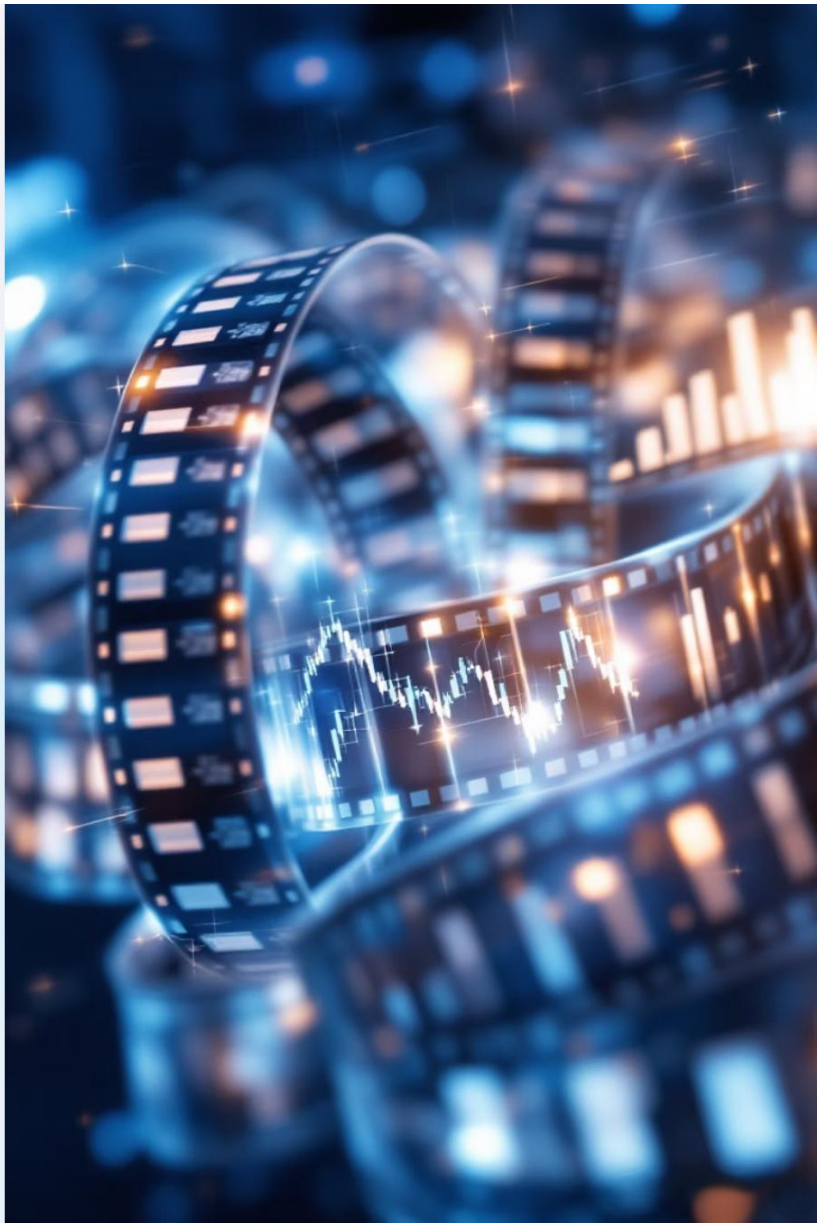
Film Industry Analysis with LSTM

CO3029 - Data Mining



Team member


1
Huỳnh Huy Mân
2
Đoàn Nhật Tiến
3
Bùi Thanh Tùng
4
Trần Nguyễn Anh Khoa

Project Overview

Industry Evolution

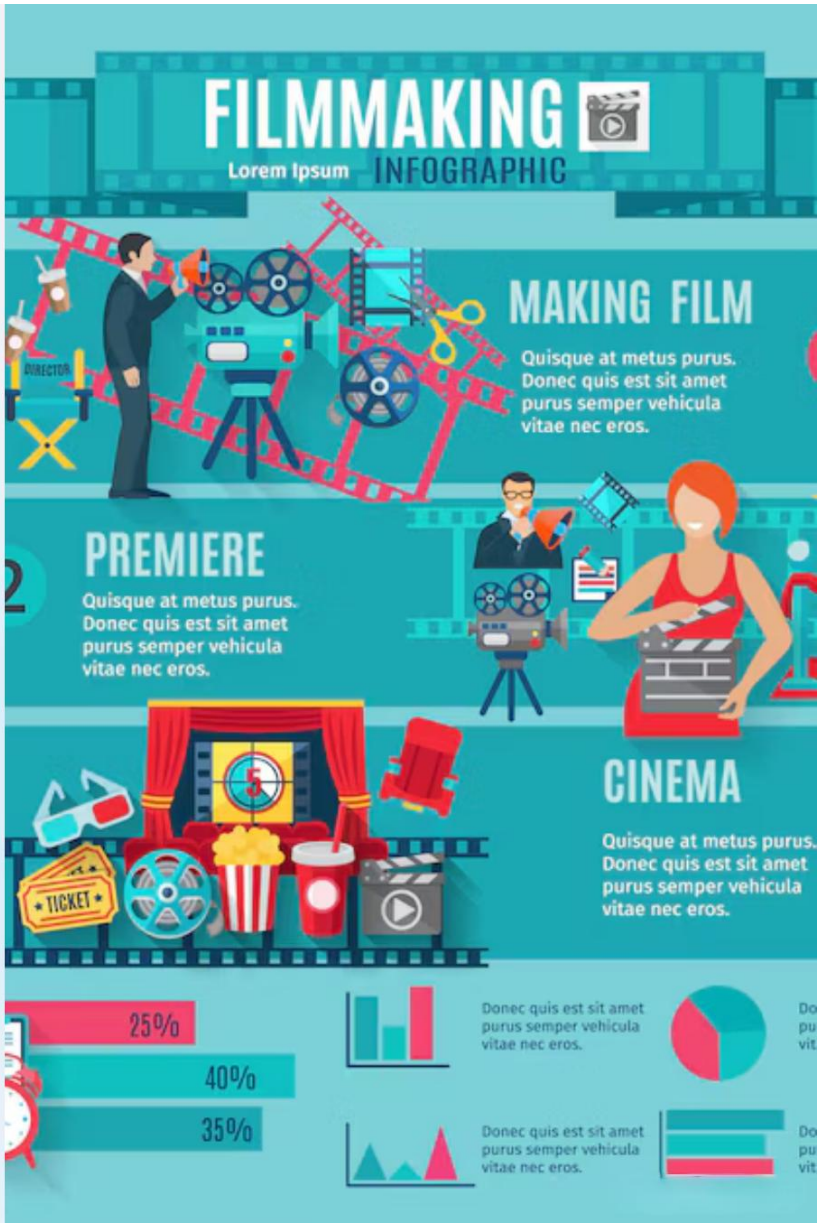
From silent films to digital streaming platforms

Data-Driven Insights

Using TMDB dataset to analyze patterns

LSTM Application

Predicting future trends with neural networks



LSTM Architecture

1

Recurrent Neural Networks

Process sequential data but struggle with long dependencies

2

LSTM Solution

Special gates control information flow

3

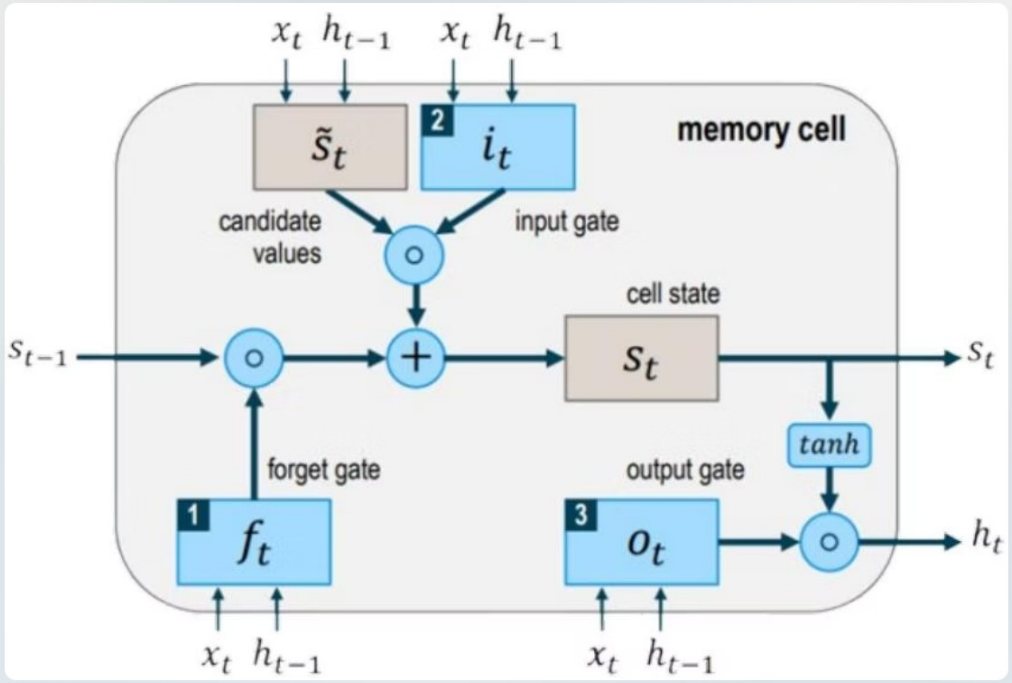
Memory Cells

Forget, input, and output gates manage data retention

4

Time Series Analysis

Perfect for year-by-year film data patterns



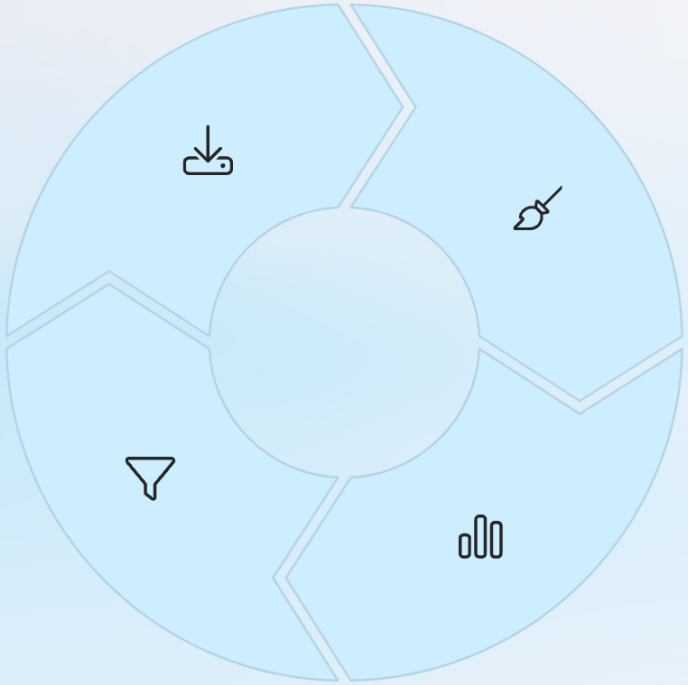
Data Processing Workflow

Data Loading

Import CSV files from TMDB database

Data Cleaning

Remove duplicates, handle missing values



Feature Engineering

Extract year/month, group by categories

Exploratory Analysis

Visualize distributions and relationships

#	Column	Non-Null Count	Dtype
0	id	1208039 non-null	int64
1	title	1208026 non-null	object
2	vote_average	1208039 non-null	float64
3	vote_count	1208039 non-null	int64
4	status	1208039 non-null	object
5	release_date	990151 non-null	object
6	revenue	1208039 non-null	int64
7	runtime	1208039 non-null	int64
8	adult	1208039 non-null	bool
9	backdrop_path	315175 non-null	object
10	budget	1208039 non-null	int64
11	homepage	126977 non-null	object
12	imdb_id	619329 non-null	object
13	original_language	1208039 non-null	object
14	original_title	1208026 non-null	object
15	overview	953685 non-null	object
16	popularity	1208039 non-null	float64
17	poster_path	813313 non-null	object
18	tagline	169301 non-null	object
19	genres	710289 non-null	object
20	production_companies	537058 non-null	object
21	production_countries	658056 non-null	object
22	spoken_languages	679316 non-null	object
23	keywords	318649 non-null	object

Data Shape

- **Rows:** ~1.2 million
- **Columns:** 24
- Includes movie info like title, vote_average, budget, revenue, release_date, etc.
- Some columns have missing values (e.g. homepage, overview, genres)
- Data types: integers, floats, strings, and boolean

Remove unnecessary columns

'backdrop_path', 'homepage', 'imdb_id', 'original_language', 'original_title', 'overview', 'popularity', 'poster_path', 'title' 'production_companies', 'production_countries', 'spoken_languages', 'keywords', 'tagline',

Remove duplicated title

id	title	vote_avera	vote_count	status	release_dat	revenue	runtime	adult	backdrop_path	budget	homepage	imdb_id	origin	original_title
27205	Inception	8.364	34495	Released	7/15/2010	825532764	148	FALSE	/8ZTVqvKDQ8emSGUEMjsS4yHAwrp.jpg	160000000	https://www.warne	tt1375666	en	Inception
157336	Interstellar	8.417	32571	Released	11/5/2014	701729206	169	FALSE	/pbrkL804c8yAv3zBZR4QPEafpAR.jpg	165000000	http://www.interste	tt0816692	en	Interstellar
155	The Dark Knight	8.512	30619	Released	7/16/2008	1.005E+09	152	FALSE	/nMKdUUepR0i5zn0y1T4CsSB5chy.jpg	185000000	https://www.warne	tt0468569	en	The Dark Knigh
19995	Avatar	7.573	29815	Released	12/15/2009	2.924E+09	162	FALSE	/vL5LR6WdxWPjLPFRLe133jXWsh5.jpg	237000000	https://www.avatar	tt0499549	en	Avatar
24428	The Avengers	7.71	29166	Released	4/25/2012	1.519E+09	143	FALSE	/9BBTo63ANSmhC4e6r62OJFuK2GL.jpg	220000000	https://www.marve	tt0848228	en	The Avengers
293660	Deadpool	7.606	28894	Released	2/9/2016	783100000	108	FALSE	/en971MEXui9diirXlogOrPKmsEn.jpg	58000000	https://www.20thce	tt1431045	en	Deadpool
299536	Avengers: Infinity	8.255	27713	Released	4/25/2018	2.052E+09	149	FALSE	/mDfJG3LC3Dqb67AZ52x3Z0jU0uB.jpg	300000000	https://www.marve	tt4154756	en	Avengers: Infin
550	Fight Club	8.438	27238	Released	10/15/1999	100853753	139	FALSE	/hZkgoQYus5vegHoetLkCJzb17zJ.jpg	63000000	http://www.foxmov	tt0137523	en	Fight Club
118340	Guardians of the	7.906	26638	Released	7/30/2014	772776600	121	FALSE	/uLtVbjvS1O7gXL8lUOwsFOH4man.jpg	170000000	http://marvel.com/	tt2015381	en	Guardians of t
680	Pulp Fiction	8.488	25893	Released	9/10/1994	213900000	154	FALSE	/suaEOtk1N1sgg2MTM7oZd2cfVp3.jpg	8500000	https://www.miram	tt0110912	en	Pulp Fiction
13	Forrest Gump	8.477	25409	Released	6/23/1994	677387716	142	FALSE	/qdIMHd4sEfJSckfVJfKQvisL02a.jpg	55000000	https://www.param	tt0109830	en	Forrest Gump
671	Harry Potter and t	7.916	25379	Released	11/16/2001	976475550	152	FALSE	/hziiv14OpD73u9gAak4XDDfBKa2.jpg	125000000	https://www.warne	tt0241527	en	Harry Potter a
1726	Iron Man	7.64	24874	Released	4/30/2008	585174222	126	FALSE	/cyecB7godJ6kNHGONFjUyVN9OX5.jpg	140000000	https://www.marve	tt0371746	en	Iron Man
68718	Django Unchainec	8.171	24672	Released	12/25/2012	425368238	165	FALSE	/5Lbm0gpFDRAPlV1Cth6ln9iL1ou.jpg	100000000	http://www.unchair	tt1853728	en	Django Unchai
278	The Shawshank R	8.702	24649	Released	9/23/1994	28341469	142	FALSE	/kXfqcdQKsToO0OUXHcrrNCHDBzO.jpg	25000000		tt0111161	en	The Shawshan
299534	Avengers: Endgam	8.263	23857	Released	4/24/2019	2.8E+09	181	FALSE	/7RyHsO4yDXtBv1zUU3mTpHeQ0d5.jpg	356000000	https://www.marve	tt4154796	en	Avengers: End
603	The Matrix	8.206	23815	Released	3/30/1999	463517383	136	FALSE	/oMsxZEvz9a708d49b6UdZK1KAo5.jpg	63000000	http://www.warner	tt0133093	en	The Matrix
597	Titanic	7.9	23637	Released	11/18/1997	2.264E+09	194	FALSE	/rzdPqYx7Um4FUZeD8wpXqjAUcEm.jpg	200000000	https://www.param	tt0120338	en	Titanic
475557	Joker	8.168	23425	Released	10/1/2019	1.074E+09	122	FALSE	/hO7KbdvGOtDdeg0W4Y5nKEHeDDh.jpg	55000000	http://www.jokerm	tt7286456	en	Joker
120	The Lord of the Ri	8.402	23323	Released	12/18/2001	871368364	179	FALSE	/x2RS3uTcsJJ9IfjNPcgDmukoEcQ.jpg	93000000	http://www.lordoftl	tt0120737	en	The Lord of th
122	The Lord of the Ri	8.474	22334	Released	12/1/2003	1.119E+09	201	FALSE	/2u7zbn8EudG6kLlBzUYqP8RyFU4.jpg	94000000	http://www.lordoftl	tt0167260	en	The Lord of th
11324	Shutter Island	8.2	22318	Released	2/14/2010	294800000	138	FALSE	/2nqsOT2AqPkTW81bWaLRtjgjqVM.jpg	80000000	http://www.shutter	tt1130884	en	Shutter Island
106646	The Wolf of Wall	8.035	22222	Released	12/25/2013	392000000	180	FALSE	/63y4XSVTZ7mRzAzkqwi3o0ajDZZ.jpg	100000000	http://www.thewolf	tt0993846	en	The Wolf of W
99861	Avengers: Age of I	7.276	21754	Released	4/22/2015	1.405E+09	141	FALSE	/6YwkGolwdOMNpbTOmLjoehlVWs5.jpg	365000000	http://marvel.com/	tt2395427	en	Avengers: Age
271110	Captain America:	7.4	21541	Released	4/27/2016	1.155E+09	147	FALSE	/wdwcOBMkt3zmPQuEMxB3FUtMio2.jpg	250000000	https://www.marve	tt3498820	en	Captain Ameri
49026	The Dark Knight R	7.777	21335	Released	7/17/2012	1.081E+09	165	FALSE	/c3OHQncTAnKFhdOTX7D3LTW6son.jpg	250000000	http://www.thedark	tt1345836	en	The Dark Knigh
68721	Iron Man 3	6.928	21064	Released	4/18/2013	1.216E+09	130	FALSE	/aFTYFqrWp4RS46Twm87I5e0ItYb.jpg	200000000	https://www.marve	tt1300854	en	Iron Man 3

-Số dòng riêng biệt:

title	
Home	166
Untitled	129
Mother	106
Alone	104
The Gift	84
	...
Khajuraho	2
Horny Neighbours	2
Bröllopet på Ulfåsa	2
The Long Ride Home	2
Tickets Please	2

#	Column	Non-Null Count	Dtype
0	id	1033579 non-null	int64
1	vote_average	1033579 non-null	float64
2	vote_count	1033579 non-null	int64
3	status	1033579 non-null	object
4	release_date	849115 non-null	object
5	revenue	1033579 non-null	int64
6	runtime	1033579 non-null	int64
7	adult	1033579 non-null	bool
8	budget	1033579 non-null	int64
9	genres	604029 non-null	object

Preprocessing data

1. Xử lý giá trị thiếu trong các cột như status và genres.

2. Chuyển đổi cột release_date thành kiểu datetime.

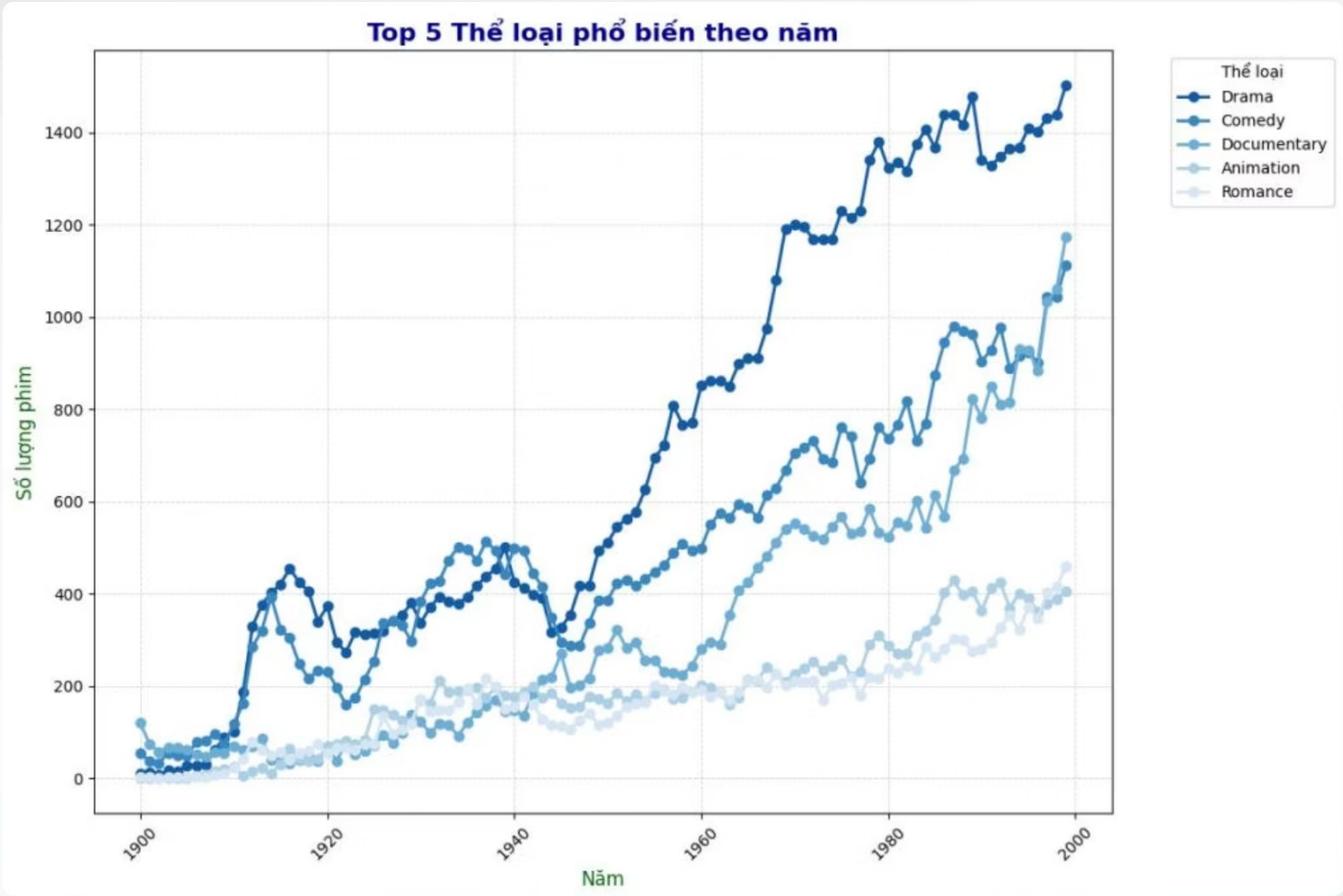
3. Tách cột genres thành danh sách (list).

4. Thêm các cột year và month từ release_date.

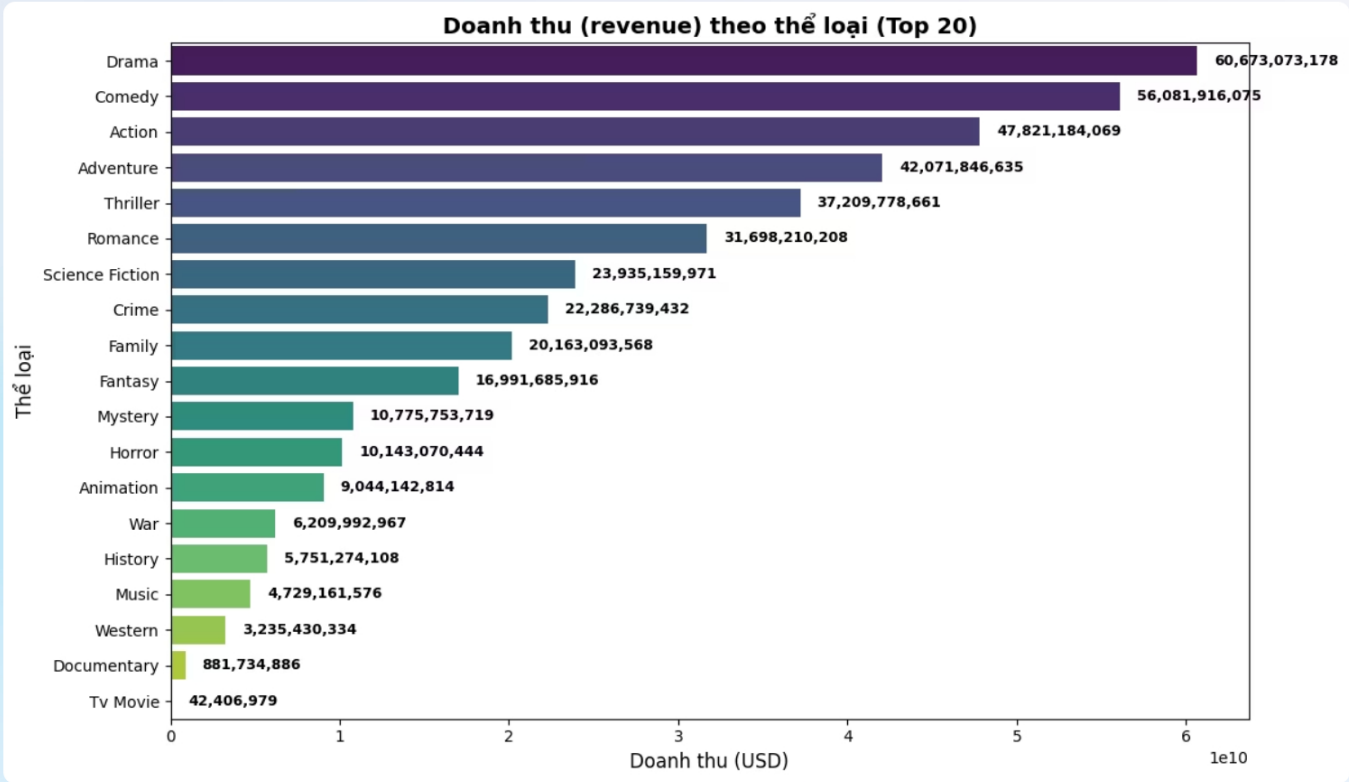
5. Giữ lại các bộ phim có năm phát hành trong thế kỷ 20 (1900-1999)

	id	vote_average	vote_count	status	release_date	revenue	runtime	adult	budget	genres
0	27205	8.364	34495	Released	2010-07-15	825532764	148	False	160000000	Action, Science Fiction, Adventure
1	157336	8.417	32571	Released	2014-11-05	701729206	169	False	165000000	Adventure, Drama, Science Fiction
2	155	8.512	30619	Released	2008-07-16	1004558444	152	False	185000000	Drama, Action, Crime, Thriller
3	19995	7.573	29815	Released	2009-12-15	2923706026	162	False	237000000	Action, Adventure, Fantasy, Science Fiction
4	24428	7.710	29166	Released	2012-04-25	1518815515	143	False	220000000	Science Fiction, Action, Adventure

Film Genre Analysis



Total Revenue by Genre



Drama and Comedy have high total revenue, but low profit per film.

Sci-fi and Adventure top the chart in profitability.

Steady Growth

Number of films released increased steadily.

Break - through

A significant spike in film releases around 1912.

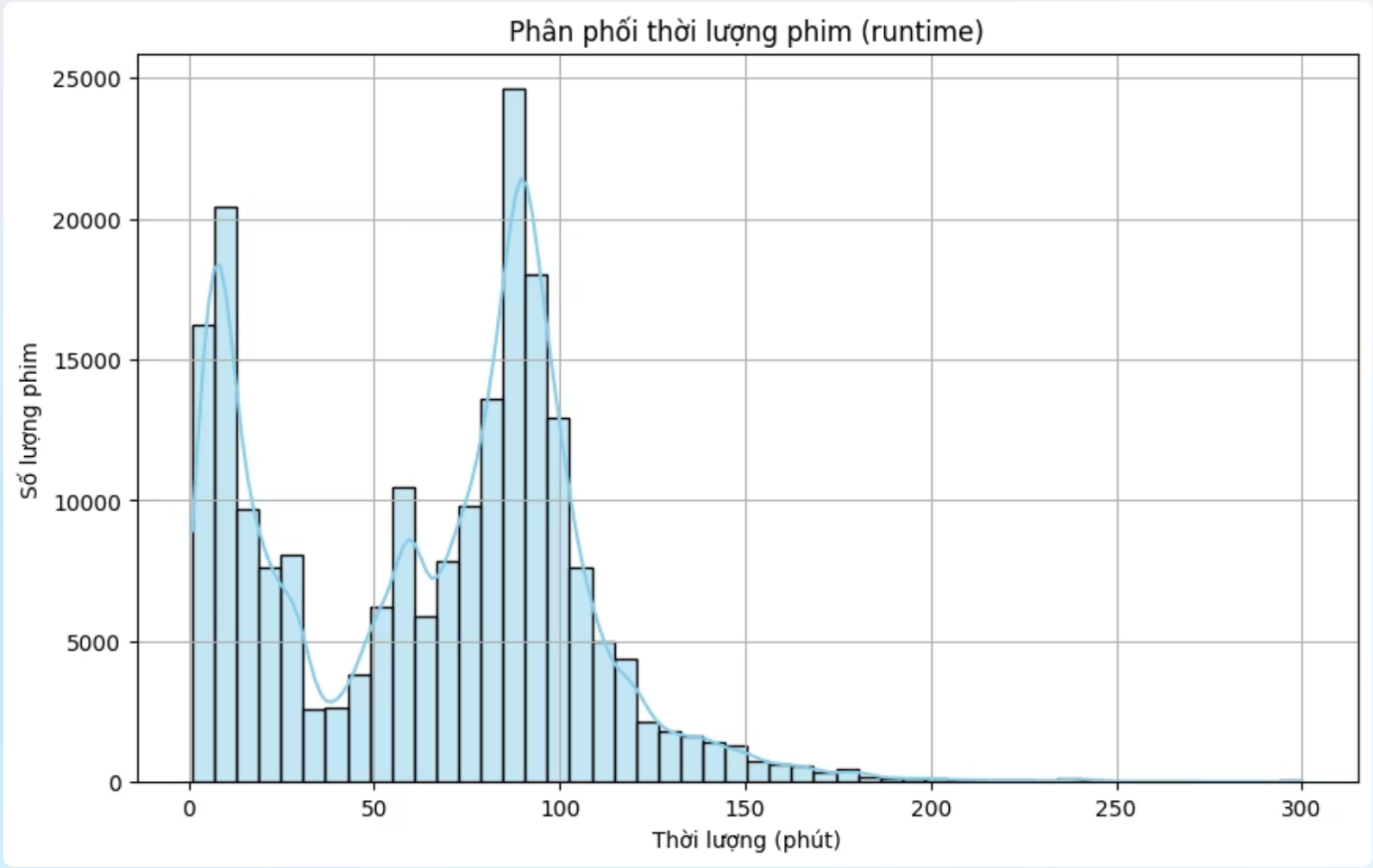
Decline

Slowdown due to World War I (WWI) and a sharp drop during World War II (WWII).

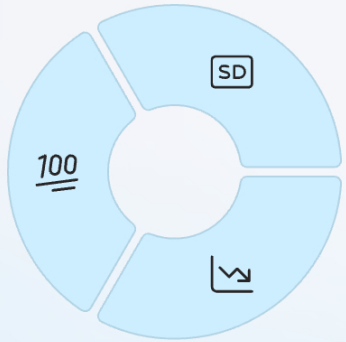
Recovery and Growth

Post-WWII, the number of films released increased steadily again.

Film Releases Over Time



Typical Runtime:



– Mostly short clips (a few mins) or full films

– 15–100 min range is becoming "standard"

– Longer films are rare and scattered

Runtime Growth Over Time

Steady Growth

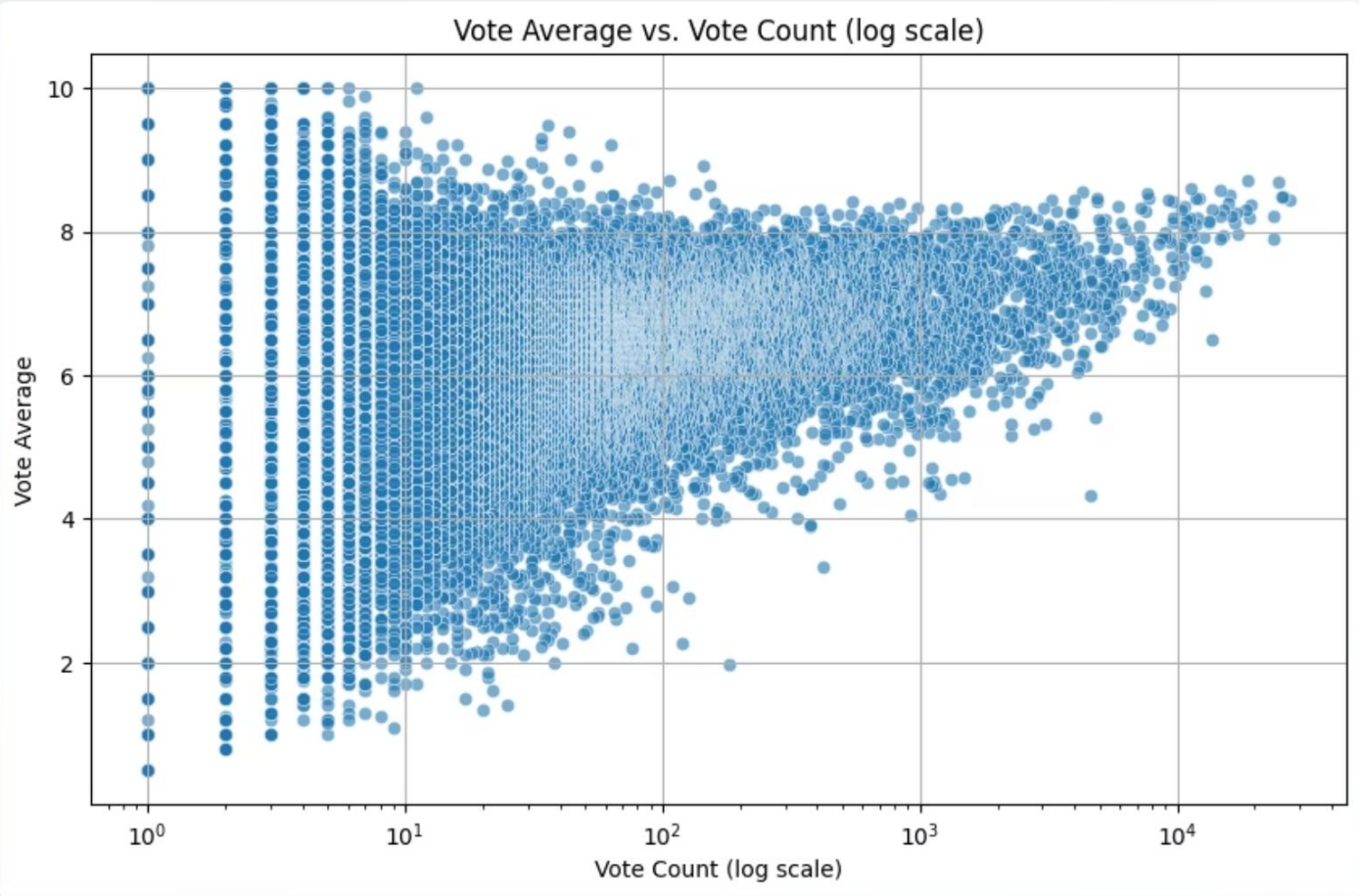
Increase reaching around 80 minutes

Evolving Trends

Reflects changing audience expectations

Early Leap

A big jump occurred in the 1910s



– Vote Count:

Higher counts lead to more accurate results

Vote Average:

– Typically falls between 6–7

– As the count increases, the rating becomes more reliable and converges toward a more accurate value.

Monthly Distribution

The peak release season is mainly in January. What is the reason for this?

Removed January 1st

Financial Trends

1970s

Growth Begins

Budgets and revenues start upward trend

1980s

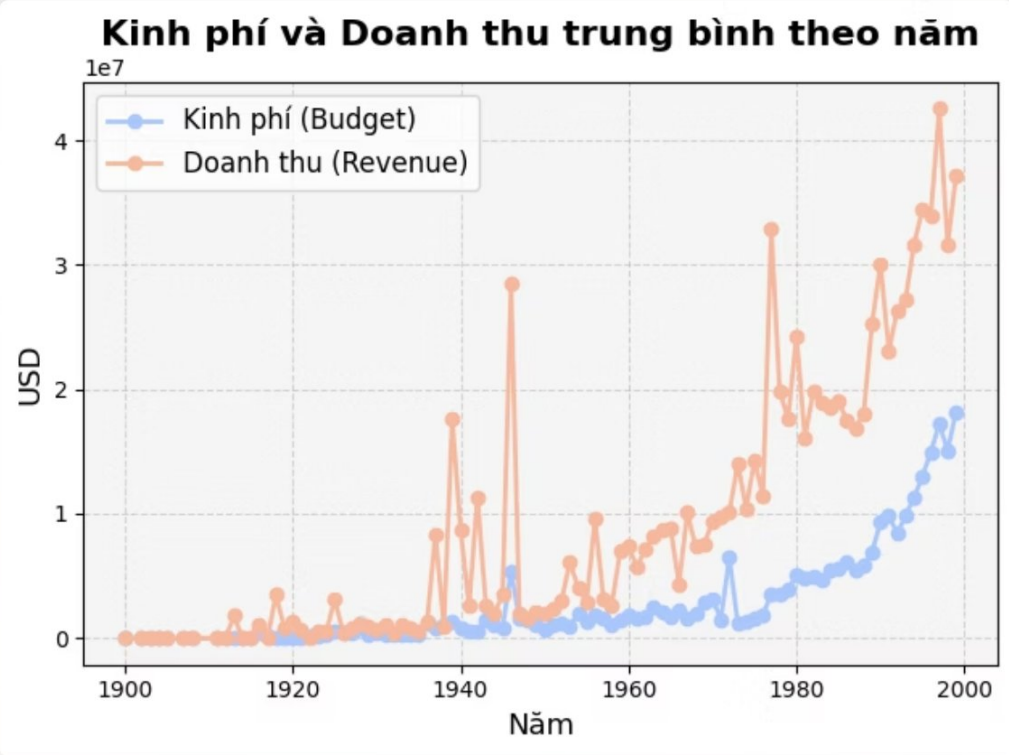
Blockbuster Era

Major increase in production investments

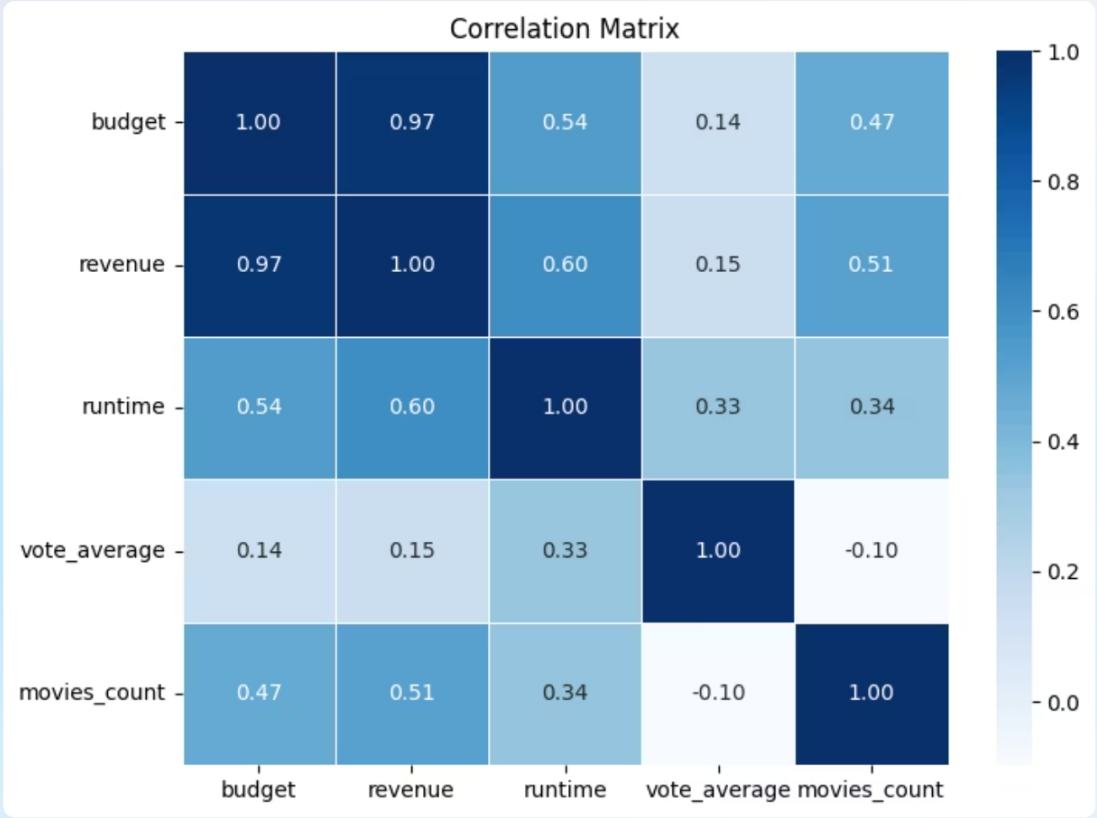
1990s

Peak Growth

Highest revenue expansion in 20th century



Feature and model




```
Input (look_back=3, features=3)
        |
  ┌──────────┐
  |   LSTM   | (64 units, tanh)
  └──────────┘
        |
  ┌──────────┐
  | Dropout  | (rate=0.2)
  └──────────┘
        |
  ┌──────────┐
  | Dense    | (32 units, ReLU)
  └──────────┘
        |
  ┌──────────┐
  | Dense    | (1 unit - Linear output)
  └──────────┘
```

Training model and validation

1 Optimizer

- Adam (lr=0.001),
- Loss: MSE,
- Metric: MAE

2 Callbacks

- EarlyStopping (patience=8),
- ReduceLROnPlateau (factor=0.2, min_lr=1e-5)

3 Early stopping

- At epoch 38
- Best val_loss: 0.1015 (epoch 30),
- VVal_MAE ≈ 0.2812
- Last learning rate: 4e−5

4 Evaluate results

- MSE: 5.02e12
- MAE: 1.8e6
- R^2: -62.92

LSTM Prediction Results

Training Data

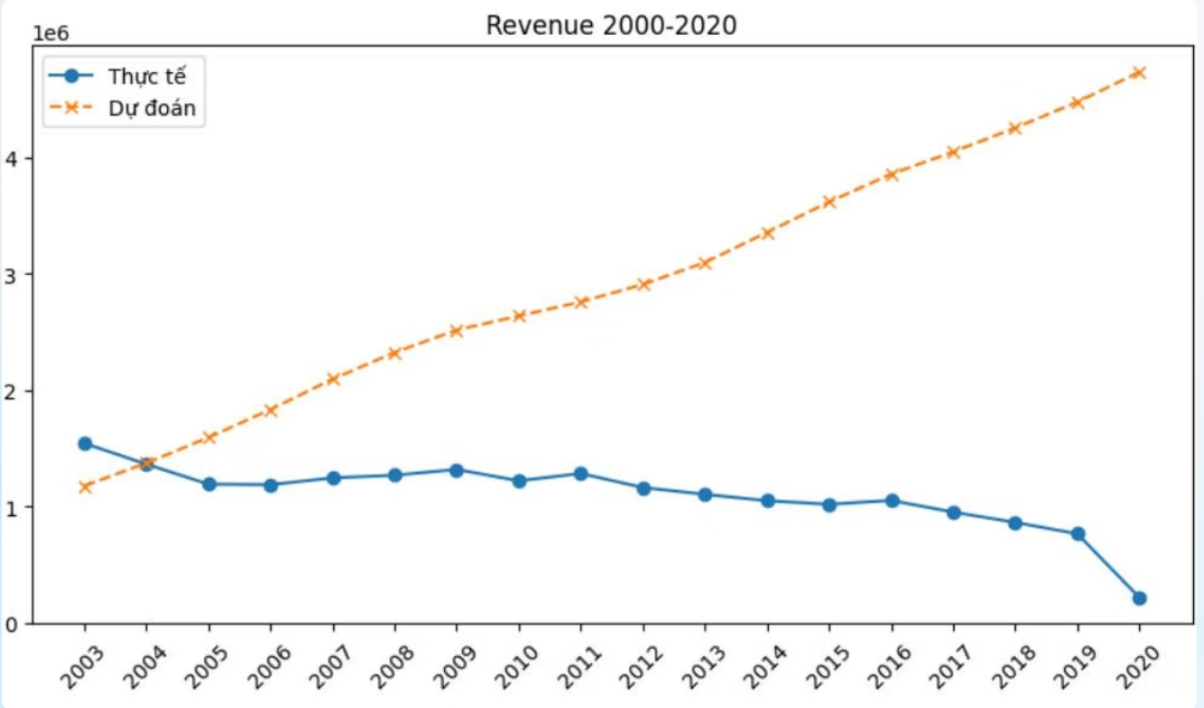
20th century film data (1900-1999)

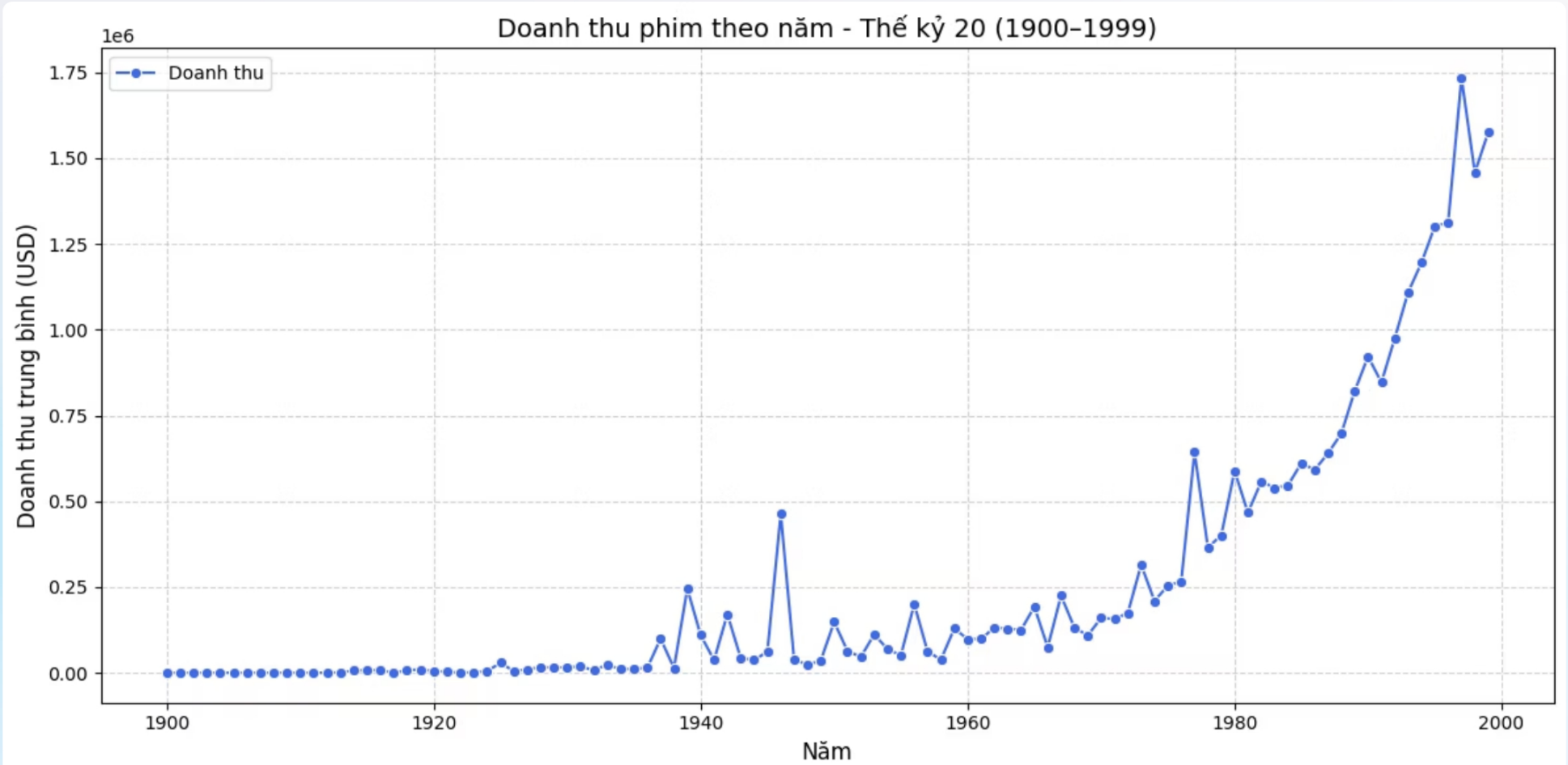
Model Building

LSTM architecture with correlation analysis

Prediction

Revenue forecasting for 2000-2020




Doanh thu phim theo năm - Thế kỷ 20 (1900–1999)
1e6
Doanh thu
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
Doanh thu trung bình (USD)
1900
1920
1940
1960
1980
2000
Năm

Conclusions & Future Work

Model Improvements

- Advanced deep learning
- Content analysis with NLP
- Recommendation systems

Future Directions

- Expand data sources
- Regional analysis
- Cast/crew impact study

Key Findings

- Drama and Comedy most produced
- Adventure/Sci-Fi highest revenue
- 90-120 minute optimal runtime

THANK YOU FOR YOUR LISTENING!

Any question could help us improve ourselves.