

1. Team Information

Team Name: The Coach

Name	USC ID	GitHub Link
Tianqi Qiu	1716906139	https://github.com/tianqi0301/DSCI560
Shida Yan	5725964711	https://github.com/ShidaYan/dsci-560
Herun Kan	7222919427	https://github.com/herunkan/Data-Science-Professional-Practicum

Demo link: https://youtu.be/E4cwGM_igWQ

2. Domain Rationale

Target Group

The target group for this project is small fitness accountability groups, such as friends, classmates, or gym partners who are working toward shared fitness goals. These groups typically already use group chats to communicate and rely on social motivation to stay consistent.

Problems Faced by the Group

Although group chats are convenient, they also create several challenges:

- Motivation tends to drop over time, leading to missed workouts
- Fitness information is spread across multiple apps and platforms
- Important messages like workout plans or reminders get lost in casual conversation
- There is no easy way to track group-level progress or participation

Because of these issues, many people struggle to maintain consistency even when they have a supportive group.

Limitations of Existing Tools

Current fitness tools and chatbots do not fully address these problems. Most fitness apps are designed for individual users and do not support group accountability or shared progress tracking. General-purpose chatbots can answer fitness-related questions, but their responses are often generic and not tailored to the group's goals or activity level. Additionally, these tools lack awareness of group context and cannot automate simple tasks such as reminders, summaries, or workout logging. As a result, users often rely on multiple apps, which increases effort instead of simplifying the experience.

Why It Works Well

A chat-based AI assistant fits naturally into how fitness groups already communicate. Instead of introducing a new platform, the assistant becomes part of the existing group chat. Users can interact with it through simple messages to log workouts, ask fitness questions, or request reminders. By understanding group context and using domain-specific fitness data, the assistant can provide relevant responses, automate routine tasks, and summarize group activity. This approach reduces friction while improving accountability and consistency within the group.

3. Brainstormed Domain Options:

Fitness Accountability Group (Final pick)
This domain focuses on a small group of people who want to stay consistent with their fitness goals. Group chats are commonly used to share workouts, motivate each other, and ask fitness-related questions. An AI assistant can help automate workout logging, reminders, and answer basic fitness questions.

Travel Planning Group
This domain focuses on friends or families planning trips together. Group chats are used to discuss destinations, share itineraries, and coordinate bookings. An AI assistant could help organize plans, suggest activities, and answer travel-related questions.

Gaming Team Chat
This domain focuses on online gaming teams or guilds. Group chats are used to coordinate play sessions and strategies. An AI assistant could manage schedules, summarize strategies, and track team performance.

Classroom Study Group
This domain would support students working together in study groups. Common needs include sharing resources, asking questions, and organizing study schedules. An AI assistant could help summarize topics and answer common questions.

Apartment Roommates Group
This domain supports roommates living together. Group chats are commonly used to discuss chores, bills, and shared responsibilities. An AI assistant could help track expenses, send reminders, and resolve scheduling conflicts.

Language Learning Support Group
This domain targets people learning a new language together. Group chats are used to practice vocabulary and ask grammar questions. An AI assistant could provide explanations, quizzes, and daily practice prompts.

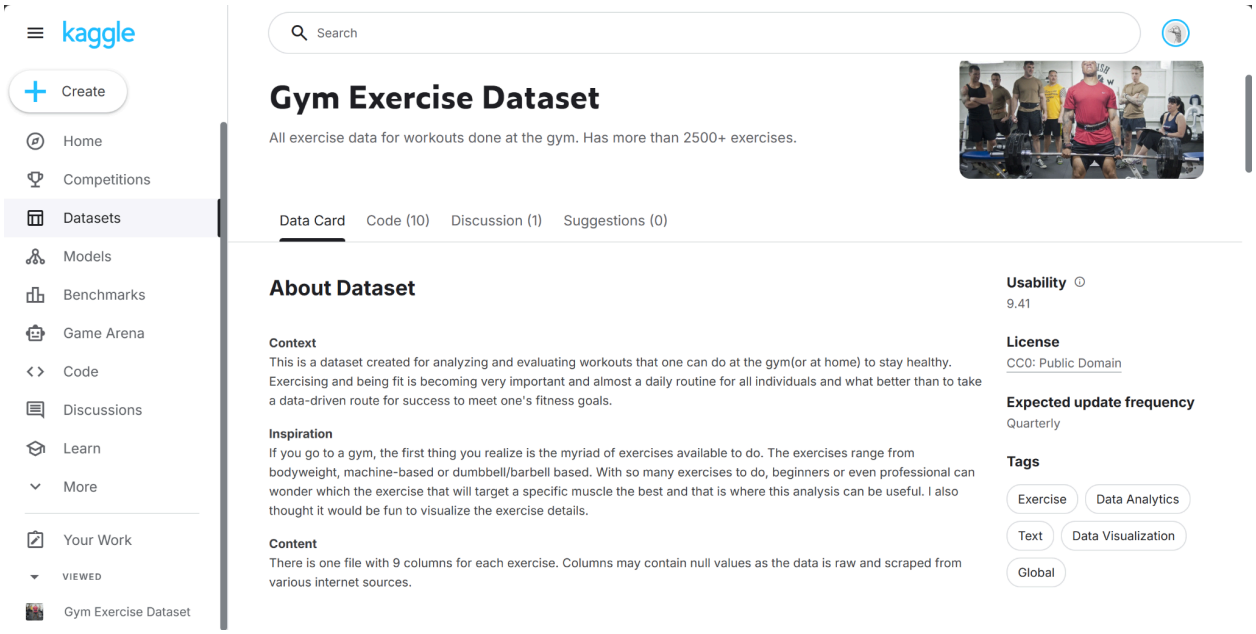
4. Data Sources and Data Collection

This section describes the publicly available data sources used in this project. Three different data types were collected to support the fitness group chat AI assistant: structured CSV data, unstructured forum text, and scientific PDF documents. This project was implemented using Python and several commonly used data processing libraries. Pandas was used for loading and analyzing structured CSV data. Requests

and BeautifulSoup were used to simulate web scraping and extract text-based forum content. Pdfplumber was used to extract textual information from PDF-based scientific articles. These tools were selected because they are lightweight, well-documented, and suitable for handling heterogeneous data sources commonly encountered in real-world applications.

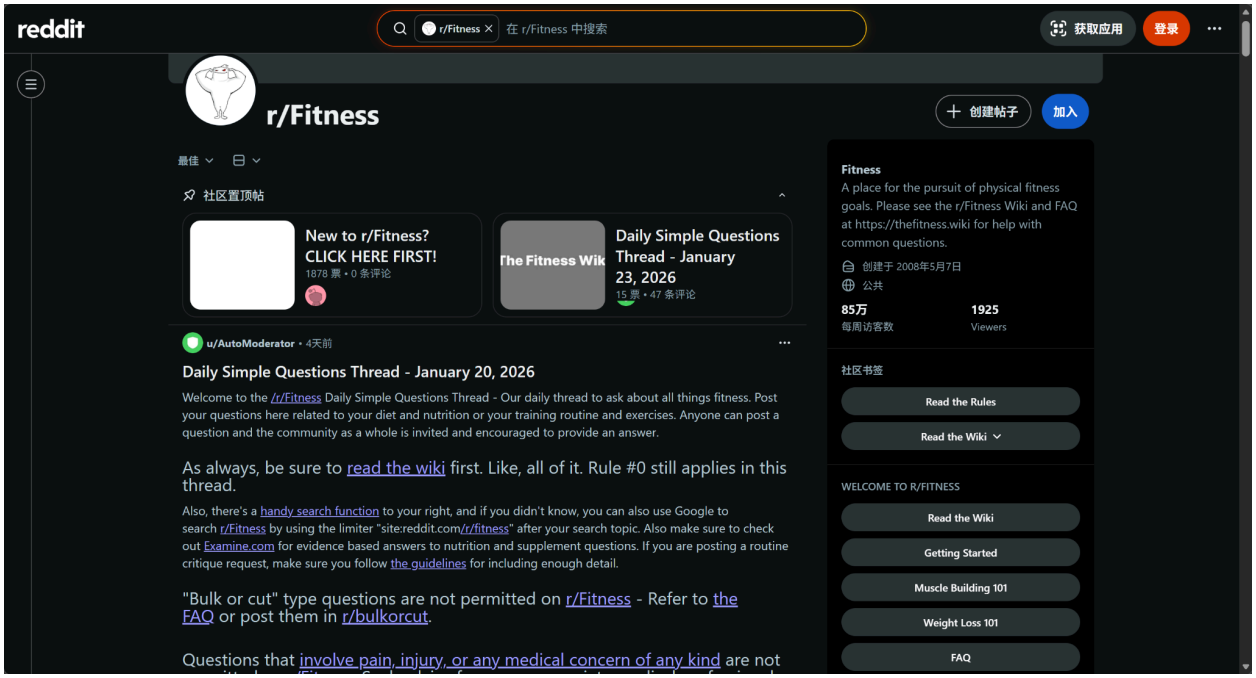
The first data source is a structured exercise database obtained from Kaggle.
Exercise Database (CSV)
Source: https://www.kaggle.com/niharika41298/gym-exercise-data

This dataset contains detailed information about gym exercises, including exercise names, primary muscle groups, required equipment, difficulty levels, and recommended training parameters. The structured nature of the data makes it suitable for basic analysis and supports features such as workout recommendations and exercise categorization.



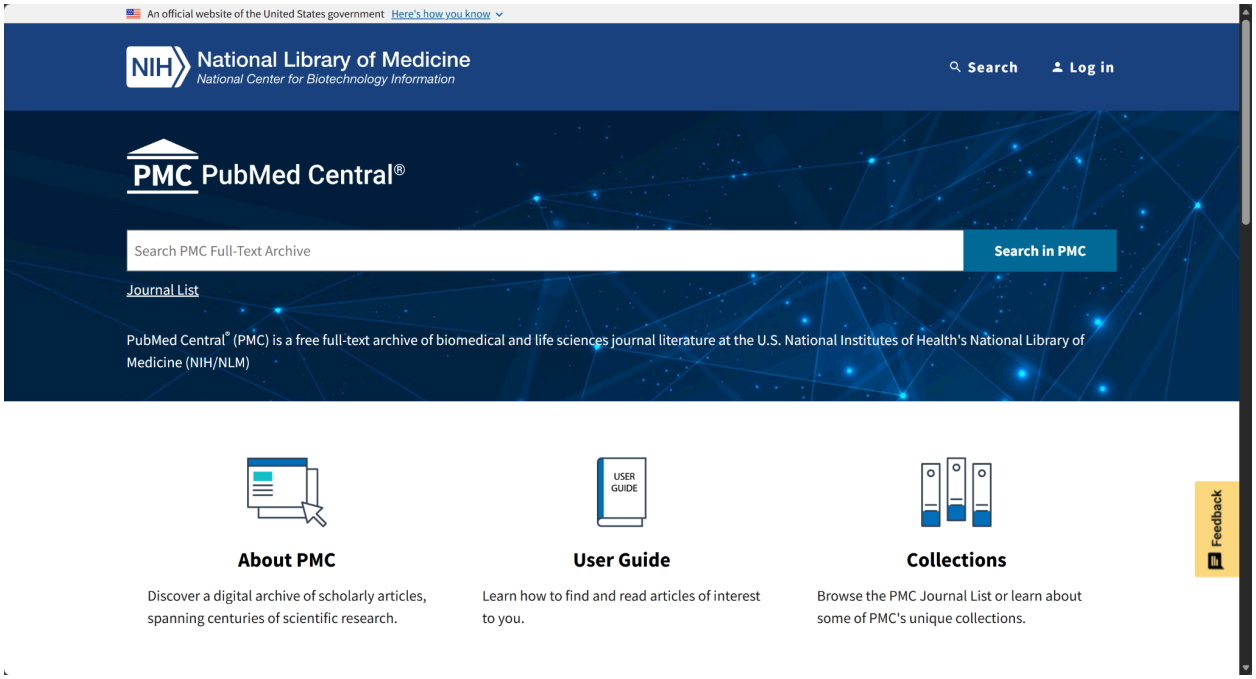
The second data source consists of unstructured text collected from online fitness forum discussions.
Reddit - r/Fitness (Web Scraping)
Source: https://www.reddit.com/r/Fitness/

Forum content from Reddit's r/Fitness community represents real-world fitness discussions posted by a large and active user base. The subreddit includes questions and conversations related to workout routines, strength training, cardio, nutrition, injury prevention, and beginner guidance. This unstructured text reflects informal, conversational language and practical concerns shared by everyday users, making it a valuable source for capturing realistic conversational patterns. Collecting data from r/Fitness helps the chatbot better understand common fitness-related inquiries and improves its ability to generate relevant, context-aware responses in a group chat setting.



The third data source is composed of scientific research articles obtained from PubMed Central.
Exercise Science Journals (PDF)
Source: https://www.ncbi.nlm.nih.gov/pmc/

These peer-reviewed articles provide evidence-based insights into exercise science topics such as muscle hypertrophy, protein requirements, biomechanics, and training safety. PDF documents were selected to satisfy the requirement of processing document-based data and to ensure that chatbot responses can be grounded in scientific literature.



5. Data Exploration and Processing

A Python script named data_exploration.py was developed to collect, process, and explore data from three selected sources: a structured exercise database, fitness-related forum discussions from Reddit, and exercise science research articles from PubMed Central. The script was designed to handle heterogeneous data formats and convert them into structured CSV files suitable for analysis and future chatbot integration. Console outputs were generated at each stage to verify successful execution and summarize key characteristics of the data.

The script begins by initializing the data collection workflow and creating a dedicated directory for storing all extracted datasets. It logs the domain context and selected data sources to ensure transparency and reproducibility of the data exploration process.

The first dataset processed is the exercise database obtained from Kaggle. The script loads the CSV file and performs basic exploratory analysis, including calculating the total number of records and columns, inspecting column names, and displaying sample rows. This analysis helps validate data integrity and confirms the suitability of the dataset for structured exercise lookup and categorization.

```
======================================================================
SOURCE 1: LOCAL EXERCISE DATABASE (CSV)
======================================================================
✓ Successfully loaded 2918 records
✓ Columns: 9

Column names: ['Unnamed: 0', 'Title', 'Desc', 'Type', 'BodyPart', 'Equipment', 'Level', 'Rating', 'RatingDesc']

First 5 rows:
   Unnamed: 0                        Title  ... Rating RatingDesc
0           0       Partner plank band row  ...    0.0        NaN
1           1  Banded crunch isometric hold  ...    NaN        NaN
2           2         FYR Banded Plank Jack  ...    NaN        NaN
3           3                 Banded crunch  ...    NaN        NaN
4           4                        Crunch  ...    NaN        NaN

[5 rows x 9 columns]
```

The second data source consists of fitness forum discussions collected from Reddit's r/Fitness community using the Reddit JSON API. The script retrieves recent posts, verifies successful responses, and stores post titles and metadata in a structured CSV format. This step captures real-world fitness questions and conversational patterns, which are useful for understanding how users naturally discuss fitness-related topics in online communities.

```
======================================================================
SOURCE 2: REDDIT r/FITNESS (JSON API)
======================================================================

Trying: https://www.reddit.com/r/Fitness.json
Response status: 200
Found 25 posts in JSON data
  ✓ Post 1: New to r/Fitness? CLICK HERE FIRST!...
  ✓ Post 2: Daily Simple Questions Thread - January 23, 2026...
  ✓ Post 3: Physique Phriday...
  ✓ Post 4: Daily Simple Questions Thread - January 22, 2026...
  ✓ Post 5: Rant Wednesday - January 21, 2026...
  ✓ Post 6: Daily Simple Questions Thread - January 21, 2026...
  ✓ Post 7: Daily Simple Questions Thread - January 20, 2026...
  ✓ Post 8: Moronic Monday - Your weekly stupid questions thread...
  ✓ Post 9: Victory Sunday...
  ✓ Post 10: Daily Simple Questions Thread - January 18, 2026...
  ✓ Post 11: Gym Story Saturday...
  ✓ Post 12: Daily Simple Questions Thread - January 17, 2026...
  ✓ Post 13: Daily Simple Questions Thread - January 16, 2026...
  ✓ Post 14: Physique Phriday...
  ✓ Post 15: Monthly Fitness Pro-Tips Megathread...
  ✓ Post 16: Daily Simple Questions Thread - January 15, 2026...
  ✓ Post 17: Rant Wednesday - January 14, 2026...
  ✓ Post 18: Daily Simple Questions Thread - January 14, 2026...
  ✓ Post 19: Daily Simple Questions Thread - January 13, 2026...
  ✓ Post 20: Moronic Monday - Your weekly stupid questions thread...

✓ Successfully scraped 20 posts
✓ Saved to: collected_fitness_data\reddit_posts.csv
```

The third dataset is derived from exercise science research articles obtained from PubMed Central. The script queries the PubMed API using fitness-related keywords, retrieves article identifiers, and extracts metadata such as titles and publication sources. This process enables scientific literature to be incorporated into a machine-readable format while preserving its academic context and credibility.

```
======================================================================
SOURCE 3: PUBMED RESEARCH ARTICLES (API)
======================================================================

Searching PubMed for: 'resistance training'
  Found 5 article IDs
  ✓ PMID 32457216: Resistance Training for Children and Adolescents....
  ✓ PMID 27102172: Effects of Resistance Training Frequency on Measures of Musc...
  ✓ PMID 36497565: The Effect of Resistance Training on the Rehabilitation of E...
  ✓ PMID 31343601: Resistance Training for Older Adults: Position Statement Fro...
  ✓ PMID 35055695: A Review on Aging, Sarcopenia, Falls, and Resistance Trainin...

Searching PubMed for: 'exercise physiology'
  Found 5 article IDs
  ✓ PMID 32658037: Exercise Progression to Incrementally Load the Achilles Tend...
  ✓ PMID 29129214: Exercise and Older Adults....
  ✓ PMID 29098620: Clinical Evidence of Exercise Benefits for Stroke....
  ✓ PMID 30642252: Effects of three home-based exercise programmes regarding fa...
  ✓ PMID 25305061: Exercise and cardiovascular risk in patients with hypertensi...

Searching PubMed for: 'strength training'
  Found 5 article IDs
  ✓ PMID 30525318: [Strength training in children and adolescents: benefits, ri...
  ✓ PMID 35564764: Strength Training in Swimming....
  ✓ PMID 28490537: Adaptations to Endurance and Strength Training....
  ✓ PMID 27692740: Strength training for plantar fasciitis and the intrinsic fo...
  ✓ PMID 23914932: Optimizing strength training for running and cycling enduran...

✓ Successfully fetched 15 articles
✓ Saved to: collected_fitness_data\pubmed_articles.csv
```

6. Dataset Integration

After completing data exploration and processing, all extracted datasets were standardized and stored in a unified directory. The script loads the exercise database, Reddit forum posts, and PubMed research articles and verifies successful integration by reporting record counts for each source. A summary table is generated to confirm that all data sources were collected successfully and saved in CSV format.

The final integrated dataset includes structured exercise records, real-world forum discussions, and evidence-based scientific articles. This unified dataset serves as a foundation for building a domain-specific knowledge base that supports both conversational responses and automated group-level services in the proposed fitness group chat AI assistant.

```
======================================================================
COLLECTION SUMMARY
======================================================================

Collection Results:
- Total sources attempted: 3
- Successful collections: 3/3
- Total records collected: 2953

Detailed breakdown:
           source              filename  records  columns  size_kb    status
Exercise Database exercise_database.csv     2918        9   657.27 Success ✓
     Reddit Posts      reddit_posts.csv       20       10     7.86 Success ✓
  Pubmed Articles   pubmed_articles.csv       15        9     4.32 Success ✓

✓ Summary saved to: collected_fitness_data\collection_summary.csv
```

7. Limitations of Existing Chatbots

Most existing chatbots are designed for general-purpose, one-on-one interactions and lack awareness of group context. In fitness-related applications, this often results in generic responses that do not account for shared goals, collective progress, or historical group behavior. These systems typically cannot track workouts over time, summarize group activity, or automate routine tasks such as reminders and logging.

In addition, many chatbots rely on broad or unverified knowledge sources, which limits the accuracy and reliability of their responses. The lack of integration between structured exercise data, real user discussions, and scientific literature further reduces their usefulness in domains where correctness and safety are important.

8. How the Dataset Improves Chatbot

The dataset developed in this project improves chatbot performance by combining structured exercise data, real-world forum discussions, and peer-reviewed scientific literature. Structured data enables precise exercise recommendations, while forum content captures common user questions and conversational patterns. Scientific articles provide evidence-based grounding for responses related to training and nutrition.

By integrating these complementary sources, the chatbot can deliver more accurate, context-aware, and domain-specific responses. This approach improves both the relevance and correctness of interactions, particularly in a group fitness setting where accountability and safety are critical.

9. Challenge and Future Work

During the development of this lab, several practical challenges were encountered while working with real-world data sources. The collected datasets varied significantly in structure and quality, which required additional effort to standardize formats and ensure consistency across sources. In particular, forum-based text data was noisy and unstructured, making it more difficult to organize compared to structured CSV data. Processing PDF-based scientific articles also required careful handling to extract meaningful textual content while preserving relevant metadata.

Key challenges encountered during the lab include:

- Inconsistent data formats across CSV, web text, and PDF sources
- Noise and variability in forum discussion content
- The need to balance data completeness with relevance when extracting information from PDFs

The primary direction for future work is the implementation of the actual fitness group chat AI assistant. The unified dataset created in this lab provides the foundation for this system. Future development would focus on integrating the dataset into a chatbot capable of responding to fitness-related questions, tracking group activity, and supporting accountability within group chats. This would allow for direct evaluation of the dataset's effectiveness in improving response relevance, correctness, and group-level interaction.