Movie Recommendation System

TEAM: MovieRec

Dan Ahimbisibwe | Danish Puri | Emmanuel Owusu-Ampaw | Jiawei Zhao | Shiyue Zhu

TABLE OF CONTENTS

TABLE OF CONTENTS

TABLE OF CONTENTS

GOAL

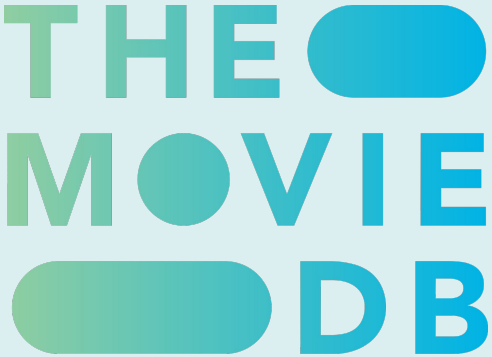
- Build a Movie Recommendation System using the TMDB 5000 dataset.

What Is a Recommender System?

- Software that predicts what a user will like next and ranks items (movies, songs, products) to personalize the experience.
- Uses behavioral signals (ratings, clicks, watch time) and item features (genres, cast, text)

About the Dataset

- TMDB(The Movie Database) — an open-source platform for movie metadata like ratings, cast, and genres.
- Contains 4,802 movies, 40+ features
- Includes budget, revenue, profit, cast, crew, and genres
- Movies range from 1910s–2010s, mostly 2000s onward

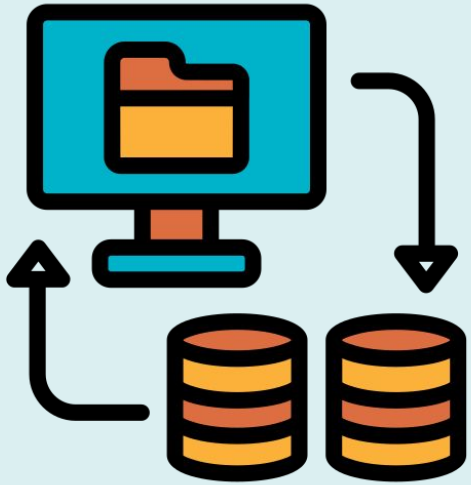


Dataset Exploration

- Loaded and examined the structure of both TMDB(The Movie Database) and credits datasets.
- Assessed data quality by identifying missing values, zero values, and duplicates.
- Analyzed statistical distributions of key features like budget, revenue, and ratings.
- Created improved visualizations using log-scale transformations for skewed data.
- Explored genre patterns and temporal trends across different decades.
- Conducted correlation analysis to understand feature relationships.
- Identified top-performing movies across different metrics.
- Discovered key insights about budget vs ratings and movie distribution patterns.
- This exploration guided our data cleaning and feature engineering strategies.

Data Quality Issues

- Identified significant missing values in the homepage field, with 64.4% of movies lacking homepage information.
- Tagline field had missing values in 17.6% of movies, affecting content-based recommendation features.
- Discovered zero values in critical financial columns that likely represented missing data rather than actual zeros.
- Budget column contained 1,037 zero values that needed to be treated as missing data.
- Revenue column had 1,427 zero values that required special handling for accurate financial analysis.
- Runtime column showed 35 movies with zero-minute duration, which were clearly missing data entries.
- Found JSON-formatted data in multiple columns that needed parsing to extract useful information.
- The release_date column contained one missing value that required removal to maintain data integrity.
- Data type inconsistencies existed where dates were stored as strings instead of datetime objects.
- The dataset required comprehensive cleaning to ensure all features were properly formatted for analysis.



Data Cleaning Steps

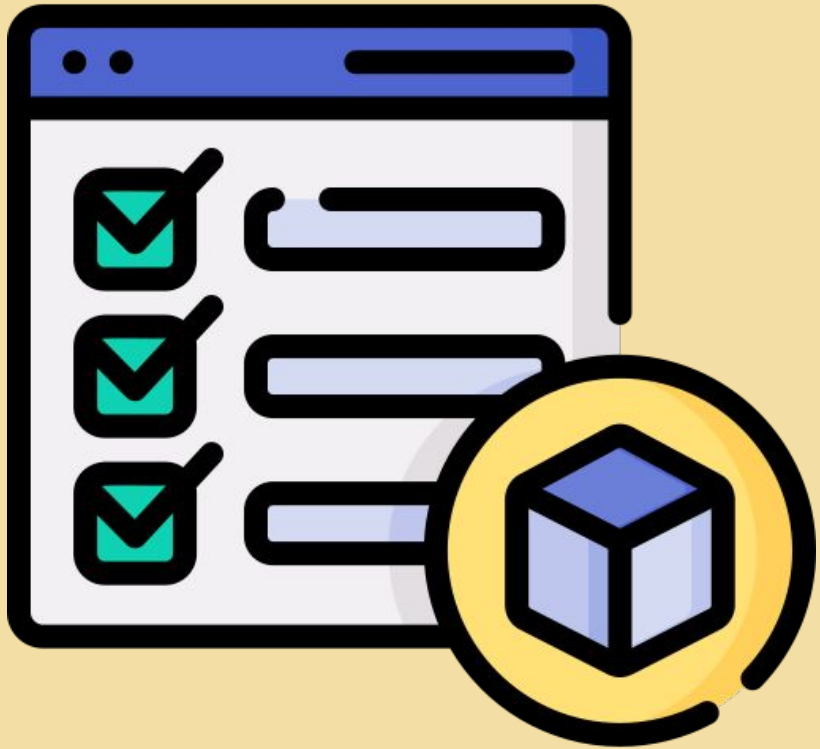
- Replaced zero values with NaN in budget, revenue, and runtime columns.
- Filled missing values in homepage and tagline with placeholder text.
- Removed movies with missing release dates.
- Converted release_date from string to datetime format.
- Parsed JSON data in genres, keywords, and production columns.
- Extracted director and actor information from credits data.
- Merged movies and credits datasets using movie ID.
- Created profit, year, decade, and genre count features.
- Generated combined text features for content filtering.
- Calculated normalized popularity scores.
- Performed final quality checks on cleaned dataset.

Parsing JSON & Merging

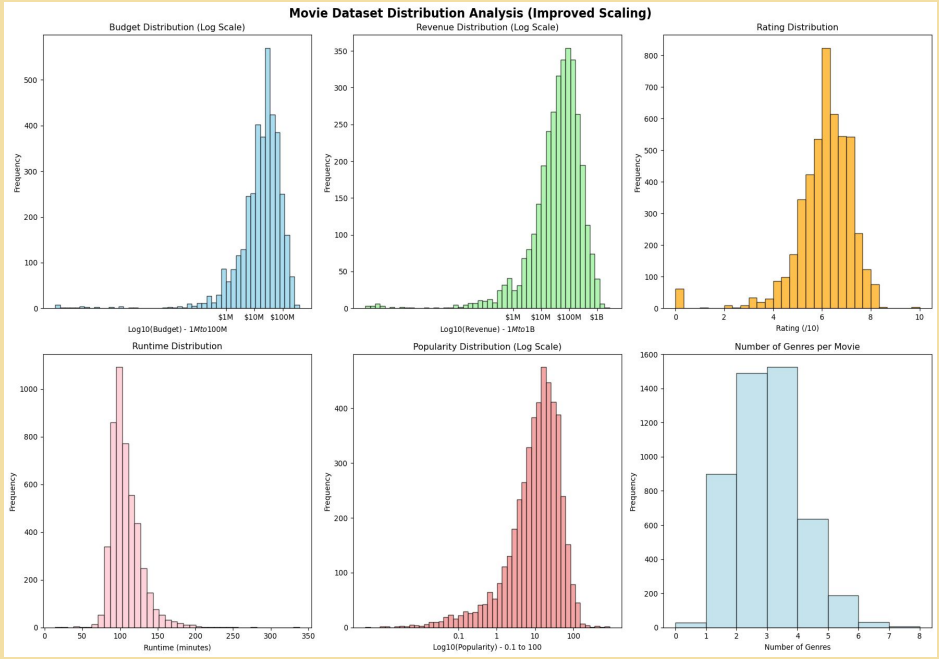
- Parsed JSON data in genres, keywords, and production columns using ast.literal_eval().
- Extracted genre names, keywords, and production information from nested structures.
- Processed credits data to extract director names and top actors.
- Merged movies and credits datasets using movie ID as primary key.
- Combined 4,802 movie records with cast and crew information.
- Handled duplicate column names during the merge process.
- Created comprehensive 40-feature dataset for recommendation systems.
- All JSON data converted to usable list and string formats.

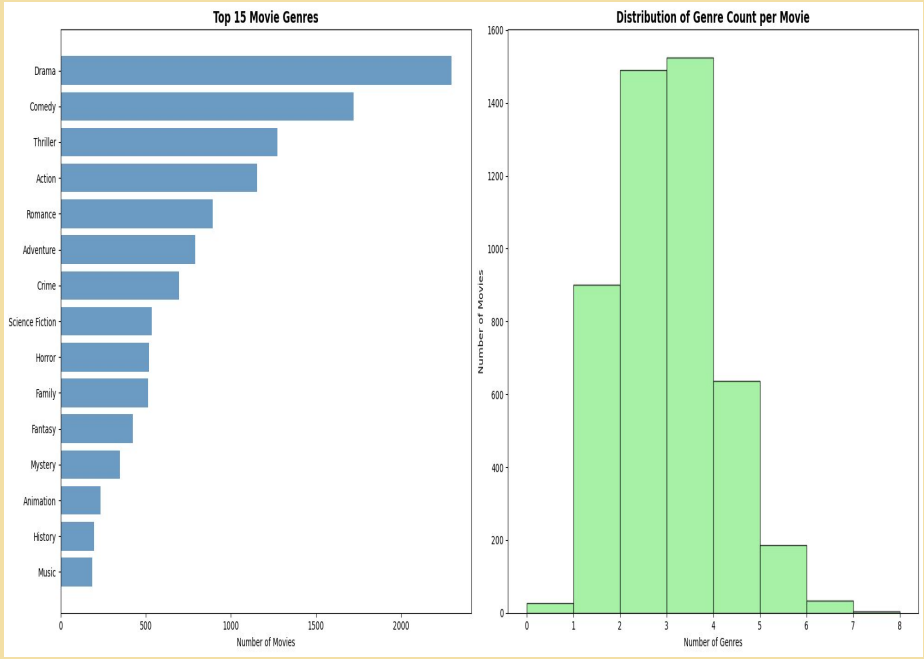
Feature Creation

- Created profit and profit margin features from budget and revenue data.
- Extracted year and decade features from release dates.
- Generated normalized popularity scores for better scaling.
- Created genre count feature to measure movie diversity.
- Built combined text features from overview, tagline, and title.
- Added director and top actors information from credits data.
- Generated features supporting content-based and collaborative filtering.
- Created comprehensive 40+ feature dataset for recommendation systems.

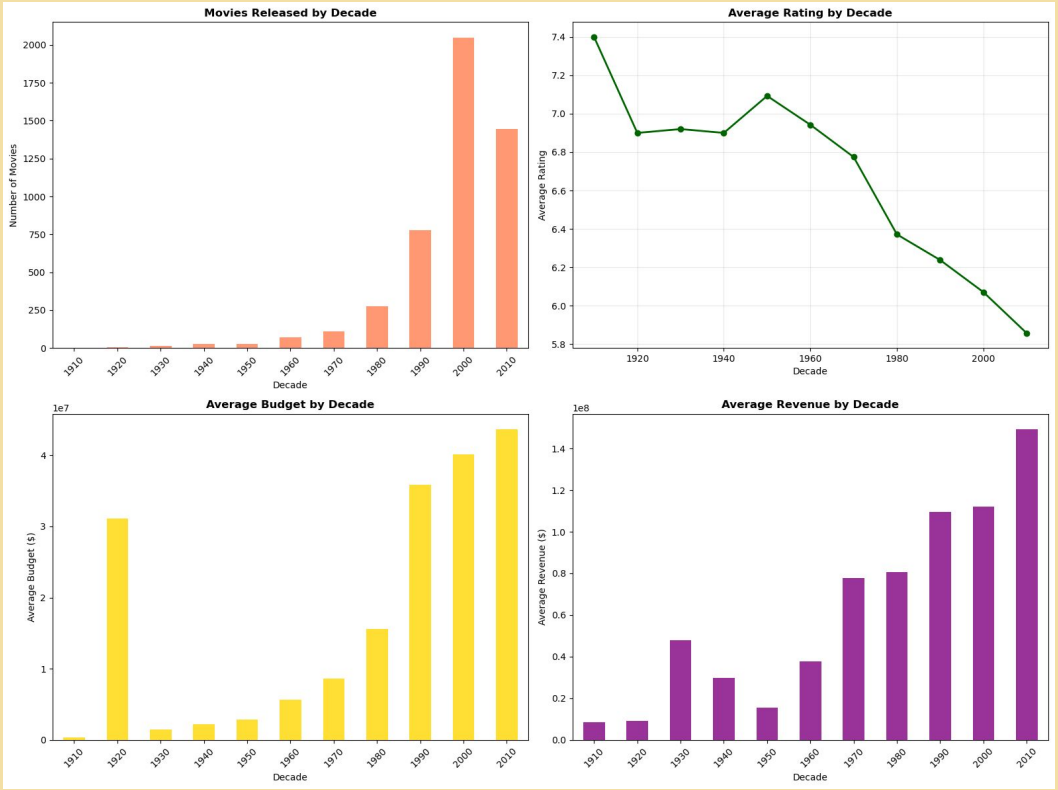


Visual Insights (1)

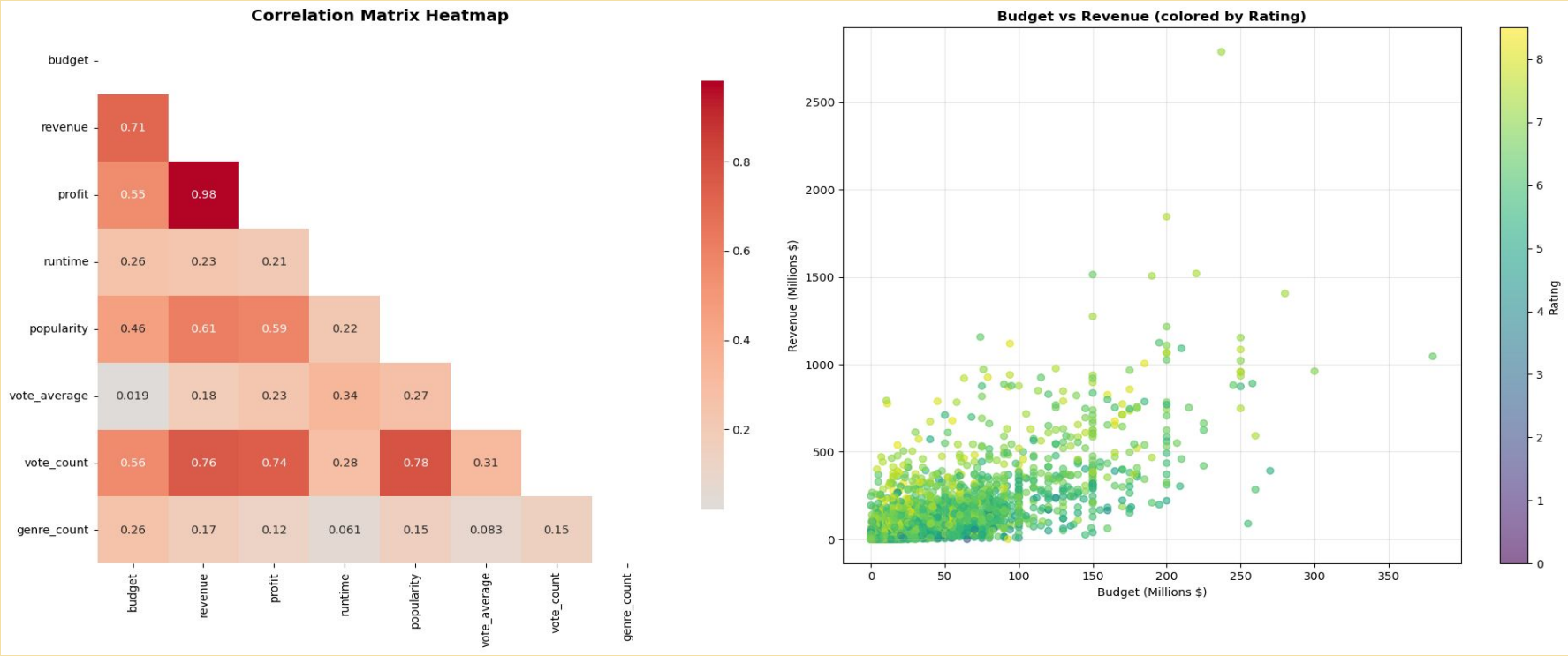




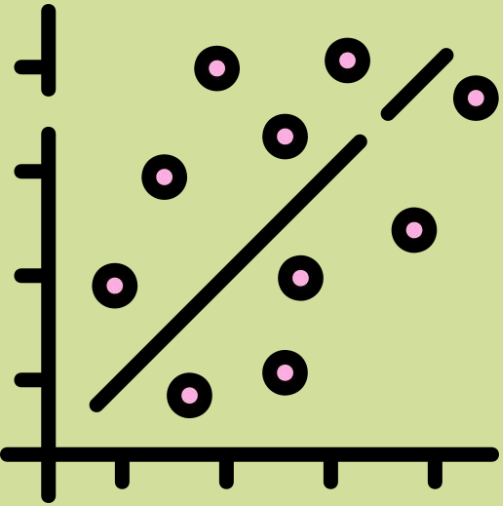
Visual Insights (2)



Visual Insights (3)



Correlations & Top Movies



- Budget vs Revenue correlation = 0.70
- Top-grossing: Avatar, Titanic, Avengers
- Top-rated: Shawshank Redemption, The Godfather

Key Takeaways & Next Week

- Data cleaned and ready for modeling
- Insights visualized successfully

THANK YOU