IMDB MOVIES ANALYSIS

BY

Rahul Patil

Table of Contents

Project Description:

- Excellent film analysis will explain how a film has been made: which filmmaking techniques have been chosen and why, how the visual storytelling supports the narrative, and the effect that filmmaking elements have on the viewer.
- It helps producer, director, and investor take better decision while making any new film as identifying which is most famous actor, which actor movies people watch most, what kind of story people like to watch.

Approach:

- To successfully carry out this project we are going to use **SIX STEP** of Data Analysis Process i.e (Ask, Prepare, Process, Analyze, Share, Act)
- Ask step include asking right set of question which justify goal and give motivation to carry out analysis

- We have following set of question (reasons) to justify goal of this project.

  - Find the movies with the highest profit?
  - Find IMDB Top 250 movies?
  - Find out the top 10 directors?
  - Find popular genres?
  - Find the critic-favorite and audience-favorite actors
- Prepare: We have data in excel format which need to first clean, transform and load into correct format to make it suitable for analysis purpose.

- This step includes selecting right data, tools, data source to make project successful
- Process: Data we have in excel format we need to clean data such as removing null values, identifying data type, removing outliers and unnecessary columns which affect the analysis.
- Analyze: We are using excel itself which come with inbuilt statistical formulae and visualization tool to analyze data to draw insight.
- Share: we are showing data obtain from analysis in the form of row and column as well as chart wherever required for better and easy understanding.
- Act: Step include taking decision based on insight opt from this project.

Tech-Stack Used:

- We are going to use excel which come with inbuilt statistical formulae and visualization tool to analyze data to draw insight.
- Excel come with power query which is suitable for ETL process
- Pivot table help to perform quick ad-hoc analysis.

◆ # Insights:

Let’s first understand data

Open another excel sheet click on data tab user power query editor to Extract, load, and transform data.

Total number of record = 5041
After removing null records total record count = 3884
77% data is left after removing null values

Total number of columns = 28

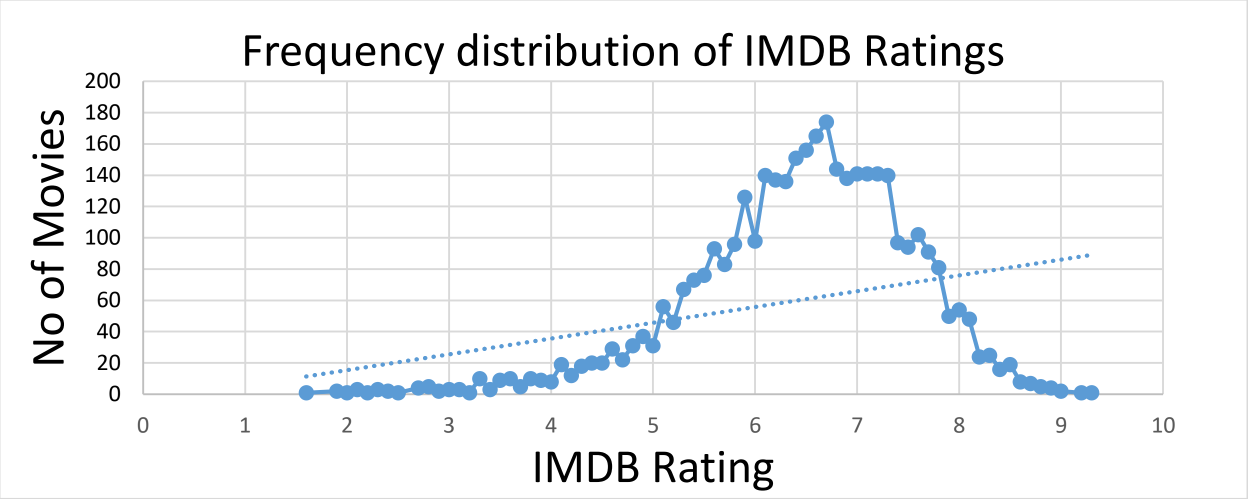
We don’t require all the features of our analysis so we will drop the unnecessary features.

We mostly be analyzing the movies with respect to the ratings, gross collection, popularity of movies, etc.
So it is advised to drop the following columns.

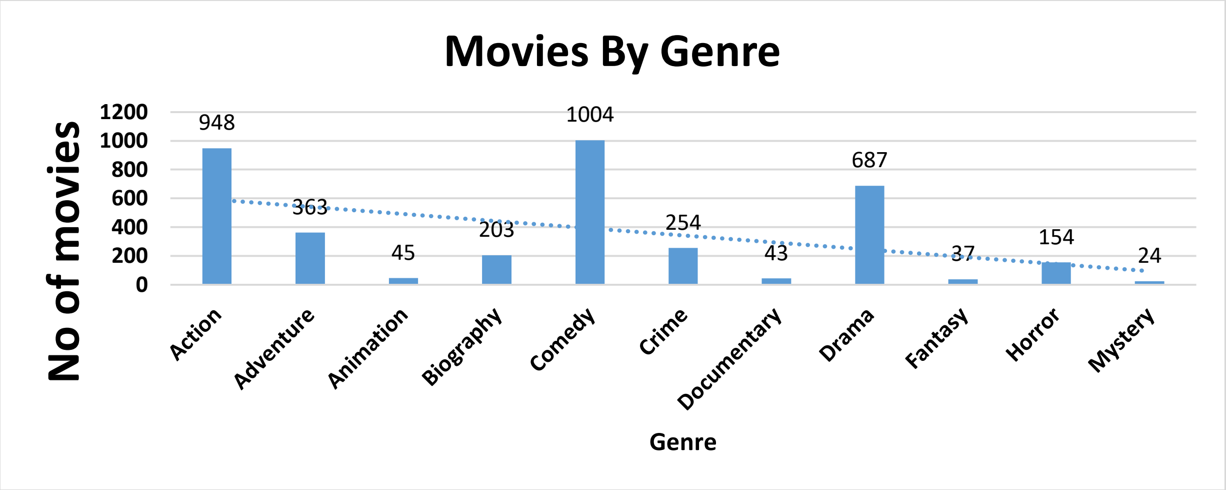
- color
- director_facebook_likes
- actor_1_facebook_likes
- actor_2_facebook_likes
- actor_3_facebook_likes
- actor_2_name
- cast_total_facebook_likes
- actor_3_name
- duration
- facenumber_in_poster
- content_rating
- country
- movie_imdb_link
- aspect_ratio
- plot_keywords

1.General observations

The below observations suggest that most of the movies have a rating of 5 or more and 8 or less.



The data set comprised of movie data, most of which were comedy, action or drama.



2.Find the movies with the highest profit?

Create new column called profit which is difference between gross and budget

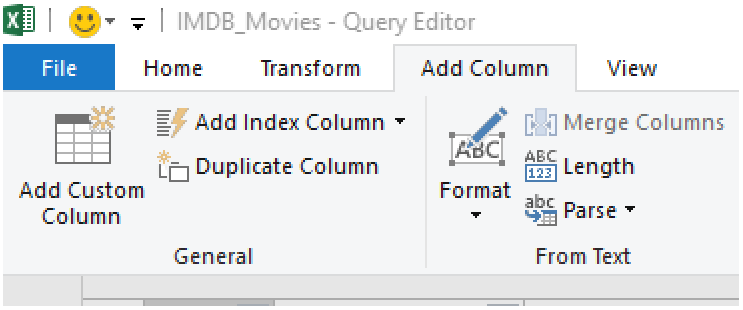
We can use following formula

=[@gross]-[@budget]

Or PowerQuery to add custom column

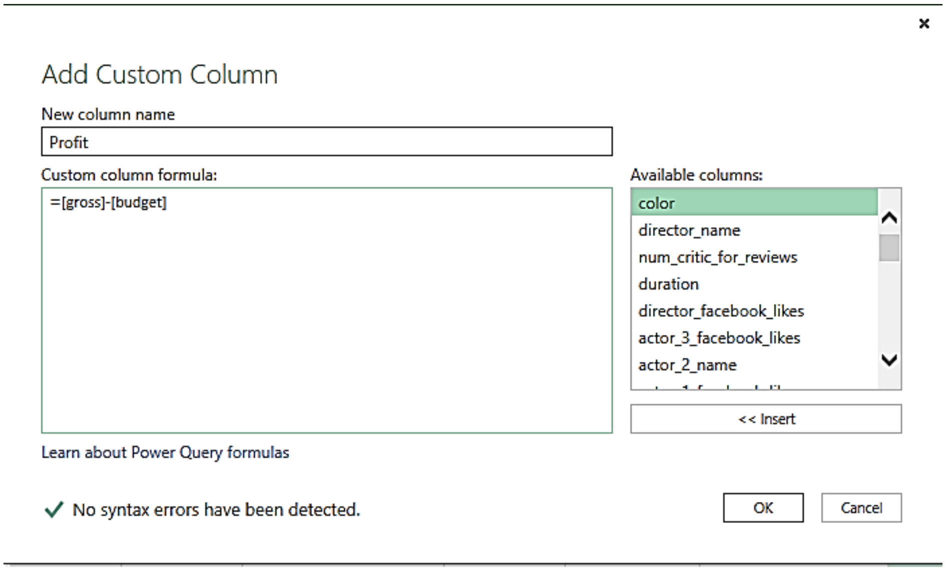
1.click on Add Column tab

2.Then click on Add Custom Column



3.Following dialogue box will open

4.Use following formula to calculate profit and click on ok.



Add Custom Column

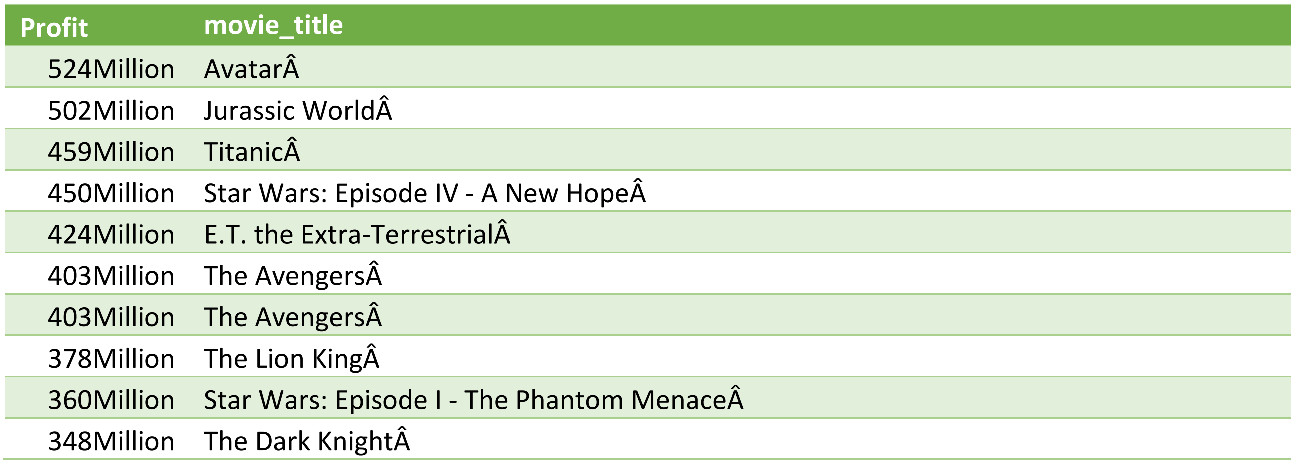
New column name: Profit

Custom column formula:

```
=[gross]-[budget]
```

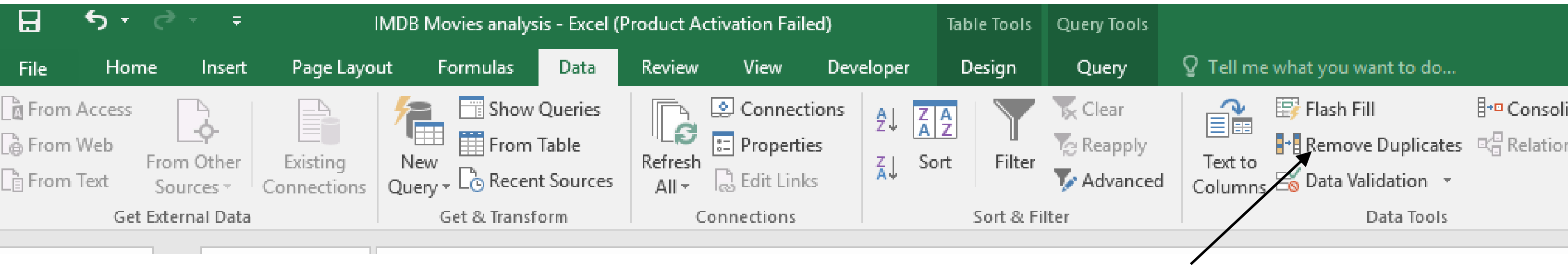
✓ No syntax errors have been detected.

After sorting Profit columns, we can see that there are some duplicates values.
We will drop the duplicate record before moving further

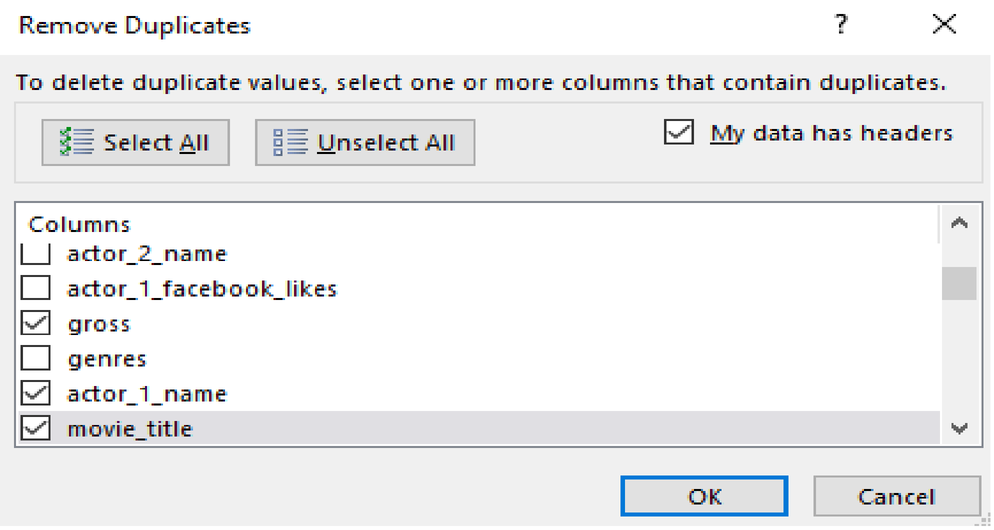


Profit	movie_title
524Million	AvatarÂ
502Million	Jurassic WorldÂ
459Million	TitanicÂ
450Million	Star Wars: Episode IV - A New HopeÂ
424Million	E.T. the Extra-TerrestrialÂ
403Million	The AvengersÂ
403Million	The AvengersÂ
378Million	The Lion KingÂ
360Million	Star Wars: Episode I - The Phantom MenaceÂ
348Million	The Dark KnightÂ

1.click on data tab select Remove Duplicates option.



2.Following pop up will open, follow the message in box to delete duplicates rows
3.Here select movies_title, actor_1_names and gross to delete duplicate rows and click on ok.



Remove Duplicates

To delete duplicate values, select one or more columns that contain duplicates.

Select All | Unselect All | ☑ My data has headers

Columns
- ☐ actor_2_name
- ☐ actor_1_facebook_likes
- ☑ gross
- ☐ genres
- ☑ actor_1_name
- ☑ movie_title

4.Now sort the data again by profit columns and we will get top 10 profitable movies.

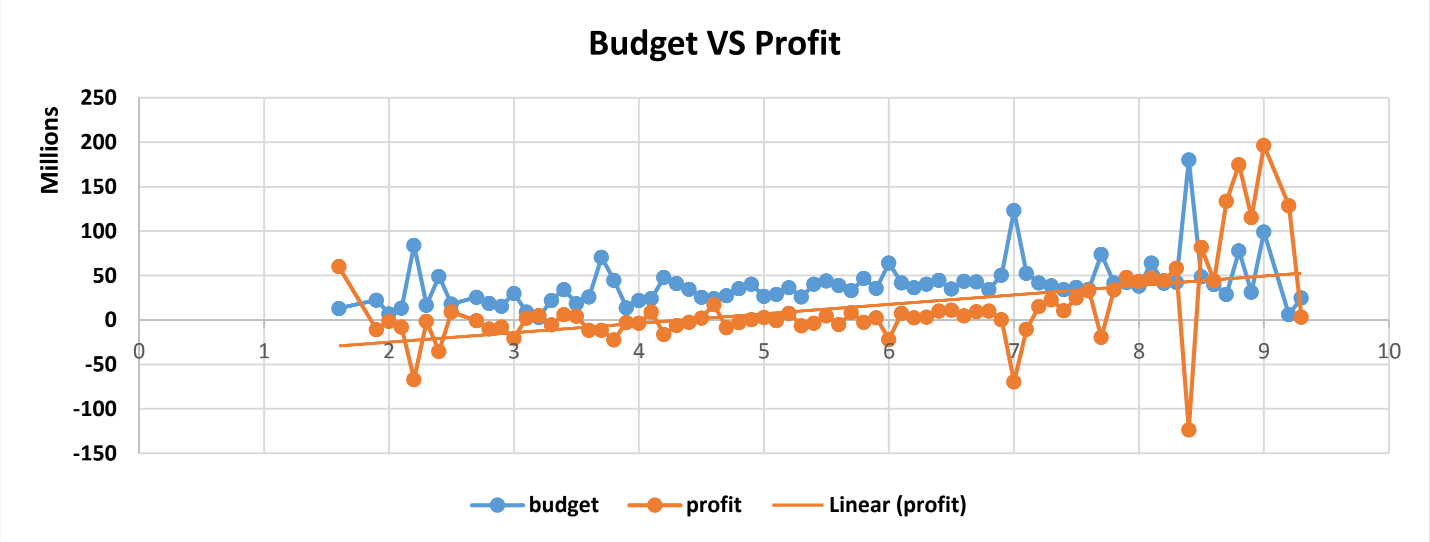
Profit	movie_title
524Million	AvatarÂ
502Million	Jurassic WorldÂ
459Million	TitanicÂ
450Million	Star Wars: Episode IV - A New HopeÂ
424Million	E.T. the Extra-TerrestrialÂ
403Million	The AvengersÂ
378Million	The Lion KingÂ
360Million	Star Wars: Episode I - The Phantom MenaceÂ
348Million	The Dark KnightÂ
330Million	The Hunger GamesÂ

2.1. Do good IMDB ratings guarantee a movie's success at the box office?

From what was analysed and found in the data, the answer to the above question seems to be "Yes" but that comes with a certain set of conditions. From what we can observe below, it seems that a movie on average needs to be atleast rated "6" by IMDB for it to break even, and trend shows exponential growth with ratings going up from here onwards.

We did find outliers in both the extreme ends of this graph, where a movie with rating over 9 didn't make profits and a movie with 1.5 rating making 50 million worth of profits. The former being "Shawshank Redemption" and the Latter being Justin Beiber's "Never say never".

There are various social reasons that contributed to these extreme situations that speak a different language than what the majority of data represents.

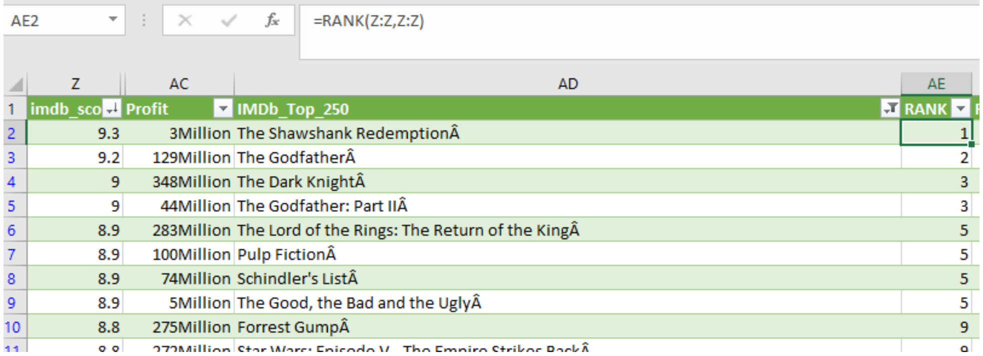


To summarise Mathematically, 91 % of the movies that have been rated over 5 have had positive profits which suggests that IMDB ratings and movie's success (Net profit) are positively correlated.

Also point to be noted that as budget increase Net profit tends to descrease.

3.Find IMDB Top 250 movies?

- Sort the **imdb_score** column in descending order
- Create new column name called **IMDb_Top_250** by entering following formula in new column. =IF([num_voted_users]>25000,[movie_title],"")
- Rank movies by using RANK function

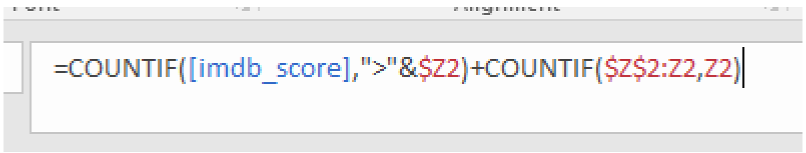


AE2

=RANK(Z:Z,Z:Z)

	Z	AC	AD	AE
1	imdb_sco	Profit	IMDb_Top_250	RANK
2	9.3	3Million	The Shawshank RedemptionÂ	1
3	9.2	129Million	The GodfatherÂ	2
4	9	348Million	The Dark KnightÂ	3
5	9	44Million	The Godfather: Part IIÂ	3
6	8.9	283Million	The Lord of the Rings: The Return of the KingÂ	5
7	8.9	100Million	Pulp FictionÂ	5
8	8.9	74Million	Schindler's ListÂ	5
9	8.9	5Million	The Good, the Bad and the UglyÂ	5
10	8.8	275Million	Forrest GumpÂ	9
11	8.8	272Million	Star Wars: Episode V - The Empire Strikes BackÂ	9

We can see RANK function gives same rank to same values but skip the continuity



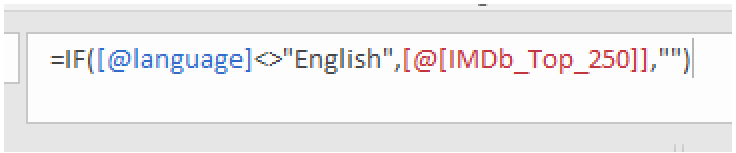
=COUNTIF([imdb_score],">"&$Z2)+COUNTIF($Z$2:Z2,Z2)

Above formula can be use to maintain continuity in RANK & we will get Top 250 IMDB Movies

IMDb_Top_250	RANK2
The Shawshank RedemptionÂ	1
The GodfatherÂ	2
The Dark KnightÂ	3
The Godfather: Part IIÂ	4
The Lord of the Rings: The Return of the KingÂ	5
Pulp FictionÂ	6
Schindler's ListÂ	7
The Good, the Bad and the UglyÂ	8
Forrest GumpÂ	9
Star Wars: Episode V - The Empire Strikes BackÂ	10
The Lord of the Rings: The Fellowship of the RingÂ	11
InceptionÂ	12

3.1. Find Top Foreign Movies?

- Use following formula to find out Foreign Movies



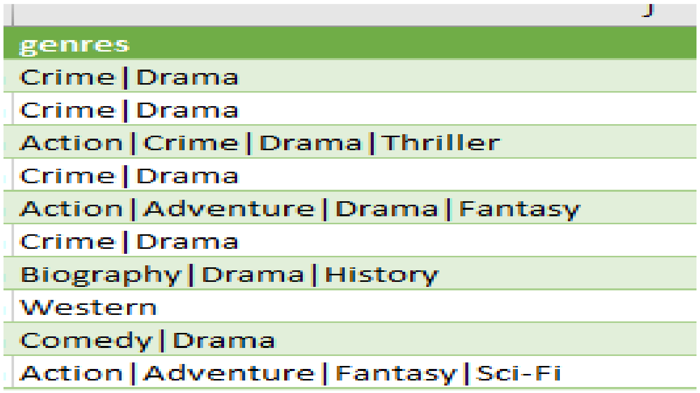
```
=IF([@language]<>"English",[@[IMDb_Top_250]],"")
```

language	Top foreign movies
Italian	The Good, the Bad and the UglyÂ
Portuguese	City of GodÂ
Japanese	Seven SamuraiÂ
Japanese	Spirited AwayÂ
German	The Lives of OthersÂ
Persian	Children of HeavenÂ
Persian	A SeparationÂ
Korean	OldboyÂ
German	Das BootÂ
Telugu	Baahubali: The BeginningÂ
French	AmÃ©lieÂ
Japanese	Princess MononokeÂ
Danish	The HuntÂ
German	MetropolisÂ
German	DownfallÂ
Spanish	Pan's LabyrinthÂ
Spanish	The Secret in Their EyesÂ
French	IncendiesÂ
Japanese	Howl's Moving CastleÂ
Spanish	Amores PerrosÂ

4.Find out the top 10 directors?

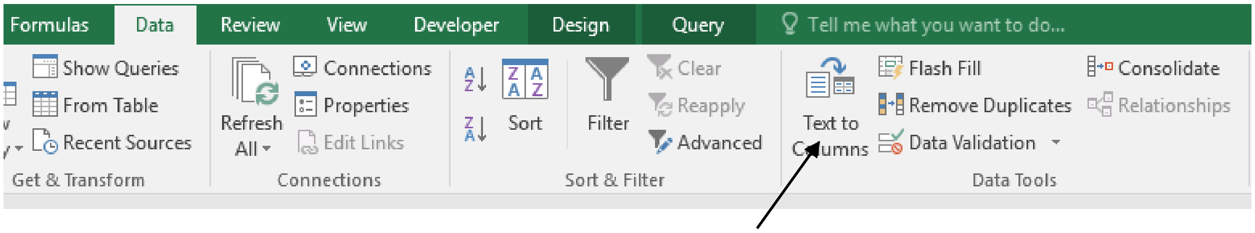
Top 10 directors	Average of imdb_score
Charles Chaplin	8.6
Tony Kaye	8.6
Alfred Hitchcock	8.5
Damien Chazelle	8.5
Majid Majidi	8.5
Ron Fricke	8.5
Sergio Leone	8.433333333
Christopher Nolan	8.425
Marius A. Markevicius	8.4
Richard Marquand	8.4
S.S. Rajamouli	8.4

5.Find popular genres?

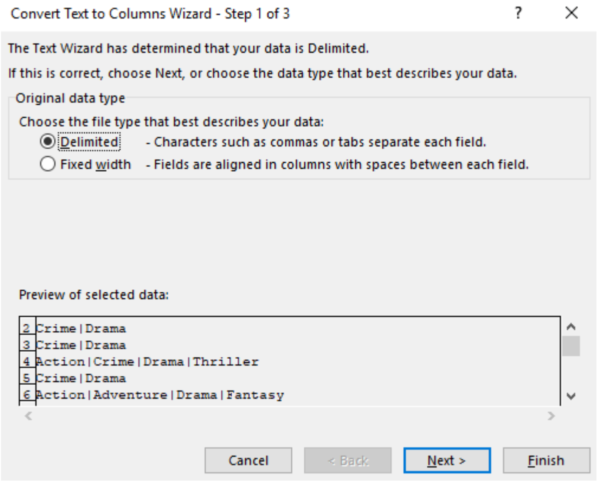


genres
Crime|Drama
Crime|Drama
Action|Crime|Drama|Thriller
Crime|Drama
Action|Adventure|Drama|Fantasy
Crime|Drama
Biography|Drama|History
Western
Comedy|Drama
Action|Adventure|Fantasy|Sci-Fi

- We can see genres column contain more than 1 genre by using separator |.
- Split the genres by using following step.
- Select column and click on data tab
- Select Text to Columns



- New dialogue box will open click select delimited click on next.



Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

Choose the file type that best describes your data:

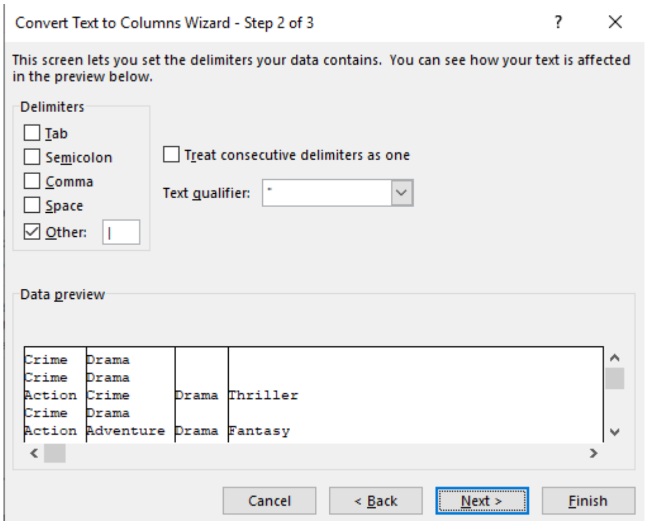
● Delimited - Characters such as commas or tabs separate each field.

○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
2 Crime|Drama
3 Crime|Drama
4 Action|Crime|Drama|Thriller
5 Crime|Drama
6 Action|Adventure|Drama|Fantasy
```

- Select other type | symbol in it & click on next and then click on finish.



Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

- ☐ Tab
- ☐ Semicolon
- ☐ Comma
- ☐ Space
- ☑ Other: |

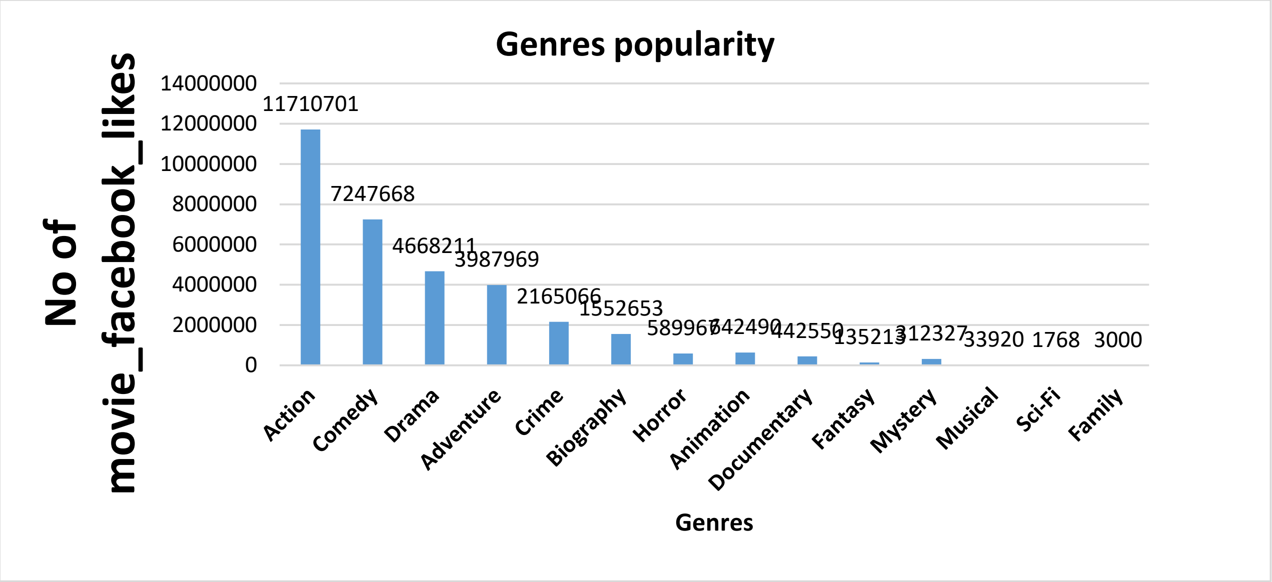
☐ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
Crime   Drama
Crime   Drama
Action  Crime            Drama   Thriller
Crime   Drama
Action  Adventure        Drama   Fantasy
```

- Most of the action, drama & comedy movies people usually likes to watch
- Following are the most popular genres among the people



6.Find the critic-favorite and audience-favorite actors?

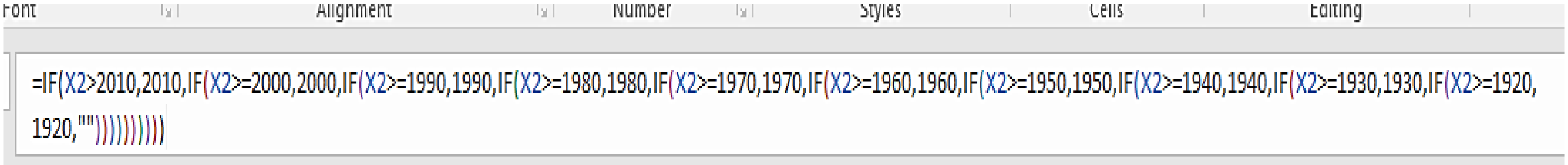
- Create three new columns namely, Meryl_Streep, Leo_Caprio, and Brad_Pitt which contain the movies in which the actors: 'Meryl Streep', 'Leonardo DiCaprio', and 'Brad Pitt' are the lead actors.
- Filter data for this 3 actor
- Copy filter data in new sheet
- Create pivot table

Actor	Average of num_critic_for_reviews	Average of num_user_for_reviews
Leonardo DiCaprio	344.45	956.65
Brad Pitt	252.13	784.94
Meryl Streep	191.22	319.67

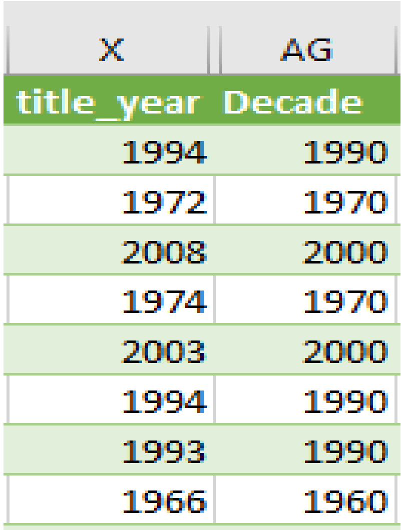
Leonardo has aced both the lists!

6.1. Observe the change in number of voted users over decades using a bar chart.

- Use nested if formula to find decade

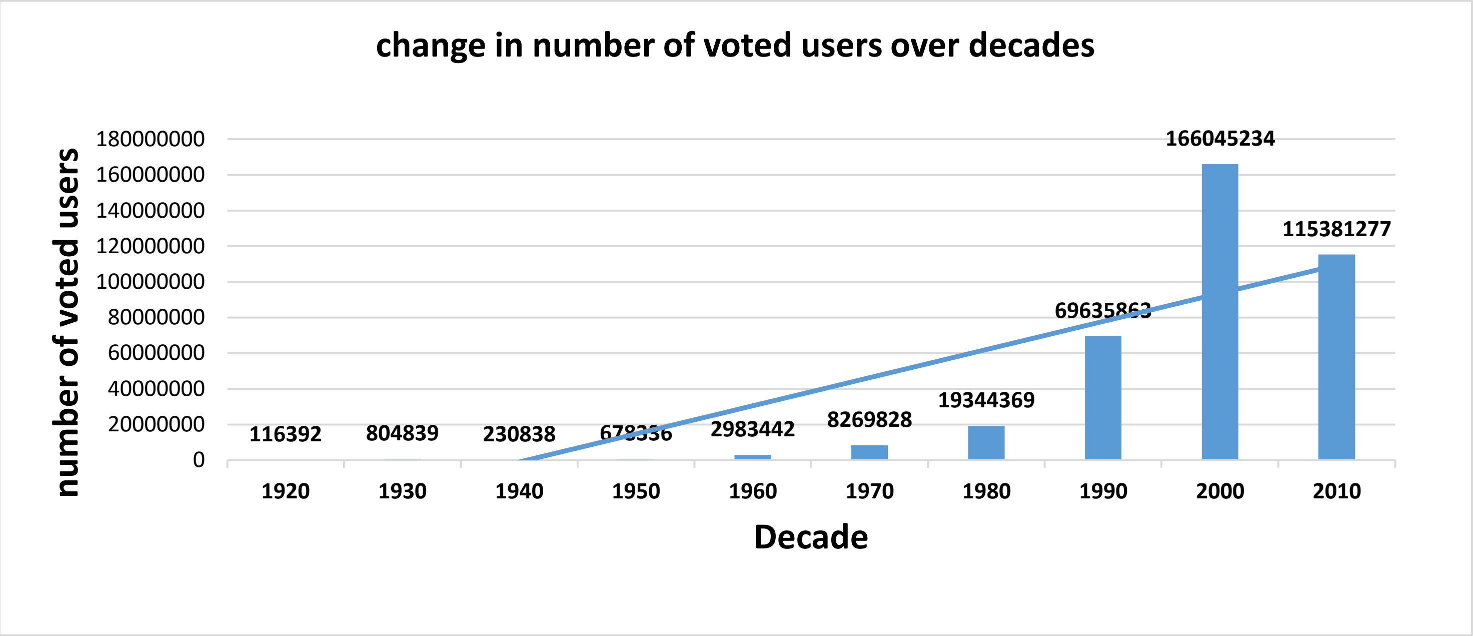


```
=IF(X2>2010,2010,IF(X2>=2000,2000,IF(X2>=1990,1990,IF(X2>=1980,1980,IF(X2>=1970,1970,IF(X2>=1960,1960,IF(X2>=1950,1950,IF(X2>=1940,1940,IF(X2>=1930,1930,IF(X2>=1920,
1920,""))))))))))
```



X	AG
title_year	Decade
1994	1990
1972	1970
2008	2000
1974	1970
2003	2000
1994	1990
1993	1990
1966	1960

- Number of user voted has been increased over the decade but after the 2000 it started to decrease



Result:

- In this project we understand how to use SIX step process of Data analysis.
- Learnt to use why-why analysis technique to define new problem and find the solution.
- In this project we learn to use excel power query and different excel function like countif, nested if, data sorting and filtering, duplicate values removal,text spliting.
- In this project we learn to use excel pivot to perform ad-hoc analysis and different chart and graph to represent data.

- **Key project insight are:**
  - 91 % of the movies that have been rated over 5 have had positive profits which suggests that IMDB ratings and movie’s success (Net profit) are positively correlated.
  - As budget increase Net profit tends to decrease.

- In top 10 director only 1 south Indian director S.S. Rajamouli rest are foreigner.
- Most of the action, drama & comedy movies people usually likes to watch
- Leonardo DiCaprio is most favorite actor of people.
- Number of user voted has been increased over the decade but after the 2000 it started to decrease.