



IMDB Movie Analysis

KOVURU VENKATA CHANDANA MYTHRI

PROJECT DESCRIPTION

The objective of this project is to perform data analytics on a dataset containing various columns of different IMDb movies. The aim is to extract insights and patterns from the dataset that can help in understanding the factors that contribute to the success or failure of a movie.

Approach

To achieve this objective, we will start with data cleaning and preprocessing. This will involve handling missing values, removing duplicates, and correcting errors in the dataset. Next, we will perform exploratory data analysis (EDA) to gain insights into the data. EDA will involve visualizing the data using graphs and charts to identify trends and patterns.

Tech-Stack

EXCEL,SQL, Word, Ppt

1. Clean Data:

https://drive.google.com/file/d/1Zk3-afRpByJ71loc7bGloB1X1hWuCbVW/view?usp=share_link

director_name	movie_title	actor_1_name	budget	gross	genres	title_year	language	num_vo	num_cr	num_user_for_reviews	movie_id	imdb_score
James Cameron	Avatar	CCH Pounder	237000000	760505847	Action Ad	2009	English	886204	723	3054	http://www	7.9
Gore Verbinski	Pirates of the Caribbean	Johnny Depp	300000000	309404152	Action Ad	2007	English	471220	302	1238	http://www	7.1
Sam Mendes	Spectre	Christoph Waltz	245000000	200074175	Action Ad	2015	English	275868	602	994	http://www	6.8
Christopher Nolan	The Dark Knight Rises	Tom Hardy	250000000	448130642	Action Thi	2012	English	1144337	813	2701	http://www	8.5
Doug Walker	Star Wars: Episode VII -	Doug Walker	0	0	Document	0	English	8	0	0	http://www	7.1
Andrew Stanton	John Carter	Daryl Sabara	263700000	73058679	Action Ad	2012	English	212204	462	738	http://www	6.6
Sam Raimi	Spider-Man 3	J.K. Simmons	258000000	336530303	Action Ad	2007	English	383056	392	1902	http://www	6.2
Nathan Greno	Tangled	Brad Garrett	260000000	200807262	Adventure	2010	English	294810	324	387	http://www	7.8
Joss Whedon	Avengers: Age of Ultron	Chris Hemsworth	250000000	458991599	Action Ad	2015	English	462669	635	1117	http://www	7.5
David Yates	Harry Potter and the Ha	Alan Rickman	250000000	301956980	Adventure	2009	English	321795	375	973	http://www	7.5
Zack Snyder	Batman v Superman: Da	Henry Cavill	250000000	330249062	Action Ad	2016	English	371639	673	3018	http://www	6.9
Bryan Singer	Superman Returns	Kevin Spacey	209000000	200069408	Action Ad	2006	English	240396	434	2367	http://www	6.1
Marc Forster	Quantum of Solace	Giancarlo Giannini	200000000	168368427	Action Ad	2008	English	330784	403	1243	http://www	6.7
Gore Verbinski	Pirates of the Caribbean	Johnny Depp	225000000	423032628	Action Ad	2006	English	522040	313	1832	http://www	7.3
Gore Verbinski	The Lone Ranger	Johnny Depp	215000000	89289910	Action Ad	2013	English	181792	450	711	http://www	6.5
Zack Snyder	Man of Steel	Henry Cavill	225000000	291021565	Action Ad	2013	English	548573	733	2536	http://www	7.2
Andrew Adamson	The Chronicles of Narnia	Peter Dinklage	225000000	141614023	Action Ad	2008	English	149922	258	438	http://www	6.6
Joss Whedon	The Avengers	Chris Hemsworth	220000000	623279547	Action Ad	2012	English	995415	703	1722	http://www	8.1
Rob Marshall	Pirates of the Caribbean	Johnny Depp	250000000	241063875	Action Ad	2011	English	370704	448	484	http://www	6.7
Barry Sonnenfeld	Men in Black 3	Will Smith	225000000	179020854	Action Ad	2012	English	268154	451	341	http://www	6.8
Peter Jackson	The Hobbit: The Battle o	Aidan Turner	250000000	255108370	Adventure	2014	English	354228	422	802	http://www	7.5
Marc Webb	The Amazing Spider-Man	Emma Stone	230000000	262030663	Action Ad	2012	English	451803	599	1225	http://www	7
Ridley Scott	Robin Hood	Mark Addy	200000000	105219735	Action Ad	2010	English	211765	343	546	http://www	6.7
Peter Jackson	The Hobbit: The Desolat	Aidan Turner	225000000	258355354	Adventure	2013	English	483540	509	951	http://www	7.9
Chris Weitz	The Golden Compass	Christopher Lee	180000000	70083519	Adventure	2007	English	149019	251	666	http://www	6.1
Peter Jackson	King Kong	Naomi Watts	207000000	218051260	Action Ad	2005	English	316018	446	2618	http://www	7.2
James Cameron	Titanic	Leonardo DiCaprio	200000000	658672302	Drama Ro	1997	English	793059	315	2528	http://www	7.7
Anthony Russo	Captain America: Civil W	Robert Downey Jr.	250000000	407197282	Action Ad	2016	English	272670	516	1022	http://www	8.2

2. Movies with highest profit are given below along with sql queries run to find them

The screenshot shows a PostgreSQL database interface. On the left is a sidebar with a tree view of the database structure. The 'public' schema is expanded, showing various objects like Aggregates, Collations, Domains, FTS Configurations, FTS Dictionaries, FTS Parsers, FTS Templates, Foreign Tables, Functions, Materialized Views, Operators, Procedures, Sequences, and Tables (1). The 'movies' table is selected under Tables (1).

The main window displays a SQL query in the 'Query' tab:

```
1
2 SELECT * FROM movies
3 ORDER BY profit DESC
4 LIMIT 10;
5
```

Below the query editor is the 'Data Output' tab, which shows the results of the query in a table format. The table has 10 columns: director_name, movie_title, actor_1_name, budget, gross, genres, and title_year. The results are ordered by profit in descending order, with the top 10 movies listed.

	director_name	movie_title	actor_1_name	budget	gross	genres	title_year
1	James Cameron	Avatar	CCH Pounder	237000000	760505847	Action Adventure Fantasy Sci-Fi	2009
2	Colin Trevorrow	Jurassic World	Bryce Dallas Howard	150000000	652177271	Action Adventure Sci-Fi Thriller	2015
3	James Cameron	Titanic	Leonardo DiCaprio	200000000	658672302	Drama Romance	1997
4	George Lucas	Star Wars: Episode IV - A New Hope	Harrison Ford	11000000	460935665	Action Adventure Fantasy Sci-Fi	1977
5	Steven Spielberg	E.T. the Extra-Terrestrial	Henry Thomas	10500000	434949459	Family Sci-Fi	1982
6	Joss Whedon	The Avengers	Chris Hemsworth	220000000	623279547	Action Adventure Sci-Fi	2012
7	Joss Whedon	The Avengers	Chris Hemsworth	220000000	623279547	Action Adventure Sci-Fi	2012
8	Roger Allers	The Lion King	Matthew Broderick	45000000	422783777	Adventure Animation Drama Family Musical	1994
9	George Lucas	Star Wars: Episode I - The Phantom Menace	Natalie Portman	115000000	474544677	Action Adventure Fantasy Sci-Fi	1999
10	Christopher Nolan	The Dark Knight	Christian Bale	185000000	533316061	Action Crime Drama Thriller	2008

At the bottom of the interface, a status bar indicates 'Total rows: 10 of 10', 'Query complete 00:00:00.000', and 'Ln 1, Col 1'.

3. IMDB top 250 movies

Casts

Catalogs

Event Triggers

Extensions

Foreign Data Wrappers

Languages

Publications

Schemas (1)

public

Aggregates

Collations

Domains

FTS Configurations

FTS Dictionaries

FTS Parsers

FTS Templates

Foreign Tables

Functions

Materialized Views

Operators

Procedures

Sequences

Tables (1)

movies

Columns

Constraints

Indexes

RLS Policies

Rules

Query

Query History

1

2

3

4

SELECT

movie_title,

imdb_score,

Rank

FROM

movies

WHERE

IMDb_Top_250 = TRUE

ORDER BY

Rank;

Data Output

Messages

Notifications

+

-

↕

↗

↘

↙

↚

↛

	movie_title text	imdb_score numeric	rank integer
1	The Shawshank Redemption	9.3	1
2	The Godfather	9.2	2
3	The Godfather: Part II	9	3
4	The Dark Knight	9	3
5	Fargo	9	3
6	12 Angry Men	8.9	4
7	The Lord of the Rings: The Return of the King	8.9	4
8	Pulp Fiction	8.9	4
9	Schindler's List	8.9	4
10	The Good, the Bad and the Ugly	8.9	4
11	The Lord of the Rings: The Fellowship of the Ring	8.8	5
12	Forrest Gump	8.8	5
13	It's Always Sunny in Philadelphia	8.8	5
14	Fight Club	8.8	5
15	Inception	8.8	5
16	Daredevil	8.8	5
17	Star Wars: Episode V - The Empire Strikes Back	8.8	5

Total rows: 1000 of 5009

Query complete 00:00:00.116

Ln 5, Col 1

This is list of non English movies in the top 250 list

The screenshot shows a database management interface. On the left is a sidebar with a tree view of database objects. The 'public' schema is expanded, and the 'movies' table is selected. The main area displays a SQL query in the 'Query' tab and its results in the 'Data Output' tab.

Query:

```
1 SELECT movie_title, language, imdb_score
2 FROM movies
3 WHERE IMDb_Top_250 = TRUE AND language != 'English'
4 ORDER BY imdb_score DESC;
```

Data Output:

	movie_title	language	imdb_score
1	The Good, the Bad and the Ugly	Italian	8.9
2	Seven Samurai	Japanese	8.7
3	Gomorrah	Italian	8.7
4	City of God	Portuguese	8.7
5	Spirited Away	Japanese	8.6
6	The Lives of Others	German	8.5
7	Airlift	Hindi	8.5
8	Children of Heaven	Persian	8.5
9	Samsara	None	8.5
10	Princess Mononoke	Japanese	8.4
11	Amélie	French	8.4
12	Oldboy	Korean	8.4
13	A Separation	Persian	8.4
14	Baahubali: The Beginning	Telugu	8.4
15	Rang De Basanti	Hindi	8.4
16	Das Boot	German	8.4
17	The Usual Suspects	English	8.3

Total rows: 324 of 324 Query complete 00:00:00.063 In 5, Col 1

4. Best director

The screenshot shows a database management interface. On the left is a sidebar with a tree view of database objects. The 'public' schema is expanded, showing various objects like Aggregates, Collations, Domains, FTS Configurations, FTS Dictionaries, FTS Parsers, FTS Templates, Foreign Tables, Functions, Materialized Views, Operators, Procedures, Sequences, Tables (1), and 'movies'. The 'movies' table is selected. The main area has a 'Query' tab with the SQL statement: `SELECT * FROM top_directors;`. Below the query is a 'Data Output' tab showing the results of the query. The results are displayed in a table with two columns: 'director_name' (text) and 'mean_imdb_score' (numeric). The table contains 10 rows of data. At the bottom, a status bar indicates 'Total rows: 10 of 10' and 'Query complete 00:00:00.126'.

	director_name	mean_imdb_score
1	John Blanchard	9.500000000000000
2	Cary Bell	8.700000000000000
3	Mitchell Albieri	8.700000000000000
4	Sadyk Sher-Niyaz	8.700000000000000
5	Charles Chaplin	8.600000000000000
6	Mike Mayhall	8.600000000000000
7	Damien Chazelle	8.500000000000000
8	Majid Majidi	8.500000000000000
9	Raja Menon	8.500000000000000
10	Ron Fricke	8.500000000000000

Total rows: 10 of 10 Query complete 00:00:00.126 Ln 1, Col 1

5. Finding the most popular genres

The screenshot shows a database management interface. On the left is a tree view of the database schema. The 'public' schema is expanded, showing various objects. The 'movies' table is selected. The main pane displays a SQL query in the 'Query' tab. The query is:

```
1 SELECT genres, COUNT(*) AS count
2 INTO popular_genres
3 FROM movies
4 GROUP BY genres
5 ORDER BY count DESC
6 LIMIT 10;
7
```

 Below the query editor are tabs for 'Data Output', 'Messages', and 'Notifications'. The 'Messages' tab is active, showing the message: 'Query returned successfully in 54 msec.' At the bottom, a status bar indicates 'Total rows: 10 of 10' and 'Query complete 00:00:00.054'. The bottom right corner shows 'Ln 6 Col 10'.

Query Query History

```
1 SELECT genres, COUNT(*) AS count
2 INTO popular_genres
3 FROM movies
4 GROUP BY genres
5 ORDER BY count DESC
6 LIMIT 10;
7
```

Data Output Messages Notifications

SELECT 10

Query returned successfully in 54 msec.

Total rows: 10 of 10 Query complete 00:00:00.054 Ln 6 Col 10

Foreign Data Wrappers

Languages

Publications

Schemas (1)

public

Aggregates

Collations

Domains

FTS Configurations

FTS Dictionaries

FTS Parsers

FTS Templates

Foreign Tables

Functions

Materialized Views

Operators

Procedures

Sequences

Tables (3)

movies

popular_genres

top_directors

Trigger Functions

Types

Views

Subscriptions

Login/Group Roles

Tablespaces

Query Query History

1 SELECT * FROM public.popular_genres

2

Data Output Messages Notifications

≡

▼

📄

🗑️

🔍

⬇️

📈

	genres	count
	text	bigint
1	Drama	236
2	Comedy	209
3	Comedy/Drama	191
4	Comedy/Drama/Romance	187
5	Comedy/Romance	158
6	Drama/Romance	152
7	Crime/Drama/Thriller	101
8	Horror	71
9	Action/Crime/Drama/Thriller	68
10	Action/Crime/Thriller	65

Total rows: 10 of 10 Query complete 00:00:00.160

✓ Successfully run. Total query runtime: 160 msec. 10 rows affected. ✕

6. Finding the critic-favorite and audience-favorite actors Created 3 tables found out who is most liked by public and the critics

The screenshot displays a database management interface. On the left, a tree view shows the database structure, with 'Tables (3)' expanded under the 'public' schema. The three tables listed are 'movies', 'popular_genres', and 'top_directors'. The main area shows a SQL query editor with the following code:

```
1 ALTER TABLE movies ADD COLUMN Meryl_Streep TEXT;  
2 ALTER TABLE movies ADD COLUMN Leo_Caprio TEXT;  
3 ALTER TABLE movies ADD COLUMN Brad_Pitt TEXT;  
4  
5 UPDATE movies SET Meryl_Streep = movie_title  
6 WHERE actor_1_name = 'Meryl Streep';  
7  
8 UPDATE movies SET Leo_Caprio = movie_title  
9 WHERE actor_1_name = 'Leonardo DiCaprio';  
10  
11 UPDATE movies SET Brad_Pitt = movie_title  
12 WHERE actor_1_name = 'Brad Pitt';  
13
```

Below the query editor, the 'Messages' tab is active, showing the execution status: 'UPDATE 18' and 'Query returned successfully in 48 msec.' The bottom status bar indicates 'Total rows: 10 of 10', 'Query complete 00:00:00.040', and 'Ln 13, Col 1'.

Query

Query History

```

1 ALTER TABLE movies ADD COLUMN decade TEXT;
2
3 UPDATE movies SET decade = CONCAT(SUBSTRING(CAST(title_year as text), 1, 3), '0s');
4
5 SELECT decade, SUM(num_voted_users) AS total_users
6 FROM movies
7 GROUP BY decade
8 ORDER BY decade;
9

```

Data Output

Messages

Notifications

	decade	total_users
	text	numeric
1	00s	3131768
2	1910s	10718
3	1920s	128672
4	1930s	984397
5	1940s	1211888
6	1950s	1638504
7	1960s	5153052
8	1970s	11312705
9	1980s	23176169
10	1990s	70633270
11	2000s	178354686
12	2010s	126202706

✓ Query returned successfully in 48 msec. ✕

Total rows: 12 of 12 Query complete: 00:00:00.000

Query

Query History

```

1 SELECT actor_1_name, AVG(num_critic_for_reviews) AS avg_critic_reviews, AVG(num_user_for_reviews) AS avg_user_reviews
2 FROM movies
3 GROUP BY actor_1_name
4 ORDER BY avg_critic_reviews DESC, avg_user_reviews DESC
5 LIMIT 5;
6

```

Data Output

Messages

Notifications

Languages

Publications

Schemas (1)

public

Aggregates

Collations

Domains

FTS Configurations

FTS Dictionaries

FTS Parsers

FTS Templates

Foreign Tables

Functions

Materialized Views

Operators

Procedures

Sequences

Tables (3)

movies

popular_genres

top_directors

Trigger Functions

Types

Views

Subscriptions

Login/Group Roles

Tablespaces

1 ALTER TABLE movies ADD COLUMN Combined TEXT;

2

3 UPDATE movies SET Combined = CONCAT(Meryl_Streep, ', ', Leo_Caprio, ', ', Brad_Pitt);

4

Data Output

Messages

Notifications

UPDATE 5043

Query returned successfully in 48 msec.

✓ Query returned successfully in 48 msec. ✕

Total rows: 10 of 10 Query complete 00:00:00.048 Ln 4, Col 1

RESULTS:- All the results obtained are listed above