CS7DS4 Data Visualization 2019-20

Assignment 3

Declaration:
"I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.
I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write."

Student Name: Vivek Kumar
Student No: 19300755

Introduction

In this visualization assignment, Movies dataset is used to build an interactive dashboard in Tableau which provide better insight of data. The visualization is an explanatory, which focuses on data and present answers to some frequent queries related to movies like its budget, director, star, rating, etc. I have created three dashboard based on movie count, budgets and film studies.

Description

Dataset:

The dataset is taken from kaggle website (https://www.kaggle.com/danielgrijalvas/movies) which contains chunk of movies released from the year 1986 to 2016 in different countries based on different genre. The actual source of this dataset is IMDB.com from where the data is scraped. There are altogether 6820 movies in the dataset (1986-2016).

The features presented in dataset are:

Budget: the budget of a movie	*Company*: the production company
Director: the director	*Released*: release date (YYYY-MM-DD)
Country: country of origin	*Genre*: main genre of the movie.
Name: name of the movie	*Gross*: revenue of the movie
Rating: rating of the movie (R, PG, etc.)	*Runtime*: duration of the movie
Votes: number of user votes	*Year*: year of release
Score: IMDb user rating	*Star*: main actor/actress

- *Total_movies:* added to list of features to count the movies released year wise.

Snapshot of dataset:

budget	company	country	director	genre	gross	name	rating	released	runtime	score	star	votes	writer	year
6000	Next Wav	UK	Christoph	Crime	48482	Following	R	11/5/1999	69	7.6	Jeremy Th	72113	Christoph	1998
7000	Columbia	USA	Robert Ro	Action	2040920	El Mariach	R	2/26/1993	81	6.9	Carlos Gal	55226	Robert Ro	1992
7000	ERBP	USA	Shane Car	Drama	424760	Primer	PG-13	5/27/2005	77	6.9	Shane Car	81468	Shane Car	2004
15000	Can I Wat	USA	Rose Trocl	Drama	2408311	Go Fish	R	7/8/1994	83	5.6	V.S. Brodi	1742	Rose Trocl	1994
15000	Solana Fil	USA	Oren Peli	Horror	1.08E+08	Paranorm	R	10/16/2009	86	6.3	Katie Feat	195668	Oren Peli	2007
20000	Professor	USA	Ben Cocci	Drama	7221	Zero Day	UNRATED	1/23/2004	92	7.3	Cal Robert	3594	Ben Cocci	2003
22769	Desperate	USA	Gregg Ara	Comedy	692585	The Living	UNRATED	8/21/1992	81	6.5	Mike Dytr	2140	Gregg Ara	1992

Visualization Task / Encoding channels:

In order to have better insight of data, I have created 3 visualization dashboards which covers different tasks on same dataset in more appealing and effective way. The visualization task covered are Distinguish, Categorize and Rank. The movie data is distinguished based on different genre, year of release, budget, IMDb rating, casting and revenue. The data is categorized into different sets based on its company, genre, year, etc. As the list of company is huge, I have ranked only top 25 company which gave the most number of movies.
Visualization Encoding channels covered are size, position and colour as it covers almost all characteristic of channel used like distinguish, categorize and rank.

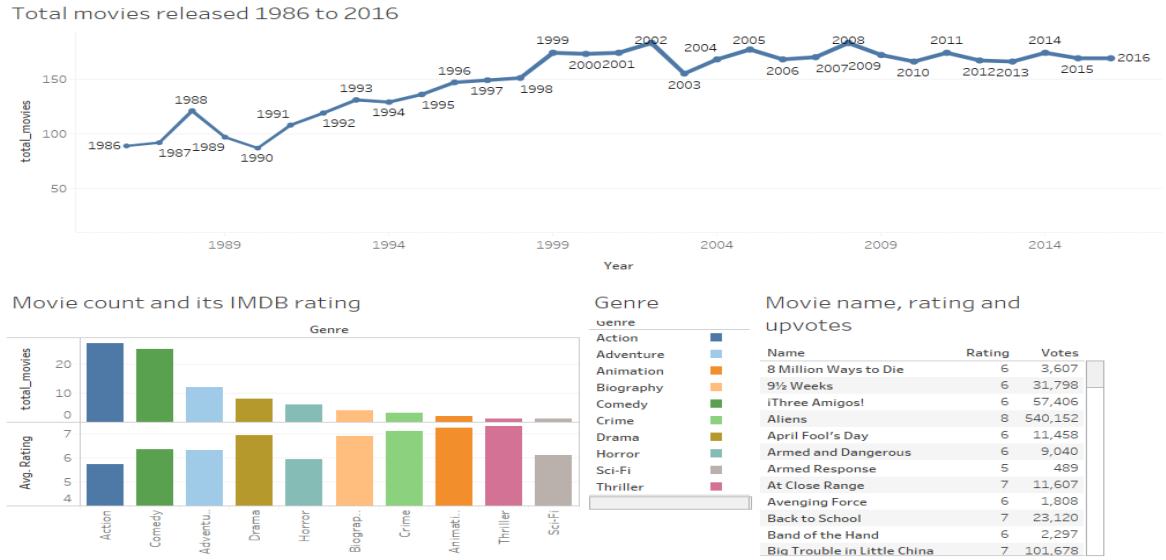
Approach:

In order to have an explanatory visualization of my data, I created particular set of question. Keeping them in mind, my visualization covered all aspect of features. The question framed are:

- Total list of movies released every year based on their genre and country.
- Country generating most revenue overall and genre wise.
- Genre of movie liked in a particular country.
- Film Studio producing maximum number of movies.
- Star that is casted most and different director with whom he worked with
- Least liked or produced Genre of movie

Dashboard 1:

Task: To analyze total movies released yearly from 1986 to 2016 based on genre, rating, etc.
Encoding Channels: Color, size, position

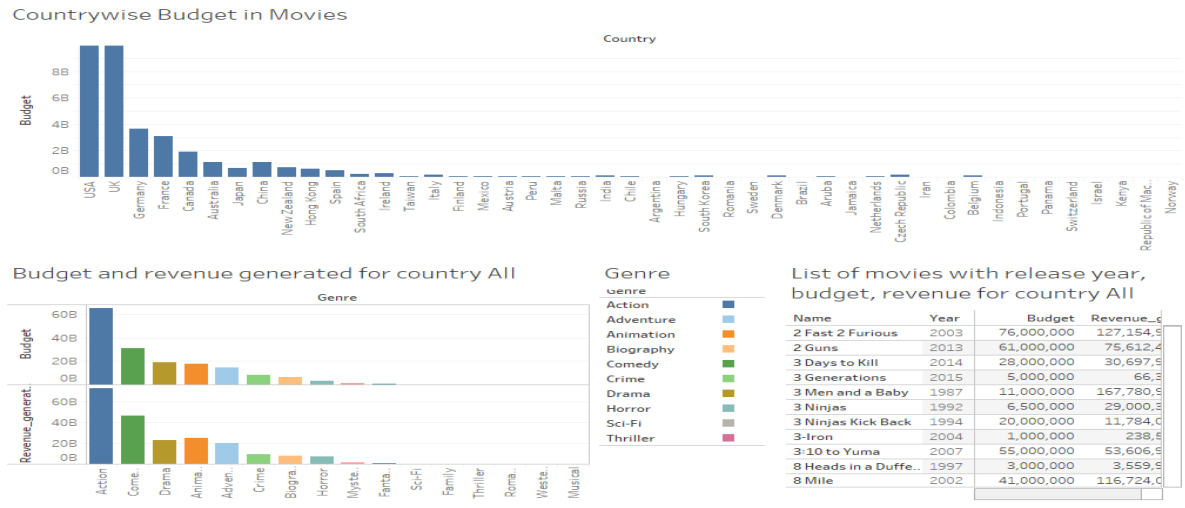


The first line graph shows total movies released in a particular year from 1986 to 2020, which can be further drilled down to total count of genre of movies in those years and the average IMDb rating of that genre. This data is presented in form of bar graph with different color depicting different genre of movie, and tabular column which show the list of movie, its rating and total upvotes received from user.
The thickness of line in line graph depicts total movies released in that year. Marks are labelled as year. In addition to this, we can also, view total movies count year wise based on the genre of movie by selecting particular genre from the list provided which will update the line, bar and tabular data likewise. As the dashboard is interactive, we can modify the condition to get an updated visualization of result, providing a better insight to data.

For example, I want all the list of movies released in year 1988 on genre comedy with thier IMDb rating. To have this filtered set of result, we will select 1988 from the line chart which will update all the 121 movies based on their genre in bar graph. From the bar chart, we can see total of 45 movies are released in genre comedy which upon selecting will update the tabular chart showing all the movies with their rating and votes. In this dashboard we can visualize different question related to movies count based on year, genre or rating.

Dashboard 2:

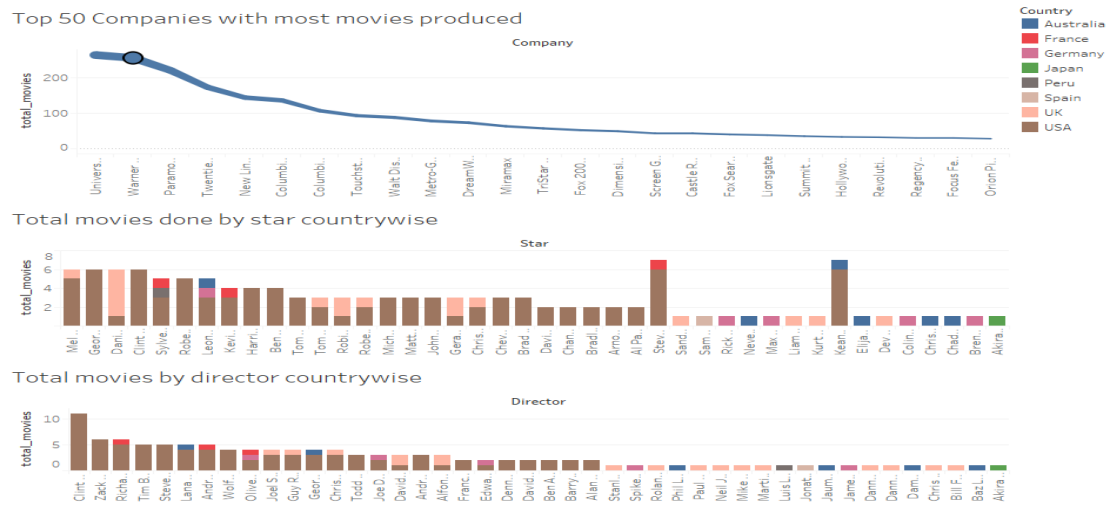
Task: Analyze total budget spent and revenue generated countrywise overall and genrewise
Encoding Channels: Color, size, position



It analyze the total budget spent, revenue generated, total movies released, etc. by different countries along with the genre. The filter used here is on countries and genre. The first bar/line graph shows budget spent (line chart) and revenue generated (bar graph) by movies country wise and upon hovering over the bar we can see total movie produced by that particular country. This can be further drilled down to analyzing budget and revenue generated based on the genre. This data is presented in form of bar graph with different color depicting different genre of movie, and tabular column which show the list of movie, its release date, budget and revenue generated. In addition to this, we can also, view total movies produced countrywise based on the genre of movie by selecting particular genre from the list provided which will update the bar and tabular data likewise.
For example, I can see that USA has produced the maximum number of movies i.e. 3726 in these year with 892 movies on action genre spending budget of 55,033,253,376 and generating revenue of 63,405,545,179 (viewed in tabular data)

Dashboard 3:

Task: To analyze total movies produced by different film studios and stars casted.
Encoding Channels: Color, size, position



Here, we analyzed the total movies produced by different film studios, total movies stars worked and total movies done by directors in different countries. The filter used here is on companies, stars and directors. The first line graph shows total movies produced by different film studios (viewed in thickness of line), which can be further drilled down to list of star casted in them along with directors who produced the movie with color variation showing different countries. Upon selecting the particular director, we can see the star which worked with them and vice versa. Different color in bar graph shows different countries where they directed the movie. We can get the details on hovering over the bars.
For example, I can see that Universal Pictures has produced the maximum number of movies i.e. 265, which cased star Van Diesel in 6 of the movies which are directed by four different directors (Justin Lin, Rob Cohen, James Wan and David Twohy) and all these movies are done in USA.

Conclusion:

We can answer various question related to move using this visualization. Few insight from these visualization are:

- USA produced maximum number of movies.
- Universal film studio has produced maximum number of movies
- Overall, Bibliography genre has the highest rating
- Naomi Watts did maximum number of movies i.e. 11
- Woody Allen directed maximum movies i.e. 30, and in USA alone its 26.
- Movies in action genre is most produced with maximum count in USA i.e. 892

References

[1].Dataset: https://www.kaggle.com/danielgrijalvas/movies

[2].The tool used for visualization is Tableau