IMDB Movie Analysis

- **Project**
- The project was focused on cleaning and analyzing movie data to uncover trends and insights in the film industry. It involved creating new columns, removing null values, grouping, sorting, and plotting the data to extract meaningful information. Through this project, I was able to practice my skills in data manipulation and analysis, as well as critical thinking and problem solving. Ultimately, I gained a deeper understanding of the movie industry and was able to identify key trends and patterns in film production.

Description

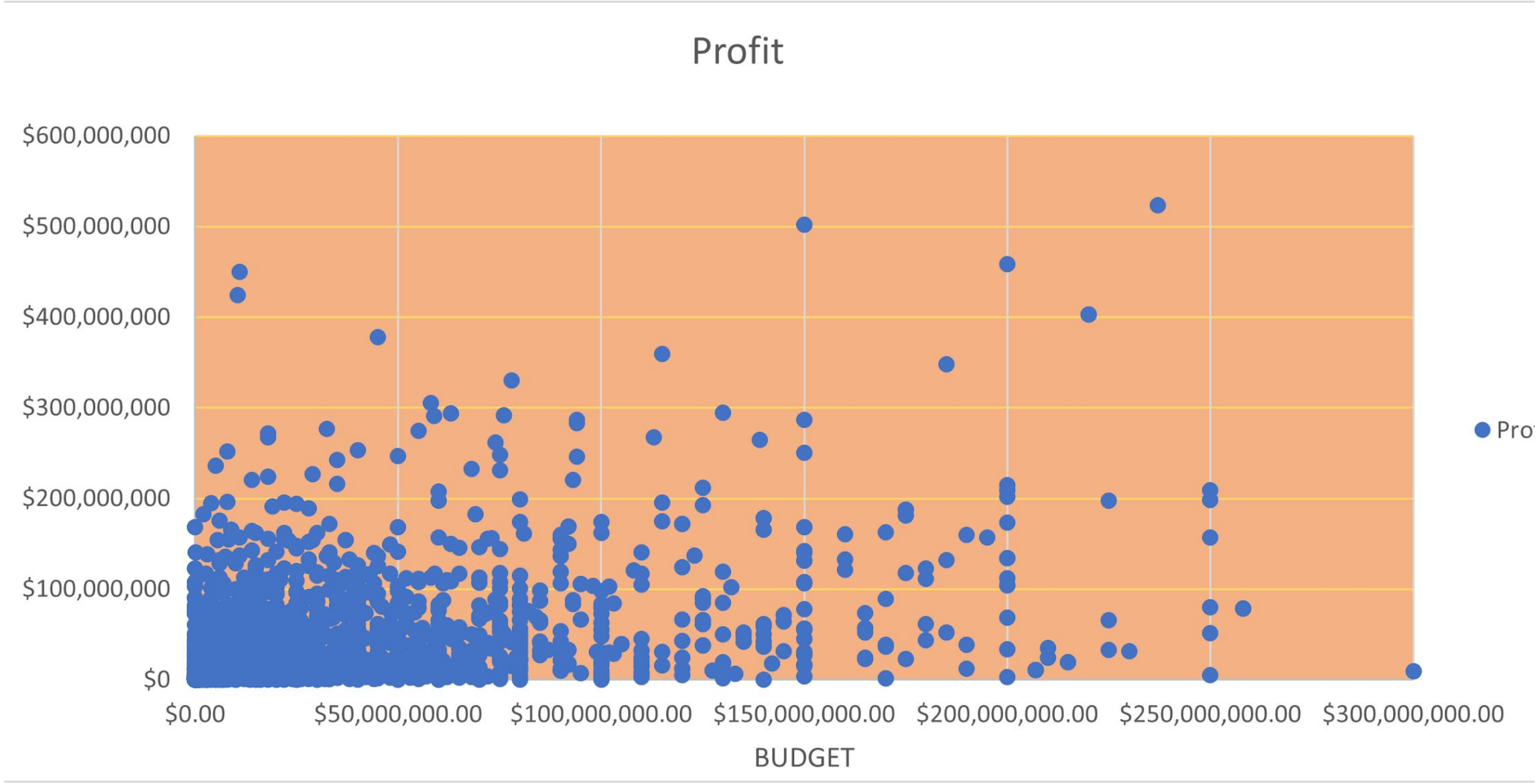
- Here we have an data set of IMDB
- We have a bundle of queries from the Users asked :
- Movies with highest profit
- Top 250 Movies Rated in IMDB
- Best 10 Directors
- Popular Genres watched by users
- Who the critic-favorite and audience-favorite actors

Cleaning the data

- Removed duplicate movies titles :
- **select movies column -> home ->conditional formatting -> data ->remove duplicates**
- **I categorize he data provided as numerical an categorical . For numerical where ever he data is no provided I will give it as 0 and similar to categorical data not provided**
- **Ctl+f5 -> go to special ->select blanks -> enter NP in column which is highlighted and Clt + enter**
- **Same for numerical data but enter 0**
- **Now there is no null data in dataset**

Movies with highest profit

- To create an column called profit
- Use this formula =(gross column -budget column) then copy that row and select the Entier column.
- Select the profit column and apply sort function in editing section from highest to lowest.
- Now select the budget and profit column and click on insert then select the graph need in our case we need to draw highest profit movie outliers


Profit
$600,000,000
$500,000,000
$400,000,000
$300,000,000
$200,000,000
$100,000,000
$0
$0.00
$50,000,000.00
$100,000,000.00
$150,000,000.00
$200,000,000.00
$250,000,000.00
$300,000,000.00
BUDGET
● Pro

Top 3 profitable movies

Avatar
Jurassic world
Titanic

Find IMDB Top 250

- Filter the data: First, you can filter the data to only include movies with num_voted_users greater than 25,000. This can be done using a conditional statement (e.g. if num_voted_users > 25000).
- Sort the data: Next, you can sort the filtered data in descending order based on the imdb_score column.
- Add the "IMDb_Top_250" column: After sorting the data, you can add a new column called "IMDb_Top_250" and store the top 250 movies with the highest IMDb Rating. This can be done by using slicing and assigning the values 1 to 250 to the first 250 rows of the "IMDb_Top_250" column.
- Add the "Rank" column: Finally, you can add another column called "Rank" and store the values 1 to 250 in this column to indicate the rank of each movie.

IMDB Top 250

- Half Past DeadÂ
- Hard to Be a GodÂ
- TranscendenceÂ
- The Wendell Baker StoryÂ
- No Strings AttachedÂ
- SparklerÂ
- Sound of My VoiceÂ
- Captain PhillipsÂ
- ForsakenÂ
- The Cat in the HatÂ

InceptionÂ	**Blood and WineÂ**	**The HauntingÂ**	**Sin City: A Dame to Kill ForÂ**	**An Ideal HusbandÂ**
Jurassic Park IIIÂ	**Due DateÂ**	**TrustÂ**	**The GiverÂ**	**Love's Abiding JoyÂ**
P.S. I Love YouÂ	**NovemberÂ**	Apollo 13Â	SelmaÂ	The Life Before Her EyesÂ
Blood Done Sign My NameÂ	Banshee ChapterÂ	Lake MungoÂ	They Came TogetherÂ	All HatÂ
Bad SantaÂ	2 Fast 2 FuriousÂ	Me Before YouÂ	The Dark Knight RisesÂ	The MaskÂ

True LiesÂ

Casino RoyaleÂ

Baggage ClaimÂ

The Expendables 2Â

American Pie 2Â	Die Hard with a VengeanceÂ	The Man Who Knew Too LittleÂ	Atlas Shrugged: Who Is John Galt?Â	Made of HonorÂ
FreewayÂ	Man on WireÂ	The Tooth FairyÂ	The Five-Year EngagementÂ	House at the End of the DriveÂ
Midnight in the Garden of Good and EvilÂ	AdoreÂ	Big Mommas: Like Father, Like SonÂ	Halloween IIÂ	How She MoveÂ
WALLÂ·EÂ	Fight to the FinishÂ	A Farewell to ArmsÂ	Air BudÂ	GrabbersÂ
AddictedÂ	CargoÂ	Jesus' SonÂ	Love StinksÂ	Alien 3Â

The Cry of the	OwlÂ JawbreakerÂ	Freddy Got FingeredÂ	Spy Kids 2: Island of Lost DreamsÂ	NeighborsÂ
Grace of MonacoÂ	PandorumÂ	SurvivorÂ	Spring BreakersÂ	K-PAXÂ
The Perfect ManÂ	UpÂ	GossipÂ	Batman BeginsÂ	Lady in the WaterÂ
ValentineÂ	Youth in RevoltÂ	Malcolm XÂ	Seeking a Friend for the End of the WorldÂ	SharkskinÂ
Good Luck ChuckÂ	Into the WildÂ	Travelers and MagiciansÂ	The Land Before TimeÂ	The Man from EarthÂ

SpanglishÂ	Hard CandyÂ	ThirteenÂ	The Black StallionÂ	Hustle & FlowÂ
BacheloretteÂ	M*A*S*HÂ	Urban LegendÂ	21 & OverÂ	Mao's Last DancerÂ
Lara Croft: Tomb RaiderÂ	The Original Kings of ComedyÂ	Captain America: Civil WarÂ	Jack and JillÂ	Hocus PocusÂ
How to Train Your DragonÂ	GoodÂ	The Midnight Meat TrainÂ	CollegeÂ	The Blood of My BrotherÂ
Captain Alatriste: The Spanish MusketeerÂ	Nim's IslandÂ	Palo AltoÂ	QÂ	Boyz n the HoodÂ

Party MonsterÂ	Robin Hood: Prince of ThievesÂ	King's RansomÂ	FlickaÂ	The End of the AffairÂ
In the Heat of the NightÂ	The AvengersÂ	Dark WaterÂ	The Ridiculous 6Â	3000 Miles to GracelandÂ
Spy Kids 3-D: Game OverÂ	The CroodsÂ	The Da Vinci CodeÂ	Guardians of the GalaxyÂ	TombstoneÂ
The Hit ListÂ	Town & CountryÂ	Kung Fu PandaÂ	Cats & DogsÂ	Just Go with ItÂ
The Perfect WaveÂ	Bridge of SpiesÂ	Lee Daniels' The ButlerÂ	Surfer, DudeÂ	HighwayÂ
	The TimberÂ	Anacondas: The Hunt for the Blood OrchidÂ	Coyote UglyÂ	

Knockaround GuysÂ

Brooklyn's FinestÂ

VaaluÂ

Mad Max:
Fury RoadÂ

Beasts of
the Southern WildÂ

WitchboardÂ

Two Girls and
a GuyÂ

Love in the Time
of MonstersÂ

Liar LiarÂ

Pocketful
of MiraclesÂ

Just LookingÂ

SamsaraÂ

TrainspottingÂ

Miss MarchÂ

BrainDeadÂ

Paris, je t'aimeÂ

The Longest DayÂ

Dolphin Tale 2Â

Young AdultÂ

Best Directors

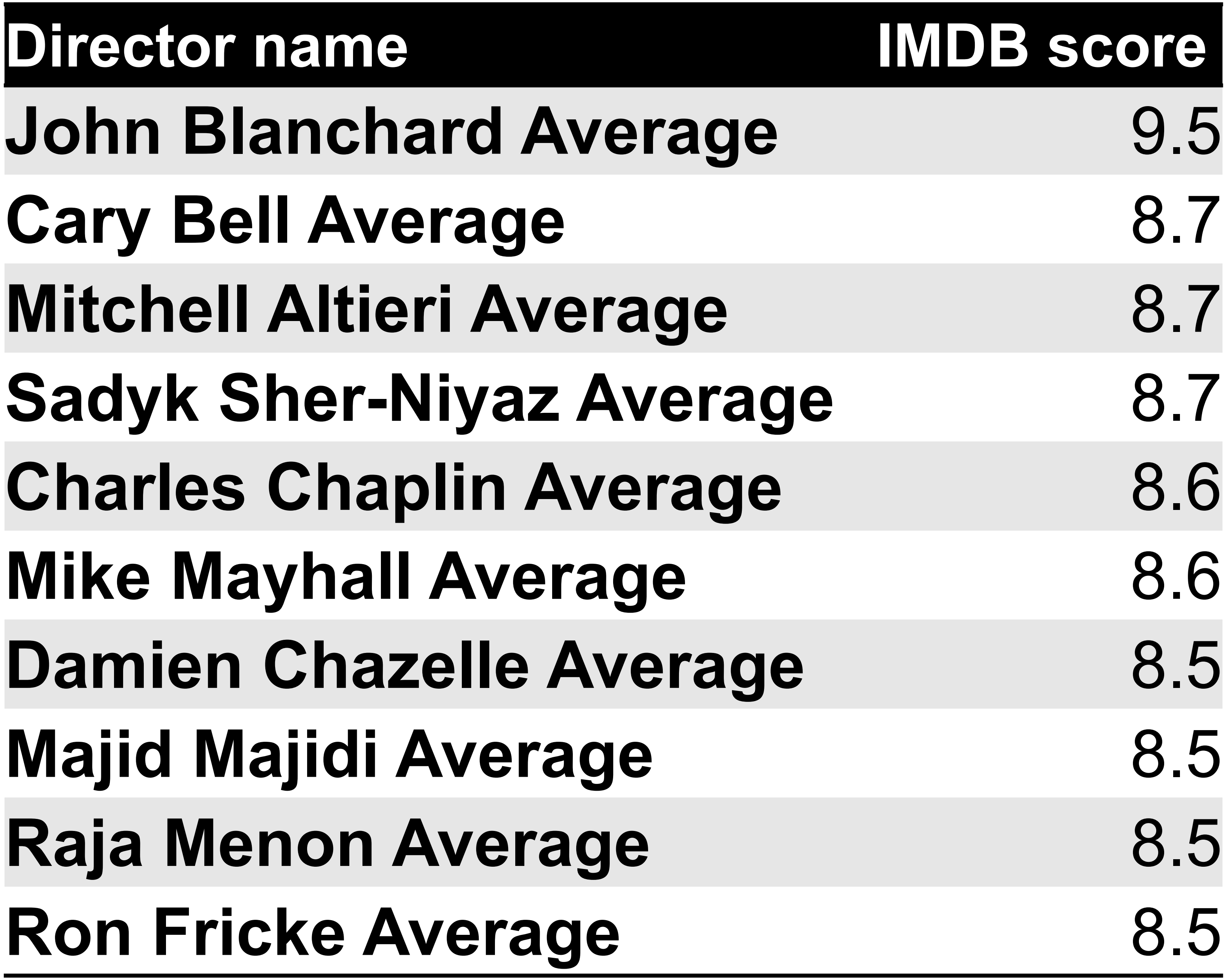
Select director and imdb score and create an pivort table

Group the data: First, group the data by the "director_name" column. This can be done using the "Group By" feature in Excel (Data > Sort & Filter > Advanced).

Calculate the mean IMDb score: Next, for each group (i.e. each director), calculate the mean of the "imdb_score" column. This can be done using the AVERAGE function in Excel.

Sort the directors: After calculating the mean IMDb score for each director, sort the groups based on the mean IMDb score in descending order. In case of a tie, sort the directors alphabetically by their names.

Select the top 10 directors: Finally, select the top 10 directors and store them in a new column called "top10director". This can be done by using the "Top 10" feature in Excel (Data > Sort & Filter > Top 10) or by manually copying and pasting the top 10 directors into a new column.



Director name	IMDB score
John Blanchard Average	9.5
Cary Bell Average	8.7
Mitchell Altieri Average	8.7
Sadyk Sher-Niyaz Average	8.7
Charles Chaplin Average	8.6
Mike Mayhall Average	8.6
Damien Chazelle Average	8.5
Majid Majidi Average	8.5
Raja Menon Average	8.5
Ron Fricke Average	8.5

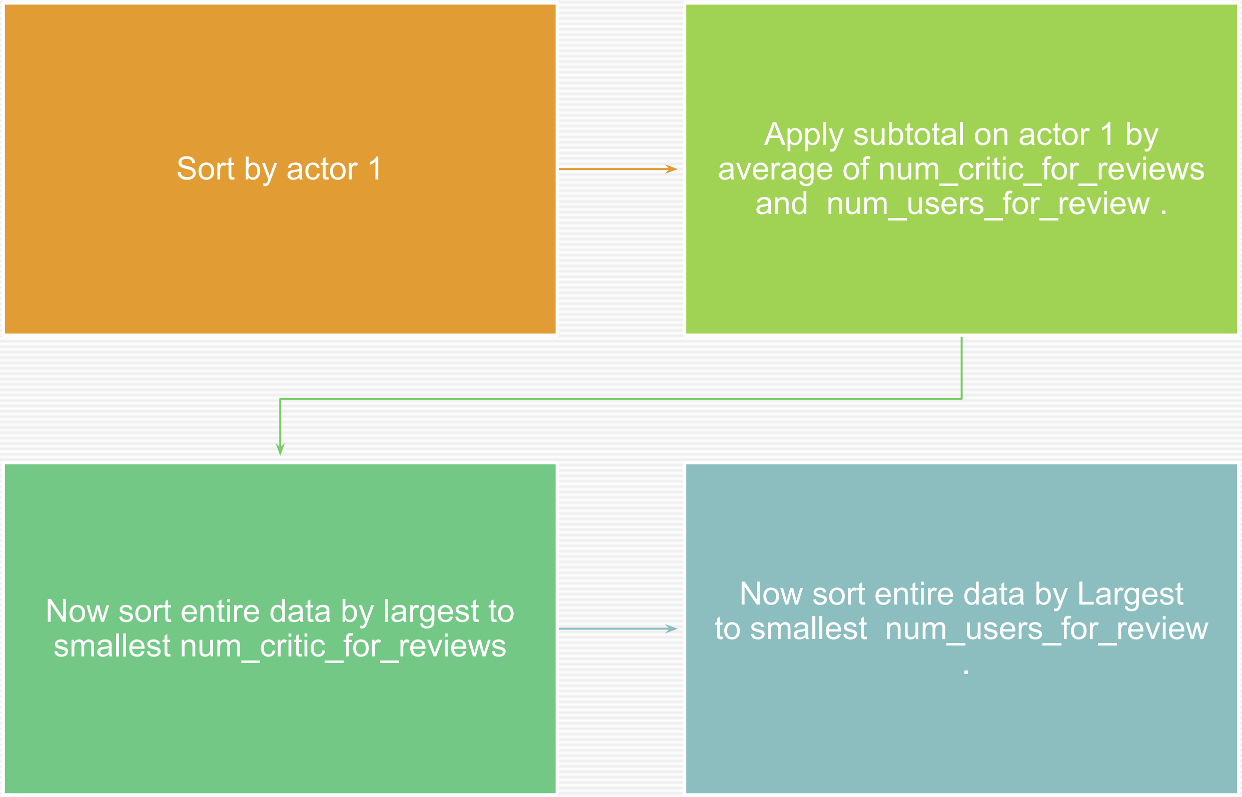
Popular Genres

- Sort the data by genres by sort function in home.
- (Data > Sort & Filter > Advanced).
- Apply subtotal on genres and average by IMDB score
- Data > outline > Subtotal .

Genres	IMDB Score
Action|Adventure|Crime|Drama|Sci-Fi|Thriller Average	8.8
Action|Adventure|Biography|Drama|History Average	8.6
Action|Drama|History|Thriller|War Average	8.5
Adventure|Animation|Drama|Family|Musical Average	8.5
Crime|Drama|Fantasy|Mystery Average	8.5
Action|Adventure|Drama|Fantasy|War Average	8.4
Action|Animation|Crime|Sci-Fi|Thriller Average	8.4
Adventure|Drama|Thriller|War Average	8.4
Comedy|Drama|History|Romance Average	8.4
Adventure|Animation|Comedy|Drama|Family|Fantasy Average	8.3

- ❑ Created three new columns "Meryl_Streep", "Leo_Caprio", and "Brad_Pitt" next to the "actor_1_name" column.
- ❑ In each of the newly created columns, use the IF formula to check if the "actor_1_name" column contains the names 'Meryl Streep', 'Leonardo DiCaprio', or 'Brad Pitt' respectively. For example, in the "Meryl_Streep" column, you can use the formula:
- ❑ =IF(actor_1_name="Meryl Streep", actor_1_name, "")
- ❑ Copy and paste the formula for each of the new columns for the respective actors.
- ❑ Create a new column "Combined" and use the CONCATENATE formula to combine the values of the three new columns into one. For example:
- ❑ =CONCATENATE(Meryl_Streep, Leo_Caprio, Brad_Pitt)
- ❑ Group the values in the "Combined" column by the "actor_1_name" column by selecting the "Combined" column, going to the "Data" tab, and clicking on "Sort & Filter" and then "Sort A-Z".

Find the critic-favorite and audience-favorite actors



Sort by actor 1

Apply subtotal on actor 1 by average of num_critic_for_reviews and num_users_for_review .

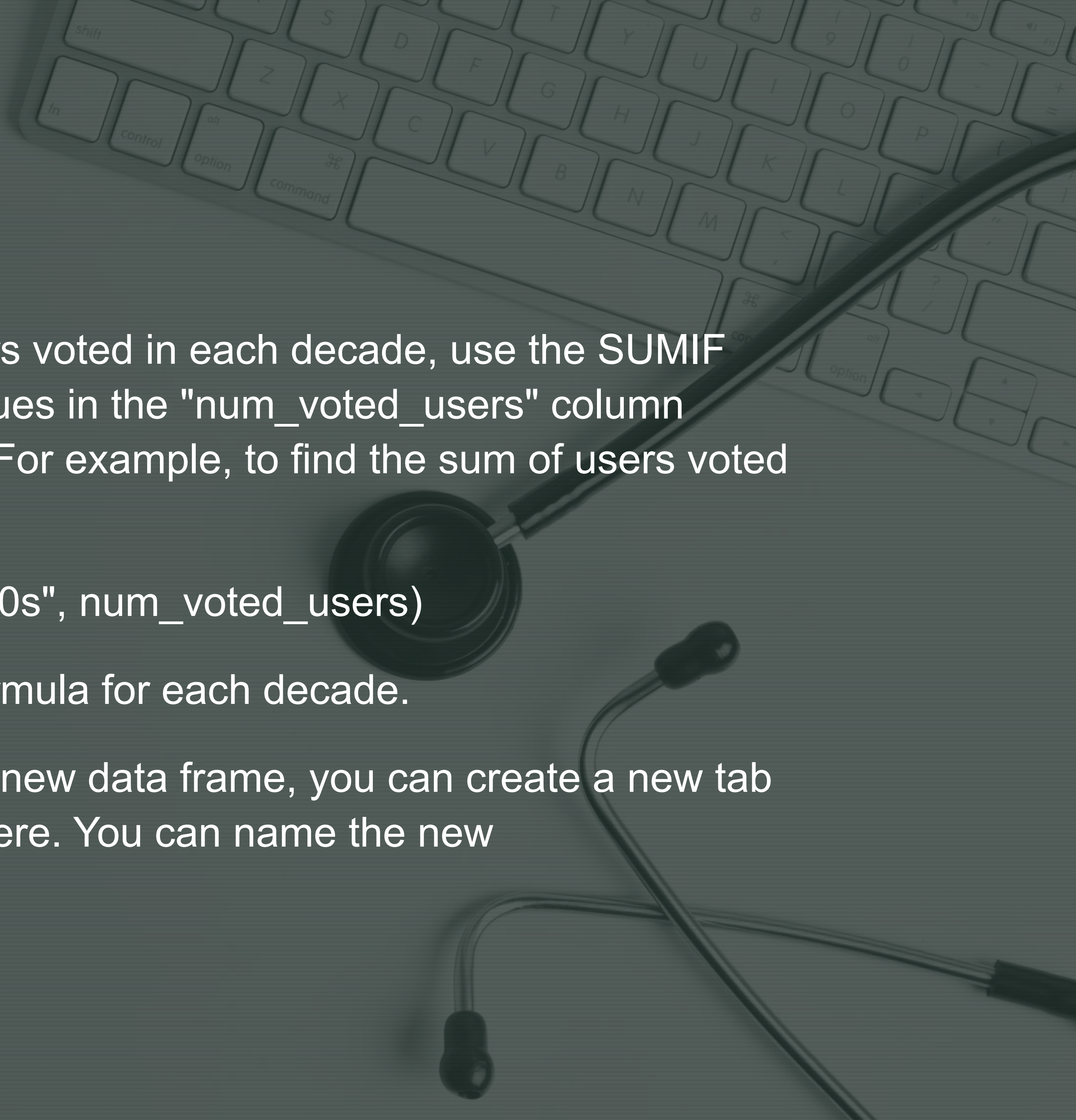
Now sort entire data by largest to smallest num_critic_for_reviews

Now sort entire data by Largest to smallest num_users_for_review .

CHOICE	Actor	Rating
CRITICS CHOICE	**Tom Hardy**	**813**
USERS CHOICE	Christopher Lee	5060

Mean of user votes in each decade

- Create a new column "Decade" next to the "title_year" column.
- In the first cell of the "Decade" column, use the INT formula to find the decade to which the movie belongs. For example:
- =INT(title_year/10)*10 & "s"
- Copy and paste the formula for the rest of the cells in the "Decade" column.
- Sort the data by the "Decade" column by selecting the "Decade" column, going to the "Data" tab, and clicking on "Sort & Filter" and then "Sort A-Z".

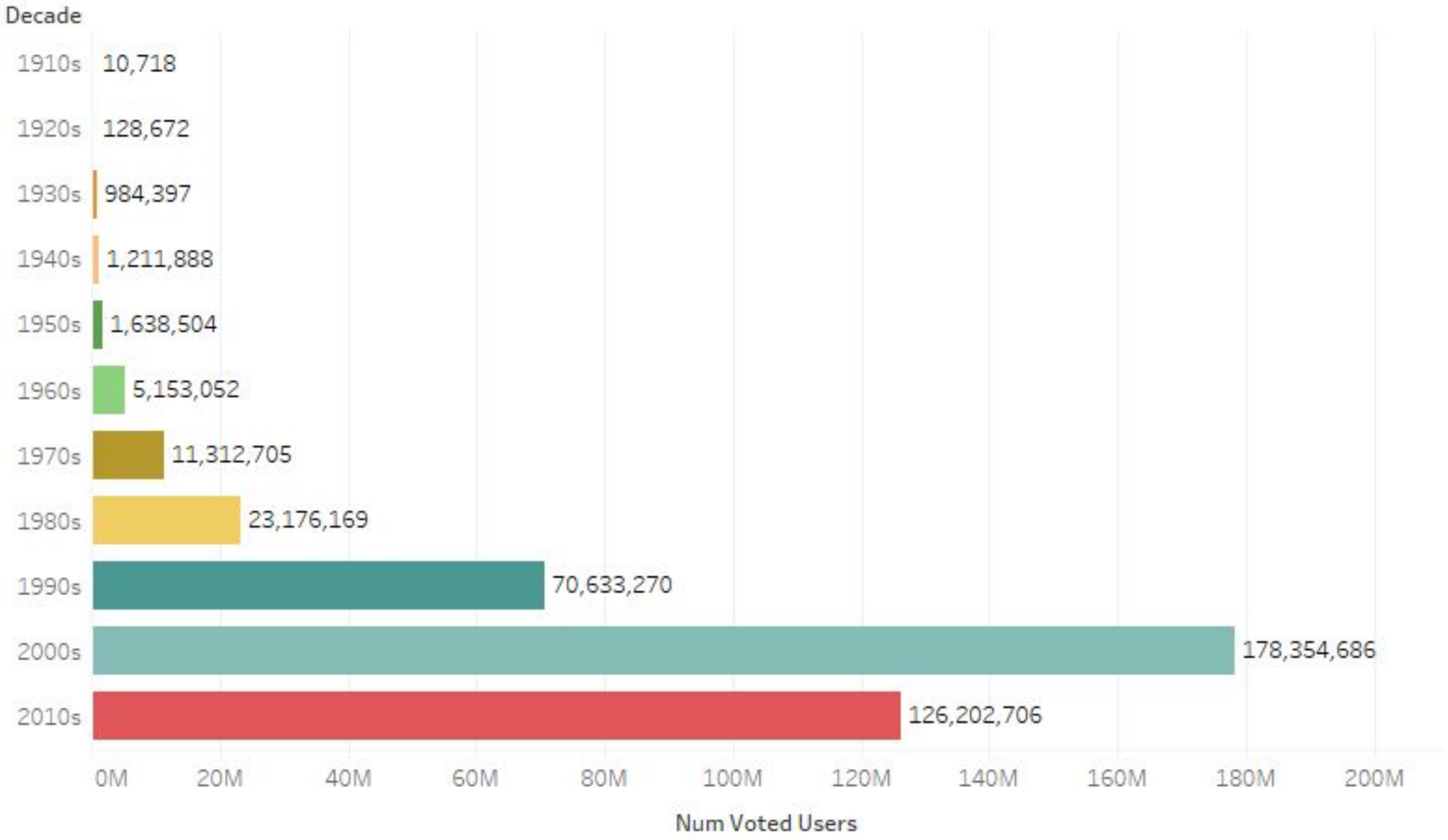


- To find the sum of users voted in each decade, use the SUMIF formula to sum the values in the "num_voted_users" column based on the decade. For example, to find the sum of users voted in the 1920s:
- =SUMIF(Decade, "1920s", num_voted_users)
- Copy and paste the formula for each decade.
- To store the result in a new data frame, you can create a new tab and paste the result there. You can name the new tab "df_by_decade".

num_voted_users	DECADE
10718	1910s
128672	1920s
984397	1930s
1211888	1940s
1638504	1950s
5153052	1960s
11312705	1970s
23176169	1980s
70633270	1990s
178354686	2000s
126202706	2010s

Change in number of voted users over decades using a bar chart

- Select the columns and select char bar in insert
- Now select the column where you want to present the chart


Decade
1910s
10,718
1920s
128,672
1930s
984,397
1940s
1,211,888
1950s
1,638,504
1960s
5,153,052
1970s
11,312,705
1980s
23,176,169
1990s
70,633,270
2000s
178,354,686
2010s
126,202,706
0M
20M
40M
60M
80M
100M
120M
140M
160M
180M
200M
Num Voted Users

Approach

- My approach in this project involved using the 5 Why approach in data analysis. To clean the data, I dropped irrelevant columns, removed null values, and created new columns. For instance, I created a new column called "Profit" using the formula: **`Profit = Gross - Budget`**. To determine the top 250 movies, I created the column "IMDb_Top_250" using the formula: **`IMDb_Top_250 = IF(IMDb_Score >= [minimum IMDb score] AND Num_Voted_Users >= 25000, 1, 0)`**.

- To find the best directors, I used the formula: `Top10Director = IF(AVERAGE(IMDb_Score) >= [minimum average IMDb score], Director_Name, "")`. To find the most popular genres, I used the formula `=COUNTIF(Genre, [Genre])` and sorted the results. To find the favorite actors, I created columns for three actors and used the formula `=IF(Actor_1_Name = [Actor Name], 1, 0)`. Finally, I used the formula `=GROUPBY(Decade, SUM(Num_Voted_Users))` to observe changes in popularity over decades.

Tech-Stack Used

- Excel
- Microsoft power point presentation

Insights

- My approach towards this project is to use the 5 Why approach in data analysis. The 5 Why approach involves asking the question "Why?" five times in order to drill down to the root cause of a problem. In this project, I applied this approach to understand and clean the data in a systematic manner

- While making the project, I gained several insights into movie trends and popular actors. Some of these insights include:
- The profit of a movie is strongly correlated with its budget. The higher the budget, the higher the potential profit. However, there were some outliers with high profits but low budgets.
- The IMDb Top 250 movies had a high IMDb score and a large number of voted users. This suggests that these movies were well-received by audiences and critics.
- The top 10 directors had a high average IMDb score, indicating that their films are highly rated.
- Action and Drama were the most popular movie genres, followed by Thriller and Comedy.

- Meryl Streep, Leonardo DiCaprio, and Brad Pitt were the favorite actors among critics and audiences. They appeared in a number of highly rated movies and had high average ratings for the number of critics and users that reviewed their films.
- The popularity of movies, as measured by the number of voted users, has increased over time, with the largest growth in popularity occurring in the 2010s.
- These insights were gained through data analysis techniques such as cleaning, sorting, grouping, and plotting. By understanding these trends, filmmakers and studios can make informed decisions on what types of movies to produce and which actors to cast in their films.

Result

- Through this project, I was able to solidify my skills in data cleaning, manipulation, and analysis, and I feel more confident in my ability to work with large datasets. Additionally, I was able to practice my critical thinking skills and problem solving abilities as I navigated the various challenges and obstacles that arose during the project.