BAIS:3200

Movie Database Final Report

Gavin Kafkakis, Joseph Kim, Jake Thomas, Ramsey Khayal, Tyler Tobin

Introduction

Project Overview:

Hollywood production studios are worth billions of dollars, and the actors who star in the films get paid millions for their work. For our project, we wanted to distill the elements of a successful blockbuster movie. Through close analysis of data gathered in the last five years, we came to conclusions surrounding specific questions we had about blockbuster success. We will look at movie ratings, revenue, actors, and directors to find out how much these factors impact making a blockbuster movie.

Data

### Data Description:

Our dataset came from Kaggle. It contains information about the top 10,000 movies listed on IMDB. IMDB is one of the world's largest databases containing information about movies, TV shows, and celebrities (Actors/Actresses/Directors). Additionally, the dataset has other attributes and features such as the movie's name, release year, rating, gross sales, and the directors/stars involved in the film.

Data Dictionary:

Attribute	Type	Description
MovieName	Text	Name of the Movie
YearOfRelease	Numeric	Year the movie came out
RunTimeInMinutes	Numeric	Total runtime of the movie
Rating	Numeric	IMDB user rating.
MetaScore	Numeric	Average rating from different critics
GrossIncome	Numeric	Gross income of the movie
Votes	Numeric	# of votes on IMDB
Genre	Text	Genre of the movie
Certification	Text	Certification or rating of the movie
Directors	Text	Directors of the movie
Actors	Text	Main cast or actors
Job	Text	Actor or Director

Conceptual Model, Entity Relationship Diagram (ERD):
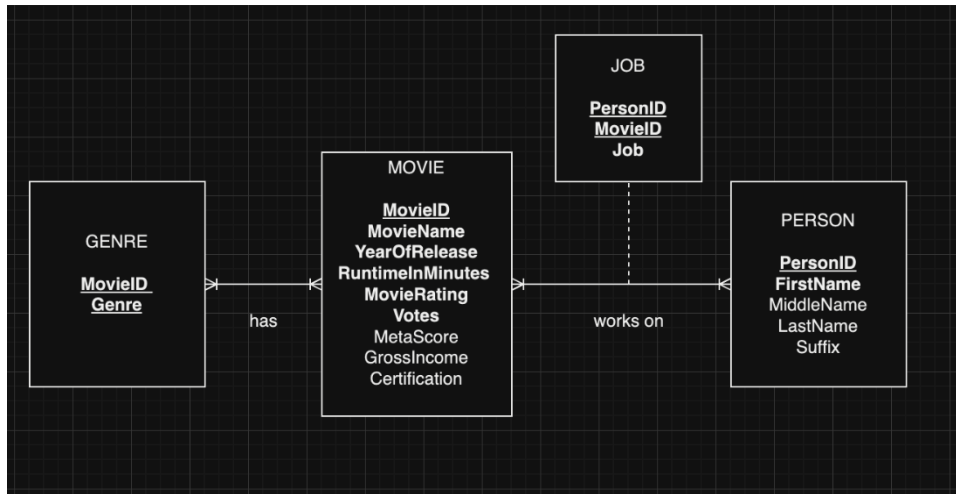
Logical Model: Graphical Relational Schema:
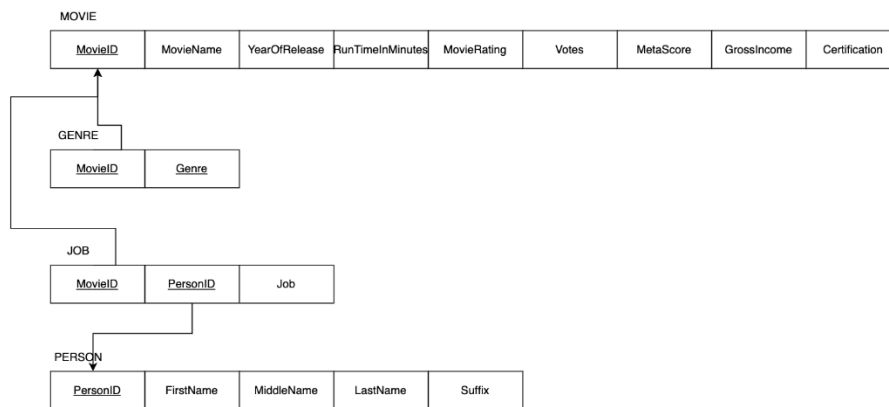
Data Cleaning Explanation:

When cleaning the data, we used a variety of functions in Excel to prepare the data to import into Oracle Apex.

First, we separated the concatenated lists using the text-to-column function in Excel using a delimiter of commas. The columns separated were genres, actors, and director's columns.

Next, we created a unique MovieID primary key for each movie and then associated the MovieID with multiple genres. We then sorted the columns alphabetically to remove the spaces between movies and then copied each different row to the bottom of the main columns. This ensured that the table was in its first normal form. This completed the Genre Table.

Next was the personal table. To ensure that we did not have multivariable attributes, we again used the text-to-column function in Excel using a delimiter of spaces. Text-to-column separated, first, middle, and last names and any suffixes like Jr. or II. We then set up a filter to filter the rows with 4 names, this would only include data that included first, middle, and last names and any suffixes. We then sorted the rows with one, two and three and separated the names into the proper columns. Any row with two names was a first and last name input, but because we delimited by spaces in the text-to-column function this put the last name into the middle name column. We moved the second name in a row of 2 to the correct last name designation.

After this, we sorted the names numerically to enable the VLOOKUP function. I used VLOOKUP to associate the correct MovieID and PersonID with the occupation, completing the Job table.

Database Implementation

SQL Commands:

To finally put our cleaned dataset into APEX Oracle we used the following CREATE TABLE codes:

<u>MOVIES:</u>

For our dataset we needed to create the MOVIES table first:

CREATE TABLE MOVIE (

  MovieID NUMBER(10) not null,

  MovieName VARCHAR2(150) not null,

  YearOfRelease number(4) not null,

  RunTimeInMinutes number(5) not null,

  MovieRating number(4, 2) not null,

  Votes number(20) not null,

  MetaScore varchar2(20),

  GrossIncome varchar2(20),

  Certification VARCHAR2(150),

  constraint movie_pk primary key (MovieID)

);

<u>Insert Command</u>

Insert into movie (movvieid, moviename, yearofrelease, runtimeinminutes, movierating, votes)

Values (00001, Daman, 2022, 121, 9, 13428) ;

<u>PERSON:</u>

Keeping up with our schema we built the PERSON table second:

CREATE TABLE PERSON (

  PersonID number(10) not null,

  FirstName VARCHAR2(150) not null,

  MiddleName VARCHAR2(150),

  LastName VARCHAR2(150),

  Suffix VARCHAR2(150),

  constraint person_pk primary key (PersonID)

);

Insert Command

Insert into person (personid, firstname, lastname)

Values (00001, 50, Cent) ;

GENRE:

Our third table created was the GENRE table:

CREATE TABLE GENRE (

   MovieID number(10) not null,

   Genre VARCHAR2(150) not null,

   constraint genre_pk primary key (MovieID, Genre),

   constraint movie_fk foreign key (MovieID)

   references MOVIE (MovieID)

);

Insert Command

Insert into genre (MovieID, genre)

Values (00132, Romance) ;

JOB:

CREATE TABLE JOB (

   PersonID number(10) not null,

   MovieID number(10) not null,

   Job VARCHAR2(150) check (Job in ('Actor', 'Director')) not null,

   constraint job_pk primary key (PersonID, MovieID),

   constraint person_fk foreign key (PersonID) references PERSON (PersonID),

   constraint job_fk foreign key (MovieID) references MOVIE (MovieID)

);

Insert Command

Insert into job (personid, movieid, job)

Values (03000, 00165, Actor) ;

Analysis

For our analysis we wanted to find how the features that we selected affected how successful a movie becomes. We wanted to find the best actors and directors within the last five years that made the best movies. Our goal was to find out what film culture considered to be most interesting in the past five years and see if we can determine the upcoming movie trends.

Question 1 – Genre: Which Genres have the highest average rating?

The genre of a movie is very important in determining its success. Pop culture's favorite genre has changed over the years, so it is important to understand what people are interested in to stay with the current movie trends. In this question we hope to find out what genre people liked the most and if there is a genre that is rising in popularity.

This is the following SQL query used to find this answer:

select genre, round(avg(movierating), 3) as averageRating

from genre join movie

on genre.movieid = movie.movieid

group by genre

order by averagerating desc

fetch first 10 rows only ;

GENRE	AVERAGERATING
Biography	7.016
Sport	6.979
War	6.975
Western	6.94
Animation	6.923
History	6.855
Music	6.812
Musical	6.777
Drama	6.77
Crime	6.688

The range of averages within the top 10 is very minimal, as all of the genres shown are within 0.5 of each other. This lack of disparity between each genre leads us to believe there may not be a strong correlation between a certain genre and a high rating. However, we do not have enough information to say for sure.

Question 2 – Actor/Genre: Which actor acted in the most action and drama movies?

With this question, we wanted to target a specific category just to see if they have any correlation. We thought action and drama were sometimes closely related. This question shows us if there is any correlation between these two genres and if there is a star that excels at making movies with those two genres

This is the following SQL query used to find this answer:

select person.personid, person.firstname || ' ' || person.middlename || ' ' || person.lastname || ' ' || person.suffix as FULLNAME , count(movie.movieid) as totalMovies

from movie join genre

on movie.movieid = genre.movieid

join job on movie.movieid = job.movieid join person

on person.personid = job.personid

where genre = 'Action' and job.job = 'Actor'

group by person.personid, person.firstname, person.middlename, person.lastname, person.suffix

union

select person.personid, person.firstname || ' ' || person.middlename || ' ' || person.lastname || ' ' || person.suffix as FULLNAME, count(movie.movieid) as totalMovies

from movie join genre

on movie.movieid = genre.movieid

join job on movie.movieid = job.movieid join person

on person.personid = job.personid

where genre = 'Drama' and job.job = 'Actor'

group by person.personid, person.firstname, person.middlename, person.lastname, person.suffix

order by totalMovies desc

fetch first 25 rows only ;

PERSONID	FULLNAME	TOTALMOVIES
02325	John Andreas Andersen	11
04319	Scoot McNairy	10
04896	Valerio Zanoli	10
03317	Michael Abbott Jr.	9
00129	Akshay Kumar	8
03238	Matteo Garrone	8
03000	Lukasz Kosmicki	8
02325	John Andreas Andersen	8
01321	Dwayne Johnson	8
01085	Dani Popescu	8
00051	Adam Driver	7
04798	Tom Hanks	7

After running our query, we found that the top 3 out of ten actors who have starred in both action and drama movies are John Andreas Anderson, Scoot McNairy, and Valerio Zanoli. This query shows that if a studio is interested in creating an action and drama movie, they should look into casting these actors because they have a lot of experience in creating and acting in these movies. Although we cannot know the quality of their performances just that they have a lot of experience working in these genres.

Question 3 – Movie: What is the movie with the highest gross income in the last five years? How does the rating of the movie compare to the gross income of the movie?

We wanted to find the best movie that came out from the last five years because that will show us what ultimately the trend was for the past five years. After finding this information, we can look at the directors and actors that starred in it to see if they had any impact on the movie. It is important to look at the past data to be able to formulate what future data might look like.

This is the following SQL query used to find this answer:

select movie.moviename, movierating, to_char(grossincome, '$999,999,999,999') as grossincome

from movie

where grossincome is not null

order by grossincome desc ;

MOVIENAME	MOVIERATING	GROSSINCOME
Avengers: Endgame	8.4	$858,370,000
Spider-Man: No Way Home	8.2	$804,750,000
Top Gun: Maverick	8.3	$718,730,000
Black Panther	7.3	$700,060,000
Avengers: Infinity War	8.4	$678,820,000
Avatar: The Way of Water	7.6	$659,680,000
Incredibles 2	7.6	$608,580,000
The Lion King	6.8	$543,640,000
Star Wars: The Rise Of Skywalker	6.4	$515,200,000
Frozen II	6.8	$477,370,000
Black Panther: Wakanda Forever	6.7	$453,720,000
Toy Story 4	7.7	$434,040,000

The result of our query showed us that the movie with the highest gross income in the last five years is Avengers: Endgame with $858,370,000. Avengers: Endgame also has the highest IMDB rating out of movies that have a gross income. We can see also that the highest grossing movies tend to have higher IMDB ratings, but we cannot assume that the two variables are correlated based on this analysis alone.

Question 4 – Actors/Actresses: Which are the top 10 actors that starred in the most movies?

Actors and actresses are the ones that bring these films to life, they are the ones that people ultimately buy tickets to come to see so they are one the most important components to creating a successful film. With this question, we seek to find out which actors people are using their hard-earned money to go see. This is important because we can see what actors can make the biggest impact and see if any rising stars are coming up.

This is the following SQL query used to find this answer:

select person.personid, person.firstname || ' ' || person.middlename || ' ' || person.lastname || ' ' || person.suffix as FULLNAME, count(movie.movieid) as totalMovies

from movie join job

on movie.movieid = job.movieid join person

on person.personid = job.personid

where job.job = 'Actor'

group by person.personid, person.firstname, person.middlename, person.lastname, person.suffix

order by totalMovies desc

fetch first 10 rows only ;

PERSONID	FULLNAME	TOTALMOVIES
02325	John Andreas Andersen	26
03317	Michael Abbott Jr.	18
00879	Chiwetel Ejiofor	18
01085	Dani Popescu	14
03000	Lukasz Kosmicki	14
04896	Valerio Zanoli	13
00129	Akshay Kumar	13
04319	Scoot McNairy	12
03238	Matteo Garrone	12
00515	Ayushmann Khurrana	11

This query tells us the top 10 actors that have starred in most movies. We found out that the top 3 out of 10 actors are John Andreas Anderson, Michael Abbott Jr, and Chiwetel Ejiofor. John Andreas Anderson also was first in our query in question 2. If a movie studio was interested in creating a film, they should look at these 10 actors because they have a lot of experience acting in movies. However, this query does not answer whether these actors would fit the role that is required but it is a good starting point to look for potential casting.

Question 5 – Directors: What are the average ratings of each of the top 10 directors?

Directors play a crucial role in making a movie. The actors might make the movie come alive, but directors are the ones who have the final say and are in control of the movie's artistic direction. We wanted to see what directors were making the best movies in the industry and what pop culture thinks about these directors' artistic views and creative style. We needed to filter out the directors with only one movie in our database because their average rating was not an average, but rather just the rating of that one movie, which heavily skewed our results.

This is the following SQL query used to find this answer:

select person.personid, person.firstname || ' ' || person.middlename || ' ' || person.lastname || ' ' || person.suffix as FULLNAME, round(avg(movierating),2) as averageMovieRating

from movie join job

on movie.movieid = job.movieid join person

on person.personid = job.personid

where job = 'Director'

having count(job.movieid) > 1

group by person.personid, person.firstname, person.middlename, person.lastname, person.suffix

order by averageMovieRating desc

fetch first 10 rows only ;

PERSONID	FULLNAME	AVERAGEMOVIERATING
01831	Haruo Sotozaki	8.63
04921	Vetrimaaran	8.4
04977	Vivek Agnihotri	8.35
03128	Mari Selvaraj	8.35
03860	Prashanth Neel	8.25
02953	Lokesh Kanagaraj	8
04309	Sashi Kiran Tikka	7.95
01728	Gowtam Tinnanuri	7.9
04601	Sukumar	7.9
04211	Saeed Roustayi	7.85

10 rows returned in 0.00 seconds   Download

After running our queries, we find that the majority of the directors that have the highest average rating are not from western culture. The top director with the highest average rating is Haruo Sotozaki, he is a Japanese director that main directs animated movies. The rest of the top 10 are all from the eastern culture mainly from India. We found out that Indian movie culture is a big market that can be investigated. From the results of this query, we should look deeper into the film culture of other countries.

Web Application

Link to Web Application:
https://apex.oracle.com/pls/apex/r/group_7_project/project-application/home?session=110664120379877

Web Application Creation:

In creating our web application, we took many steps to make a functional, appealing application that displays our findings deliberately and effectively. First, we disabled the log in page, to ensure that anybody could access our application without having to log in. We then designed our home page using HTML coding techniques to make headers, sections and an embedded link to our data source. Our last step for our home page was to add a picture that is relevant to our subject matter.

Our next step was to create pages with interactive reports for each of the tables we created. These pages can be easily accessed under the table section in the navigation bar. The implementation of these reports was relatively simple in that all we needed to do was use a query that selected all the data from each table and inserted it into an interactive report.

Finally, we added our question pages which included static tables for each and charts for two of the questions. We added brief descriptions of each of the questions and answers and utilized the underlining HTML function in these descriptions. To add the bar chart and scatterplot, we inserted the two queries into the chart and arranged the axes, labels, and title to satisfy our needs. The bar chart and scatterplot both visually illustrate what can be seen in the table next to them, which makes our data more interpretable for our users.

Home Page:
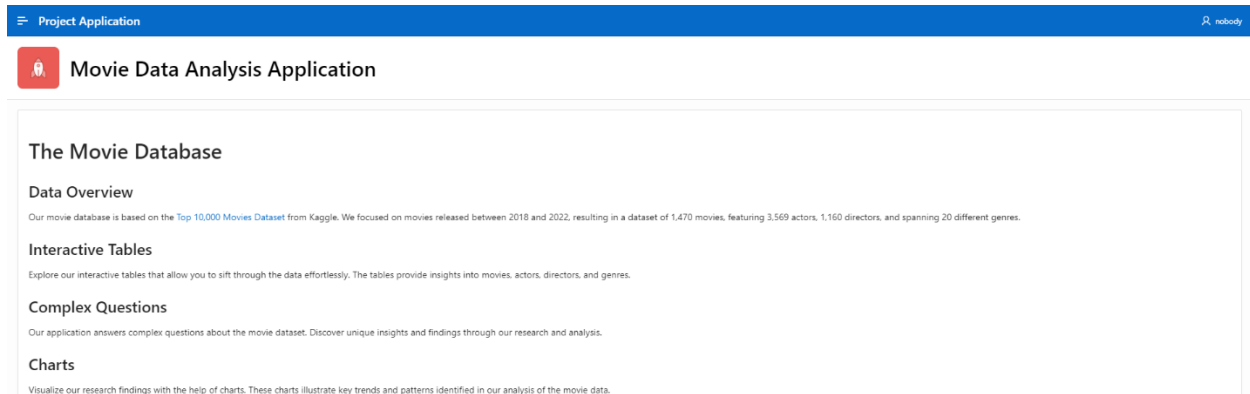
On the home page we included an overview of our website and the data we used to create the webpage. We outline what can be found on the other pages of the site.
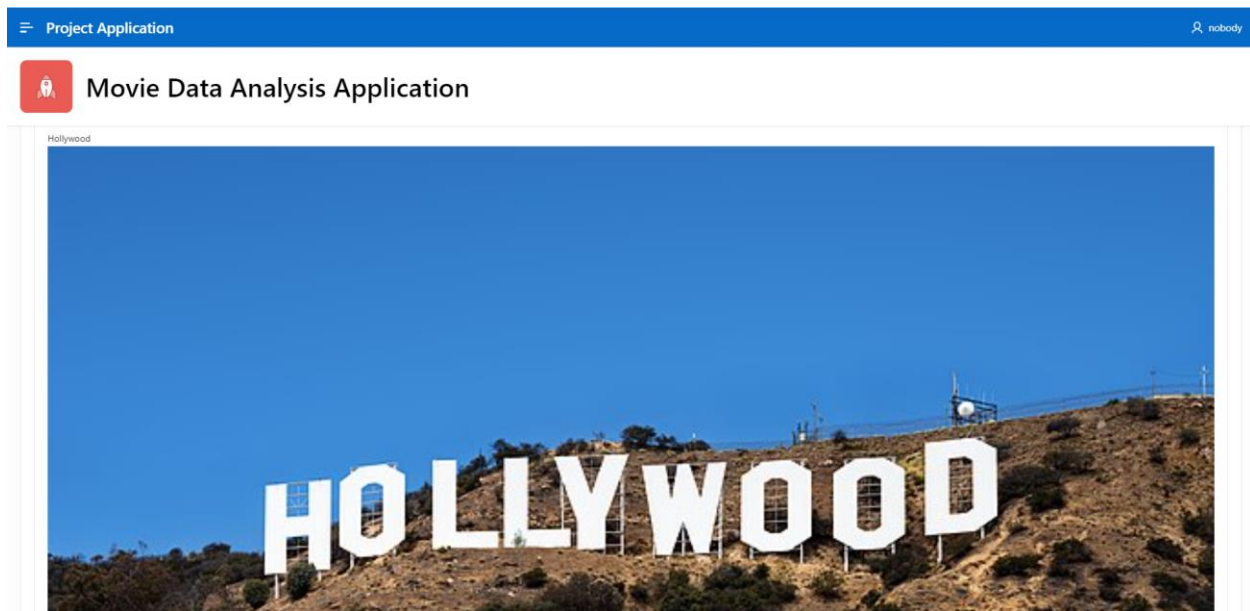
Additionally on the home page we added a picture of Hollywood to entice the reader to dive deeper into the site.

Hollywood is a staple landmark of famous actors and because out data being about actors, we thought it would engage our target audience.

We then added all of our table pages as interactive reports, so that the user can view all of the data and create filters that they would want to. We then added a little description of each of the tables outlining the number of rows and the different columns that are in each of the tables. Each table will have their own tab so that the user is able to make sure that they can see all of the tables that were used throughout the report.

Movie Table Page:

Movie Table

Movie ID	Movie Name	Year of Release	Runtime (Minutes)	Movie Rating	Votes	MetaScore	Gross Income	Certification
00015	Major	2022	150	8.1	31161			
00016	Hridayam	2022	172	8.1	14339			
00017	Karthikeya 2	2022	150	8	63779			
00018	Thiruchitrambalam	2022	131	7.9	12267			
00027	Gangubai Kathiawadi	2022	152	7.8	59551			TV-MA
00028	A Thursday	2022	128	7.7	24883			
00035	Bheeshma Parvam	2022	144	7.6	12532			
00036	Burning Days	2022	129	7.6	10175			
00039	Mahaan	2022	163	7.6	17086			
00040	Ponniyin Selvan: Part I	2022	167	7.6	35344			Not Rated
00041	Chup	2022	135	7.6	24227			Not Rated
00043	Sharmaji Namkeen	2022	119	7.5	10993			
00047	Dasvi	2022	125	7.5	56147			
00048	Khuda Haafiz Chapter 2 Agni Pariksha	2022	144	7.4	14683			
00049	Entergalactic	2022	93	7.4	13767			TV-MA
00050	Jhund	2022	176	7.4	25554			
00052	Freddy	2022	124	7.4	32713			
00055	Monica, O My Darling	2022	129	7.4	19768			TV-MA
00060	Jersey	2022	170	7.3	13999			
00062	She Said	2022	129	7.3	34803			R
00069	Rescued by Ruby	2022	90	7.2	10059			TV-G
00078	Vikram Vedha	2022	160	7.1	32721			Not Rated

Genre Table Page:

Genre Table

The GENRE table maps each movie to the one or many genres that it belongs to. It consists of 3,879 rows and maps each Movie ID to the genres that each movie is a part of. Every movie must have one genre but can also have many genres.

Movie ID	Genre
00132	Romance
00133	Action
00133	Adventure
00133	Drama
00134	Action
00134	Adventure
00134	Comedy
00135	Action
00135	Comedy
00135	Thriller
00136	Drama
00136	Thriller
00137	Comedy
00137	Drama
00138	Drama
00138	Horror
00138	Thriller
00139	Comedy

Person Table Page:

Job Table Page:

We then started to work on our pages for our questions. In our first question, we were attempting to figure out what the genres with the highest average ratings were. We added a description on the top that explained the question and an explanation of the result that we got for that question. After that we added in the table that showed the user all of the genres and what their average ratings were in a table format. As well as a bar graph that will show the user in a graph format the highest average rating for each of the top 10 genres from the highest to the lowest.

Question 1 Page:

Q-1

Question: Which genre has the highest average IMDB rating? What are the top 10 highest-rated genres on average?
Answer: The highest-rated genre on average is Biography. The second highest-rated genre is Sport and the third is War. The top 10 highest-averaging genres can be seen in the bar graph below. The range of averages within the top 10 is very minimal, as all of the genres shown are within 0.5 of each other. This lack of disparity between each genre leads us to believe there may not be a strong correlation between a certain genre and a high rating. However, we do not have enough information to say for sure.

Top 10 Genres

Genre	Average Rating
Biography	7.016
Sport	6.979
War	6.975
Western	6.94
Animation	6.923
History	6.855
Music	6.812
Musical	6.777
Drama	6.77
Crime	6.688

1 - 10

Graph of Genres With Highest Average IMDB

Genre	Average IMDB Rating
Biography	7.016
Sport	6.979
War	6.975
Western	6.940
Animation	6.923
History	6.855
Music	6.812
Musical	6.777
Drama	6.770
Crime	6.688
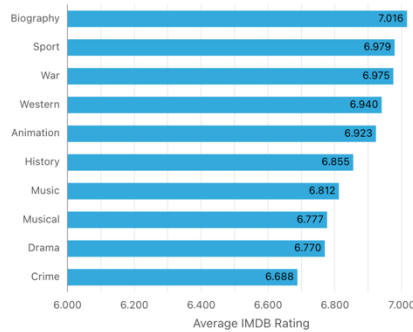
We then moved on to finding out the actors that acted in the most drama and action movies, as these are popular genres. We added the table on the right side that will show all of the actors in the top 25 so that the user is able to see each one and how many they acted in. As well as on the top of the page we added a description that tells the question and a quick explanation of the results that we ended up getting from this question.

Question 2 Page:

Q2 - Actors

Actors Acting in Drama and Action Genres

Question: Which actor acted in the most Drama and Action movies? Who else was in the top 25?
Answer: The actor that acted in the most drama and action movies was John Andreas Anderson. Then the following after that are the rest of the top 25 actors in these kinds of genres.

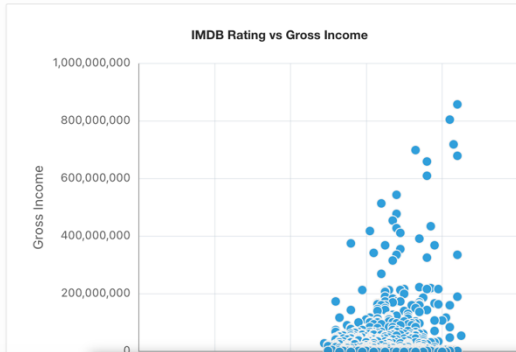| Person-ID	Actor Name	Total Action/Drama Movies
02325	John Andreas Andersen	11
04319	Scoot McNairy	10
04896	Valerio Zanoli	10
03317	Michael Abbott Jr.	9
00129	Akshay Kumar	8
03238	Matteo Garrone	8
03000	Lukasz Kosmicki	8
02325	John Andreas Andersen	8
01321	Dwayne Johnson	8
01085	Dani Popescu	8
00051	Adam Driver	7
04798	Tom Hanks	7
04184	Ryan Reynolds	7
03579	Nicole Kidman	7
03317	Michael Abbott Jr.	7

After that, we moved on to our next question; movies with the highest grossing income in the past 5 years. We put a description of the question and made sure to mention the filter that we had on the data, which is that we only tracked movies that had a gross income listed. Next, we added a database table that showed the gross income and IMDB ratings of the movies that had a gross income, in order of the movies with the highest gross income in the past 5 years. Finally, we added a scatter plot that would allow the user to interact with the data and see where each movie will show up when compared to rating and the amount of income it ended up getting.

Question 3 Page:

Q-3 - Movies

Question: What is the movie with the highest gross income in the last five years? How does the rating of the movie compare to the gross income of the movie? (Only includes movies with an available gross income)

Answer: The movie with the highest gross income in the last five years is Avengers: Endgame with $858,370,000. Avengers: Endgame also has the highest IMDB Rating out of movies that have a gross income. The rest of the movie's gross income and IMDB ratings can be seen in the table to the right ordered by gross income, or the scatter plot below.

	IMDB Rating	Gross Income
Avengers: Endgame	8.4	$858,370,000
Spider-Man: No Way Home	8.2	$804,750,000
Top Gun: Maverick	8.3	$718,730,000
Black Panther	7.3	$700,060,000
Avengers: Infinity War	8.4	$678,820,000
Avatar: The Way of Water	7.6	$659,680,000
Incredibles 2	7.6	$608,580,000
The Lion King	6.8	$543,640,000
Star Wars: The Rise Of Skywalker	6.4	$515,200,000
Frozen II	6.8	$477,370,000
Black Panther: Wakanda Forever	6.7	$453,720,000
Toy Story 4	7.7	$434,040,000
Captain Marvel	6.8	$426,830,000
Jurassic World: Fallen Kingdom	6.1	$417,720,000
Doctor Strange in the Multiverse of Madness	6.9	$411,330,000
Spider-Man: Far from Home	7.4	$390,530,000
Jurassic World Dominion	5.6	$376,850,000
Minions: The Rise of Gru	6.5	$369,700,000

IMDB Rating vs Gross Income

Our fourth question reached at the question of which actors in our database from the past couple of years ended up acting in the most movies. We added our question onto the top and then returned the table from the query as an interactive result in the page to showcase to the user the entire top 10 actors in most movies from this time period.

Question 4 Page:

Q-4 - Actors

Actors That Starred In The Most Movies

Question: Which actor starred in the most movies in the last 5 years? Who are the top 10?

Answer: John Andreas Andersen starred in 26 movies from 2018 through 2022, which was the most out of the actors in our database. Michael Abbott Jr. and Chiwetel Ejiofor tied for second with 18 movies each. The rest of the top 10 can be seen in the chart below.

Person ID	Full Name	Total Movies
02325	John Andreas Andersen	26
03317	Michael Abbott Jr.	18
00879	Chiwetel Ejiofor	18
01085	Dani Popescu	14
03000	Lukasz Kosmicki	14
04896	Valerio Zanoli	13
00129	Akshay Kumar	13
04319	Scoot McNairy	12
03238	Matteo Garrone	12
00515	Ayushmann Khurrana	11

Our final question and page looked at the directors with the highest average rating throughout our time period. We added our description onto the top and made sure to explain what our filter was on the query, which was that we were looking at directors that had more than 1 movie. After that, we added in our interactive report showcasing the top 10 directors that had the highest average rating for the time period.

Question 5 Page:

Q-5 - Directors

Directors With The Highest Average Rating

Question: Which director with more than 1 movie, has the highest average rating? Who else is in the top 10?

Answer: The director with the highest average rating is Haruo Sotozaki with 8.63. The director with the second-highest average rating is Vetrimaaran with 8.4. The rest of the top 10 are seen in the table to the right. We decided to filter out the directors who only directed 1 movie in our database because the results were very skewed when we included them.

Person ID	Full Name	Average Movie Rating
01831	Haruo Sotozaki	8.63
04921	Vetrimaaran	8.4
04977	Vivek Agnihotri	8.35
03128	Mari Selvaraj	8.35
03860	Prashanth Neel	8.25
02953	Lokesh Kanagaraj	8
04309	Sashi Kiran Tikka	7.95
01728	Gowtam Tinnanuri	7.9
04601	Sukumar	7.9
04211	Saeed Roustayi	7.85

Conclusion

In summation, the movie database and application that we have created explores and analyzes the trends of movies from the last 5 years. Starting with the dataset from Kaggle, we utilized our data cleaning skills to transform the data into 4 tables that we could query and analyze. We examined many variables such as a movies genre, director(s), and actors that worked on the movie. Our application has made this information presentable and interpretable by users who are curious about recent movie data.

Citations

Hollywood Image:
https://commons.wikimedia.org/w/index.php?search=hollywood&title=Special:MediaSearch&go=Go&type=image

Kaggle Data:
https://www.kaggle.com/datasets/dk123891/10000-movies-data?select=data.csv

Camera Icon:
https://commons.wikimedia.org/wiki/File:Video_Camera_Icon.svg