# CSE 587 Data Intensive Computing
# Term Project
# Communicating the Results of Data Analytics

Due Date: May 7th 2017

Group
Arnav Ahire [5020 8006] UBIT-arnavane
Ritika [5020 6346] UBIT-ritika

## Dataset used:

*Name:* IMDB 5000 Movie Dataset

*Source*: https://www.kaggle.com/deepmatrix/imdb-5000-movie-dataset

*File:* imdb-5000-movie-dataset.zip
This zipped folder contains file movie_metadata.csv which we will use for this project.

*Attributes Used:*

1. Director Name: Name of the director.
2. Director Facebook Likes: The count of 'Likes' received by director on Facebook.
3. Actor 1 Facebook Likes: The count of 'Likes' received by the lead actor on Facebook.
4. Gross: Gives the gross income earned by a movie as number.
5. Genres: Gives the names of the Genres a movie falls into.
6. Actor 1 Name: Gives the name of the lead actor.
7. Plot keyword: The '|' separated keywords that describe the plot of the movie.
8. Num User For Reviews: Total number of users that reviewed the film.
9. Num Critics For Reviews: Total number of critics that reviewed the film.
10. Country: The name of the country to which the movie belongs.
11. Imdb Score: Gives the IDMB score in decimals.
12. Movie Facebook Likes: The count of 'Likes' received by movie on Facebook.
13. Title Year: The year in which the movie was released.

The data set also contains other fields like Budget, Content Rating, Aspect Ratio, Color etc. However, we have not used these fields for analysis so we won't be discussing about them.

## Work Environment Used:

Name: Tableau 10.2

## Problem to be solved and it's scope:

We have always been a hardcore fan of movies and as group mates we wanted to do something related to movies as that is something that excites us. That's when we decided to go with this dataset and go with the question that we always had in our minds: *'What could be the factor that makes a film so successful?'* So throughout our analysis we tried to figure out what factors could contribute to making a film successful. In order to do that we took the data set and gave it to Tableau.

## Inputs to Tableau:

File: movie_metadata.csv (present in imdb-5000-movie-dataset.zip)

## Tableau Workbook:

Name: MovieDataset

## Sheets Used: (Sheet Names)

1. Movie Success
2. Total Viewership
3. Movies with good IMDB score but low Gross Income in US and UK
4. Avg. IMDB score for Movies by Genre
5. Other Factors Check
6. Plot Keyword vs Gross
7. Predicted Total Gross of Movie industry
8. Predicted Max IMDB score for any Movie
9. Predicted Avg Number of Users for Reviews
10. Regions based on Movie Success
11. Top 10 Successful Directors

Dashboards Used: (Dashboard Names)

1. Movie Success Dashboard
2. Total Viewership Dashboard
3. Outliers for US and UK Dashboard
4. IMDB vs Genre Dashboard
5. Other Factors Check Dashboard
6. Prediction Dashboard
7. Interesting Facts Dashboard

Story:

Name: What factors affect the success of a film?

Assumption: Success of the film is measured by the Gross income of the Movie.

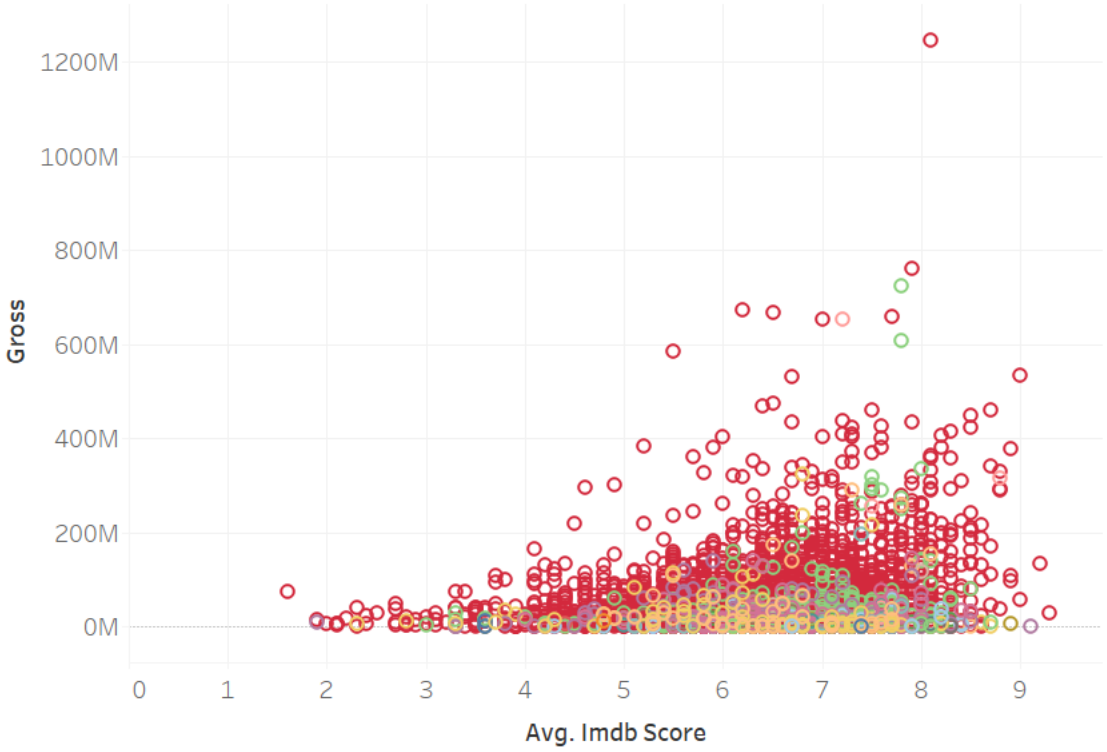
*Story-point 1:*

Dashboard Used: Movie Success Dashboard

Sheets Used: Movie Success

What We Did:

1. We created a plot of Avg. Imdb Score vs Gross to see how does Imdb score affect the Gross income of a movie.



Observation:

1. The Gross income of a movie increases with increase in IMDB score. However, there are outliers to this. As we go above a score of 5.0 in IMDB, we see that there are a lot of movies who still have less Gross income which includes most of the films from countries other than US and UK and also a lot of films from US and UK as well.

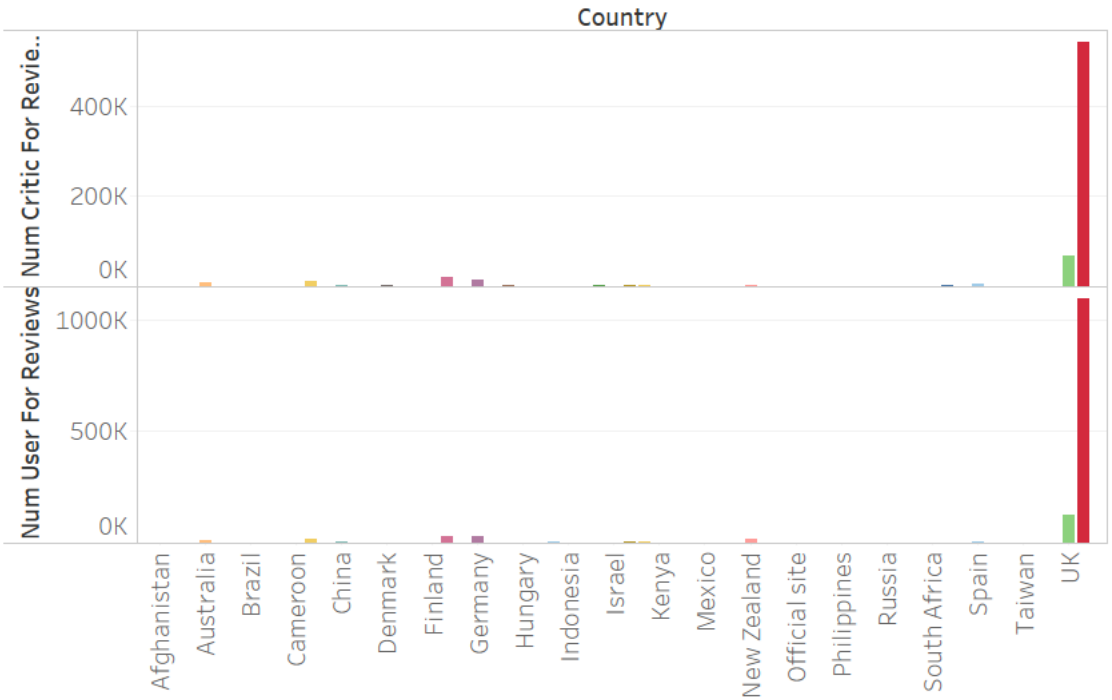
*Story-point 2:*

Dashboard Used: Total Viewership Dashboard

Sheets Used: Total Viewership

What We Did:

1. In order to understand the outliers obtained in the previous graph, we thought of observing the viewership in each country to see if the viewers count affects the movie's success.
2. Hence we created a plot of Country vs Num User For Reviews and Num Critics For Reviews. This gave us the information of all the critics and general audiences in a country that review a film.




Observation:

1. The number of general users that review the film for all other countries is significantly less than that of US. Only UK has the second highest number of users that review the film. For other countries since very few people have reviewed the film, very less people must be having the knowledge about the

film and so many of them might not have seen it and so it's possible that the film went unnoticed.

2. Still we know that there are several films from US and UK who have earned less. We will analyze this ahead.

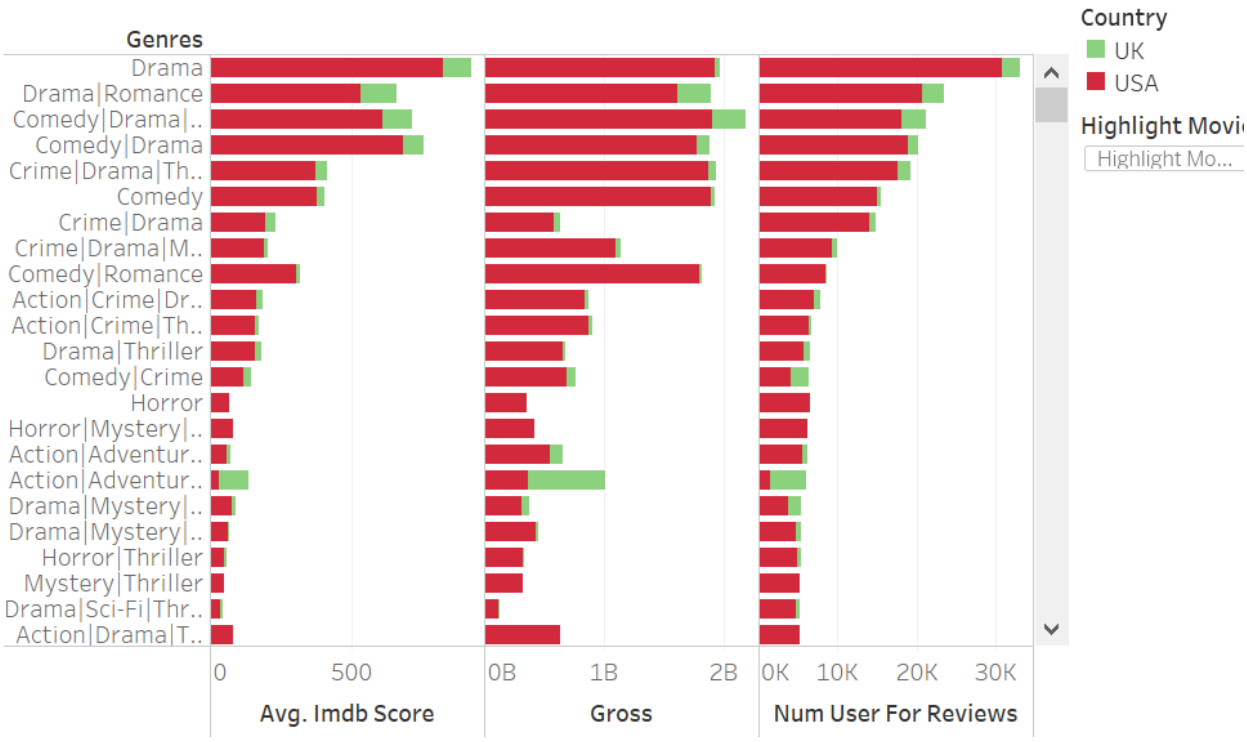
*Story-point 3:*

Dashboard Used: Outliers for US and UK Dashboard

Sheets Used: Movies with good IMDB score but low Gross Income in US and UK

What We Did:

1. In order to analyze the situation where the movies from US and UK were failing to earn enough income, from the total set of points in our first graph, we selected those points that were corresponding to more IMDB values and less Gross and filtered only US and UK related data and we created a plot of Avg. Imdb Score, Gross, Num User for Reviews vs Genre of the movies. We did this because we wanted to see whether the film's Genre is affecting the film's grossing in countries like UK and US.



Observations:

1. Most of the movies that failed to earn enough despite of good IMDB scores belonged to the first 6 genres, the common genres being Comedy, Drama, Romance and Crime.
2. We will analyze this further.

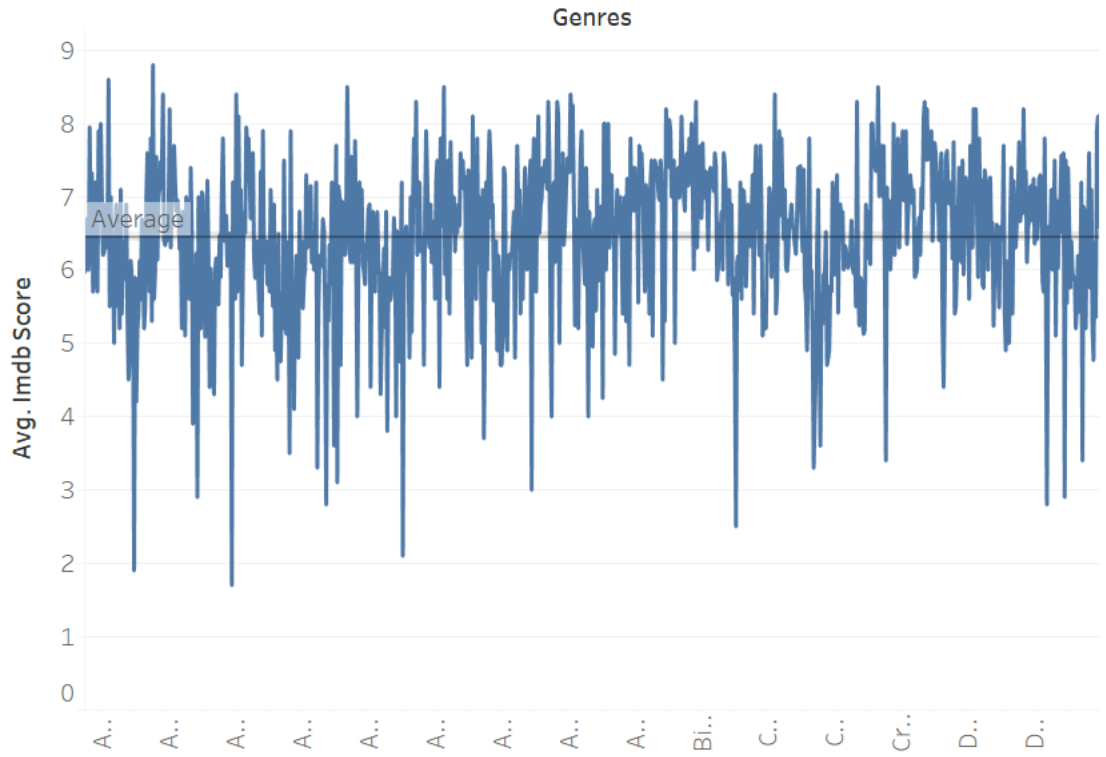
*Story-point 4:*

Dashboard Used: IMDB vs Genre Dashboard

Sheets Used: Avg. IMDB score for Movies By Genre

What We Did:

1. We created a plot of Genre vs IMDB score in order to understand if there is any relation between them.
2. We then computed the average IMDB score for all the genres and plotted it as a line using Model 'Average with 95 % CI'. So it came out to be 6.452.



Observations:

1. By adjusting the values for our genres i.e Comedy, Drama, Crime and Romance we found out that all of them had an average IMDB of around 6.5 to 7.2 which was close to out average.
2. Thus we can deduce that the genres were pretty average, which caused the films in US and UK to not earn so much despite of having a good IMDB score.

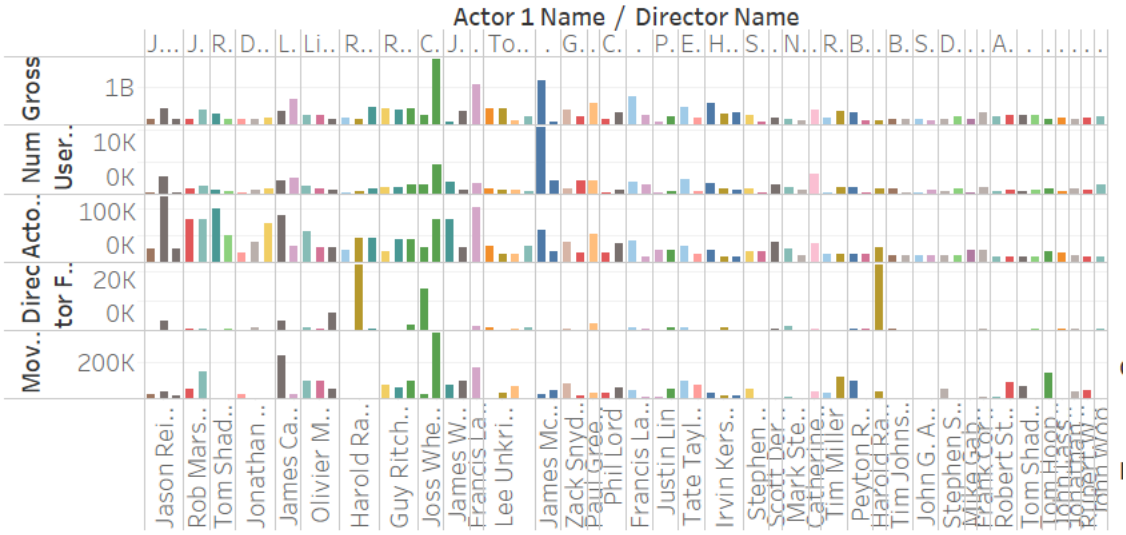
*Story-point 5:*

Dashboard Used: Other Factors Check Dashboard

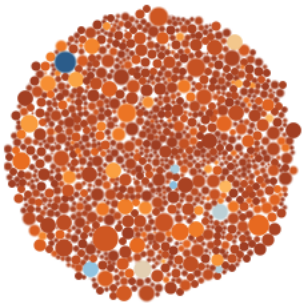
Sheets Used: Other Factors Checked, Plot Keyword vs Gross

What We Did:

1. We wanted to see if other factors also significantly affect the success of a film and hence we created two different plots.
2. In the first plot we check for various actor-director combinations and for them we check how similar is the graph of Gross with other graphs such as Num User for Reviews, Actor1 Facebook Likes, Director Facebook Likes, Movie Facebook Likes.
3. In second we created a bubble chart to see if plot keywords affect the movie income.




## Plot Keyword vs Gross



Here we checked for the different factors other than IMDB score and viewership that could affect the film success. First we checked whether the popularity of the lead actor/director/movie(franchise) affects a movie's income, however we could also observe here that the graph of Gross was similar to that of average number of user reviews. Similarly we tried to see if any keywords that are related to movie plot affect the movie's success.

Observations:

1. In the first graph, although Actor 1 Facebook Likes is a good factor, it's not that effective since it lacks consistency.
2. Also Director Facebook Likes is not a good factor which is clearly evident. However, Movie Facebook Likes is a good factor, and the best is Num User for Reviews.
3. In the second graph (bubble chart) we created Plot Keyword vs Graph plot where we tried to see if the particular keywords related to plot of the movie actually affects it's grossing.

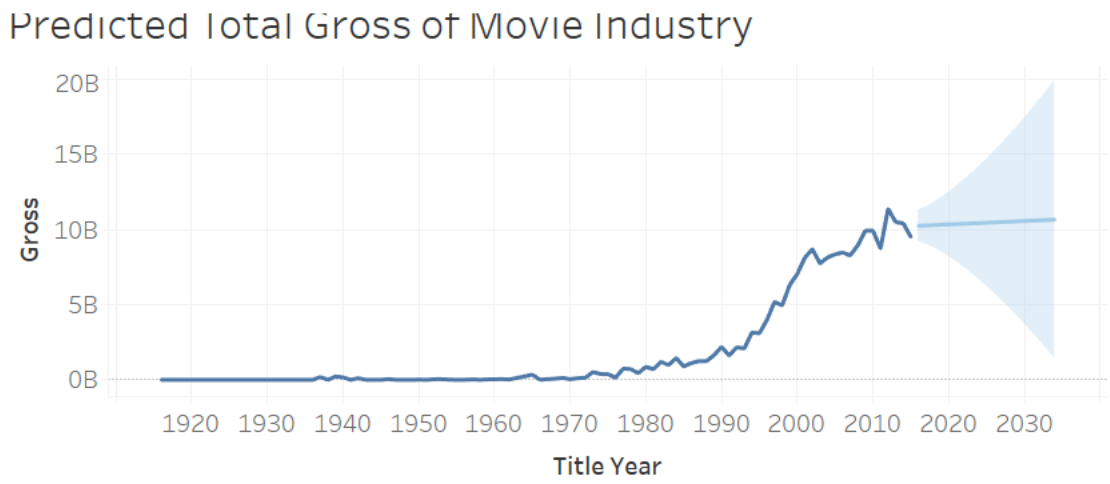
*Story-point 6:*

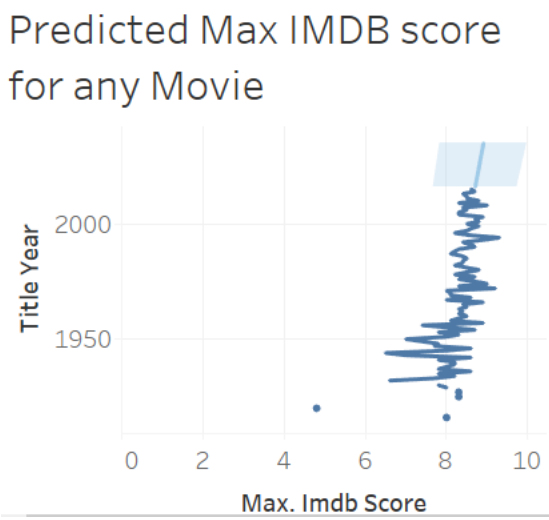
Dashboard Used: Prediction Dashboard

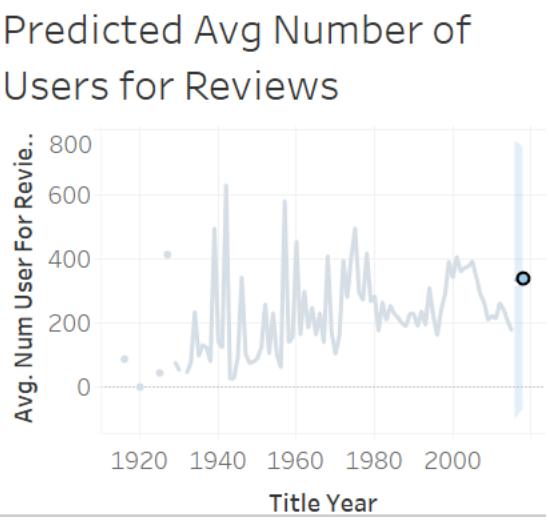
Sheets Used: Predicted Total Gross of Movie industry,
Predicted Max IMDB score for any Movie,
Predicted Avg Number of Users for Reviews

What We Did:

1. Based on the observations and analysis, we found out that the important factors to be predicted by our system should be related to Gross income of the movie industry, the Maximum IMDB score which will help us obtain the information of how successful our future films are going to be and last but not the least, the average viewership in future.
2. So we used the 'Forecast' feature of Tableau for this purpose and we set it on each graph respectively.




Observations:

1. We got the predictions as we can see in the above plots.
2. We can see positive estimates for each case where-in we will be noticing big Gross incomes that will be earned by movie industry, along with improved IMDB scored movies that have increased viewership than before.

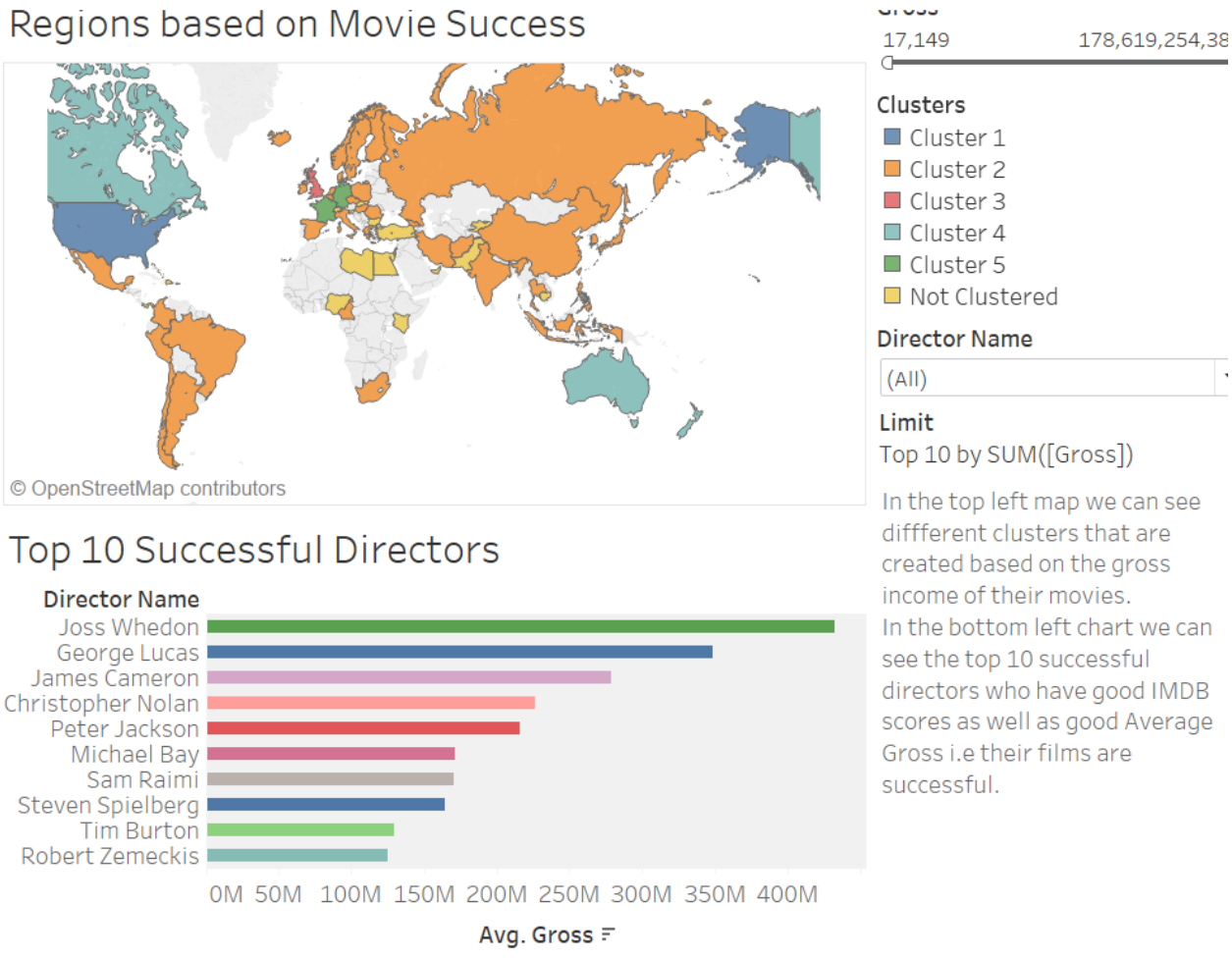
*Story-point 7:*

Dashboard Used: Interesting Facts Dashboard

Sheets Used: Regions based on Movie Success, Top 10 Successful Directors

What We Did:

1. In the first part we created different clusters in the world map according to the total gross income earned by the movies in those clusters.
2. We used the 'Cluster' feature of Analytics tab in Tableau for this.
3. In the second part we created a list of top 10 Hollywood Directors that direct movies that not only receive good IMDB rating but also are highly successful and earn enough.



Observations:

1. In the first part we get a beautiful map where-in we can clearly specify the regions that have more successful films than others.
2. In the second part we get a list of directors that are responsible for creating successful films that are not only well-reviewed but have earned well and hence are successful.

## Conclusion:

We performed the following:

1. We started with finding a relation with Gross income and a movie's IMDB rating.
2. Then we went ahead to understand the outliers in this relations and in finding those outliers we discovered another important factor i.e Num User for Reviews which tells us that the number of people who reviewed it affected the gross income of the movie.
3. We tried to find out other factors that could contribute to a film's success.
4. We finally created a prediction for the important factors such as IMDB score, Num User for Reviews and also the gross income of the movie industry.
5. We then checked for some interesting facts related to or data set where we got to see the features of Tableau such as 'Cluster' as well.

Thus we conclude by saying that the prime factors affecting the success of the movies are:

1. IMDB Score
2. Num User For Reviews

## Project Link:

https://public.tableau.com/profile/arnav.ahire#!/vizhome/MovieDataset/Story