Insights & Data Analysis Report of Movies_Metadata.csv Dataset

By – Harshit Khare

The Dataset Contains a total of **5043 rows** and **28 columns**.

A significant amount of Data is either NULL or NA. To counter this problem, I took help of Python's inbuilt Library "**PANDAS**"

```
Jupyter QtConsole 4.3.1
Python 3.6.3 |Anaconda custom (64-bit)| (default, Oct 15 2017, 03:27:45) [MSC v.1900 64 bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 6.1.0 -- An enhanced Interactive Python. Type '?' for help.

In [1]: import pandas as pd

In [2]: pwd
Out[2]: 'C:\\Users\\HARSHIT'

In [3]: cd Downloads
C:\Users\HARSHIT\Downloads

In [4]: movies = pd.read_csv('movie_metadata.csv')

In [5]: movies.shape
Out[5]: (5043, 28)

In [6]: movies.isnull().sum()
Out[6]:
color                         19
director_name                104
num_critic_for_reviews        50
duration                      15
director_facebook_likes      104
actor_3_facebook_likes        23
actor_2_name                  13
actor_1_facebook_likes         7
gross                        884
genres                         0
actor_1_name                   7
movie_title                    0
num_voted_users                0
cast_total_facebook_likes      0
actor_3_name                  23
facenumber_in_poster          13
plot_keywords                153
movie_imdb_link                0
num_user_for_reviews          21
language                      12
country                        5
content_rating               303
budget                       492
title_year                   108
actor_2_facebook_likes        13
imdb_score                     0
aspect_ratio                 329
movie_facebook_likes           0
dtype: int64

In [7]: moviesnew = movies.fillna(" ")
```

```
In [8]: moviesnew.isnull().sum()
Out[8]:
color                      0
director_name              0
num_critic_for_reviews     0
duration                   0
director_facebook_likes    0
actor_3_facebook_likes     0
actor_2_name               0
actor_1_facebook_likes     0
gross                      0
genres                     0
actor_1_name               0
movie_title                0
num_voted_users            0
cast_total_facebook_likes  0
actor_3_name               0
facenumber_in_poster       0
plot_keywords              0
movie_imdb_link            0
num_user_for_reviews       0
language                   0
country                    0
content_rating             0
budget                     0
title_year                 0
actor_2_facebook_likes     0
imdb_score                 0
aspect_ratio               0
movie_facebook_likes       0
dtype: int64

In [9]: moviesnew.to_csv('moviesnew.csv' , index=False)
```

Now, the Dataset is free from any NULL or NA values & can be used for Data Visualization and Analysis.

To visualize the Dataset, I used a very useful software called **Tableau Public**.

Here are some of the insights I gained after visualizing some of the relations in the Dataset **:**

1. Director and Gross Amount Relation

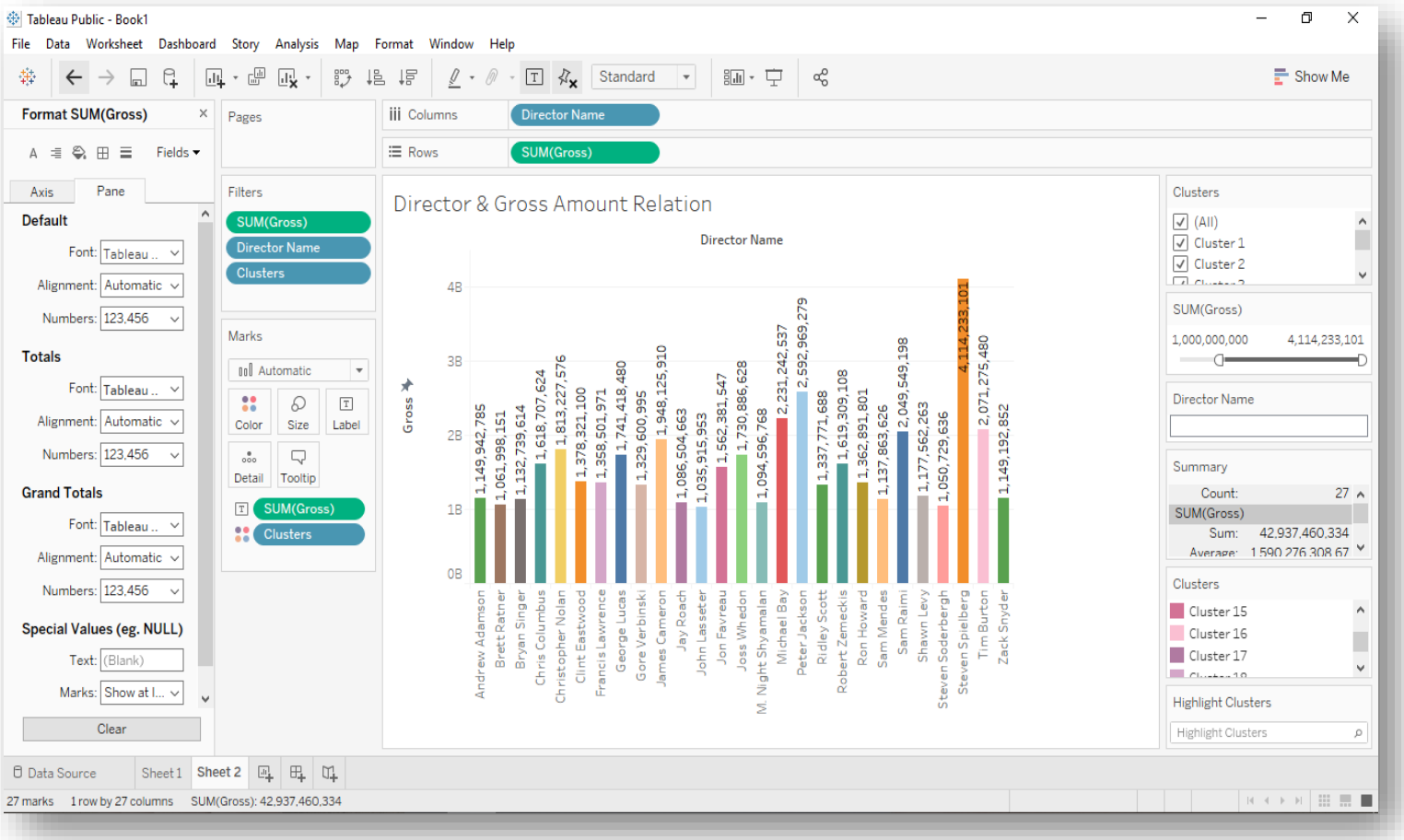
Relation : Director Name and Total Gross Amount.

Visualization By : Column Chart.

Max Gross Amount by : Director Steven Spielberg ($4,114,233,101)

Lowest Gross Amount by : Director Ekachai Uekrongtham ($162)

Insights : Only 27 Directors out of a total of 1879 director have a total gross amount above $1 billion with highest Gross Amount ($4,114,233,101) being of Director Steven Spielberg and Lowest gross Amount ($1,035,915,953) being of Director John Lasseter.



2. Year and Gross Amount Relation

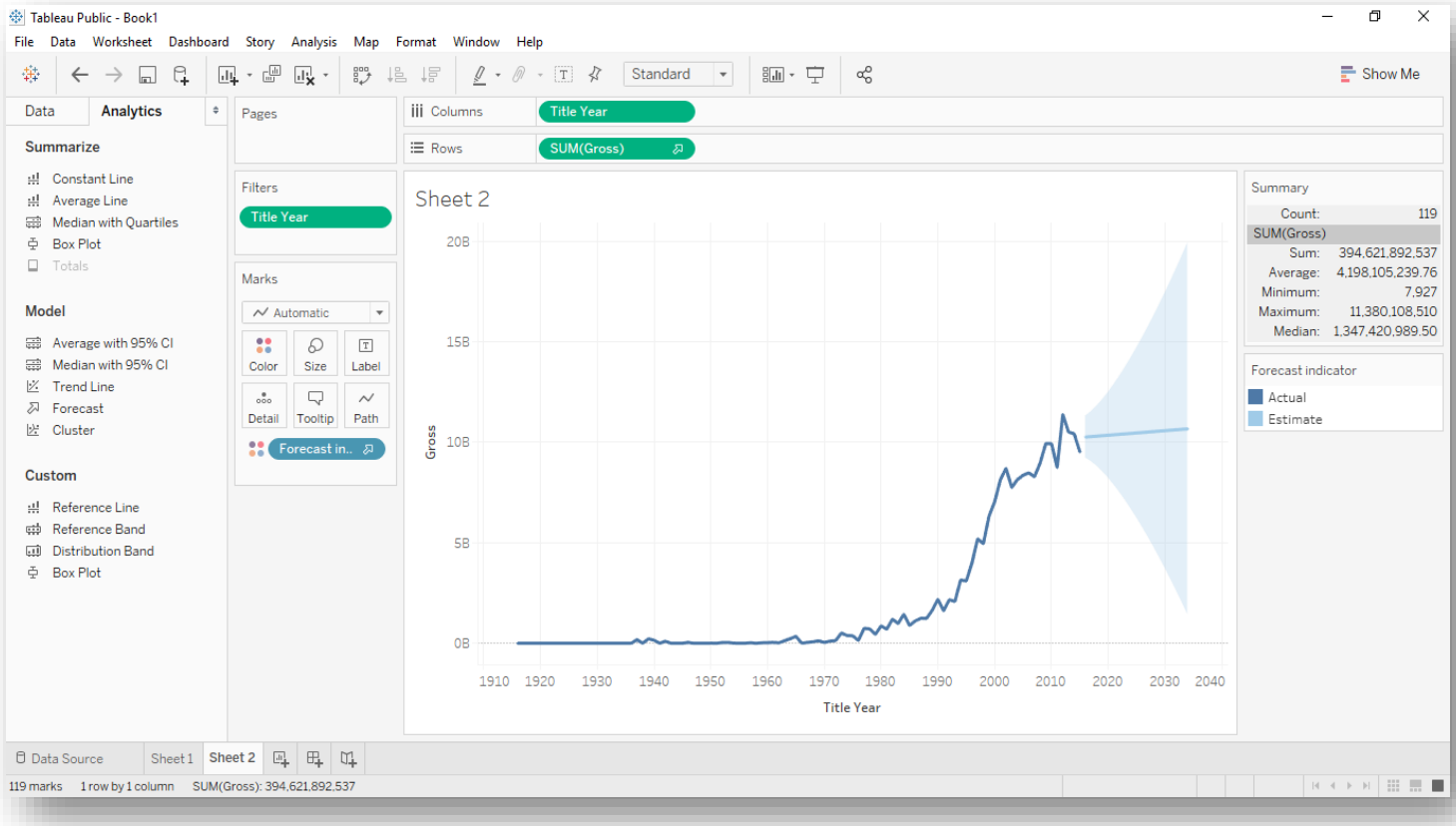
Relation : Year of Release and Total Gross Amount.

Visualization By : Line Chart.

Max Gross Amount in the year : 2012 ($11,380,108,510)

Lowest Gross Amount in the year : 1947 ($7927)

Insights : The Relation Depicts a rise in Total Gross Amount of Movies Released after 1990's with the year 2012 being the Highest Grossing Total for the movies at $11,380,108,510 total gross amount. There is a slight declination after the year 2012 but it does not affect the trend forecast much. The Trend Forecast predicts the gross amount to lie between $10 bn to $11 bn between years 2016 and 2034 with maximum High and Lowest Downfall of $20 bn and $2 bn respectively.



3. Content Rating and IMDb Score Relation

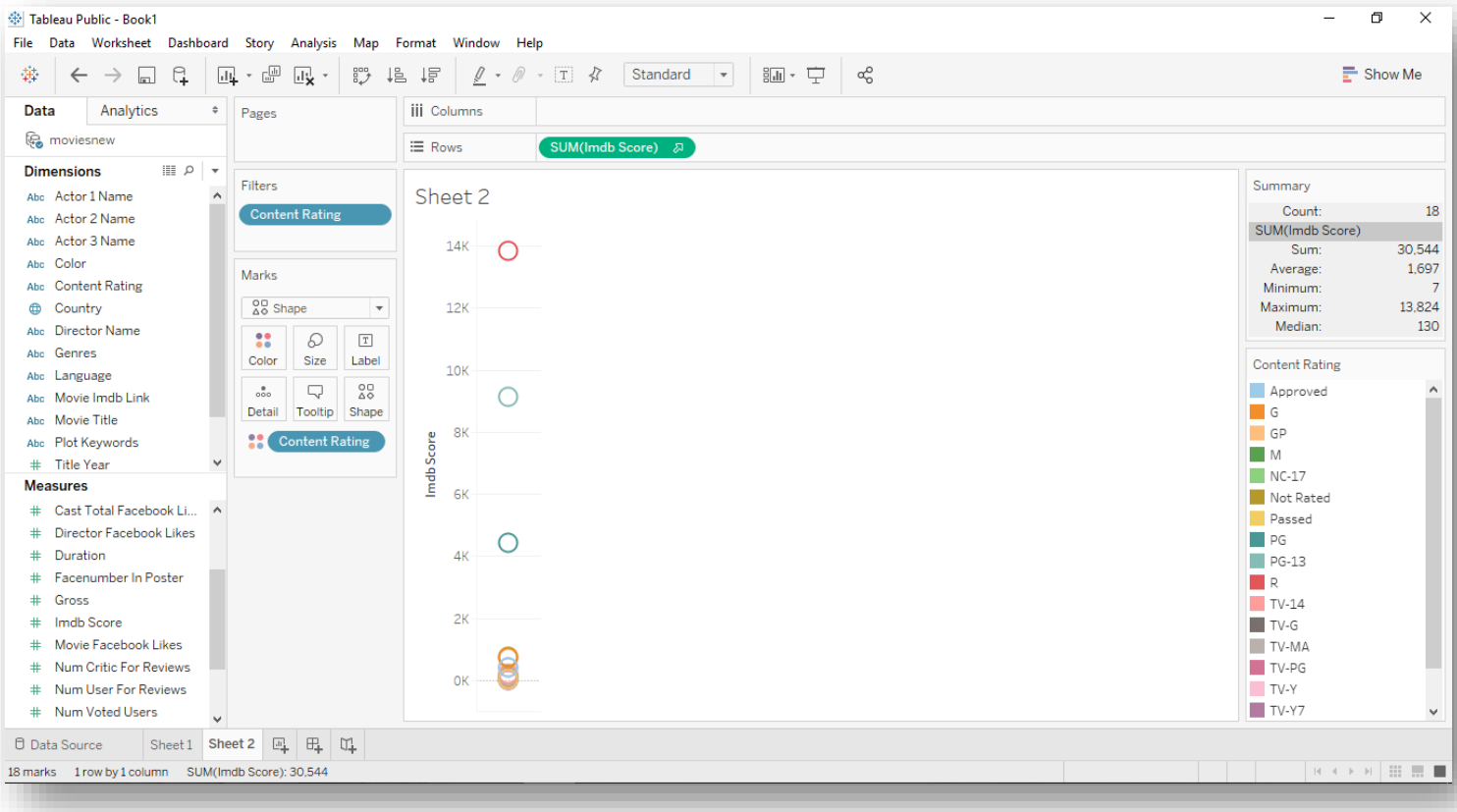
Relation : Content Rating and IMDb Score (Sum)

Visualization By : Circle Views.

Max IMDb Score is of the Rating : R (13824)

Lowest IMDb Score is of the Rating : TV-Y7 (7)

Insights : The Relation Depicts that most of the content ratings have a low IMDb score (Sum) between 7 to 731. The 3 Topmost IMDb score is of the rating R with a score of 13824, followed by PG-13 with a score of 9142 and lastly the rating PG with a score of 4412.



4. Facenumber in Poster and Gross Amount Relation

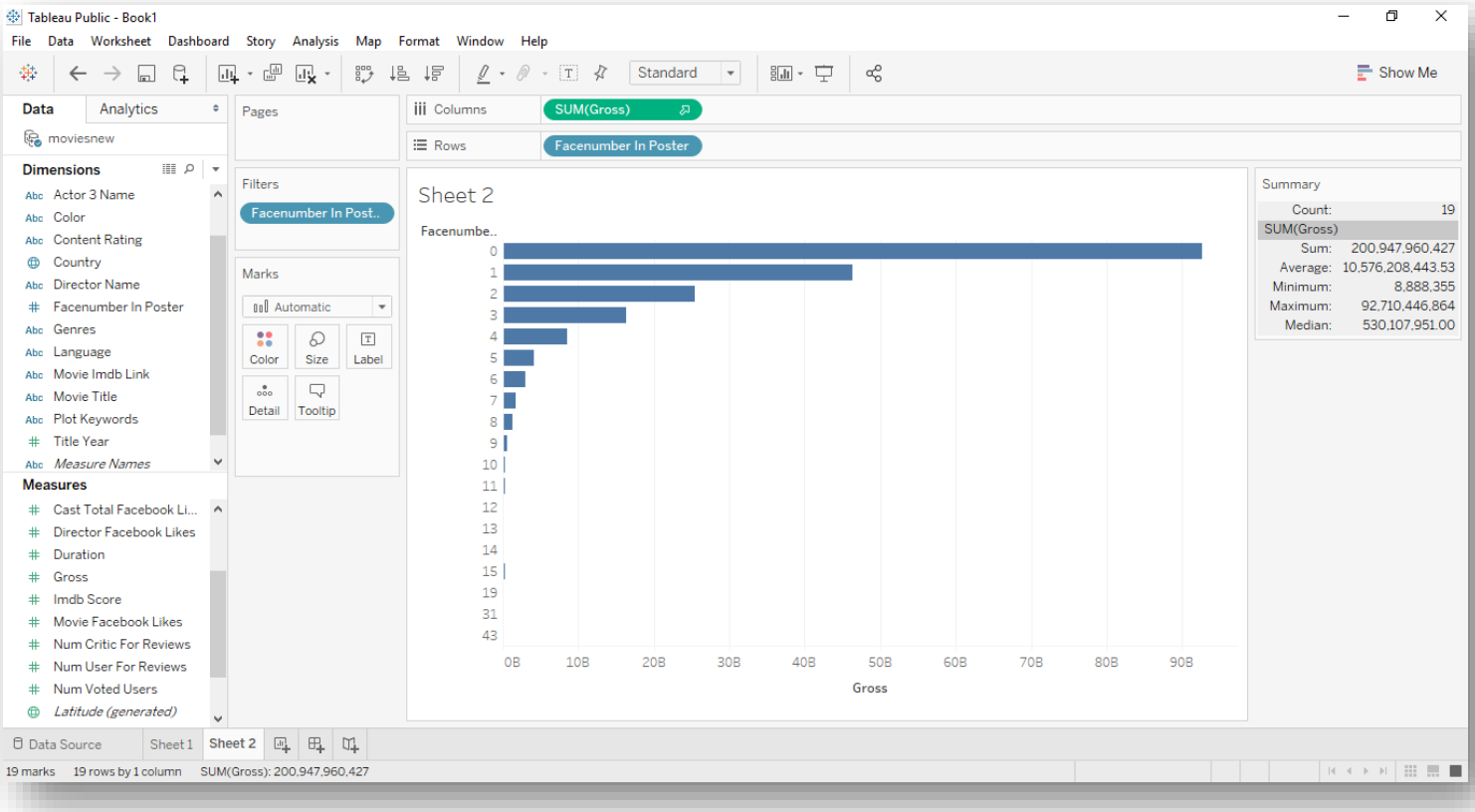
Relation : Facenumber in Poster and Total Gross Amount

Visualization By : Bar Graph.

Max Gross Amount is of Posters having Facenumber : 0 ($92,710,446,864)

Lowest Gross Amount is of Posters having Facenumber: 19 ($8,888,355)

Insights : The Relation Depicts that Movies whose posters have no faces are having the highest total gross amount. The general trend is the total gross amount keeps on decreasing as the number of faces in the poster increases. Valuable insight which can be gained from this relation is that the movie with less number of faces on its poster is supposed to have a high total gross amount.



5. Movie Title and Number of records Relation

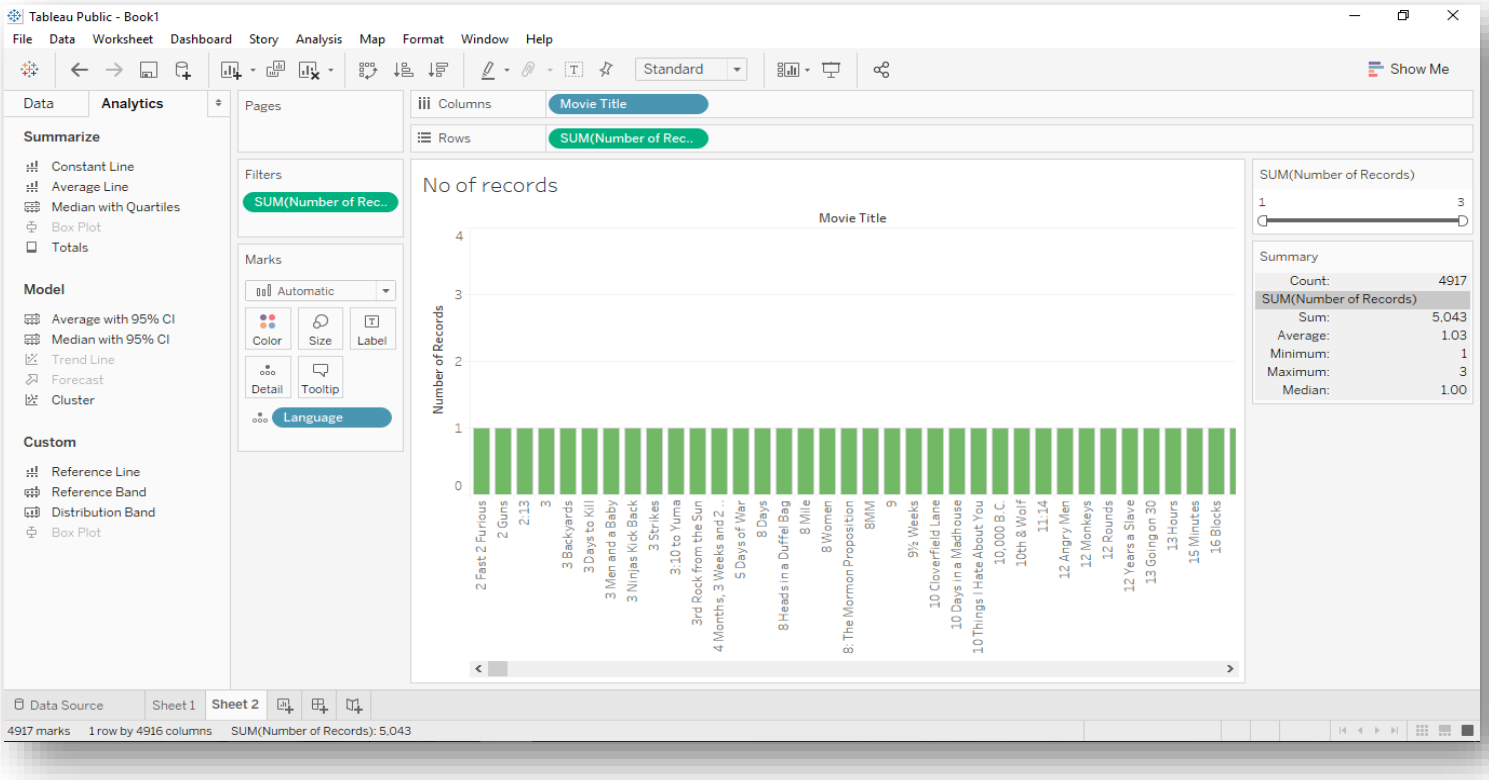
Relation : Movie Title and Number of Records (Calculated Field)

Visualization By : Column Chart.

Max Records won by Movies : 3

Min Record won by Movies : 1

Insights 1: All the Movies have won atleast 1 award. A total of 4917 movies have won atleast 1 Award.



Insights 2: A total of 119 movies have won atleast 2 Awards. A significant drop in number of movies when number of awards won is 2.

Insights 3: Only 7 movies have won 3 Awards. These Movies include Ben-Hur, Halloween. Home, King-Kong, Pan, The Fast & The Furious and Victor Frankenstein. This is a very Valuable Insight as it depicts that though there were 4917 Movies that have won atleast 1 award but only 7 have managed to win 3 awards

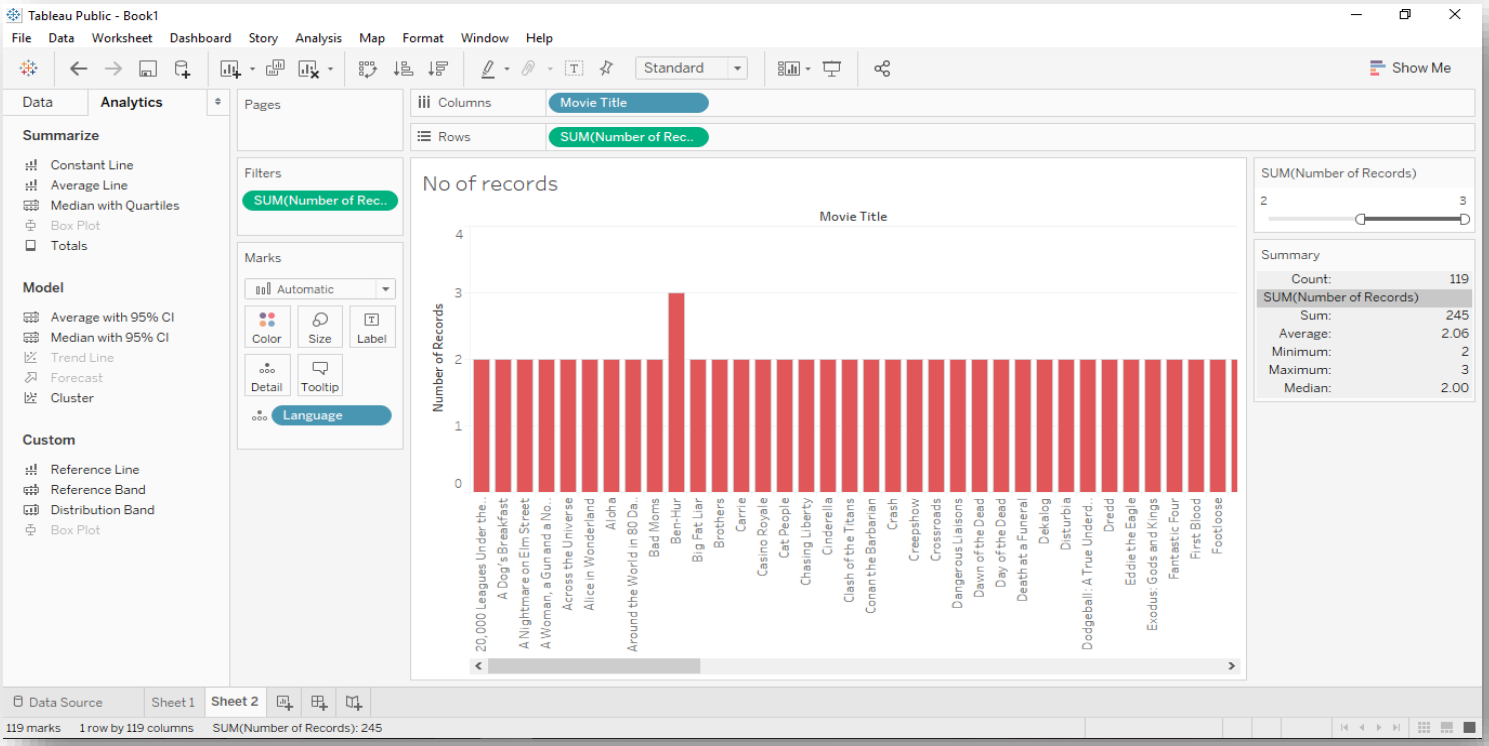


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Window Help
Standard
Show Me
Data Analytics
Summarize
Constant Line
Average Line
Median with Quartiles
Box Plot
Totals
Model
Average with 95% CI
Median with 95% CI
Trend Line
Forecast
Cluster
Custom
Reference Line
Reference Band
Distribution Band
Box Plot
Pages
Filters
SUM(Number of Rec..
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Language
Columns
Movie Title
Rows
SUM(Number of Rec..
No of records
Movie Title
Number of Records
20,000 Leagues Under the..
A Dog's Breakfast
A Nightmare on Elm Street
A Woman, a Gun and a No..
Across the Universe
Alice in Wonderland
Aloha
Around the World in 80 Da..
Bad Moms
Ben-Hur
Big Fat Liar
Brothers
Carrie
Casino Royale
Cat People
Chasing Liberty
Cinderella
Clash of the Titans
Conan the Barbarian
Crash
Creepshow
Crossroads
Dangerous Liaisons
Dawn of the Dead
Day of the Dead
Death at a Funeral
Dekalog
Disturbia
Dodgeball: A True Underd..
Dredd
Eddie the Eagle
Exodus: Gods and Kings
Fantastic Four
First Blood
Footloose
SUM(Number of Records)
Summary
Count: 119
SUM(Number of Records)
Sum: 245
Average: 2.06
Minimum: 2
Maximum: 3
Median: 2.00
Data Source
Sheet 1
Sheet 2
119 marks
1 row by 119 columns
SUM(Number of Records): 245

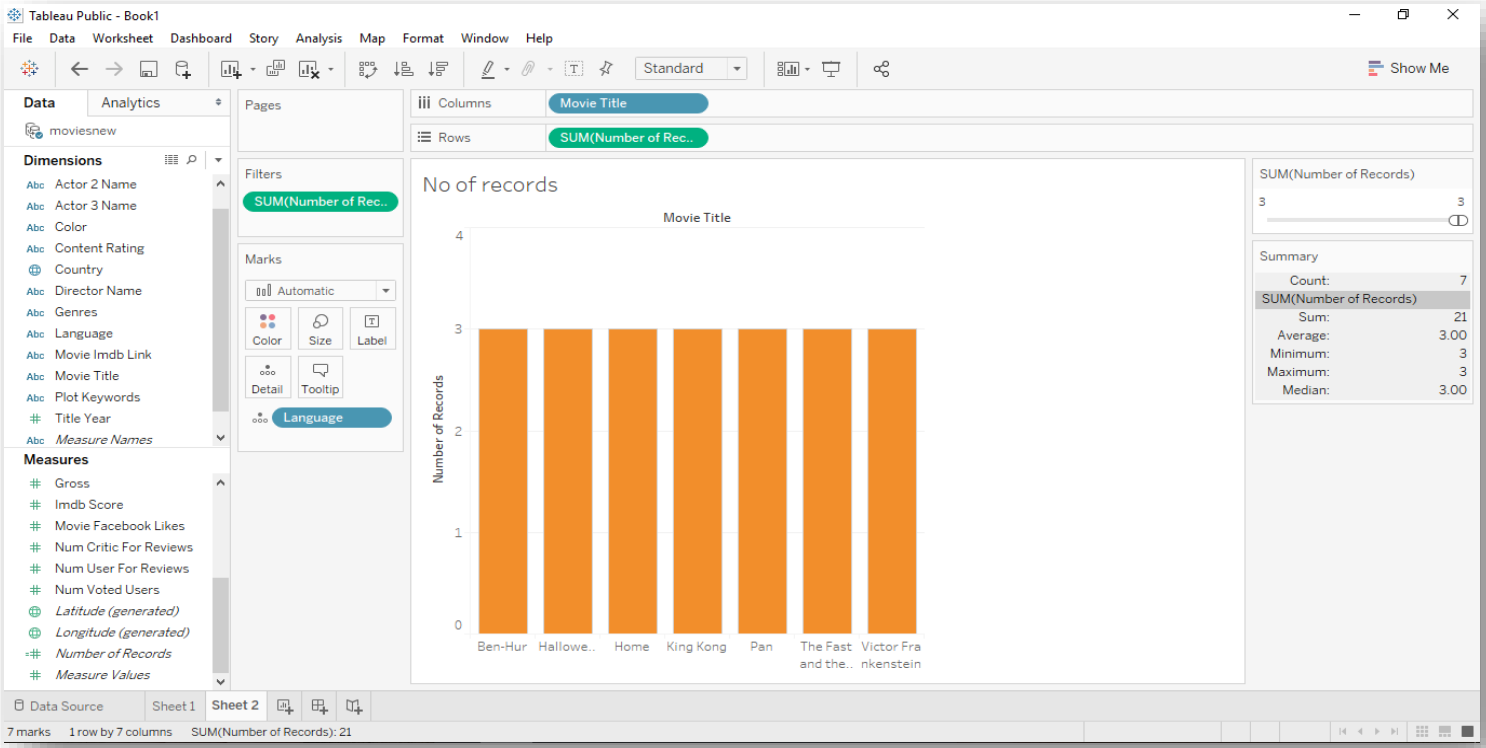


Tableau Public - Book1
File Data Worksheet Dashboard Story Analysis Map Format Window Help
Standard
Show Me
Data Analytics
moviesnew
Dimensions
Actor 2 Name
Actor 3 Name
Color
Content Rating
Country
Director Name
Genres
Language
Movie Imdb Link
Movie Title
Plot Keywords
Title Year
Measure Names
Measures
Gross
Imdb Score
Movie Facebook Likes
Num Critic For Reviews
Num User For Reviews
Num Voted Users
Latitude (generated)
Longitude (generated)
Number of Records
Measure Values
Pages
Filters
SUM(Number of Rec..
Marks
Automatic
Color
Size
Label
Detail
Tooltip
Language
Columns
Movie Title
Rows
SUM(Number of Rec..
No of records
Movie Title
Number of Records
Ben-Hur
Hallowe..
Home
King Kong
Pan
The Fast and the..
Victor Frankenstein
SUM(Number of Records)
Summary
Count: 7
SUM(Number of Records)
Sum: 21
Average: 3.00
Minimum: 3
Maximum: 3
Median: 3.00
Data Source
Sheet 1
Sheet 2
7 marks
1 row by 7 columns
SUM(Number of Records): 21

•

6. Budget and Gross Amount Relation

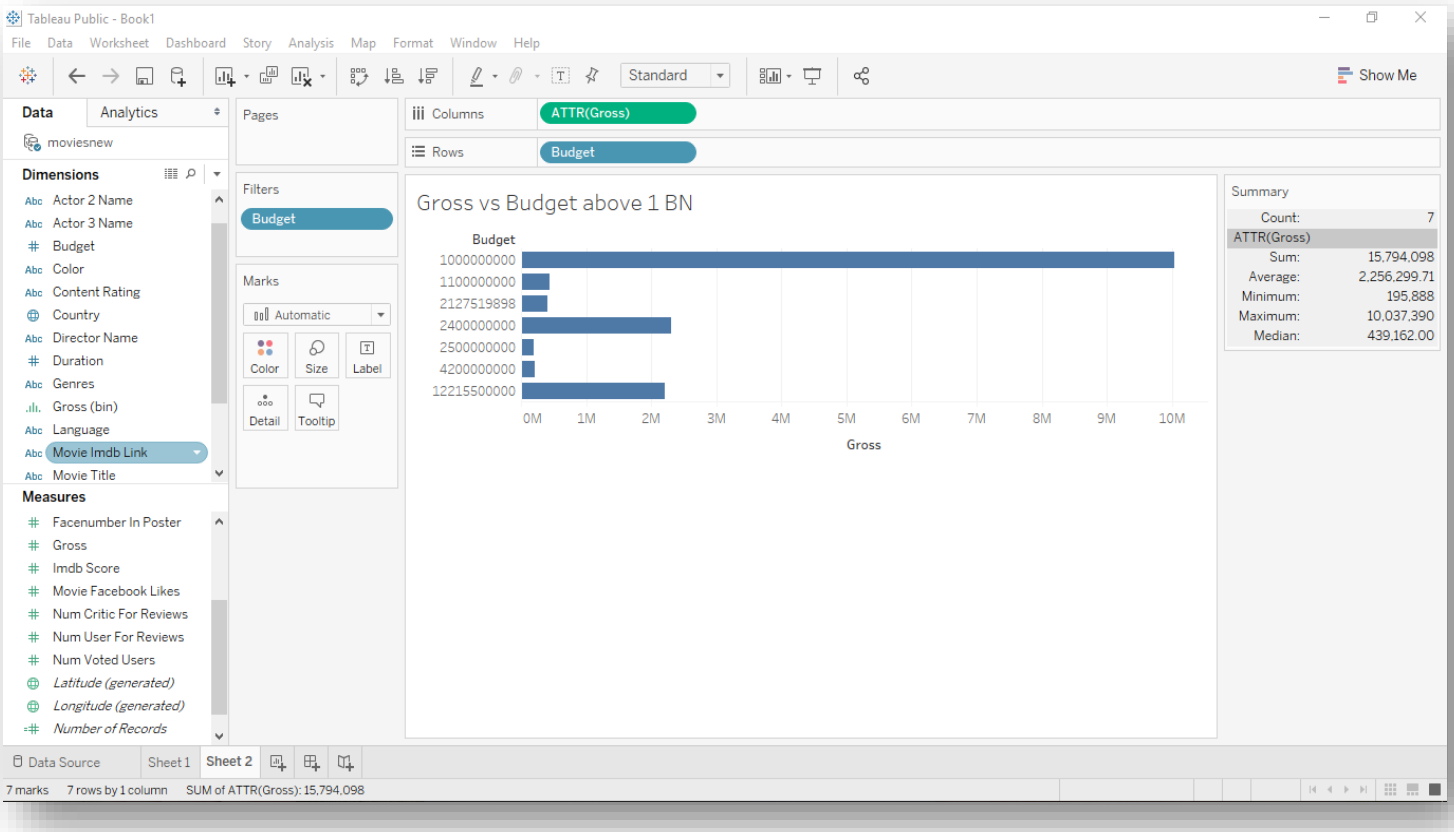
Relation : Budget of Movies (Above $1 bn) and Total Gross Amount

Visualization By : Bar Graph.

Max Gross Amount is of the movie having budget : $1,000,000,000

Lowest Gross Amount is of the movie having budget: $2,500,000,000

Insights : The Relation Depicts that Movies that have budget equal to $1 billion are having the highest gross amount. The Movies having budget $2,400,000,000 and $12,215,500,000 have high gross amount too but still very less than movies having budget $1,000,000,000.



7. IMDb Score and Gross Amount Relation

Relation : IMDb Score and Total Gross Amount

Visualization By : Column Chart.

Max Gross Amount is of IMDb Score : 6.7 ($9,981,888,384)

Lowest Gross Amount is of IMDb Score : 9.1 ($894,186)

Insights : The Relation Depicts that Movies that have IMDb Score between 5.4 and 8.1 are having a higher total gross amount as compared to movies having other scores. A valuable insight which can be gained is that though the movies having a low IMDb Score do not have a high total gross amount, the movies with a very high rating also do not have high total gross amount.