ITMO

THE MOVIE DATABASE ANALYSIS

From Collection to Insight

COURSE:
BIG DATA & TECHNOLOGIES

PRESENTED BY:
SYED MUHAMMAD SHAJEE RAZA
ABU BAKAR

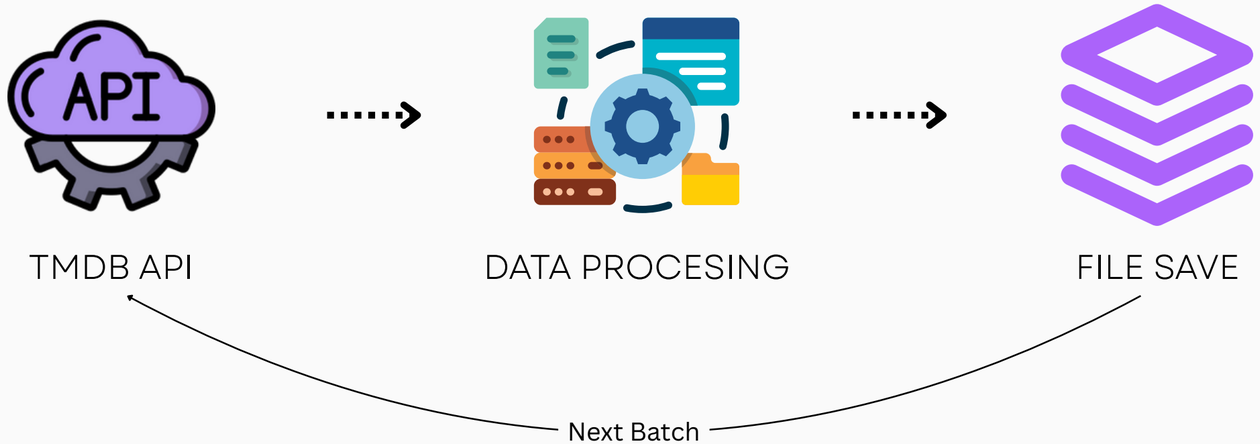
SYSTEM OVERVIEW

OBJECTIVE

Leverage big data technologies to analyze TMDB movie records, identify trends, and build a predictive model forecasting popular genres for 2026 using machine learning.

Why TMDB?

TMDB (The Movie Database) is used because it provides rich, accurate, and up-to-date data about movies, TV shows, and cast details. It's a reliable source for building entertainment-related applications or analyses.



DATA GATHERING OVERVIEW

Data Gathering from TMDB API – Foundation for Big Data Processing

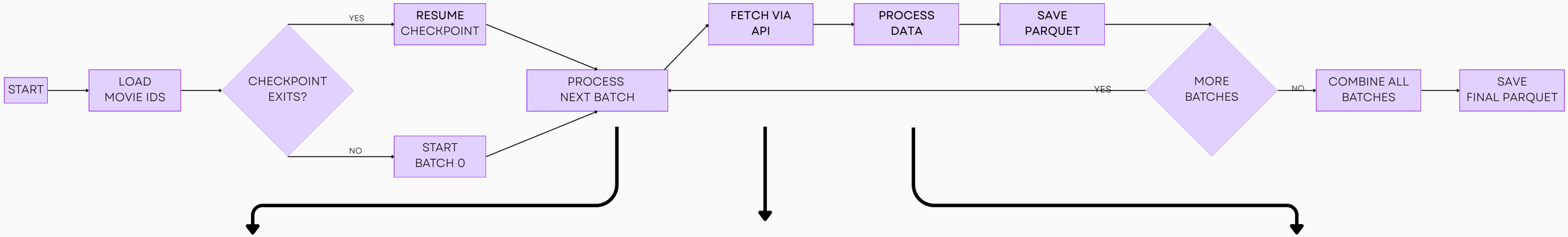
From raw movie data to meaningful insights – this **5Vs** breakdown shows how we turned millions of TMDB records into a clean, powerful dataset ready for analytics, recommendations, and smart decision-making.

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE
Extracted: 15 Million Movie IDs	Collected: Diverse Fields	Ensuring up-to-date (till this week) & relevant data	Robust Error Handling HTTP errors, Network timeouts,	Supports analytics, recommendations, and business decisions
Retrieved: 0.4 Million Movies	Metadata: Title, Cast, Crew Roles	Rate Limit Handling, Rate limit of 40 requests per 10 seconds	Retry Logic, Response Filtering	Enables meaningful insights into movie trends, user preferences, and performance metrics
Stored: Parquet Format	Date, Budget, Revenue, Popularity, Vote Count	Parallel Execution for Throughput, via Thread Pooling	Deduplication, Final Validation	High-quality, curated dataset ready for downstream use in BI, ML, or reporting tools

TECHNOLOGIES & TASKS

Data Gathering: Tools, Technologies & Methodology | Readiness

Methodology



RateLimitedSession:	**fetch_movie_details:**	**process_movie_data:**
• Initiolaze with retry strategy • Track request times • Check rates limits • Sleep if needed • Execute request with retries	Make API request for movie data handle errors (404, rate limits, timeouts)	• Extracted info • Process Info • Distribute Info • Save Info

MIDDLEWARE INFRASTRUCTURE

Why Apache Spark?

- Dataset Size: ~400K movies with ~20+ columns, including financial and textual metadata
- EFFICIENT PROCESSING OF LARGE PARQUET FILES
- DISTRIBUTED COMPUTING FOR FAST TRANSFORMATIONS (E.G., FILTERING, AGGREGATING)

Data Pipeline Overview:

[PARQUET FILES] → [SPARK DATAFRAME] → [CLEANED DATA] → EXPLODING & AGGREGATING DATA → VISUALIZATIONS

Data Overview:

TOTAL ROWS: ~400,000
TOTAL COLUMNS: 22

DATA PREPROCESSING USING SPARK

STEP	What we did
Date Parsing	Converted release_date to proper date format and extracted year
Null Handling	Filtered out invalid budget/revenue/genre values
Outlier Removal	Removed future years (e.g., 2030), zero-budget movies
Genre/Cast Splitting	Used split()+ explode()to normalize genres and actors
Derived Metrics	Created profit and ROI fields for financial analysis

SCHEMA

```
root
 |-- id: long (nullable = true)
 |-- title: string (nullable = true)
 |-- original_title: string (nullable = true)
 |-- release_date: string (nullable = true)
 |-- runtime: long (nullable = true)
 |-- budget: long (nullable = true)
 |-- revenue: long (nullable = true)
 |-- popularity: double (nullable = true)
 |-- vote_average: double (nullable = true)
 |-- vote_count: long (nullable = true)
 |-- imdb_id: string (nullable = true)
 |-- genres: string (nullable = true)
 |-- production_companies: string (nullable = true)
 |-- production_countries: string (nullable = true)
 |-- spoken_languages: string (nullable = true)
 |-- cast: string (nullable = true)
 |-- director: string (nullable = true)
 |-- director_of_photography: string (nullable = true)
 |-- writers: string (nullable = true)
 |-- producers: string (nullable = true)
 |-- music_composer: string (nullable = true)
 |-- __index_level_0__: long (nullable = true)

Total Rows: 398260
Total Columns: 22
```

TMDB MOVIE DATASET OVERVIEW

```
========================================
• Total Rows (Movies): 398260
• Total Columns: 22
• Titles Available: 398260 (100.00%)
• Contains financial data (budget/revenue): True
• Contains genre & cast information: True
========================================
```

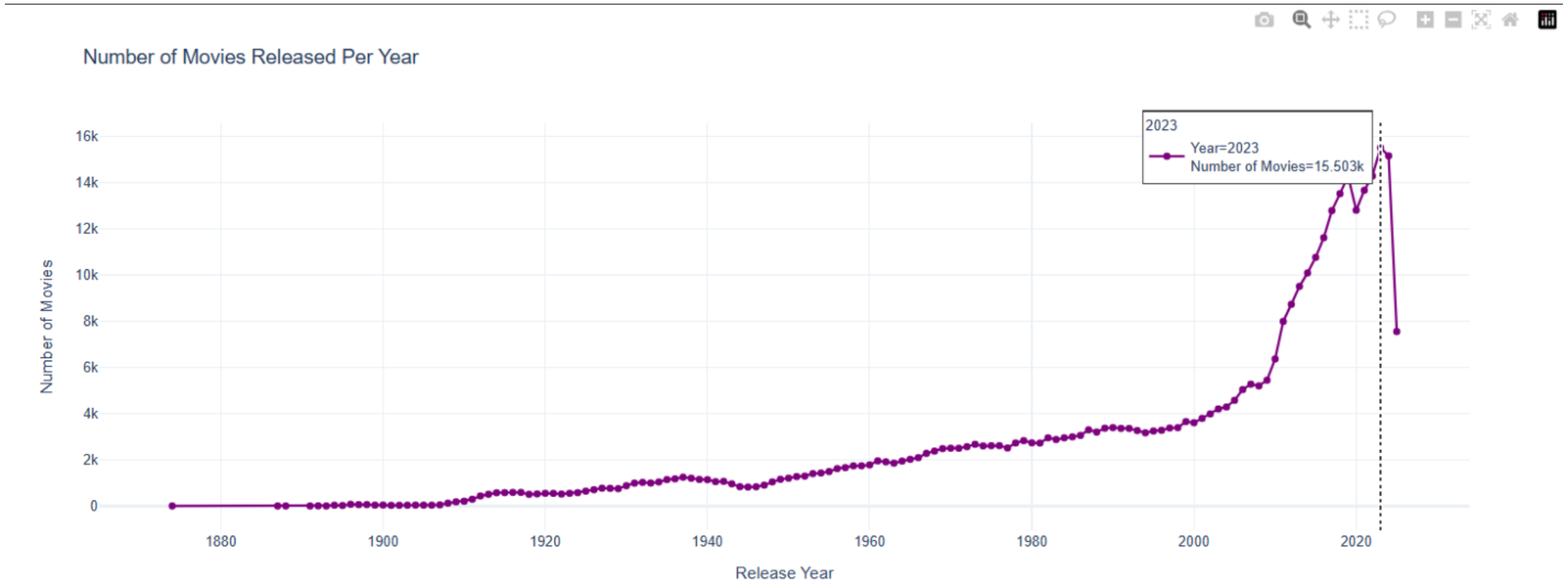
VISUALIZATIONS

Top 10 Years by Movie Count:

Top 10 Years by Movie Count:

release_year	movie_count
2023	15503
2024	15150
2019	14289
2022	14280
2021	13662
2018	13514
2020	12799
2017	12781
2016	11609
2015	10760

only showing top 10 rows

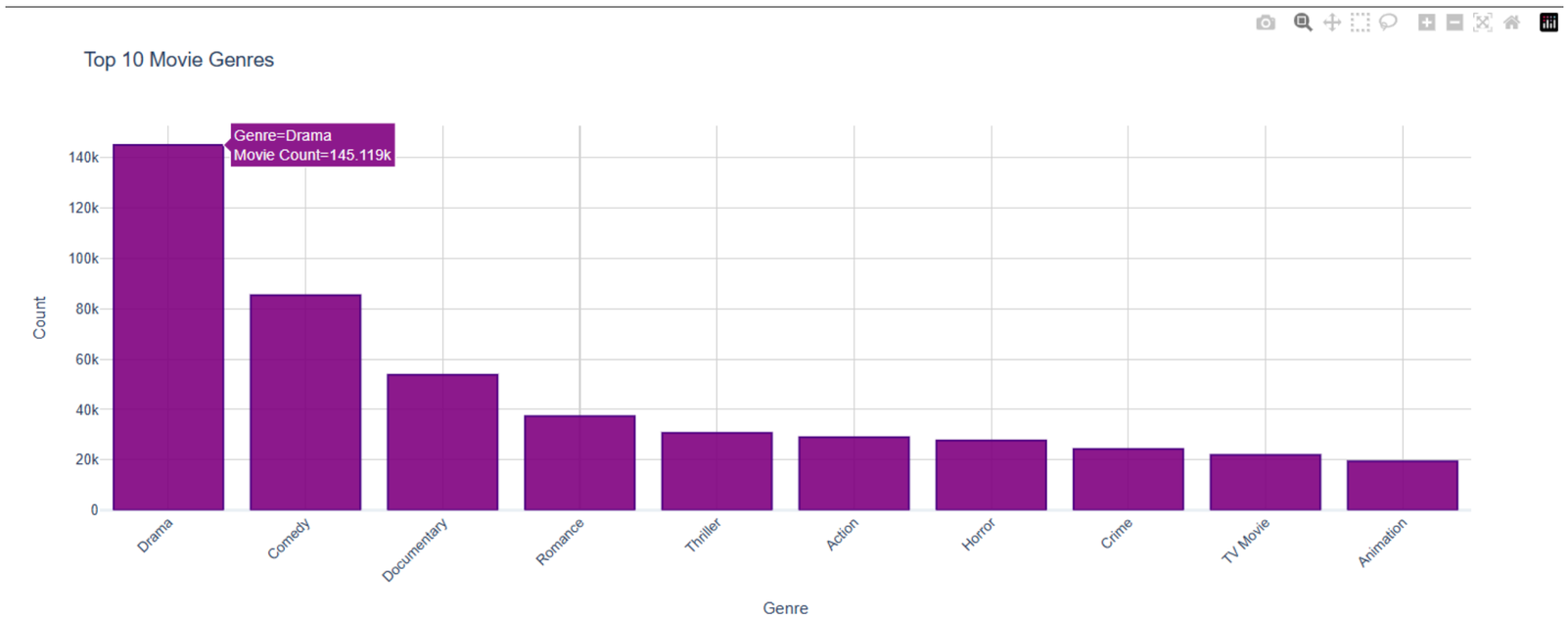


VISUALIZATIONS

Top Genres by Count

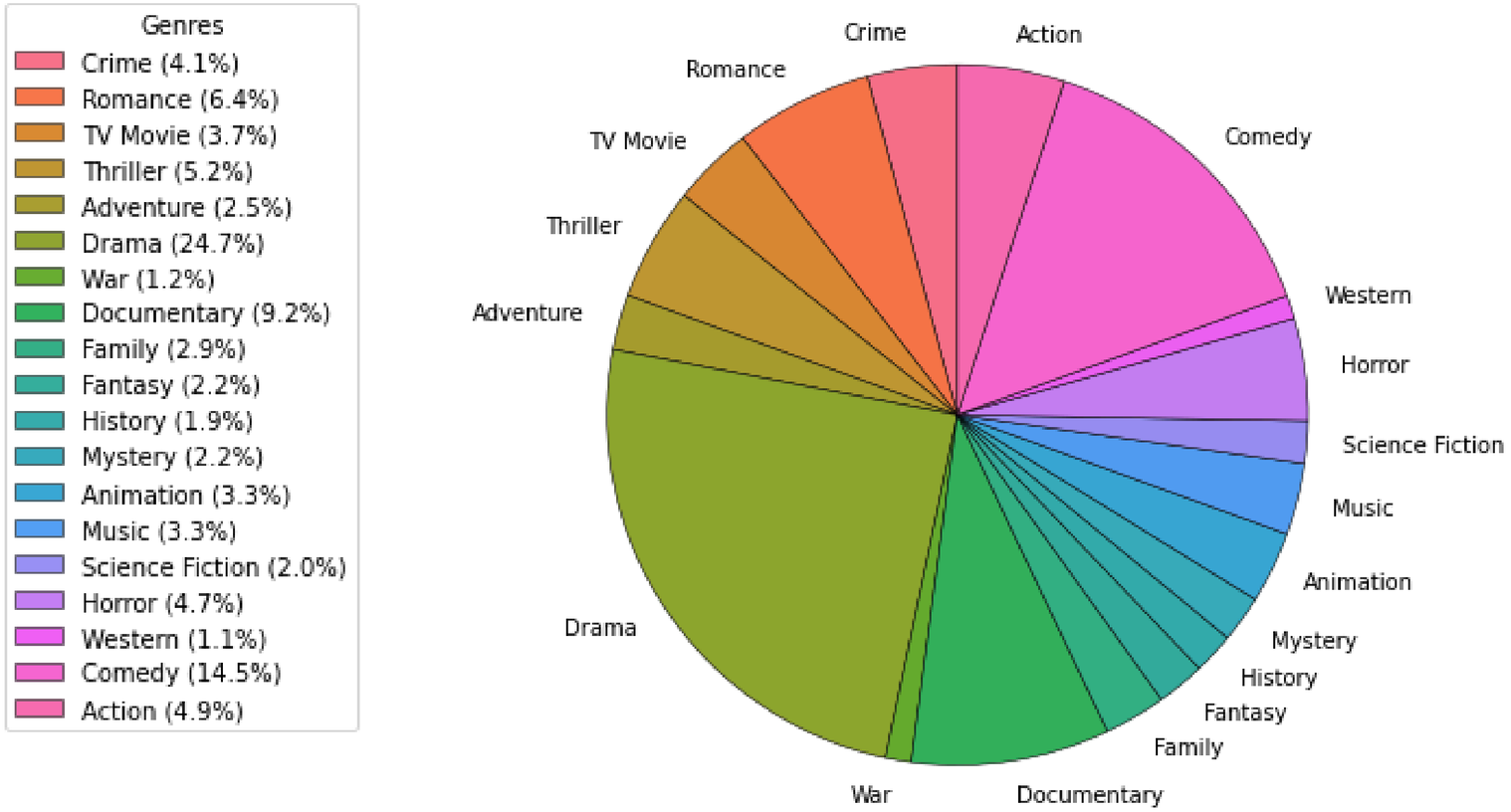
genre	total
Drama	145119
Comedy	85413
Documentary	53778
Romance	37342
Thriller	30707
Action	28979
Horror	27692
Crime	24294
TV Movie	21933
Animation	19409

only showing top 10 rows



VISUALIZATIONS

Genre Proportions by Movie Count

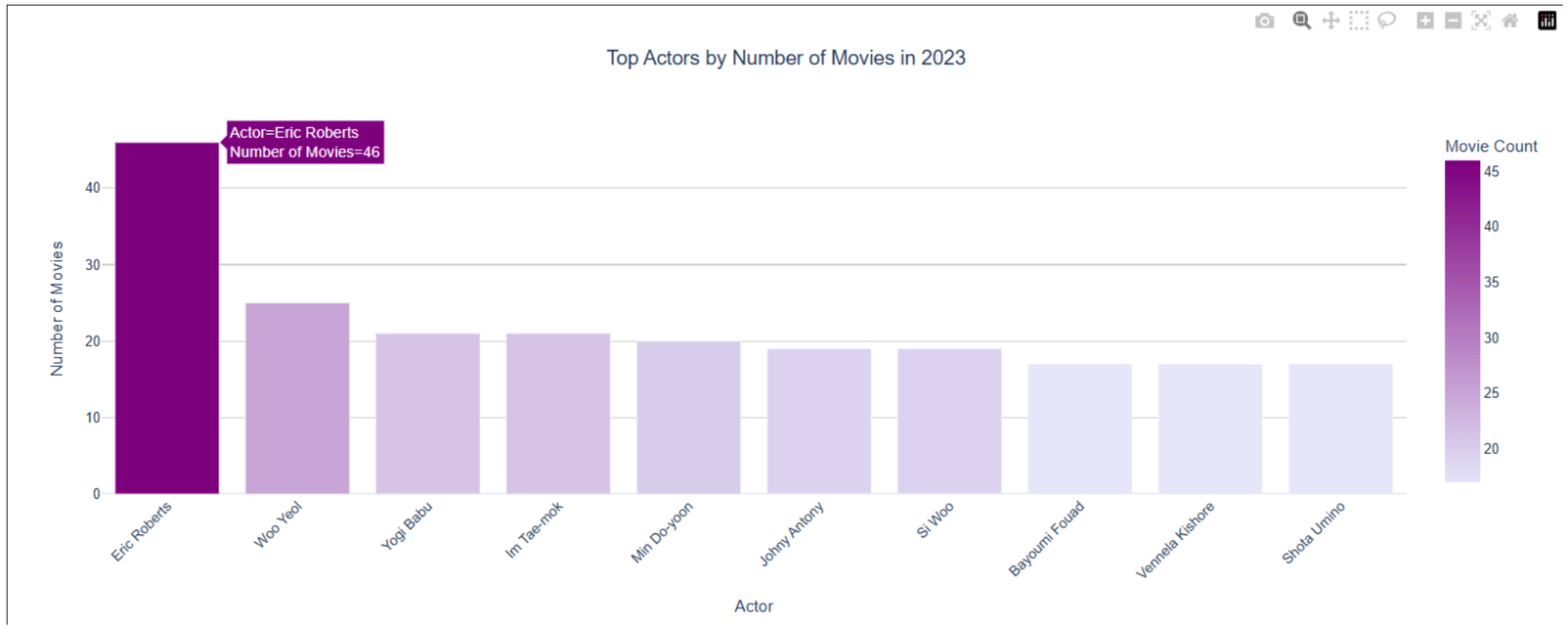


VISUALIZATIONS

which actor appeared in the most movies in 2023

cast_member	movie_count
Eric Roberts	46
Woo Yeol	25
Yogi Babu	21
Im Tae-mok	21
Min Do-yoon	20
Johny Antony	19
Si Woo	19
Bayoumi Fouad	17
Shawn C. Phillips	17
Shota Umino	17

only showing top 10 rows

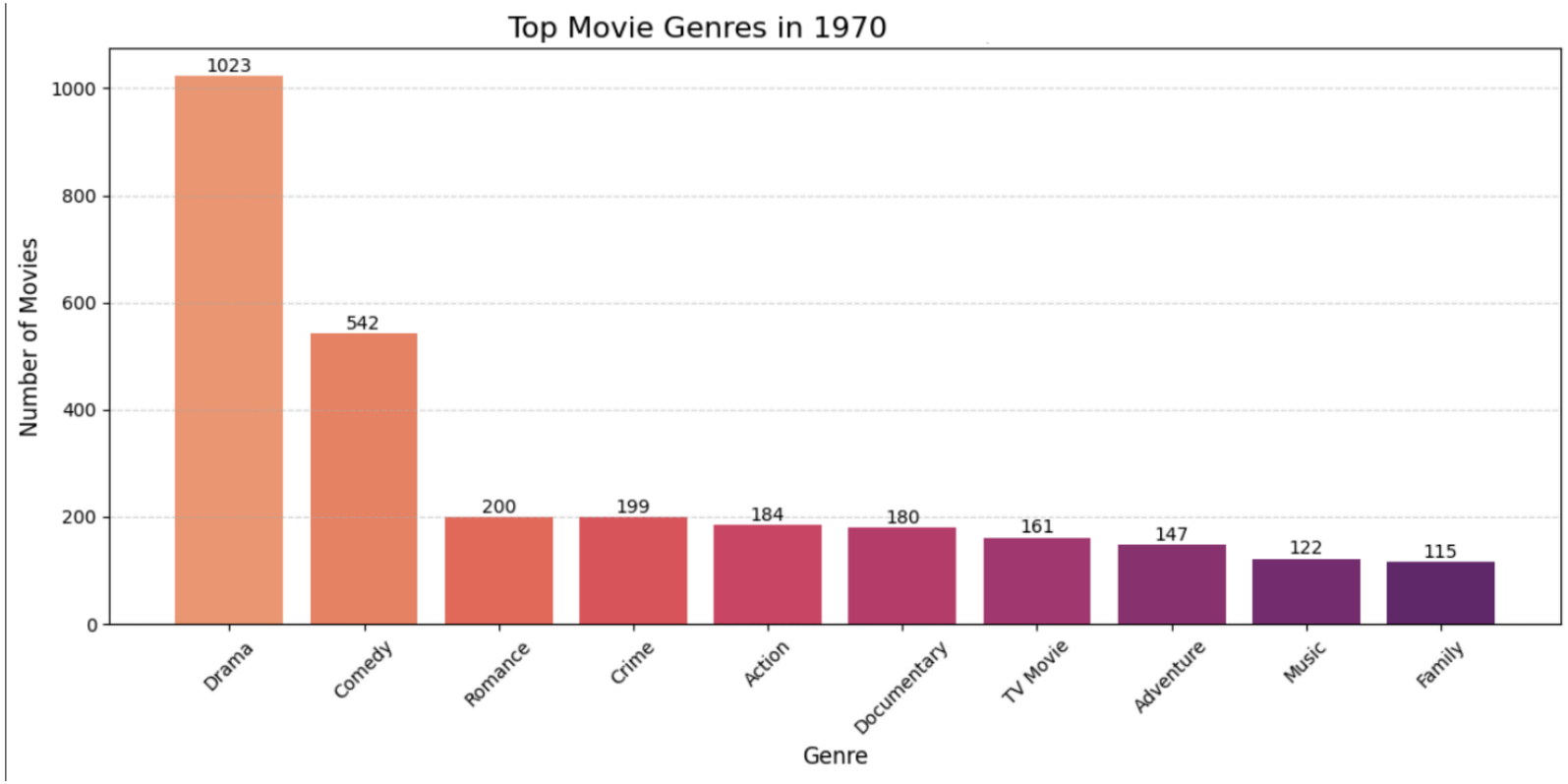


VISUALIZATIONS

Genre Popularity Over Time

release_year	genre	movie_count
1970	Action	184
1970	Adventure	147
1970	Animation	95
1970	Comedy	542
1970	Crime	199
1970	Documentary	180
1970	Drama	1023
1970	Family	115
1970	Fantasy	68
1970	History	72
1970	Horror	108
1970	Music	122
1970	Mystery	59
1970	Romance	200
1970	Science Fiction	41
1970	TV Movie	161
1970	Thriller	111
1970	War	89
1970	Western	72
1971	Action	214

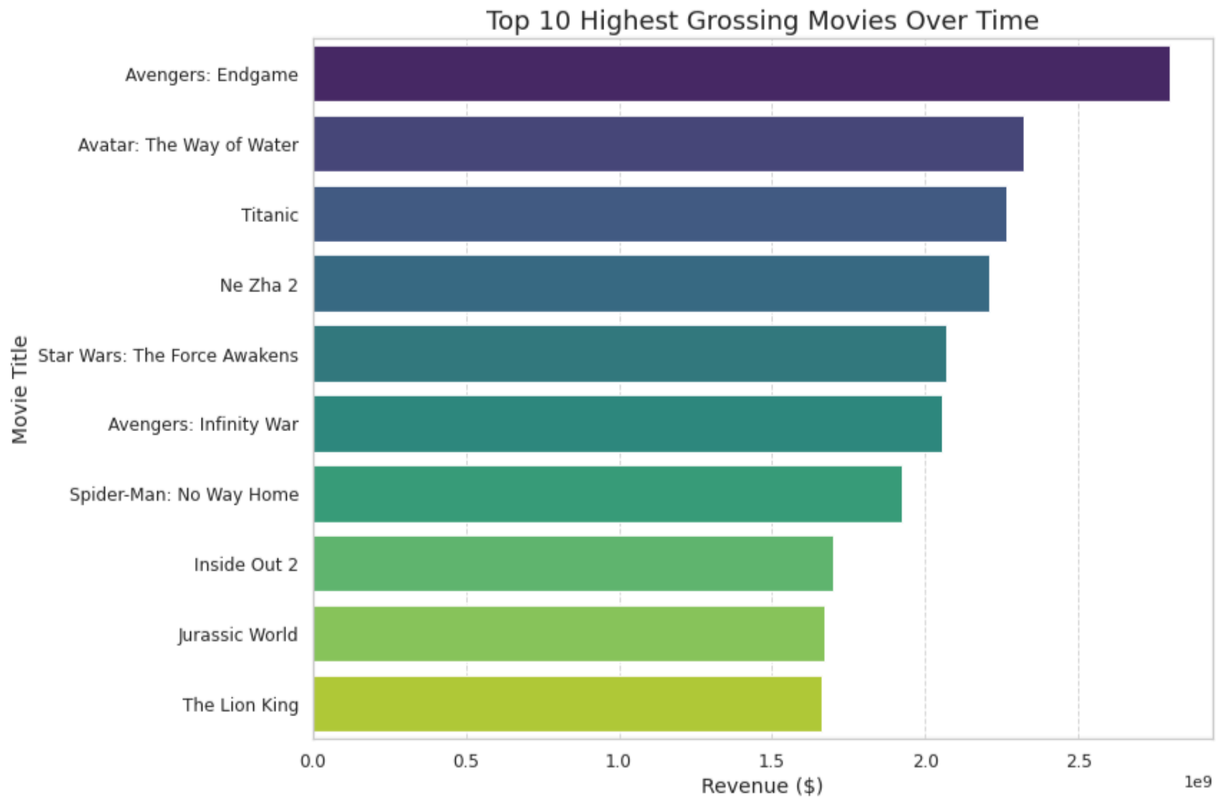
only showing top 20 rows



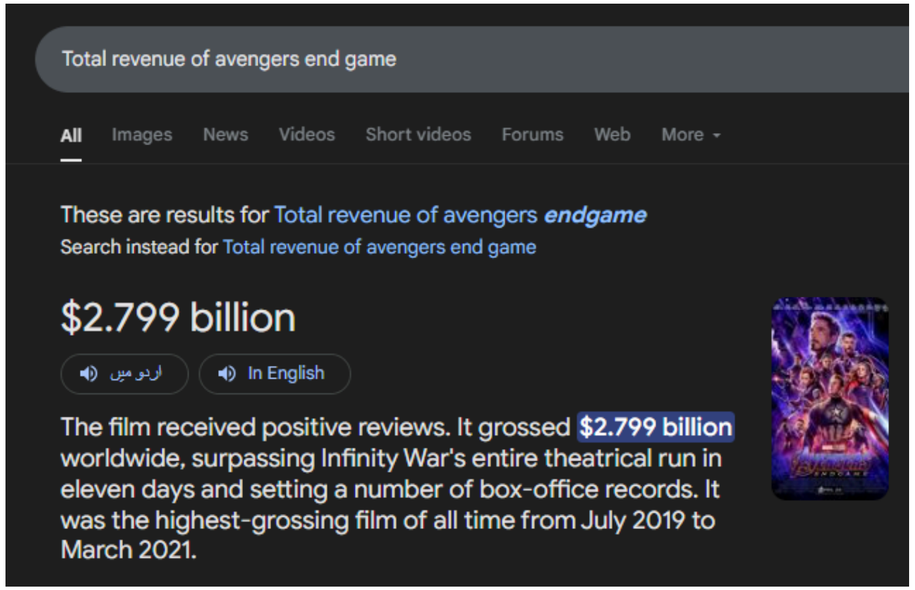
Note: In the code, while plotting, you have the option to select and plot the data for any year you want.

VISUALIZATIONS

Top movies by revenue over time




```
+--------------------+------------+-------------+
|               title|release_year|total_revenue|
+--------------------+------------+-------------+
|   Avengers: Endgame|        2019|   2799439100|
|Avatar: The Way o...|        2022|   2320250281|
|             Titanic|        1997|   2264162353|
|           Ne Zha 2|        2025|   2208800000|
|Star Wars: The Fo...|        2015|   2068223624|
|Avengers: Infinit...|        2018|   2052415039|
|Spider-Man: No Wa...|        2021|   1921847111|
|        Inside Out 2|        2024|   1698863816|
|      Jurassic World|        2015|   1671537444|
|       The Lion King|        2019|   1662020819|
+--------------------+------------+-------------+
only showing top 10 rows
```



VISUALIZATIONS

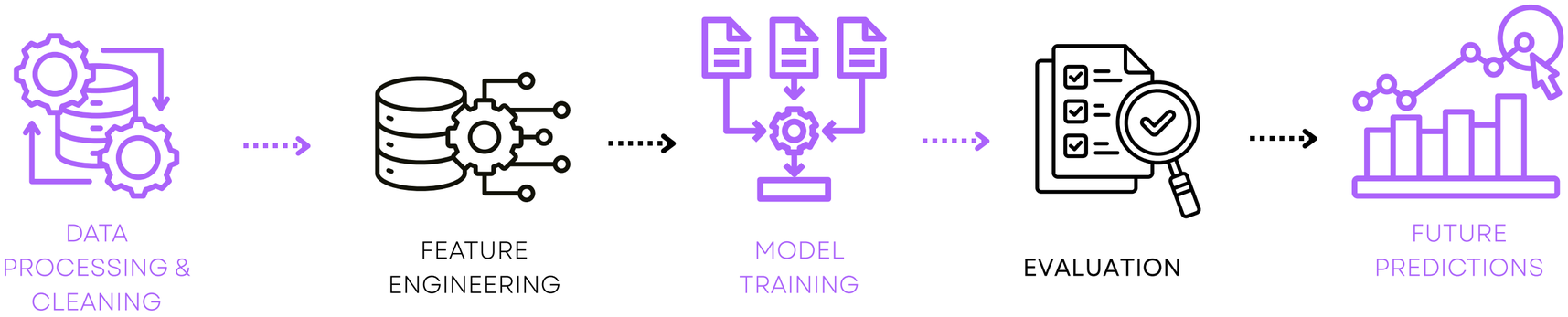
Budget vs. ROI (Big Data Stats)

	title	budget	revenue	roi	release_year
0	Titanic	$200,000,000	$2,264,162,353	1,032%	1997
1	Jurassic World	$150,000,000	$1,671,537,444	1,014%	2015
2	Barbie	$145,000,000	$1,445,638,421	897%	2023
3	Frozen II	$150,000,000	$1,453,683,476	869%	2019
4	Spider-Man: No Way Home	$200,000,000	$1,921,847,111	861%	2021
5	Top Gun: Maverick	$170,000,000	$1,488,732,821	776%	2022
6	Frozen	$150,000,000	$1,274,219,009	749%	2013
7	Inside Out 2	$200,000,000	$1,698,863,816	749%	2024
8	Star Wars: The Force Awakens	$245,000,000	$2,068,223,624	744%	2015
9	Star Wars: Episode I - The Phantom Menace	$115,000,000	$924,317,558	704%	1999

TOP 10 MOVIES BY ROI WITH BUDGET > 100M

GENRE POPULARITY PREDICTION PIPELINE

ML Pipeline



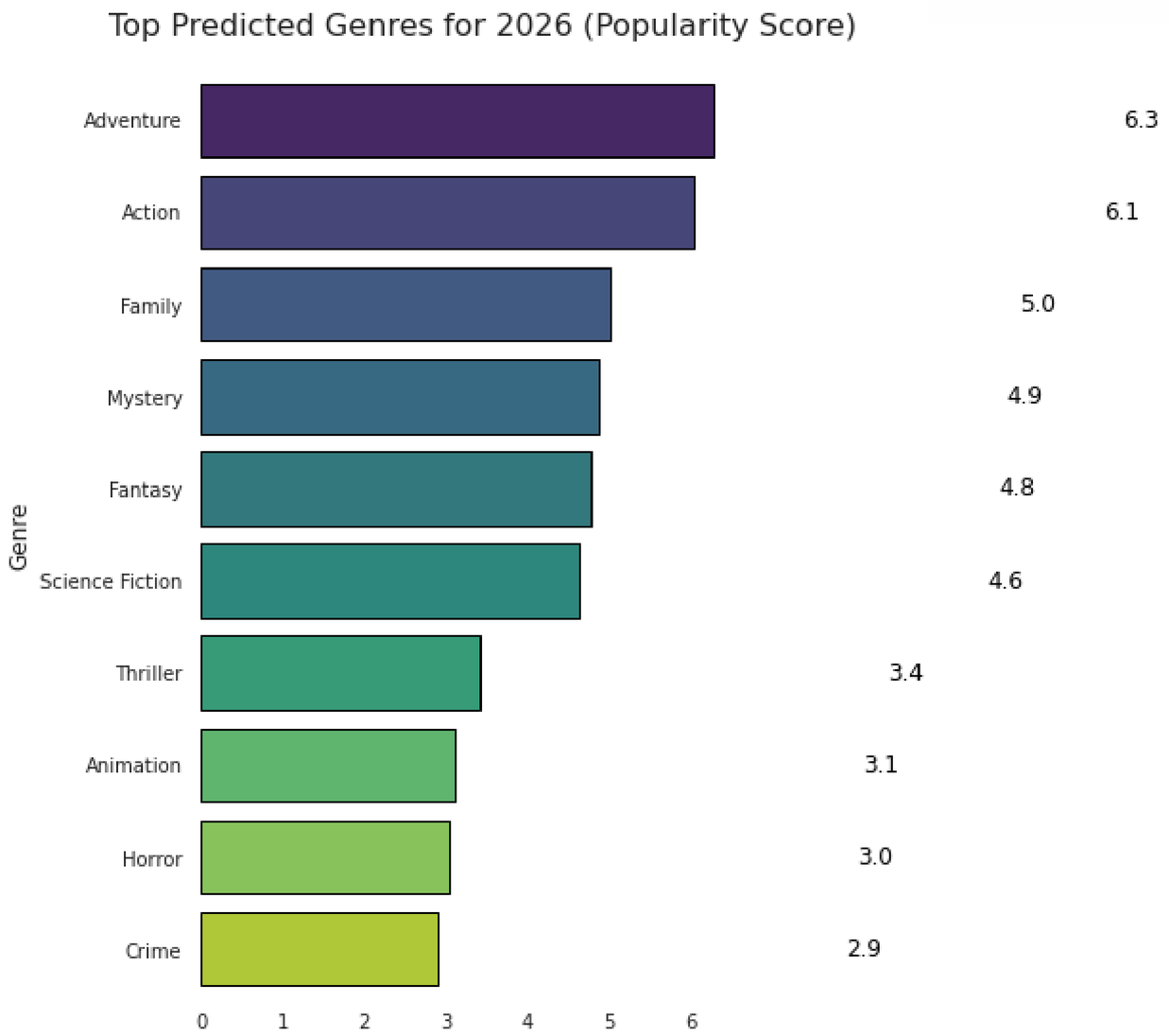
WHAT WE DID:

- **DATA CLEANING:** PROCESSED ~400K TMDB RECORDS—PARSED DATES, FILTERED YEARS, AND EXPLODED GENRES INTO ROWS.
- **FEATURE ENGINEERING (PYSPARK ML):** ENCODED GENRES (STRINGINDEXER) AND MERGED FEATURES (VECTORASSEMBLER) FOR ML INPUT.
- **MODEL TRAINING (SPARK MLLIB):** TRAINED A RANDOM FOREST MODEL (50 TREES) ON GENRE/YEAR DATA, VALIDATED WITH RMSE.
- **PREDICTIONS & VIZ (PANDAS/SEABORN):** PREDICTED 2026 GENRE POPULARITY AND VISUALIZED THE TOP 10 GENRES WITH STYLED BAR PLOTS.

RESULTS

MODEL RMSE = 3.63

"OUR OBJECTIVE WAS TO LEVERAGE BIG DATA TECHNOLOGIES TO ANALYZE TMDB MOVIE RECORDS, IDENTIFY TRENDS, AND BUILD A PREDICTIVE MODEL FORECASTING POPULAR GENRES FOR 2026. WITH THESE RESULTS, WE'VE SUCCESSFULLY EXECUTED THE END-TO-END PIPELINE—FROM DATA PROCESSING TO ACTIONABLE PREDICTIONS."



THANK YOU FOR YOUR TIME!

ITMO more than a UNIVERSITY