Mining Minds

MINING IMDB FOR MOVIE TRENDS

Team Members:

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

Mentor: Ravi Kumar

IMDb

DATA MINING

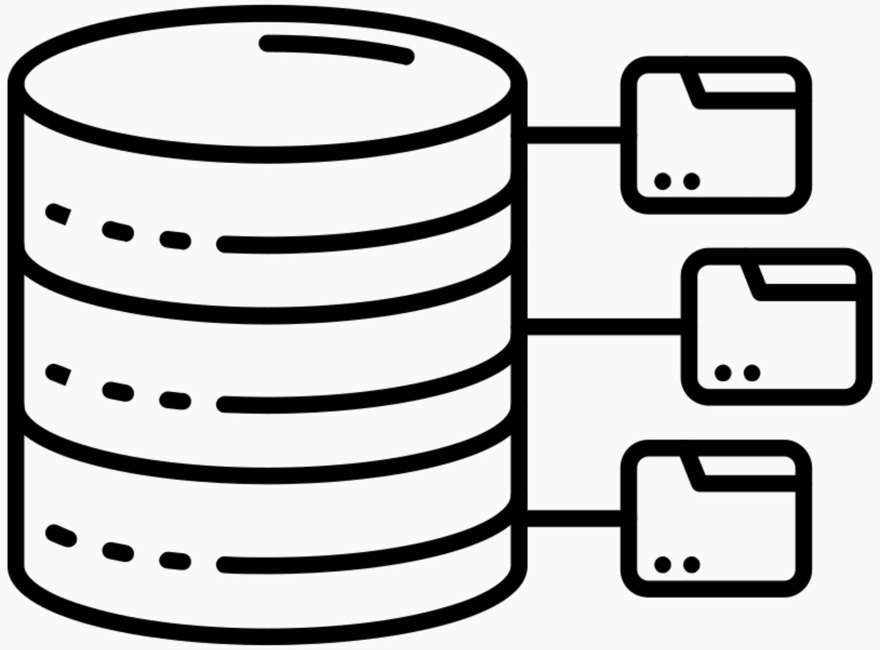
PROBLEM STATEMENT, GOAL & DATASET

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.

OBJECTIVES

- Genre & Trend Analysis
- Missing Value & Outlier Study
- Title-Type Distribution Study
- Predictive Rating Modelling
- Interactive Streamlit Visualization



DATASET

- Source: **IMDb Non-Commercial Datasets**
- Files Used: title.basics, title.ratings, title.crew, title.akas
- Purpose: Merged for unified analysis

WORK PLANNING, TEAM ROLES & GITHUB WORKFLOW

TEAM ROLES (PHASE 2)

Member	Role	Focus Area
Prashant	Lead	Workflow planning, task coordination, predictive analysis
Prerak	Research Lead	Refining research questions, research methodology, Streamlit app development
Vivekananda	Data Lead	Complete EDA, descriptive analysis, insight generation

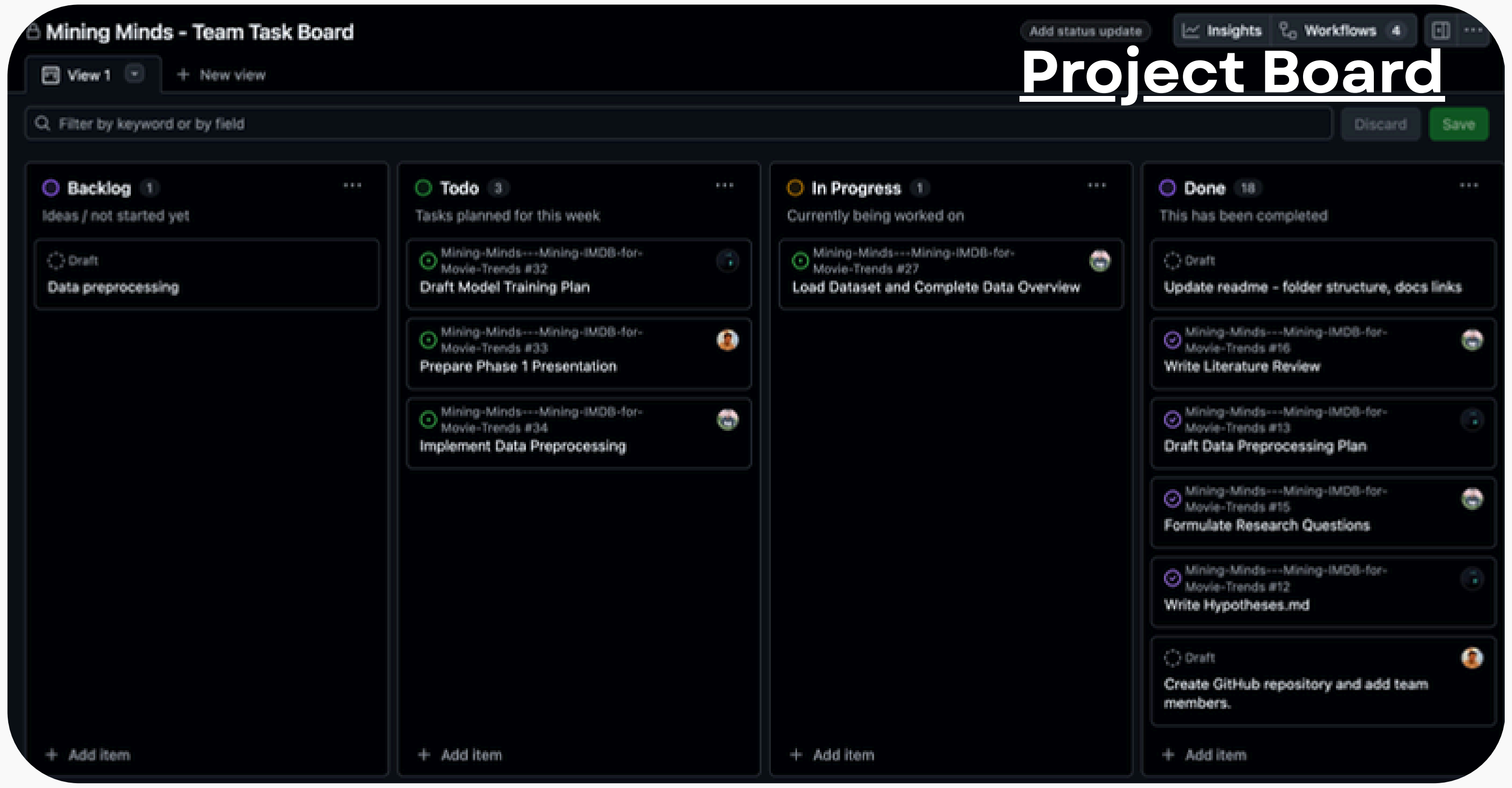
LEADERSHIP ROTATION PLAN

- 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



Project Board

Mining Minds - Team Task Board

Backlog 1
Ideas / not started yet

- Draft
  Data preprocessing

+ Add item

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

+ Add item

In Progress 1
Currently being worked on

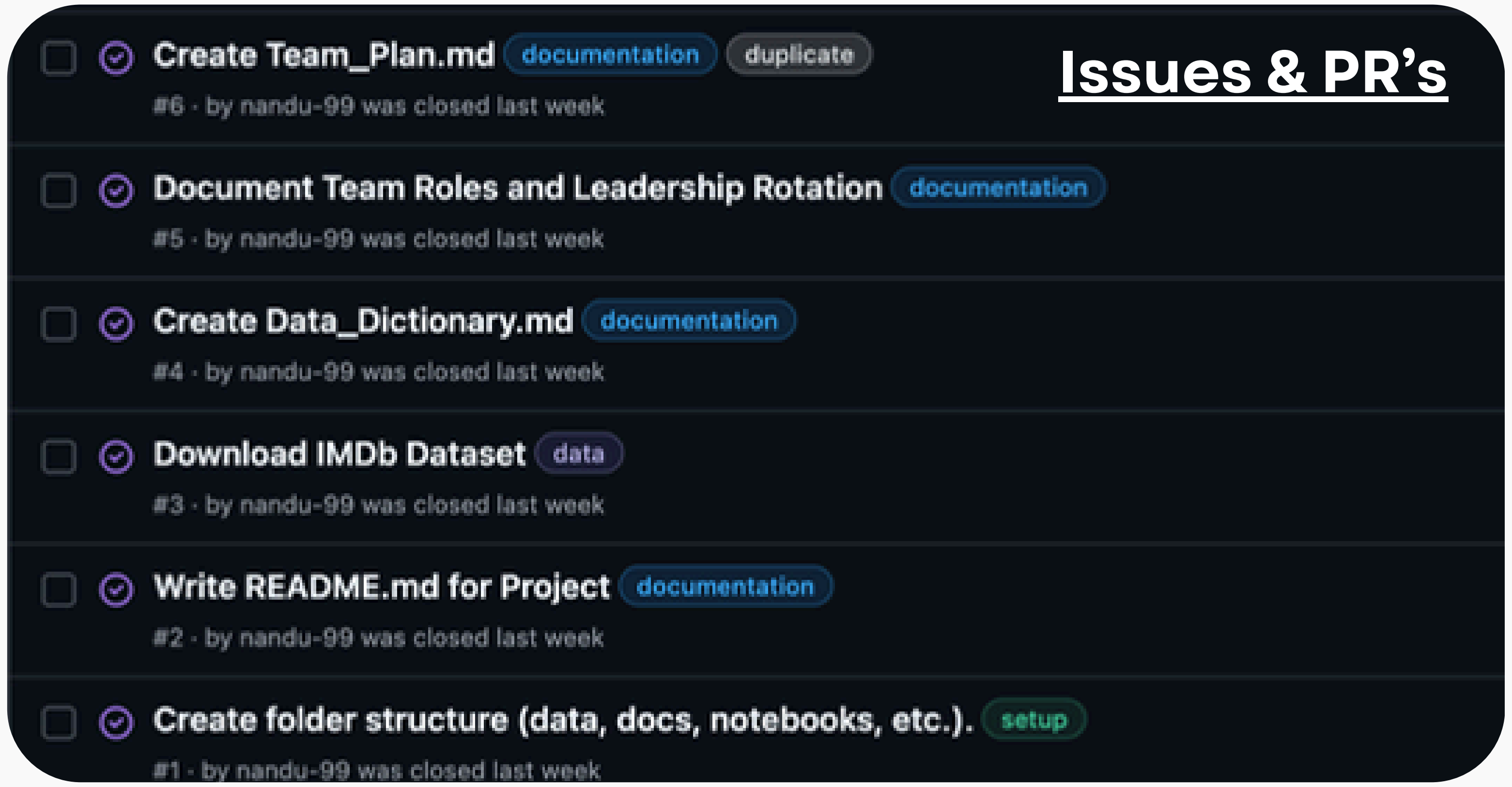
- Mining-Minds---Mining-IMDB-for-Movie-Trends #27
  Load Dataset and Complete Data Overview

+ Add item

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.

+ Add item



Issues & PR’s

□ Create Team_Plan.md `documentation` `duplicate`
#6 · by nandu-99 was closed last week

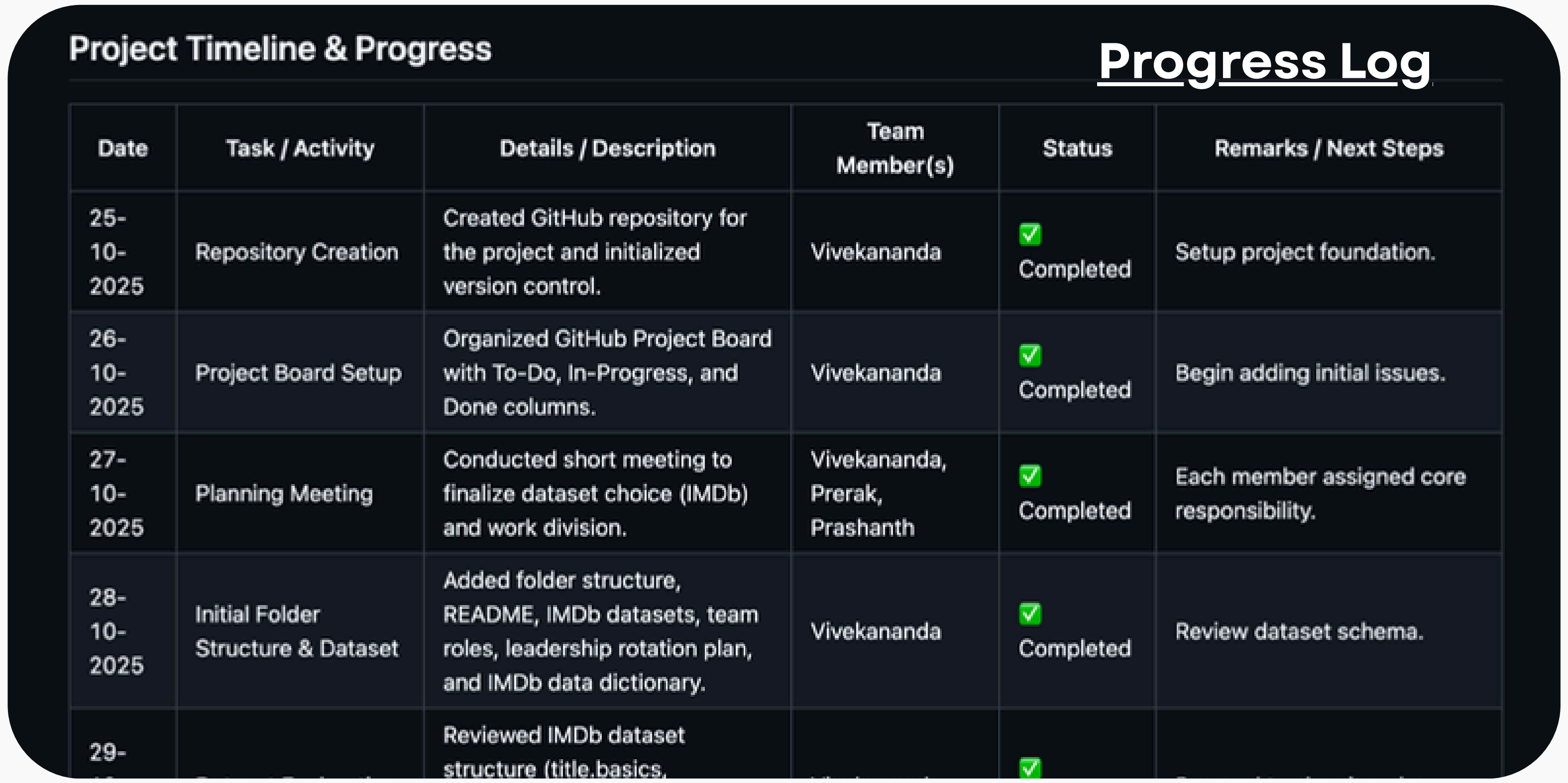
□ Document Team Roles and Leadership Rotation `documentation`
#5 · by nandu-99 was closed last week

□ 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

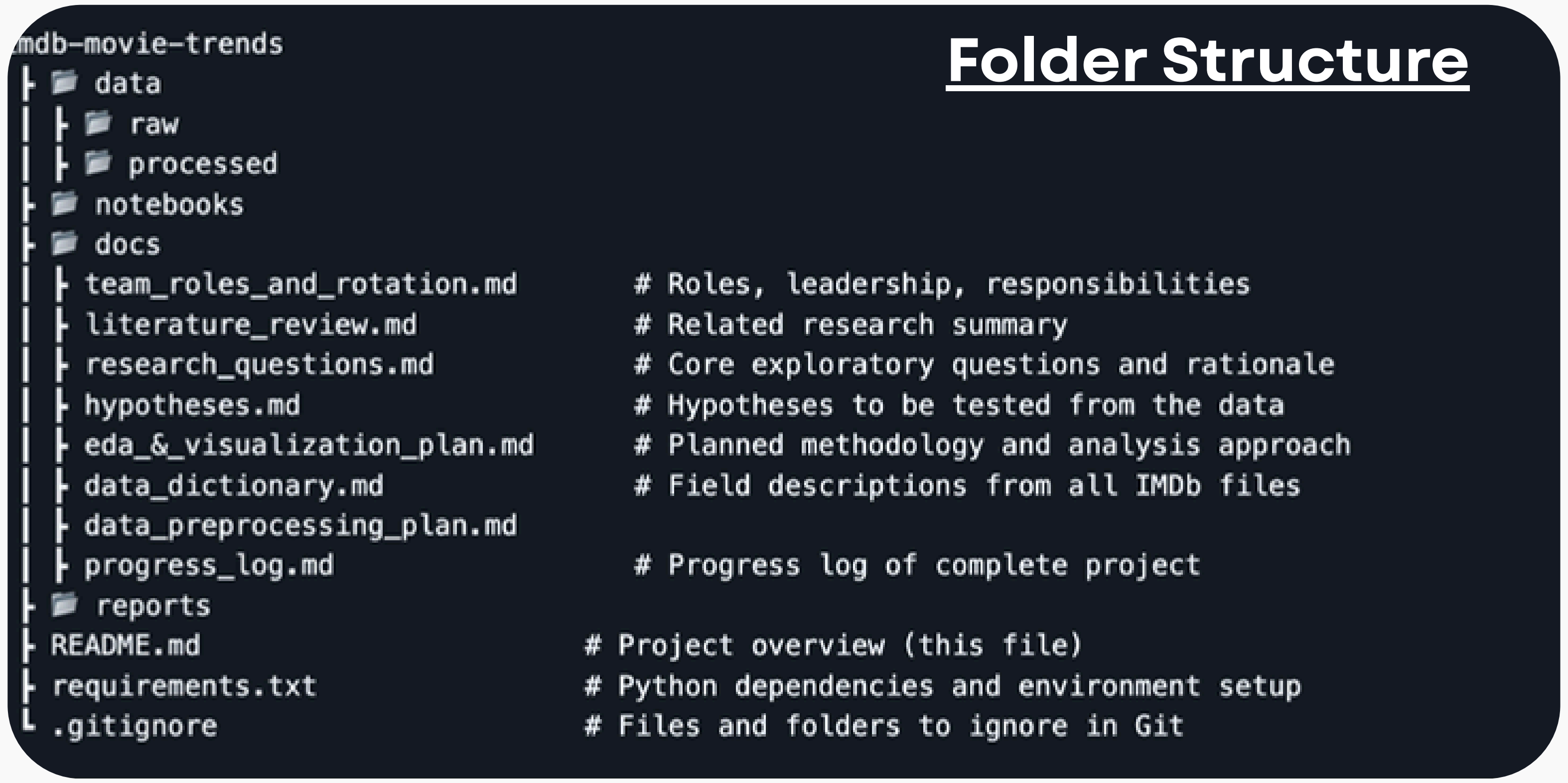
□ Create folder structure (data, docs, notebooks, etc.). `setup`
#1 · by nandu-99 was closed last week



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
```

LITERATURE REVIEW

Paper	Dataset / Scope	Methods Used	Key Insights	Gap Identified
Bahraminasr & Vafaei-Sadr (2020) — *IMDb Data from Two Generations (1979–2019)*	79,000+ movies (1979–2019) with ratings, votes, genres, certificates, languages, countries	Statistical analysis + basic ML	• Metadata influences ratings (genre, MPAA rating, etc.)• Clear temporal trends in ratings & production volume• Regional bias: US vs non-US ratings differ	No analysis of individual actor/director impact; limited modelling depth
Movie Popularity Classification using C4.5, PART & Correlation Coefficient	IMDb-style attributes (genre, duration, year, directors, actors, etc.) approx 1,000 titles (Pre-Release Between 2000 to 2011)	C4.5, PART decision rules, correlation	• Content attributes can classify popularity with reasonable accuracy• Simple rule-based models capture patterns in viewer preference	Does not explore temporal trends or multi-genre effects and uses limited data.

UPDATED RESEARCH QUESTIONS

- Foundational Analysis: Tracking Movie Genre Popularity and Quality
- TV Show "Rating Decay": Analyzing Quality Over Time
- The "Genre Hybridity" Paradox: Complexity vs. Audience Engagement
- The "Creative Hierarchy": Director vs. Writer vs. Cast Impact

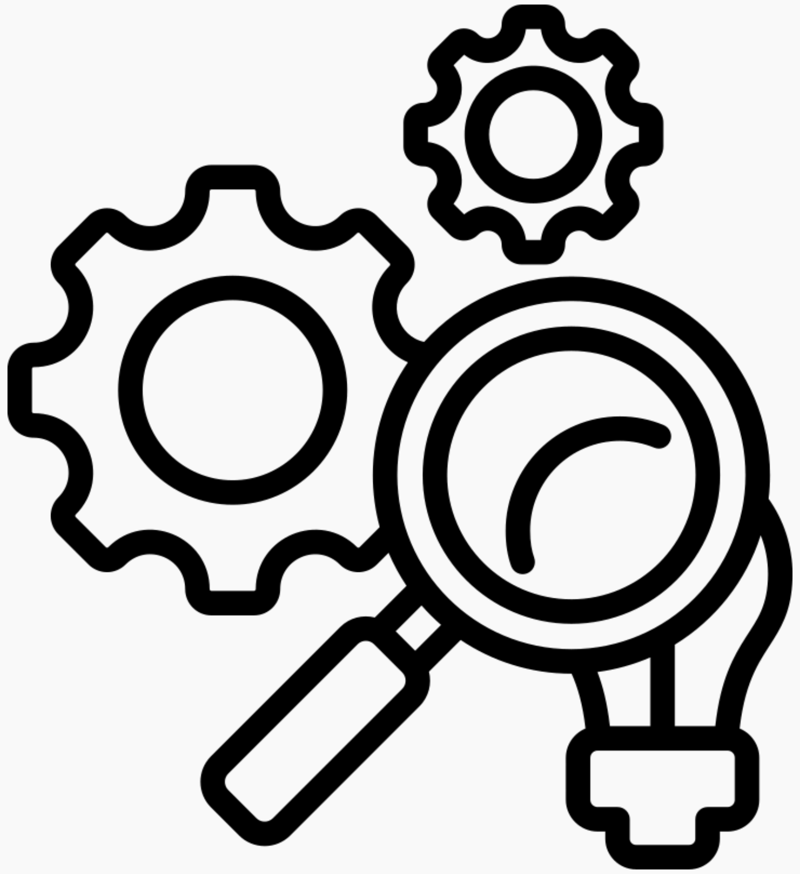
UPDATED HYPOTHESES

- H1: Do major movie genres show significant long-term trends in popularity and quality over the last three decades?
- H2: Do long-running TV shows exhibit “rating decay” as seasons progress?
- H3: Does genre hybridity (multi-genre movies) lead to lower or higher audience ratings compared to single-genre films?
- H4: Does the creative hierarchy (writer, director, cast) vary in importance across different genres when predicting movie success?

METHODOLOGY

RESEARCH METHODOLOGY

- **Design:** Quantitative, exploratory & predictive data-mining approach
- **Data Source:** IMDb datasets (title.basics, title.ratings, title.akas, title.crew, title.episode, title.principals, name.basics)
- **Data Processing:** Cleaned, standardized, filtered movies; merged tables via keys (tconst, nconst, parentTconst)
- **Modeling:** Classification to categorize average ratings into Low / Medium / High
- **Features:** Title type, year, runtime, genre, director/writer/cast profiles, vote counts
- **Analysis Techniques:** Aggregation, time-series trends, genre interaction; visualizations via line charts, bar charts, scatter plots, boxplots
- **Outcome:** Scalable framework to analyze genre evolution, talent impact & rating patterns



EXPLORATORY DATA ANALYSIS (EDA) & DESCRIPTIVE ANALYSIS

- **Dataset Overview:** 12M+ titles, 9 columns; key columns (titleType, primaryTitle, isAdult) complete
- **Missing Values:** endYear 98.7%, runtimeMinutes 64.6%, startYear 12%
- **Outliers:** startYear outliers (~14%) removed; runtime outliers ignored
- **Sampling:** Stratified sample of 200k titles for analysis
- **TitleType Distribution:** Movies, Shorts, TV Episodes; distribution maintained in sample
- **Genre Analysis:** Top genres – Drama, Comedy, Action, Short, Documentary; genre pairs analyzed
- **Temporal Trends:** Movie releases growing; genre popularity (Drama, Comedy, Action) rising over decades
- **Adult Content:** Mostly non-adult; slightly higher in TV Episodes
- **TitleType vs Genre:** Shorts → Short, Movies & TV Episodes → Drama
- **Key Takeaways:** Drama & Comedy dominate; Shorts mostly "Short" genre; genre popularity increasing; adult content rare

ANALYSIS

MOVIE GENRE ANALYSIS: POPULARITY & QUALITY SINCE 1995

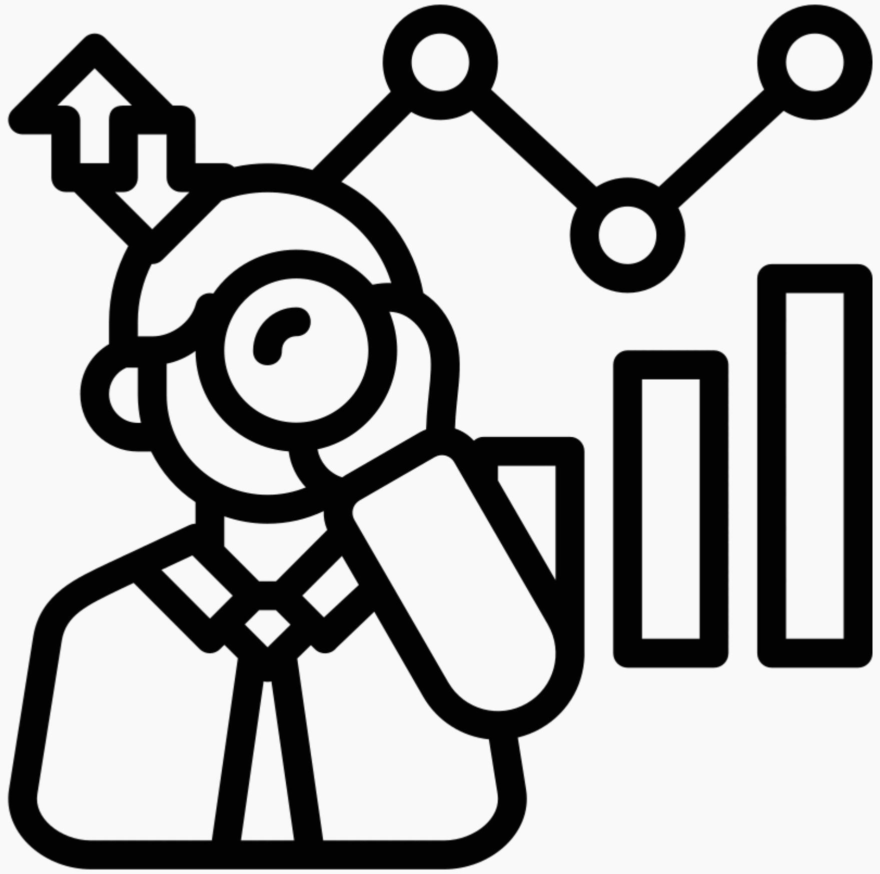
- **Question:** How have genre popularity (votes) and quality (ratings) changed over time?
- **Analysis Steps:**
  - Filtered movies (1995+), expanded multi-genre titles
  - Stratified sample (~100k) with ratings added
  - Calculated median rating & votes per genre per year
- **Key Insights:**
  - Quality Stable: Median ratings mostly steady
  - Popularity Declining: Votes more volatile
  - Documentaries: Highest quality, lowest popularity
  - Popular Genres: Mystery, Adventure, Crime → moderate quality
  - Horror: Lowest-rated but moderate interest
  - Drama: Most produced (~24% of movies)

NOTEBOOK

ANALYSIS

PREDICTIVE MODELING SUMMARY

- **Objective:** Multi-class classifier to predict movie success tiers (Low, Medium, High).
- **Data Prep:** Merged IMDb datasets; engineered features (log votes, one-hot genres).
- **Target:** Success tiers – Low (≤3.0), Medium (<7.0), High (≥7.0).
- **Sampling:** Stratified sample (~100k rows) preserving movie type distribution.
- **Modeling:** Logistic Regression, Decision Tree, Random Forest (best ~74% accuracy).
- **Key Drivers:** Vote count (audience engagement) strongest predictor.
- **Impact:** Early buzz predicts perceived quality; helps pre-release risk assessment.



DASHBOARD

INTERACTIVE DASHBOARD: DIRECTOR MOVIE EXPLORER

- **Purpose:** Explore a director’s top 3 movies based on IMDb ratings
- **Data Used:** Filtered IMDb dataset (directors with ≥3 movies)
- **Features:**
  - Dropdown to select director
  - Displays top 3 movies with rating
  - Interactive plots & tables
- **Tech Stack:** Python, Streamlit, DuckDB, Plotly
- **Outcome:** Quick insights into director performance and movie quality trends

THANK YOU

[GET READY FOR FINAL PHASE]