INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT

(Project Semester August-December 2019)

DATA ANALYSIS ON MOVIES DATASET

Submitted by: Saurabh Upadhyay

Registration No: 11714489

Under the Guidance of Ms. Vasudha

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

Lovely Professional University, Phagwara

INDEX

CERTIFICATE

This is to certify that Saurabh Upadhyay Gaur bearing Registration no. 11714489 has completed project for INT 217 project titled, “Analysis on Movies Dataset” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original development, effort, and study.

Miss Vasudha
Assistant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab.

Date: 10/13/19

DECLARATION

I, Saurabh Upadhyay, student of BTECH under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 11/13/2019 Signature:

Registration No. 11714489 Name of the student: Saurabh Upadhyay

ACKNOWLEDGEMENT

I would like to thank our teacher who gave us this opportunity to express our knowledge with this project. I would also thank her for guiding us in the entire phase of this project completion. Not only her but I would like to thank this university for providing us with opportunities to shine and showcase our talent. Special thanks to the friends who helped me to complete this project.

INTRODUCTION

Microsoft Excel is a spreadsheet program that is used to record and analyze numerical data. Think of a spreadsheet as a collection of columns and rows that form a table. Alphabetical letters are usually assigned to columns and numbers are usually assigned to rows. The point where a column and a row meet is called a cell. The address of a cell is given by the letter representing the column and the number representing a row. Let's illustrate this using the following image.

We all deal with numbers in one way or the other. We all have daily expenses which we pay for from the monthly income that we earn. For one to spend wisely, they will need to know their income vs. expenditure. Microsoft Excel comes in handy when we want to record, analyze and store such numeric data.

Tableau

Tableau Software is an American computer software company headquartered in Seattle, WA, USA. It generates interactive data visualization products which focused on BI. The company was established at Stanford University's Department of Computer Science between 1997 and 2002.

Tableau Desktop is a data visualization application to facilitate you to examine virtually any kind of structured data and generate highly interactive, beautiful graphs, dashboards, and reports within minutes. Once a quick installation, you can tie to virtually any data source from spreadsheets to data warehouses and display information in several graphic perspectives. Designed to be easy to utilize, you'll be working more rapidly than ever before.

Movies, also known as films, are a type of visual communication which uses moving pictures and sound to tell stories or teach people something. People in every part of the world watch movies as a type of entertainment, a way to have fun. For some people, fun movies can mean movies that make them laugh, while for others it can mean movies that make them cry or feel afraid.

Most movies are made so that they can be shown on big screens at movie theatres and at home. After movies are shown on movie screens for a period of weeks or months, they may be marketed through several other media. They are shown on pay television or cable television, and sold or rented on DVD disks or videocassette tapes, so that people can watch the movies at

home. You can also download or stream movies. Older movies are shown on television broadcasting stations.

An actor is a person who portrays a character in a performance (also actress; see below). The actor performs "in the flesh" in the traditional medium of the theatre or in modern media such as film, radio, and television. The analogous Greek term is ὑποκριτής (hupokritḗs), literally "one who answers". The actor's interpretation of their role—the art of acting—pertains to the role played, whether based on a real person or fictional character. Interpretation occurs even when the actor is "playing themselves", as in some forms of experimental performance art.

A film director is a person who directs the making of a film. A film director controls a film's artistic and dramatic aspects and visualizes the screenplay (or script) while guiding the technical crew and actors in the fulfilment of that vision. The director has a key role in choosing the cast members, production design, and the creative aspects of filmmaking. Under European Union law, the director is viewed as the author of the film.

The film director gives direction to the cast and crew and creates an overall vision through which a film eventually becomes realized or noticed. Directors need to be able to mediate differences in creative visions and stay within the budget.

IMDb (Internet Movie Database) is an online database of information related to films, television programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, fan and critical reviews, and ratings. An additional fan feature, message boards, was abandoned in February 2017. Originally a fan-operated website, the database is owned and operated by IMDb.com, Inc., a subsidiary of Amazon.

As of May 2019, IMDb has approximately 6 million titles (including episodes) and 9.9 million personalities in its database,[2] as well as 83 million registered users.

IMDb began as a movie database on the Usenet group "rec.arts.movies" in 1990 and moved to the web in 1993.

The following is a project named Analysis on Movies Dataset done with the help of MS Excel and Tableau. Tableau was simply used to clean the data set and Excel was used to draw analysis on the various data available. The dataset contains 28 columns and 5000 rows. The dataset contains records like movie name, actor and actress name, directors name, number of likes they received

on Facebook by viewers, IMDB score and many more. The analysis that could be made are like best actor. best supporting actor, best director based on factors like number of like received on Facebook or IMDB score or the amount that their movies grossed in the international market.

OBJECTIVES/SCOPE OF THE ANALYSIS

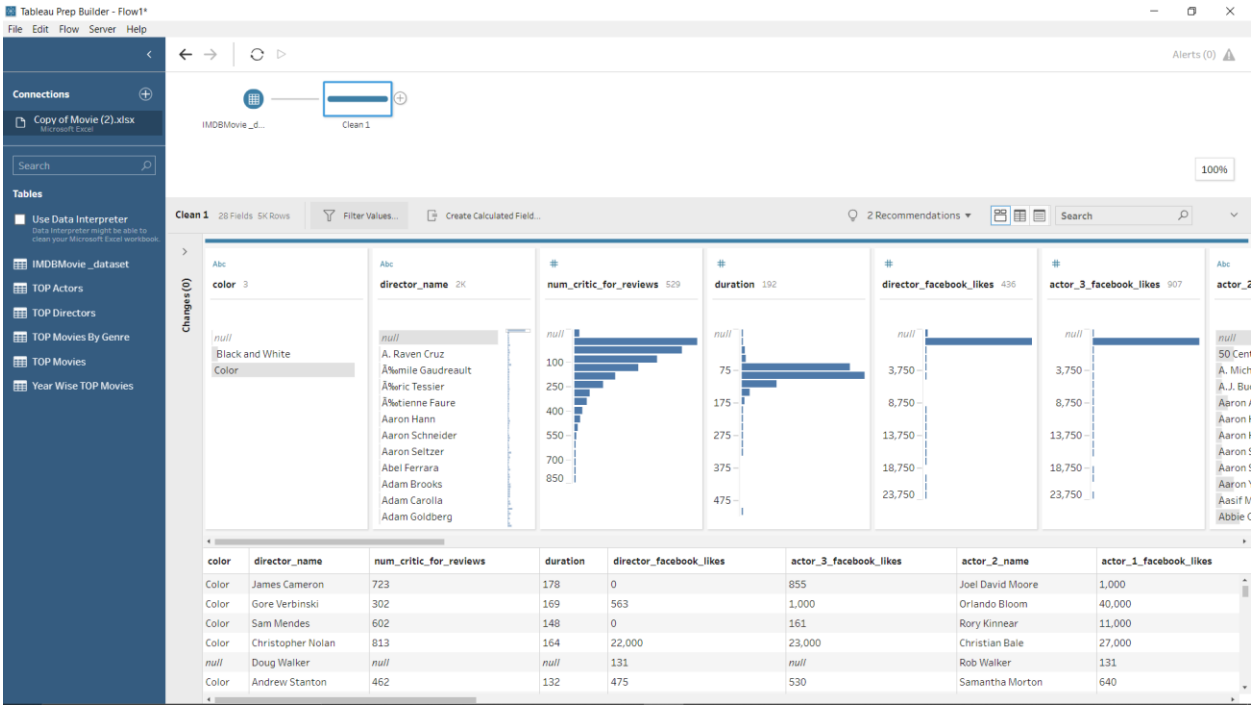
The objective of the analysis is to implement the knowledge of excel and tableau learned throughout the semester in a practical manner to clearly test how good command we have over data management. Objective for the project is to clean the data set, find the top ten movies, director, actor, supporting actor, based on Facebook likes, IMDB Score and gross figure. Same conclusion is also drawn for genre wise movies and also year wise movies. Then to draw graphs and charts like pie chart, line chart, bar diagram for each one of them. Make best use of pivot table to handle the large dataset. And at the end draw dashboard to bring the necessary conclusion altogether by finding out who are the top ten in their respective fields

Source of dataset: Kaggle.com

Link: https://www.kaggle.com/carolzhangdc/imdb-5000-movie-dataset

ETL PROCESS

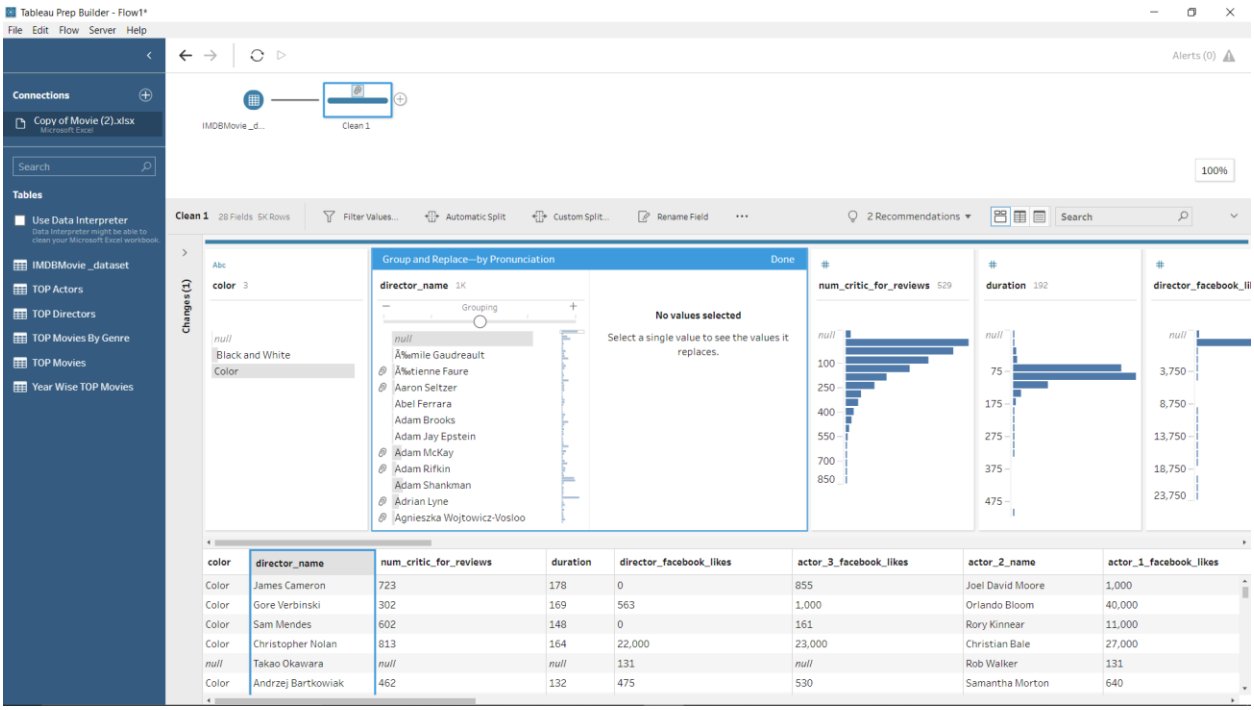
Step 1



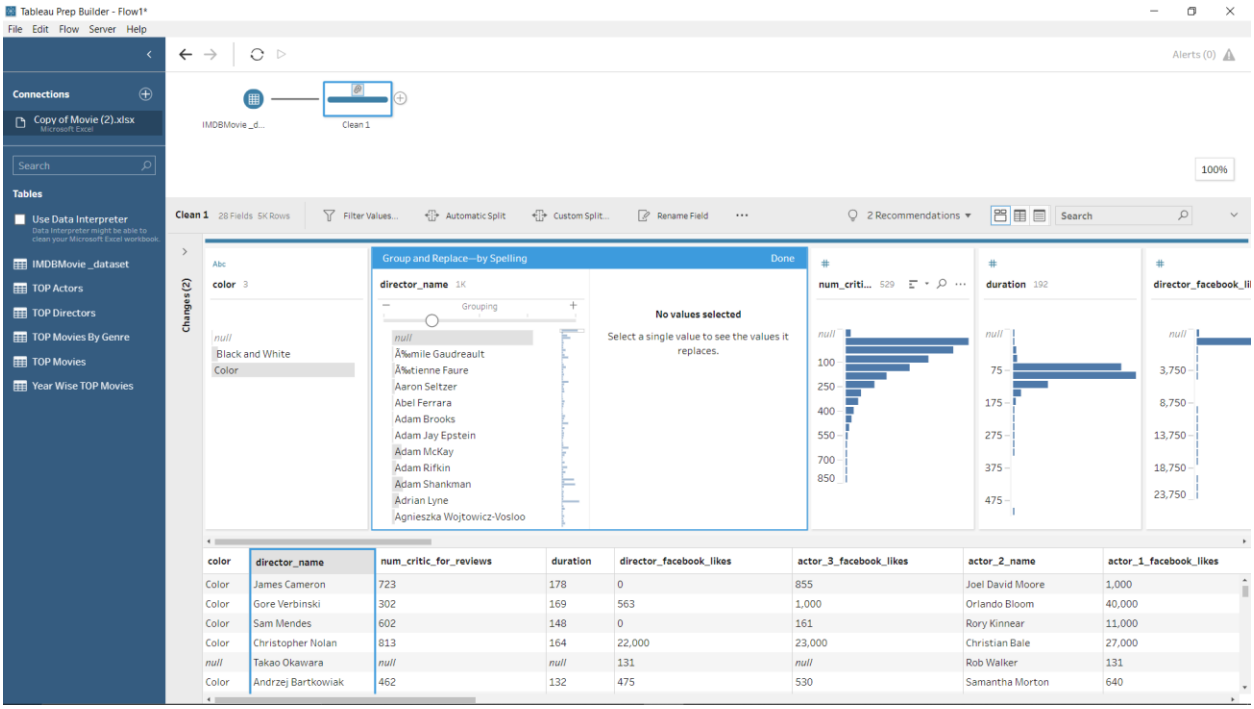
This data set is a meta dataset and could not be directly used. Hence it could not be directly used. Some cleansing was necessary. I just copied the entire data and saved to a new file.

Step 2

Records contained speckling mistakes. Like

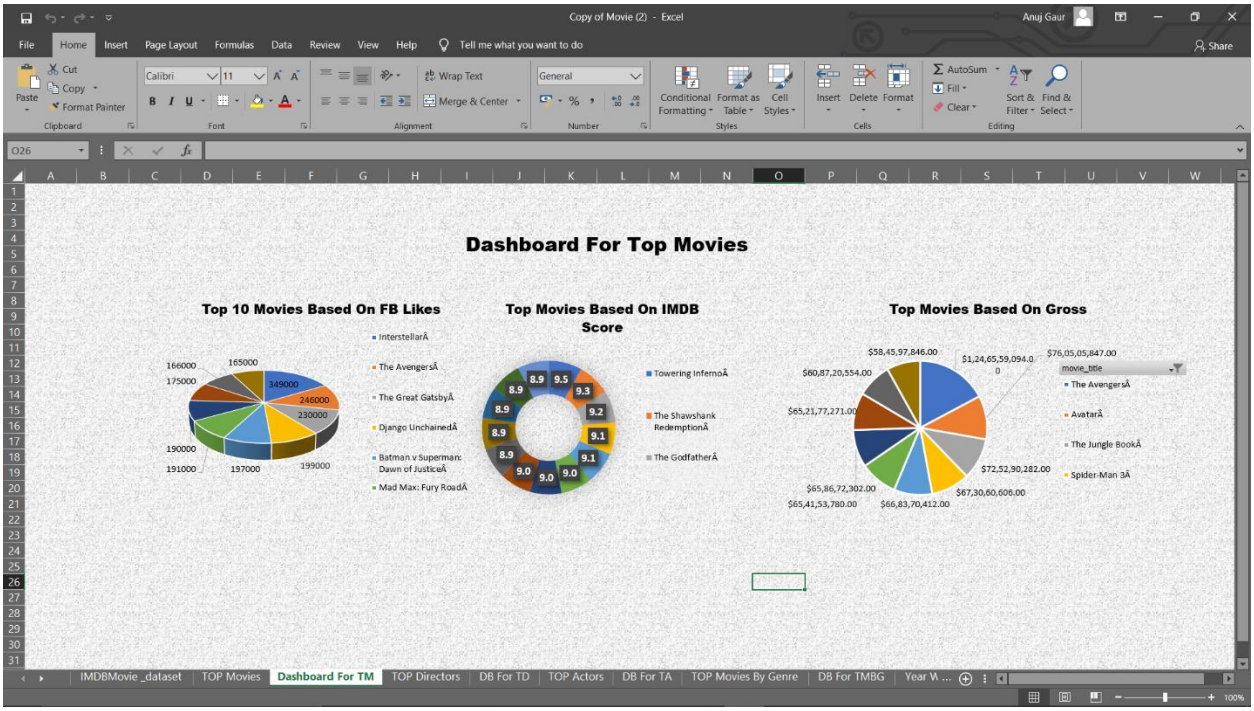


Spelling was corrected like in above movie title column. The letter Â was unnecessarily added. It was removed using excels remove and replace function. Now there are no Â on movie title.



Analysis on dataset

1. Top 10 movies in the world



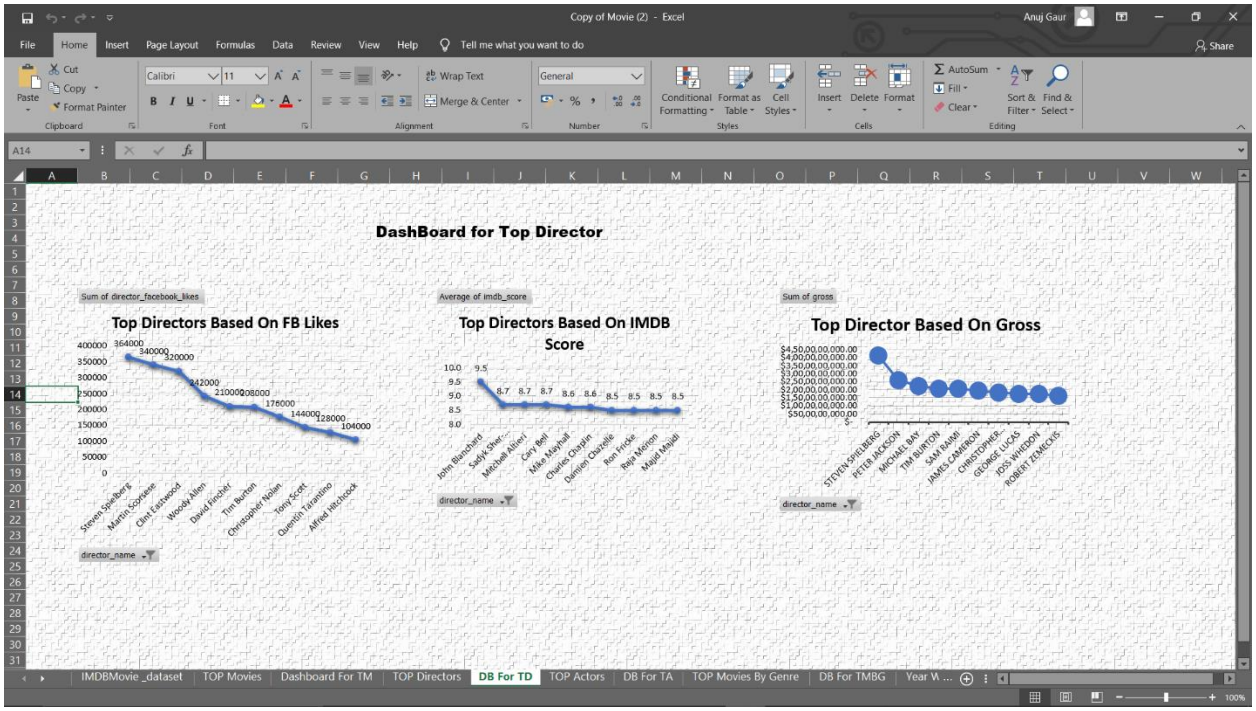
Description:

The above charts show the top 10 best movies based on likes received on Facebook, IMDB score and amount the movie has grossed. The pie charts have been used to make the analysis.

Conclusion:

The analysis shows that Interstellar is the most liked movie, Towering Inferno is the highest rated movie on IMDB platform and by the amount of money the movie earned The Avengers take the top spot.

2.Top 10 Directors in the world



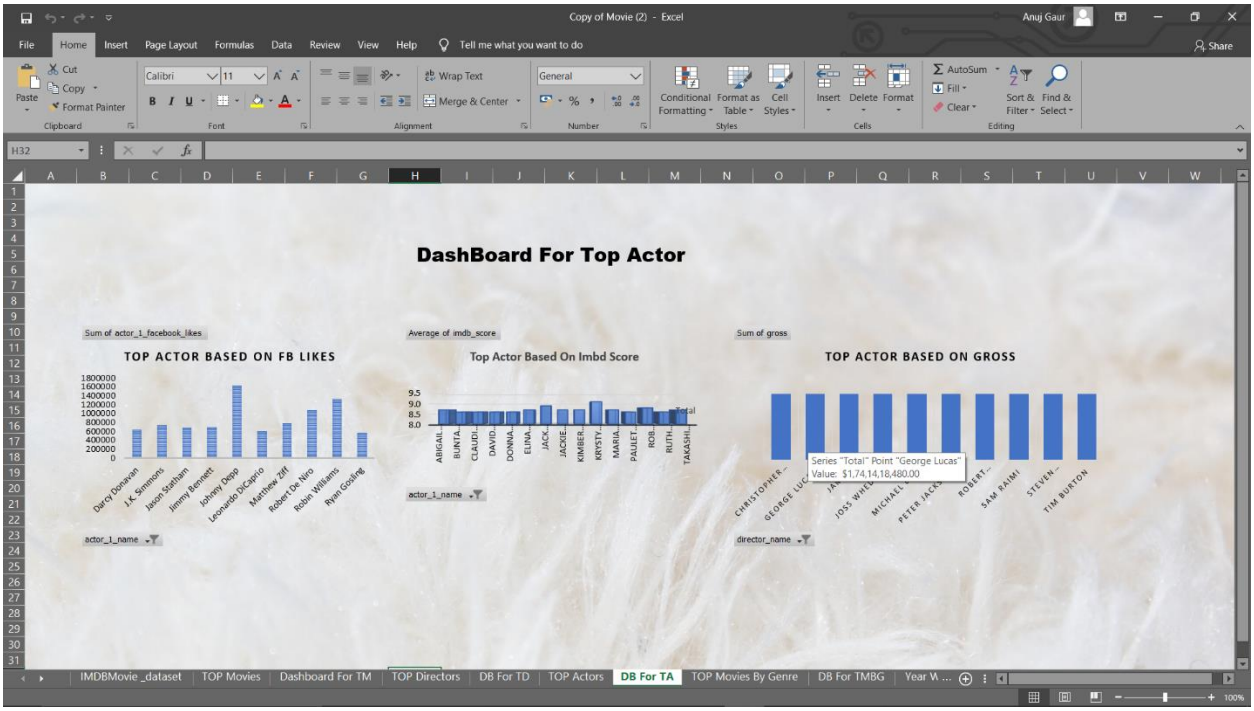
Description:

The above charts show the top 10 best movies based on likes received on Facebook, IMDB score and amount the movie has grossed. The line charts have been used to make the analysis.

Conclusion:

The analysis shows that Steven Spielberg is the most liked Director, John Blanchard is the highest rated director on IMDB platform and by the amount of money the movie earned Steven Spielberg takes the top spot.

3.Top 10 Lead Actors in the world



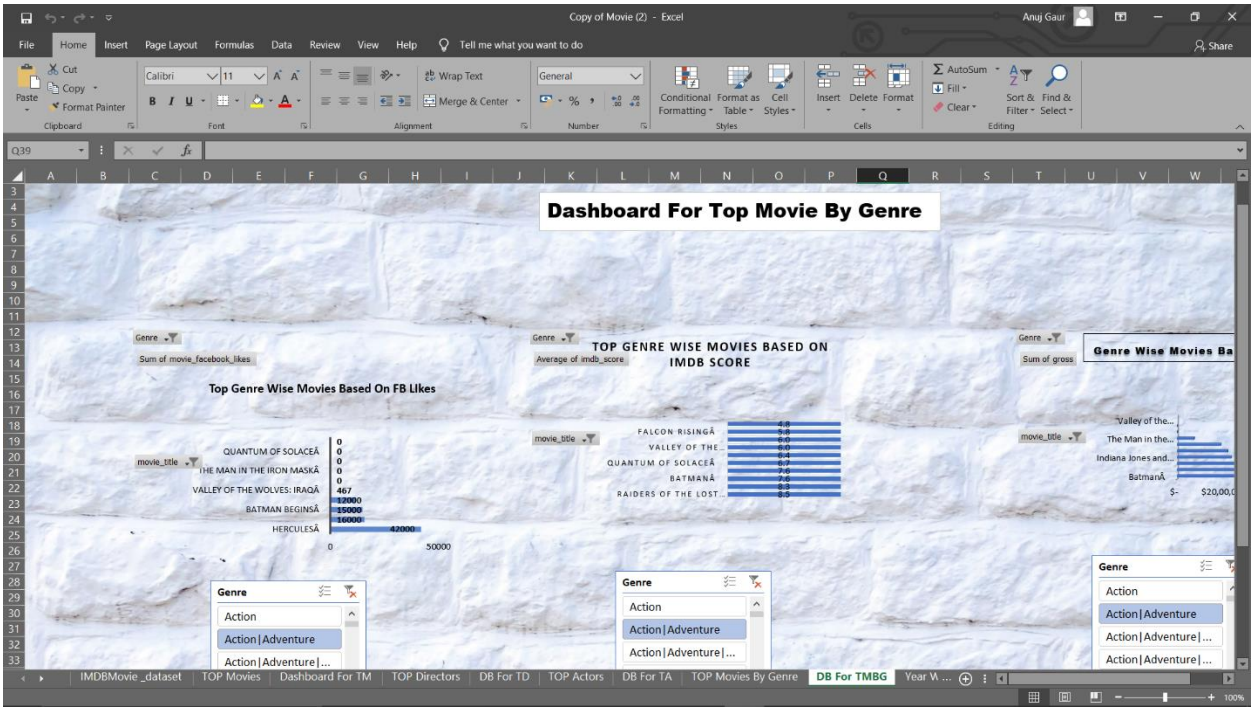
Description:

The above charts show the top 10 best Actors based on likes received on Facebook, IMDB score and amount the movie has grossed. The Column charts have been used to make the analysis.

Conclusion:

The analysis shows that Johnny Depp is the most liked Actor, Krystyna Janda is the highest rated Actor on IMDB platform and by the amount of money the movies earned Steven Spielberg takes the top spot.

4. Top Genre wise movies in the world



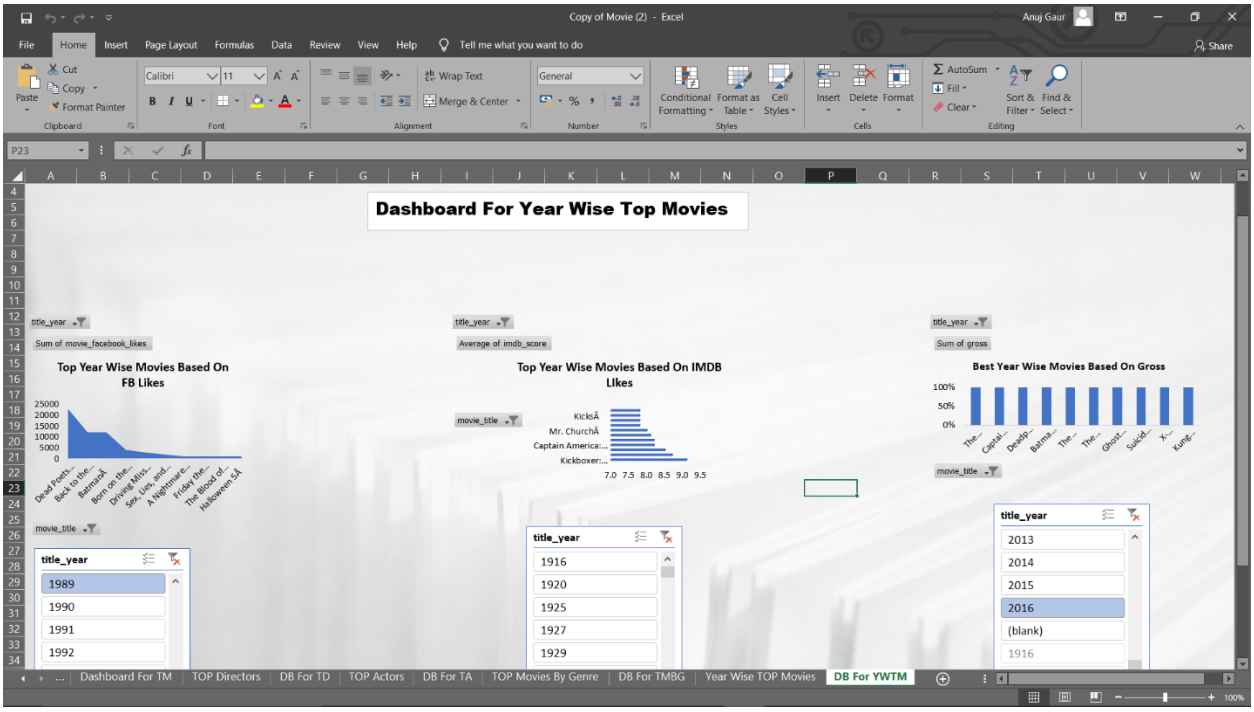
Description:

The above charts show the top genre wise movies based on likes received on Facebook, IMDB score and amount the movie has grossed. The Bar charts have been used to make the analysis.

Conclusion:

The analysis shows that different movies rank in all different categories. Like Hercules is most liked action-adventure movie while The man with iron fist is most liked action movie. The slicer added lets us choose the genre.

5. Top Year wise movies in the world



Description:

The above charts show the top movies for each year based on likes received on Facebook, IMDB score and amount the movie has grossed. The area charts have been used to make the analysis.

Conclusion:

The analysis shows that different movies rank in all different Years. Like Running Forever is the highest rated movie of the year 2015 while Kickboxer Vengeance is the highest rated movie in 2016 from IMBD score. The slicer lets us to choose the year.

CONCLUSION

The best movies are those who are well liked by people over various platform, highly rated and at the same time earn their maker's huge profit. Best director cannot be one for different categories different director excel. Same is the condition for actors and actresses. One may be excellent in crime thriller other may be best in comedy. Hence only by categorizing respective directors, actors with their respective field of work only then we can find top professionals. The analysis also shows that drama and thriller movies have clearly lesser budget and earn lesser than that of comedy action movies. Comedy and action movies have huge fan base while drama and thriller movies are only liked by people with deep mindset. These are few conclusions from the project.

BIBLIOGRAPHY

Wikipedia

Dataset from www.kaggle.com

Use tableau prep

Used MS EXCEL 2016

Information from Google and Wikipedia

Learnt from You Tube video