NTNU
Kunnskap for en bedre verden

TDT4225 - Large, distributed data sets

Assignment 3 - TDT4225

Authors:
Ola Munthe Vassbotn
Sasha Elizabeth Landell-Mills
Sepanta Jamshid Ganjei

31st October 2025

Table of Contents

1 Introduction

In this assignment, the datasets movies_metadata.csv, credits.csv, keywords.csv, ratings.csv, and links.csv were utilized for exploratory data analysis (EDA), database setup, and query execution. These datasets originate from Kaggle's The Movies Dataset, which is divided into multiple files with descriptive names.

The primary objective of the EDA was to obtain an understanding of the data, identify potential inconsistencies, and perform the necessary cleaning to ensure data quality for use in the database. Following the data preparation, the cleaned datasets were imported into a MongoDB database, where an appropriate schema design was developed based on the structure of the available data and the queries to be performed.

Finally, a series of queries were executed on the MongoDB database to extract relevant insights. The results of these queries are presented in Section 6.

2 Datasets before cleaning

There was no specific business objective, trend, or common pattern we aimed to explore through the queries after conducting the EDA. Therefore, the main purpose of the EDA was to gain an overall understanding of the dataset. This included identifying its structure, main features, and any preliminary patterns that could inform future analysis. The datasets we looked at in the EDA were: movies_metadata.csv, credits.csv, keywords.csv, links.csv and ratings.csv. The Python script used to find these results can be found in the directory eda.

2.1 EDA of movies_metadata.csv

List of features and type in movies_metadata.csv can be seen in Table 1. The proportion of missing and zero values for each feature in movies_metadata.csv is summarized in Table 2. Since this is a large dataset with many different variables, every notable variable has been explored in its own subsection.

Number of rows in movies_metadata.csv: 45572
Number of columns in movies_metadata.csv: 24

Table 1: Features and types in movies_metadata.csv file before cleaning.

Feature	Type
adult	object
belongs_to_collection	object
budget	object
genres	object
homepage	object
id	object
imdb_id	object
original_language	object
original_title	object
overview	object
popularity	object
poster_path	object
production_companies	object
production_countries	object
release_date	object
revenue	float64
runtime	float64
spoken_languages	object
status	object
tagline	object
title	object
video	object
vote_average	float64
vote_count	float64

Table 2: Missing and zero values per feature in movies_metadata.csv before cleaning.

Feature	Missing Values (%)	Zero Values (%)
belongs_to_collection	40972 (90.12%)	0
homepage	37684 (82.88%)	0
imdb_id	17 (0.04%)	0
original_language	11 (0.02%)	0
overview	954 (2.10%)	0
popularity	5 (0.01%)	0
poster_path	386 (0.85%)	0
production_companies	3 (0.01%)	0
production_countries	3 (0.01%)	0
release_date	87 (0.19%)	0
revenue	6 (0.01%)	38052 (83.69%)
runtime	263 (0.58%)	1558 (3.43%)
spoken_languages	6 (0.01%)	0
status	87 (0.19%)	0
tagline	25054 (55.10%)	0
title	6 (0.01%)	0
video	6 (0.01%)	0
vote_average	6 (0.01%)	2998 (6.59%)
vote_count	6 (0.01%)	2899 (6.38%)

2.1.1 status

The unique status values where found to be: Released, Rumored, Post Production, In Production, Planned, Canceled. The distribution of this variable within the dataset is shown in Table 3.

Table 3: Overview of movie production statuses from `movies_metadata.csv` before cleaning.

Status	Number of Movies	Percentage
Released	45,014	99.01%
Rumored	230	0.51%
Post Production	98	0.22%
In Production	20	0.04%
Planned	15	0.03%
Canceled	2	0.00%

2.1.2 revenue

Max revenue, min revenue, average, and median revenue of the movies in `movies_metadata.csv` is shown in Table 4 b.

Table 4: Revenue statistics for movies in `movies_metadata.csv` before cleaning.

Statistic	Value
Mean	11,209,348.54
Median	0.00
Max	2,787,965,087.00
Min	0.00

2.1.3 budget

The mean, median, min, and max budget of all movies in the `movies_metadata.csv` is shown in Table 5.

Table 5: Budget statistics for budget variable in the `movies_metadata.csv` before cleaning

Statistic	Value
Mean	4,224,578.81
Median	0.00
Max	380,000,000.00
Min	0.00

To explore the relationship between a movie's budget and revenue a scatterplot was made with the two variables, as shown in Figure 1.

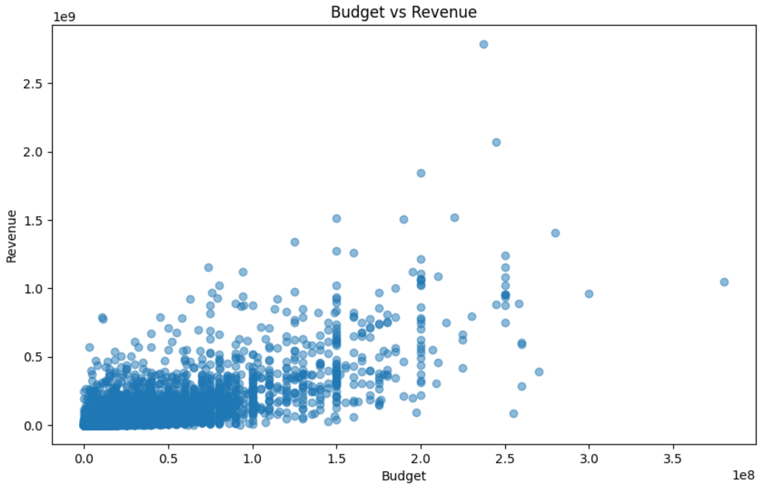

Figure 1: Relationship between movie budget and revenue from the `movies_metadata.csv` data before cleaning, shown as a scatterplot.

2.1.4 runtime

The runtime variable was explored as well. A general overview of the distribution is shown in Table 6. The frequency of movies with different runtimes is shown in a histogram, Figure 2, where all the movies from `movies_metadata.csv` are divided into 25-minute bins.

Table 6: Runtime statistics for `movies_metadata.csv` before cleaning

Statistic	**Value**
Max	1,256.00 minutes
Min	0.00 minutes
Mean	94.13 minutes
Median	95.00 minutes
Movies with runtime above 300 minutes	108
Movies with 0 runtime	1,558

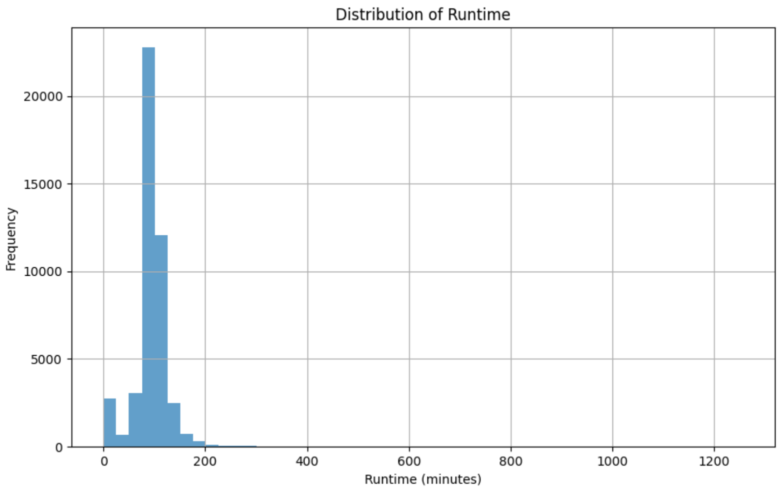

Figure 2: Runtime of movies divided into 25.12 minute bins from the `movies_metadata.csv`, before cleaning

Because of this wide distribution in Figure 2, we wanted to look the distribution of runtimes, without the extremal points. This was achieved by capping the data at the 99th percentile, and plotting it into a histogram. This is shown in Figure 3, where movies are divided into 3.7 minute bins.

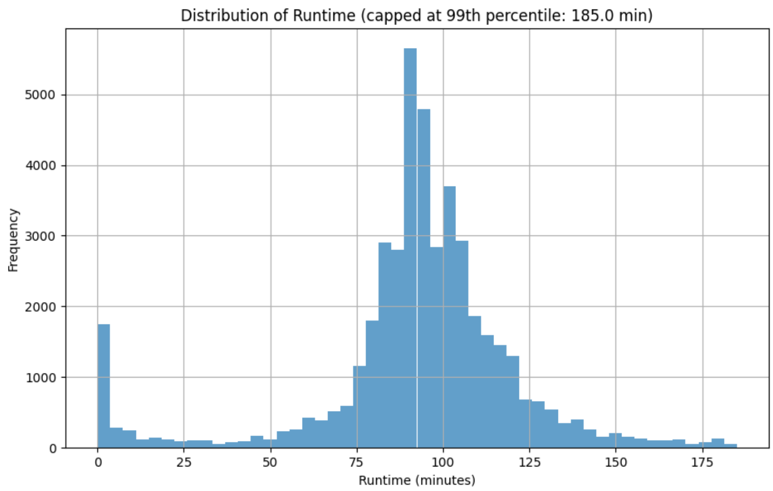

Figure 3: Runtime of 99th percentile of movies divided into 3.7 minute bins from the `movies_metadata.csv`, before cleaning

2.1.5 genre

Number of different genres in `movies_metadata.csv` before cleaning: 32

As shown in Table 7, the average and median runtimes for each genre, along with the number of movies, are summarized.

Table 7: Average and Median Runtime for Each Genre

Genre	Average Runtime (min)	Median Runtime (min)	Number of Movies
History	124.15	113	1,395
War	112.53	106	1,320
Drama	103.05	100	20,210
Romance	102.44	100	6,723
Adventure	101.41	98	3,490
Action	100.77	98	6,585
Foreign	100.57	98	1,622
Crime	99.91	98	4,298
Music	99.18	98	1,597
Thriller	98.56	96	7,613
Mystery	98.30	96	2,465
Western	97.47	95	1,042
Science Fiction	93.99	92	3,038
Fantasy	93.91	94	2,302
TV Movie	93.38	98	756
Comedy	91.40	94	13,095
Horror	89.81	92	4,670
Family	87.48	90	2,759
Documentary	87.15	87	3,920
Animation	64.27	75	1,930
Aniplex	nan	nan	0
BROSTA TV	nan	nan	0
Carousel Productions	nan	nan	0
GoHands	nan	nan	0
Mardock Scramble Production Committee	nan	nan	0
Odyssey Media	nan	nan	0
Pulser Productions	nan	nan	0
Rogue State	nan	nan	0
Sentai Filmworks	nan	nan	0
Telescene Film Group Productions	nan	nan	0
The Cartel	nan	nan	0
Vision View Entertainment	nan	nan	0

2.1.6 release_date

To better understand how the release_date data was distributed, a histogram of the release dates of movies by year was created (see Figure 4).

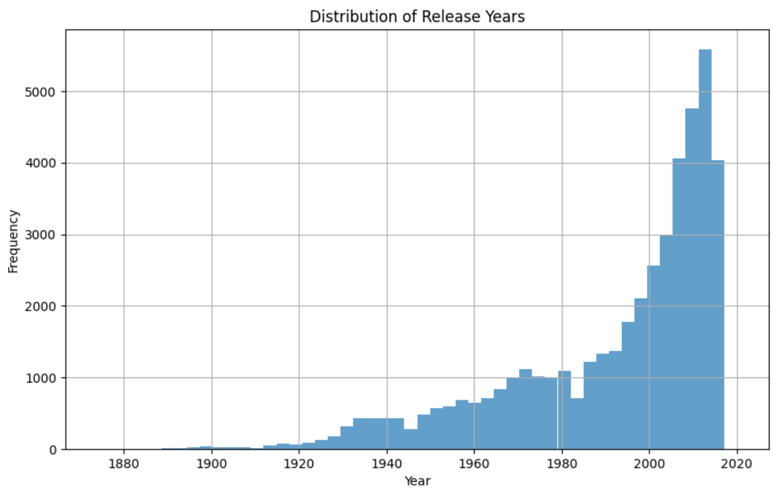

Figure 4: Number of movies released each year from `movies_metadata.csv`, before cleaning

2.1.7 production_companies

Number of unique production companies: 23537

Top 3 production companies:
Warner Bros : 1250 movies
Metro-Goldwyn-Mayer (MGM): 1076 movies
Paramount Pictures: 1003 movies

2.1.8 production_countries

Number of unique production countries: 160

Top 3 production countries:
United States of America: 21153 movies
United Kingdom: 4094 movies
France: 3940 movies

In Figure 5, the number of movies produced in the top 10 countries are shown.

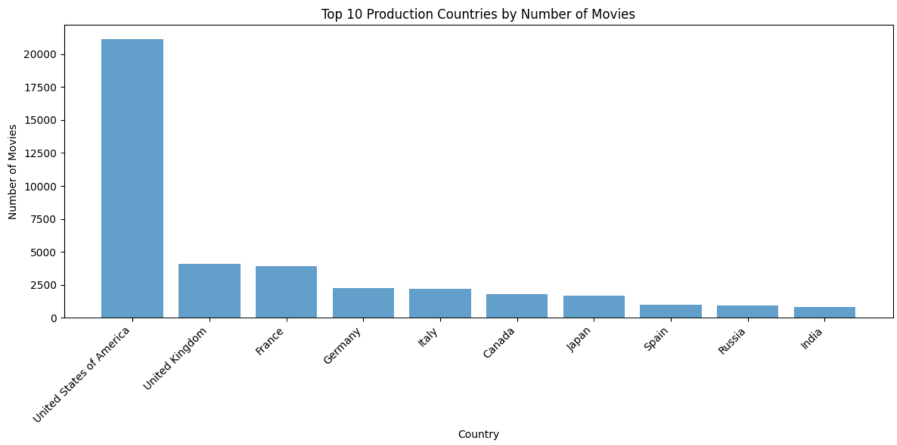

Figure 5: Number of movies produced in the top 10 production countries from `movies_metadata.csv`, before cleaning.

2.1.9 spoken_languages

Number of unique spoken languages: 75
Average number of movies per spoken language: 710.67

Top 3 spoken languages:
English: 28745 movies
Français: 4196 movies
Deutsch: 2625 movies

2.1.10 Vote_count & vote_average

Average vote_average: 5.62
Average vote_count: 109.90 votes

Figure 6 shows that most movies have an average rating between 5 and 7, while a small cluster with 0 ratings.

Figure 7 reveals a highly skewed distribution where most movies have very few votes, and only a few popular titles have thousands.

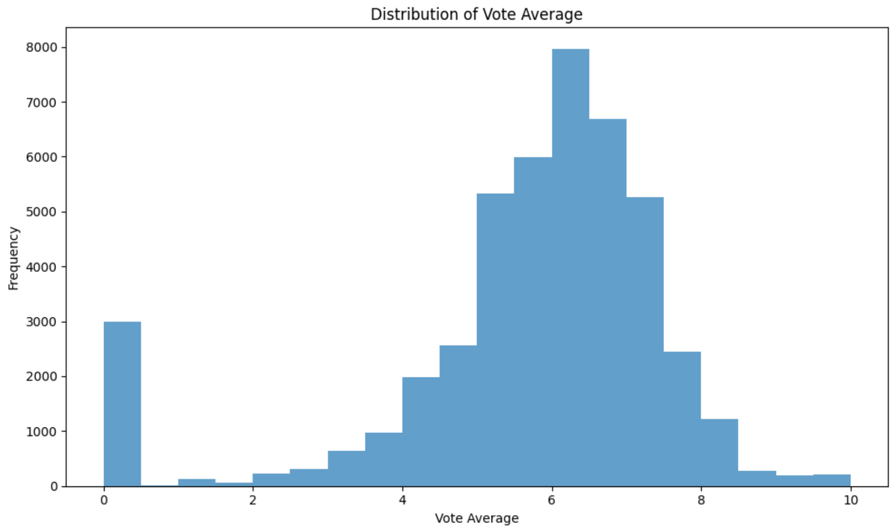

Figure 6: Distribution of vote average in `movies_metadata.csv` before cleaning

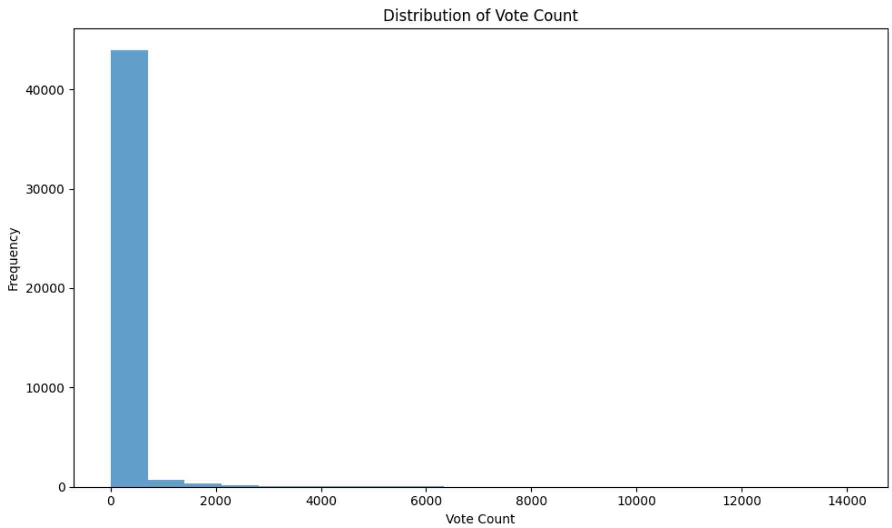

Figure 7: Distribution of vote count in `movies_metadata.csv` before cleaning.

2.1.11 belongs_to_collection

Number of unique collections in `movie_metadata.csv`: 1695

Top 10 collections by number of movies:
The Bowery Boys: 29
Totò: 27
James Bond: 26
Zatôichi: The Blind Swordsman: 26

Carry On: 25
Pokémon: 22
Charlie Chan (Sidney Toler): 21
Godzilla (Showa): 16
Uuno Turhapuro: 15
Charlie Chan (Warner Oland): 15

Average number of movies in a collection: 2.65

2.2 EDA of `credits.csv`

List of features, and their types, in `credits.csv` can be seen in Table 8.

Number of rows: 45476
Number of columns: 3
Cast: 2418 rows with null value or empty arrays
Crew: 771 rows with null value or empty arrays

Table 8: Features and types in `credits.csv` file before cleaning.

Feature	**Type**
cast	object
crew	object
id	int64

2.3 EDA of `keywords.csv`

The attributes and their corresponding data types in the `keywords.csv` dataset are shown in Table 9. The file contains 2 columns and 46419 rows. There where no missing values for movies or keywords.

Keywords had 14795 rows with null or empty arrays.
There where 0 rows that have missing or zero values for ID.

Table 9: Attributes and data types in `keywords.csv` before cleaning.

Attribute	**Type**
id	int64
keywords	object

2.4 EDA of `ratings.csv`

The features and their corresponding data types in the `ratings.csv` file are presented in Table 10. No missing or zero values were found for any of the variables.

Number of rows in `ratings.csv`: 26024289
Number of columns in `ratings.csv`: 4
Average rating: 3.528
Median rating: 3.5
Total number of users: 270 896
Highest rating: 5

lowest rating: 0.5

Table 10: Features and data types in `ratings.csv` before cleaning

Feature	Type
userId	int64
movieId	int64
rating	float64
timestamp	int64

2.5 EDA of `links.csv`

The attributes and their corresponding data types in the `links.csv` file are listed in Table 11 for the attributes in `links.csv`. The number of rows where found to be 45843, and there where 3 columns. The only missing values were in the tmdbId row. For tmdbId 219 (0.48%) rows had null values.

Table 11: Attributes and data types in `links.csv` before cleaning

Attribute	Type
movieId	int64
imdbId	int64
tmdbId	float64

3 Cleaning the Data

The data was cleaned to address missing values and inconsistencies, to ensure the query results were precise and accurate. The datasets not mentioned have not been altered. The cleaning was performed in the order listed.

The python code for cleaning the data can be found at `data_cleaning.py`.

3.1 Cleaning `movies_metadata.csv`

Firstly, all movies with a runtime of 0 were assigned the average runtime for their genre. This did not change the number of rows. This was based on the assumption that a movie is not a movie if it has runtime below 1 minute.

All movies with a different status than `released` were also removed. This was based on the assumption that in the queries, only released movies were of interest. This led to a reduction of 452 rows in `movies_metadata.csv`.

The following columns were removed from `movies_metadata.csv` as they served no purpose for the queries later: homepage, original_title, overview, popularity, poster_path, status, tagline, and video.

All rows with genres: "Aniplex", "BROSTA TV2", "Carousel Productions", "GoHands", "Mardock Scramble Production Committee", "Odyssey Media", "Pulser Productions ", "Rogue State", "Sentai Filmworks ", "Telescene Film Group Productions", "The Cartel", and "Vision View Entertainment" were removed.

Rows before cleaning: 45466
Rows removed: 452

Rows after cleaning: 45014

3.1.1 Cleaning `movies_metadata.csv` using `ratings.csv`

Several duplicate rows were identified in `movies_metadata.csv`. These duplicates contained identical information, except for differences in the vote_count field. To address this inconsistency, `links.csv` was used to connect `movies_metadata.csv` and `ratings.csv` via the different IDs present in the datasets. Using this, we calculated the correct vote variables and updated the corresponding rows. Following this, the duplicate rows in `movies_metadata.csv` were removed, resulting in a smaller dataset.

Rows removed: 29
Rows after cleaning: 44985

3.2 Cleaning `credits.csv`

All rows with no crew information were removed. This was due to the assumption that one cannot make a movie without any crew, and therefore, these entries were invalid.

Rows before cleaning: 45476
Rows removed: 771
Rows after cleaning: 44705

3.3 Cleaning `keywords.csv`

All rows where either movie ID or keywords were empty were removed.

Rows before cleaning: 46419
Rows removed: 14795
Rows after cleaning: 31624

3.4 Cleaning `links.csv`

`links.csv` works as a connector between the `movies_metadata.csv` and `ratings.csv` as the file maps movieIDs to TMDB and IMDB IDs. All rows with at least one missing attribute in links.csv were removed.

Rows before cleaning: 45843
Rows removed: 219
Rows after cleaning: 45843

4 Data after cleaning

4.1 Variables in `movies_metadata.csv` after cleaning

The features / columns in `movies metadata_csv` after cleaning are shown in Table 12.

Table 12: Features and types in `movies_metadata.csv` file after cleaning.

Feature	**Type**
belongs_to_collection	object
budget	object
genres	object
id	object
imdb_id	object
original_language	object
production_companies	object
production_countries	object
release_date	object
revenue	float64
runtime	float64
spoken_languages	object
title	object
vote_average	float64
vote_count	float64

4.2 Vote variables in `movies_metadata.csv` after cleaning

In Table 13 the statistics for the vote variables vote_avarage and vote_count are shown after cleaning.

Table 13: Summary statistics for vote variables after cleaning

Statistic	Value
Average vote_average	3.2773
Average vote_count	577.43

4.3 Revenue variables in `movies_metadata.csv` after cleaning

In Table 14 the statistics for revenue in cleaned `movies_metadata.csv` is shown.

Table 14: Summary statistics for revenue variable after cleaning

Statistic	Value
Mean	11,322,295.80
Median	0.00
Max	2,787,965,087
Min	0

4.4 Genre variable in `movies_metadata.csv` after cleaning

After cleaning, the dataset contains movies across 20 different genres. Table 15 shows the average runtime and movie count for each genre in the cleaned dataset.

Table 15: Average runtime by genre after cleaning

Genre	Avg Runtime (min)	Movie Count
History	126.0	1,388
War	113.9	1,314
Drama	105.0	20,004
Foreign	104.7	1,586
Romance	104.2	6,651
Adventure	103.0	3,462
Action	102.6	6,533
Crime	101.5	4,274
Music	100.8	1,586
Thriller	100.5	7,557
Mystery	99.9	2,451
Western	98.8	1,038
TV Movie	96.7	748
Science Fiction	95.4	3,004
Fantasy	95.2	2,279
Comedy	94.5	12,987
Horror	91.3	4,632
Documentary	90.3	3,860
Family	89.4	2,729
Animation	65.3	1,908

5 MongoDB database

For this project, the team set up a MongoDB database based on the cleaned data in `movies_metadata.csv`, `credits.csv`, `keywords.csv`, `ratings.csv` and `links.csv`. Then 4 diffrent collections were created within the DB: credits, movies, people, ratings.

The credits collection consists of the same rows as `credits.csv`. The movies collection consists of the same rows as `movies_metadata.csv`, with the keywords from `keywords.csv` added as a field. The ratings collection consists of the same rows as `rating.csv`. The people collection was created from the `credits.csv` with fields; id, name and gender. This collection was made to make information about both crew and actors more readily available for the queries, as many of the queries ask about people. The code responsible for setting up the database can be found in `setup_mongodb.py`.

6 Queries

In this section, queries were executed on the database to answer the questions presented. Each query corresponds to a file in the delivered materials, following the format `query[NUMBER].py`.

6.1 Query 1: Top Directors by Median Revenue

Considering only crew members with job = Director, this query identifies the 10 directors with ≥ 5 movies who have the highest median revenue, along with their movie count and mean vote average.

Table 16: Top 10 Directors (≥ 5 movies) with Highest Median Revenue

Director	**Movies**	**Median Revenue**	**Mean Vote Avg**
George Lucas	7	$649,398,328.00	3.48
Francis Lawrence	6	$619,388,635.50	3.46
David Yates	10	$583,042,696.50	3.67
Chris Renaud	5	$543,513,985.00	3.45
Eric Darnell	5	$532,680,671.00	6.34
Tom McGrath	5	$532,680,671.00	3.30
Carlos Saldanha	7	$484,635,760.00	3.14
Andrew Adamson	6	$452,030,315.50	3.42
Michael Bay	13	$449,220,945.00	3.02
Brad Bird	6	$416,438,570.00	3.68

6.2 Query 2: Frequently Co-Starring Actor Pairs

This query identifies actor pairs who have co-starred in ≥ 3 movies together, reporting their number of co-appearances and average movie vote average.

Table 17: Actor Pairs with ≥3 Co-Starring Movies (Top 50 by Co-Appearances, Avg Vote)

Actor 1	**Actor 2**	**Co-Appearances**	**Avg Vote**
Huntz Hall	Leo Gorcey	35	3.43
Charlie Chaplin	Edna Purviance	33	3.12
Mayumi Tanaka	Masako Nozawa	31	3.14
Oliver Hardy	Stan Laurel	30	3.44
Masako Nozawa	Naoko Watanabe	30	3.10
Masako Nozawa	Hiromi Tsuru	29	3.13
Lou Costello	Bud Abbott	27	3.36
Jeff Bennett	Rob Paulsen	27	2.97
Grey Griffin	Frank Welker	26	3.05
Raymond Burr	Barbara Hale	25	4.36
John Wayne	Paul Fix	24	3.25
Toshio Furukawa	Masako Nozawa	24	3.12
Masako Nozawa	Daisuke Gouri	24	3.02
Frank Welker	Jeff Bennett	22	2.94
Masako Nozawa	Ryou Horikawa	21	3.24
Frank Mayo	Jack Mower	21	3.16
Jim Cummings	Frank Welker	21	3.14
Peter Cushing	Christopher Lee	21	2.90
Jim Cummings	Jeff Bennett	21	2.83
Bernard Gorcey	Leo Gorcey	20	4.42
Charlie Chaplin	Henry Bergman	20	3.43
Masako Nozawa	Masako Nozawa	20	3.17
William R. Moses	Barbara Hale	19	5.07
Bernard Gorcey	Huntz Hall	19	4.49
Toshirō Mifune	Takashi Shimura	19	3.83
Buster Keaton	Joe Roberts	19	3.50
Simon Yam	Lam Suet	19	3.39
Charlie Chaplin	John Rand	19	3.30
Sammo Hung	Yuen Biao	19	3.22
Jackie Chan	Yuen Biao	19	3.13
Kōhei Miyauchi	Masako Nozawa	19	3.12
Charlie Chaplin	Albert Austin	18	3.36
Frank Welker	Mindy Cohn	18	3.04
Adam Sandler	Allen Covert	18	2.85
Kenneth Williams	Charles Hawtrey	18	2.04
Bess Flowers	Harold Miller	17	3.69
Bing Crosby	Bob Hope	17	3.53
Jackie Chan	Ken Lo	17	3.38
Cheech Marin	Tommy Chong	17	3.28
John DiMaggio	Tom Kenny	17	3.13
Jason Mewes	Kevin Smith	17	3.12
Grey Griffin	Mindy Cohn	17	3.04
Charlie Chaplin	Leo White	17	2.98
Kenneth Williams	Joan Sims	17	2.11
Charles Hawtrey	Joan Sims	17	2.00
Raymond Burr	William R. Moses	16	4.89
Leo Gorcey	David Gorcey	16	4.28
Huntz Hall	David Gorcey	16	4.10
Irving Bacon	Bess Flowers	16	3.83
John Ridgely	Jack Mower	16	3.34

6.3 Query 3: Actors with Widest Genre Breadth

This query identifies the top 10 actors (with $\geq$ 10 credited movies) that have appeared in the widest variety of genres.

List the top 10 actors (with $\geq$ 10 credited movies) that have the widest genre breadth. Report the actor, the number of distinct genres they've appeared in, and up to 5 example genres.

Table 18: Top 10 Actors ($\geq$ 10 movies) with Widest Genre Breadth

Actor	**Movies**	**Genres**	**Example Genres**
Christopher Lee	379	20	Thriller, Fantasy, Action, Adventure, Crime
Donald Sutherland	278	20	Family, Foreign, War, Documentary, Horror
Christopher Walken	255	20	Fantasy, Action, Crime, Adventure, Romance
Liam Neeson	224	20	Adventure, Crime, Romance, History, Mystery
Keith David	223	20	History, Crime, Horror, Adventure, Romance
Dennis Hopper	213	20	TV Movie, Comedy, Romance, Adventure, Crime
Jim Broadbent	204	20	Thriller, Drama, Animation, Music, TV Movie
Charlton Heston	194	20	Thriller, Drama, Animation, Music, History
James Earl Jones	181	20	Drama, Action, Animation, Music, Thriller
Ned Beatty	176	20	Western, Foreign, Family, War, Horror

6.4 Query 4: Top Film Collections by Revenue

This query identifies the top 10 film collections (with $\geq$ 3 movies) that have the largest total revenue.

For film collections (belongs_to_collection.name not null) with $\geq$ 3 movies, which 10 collections have the largest total revenue? For each, report movie count, total revenue, median vote_average, and the earliest $\rightarrow$ latest release date.

Table 19: Top 10 Film Collections ($\geq$ 3 movies) with Largest Total Revenue

Collection	**Movies**	**Total Revenue**	**Med. Vote**	**Date Range**
Harry Potter	8	7,707,367,425	3.78	2001-11-16 $\rightarrow$ 2011-07-07
Star Wars	8	7,434,494,790	3.86	1977-05-25 $\rightarrow$ 2016-12-14
James Bond	26	7,106,970,239	3.43	1962-10-04 $\rightarrow$ 2015-10-26
The Fast and the Furious	8	5,125,098,793	3.19	2001-06-22 $\rightarrow$ 2017-04-12
Pirates of the Caribbean	5	4,521,576,826	3.41	2003-07-09 $\rightarrow$ 2017-05-23
Transformers	5	4,366,101,244	2.77	2007-06-27 $\rightarrow$ 2017-06-21
Despicable Me	6	3,691,070,216	3.28	2010-07-08 $\rightarrow$ 2017-06-15
The Twilight	5	3,342,107,290	2.41	2008-11-20 $\rightarrow$ 2012-11-13
Ice Age	5	3,216,708,553	3.29	2002-03-10 $\rightarrow$ 2016-06-23
Jurassic Park	4	3,031,484,143	3.17	1993-06-11 $\rightarrow$ 2015-06-09

6.5 Query 5: Runtime Trends by Decade and Genre

This query analyzes median runtime and movie count by decade and primary genre (first element in genres array).

Table 20: Primary Genre by Decade (Highest Movie Count) with Median Runtime

Decade	Primary Genre	Movies	Median Runtime (min)
1870s	Documentary	2	1.0
1880s	Documentary	4	1.0
1890s	Documentary	27	1.0
1900s	Comedy	17	3.0
1910s	Comedy	54	28.5
1920s	Drama	163	90.0
1930s	Drama	380	85.0
1940s	Drama	467	98.0
1950s	Drama	633	99.0
1960s	Drama	718	102.0
1970s	Drama	858	102.0
1980s	Drama	852	105.0
1990s	Drama	1444	104.0
2000s	Drama	3086	101.0
2010s	Drama	3132	100.0

6.6 Query 6: Gender Representation in Top Billed Cast

This query calculates the average proportion of female cast members in the top 5 billed positions, aggregated by decade.

Table 21: Average Female Proportion in Top 5 Cast by Decade

Decade	Movie Count	Avg Female Proportion
1870s	1	1.0000 (100.00%)
1890s	2	0.5000 (50.00%)
1900s	11	0.5000 (50.00%)
1910s	142	0.3414 (34.14%)
1920s	368	0.3284 (32.84%)
1930s	1239	0.3691 (36.91%)
1940s	1425	0.3581 (35.81%)
1950s	1985	0.3370 (33.70%)
1960s	2363	0.3249 (32.49%)
1970s	3101	0.3113 (31.13%)
1980s	3511	0.3196 (31.96%)
1990s	5076	0.3392 (33.92%)
2000s	9411	0.3630 (36.30%)
2010s	10229	0.3774 (37.74%)

6.7 Query 7: Top Noir Films by Rating

This query identifies movies matching "noir" or "neo-noir" in overview or tagline (with vote_count ≥ 50) that have the highest vote average.

Table 22: Top 20 'Noir' Movies (vote_count ≥ 50) by Vote Average

Title	**Year**	**Vote Avg**	**Vote Count**
Casablanca	1942	4.21	30043
The Third Man	1949	4.21	7676
Double Indemnity	1944	4.20	5607
Sunset Boulevard	1950	4.20	7930
Notorious	1946	4.17	5486
The Big Sleep	1946	4.16	6303
Chinatown	1974	4.16	18397
Touch of Evil	1958	4.16	5199
Memento	2000	4.16	40706
The Maltese Falcon	1941	4.14	14281
Shadow of a Doubt	1943	4.13	2452
Strangers on a Train	1951	4.12	5966
Blade Runner	1982	4.12	37152
Vertigo	1958	4.12	17219
High and Low	1963	4.11	754
Léon: The Professional	1994	4.08	34361
Out of the Past	1947	4.08	1230
Rebecca	1940	4.07	5374
Laura	1944	4.07	2992
The Killing	1956	4.07	2401

6.8 Query 8: Director-Actor Collaborations

This query identifies the top 20 director-actor pairs with ≥ 3 collaborations (considering only movies with vote_count ≥ 100) that have the highest mean vote average.

Which 20 director–actor pairs with ≥ 3 collaborations (same movie) have the highest mean vote_average, considering only movies with vote_count ≥ 100? Include the pair's films count and mean revenue.

Table 23: Top 20 Director–Actor Pairs (≥3 Collaborations, vote_count ≥100) by Mean Vote

Director	**Actor**	**Films**	**Mean Vote**	**Mean Revenue ($)**
Francis Ford Coppola	Roman Coppola	3	4.29	113,384,031.00
Francis Ford Coppola	John Cazale	3	4.22	99,009,750.67
Francis Ford Coppola	Robert Duvall	4	4.19	96,622,408.25
Akira Kurosawa	Daisuke Katô	4	4.17	105,912.25
Akira Kurosawa	Tatsuya Nakadai	6	4.14	723,582.33
Akira Kurosawa	Haruya Sakamoto	3	4.13	90,613.67
Akira Kurosawa	Ichirô Chiba	4	4.13	81,770.25
Alfred Hitchcock	Bess Flowers	6	4.12	19,042,342.50
Akira Kurosawa	Hiroshi Tachikawa	3	4.11	0.00
Akira Kurosawa	Senkichi Ômura	5	4.11	54,368.20
Akira Kurosawa	Shin Ôtomo	4	4.11	67,960.25
Akira Kurosawa	Toranosuke Ogawa	3	4.11	109,027.00
Akira Kurosawa	Minoru Itô	3	4.11	90,613.67
Akira Kurosawa	Haruo Suzuki	3	4.11	90,613.67
Akira Kurosawa	Shôichi Hirose	4	4.10	67,960.25
Peter Jackson	Sean Bean	3	4.10	972,181,581.00
Peter Jackson	Sean Astin	3	4.10	972,181,581.00
Peter Jackson	Dominic Monaghan	3	4.10	972,181,581.00
Peter Jackson	Viggo Mortensen	3	4.10	972,181,581.00
Peter Jackson	John Rhys-Davies	3	4.10	972,181,581.00

6.9 Query 9: Non-English Films with US Involvement

Among movies where original_language ≠ "en" but at least one production company or country is United States, which are the top 10 original languages by count? For each language, report the count and one example title.

In this query, we assume that "top" means the language with the most movies fulfilling the criteria.

Table 24: Top 10 Original Languages (Non-English, US Production)

Language	**Count**	**Example Title**
fr	111	Wings of Courage
es	71	Bitter Sugar
it	55	Frankie Starlight
de	51	Cold Fever
ja	30	Godzilla 1985
pt	14	Senseless
xx	14	Quest for Fire
nl	12	Come On, Rangers
zh	11	Eat Drink Man Woman
ru	11	Dark Eyes

6.10 Query 10: User Rating Behavior Analysis

This query analyzes user rating behavior, identifying both the most genre-diverse users and the users with highest rating variance (only users with ≥ 20 ratings).

Table 25: Top 10 Most Genre-Diverse Users (≥ 20 ratings)

User ID	**Ratings Count**	**Genre Count**	**Variance**
16	134	20	0.2087
34	261	20	1.3050
37	190	20	0.7266
46	728	20	0.6871
56	233	20	0.8557
62	419	20	0.6960
68	247	20	0.7507
79	160	20	1.0892
120	564	20	0.7936
125	266	20	1.3576

Table 26: Top 10 Highest-Variance Users (≥ 20 ratings)

User ID	**Ratings Count**	**Variance**	**Genre Count**
6694	28	5.0625	13
97817	42	5.0625	10
218479	45	5.0600	12
55744	23	5.0529	9
224431	23	5.0529	12
258054	23	5.0529	10
187349	34	5.0450	10
163370	181	5.0277	19
139287	24	5.0273	7
141972	35	5.0253	14

7 Discussion of results

In this section the data cleaning process and the data after cleaning is discussed (see Section 7.1), as well as the results from the queries run (see Section 7.2).

7.1 Analysis and cleaning

During the data cleaning process, several rows across the different `.csv` files were removed because they contained null values. These missing values resulted in incomplete information for the affected entries, making them less useful for the intended database queries. There was also a goal of reducing the dataset size, where possible, to make the queries run faster.

Each cleaning step was performed based on specific assumptions, which are described in the cleaning documentation. For example, it was assumed that a movie cannot exist without any associated crew members, leading us to remove all rows lacking crew data in `credits.csv`. Similarly, in `movies_metadata.csv`, only movies with the status `"released"` were retained, as these were considered to be the only entries relevant for the subsequent analyses and queries.

For all movies with a runtime of 0, the average runtime for their respective genre was assigned. This decision was based on the finding that nearly 2,000 movies had a recorded runtime of 0 (see Figure 2). The team was worried removing these entries could have led to biased or misleading results in later queries.

To verify the validity of this approach, several movies with a registered runtime of 0 were manually checked using Google to determine their actual runtimes. Since these movies were confirmed to be real and to have runtimes greater than 0, the team decided to impute the missing values using the average runtime for each genre. This approach was considered reasonable, as the genre-based average provided a plausible estimate and served as a form of single imputation. Although single imputation is generally not recommended unless less than 5% of the data is affected, this condition was satisfied in this case (Joseph R et al. 2018; Kaiser 2014). By applying this method, the team was able to preserve more data and retain potentially valuable information for subsequent analyses and queries.

The team recognizes this cleaning could have been done in several ways. For instance, in the future, the team could consider using the median runtime for each genre (Kaiser 2014). More complicated methods, such as assigning runtime based on director or release year or an average of different variables, could have been used. Though the team feared this would be too specific and lead to misleading results. Further, these methods where seen as overall too complicated by the team.

There were inconsistencies found between `movies_metadata.csv` and `ratings.csv` in the EDA, as discussed in Section 2.1.10. The inconsistency between `movies_metadata.csv` and `ratings.csv` highlights a common challenge when integrating datasets from multiple sources. By using `links.csv` as a bridge between the two files, it was possible to ensure consistency and maintain data integrity when updating the variables vote_average and vote_count in the `movies_metadata.csv` file based on the information in `ratings.csv`. The decision to prioritize the values from `ratings.csv` was based on the understanding that this dataset represents the original user ratings, while the vote variables in `movies_metadata.csv` may be outdated. Furthermore, based on the results of the exploratory data analysis (EDA), the team concluded that `ratings.csv` did not require additional cleaning, as no anomalies or missing values of significance were identified.

The average vote_average changed from 5.63 to 3.3, and the average vote_count changed from 109.9 to 577 after cleaning. This is a substantial change, explained by the two files having diffrent voting scales. While `ratings.csv` used a 0-5 scale, `movies_metadata.csv` used a 0-10 scale, leading to a different vote_average.

The statistics for the revenue variables did not change drastically after cleaning. The mean revenue changed from 11,209,348.54 to 11,322,295.8, while the maximum value remained unchanged at 2,787,965,087. The median and minimum also stayed stable at 0. There are no clear reasons for these minor changes, except that the removal of invalid entries and other cleaning operations in `movies_metadata.csv` may have affected the revenue variable. For example, removing invalid or null values could have led to the increase observed in the mean.

There are no significant changes in the average runtime by genre before (Table 7) and after cleaning (Table 15), except for the removal of 12 rows. These rows were removed as they appeared to represent production companies that were mistakenly categorized as genres. Furthermore, they had no movies associated with them. All genres show a slightly higher average runtime and a few fewer movies after cleaning. The reduction in the number of movies is expected, given the other data cleaning operations performed. However, there is no clear explanation for the observed increase in average runtime.

Rows with missing values in `keywords.csv` and `links.csv` were removed, as they did not provide any meaningful information for the analysis. In `links.csv`, rows containing missing values could not function as valid links between datasets and were therefore considered useless.

Similarly, rows with missing values in `keywords.csv` were removed, since keywords without an associated movie, or movies without corresponding keywords, do not contribute any analytical value.

All genres with no movies connected to them were removed. This was because these rows appeared to be production companies that were mistakenly categorized as genres in the dataset and contained no movies connected to them.

There are no graphs shown for after cleaning the dataset as there were no particular graphs of interest with regard to the queries. Only the changes seen as directly valuable for the queries were shown.

7.2 Queries

For this section, we have chosen to discuss only the results that stood out to us, as many of the results from the queries were as expected.

Since there did not appear to be a general trend, business objective, or common pattern in the queries, they primarily provided additional insights into the data. Where relevant, the query findings are compared to the results of the EDA and the cleaned data to discuss the data and verify the correctness of the findings.

7.2.1 Query 2

In Query 2, we see that the average vote for movies featuring the most common co-stars seems to be around the overall average_vote of 3.3 found in Section 4.2 in general. It is surprising there is not a higher vote average for these movies, as one might expect that if two co-stars produce good average ratings, they are more likely to be cast together again.

7.2.2 Query 3

In Query 3, all of the top 10 actors with the widest genre breadth have appeared in 20 genres. This is due to this being the total number of different genres after cleaning. Since all the actors have appeared in the same number of genres, the team interpreted "top" as the actors with the most movies. This query could also have been run with defining "top" as the highest rated actors.

7.2.3 Query 4

In Query 4, the series with the highest revenue was *Harry Potter*. The series consists of 8 movies and has a total revenue of $7,707,367,425, which is about $963,420,928 per movie. In comparison, the average revenue per movie is $11,322,295.80 (Section 4.3). This means each *Harry Potter* movie makes about 85 times more money than the average movie. However, we can see that the median revenue for movies is 0 (Section 4.3), meaning that most movies do not make any money. Therefore, we can conclude that movie revenue is not evenly distributed, and the results of the query are in line with this.

7.2.4 Query 5

From Query 5, we can observe that Drama consistently emerges as the primary genre across nearly every decade from the 1920s through the 2010s, based on the highest movie count. In the earlier decades (1870s–1910s), Documentary and Comedy briefly dominated, likely due to the short-format nature and accessibility of early filmmaking. However, starting from the 1920s, Drama became the prevailing genre and maintained this position for the remainder of the dataset.

The median runtime for Drama remains remarkably stable, typically between 100 and 105 minutes throughout the decades. This trend aligns closely with the results of the cleaned data findings, where Drama was identified as the genre with the largest number of movies (20,004) and a median runtime of approximately 100 minutes and an average runtime of 105 minutes (Table 15).

7.2.5 Query 6

In Query 6, the gender representation in the top-billed cast is shown in Table 21. There appears to be a general trend of the average female proportion decreasing over time. However, from the 1990s onward, the trend seems to shift toward a higher proportion of females appearing in the top five positions. The team found it surprising that in the 1800s, females accounted for up to half or more of the top five billed positions. They believe this can be explained by the fact that fewer films were produced overall during that period, which may have led to greater variability in the data.

7.2.6 Query 8

In Query 8, we observe that all 20 actor pairs have a mean vote_average above the overall average of 3.3 found in Section 4.2. As in Query 2, it is reasonable to assume that these pairs work well together and therefore achieve such high mean ratings. In contrast to Query 2, this theory appears to hold true here. Furthermore, many of these combinations involve the same directors, with Akira Kurosawa appearing in 12, Peter Jackson in 3, and Francis Ford Coppola in 3 of the top 20. The team does not find this surprising, as two of these directors are known for famous movie series such as The Lord of the Rings (Peter Jackson) and The Godfather (Francis Ford Coppola) (IMDB n.d.[a],[b]). These results are further supported by the fact that many of the directors have identical mean votes across several of their actor combinations. Although the team cannot determine whether these series alone produced the observed results, two directors appear to be exceptions to this pattern: Alfred Hitchcock and Akira Kurosawa.

7.2.7 Query 9

In Query 9, French (fr) was found to be the language with the most movies, excluding English with at least one production company or country is United States. This is not surprising, as French was found to be the second most spoken language in movies in the EDA in Section 2.1.9. However, it is notable that Dutch (nl) was not higher on the list, since it was found to be the third most spoken language in all movies according to the EDA in Section 2.1.9.

7.2.8 Query 10

In query 10 we can see that the users who are the most genre-diverse are not the ones with the highest variance. From this, it looks like the users who watch different genres usually rate them about the same, while users who vary the most in their ratings have not watched all the genres. From Table 26 it can be observed that many of the users have variance over 5. This is in line with the findings from EDA (Section 2.4) where the lowest rating was 0.5 and the highest 5, giving a variance of 5.06 which is the same as the highest variance the user exhibits. The calculations for variance is shown in Figure 8.

8 Feedback on assigment

We don't have any feedback on the assignment, except that it would have been useful to receive feedback on exercise 2 before handing in exercise 3.

Bibliography

IMDB (n.d.[a]). Accessed: 2025-10-30. URL: Peter%20Jackson.

— (n.d.[b]). *Francis Ford Coppola*. Accessed: 2025-10-30. URL: https://www.imdb.com/name/nm0000338/.

Joseph R, Dettori, Norvell Daniel C and Jens R Chapman (2018). 'The Sin of Missing Data: Is All Forgiven by Way of Imputation?' In: *National Library of Medecine*. DOI: 10.1177/2192568218811922. URL: https://pmc.ncbi.nlm.nih.gov/articles/PMC6293424/.

Kaiser, Jiří (2014). 'Dealing with Missing Values in Data'. In: *Journal of Systems Integration* 5.1. URL: https://www.researchgate.net/publication/304500093_Dealing_with_Missing_Values_in_Data.

Appendix

A Additional Figures

In Figure 8 the variance calculations used for query 10 is shown.

Variance of 0.5 and 5

1. Calculate the mean

$$\bar{x} = \frac{0.5+5}{2} = 2.75$$

2. Find Deviations from the mean

i) $0.5 - 2.75 = -2.25$

ii) $5 - 2.75 = 2.25$

3. Square deviations

$$(-2.25)^2 = 5.06$$

$$(2.25)^2 = 5.06$$

4. Variance

$$\sigma^2 = \frac{5.06 + 5.06}{2} = 5.06$$

Figure 8: Visualization of variance calculations for rating 0.5 and 5.