IMDB Movie Analysis

Sujal Verma
Data Analytics Trainee

Overview

IMBD is a well-known movie and series rating site for users and critics worldwide. In IMBD we can find movies or series' ratings as well as its director's and actors compiled profiles as well as financials of it. Here we're provided with IMBD's dataset for movies from 1920-2010. Which contains information about the movie, its actors, directors, budget, collection, etc. We'll go to clean the dataset and get answers to asked questions by using the Five Why method for analytics using Office 365 Excel.

Approach

For this project, first, we'll get an understanding of the given data. Then We'll clean the data as per our requirement by removing null values, deleting unnecessary columns, etc. After the cleaning, we'll use a pivot table, various functions, and charts for desired answers to the questions. We'll continue to ask Whys to data to get in-depth of the root of the problem. In the end, we'll present our answers with proper formatting in tables and graphs.

Tech- Stack Used

For this IMBD Movie Analytics project, I used the Office 365 suite's Microsoft Excel. The Office 365 suite is a comprehensive collection of products offered by Microsoft Corporation. It is a productivity-focused suite that assists people and businesses in carrying out and managing a variety of daily tasks and data..

INSIGHTS

Cleaning the data: This is one of the most important steps to perform before moving forward with the analysis. Use your knowledge learned till now to do this. (Dropping columns, removing null values, etc.)

Your task: Clean the data.

Dropping unnecessary columns.

(Color, director_facebook_likes, actor_3_facebook_likes,

actor_2_name, actor_1_facebook_likes, cast_total_facebook_likes, actor_3_name, facenumber_in_posts, plot_keywords, movie_imdb_link, content_rating, actor_2_facebook_likes, aspect_ratio, movie_facebook_likes)

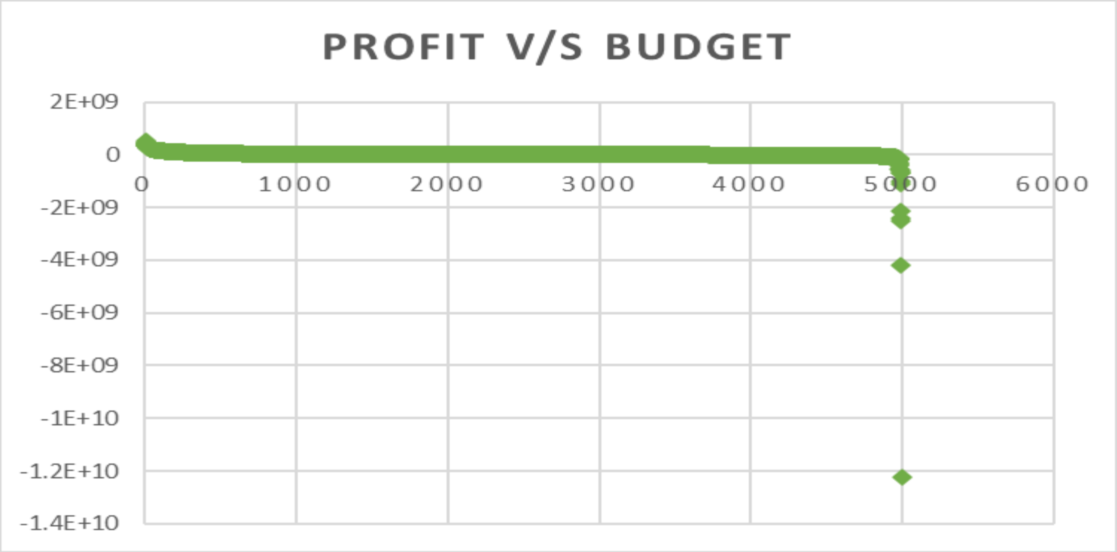
Remove Blank Cell / Null Value.

Removing Duplicate.

Movies with the highest profit: Create a new column called profit which contains the difference between the two columns: gross and budget. Sort the column using the profit column as a reference. Plot profit (y-axis) vs budget (x-axis) and observe the outliers using the appropriate chart type.

Your task: Find the movies with the highest profit.

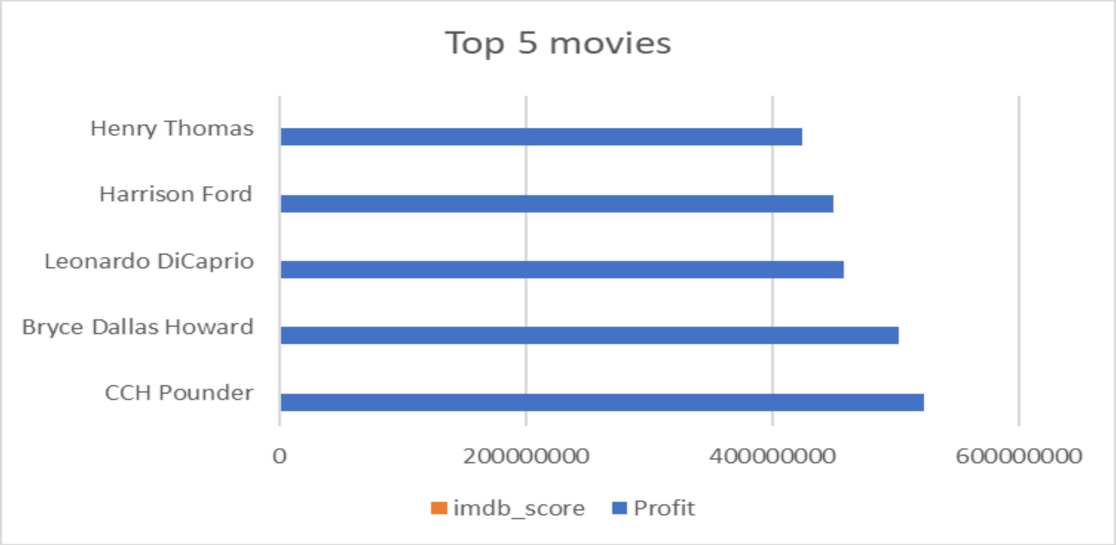
Profit Vs Budget



Top 5 Profitable Movies

director_name	genres	actor_1_name	Movie_name	language	Profit	imdb_score
James Cameron	Action|Adventure|Fantasy|Sci-Fi	CCH Pounder	AvatarÂ	English	523505847	7.9
Colin Trevorrow	Action|Adventure|Sci-Fi|Thriller	Bryce Dallas Howard	Jurassic WorldÂ	English	502177271	7
James Cameron	Drama|Romance	Leonardo DiCaprio	TitanicÂ	English	458672302	7.7
George Lucas	Action|Adventure|Fantasy|Sci-Fi	Harrison Ford	Star Wars: Episode IV - A New HopeÂ	English	449935665	8.7
Steven Spielber	Family|Sci-Fi	Henry Thomas	E.T. the Extra-TerrestrialÂ	English	42444945	7.9

g 9



C. Top 250 Movies:

Create a new column IMDb_Top_250 and store the top 250 movies with the highest IMDb Rating (corresponding to the column: imdb_score). Also make sure that for all of these movies, the num_voted_users is greater than 25,000. Also add a Rank column containing the values 1 to 250 indicating the ranks of the corresponding films.

Extract all the movies in the IMDb_Top_250 column which are not in the English language and store them in a new column named Top_Foreign_Lang_Film. You can use your own imagination also!

Your task: Find IMDB Top 250

Filter out data where num_voted_users > 25,000 using filter.

Sort the data using the imbd_score column in descending order.

Use first 250 entry for our analysis.

We'll give Rank using a Sequence Formula.

=SEQUENCE(COUNTA(G2:G251),1,1,1)

Filter out language by unselecting English. Which gives us foreign language movies in our Top 250 list.

5749

Top 250 Movies:

https://docs.google.com/spreadsheets/d/1jp2CEfhfAZuyCGyuYQ0P65SlZ2aOtb8D/edit?usp=sharing&ouid=113657493328297171808&rtpof=true&sd=true

Top 250 Foreign Language Movies:

https://docs.google.com/spreadsheets/d/1jp2CEfhfAZuyCGyuYQ0P65SlZ2aOtb8D/edit?usp=sharing&ouid=113657493328297171808&rtpof=true&sd=true

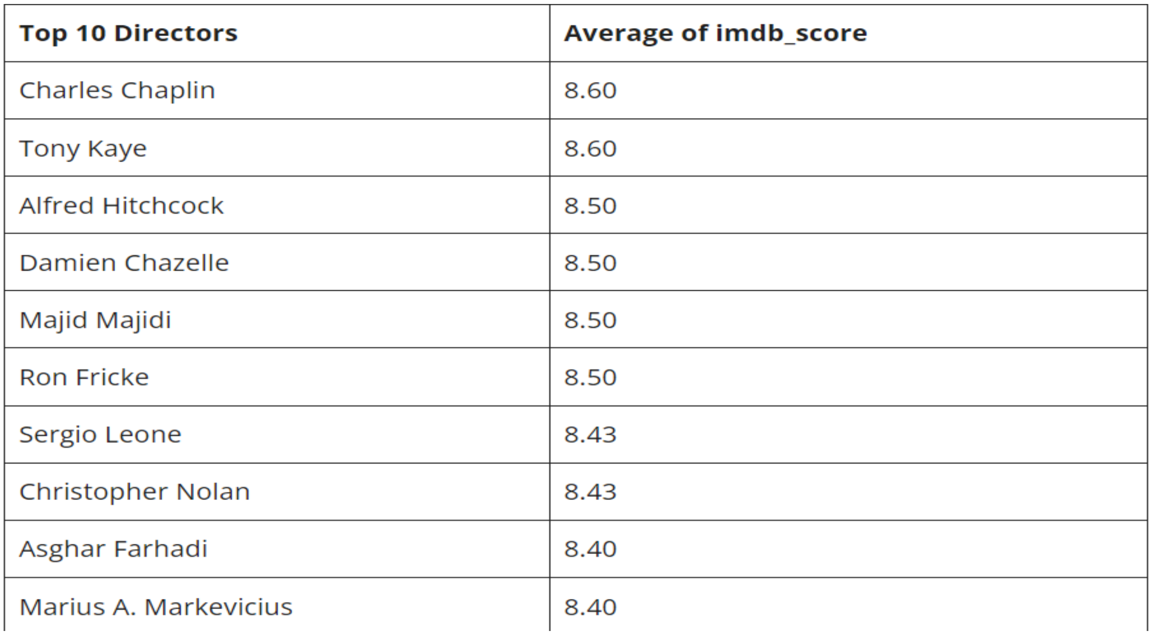
D.Top 10 Best Directors:

Group the column using the director_name column.

Find out the top 10 directors for whom the mean of imdb_score is the highest and store them in a new column top10director. In case of a tie in IMDb score between two directors, sort them alphabetically.

Your task: Find the best directors.

Using Pivot Table, Filter, and Sorting.



Top 10 Directors	Average of imdb_score
Charles Chaplin	8.60
Tony Kaye	8.60
Alfred Hitchcock	8.50
Damien Chazelle	8.50
Majid Majidi	8.50
Ron Fricke	8.50
Sergio Leone	8.43
Christopher Nolan	8.43
Asghar Farhadi	8.40
Marius A. Markevicius	8.40

Top 10 Directors

Top 10 directors:

https://docs.google.com/spreadsheets/d/1jp2CEfhfAZuyCGyuYQ0P65SlZ2aOtb8D/edit?usp=sharing&ouid=113657493328297171808&rtpof=true&sd=true

E. Popular Genres:

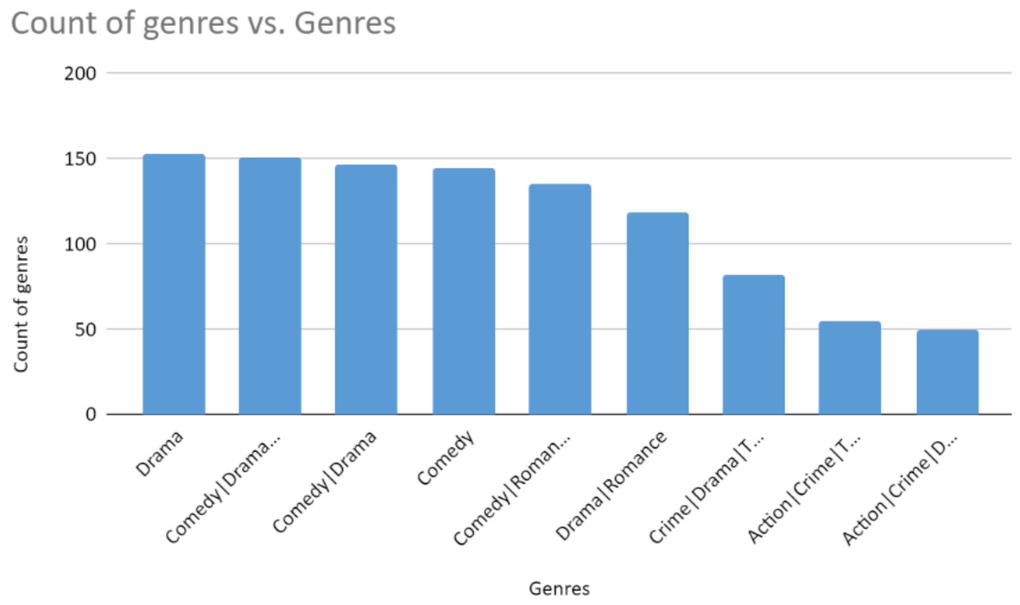
Perform this step using the knowledge gained while performing previous steps.

Your task: Find popular genres.

Using Pivot table, Filter, and Sorting.

Genres	Count of genres
Drama	153
Comedy|Drama|Romance	151
Comedy|Drama	147
Comedy	145
Comedy|Romance	135
Drama|Romance	119
Crime|Drama|Thriller	82
Action|Crime|Thriller	55
Action|Crime|Drama|Thriller	50

Most popular genres



We can see that Drama is most popular genre here.

https://docs.google.com/spreadsheets/d/1jp2CEfhfAZuyCGyuYQ0P65SlZ2aOtb8D/edit?usp=sharing&ouid=113657493328297171808&rtpof=true&sd=true

F.Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Charts: Create three new columns namely, Meryl_Streep, Leo_Caprio,and Brad_Pitt which contain the movies in which the actors: ‘Meryl Streep’, ‘Leonardo DiCaprio’, and ‘Brad Pitt’ are the lead actors. Use only the actor_1_name column for extraction. Also, make sure that you use the names ‘Meryl Streep’, ‘Leonardo DiCaprio’, and ‘Brad Pitt’ for the said extraction.

Append the rows of all these columns and store them in a new column named Combined.

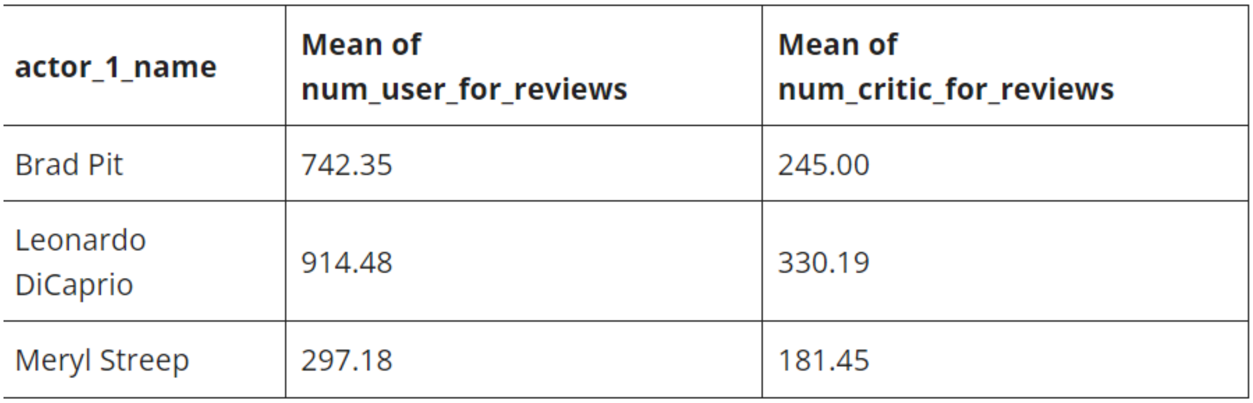
Group the combined column using the actor_1_name column.

Find the mean of the num_critic_for_reviews and num_users_for_review and identify the actors which have the highest mean.

Observe the change in number of voted users over decades using a bar chart. Create a column called decade which represents the decade to which every movie belongs to. For example, the title_year year 1923, 1925 should be stored as 1920s. Sort the column based on the column decade, group it by decade and find the sum of users voted in each decade. Store this in a new data frame called df_by_decade.

Your task: Find the critic-favorite and audience-favorite actors.

Using Pivot table.



actor_1_name	Mean of num_user_for_reviews	Mean of num_critic_for_reviews
Brad Pit	742.35	245.00
Leonardo DiCaprio	914.48	330.19
Meryl Streep	297.18	181.45

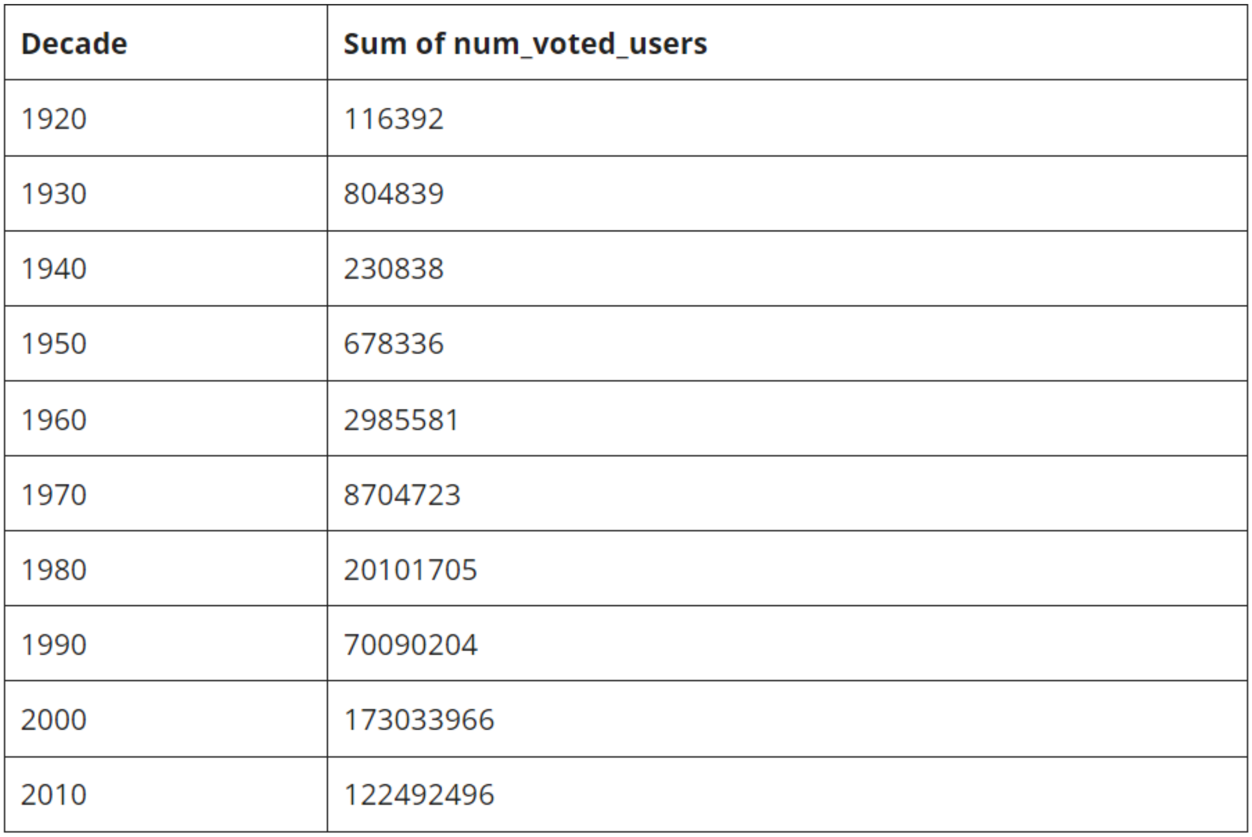
TOP 3 FAMOUS ACTORS

Mean of num_user_for_reviews Mean of num_critic_for_reviews

Here We can see that Leonardo DiCaprio is the audience's and Critic's favorite actor.

User voting by decade:

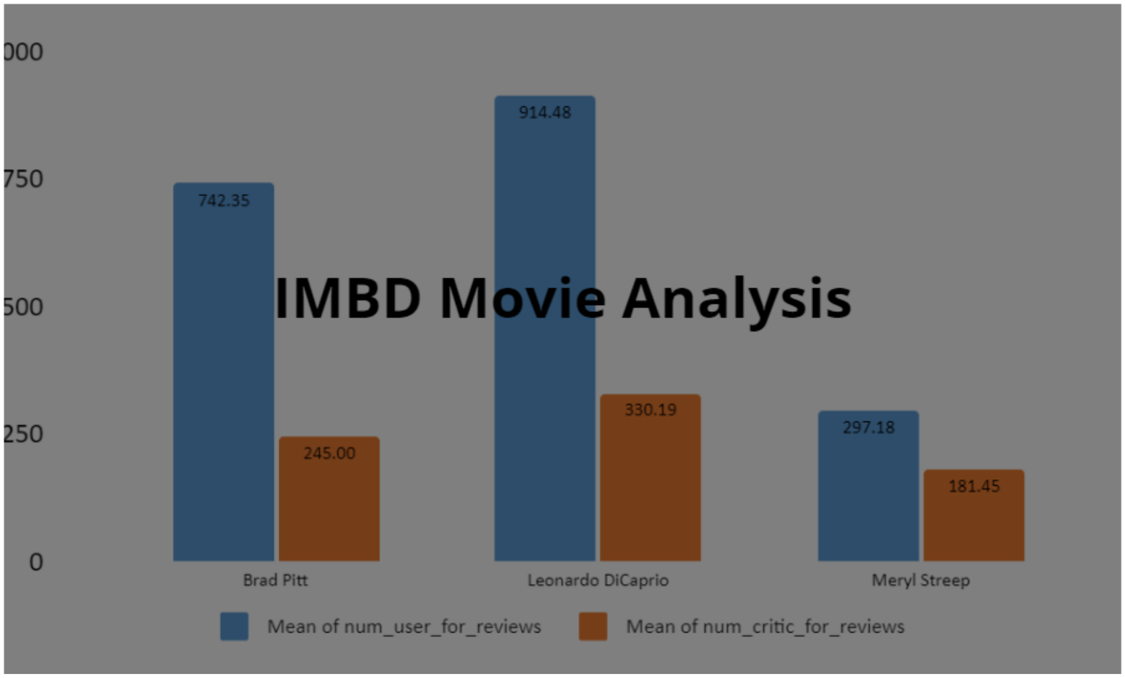
By using a pivot table.

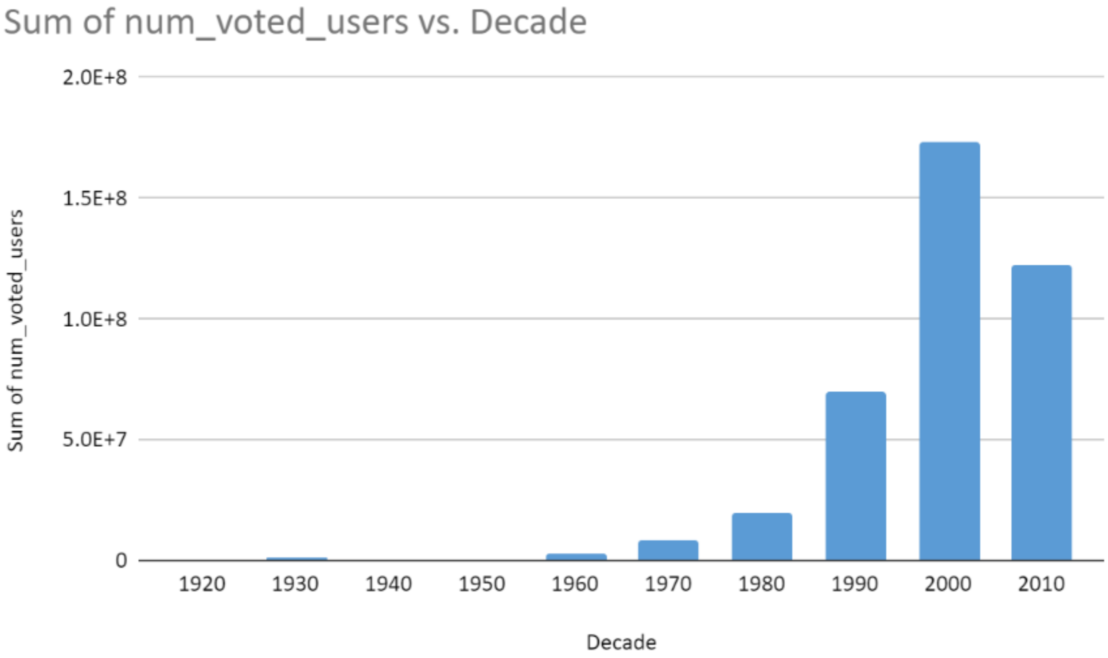


Decade	Sum of num_voted_users
1920	116392
1930	804839
1940	230838
1950	678336
1960	2985581
1970	8704723
1980	20101705
1990	70090204
2000	173033966
2010	122492496

INCREASE OF VOTED USERS

Chart:





https://docs.google.com/spreadsheets/d/1jp2CEfhfAZuyCGyuYQ0P65SlZ2aOtb8D/edit?usp=sharing&ouid=113657493328297171808&rtpof=true&sd=true

Result

In this project of IMBD Movie Analysis, I have gained various Logical, Statistics, and Technical Skills for get desired answers from the dataset.

Average, Frequency Table and Discovering Outliers are statistics concepts that help me better connect with data, offer me a thorough understanding of it, and aid in the analytics of supplied data.

My ability to apply statistics and Microsoft Excel's technical capabilities to analyze data. It speeds up data analytics tasks considerably. simplifies difficult and time-consuming calculations. I also get a sense of how the visual representation of data makes it very simple to understand through its data visualization functionality. I gain knowledge on when to utilize each visualization graph or chart based on the data and desired results.