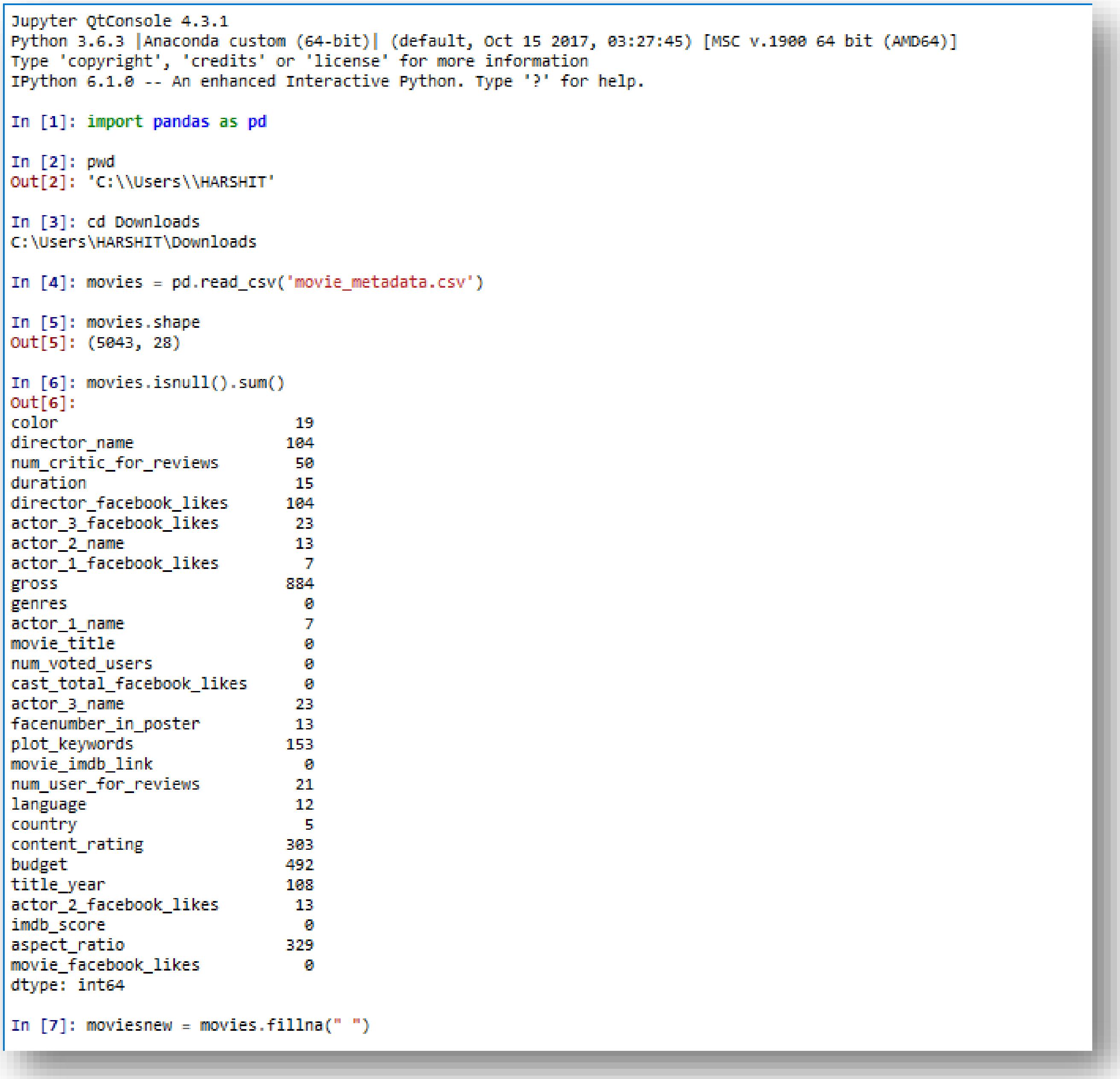
# Insights & Data Analysis Report of Movies_Metadata.csv Dataset

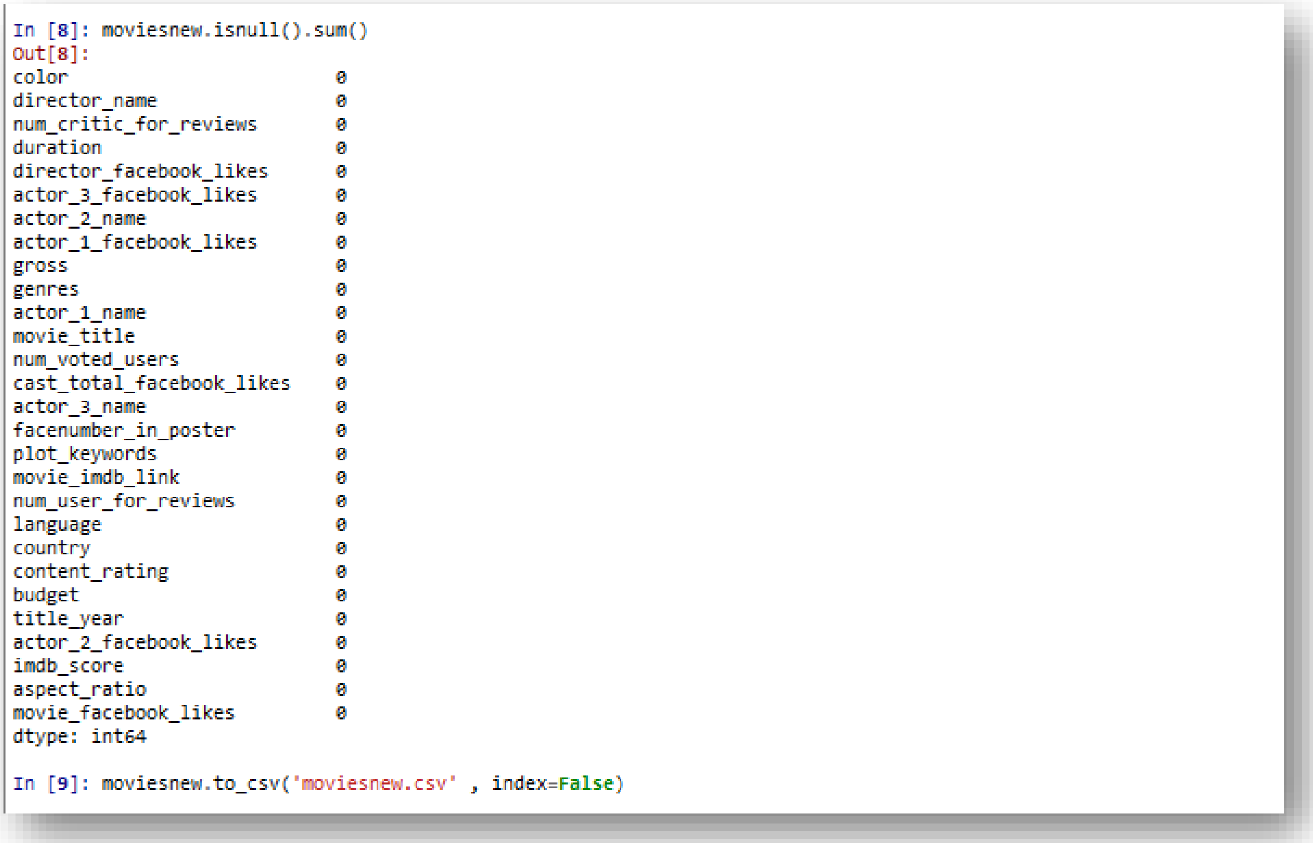
*By – Harshit Khare*

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 on 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 directors 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.

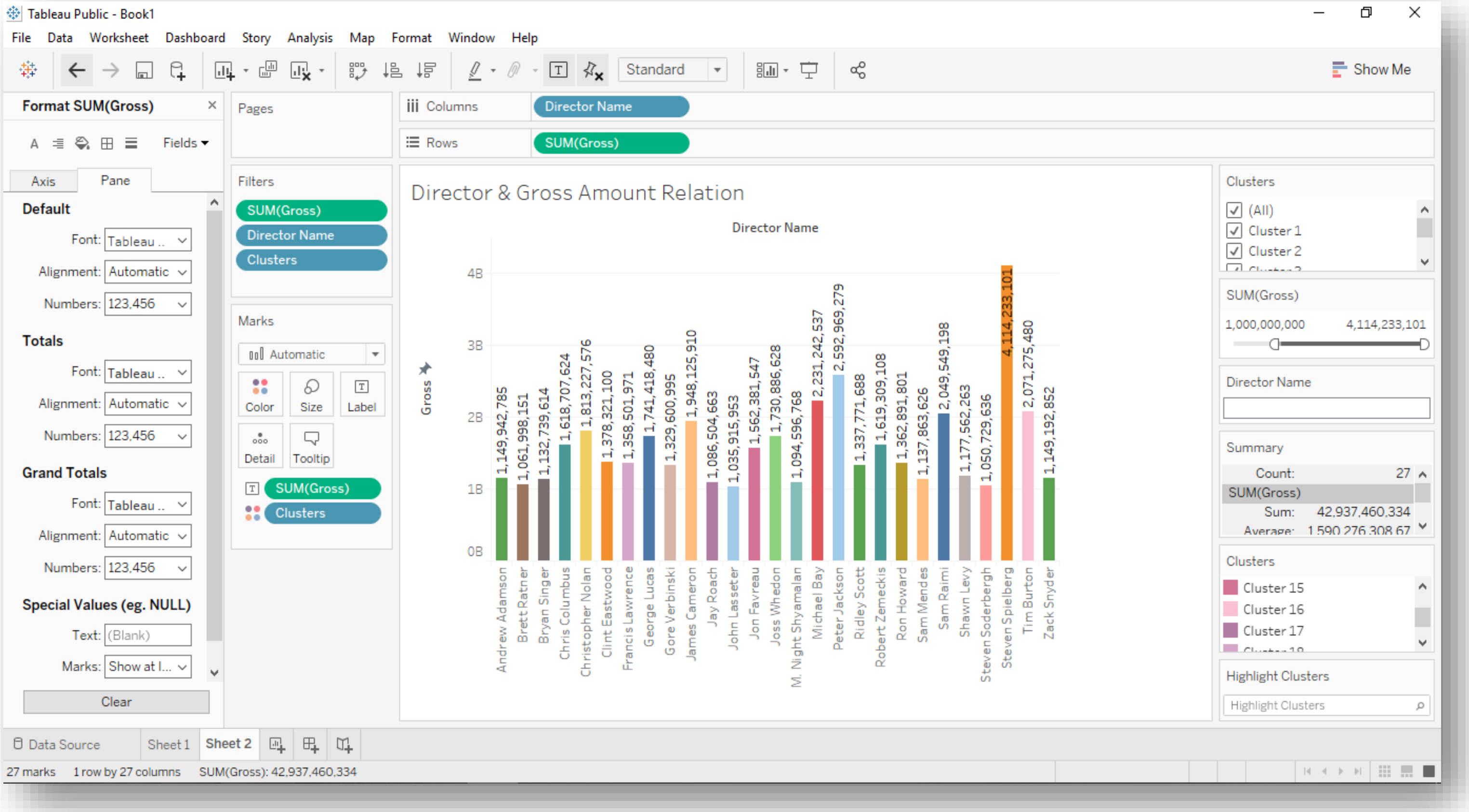


Tableau Public - Book1
File
Data
Worksheet
Dashboard
Story
Analysis
Map
Format
Window
Help
Standard
Show Me
Format SUM(Gross)
Fields
Axis
Pane
Pages
Filters
Marks
Color
Size
Label
Detail
Tooltip
Clear
Director & Gross Amount Relation
Director Name
Gross
4B
3B
2B
1B
0B
Andrew Adamson 1,149,942,785
Brett Ratner 1,061,998,151
Bryan Singer 1,132,739,614
Chris Columbus 1,618,707,624
Christopher Nolan 1,813,227,576
Clint Eastwood 1,378,321,100
Francis Lawrence 1,358,501,971
George Lucas 1,741,418,480
Gore Verbinski 1,329,600,995
James Cameron 1,948,125,910
Jay Roach 1,086,504,663
John Lasseter 1,035,915,953
Jon Favreau 1,562,381,547
Joss Whedon 1,730,886,628
M. Night Shyamalan 1,094,596,768
Michael Bay 2,231,242,537
Peter Jackson 2,592,969,279
Ridley Scott 1,337,771,688
Robert Zemeckis 1,619,309,108
Ron Howard 1,362,891,801
Sam Mendes 1,137,863,626
Sam Raimi 2,049,549,198
Shawn Levy 1,177,562,263
Steven Soderbergh 1,050,729,636
Steven Spielberg 4,114,233,101
Tim Burton 2,071,275,480
Zack Snyder 1,149,192,852
Highlight Clusters
Data Source
Sheet 1
Sheet 2

Default
Font: Tableau ..
Alignment: Automatic
Numbers: 123,456

Totals
Font: Tableau ..
Alignment: Automatic
Numbers: 123,456

Grand Totals
Font: Tableau ..
Alignment: Automatic
Numbers: 123,456

Special Values (eg. NULL)
Text: (Blank)
Marks: Show at l...

Columns: Director Name
Rows: SUM(Gross)

Filters: SUM(Gross), Director Name, Clusters

Marks: Automatic
SUM(Gross)
Clusters

Clusters
☑ (All)
☑ Cluster 1
☑ Cluster 2

SUM(Gross)
1,000,000,000 4,114,233,101

Director Name

Summary
Count: 27
SUM(Gross)
Sum: 42,937,460,334
Average: 1,590,276,308.67

Clusters
Cluster 15
Cluster 16
Cluster 17

27 marks 1 row by 27 columns SUM(Gross): 42,937,460,334

2. Year and Gross Amount Relation

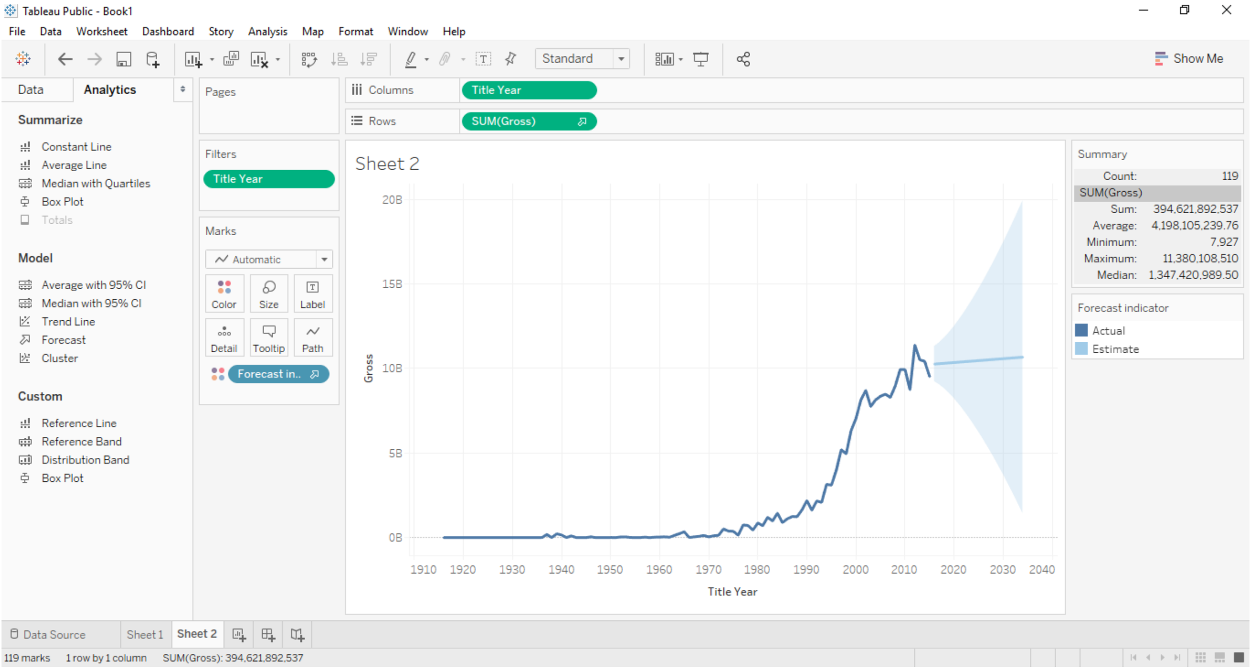
Relation : Year of Release and Total Gross Amount.

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 2040 with maximum High and Lowest Downfall of $20 bn and $2 bn respectively.



Columns: Title Year

Rows: SUM(Gross)

Summary

| | |
|---|---|
| Count: | 119 |
| SUM(Gross) | |
| Sum: | 394,621,892,537 |
| Average: | 4,198,105,239.76 |
| Minimum: | 7,927 |
| Maximum: | 11,380,108,510 |
| Median: | 1,347,420,989.50 |

119 marks 1 row by 1 column SUM(Gross): 394,621,892,537

3. Content Rating and IMDb Score Relation

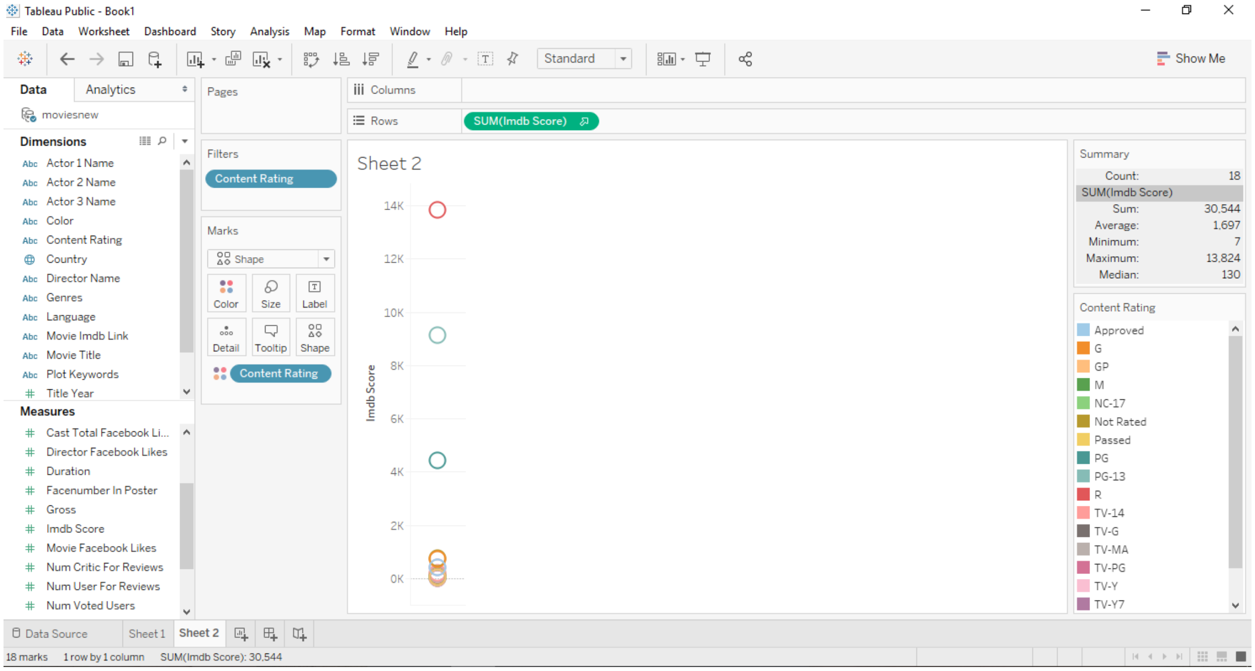
Relation : Content Rating and IMDb Score

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 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.



Columns

Rows: SUM(Imdb Score)

Filters: Content Rating

Marks: Shape

Shape: Content Rating

Summary

| | |
|---|---|
| Count: | 18 |
| SUM(Imdb Score) | |
| Sum: | 30,544 |
| Average: | 1,697 |
| Minimum: | 7 |
| Maximum: | 13,824 |
| Median: | 130 |

Content Rating

- Approved
- G
- GP
- M
- NC-17
- Not Rated
- Passed
- PG
- PG-13
- R
- TV-14
- TV-G
- TV-MA
- TV-PG
- TV-Y
- TV-Y7

18 marks 1 row by 1 column SUM(Imdb Score): 30,544

4. Facenumber in Poster and Gross Amount Relation

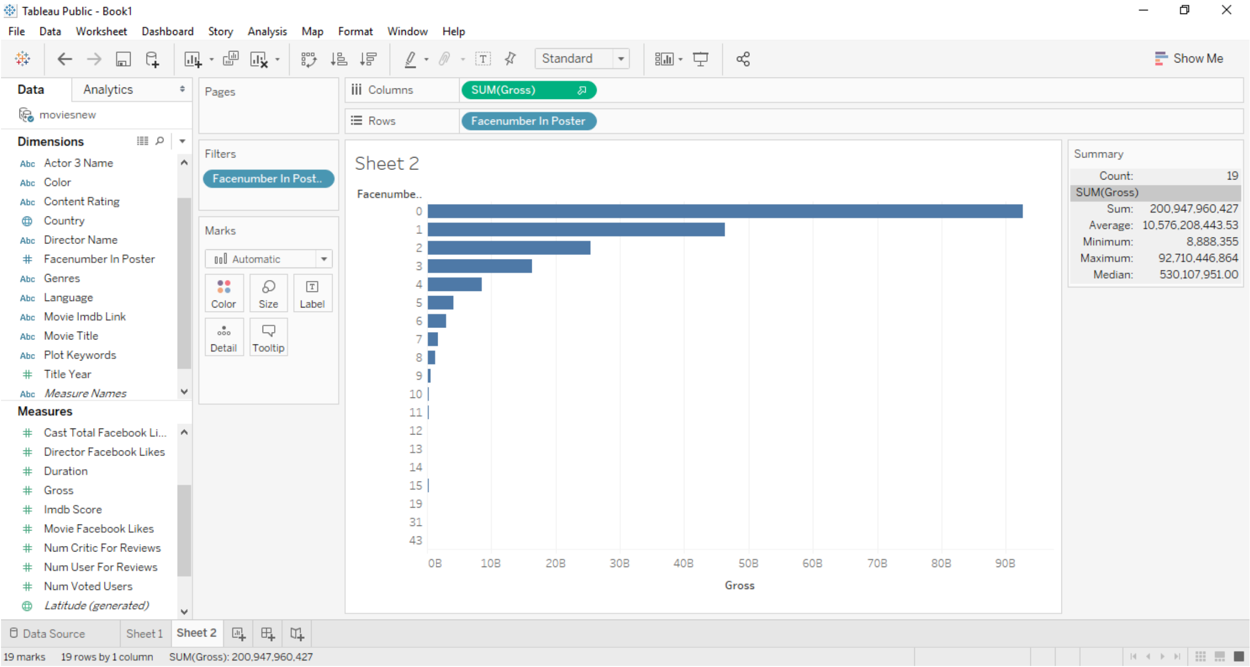
Relation : Facenumber in Poster and Total Gross Amount

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.



Columns: SUM(Gross)

Rows: Facenumber In Poster

Filters: Facenumber In Post..

Summary

| | |
|---|---|
| Count: | 19 |
| SUM(Gross) | |
| Sum: | 200,947,960,427 |
| Average: | 10,576,208,443.53 |
| Minimum: | 8,888,355 |
| Maximum: | 92,710,446,864 |
| Median: | 530,107,951.00 |

19 marks 19 rows by 1 column SUM(Gross): 200,947,960,427

5. Movie Title and Number of records Relation

Relation : Movie Title and Number of Records

Visualization By : Column Chart.

Max Records won by Movies : 3

Min Record won by Movies : 1

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.

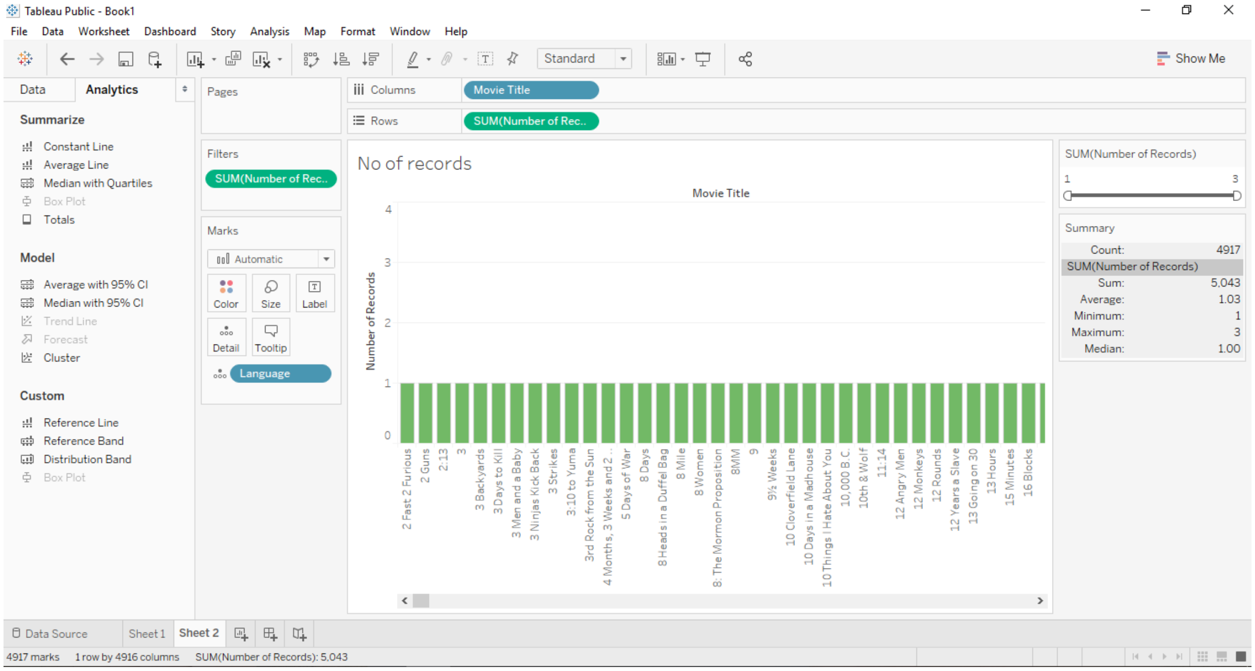


Tableau Public - Book1

Columns: Movie Title

Rows: SUM(Number of Rec..

Filters: SUM(Number of Rec..

Marks: Automatic

Language

SUM(Number of Records)

1 3

Summary

|  |  |
|---|---|
| Count: | 4917 |
| SUM(Number of Records) |  |
| Sum: | 5,043 |
| Average: | 1.03 |
| Minimum: | 1 |
| Maximum: | 3 |
| Median: | 1.00 |

4917 marks 1 row by 4916 columns SUM(Number of Records): 5,043