GitHub- Link: https://github.com/Mujtabashah4/Dataengineering-assignment-01

Group Number & Student IDs:

- Group Number: Group 7
- Student 1 ID: 24280069
- Student 2 ID: 24280052

Contributions:

- 24280069: Developed scripts for collecting data from Google Trends using Pytrends and from Reddit using the PRAW API.
- 24280052: Developed scripts to collect public data from Kaggle.
- Both: Performed data preprocessing and initial analysis with pandas, and provided a summary of key insights.

1. Overview of Our Topic

We have chosen **sports** as our topic and have specifically focused on **cricket**:

We aim to analyze trends in cricket performance and fan engagement. Given our strong interest in cricket and the **upcoming tournament** in our country, we expect to explore public interest trends and discussions surrounding cricket. We anticipate the following:

- **Increasing interest in player performance**, particularly from top-performing cricketers.
- **High engagement in discussions around match outcomes**, team strategies, and player rivalries.
- **Potential seasonal trends** in match outcomes and cricket fan activity.

2. Data Collection Process

Google Trends Data

- **Utilized the Pytrends library** to gather search interest over time for the following keywords: ["India vs Pakistan", "Champions Trophy", "Asia Cup Final", "World Cup Final"].
- **Challenges encountered**: Google Trends updated their bot detection system, making it harder to make API calls through Pytrends. To address this, we implemented a **time delay** using time.sleep(30) to avoid rate limits.

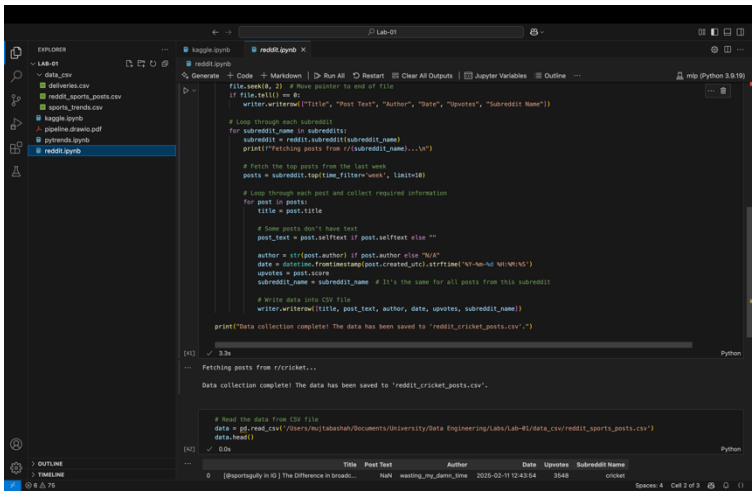
Reddit Data

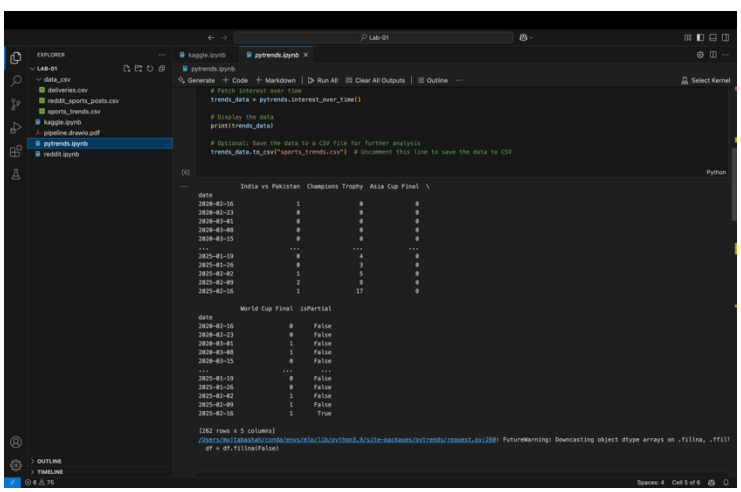
- **Utilized the PRAW library** to collect posts related to cricket.
- **Extracted metadata**: including the title, post text, author, date, upvotes, and subreddit name.
- **Challenges faced**: Some posts lacked text content, which impacted the analysis, and there were privacy concerns regarding personally identifiable information (PII).

3. Initial Observations

We used **pandas** to generate basic summaries of our datasets:

- **Google Trends**: Search interest showed spikes around major cricket events (e.g., India vs Pakistan matches, ICC tournaments).
- **Reddit Data**: High engagement on topics related to cricket performances, fan discussions, and team strategies.
- **Kaggle Data**: Increasing number of IPL-related posts and statistics, with a focus on match outcomes and player performance during the IPL season.





4. AI Product Concept

Using this data, we aim to develop an AI-driven **sports trend analysis tool** that:

- **Identifies real-time shifts** in public interest related to cricket events and performances.
- **Predicts future trends** in cricket viewership and fan engagement.

- **Highlights key discussion themes** from Reddit using **NLP**, focusing on match outcomes, player performances, and fan interactions.

5. Terms of Service & Privacy Constraints

- **Google Trends**: Data can be used for analysis but should not be redistributed without proper attribution, especially when dealing with search trends related to cricket events.
- **Reddit**: User-generated content, including cricket-related posts, cannot be stored indefinitely or republished without consent from the authors.
- **Kaggle**: Public cricket datasets may have their own licensing restrictions that must be adhered to, particularly with respect to usage in commercial or public projects.

Mitigation:

- **Store only aggregated insights** from cricket-related data, rather than keeping raw data to maintain privacy and comply with platform policies.
- **Follow API rate limits** and ensure adherence to platform-specific terms when accessing data from sources like Google Trends, Reddit, and Kaggle to avoid service disruptions or legal issues.

6. Data Quality & Challenges in Multi-Source Collection

- **Benefits:**
  - **Google Trends** provides quantitative insights into public interest and search behavior related to cricket events and players.
  - **Reddit** adds qualitative data from fan discussions, player performance comments, and match-related conversations.
  - **Kaggle** offers structured datasets, particularly for analyzing cricket match outcomes and player statistics, which can be used for validation.
- **Challenges:**
  - **Differences in update frequency** (real-time data from Google Trends vs. static Kaggle datasets).
  - **Potential discrepancies** between search interest and actual cricket match outcomes or player performances.

7. Data Storage & Integration Strategy

- **Storage**: Use a relational database (PostgreSQL) or NoSQL (MongoDB) based on the data type (e.g., match results vs. player performance statistics).
- **Integration**:
  - **Google Trends**: Store search interest data in a time-series database for efficient querying over time.
  - **Reddit**: Use text-based storage for fan discussions, with **NLP** processing to analyze match-related conversations and player mentions.
  - **Kaggle**: Store structured cricket data in a relational format for ease of analysis and validation (e.g., player statistics, match outcomes).

- **ETL Pipeline**: Automate the process of data collection, cleaning, and merging to ensure seamless integration and up-to-date datasets for analysis.