Mining Minds

MINING IMDB FOR MOVIE TRENDS

Team Members:

Pottabathini Vivekananda (230077)
Prashant Kumar (230101)
Prerak Arya (230039)

Mentor: Ravi Kumar

PROBLEM STATEMENT, GOAL & DATASET

PROBLEM STATEMENT

The movie industry is shaped by diverse factors such as genres, directors, actors, budgets, and audience preferences.
This project leverages IMDb data to explore how these factors influence a movie's ratings and popularity over time.

GOAL

- Not just descriptive – but to discover hidden patterns and insights in the movie industry.
- Analyze evolving audience and critic behavior through data.

DATASET

- Source: **IMDb Non-Commercial Datasets**
- Files Used: title.basics, title.ratings, title.crew, title.principals
- Contains metadata for 9M+ titles

WORK PLANNING, TEAM ROLES & GITHUB WORKFLOW

TEAM ROLES (PHASE 1)

Member	Role	Focus Area
Vivekananda	Lead	Repo setup, folder structure, dataset prep,contributed in all aspects.
Prashant	Research Lead	Literature review, research questions, Hypotheses
Prerak	Data Lead	Data preprocessing, Model training plan.

LEADERSHIP ROTATION PLAN

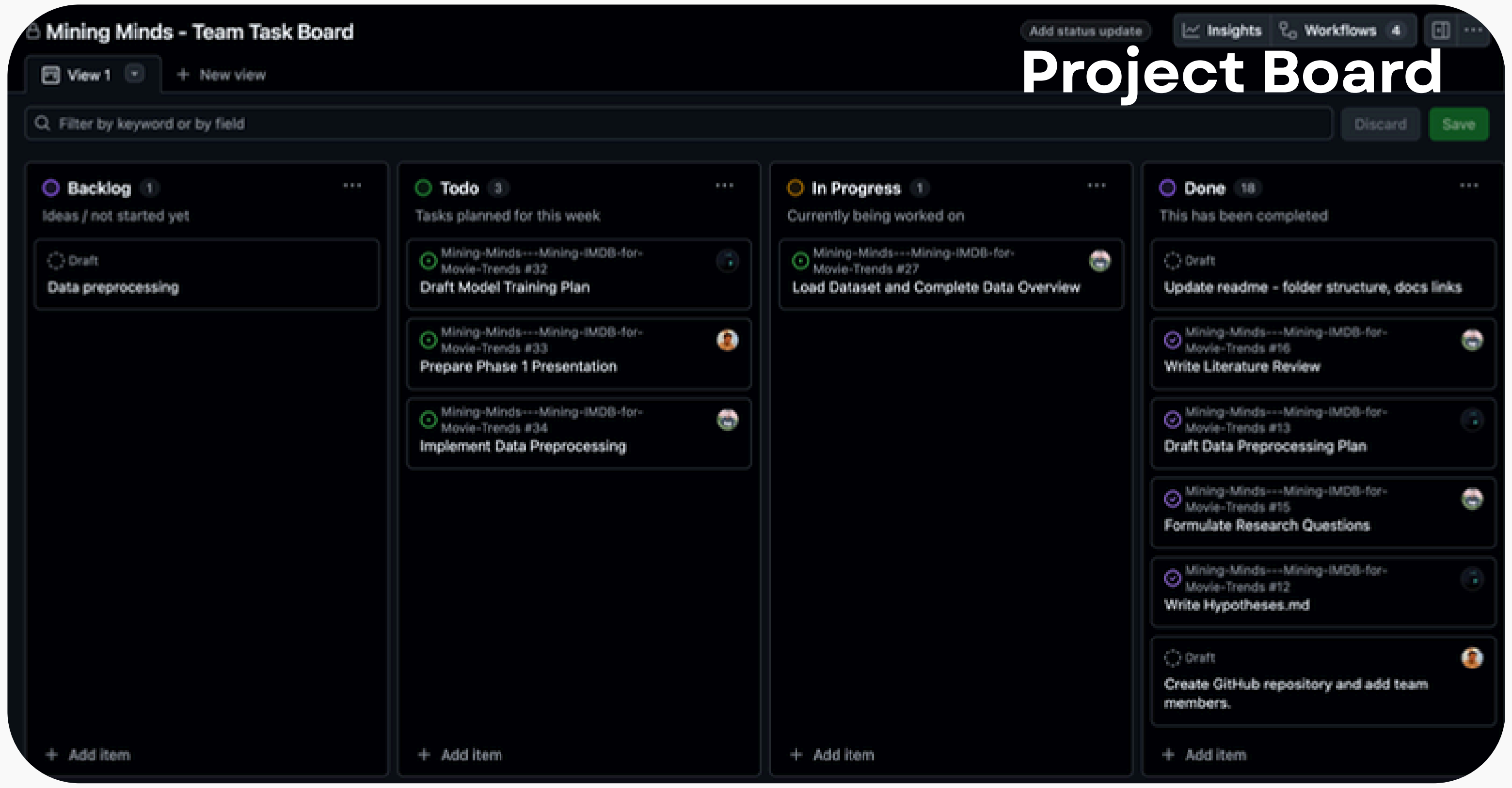
- Phase 1: Vivekananda – Planning & Setup
- Phase 2: Prashant – EDA & Modeling
- Phase 3: Prerak – Final Report & Presentation

GITHUB WORKFLOW

- Project Board → Tasks tracked as Backlog / To-Do / In-Progress / Done
- Issues → Each task documented with owner & deadline
- Branch Naming → name/issue-number-description
- Daily Review → Sync progress + PR reviews

DATA MINING

rishihood university X Newton School of Technology



Project Board

Mining Minds - Team Task Board

Backlog 1
Ideas / not started yet

- Draft
  Data preprocessing

Todo 3
Tasks planned for this week

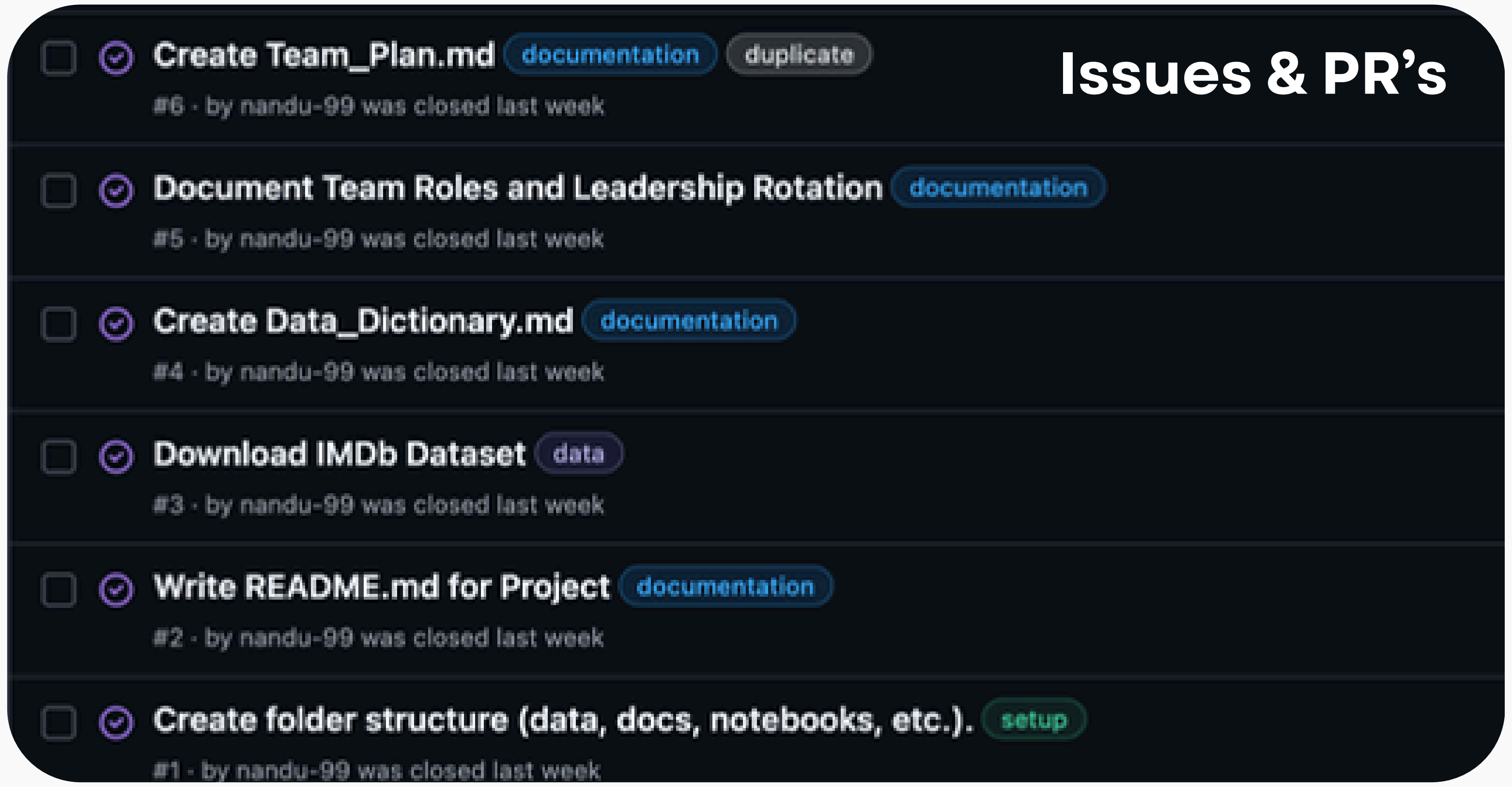
- Mining-Minds---Mining-IMDB-for-Movie-Trends #32
  Draft Model Training Plan
- Mining-Minds---Mining-IMDB-for-Movie-Trends #33
  Prepare Phase 1 Presentation
- Mining-Minds---Mining-IMDB-for-Movie-Trends #34
  Implement Data Preprocessing

In Progress 1
Currently being worked on

- Mining-Minds---Mining-IMDB-for-Movie-Trends #27
  Load Dataset and Complete Data Overview

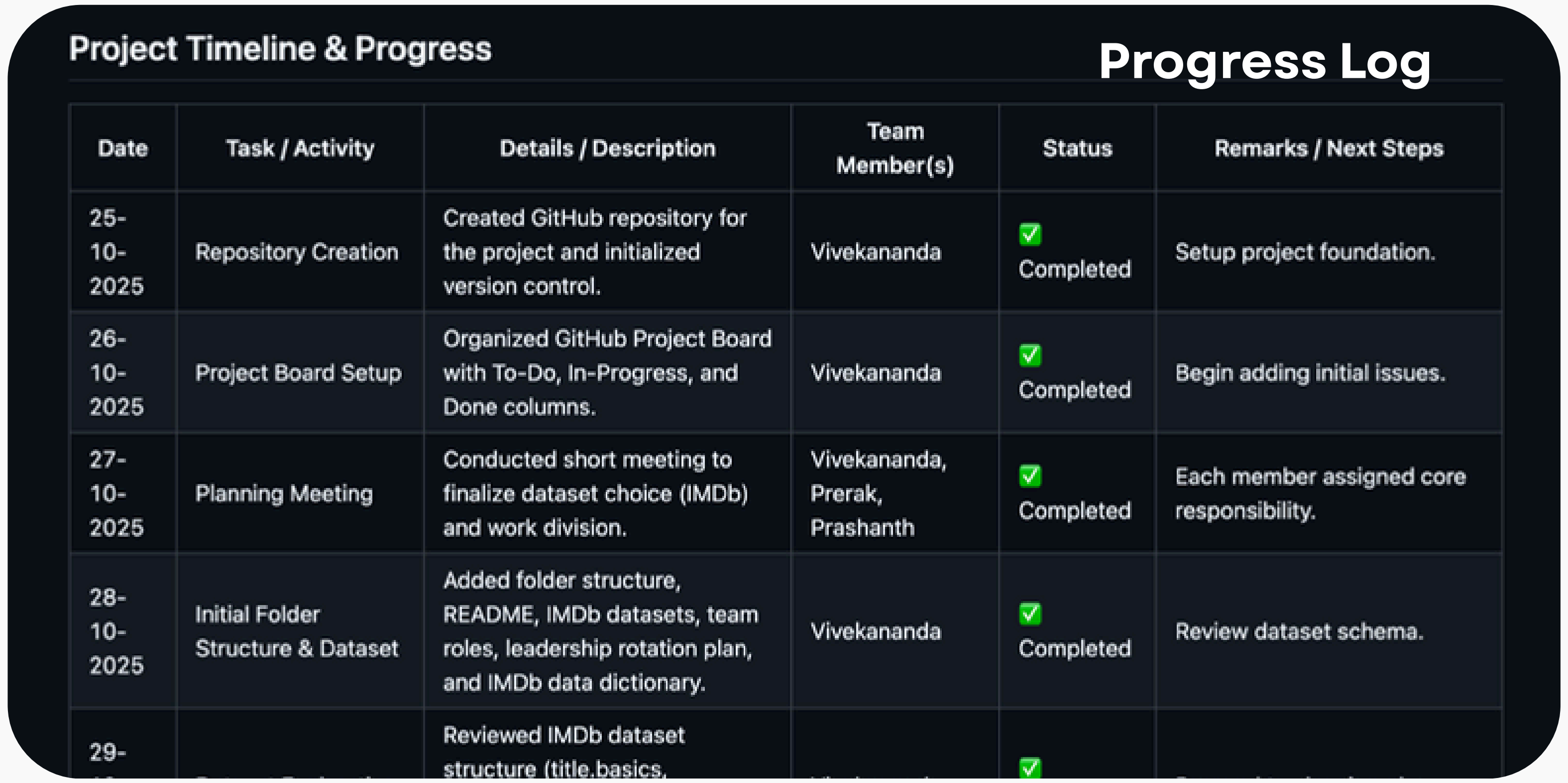
Done 18
This has been completed

- Draft
  Update readme - folder structure, docs links
- Mining-Minds---Mining-IMDB-for-Movie-Trends #16
  Write Literature Review
- Mining-Minds---Mining-IMDB-for-Movie-Trends #13
  Draft Data Preprocessing Plan
- Mining-Minds---Mining-IMDB-for-Movie-Trends #15
  Formulate Research Questions
- Mining-Minds---Mining-IMDB-for-Movie-Trends #12
  Write Hypotheses.md
- Draft
  Create GitHub repository and add team members.



Issues & PR's

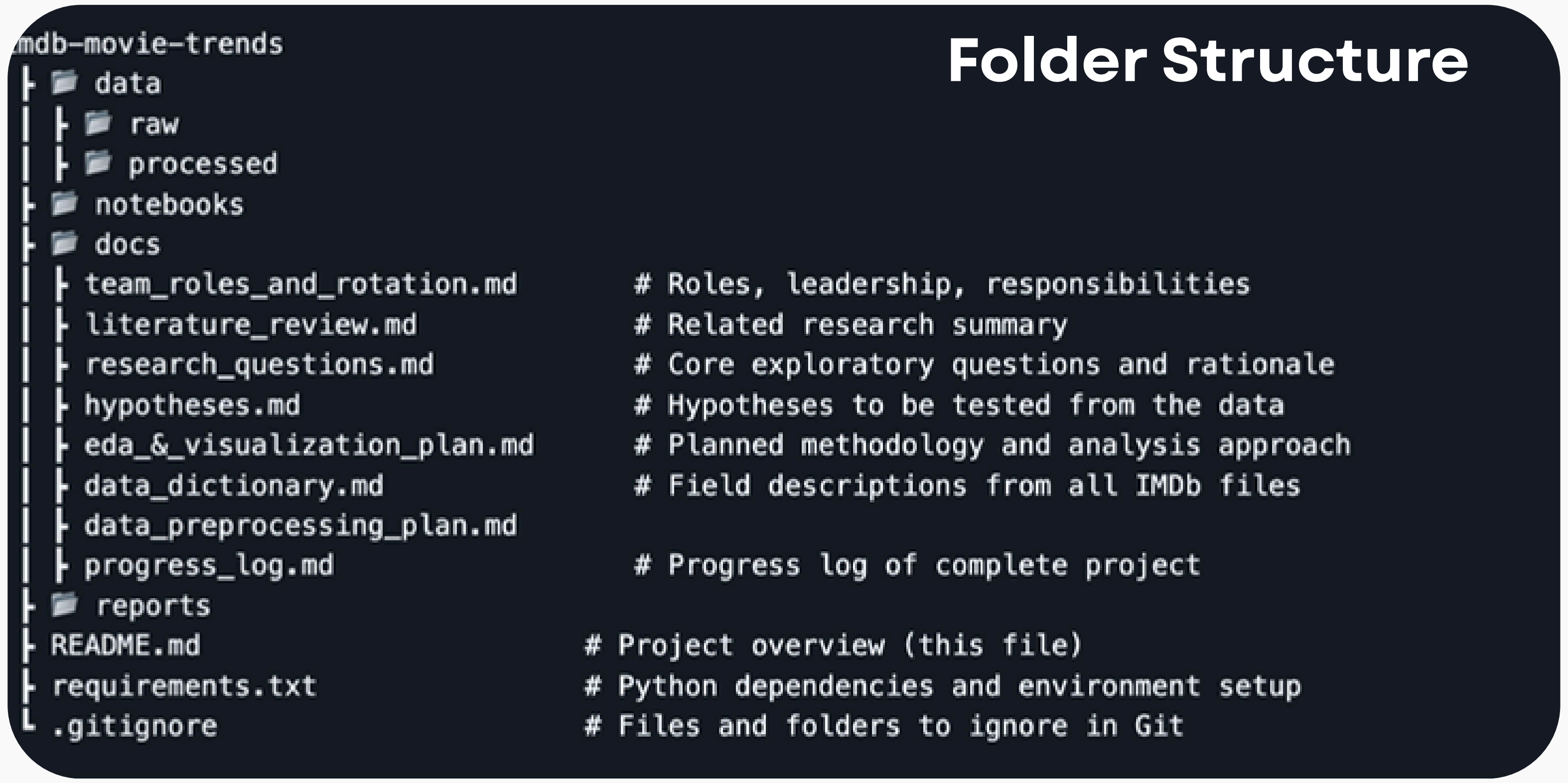
- □ Create Team_Plan.md `documentation` `duplicate`
  #6 · by nandu-99 was closed last week
- □ Document Team Roles and Leadership Rotation `documentation`
  #5 · by nandu-99 was closed last week
- □ Create Data_Dictionary.md `documentation`
  #4 · by nandu-99 was closed last week
- □ Download IMDb Dataset `data`
  #3 · by nandu-99 was closed last week
- □ Write README.md for Project `documentation`
  #2 · by nandu-99 was closed last week
- □ Create folder structure (data, docs, notebooks, etc.). `setup`
  #1 · by nandu-99 was closed last week



Progress Log

Project Timeline & Progress

Date	Task / Activity	Details / Description	Team Member(s)	Status	Remarks / Next Steps
25-10-2025	Repository Creation	Created GitHub repository for the project and initialized version control.	Vivekananda	✅ Completed	Setup project foundation.
26-10-2025	Project Board Setup	Organized GitHub Project Board with To-Do, In-Progress, and Done columns.	Vivekananda	✅ Completed	Begin adding initial issues.
27-10-2025	Planning Meeting	Conducted short meeting to finalize dataset choice (IMDb) and work division.	Vivekananda, Prerak, Prashanth	✅ Completed	Each member assigned core responsibility.
28-10-2025	Initial Folder Structure & Dataset	Added folder structure, README, IMDb datasets, team roles, leadership rotation plan, and IMDb data dictionary.	Vivekananda	✅ Completed	Review dataset schema.
29-		Reviewed IMDb dataset structure (title.basics,		✅	



Folder Structure

```
mdb-movie-trends
├ 📁 data
│ ├ 📁 raw
│ ├ 📁 processed
├ 📁 notebooks
├ 📁 docs
│ ├ team_roles_and_rotation.md          # Roles, leadership, responsibilities
│ ├ literature_review.md                # Related research summary
│ ├ research_questions.md               # Core exploratory questions and rationale
│ ├ hypotheses.md                       # Hypotheses to be tested from the data
│ ├ eda_&_visualization_plan.md         # Planned methodology and analysis approach
│ ├ data_dictionary.md                  # Field descriptions from all IMDb files
│ ├ data_preprocessing_plan.md
│ ├ progress_log.md                     # Progress log of complete project
├ 📁 reports
├ README.md                        # Project overview (this file)
├ requirements.txt                 # Python dependencies and environment setup
└ .gitignore                       # Files and folders to ignore in Git
```

Phase - 1

https://github.com/users/nandu-99/projects/2

CHALLENGE

- Planning was as important as execution.
- Defining structure, roles, and rules was tough.
- Collaborative workflow was difficult.

SOLUTION

- Team discussion to set priorities.
- Finalized folder structure.
- Introduced leadership rotation for shared responsibility.

LITERATURE REVIEW

Bahraminasr, A., & Vafaei-Sadr, A. (2020). IMDb Data from Two Generations (1979 to 2019)

About the Paper

- Largest IMDb dataset (79,000+ movies, 1979–2019)
- Includes ratings, votes, genre, certificates, languages, country, etc.
- Uses statistical + basic machine learning analysis

Key Insights (from the paper)

- Metadata influences ratings (genre, MPAA rating, etc.)
- Trends exist over time (ratings, votes, production volume)
- Regional bias: US vs. Non-US voters show different rating behavior
- Gap identified: Does not analyze individual actor/director influence

RESEARCH QUESTIONS

- Does Star Power (actor/director) influence ratings?
- Are ratings biased by production region?
- Which factors are the strongest predictors (genre, runtime, region)?
- How have genre popularity & quality evolved over decades?

HYPOTHESES

- H1: Movies with top-tier actors/directors show lower rating variance and higher ratings.
- H2: Average IMDb ratings differ significantly across production regions (US, India, South Korea, etc.).
- H3: Metadata features like genre, runtime, and region are the top predictors of rating/popularity.
- H4: Genre popularity (votes) and quality (ratings) change significantly over decades.

CHALLENGE

- Difficult to find a recent research paper relevant to IMDb data.
- Needed a paper that provided academic justification for our project.
- Had to identify a study with a clear research gap we could address.

SOLUTION

- Found Bahraminasr (2020), which analyzed metadata like region and genre.
- Identified that it lacked analysis of individual contributors (actors/directors).
- Used this gap to define our project scope and build upon their work.

DATA

- **Data Dictionary:** Provides detailed information about all IMDb dataset files, their fields, and data types. Link
- **Data Preprocessing:** Steps to clean, merge, and standardize IMDb datasets for analysis. Link
- **EDA & Visualization Plan:** Outline of exploratory data analysis and visualization strategies to uncover trends in IMDb data. Link
- **Model Training Plan:** Plan for training predictive models on IMDb data, including feature selection, algorithms, and evaluation metrics. Link

CHALLENGE

The IMDb files were very large, which made it slow and hard to combine all data.

SOLUTION

I worked with the data in small parts, kept only useful columns, and cleaned it step by step – which made it run faster and easier to analyze.

THANK YOU

[GET READY FOR PHASE-2]