**IMDB MOVIE RATINGS**

**By  
 Abhishek Shah**

1. **Data set URL:**

The dataset I am using for this project is downloaded from Kaggle website. The dataset is about IMDB movie ratings. The dataset contains 5043 movie data with 28 variables. The data is collected from IMDB website using python library called “Scrapy”. The data consists of 28 variables for 5043 movies and 4906 posters (998MB), spanning across 100 years in 66 countries. The data was uncleaned with lot of missing values as well as duplicates which were cleaned and the final cleaned dataset contains 4893 rows. The 28 variables in the dataset are as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| Movie Title | Actor\_3\_facebook\_likes | Cast\_total\_facebook\_likes | Budget |
| Color | Actor\_2\_name | Facenumber\_in\_poster | Title\_year |
| num\_critic\_for\_reviews | Actor\_2\_facebook\_likes | Plot\_keywords | Imdb\_score |
| Movie\_facebook\_likes | Actor\_1\_name | Movie\_imdb\_link | Aspect\_ratio |
| Duration | Actor\_1\_facebook\_likes | Num\_user\_for\_reviews |  |
| Director\_name | Gross | Language |  |
| Director\_facebook\_likes | Genres | Country |  |
| Actor\_3\_name | Num\_voted\_users | Content\_rating |  |

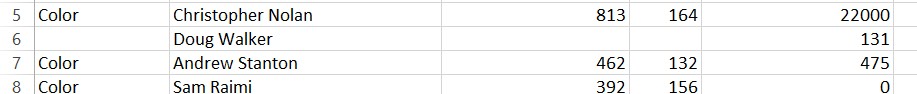
The URL of my dataset is:

<https://www.kaggle.com/deepmatrix/imdb-5000-movie-data;set/downloads/movie_metadata.csv.zip>

1. **Data Cleaning**

**Missing Values:** There were lot of missing values in my data in various columns. I checked this using Excel and applied its filters on every column. I found “Blanks” in those columns and removed those rows.

**Before:**

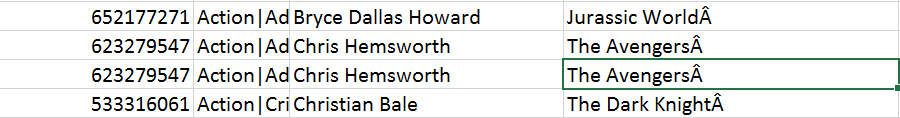


**After:**

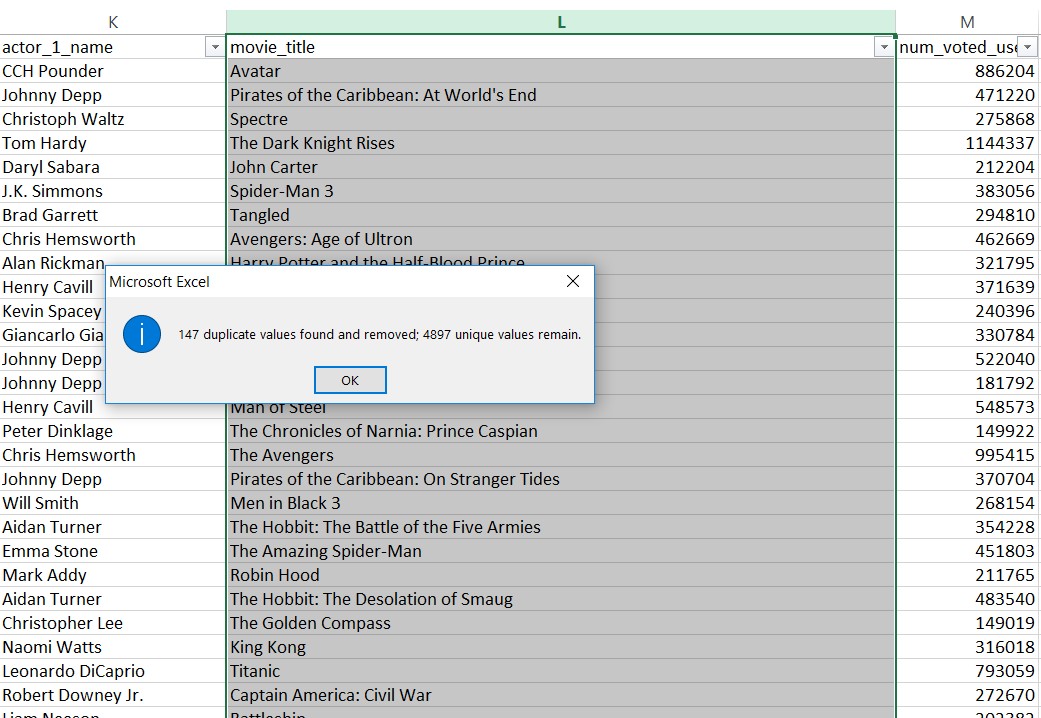


**Duplicate values:** There were duplicate values too in my dataset and I removed those using “Remove Duplicates” option in DATA tab in Excel. I found that the movie\_title coloumn had 147 redundant values and I removed those using above steps it. After removing those only 4897 unique values are remaining.

**Before:**

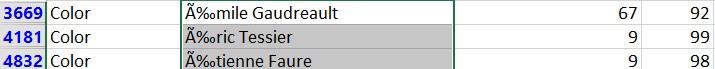


**After:**

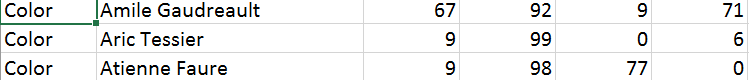


**Illegal values:** There were some values such as “A%mile Gaudreault” which was cleaned using find and replace and cleaned to “Amile Gaudreault” using Find and Replace option in Excel.

**Before:**

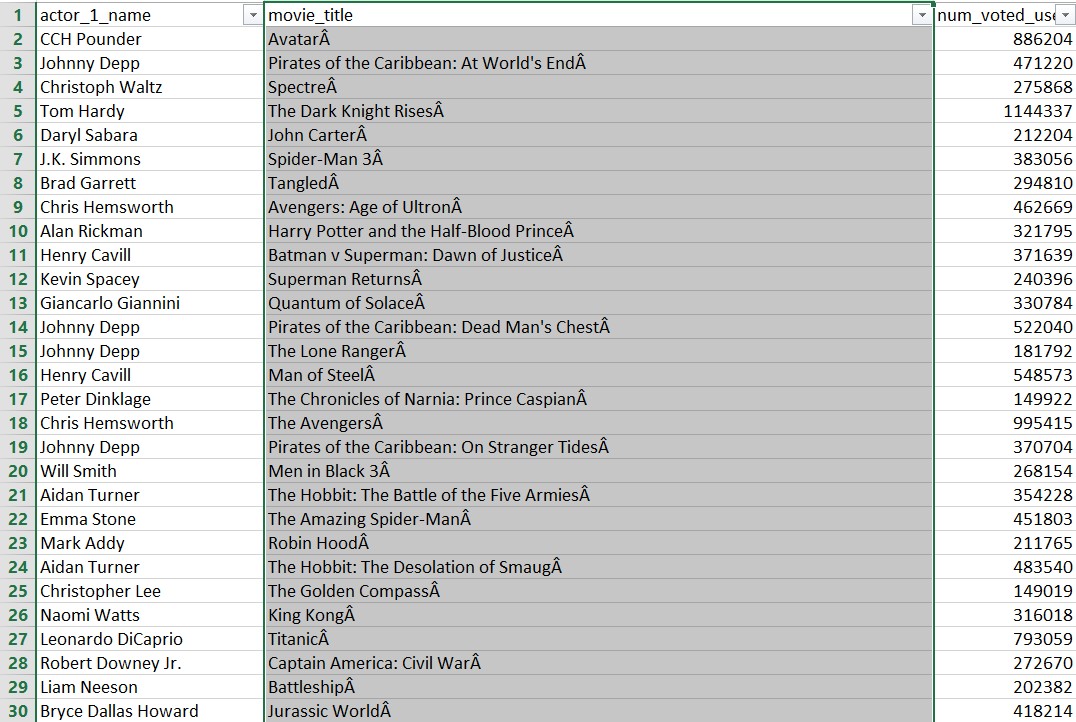


**After:**

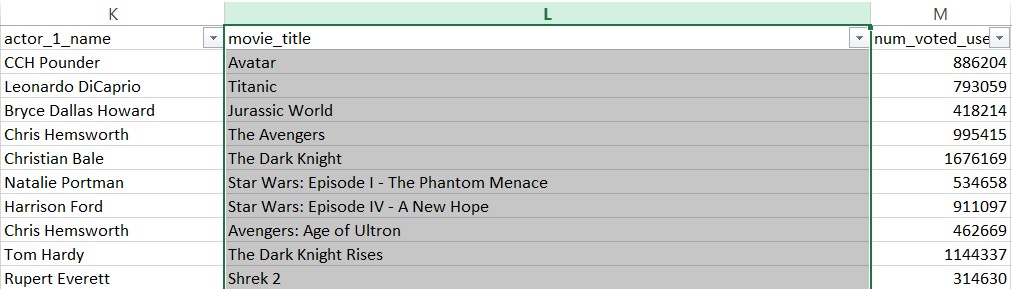


**Misspelling:** There were various Misspelled words too such as “Jurassic ParkÃ” which was converted to “Jurassic Park” using Find and Replace option in Excel.

**Before:**

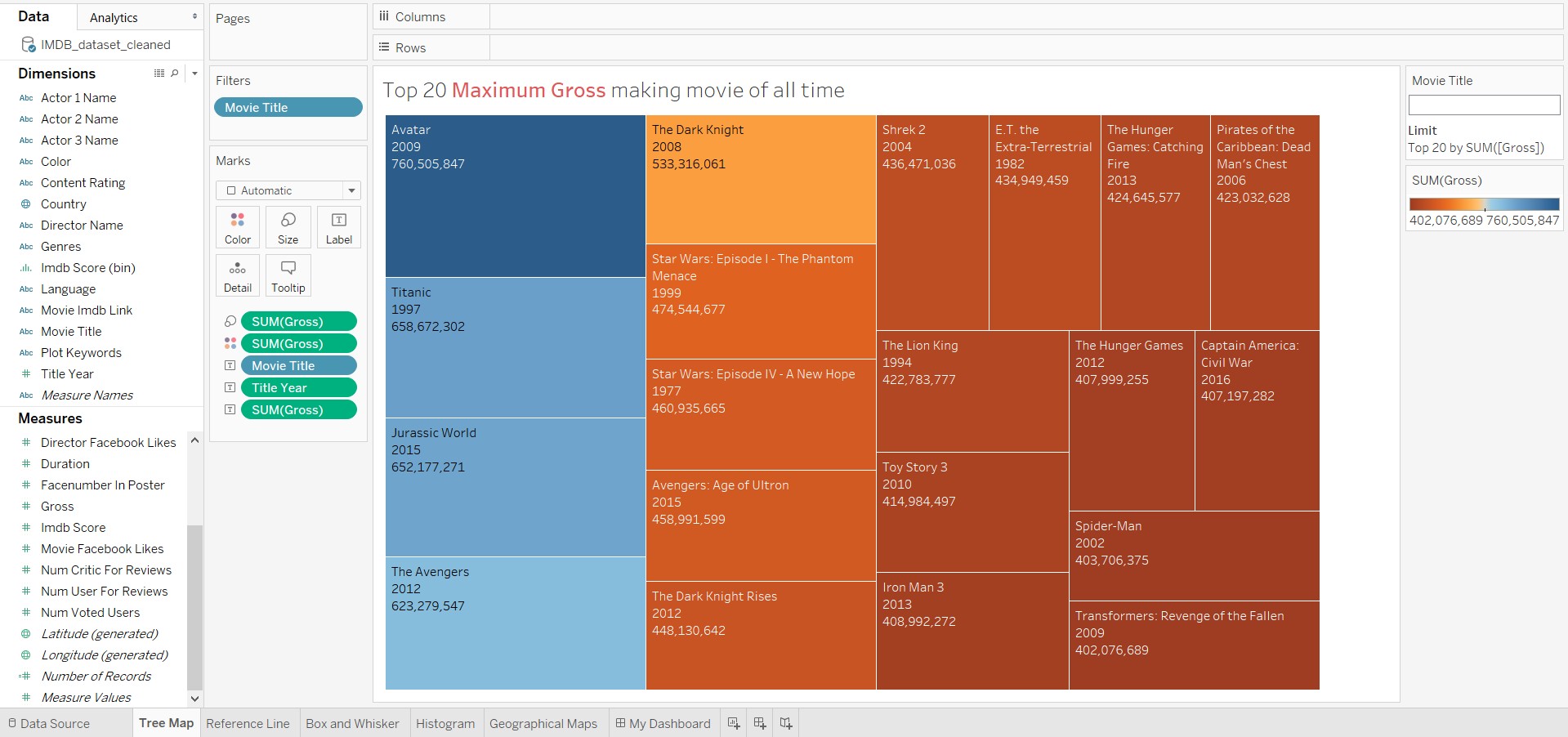


**After:**



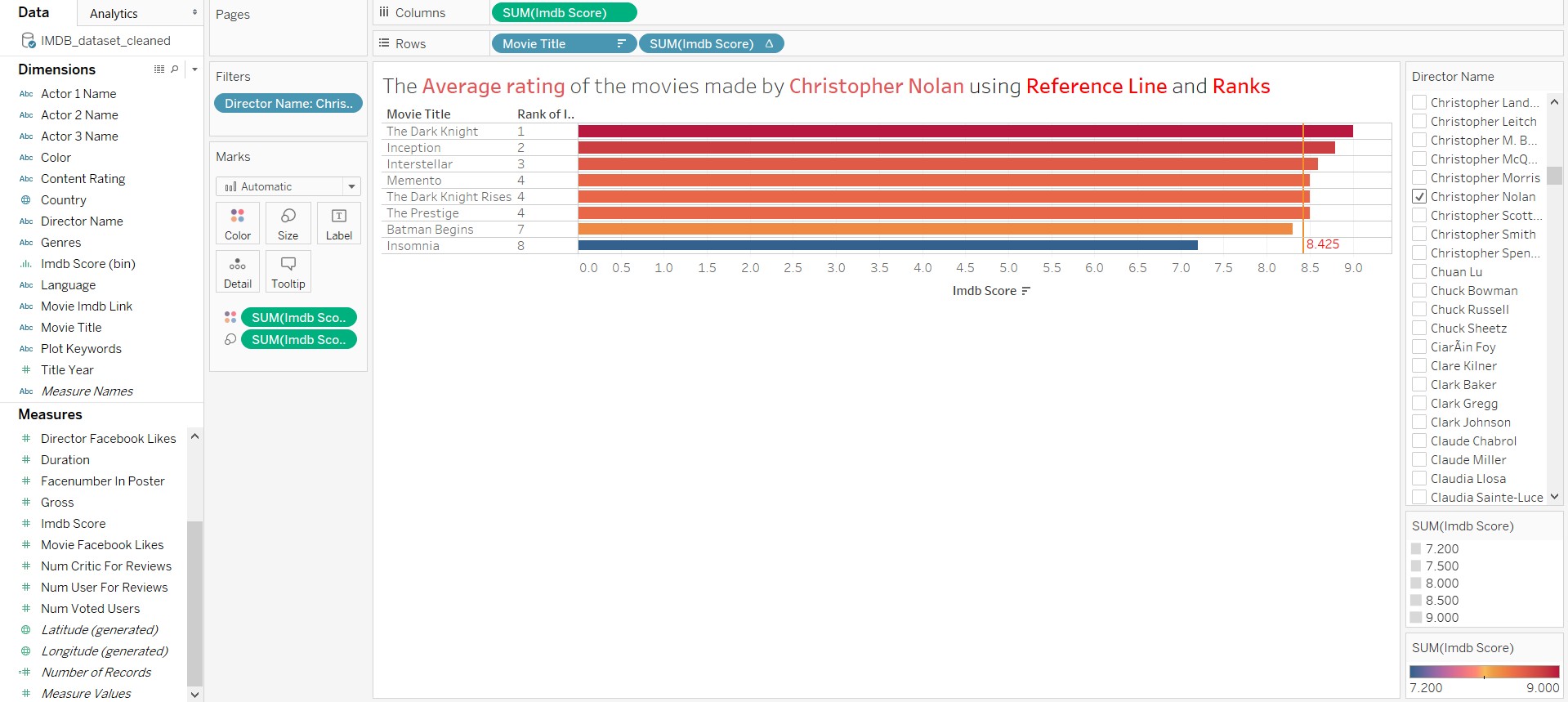
1. **Data Visualizations**
2. Top 20 maximum gross making movie of all time

I have used **Tree Map** to depict top 20 maximum gross making movies of all time. It tells us the most (depicted in Dark blue color) as well as the least (depicted in red color) gross making movie amongst top 20.



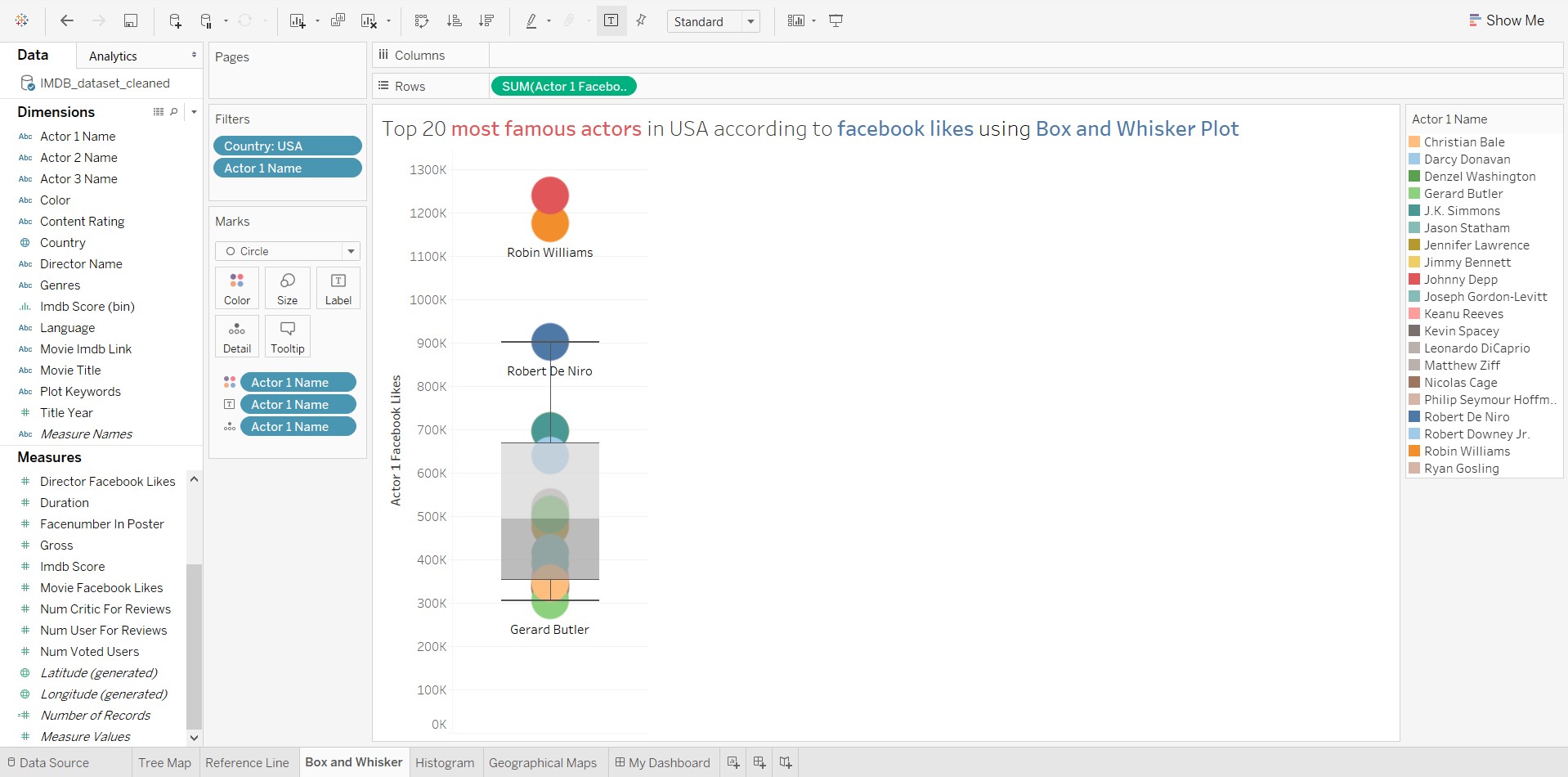
1. The average rating of the movies made by Christopher Nolan

I have used **Horizontal Bar** chart to depict the average rating of movies that were made by Christopher Nolan. The average rating is 8.425 which was calculated using **Reference Line.** I have also found out the **Rank** of movies using IMDb score which results that highest rated movie is “The Dark Knight”



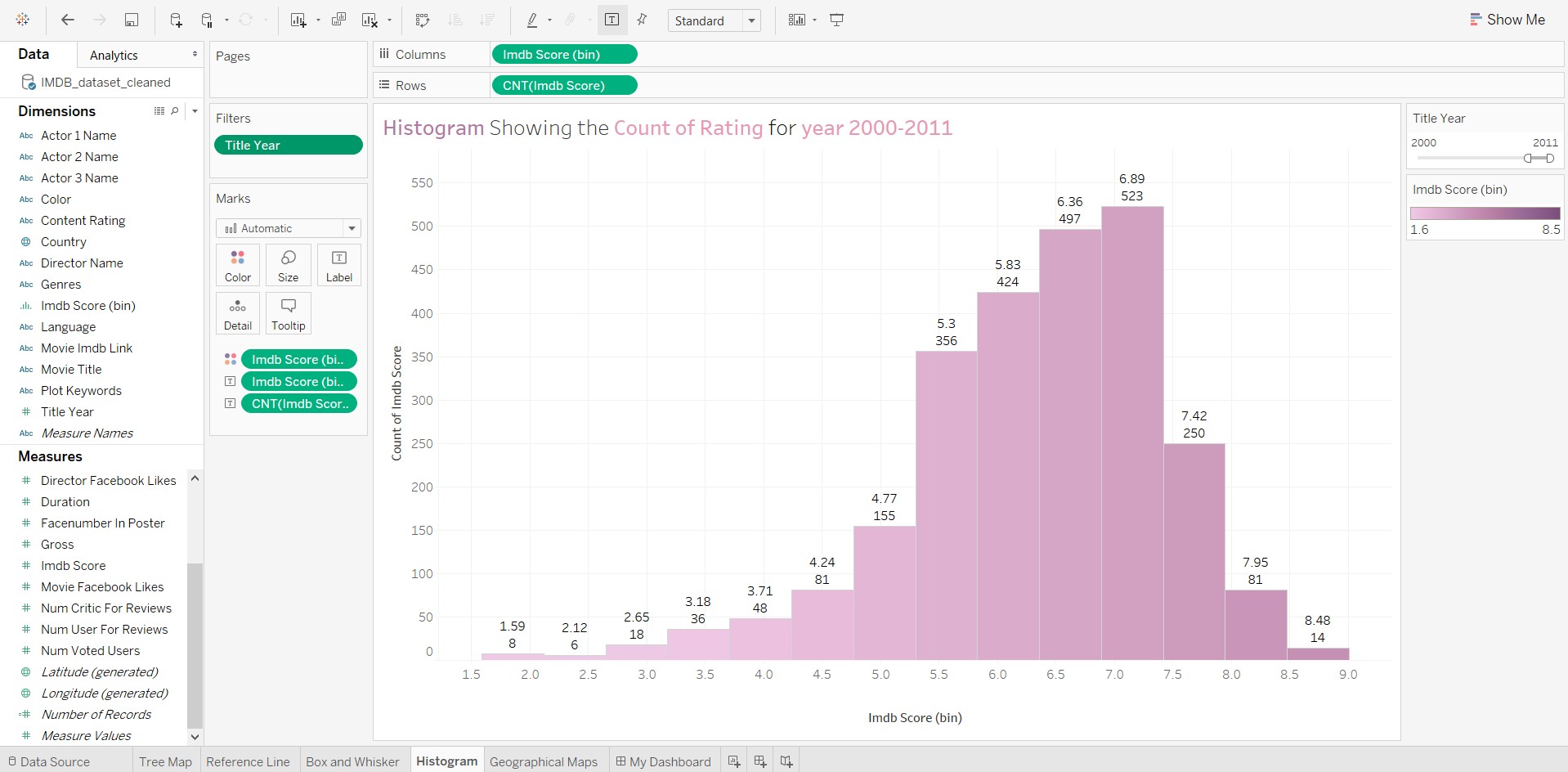
1. Top 20 most famous actors according to Facebook likes

I have used **Box and Whisker** to depict top 20 famous actors according to Facebook likes. Some of the actors such as Johnny Depp(1240000 total likes) and Robin Williams (1176000 likes) are extremes. The actor Robert De Nero is on Upper Whisker (902000 likes) and Gerard Butler on lower whisker (306000).



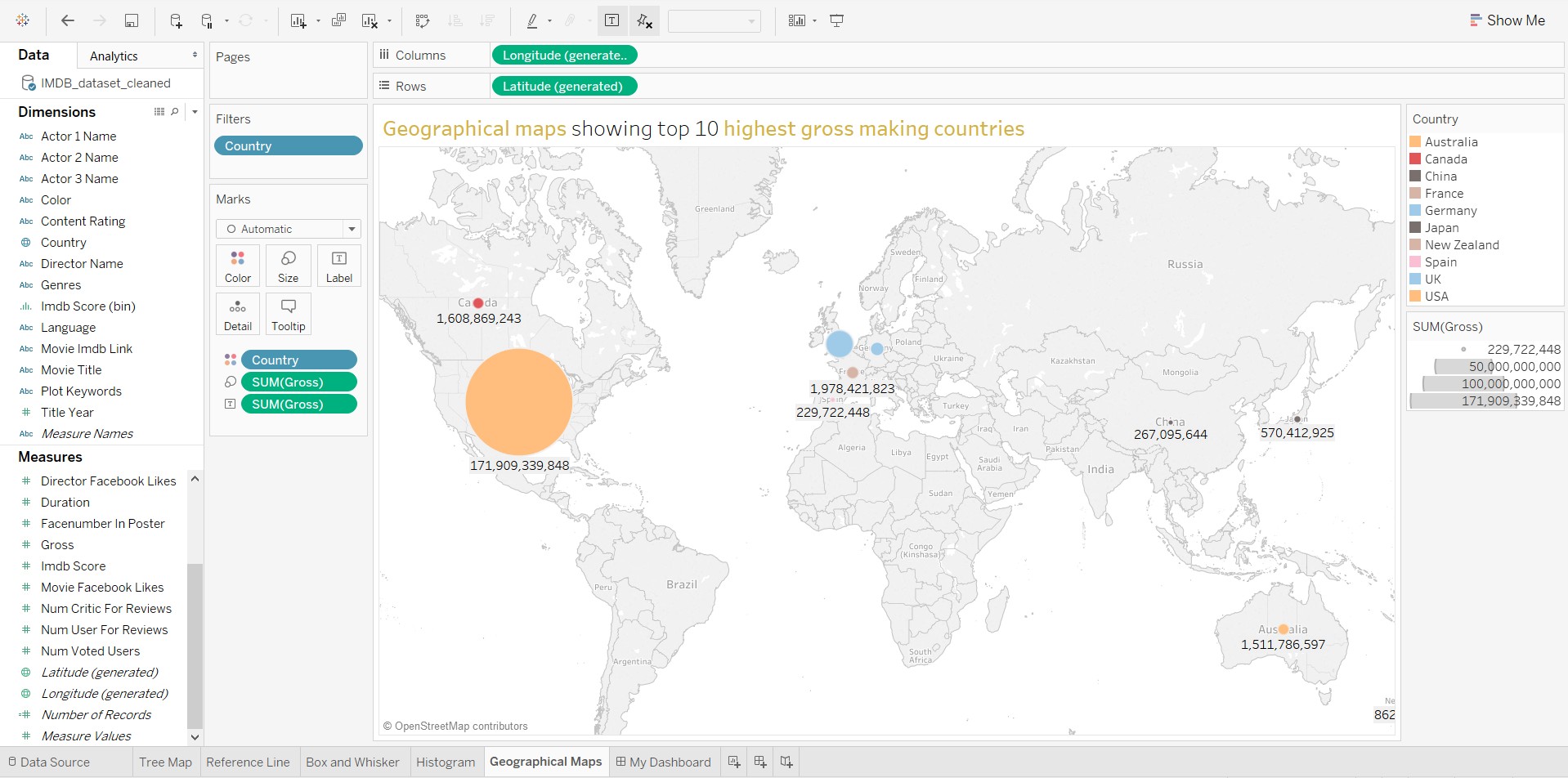
1. Count of Average Ratings from year 2000 to 2011

I have used **Histogram** to depict these counts of ratings from year 2000 to 2011. The total number of 523 movies with an average rating of 6.89 in span of 11 years whereas there are just14 movies with the highest rating of 8.48 and 8 movies with the score of 1.59.



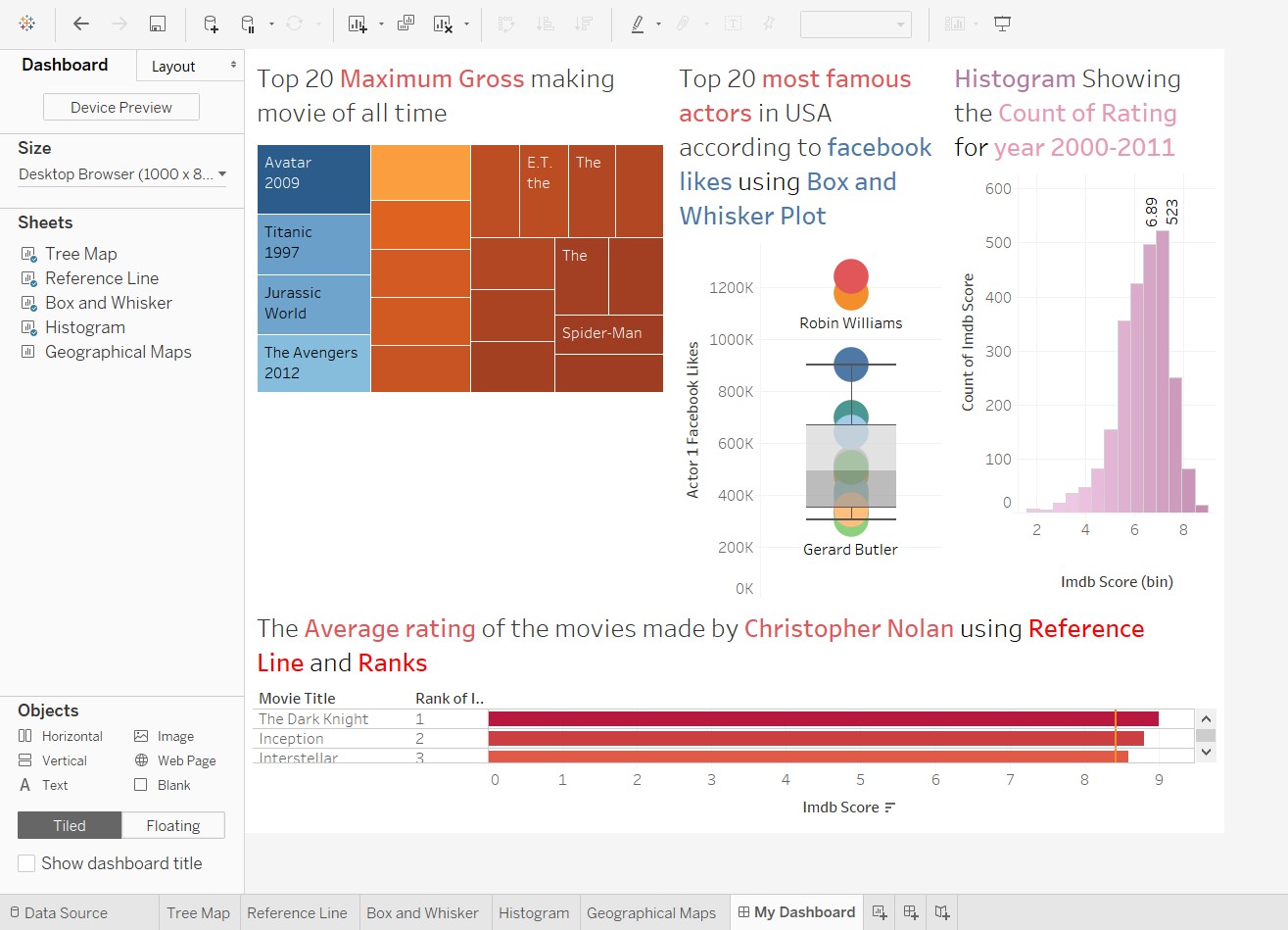
1. Top 10 countries based on total gross of the movies

I have used **Geographical** Maps to depict this. The country USA has highest gross income as 171,909,339,848 and lowest is Spain with gross income of 229,722,448.



1. **Dashboard**

I have put 4 sheets on dashboards which are **Tree Map** which depicts top 20 movies that make maximum gross income, **Box and Whisker** which explains top 20 famous actors which has highest facebook likes, **Histogram** which demonstrates count of average ratings of movies from year 2000 to 2011 and **Horizontal Bar** which depicts average ratings of movies made by Christopher Nolan.



1. **Story telling:**

**How it all started?**

Movies have always been a part of great entertainment and fun. The ratings of the movies by critics as well as people determine how popular the movie can be. A high IMDb rating makes an insignificant contribution towards a movie's success. A film's genre, stars, marketing, release schedule, and competition at the box office all matter far more than movie ratings or even reviews. Therefore, Box Office Mojo and others can fairly accurate predict a film's initial box office revenues even before it's been seen by much of the public. I have always used in the past to determine if a movie is worth watching is to use its IMDB rating. Except for few comedies (e.g. Monty Python) I have always enjoyed any movie in the IMDB TOP 250.

**Why does rating matter?**

Some people are more knowledgeable about the worth of movies. Some people are better able to differentiate a good from a bad movie. Some people are more consistent in their ratings over time. Should ratings from people who are "better" at rating movies be given more worth. One more aspect can be that the movies can open children’s eyes to new places, cultures and ideas, and parents have an important role in ensuring that experience is positive and enriching. They need the tools to decide what movies are suitable for their children to watch. From understanding how movie ratings work and have evolved with the times to helping parents find movies appropriate for their kids, we want to help make movie-going a positive experience in your family’s life.

**Questions to choose this topic:**

1. How can we tell the greatness of a movie before it is released in cinema?
2. Given that thousands of movies were produced each year, is there a better way for us to tell the greatness of movie without relying on critics or our own instincts?
3. Will the number of human faces in movie poster correlate with the movie rating?

**Currently, what are we up to?**

We are trying to analyze different aspects and top rated movies using different measures. A decent dataset which might contain hundreds of attributes from 50K or more movies might require tons of feature engineering. For comprehensive movie analysis and accurate movie ratings prediction, 28 attributes from 5000 movies should be enough.

**Top 20 movies based on Gross Profit**

This story point determines the top 20 movies amongst the whole data and uses Gross Profit as the measure. This analysis help us to determine that which movie is popular based on how much it earns. Gross variable is important because it determines that how much a will earn and be profitable.

**The average rating of the movies made by Christopher Nolan**

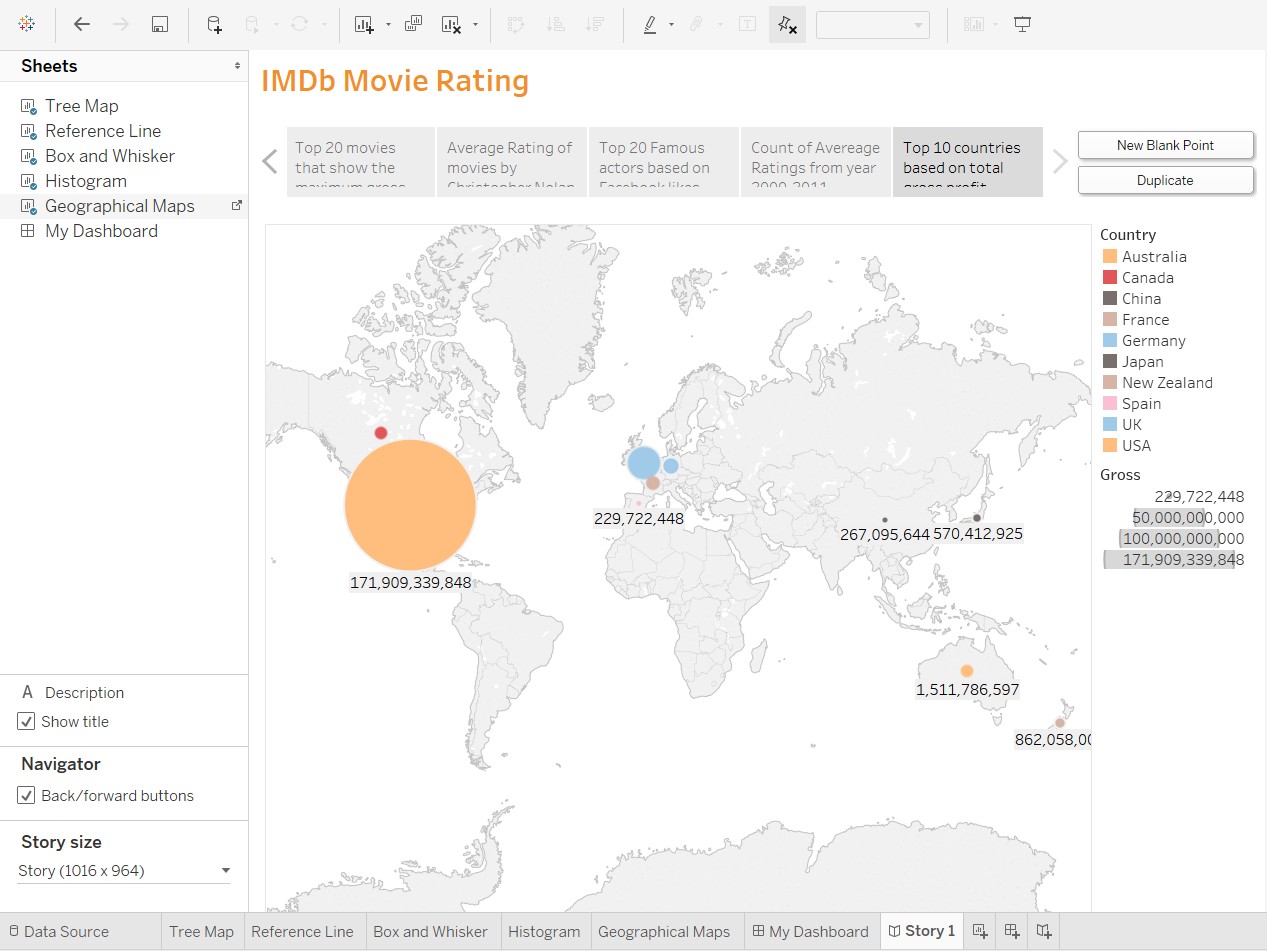
The average rating of the movies made by Christopher Nolan is 8.245. The Christopher Nolan is famous director and it helped us determining the ratings of movies that were developed by Christopher Nolan. This feature enables people to make anlysis based on directors too. This also gives us the brief idea about ranking of movies by Christopher Nolan.

**Top 20 Famous actors**

Many people are interested in knowing that which actors are most famous and their ratings or movies. The actors which has highest facebook likes are tend to be more popular and tend to make more money whose movies also become popular. So, in this story point we have determined the top 20 famous actors based on highest facebook likes as a measure.

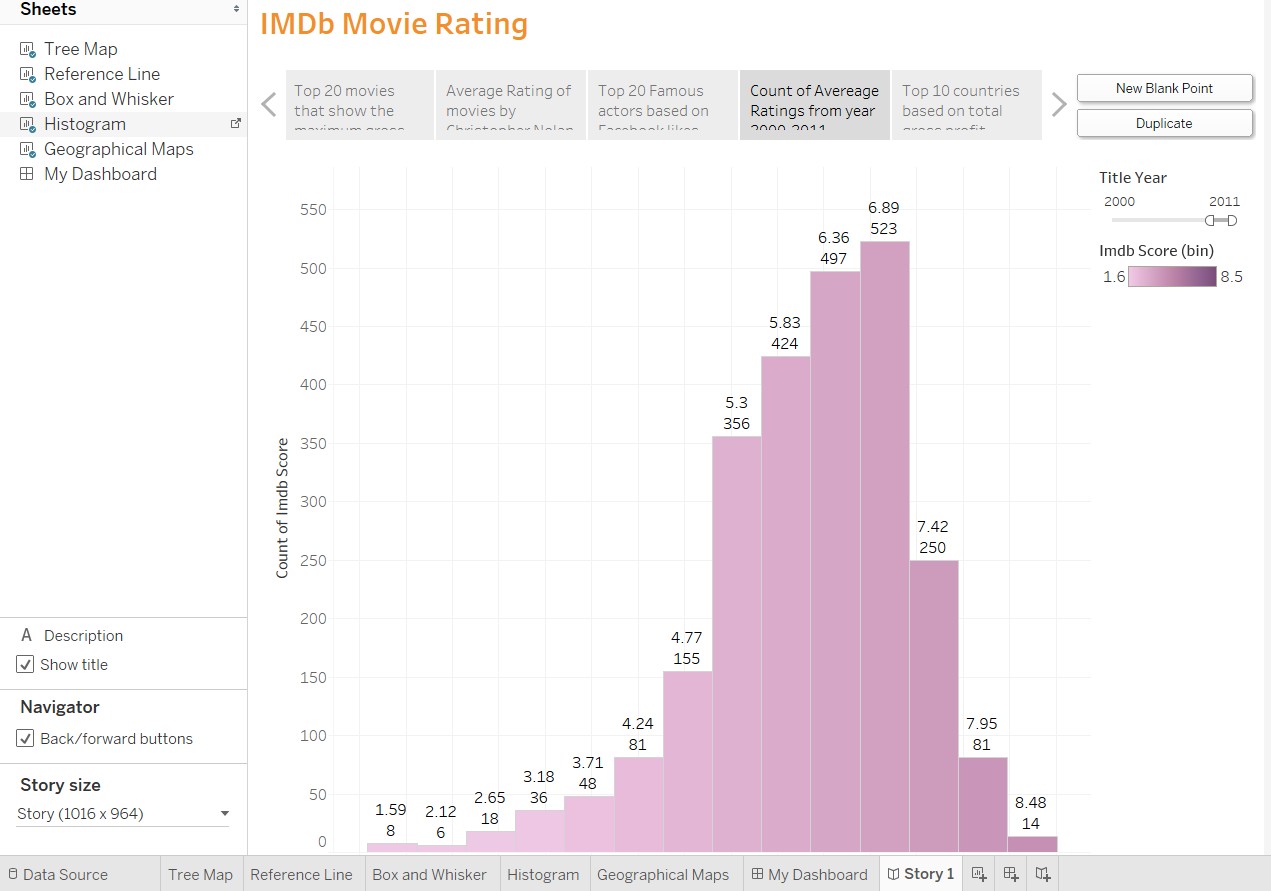
**Top 10 countries based on total gross of the movies**

The countries that have highest total gross of movies is USA with highest gross income as 171,909,339,848 and lowest is Spain with gross income of 229,722,448.



**Count of Average Ratings from year 2000 to 2011**

I have used **Histogram** to depict these counts of ratings from year 2000 to 2011. The total number of 523 movies with an average rating of 6.89 in span of 11 years whereas there are just14 movies with the highest rating of 8.48 and 8 movies with the score of 1.59.



References:

<https://www.forbes.com/sites/dorothypomerantz/2010/12/08/why-movie-ratings-matter-and-why-they-shouldnt/#2c20c56037e0>

<http://www.nytimes.com/2004/07/14/movies/study-finds-film-ratings-are-growing-more-lenient.html?_r=0>