Movie Recommendation System

Popcorn Predictors - Team #2

PRESENTED BY:
Santiago Benheim, Cherron Griffith, Ethan Iwama, Eren Kaval, Meghna Sharma, Patrick Wang
04/30/2025

About...

Goal: Recommend movies using user preferences and movie metadata

Dataset: TMDB 5000 Movie Dataset

Techniques:

- Principal Component Analysis
- Random Forest
- K-Nearest Neighbor Models

Tools:

Exploratory Data Analysis (EDA):

- Data Cleaning (pandas & numpy)
- JSON parsing for structured metadata
- Column Selection (PCA)
- Visualization (seaborn & matplotlib)

Machine Learning:

- One-Hot Encoding/TF-IDF Vectorization
- Cosine Similarity KNN
- Euclidean/Manhattan Distance KNNs

NYU

Part 1: Data Exploration and Feature Analysis

Preparation for Data Training

01

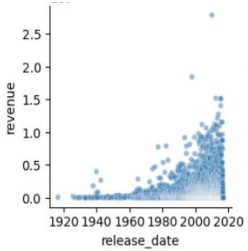
NYU

Data Cleaning

Exploratory Data Analysis (EDA) was conducted on both datasets to identify missing values, correct column data types, and uncover significant relationships relevant to our project goal.

```
movies.head()
movies.info()
movies.shape()
movies.dtypes
```

```
movie_id    0
title       0
cast        0
crew        0
dtype: int64
```



01.

Corrected column data types and extracted relevant values from columns containing lists of Python dictionaries

02.

Handled missing values by ensuring they were properly represented according to the column data type, without discarding too much data

03.

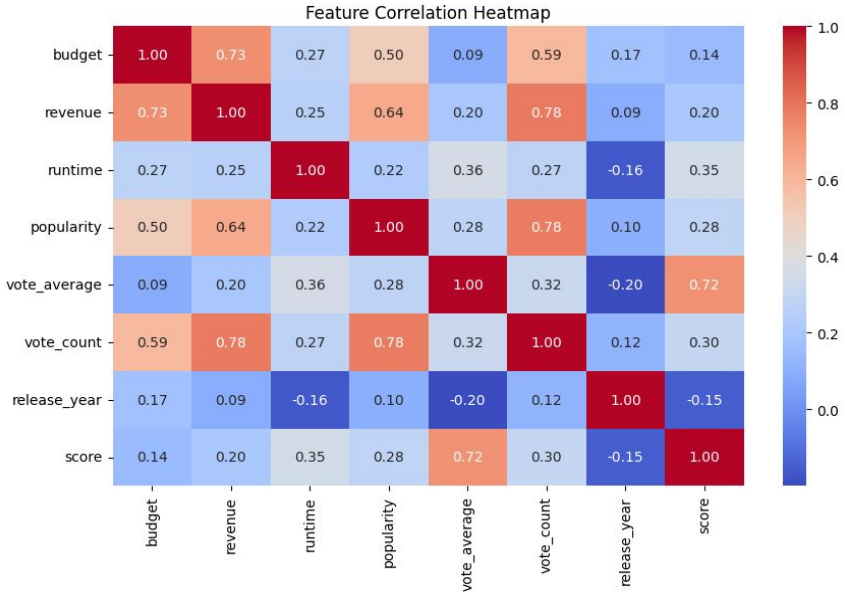
Refined the dataset to include only columns relevant to our project

04.

Created visualizations to uncover significant relationships within the data

NYU

Scoring Function



- Created a scoring function similar to the scoring metric of IMDB
    - considers both vote_average and vote_count
- Filtered out for the 95th percentile of vote_counts
    - excludes movies with low vote_counts
    - filled missing scores with 0
- Added score column to dataframe to be used for analysis

NYU

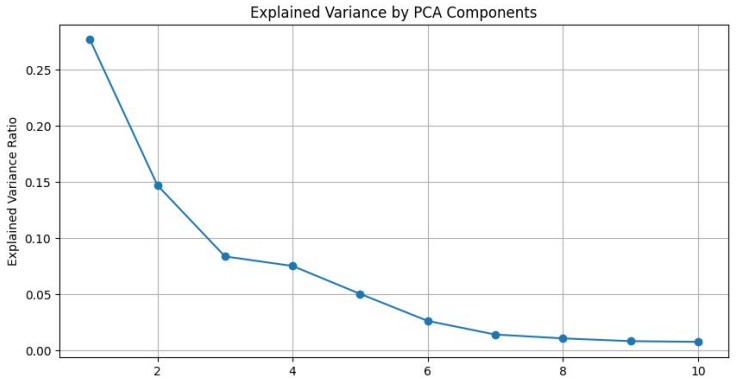
Data Preprocessing

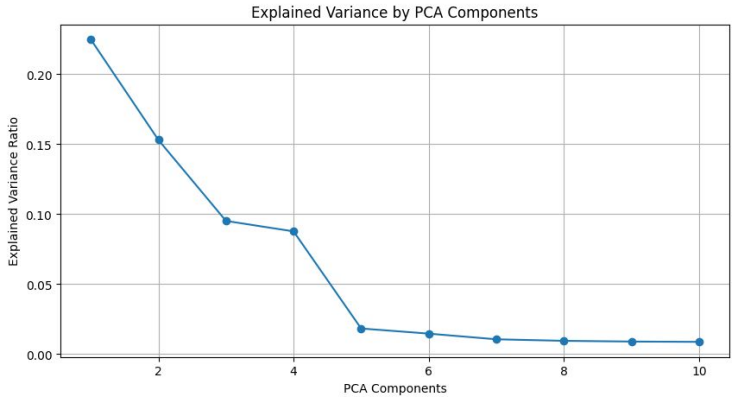
Converted categorical columns to numerical format

- **One Hot Encoding** was used on language
- **TF-IDF Vectorization** was used on genres, keywords, directors, and actors

Performed **Principal Component Analysis (PCA)** after standardizing to identify features that explain the most variance within the data

- Reduced dimensionality for improved computational efficiency and model performance
- 2 different PCA sets
  - First plot: With Revenue and Budget: Explains 70 % of variance
  - Second plot: Without Revenue and Budget:
    - Removed after Feature Relevance test with Score metric: Explains 63% of variance





NYU

Part 2: Recommendation Model Development

Evaluating Different Recommendation Models

02

NYU

Recommendation Models

Implemented **K-Nearest Neighbors (KNN)** using 3 distance metrics:

- configured these models to recommend the top 5 similar movies

Cosine Similarity:

- Measures the cosine of the angle between two vectors

Euclidean Distance:

- Calculates the straight-line distance between two vectors

Manhattan Distance:

- Computes the sum of absolute differences across dimensions

NYU

Simulation Scenarios

Each model's **accuracy metrics** were tested using simulated users in three watch history scenarios:

- Each user has randomly selected movies in their history
- Each user has randomly selected high-scored (score > 7.0) movies in their history
- Each user has randomly selected non-English movies in their history

- Each user is recommended 5 movies
- For each user, the recommended movies relevancy is determined
- **Precision, Recall, F1, and MAP** are determined based on how many recommendations are relevant
- Metrics are found for each user, then averaged for each scenario

Model Accuracy

A recommendation is determined relevant if:

- Its score is above the score threshold
- It shares ≥25% of its genres with the user's watched genres

Score threshold is determined using the mean and standard deviation of the user's watch history:

- Mean - 2*(Standard Deviation)

Metrics were determined understanding:

- **True Positives** were relevant movies recommended
- **False Positives** were non-relevant movies recommended
- **False Negatives** were relevant movies in the entire dataset not found in recommendations

NYU

Part 3: Results and Visualizations

Conclusions and Future Work

03

NYU

Example Recommendations:

```
#Some Harry Potter Movies
display(recommend_movies_cosine(['Harry Potter and the Half-Blood Prince','Harry Potter and the Order of the Phoenix','Harry Potter and the Prisoner of Azkaban']))
display(recommend_movies_euclidian(['Harry Potter and the Half-Blood Prince','Harry Potter and the Order of the Phoenix','Harry Potter and the Prisoner of Azkaban']))
display(recommend_movies_manhattan(['Harry Potter and the Half-Blood Prince','Harry Potter and the Order of the Phoenix','Harry Potter and the Prisoner of Azkaban']))
print()
```

Cosine

	title	score	genres	main_actors	director	original_language
114	Harry Potter and the Goblet of Fire	7.498761	Adventure Fantasy Family	Daniel Radcliffe Rupert Grint Emma Watson Ralp...	Mike Newell	en
22	The Hobbit: The Desolation of Smaug	7.598354	Adventure Fantasy	Martin Freeman Ian McKellen Richard Armitage K...	Peter Jackson	en
197	Harry Potter and the Philosopher's Stone	7.499008	Adventure Fantasy Family	Daniel Radcliffe Rupert Grint Emma Watson Rich...	Chris Columbus	en
262	The Lord of the Rings: The Fellowship of the Ring	7.998915	Adventure Fantasy Action	Elijah Wood Ian McKellen Cate Blanchett Orland...	Peter Jackson	en
63	The Chronicles of Narnia: The Lion, the Witch ...	6.698878	Adventure Family Fantasy	William Moseley Anna Popplewell Skandar Keynes...	Andrew Adamson	en

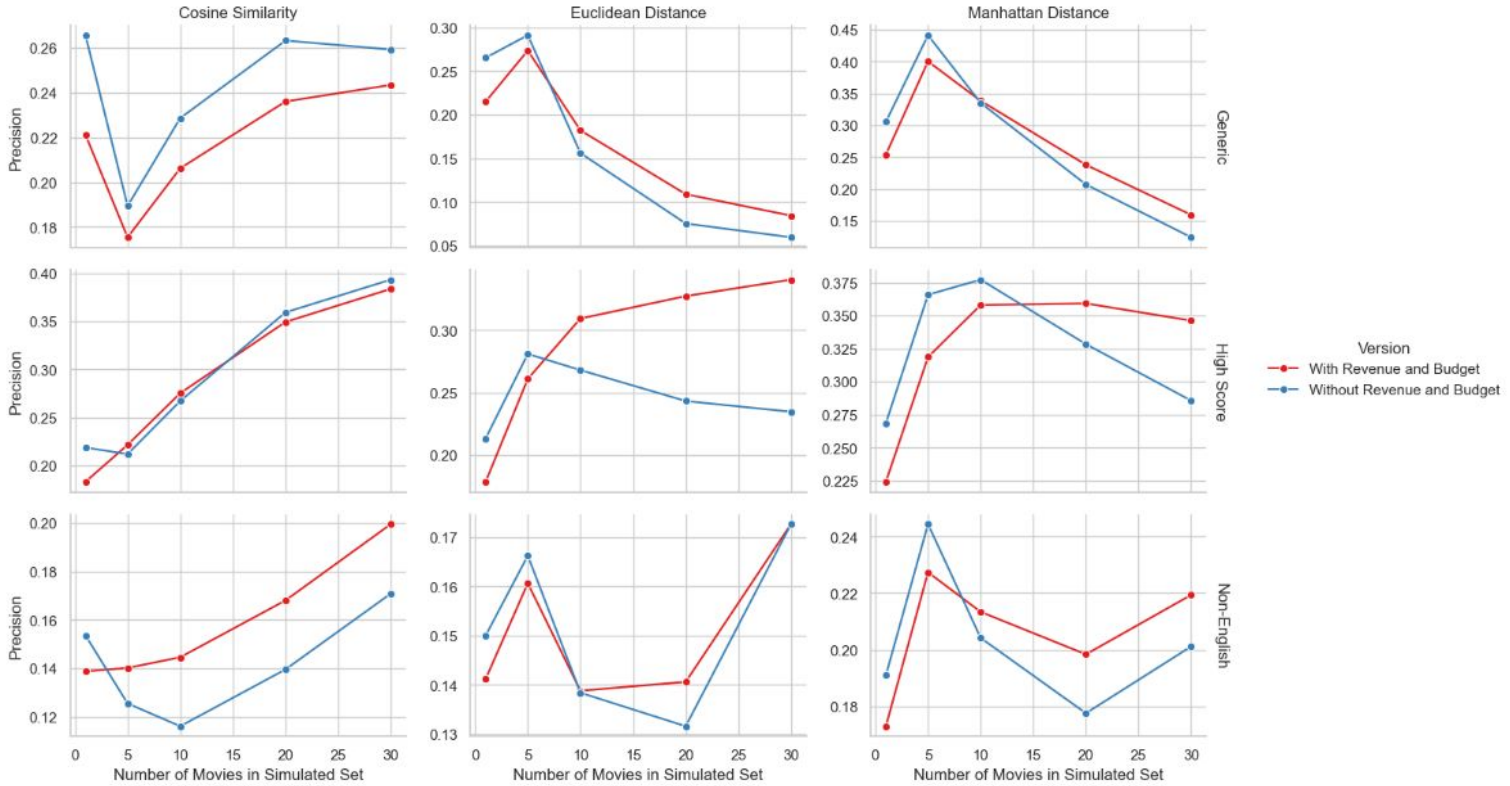
Euclidian

	title	score	genres	main_actors	director	original_language
114	Harry Potter and the Goblet of Fire	7.498761	Adventure Fantasy Family	Daniel Radcliffe Rupert Grint Emma Watson Ralp...	Mike Newell	en
277	Casino Royale	7.298457	Adventure Action Thriller	Daniel Craig Eva Green Mads Mikkelsen Judi Den...	Martin Campbell	en
63	The Chronicles of Narnia: The Lion, the Witch ...	6.698878	Adventure Family Fantasy	William Moseley Anna Popplewell Skandar Keynes...	Andrew Adamson	en
932	V for Vendetta	7.698211	Action Thriller Fantasy	Natalie Portman Hugo Weaving Stephen Rea Steph...	James McTeigue	en
183	The Hunger Games: Catching Fire	7.399007	Adventure Action Science Fiction	Jennifer Lawrence Josh Hutcherson Liam Hemswor...	Francis Lawrence	en

Manhattan

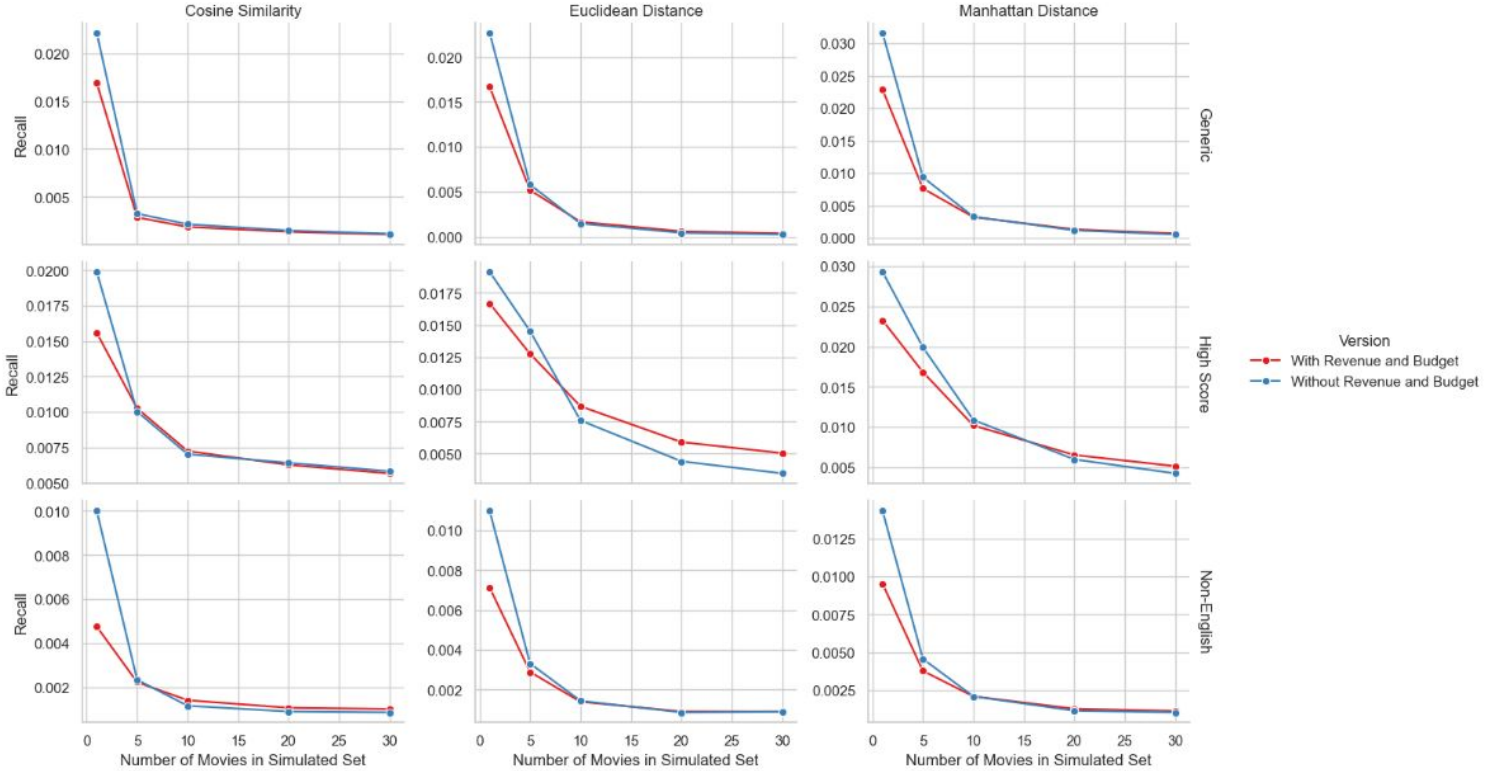
	title	score	genres	main_actors	director	original_language
114	Harry Potter and the Goblet of Fire	7.498761	Adventure Fantasy Family	Daniel Radcliffe Rupert Grint Emma Watson Ralp...	Mike Newell	en
63	The Chronicles of Narnia: The Lion, the Witch ...	6.698878	Adventure Family Fantasy	William Moseley Anna Popplewell Skandar Keynes...	Andrew Adamson	en
20	The Amazing Spider-Man	6.499704	Action Adventure Fantasy	Andrew Garfield Emma Stone Rhys Ifans Denis Le...	Marc Webb	en
22	The Hobbit: The Desolation of Smaug	7.598354	Adventure Fantasy	Martin Freeman Ian McKellen Richard Armitage K...	Peter Jackson	en
197	Harry Potter and the Philosopher's Stone	7.499008	Adventure Fantasy Family	Daniel Radcliffe Rupert Grint Emma Watson Rich...	Chris Columbus	en

Precision



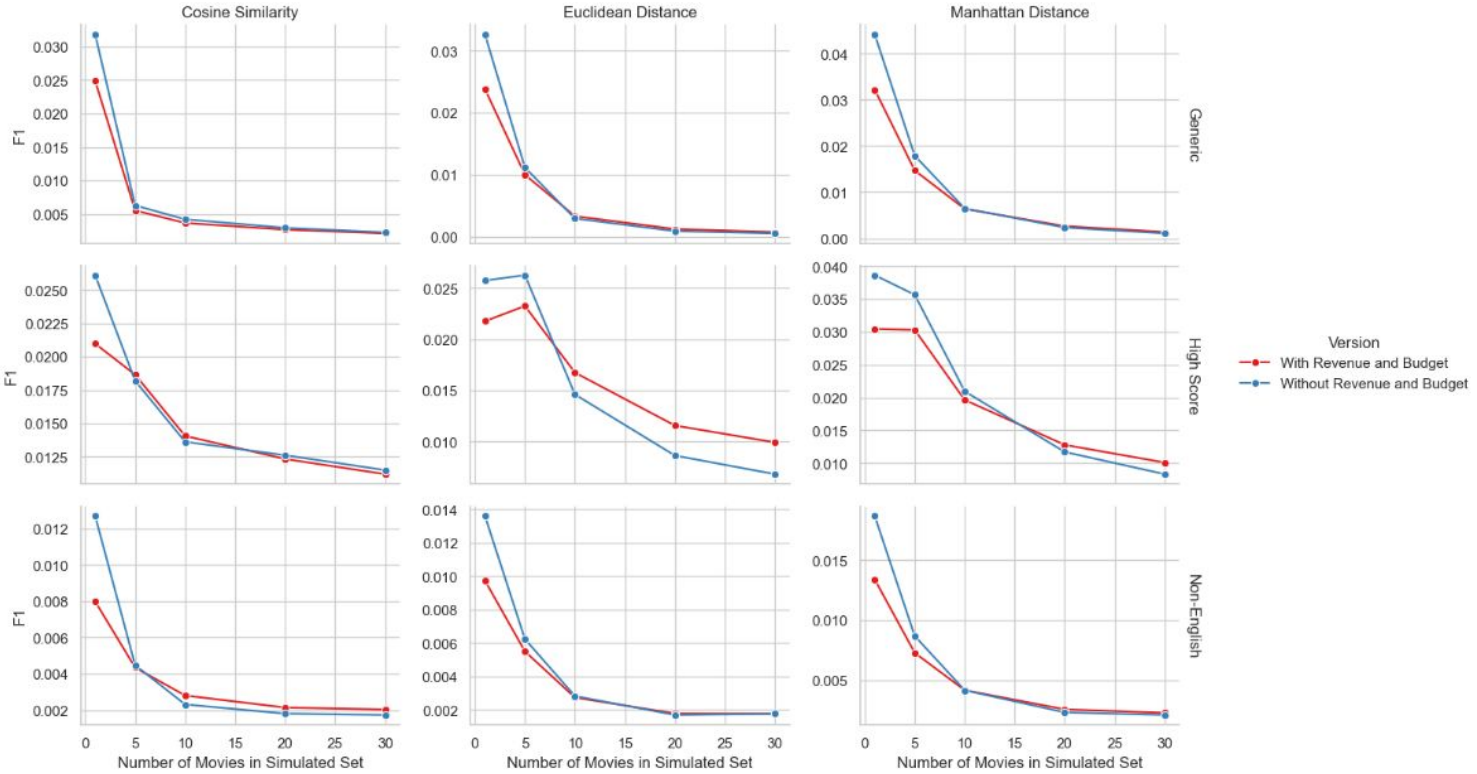
NYU

Recall

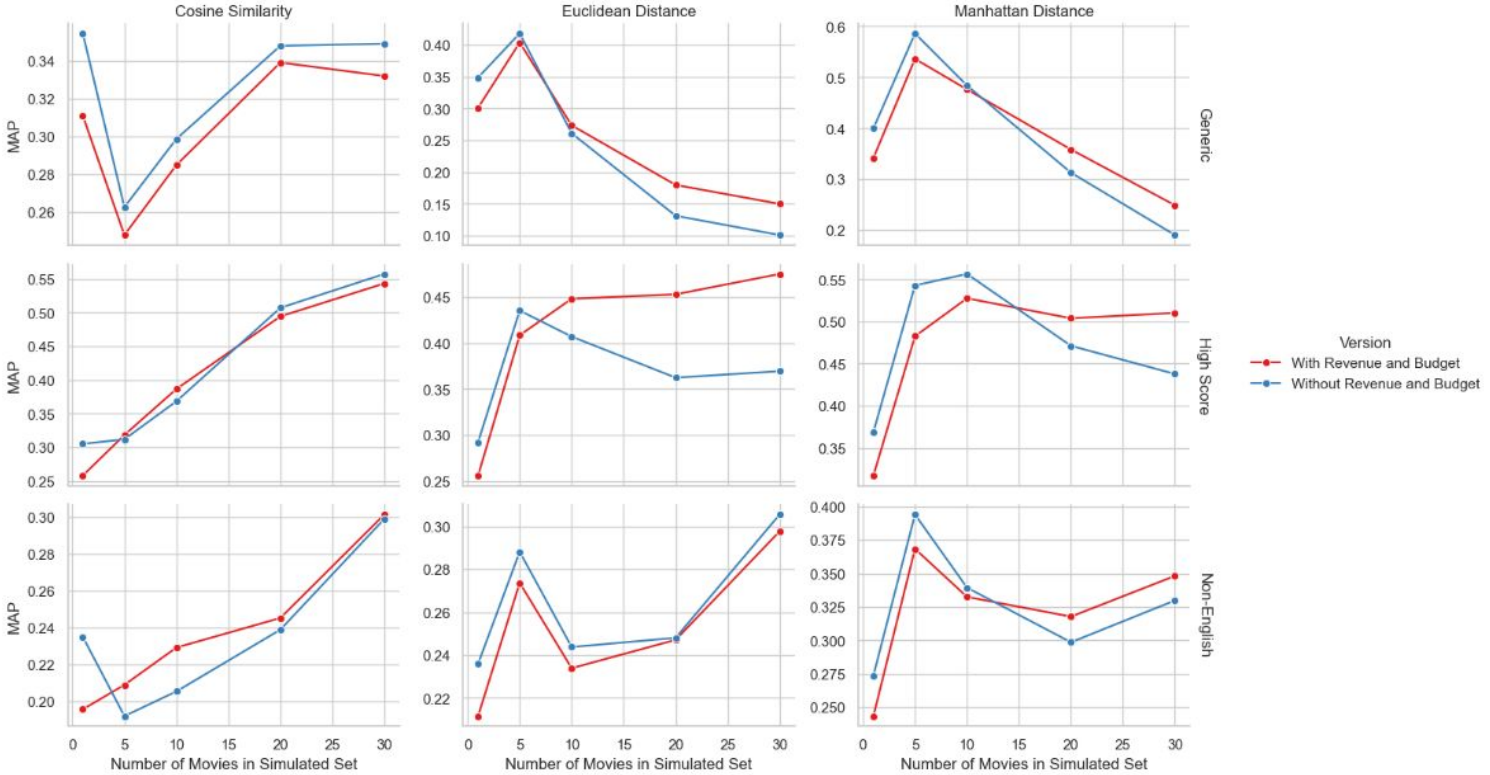

Cosine Similarity
Euclidean Distance
Manhattan Distance
Recall
Generic
High Score
Non-English
Number of Movies in Simulated Set
Version
With Revenue and Budget
Without Revenue and Budget

NYU

F1


Cosine Similarity
Euclidean Distance
Manhattan Distance
F1
Generic
High Score
Non-English
Number of Movies in Simulated Set
Version
With Revenue and Budget
Without Revenue and Budget

Mean Average Precision



NYU

Thank You!

Popcorn Predictors - Team #2

PRESENTED BY:
Santiago Benheim, Cherron Griffith, Ethan Iwama, Eren Kaval, Meghna Sharma, Patrick Wang
04/30/2025