

IMDB Movie Analysis

Ami Rai E.

OUTLINE

- ✗ Introduction
- ✗ Project Description
- ✗ Approach
- ✗ Resource used
- ✗ Data Cleaning
- ✗ Structure of the table
- ✗ Insights and Results
- ✗ Conclusion

INTRODUCTION

- ✗ IMDb (an acronym for Internet Movie Database) is an online database of information related to films, television series, podcasts, home videos, video games, and streaming content online
- ✗ It is Launched online in 1990 and a subsidiary of Amazon

PROJECT DESCRIPTION

- ✗ The dataset having various columns of different IMDB Movies are provided
- ✗ The following tasks are to be done
 - + Clean the data
 - + Find the movies with the highest profit
 - + Find IMDB Top 250
 - + Find the best directors
 - + Find popular genres
 - + Find the critic-favorite and audience-favorite actors

APPROACH

- ✗ The steps
 - + Downloading the dataset
 - + Understanding data columns and data
 - + Data cleaning
 - + Data visualization

RESOURCE USED

- ✖ Microsoft Excel
 - + For analyzing and visualizing data



DATA CLEANING

- ✗ Missing values are found in many columns like director_name, gross, language, country, etc., when filter is applied on those columns
- ✗ Those rows are deleted from the data set
- ✗ Some columns like color, duration, content, etc., are not necessary for the data analysis
- ✗ Those columns are dropped from the data set

STRUCTURE OF THE TABLE

✗ The dataset consists of the columns

✗ The sample data

- director_name
- num_critic_for_reviews
- gross
- genres
- actor_1_name
- movie_title
- num_voted_users
- num_user_for_reviews
- language
- country
- budget
- title_year
- imdb_score

	director_name	num_critic_for_reviews	gross	genres	actor_1_name	movie_title	num_voted_users	num_user_for_reviews	language	country	budget	title_year	imdb_score
1	James Cameron	723	760505847	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	3054	English	USA	237000000	2009	7.9
2	Colin Trevorrow	644	652177271	Action Adventure Sci-Fi Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	USA	150000000	2015	7
3	James Cameron	315	658672302	Drama Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	200000000	1997	7.7
4	George Lucas	282	460935665	Action Adventure Fantasy Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	USA	11000000	1977	8.7
5	Steven Spielberg	215	434949459	Family Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	USA	10500000	1982	7.9
6	Joss Whedon	703	623279547	Action Adventure Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	USA	220000000	2012	8.1
7	Joss Whedon	703	623279547	Action Adventure Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	USA	220000000	2012	8.1

RESULTS

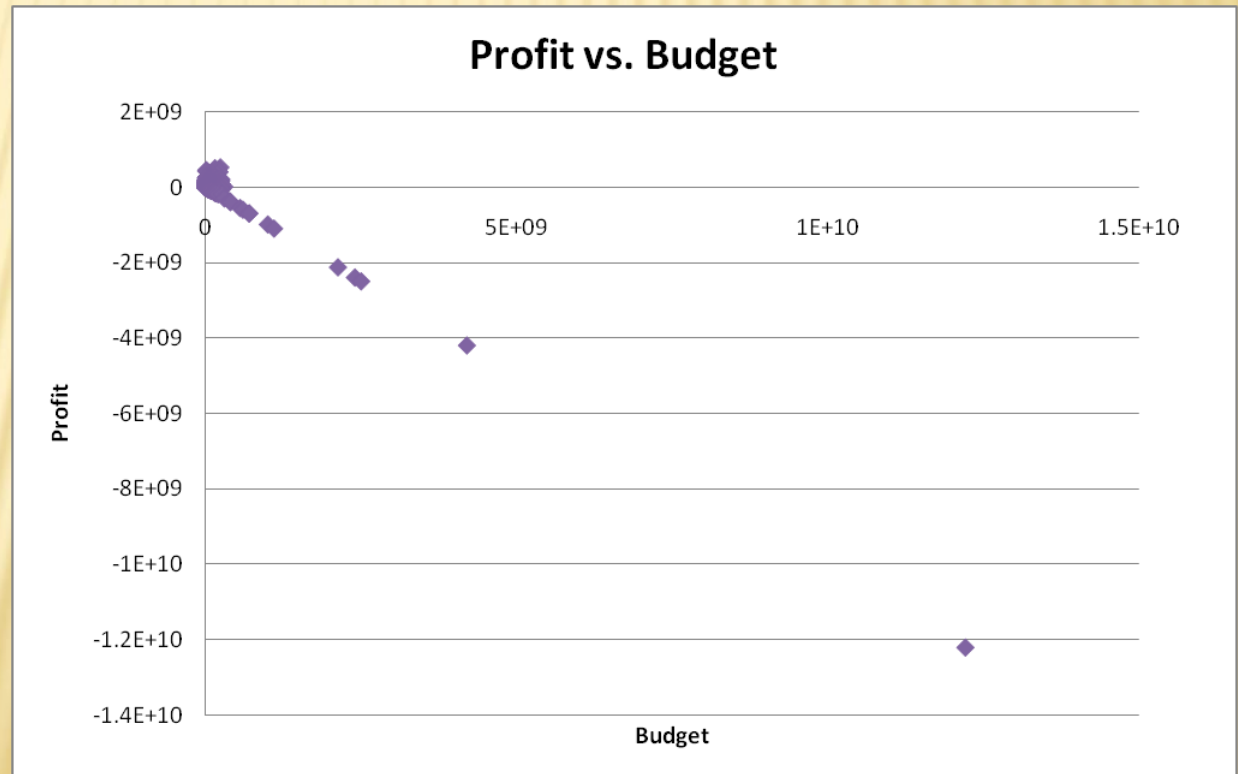
FINDING THE PROFIT

- ✗ A new column called profit which contains the difference of the two columns: gross and budget is created
- ✗ Sorted the column using the profit column as reference

	director_name	num_crit_ for_reviews	gross	genres	actor_1_name	movie_title	num_vot ed_users	num_user_f or_reviews	language	country	budget	title_y ear	imdb_ score	profit
1	James Cameron	723	760505847	Action Adventure Fantasy Sci-Fi	CCH Pounder	Avatar	886204	3054	English	USA	237000000	2009	7.9	523505847
2	Colin Trevorrow	644	652177271	Action Adventure Sci-Fi Thriller	Bryce Dallas Howard	Jurassic World	418214	1290	English	USA	150000000	2015	7	502177271
3	James Cameron	315	658672302	Drama Romance	Leonardo DiCaprio	Titanic	793059	2528	English	USA	200000000	1997	7.7	458672302
4	George Lucas	282	460935665	Action Adventure Fantasy Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New Hope	911097	1470	English	USA	11000000	1977	8.7	449935665
5	Steven Spielberg	215	434949459	Family Sci-Fi	Henry Thomas	E.T. the Extra-Terrestrial	281842	515	English	USA	10500000	1982	7.9	42449459
6	Joss Whedon	703	623279547	Action Adventure Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	USA	220000000	2012	8.1	403279547
7	Joss Whedon	703	623279547	Action Adventure Sci-Fi	Chris Hemsworth	The Avengers	995415	1722	English	USA	220000000	2012	8.1	403279547

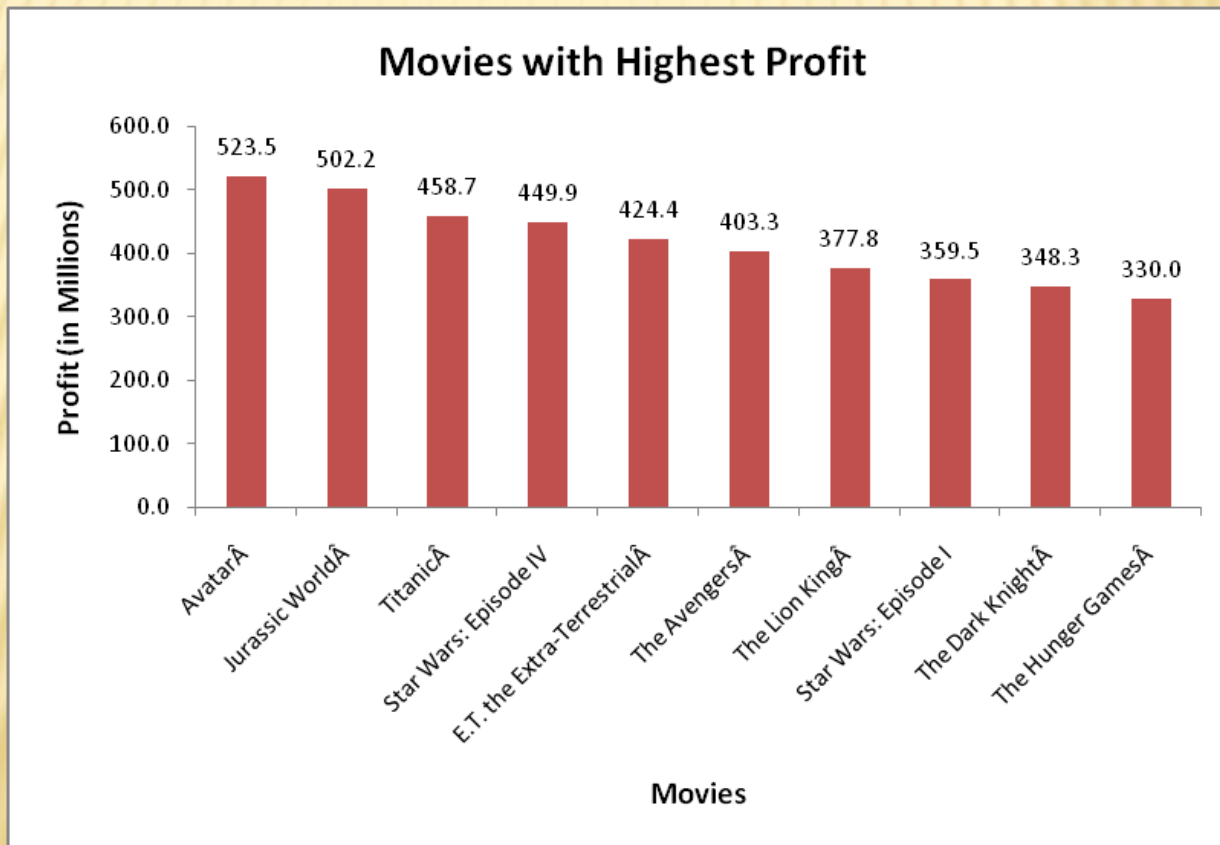
IDENTIFYING OUTLIERS

- ✗ A scatter plot with profit (y-axis) vs budget (x-axis) plotted
- ✗ One outlier is identified
- ✗ Not sure about if it is wrong data, so did not change the dataset



MOVIES WITH HIGHEST PROFIT - VISUAL

- ✗ Top 10 movies with highest profit is displayed



TOP 250 MOVIES

- ✖ Created a new column IMDb_Top_250
- ✖ Filtered the data with condition num_voted_users >25,000
- ✖ Sorted the columns based on the imdb_score
- ✖ Ranked the movies
- ✖ The top 250 movies with the highest IMDb Rating is identified

- ✖ A part of the result is displayed

Rank	imdb_score	IMDb_Top_250	language
1	9.3	The Shawshank Redemption	English
2	9.2	The Godfather	English
3	9	The Dark Knight	English
4	9	The Godfather: Part II	English
5	8.9	The Lord of the Rings: The Return of the King	English
6	8.9	Pulp Fiction	English
7	8.9	Schindler's List	English
8	8.9	The Good, the Bad and the Ugly	Italian
9	8.8	Forrest Gump	English
10	8.8	Star Wars: Episode V - The Empire Strikes Back	English
11	8.8	The Lord of the Rings: The Fellowship of the Ring	English
12	8.8	Inception	English
13	8.8	Fight Club	English
14	8.7	Star Wars: Episode IV - A New Hope	English
15	8.7	The Lord of the Rings: The Two Towers	English
16	8.7	The Matrix	English
17	8.7	One Flew Over the Cuckoo's Nest	English
18	8.7	Goodfellas	English
19	8.7	City of God	Portuguese
20	8.7	Seven Samurai	Japanese

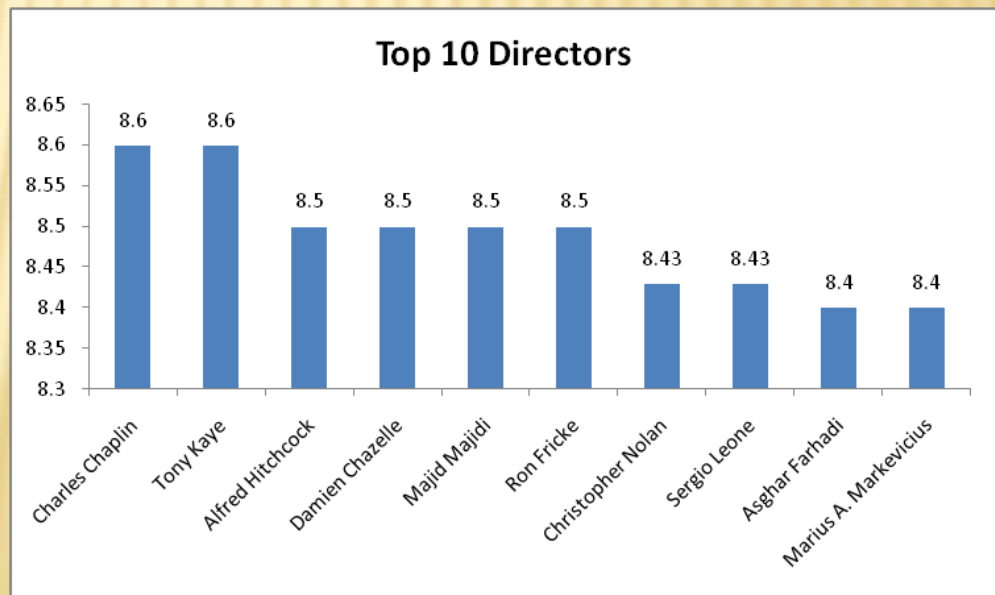
TOP FOREIGN LANGUAGE FILMS

- ✗ Extracted the movies in the IMDb_Top_250 column which are not English language and stored in another column
- ✗ Out of 250 movies, 33 are foreign languages
- ✗ part of the result is displayed

Top_Foreign_Lang_Film
The Good, the Bad and the Ugly
City of God
Seven Samurai
Spirited Away
The Lives of Others
Children of Heaven
Samsara
A Separation
Oldboy
Das Boot
Baahubali: The Beginning
Amélie
Princess Mononoke
The Hunt
Metropolis
Downfall
Pan's Labyrinth
The Secret in Their Eyes
Incendies
The Act of Killing
Howl's Moving Castle

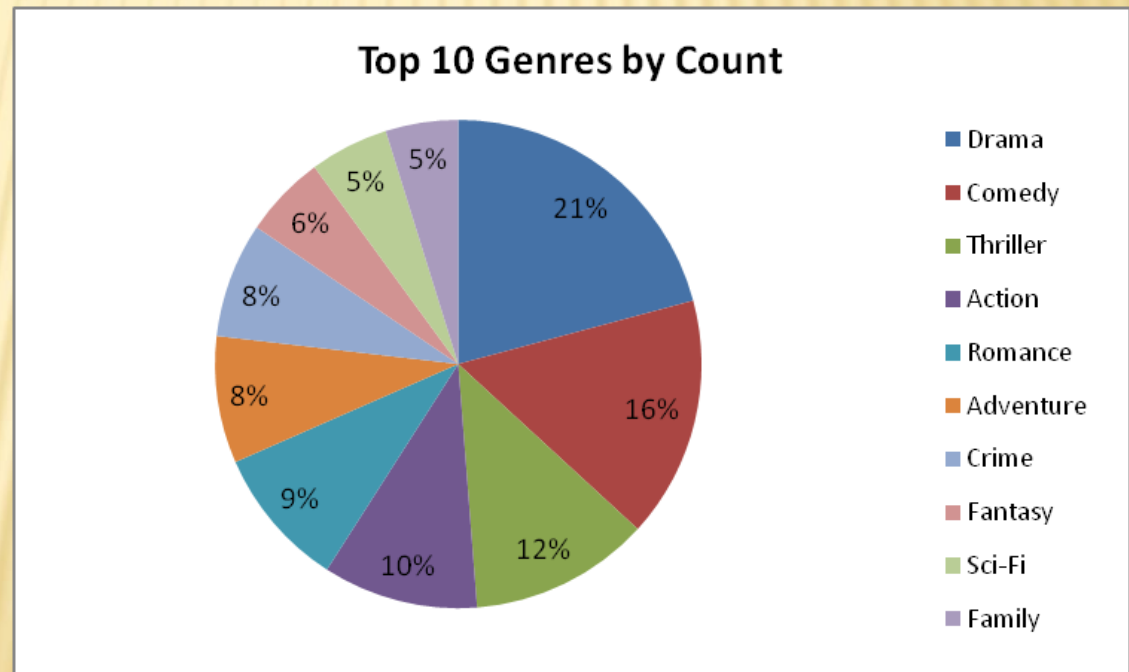
BEST DIRECTORS

- ✗ The names of directors are extracted from the director_name column (by removing duplicate elements)
- ✗ The mean of imdb_score of each director is calculated using AVERAGEIF() function
- ✗ Sorted the column with respect to imdb_score
- ✗ In case of a tie in IMDb score between two directors, names are sorted alphabetically



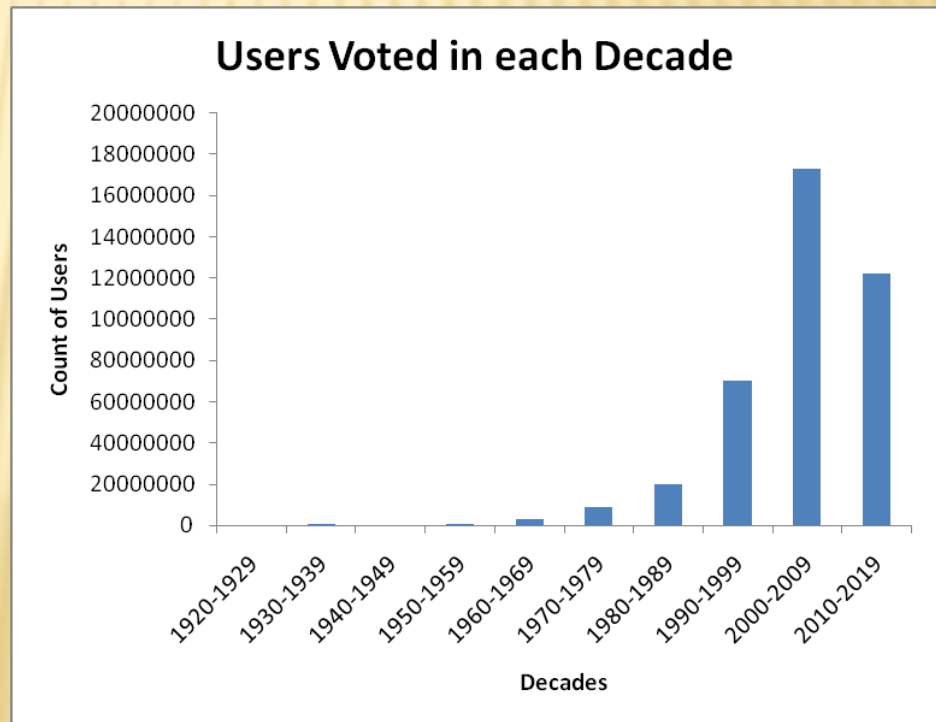
POPULAR GENRES

- ✗ The genres are separated using the 'Text to Column' option and the delimiter '|'
- ✗ The count of each genre is calculated



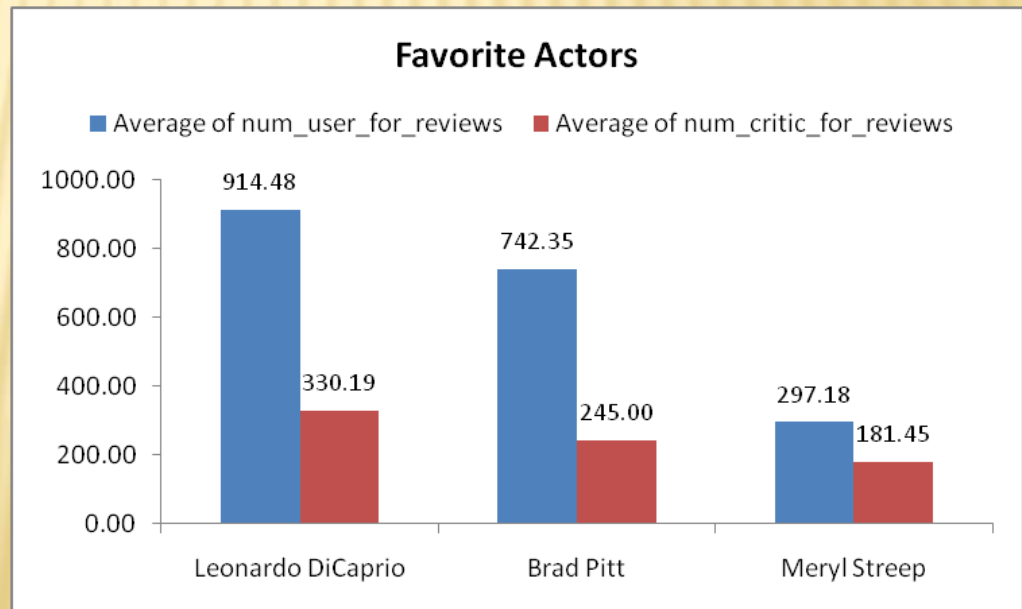
CHANGE IN NUMBER OF VOTED USERS

- ✗ Using the pivot table (Row Labels - title_year, Values - num_voted_users), the number of users voted in each year is calculated
- ✗ Grouped the years by decade (1920-1929, 1930-1939,..., 2010-2019)



FAVORITE ACTORS

- ✗ The data of following actors - Meryl_Streep, Leo_Caprio, and Brad_Pitt are taken for comparison
- ✗ The mean of the columns, num_critic_for_reviews and num_users_for_review are taken



CONCLUSION

- ✗ Avatar is the movie with the highest profit, 523.5 million
- ✗ The Shawshank Redemption is the movie with highest imdb_score of 9.3
- ✗ Out of top 250 movies, 33 are foreign languages
- ✗ Charlie Chaplin is the top director who got a average imdb_score of 8.6
- ✗ Out of 23 genres, Drama is the most popular genre
- ✗ There is increase in number of users voted from 1920 to 2016
- ✗ Leonardo DiCaprio is the most favorite actor among three actors

THANK YOU